EXPLORING NEW CHANNEL MATERIALS FOR NANOSCALE CMOS DEVICES: A SIMULATION APPROACH

A Thesis

Submitted to the Faculty

of

Purdue University

by

Anisur Rahman

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2005

To my beloved family.

ACKNOWLEDGMENTS

I would like to express my profound gratitude and deep appreciation to my PhD co-adviser, Professor Mark S. Lundstrom, who supervised my overall research and guided me throughout the course of my PhD. He also arrange funding for this research for the entire period of six years I stayed at Purdue. Without his help and guidance this research would not be possible. I would also like to thank my PhD co-adviser Professor Gerhard Klimeck for helping me to understand the tight-binding approach and appreciate its potentials. I am grateful to my Doctoral Committee member Professor Supriyo Datta for offering his extraordinary insight, which expedited the solution of some of the crucial problems of my research. I also thank Prof. Ron Reifenberger and Dr. Avik Ghosh for being in my committee and offering guidance. Thanks to my colleagues at Purdue, Dr. Zhibin Ren, Dr. Ramesh Venugopal, Dr. Jung-Hoon Rhew, Dr. Jing Guo, Dr. Jing Wang, Sayed Hasan Palash, Neophytos Neophytou, Siyuranga Koswatta, and Dr. Diego Kienle for lively discussions, insightful comments and honest opinions on my research. Thanks to Sayeef Salahuddin for helping me with the typesetting and formatting this thesis, without his help it would have been challenging to finish the thesis on time.

My heartfelt thanks to my parents, especially my mother Prof. Kaniz Maula, who supervised my home schooling during the first six academic years of my life, which served as the foundation for my education since then. My special gratitude to Professor Anisul Haque (Dept. of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology) for introducing me to the field of semiconductor device physics and their numerical simulations. Finally, my special thanks to my wife, Fariha Karim, who always stayed by my side, offered her support during the hard time with research, and did everything to uphold my spirit whenever I was down.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

ABSTRACT xii

1 Introduction 1

1.1 CMOS Scaling: From Long Channel to Nanometer Scale Devices . . . 1

1.2 Novel Channel Material CMOS: Experimental Exploration 3

1.3 Simulation Approach: Issues and Challenges 4

1.4 Outline of The Thesis 7

2 A GENERALIZED EFFECTIVE MASS APPROACH FOR N-MOSFETS 11

2.1 Introduction 11

2.2 Summary of the Overall Solution Procedure 14

2.3 Conduction-band Structure in cubic semiconductors 15

2.4 The Generalized Effective-Mass Equation and the Solution 19

2.4.1 The Quantum Confinement Problem 21

2.4.2 The Transport Problem 23

2.4.3 Simplifications in UTB Devices 25

2.5 Application to Ultra-thin-body Si and Ge MOSFETs 28

2.5.1 Transformation Matrices for the DCS 28

2.5.2 Transformation matrices for the ECS 30

2.5.3 Evaluating Effective Masses 33

2.6 Discussion 34

2.7 Summary and Conclusion 37

3 ASSESSMENT OF GERMANIUM N-MOSFETS BY NEGF SIMULATION 38

3.1 Introduction 38

Page

3.2 Approach . . . 40
3.3 Results . . . 42
3.4 Discussion . . . 46
3.5 Summary and Conclusion . . . 48
4 A TOP OF THE BARRIER BALLISTIC MODEL . . . 51
4.1 Introduction . . . 51
4.2 Device Physics of Ballistic MOSFETs . . . 52
4.3 Model . . . 57
4.4 Application to Ballistic MOSFETs . . . 61
4.4.1 Parameters for the Analytical Model . . . 61
4.4.2 Treatment of the Floating Source Potential . . . 63
4.4.3 Charge and Velocity at the Top of the Barrier . . . 65
4.5 Discussion . . . 66
4.6 Summary and Conclusion . . . 67
5 TIGHT-BINDING APPROACH TO SIMULATE NANOSCLAE MOSFETS 68
5.1 Introduction . . . 69
5.2 Semi-empirical Tight-binding Model . . . 69
5.2.1 Assumptions Leading to A Semi-empirical Model . . . 69
5.2.2 Choice of Basis . . . 72
5.3 The 3-D Tight-binding Hamiltonian in $spds^*$ Basis . . . 74
5.4 The Treatment of Strain . . . 74
5.5 Application to Finite Dimensional Systems . . . 76
5.6 Summary . . . 76
6 APPLICATION OF TIGHT-BINDING APPROACH TO UTB DG MOSFETS . . . 77
6.1 Overall Simulation Approach: Tight-binding without Band Bending . 77
6.2 Application: On the Validity of Parabolic $E(\vec{k})$ in Nanoscale Ge n-FETs and Effects of Strain Orientation in Si p-FETs . . . 79
6.2.1 Approach . . . 80

Page

6.2.2 Results and Discussions 83
6.2.3 Summary and Conclusion 89
6.3 Application: Comparing Si, Ge, GaAs and InAs as Nanoscale CMOS Channel Materials 91
6.3.1 Approach 92
6.3.2 Results and Discussion 95
6.3.3 Summary and Conclusion 102
7 APPLICATION OF TIGHT-BINDING APPROACH WITH SELF-CONSISTENT ELECTROSTATICS 103
7.1 Overall Simulation Approach: Self-consistent Electrostatics 104
7.2 Application: Mobility vs. Ballisticity in Strained Bulk single-gate COMS Devices 106
7.2.1 Background 106
7.2.2 Approach 108
7.2.3 Results and Discussion 109
7.2.4 Summary and Conclusion 112
7.3 Application: C-V Characteristics and Hole Density Profile in Strained HOI p-MOSFETs 112
7.3.1 Approach 114
7.3.2 Results and Discussion 114
7.3.3 Summary and Conclusion 118
8 SUMMARY AND FUTURE WORK 119
8.1 Summary 119
8.2 Future Work 119
LIST OF REFERENCES 122
A Solving Generalized Effective Mass Equation 132
A.1 Separable Potential 132
A.1.1 Quantum Confinement Problem 132
A.1.2 The Transport Problem 135

Page

A.2 General Potential 136
A.2.1 Quantum Confinement Problem 137
B The Top of the Barrier Ballistic Model 139
B.1 Treating Floating Boundary Condition 139
B.2 Treating Arbitrary Bandstructure 142
C The Tight-binding Hamiltonian with Strain 146
C.1 Derivation of Tight-binding Hamiltonian for Bulk Materials 146
C.2 Overlap Mtrix Elements from Slater-Koster Two-Center-Integrals . . 154
C.3 Treating Spin-orbit Coupling in Tight-binding Formalism 160
C.4 Application to Finite Dimensional Structures: UTB MOSFETs 164
C.5 Passivation of Surface States in Tight-Binding Approach 169
C.5.1 Anion Atoms at Top and Bottom Surface 169
C.5.2 Cation Atoms at Top and Bottom Surface 171
VITA 175

LIST OF TABLES

Table Page

2.1 Effective masses for (100),(111) and (110) wafers 35

2.2 Δ and Λ valley masses 36

3.1 Device parameters for the optimally designed Ge n-MOSFET 49

3.2 Performance projection for a ballistic Ge n-MOSFET. 50

LIST OF FIGURES

Figure		Page
2.1	Three orthogonal coordinate systems.	17
2.2	Model device: UTB DG MOSFET	20
2.3	CB constant energy ellipsoids along Δ.	31
2.4	CB constant energy ellipsoids along Λ.	32
2.5	Effects of two unitary transformations.	36
3.1	The ultra-thin-body, double-gate device structure	39
3.2	Three orthogonal coordinate systems, device, crystal and ellipse	40
3.3	Doping profile along channel	41
3.4	Comparing ballistic NEGF $I_D - V_G$ for Si and Ge	42
3.5	Comparing ballistic NEGF $I_D - V_D$ for Si and Ge	43
3.6	Device design optimization technique.	44
3.7	Ballistic NEGF $I_D - V_G$ for the nominal device	45
3.8	Ballistic NEGF $I_D - V_D$ for the nominal device	46
3.9	Effects of process variation.	47
3.10	Scattering Simulation for the Ge n-MOSFET.	48
4.1	The model device	52
4.2	Ballistic distribution function along channel	53
4.3	The energy resolved DOS and electron density: NEGF results	54
4.4	Energy band diagrams under high and low drain bias	54
4.5	Charge and injection velocity at top of the barrier: 2D simulation	55
4.6	Filling of the states at the top of the barrier	56
4.7	The 2D circuit model for ballistic transistors.	58
4.8	Analytical model without floating boundary condition vs. nanoMOS results	63
4.9	Analytical model with floating boundary condition vs. nanoMOS results	64

Figure Page

4.10 Charge and injection velocity behavior from the analytical model 65

6.1 Non Self-consistent Tight-binding Scheme 78

6.2 Band Bending Along Thickness is Negligible and Does Not Affect the Bandstructure 79

6.3 The UTB DG MOSFET fabricated on (100) wafer and the underlying atomic arrangement 80

6.4 The Semi-Numerical Ballistic Model Treating Arbitrary Bandstructures . 81

6.5 CB constant energy surfaces of bulk Ge 82

6.6 Tight-binding 2D DOS of an 113 AL Thick Ge Channel 82

6.7 Ballistic $I_D - V_G$ for 113 AL thick Ge MOSFET at $V_D = V_{DD}$ 83

6.8 Tight-binding 2D DOS of a 30 AL Thick Ge Channel 84

6.9 Ballistic $I_D - V_G$ for 30 AL thick Ge MOSFET at $V_D = V_{DD}$ 85

6.10 Bandstructure and 2D DOS for a 12 AL Ge (100) body 86

6.11 Ballistic $I_D - V_G$ for 12 AL thick Ge MOSFET at $V_D = V_{DD}$ 86

6.12 Biaxially and Uniaxially Strained Si p-MOSFETs 88

6.13 Quantized valence subband $E - \vec{k}$ from TB simulations of strain Si. . . . 89

6.14 Ballistic $I_D - V_G$ for Strained Si p-MOSFETs Compared 90

6.15 UTB DG MOSFET and the Underlying Atomic Arrangements 91

6.16 Formation of 2D Brillouin Zone. 92

6.17 The Top-of-the-barrier Ballistic Model 93

6.18 Si and Ge subband dispersions: 19 AL (001) Wafer 94

6.19 GaAs and InAs subband dispersions: 19 AL (001) Wafer 95

6.20 Ballistic $I_D - V_D$ at $V_G = V_{DD}$ =0.6V for UTB (19 AL) DG p- and n-MOSFETs 96

6.21 Injection velocity, υ_{inj}, vs. V_G plots for 19 AL n- and p- MOSFETs at $V_D = V_{DD} = 0.6$V. 97

6.22 Carrier density, Q_{top}, vs. V_G plots for 19 AL n- and p- MOSFETs at $V_D = V_{DD}$ =0.6V 97

6.23 CB Subband Dispersions for Thick (100 AL) Bodies 98

Figure Page

6.24 $I_D - V_D$ at $V_G = 0.6$V for Thick (100 AL) n-MOSFETs for Two Different EOTs 99

6.25 Injection velocity, v_{inj}, vs. V_G plots at $V_D = 0.6$V for thick body (100 AL) DG n-MOSFETs with two EOTs 100

6.26 Injection velocity, v_{inj}, vs. V_G plots at $V_D = 0.6$V for thick body (100 AL) DG n-MOSFETs with two EOTs 101

7.1 Self-consistent Tight-binding Scheme 105

7.2 Band Bending Along Thickness Cannot Be Ignored for Bulk and HOI MOSFETs 105

7.3 Strained MOSFET Geometries 107

7.4 Experimental Hole Mobility Data for Strained and Unstrained p-MOSFETs 108

7.5 Self-consistent Q-V and I-V Results for Single-gate p-MOSFETs 109

7.6 Ballistic Performance Ehancement in Strained p-MOSFETs 110

7.7 Ballistic Current Ratio for Strained and Unstrained n-MOSFETs 111

7.8 The HOI Structure 113

7.9 HOI Charecteristics in Subthreshold 115

7.10 HOI Charecteristics Around Threshold 116

7.11 HOI Charecteristics Above Threshold 117

7.12 Simulated Q-V and C-V characteristics for HOI 118

B.1 Treating Floating Boundary Condition. 145

ABSTRACT

Rahman, Anisur. Ph.D., Purdue University, December, 2005. Exploring New Channel Materials for Nanoscale CMOS Devices: A Simulation Approach. Major Professor: Mark Lundstrom and Gerhard Klimeck.

The improved transport properties of new channel materials, such as Ge and III-V semiconductors, along with new device designs, such as dual gate, tri gate or FinFETs, are expected to enhance the performance of nanoscale CMOS devices. Novel process techniques, such as ALD, high-κ dielectrics, and metal gates are now being used to experimentally explore such devices. New materials in the channel promise reduced series resistance and higher on-currents. The theoretical assessment of such devices is a challenge because bandstructure, arbitrary wafer orientation, quantum effects and electrostatics must all be treated. In the first part of this work, a general theoretical approach for the quantum mechanical simulation of n-MOSFETs within the Non Equilibrium Green's Function (NEGF) formalism is introduced, and its application is demonstrated by performing a scaling study for the end of the ITRS Ge device. In the second part of this work, a systematic analysis of the bandstructure effects in deeply scaled n- and p- MOSFETs with Si, Ge, GaAs and InAs channel is performed. Here, a 20 orbital $sp^3d^5s^*$-SO tight-binding model and a top-of-the-barrier quasi-2D ballistic transport model have revealed important trends in deeply scaled new channel material devices.

1. INTRODUCTION

In order to assess and compare the performances of nanoscale CMOS devices with novel channel materials (e.g. strained Si, Ge, GaAs or InAs), a simulation approach is undertaken in this work. In Sec. 1.1, a brief overview of the evolution of MOSFETs, over a period of four decades, is presented. In Sec. 1.2, the recent experimental demonstration of novel channel material CMOS devices is reviewed. In Sec. 1.3, the issues and challenges associated with the simulation of nanoscale novel channel materials are highlighted. Finally, in Sec. 1.4, the outline of this thesis is presented.

1.1 CMOS Scaling: From Long Channel to Nanometer Scale Devices

Since their first demonstration in 1960 [1], planar Si metal-oxide-semiconductor field-effect-transistors (MOSFETs) have experienced a steady, exponential downsizing of their critical dimensions. Over a period of 45 years, the printed gate lengths of the MOSFETs have been scaled down from $100\mu m$ to 35 nm [2], the later refers to the gate lengths for the 65nm technology node devices in commercial microprocessors which will be available by the end of 2005. Due to a steady improvement of their performances through scaling, MOSFETs have become the leading integrated circuit technology for high performance and low power logic applications. Over this long period of development, the technology has faced numerous challenges, which were always solved by vigorous research, ingenuous design and brilliant engineering. An excellent review of CMOS material and structural changes during the past 40 years, and their future trends, can be found in [3].

The exponential scaling down of the feature sizes, and hence the exponential increase of the transistor count in an integrated circuit, was first observed by Gordon Moore in 1965 [4, 5]. His observation, which later became known as the Moore's

Law, states that the number of transistors per integrated circuit doubles every 24 months. Moore's law has been serving as the guiding principle for the semiconductor industry for over 30 years [6,7]. According to projections by International Technology Roadmap for Semiconductors (ITRS), by the middle of next decade, the printed gate length of the MOSFETs will be less than 10nm. Since no exponential can continue forever, sustaining Moore's Law is becoming challenging for the planar bulk CMOS technology, where the key technical issue for scaling devices below 32nm node is the off-state leakage current [3]. Demonstration of planar MOSFETs with physical gate lengths as short as 5nm [8] and 15nm [9] can be found in the literature, however, due to high off-state leakage, they are not a viable option for integrated circuits. The origin of this scaling limit for the single-gate, bulk CMOS technology can be traced back to the inherent poor electrostatic design of the planar device geometry, and the poor transport properties of carriers in the silicon channels. Consequently, an intense research effort, directed toward exploring new device designs and new channel materials for the future logic technologies, has recently been undertaken.

Planar silicon-on-insulator (SOI) technology shows promise for scaling beyond the planar bulk MOSFET limit. Devices with gate lengths as short as 6 nm and channel thickness of 4.6 nm has been demonstrated recently [10]. Recent demonstration of ultra-thin-body SOI MOSFETs with channel thickness less than 1nm, or only five atomic layers, shows the maturity of this technology [11]. The electrostatic integrity of the nanoscale CMOS devices improves considerably when additional gates are included, such as for dual-gate [12, 13], tri-gate [14, 15] or FinFET devices. Due to the inherent resistance of these non-planar devices to short-channel-effects, it is widely believed that one of them will form the basic device architecture for future generations of CMOS devices.

In order to sustain Moore's law for over four decades, nearly all materials and design aspects of the original MOSFET have been lost, except the use of Si-SiO_2 material system as channel material and gate insulator. Aggressive scaling of gate oxide has reduced its thickness to a present day value of about 1nm, and any further

thinning is prohibitive due to oxide reliability issues and the exponential increase of leakage current from direct tunneling. Sustaining Moore's Law, however, depends crucially on the gate insulator scaling and consequently, replacing SiO_2 with high-κ dielectric–metal gate stack is within sight [16]. Among all semiconductors, although silicon has the poorest transport properties (mobilities and diffusion coefficients for electron and holes), its success as channel material is due to the excellent interface property of the Si-SiO_2 system. As SiO_2 will almost certainly be replaced by high-κ gate dielectric in near future, the replacement of silicon as channel material by novel materials with improved transport properties is being viewed as a promising option to continue device scaling until the middle of next decade. Research activity in this area has experienced a boom in recent years; with many new devices and material systems now proposed. A benchmarking study of such new research devices, against the existing Si technology, has recently been published in [17]. We will next review the recent experimental advancements in the area of novel channel material CMOS devices.

1.2 Novel Channel Material CMOS: Experimental Exploration

The experimental exploration of novel channel materials for CMOS devices is primarily motivated by their excellent transport properties. Their high room-temperature mobilities and saturation velocities are thought to be the key to the next generation ultra-fast, low power CMOS digital logic technology—an assumption we will examine in details in chapters 3, 6 and 7 of this thesis.

For long time, strain has been known to improve the channel transport properties of MOSFETs. Strained Si is the only *new* channel material which has recently made its way into the commercial integrated circuits. Beginning with the 90 nm technology node devices, released in 2003, leading IC industries have incorporated strained silicon, in some form, to improve the channel transport properties [2, 18–22]. Recently, substantial progress has also been made to incorporate strain in SOI

structures using bond-and-etch-back technique [23, 24]. Modulations in electron and hole mobilities with the scaling of body thicknesses in strained SOI has been reported in [25] and [26]. Also, devices fabricated on Si (110) wafer orientations has shown improved mobility characteristics over (100) devices [27–29]. Recently, similar results for (110) strained SOI MOSFETs has also been published [30].

Beyond silicon, germanium is an interesting candidate for nanoscale CMOS technology due to its excellent transport properties — two and four times bulk mobilities for electrons and holes compared to silicon, respectively. Room temperature hole mobility in a 7.5nm thick Ge quantum well has already been reported to exceed 2500 cm^2/V-sec [31]. Recently, a great progress has been made to integrate high-κ gate dielectric with Ge process and active research in this field is underway [32–35]. High performance, n- and p- channel Ge MOSFETs has been reported in [36–41].

Robust and highly manufacturable new process technologies, such as atomic layer deposition (ALD), hetero-epitaxy and metal gates, have opened the opportunity to integrate III-V semiconductors with Si technology. With their exceptionally high mobilities, III-V materials display promise for ultra-fast, very low power digital logic technology. In [42–44], using ALD Al_2O_3 as the gate insulator, GaAs MOSFETs with excellent performance was reported for the first time. Later, GaAs MOSFET with oxidized InAlP gate insulator was reported in [45]. Other III-V materials, InAs and InSb, also show great promise as novel channel material for logic technology due to their exceptionally high carrier mobilities. Recently, for the first time, InSb based Quantum Well FET was reported in [46]. Fischetti *et al.* showed in [47] that indium based semiconductor can outperform Si and Ge MOSFETs in deeply scaled MOSFETs.

1.3 Simulation Approach: Issues and Challenges

In addition to experimental exploration of nanoscale novel channel material MOSFETs, physics based simulation for such devices can offer valuable insight into their

operation and can help their design optimization. Numerical simulation not only proves valuable to guide experiments and to explain their results, it also helps to identify the strengths and the weaknesses of different approaches in the emerging field of nanoelectronics. The simulation approach has already proven useful in determining the performance limits of the Si technology by comparing their experimental performances with their ballistic performances [48]. A full 2-D simulation tool, nanoMOS 2.5, developed at Purdue University, quantum mechanically models the Si n-MOSFETs fabricated on (100) wafers.

Device modeling at nanoscale consists of self-consistently solving the Schrödinger equation and the Poisson equation. For a given potential profile, Schrödinger equation quantum mechanically calculates the carrier densities and their transmission probabilities, while the Poisson equation ensures that the charge profile is consistent with the potential profile. The most widely invoked assumption to solve the Schrödinger equation is known as the *effective mass approach*, where the slowly varying envelope of the electronic wave function is obtained by solving a Schrödinger-like *effective-mass equation* (EME). Inside the tool nanoMOS 2.5, this effective mass equation is solved within the Non-Equilibrium Green's Function (NEGF) formalism. A a coupled/decoupled mode-space approach was used, where, by expanding the electronic wavefunction in mode-space, great efficiency was achieved for solving the quantum transport problem [49–52]. Another full 2-D quantum mechanical tool, QDAME, developed at IBM, solves the effective mass equation to treat the open-boundary ballistic quantum transport problem by expanding the electronic wavefunction as a linear combination of wavefunctions which satisfy zero value and zero slope boundary conditions [53]. Both nanoMOS and QDAME employ parabolic E-k relationships for electronic bandstructure. Scaling study results for UTB nanoscale Si (100) n-MOSFETs has already been published using both nanoMOS [54–56] and using QDAME [57].

For Si n-MOSFETs fabricate on (100)-wafers, the effective mass approach has been proved to be an enormously useful technique to treat quantum transport in

an approximate manner. However, there are two limitations. First, for holes in the valence bands, where the parabolic E-k bandstructure is not valid and the heavy, light and spin-off valleys are strongly coupled, effective mass approach results in a complicated $\vec{k} \cdot \vec{p}$ description of bandstructure, which is not suitable for quantum simulation of hole transport. Second, even for electrons in the conduction band, if the principal axes of the constant energy ellipsoids are not aligned with the device axes (channel, thickness and width directions), the effective mass equation for the electrons becomes enormously complicated. This becomes a serious issue, limiting the application of effective mass equation to novel channel material devices, since for germanium n-MOSFETs, or silicon n-MOSFETs on wafer orientations other than (100), the device axes and the ellipsoid axes are no longer aligned. Consequently, the usefulness of the mode-space approach vanishes for quantum mechanical treatment of electronic transport in novel-channel material n-MOSFETs. QDAME employs a technique to address this problem by discretizing effective mass equation along the principal axis of the ellipsoid; however, this becomes a real-space approach therefore, is numerically cumbersome [58, 59]. In order to use the efficient techniques, such as mode-space approach and NEGF formalism, the first challenge for simulation of novel channel material n-MOSFET is to develop a generalized effective mass approach where the complicated Hamiltonian arising from the non-alignment of device and ellipsoid axis can be simplified.

As already pointed out, effective-mass-approach is an approximation which disregards the atomic scale fluctuation of the electronic characteristics and describes the band edge electronic properties in an approximate manner. However, as size goes down, the behavior of the electronic states in nanoscale CMOS devices become increasingly sensitive to all sorts of *microscopic phenomena*: atomic-scale fluctuations, local bond distortions, alloy effects, structure of the interfaces, quantum tunneling and energy quantization. An improved modeling of such effects is not possible within effective mass approach and a full band atomistic treatment is necessary to address them. Semi-empirical tight-binding approach [60–62], a full band technique, proves

extremely useful for atomistic treatment of nanoscale devices with any materials, provided that the tight-binding parameters for the material is known in advance. It correctly captures full-band effects, such as valley splitting in a nanostructure [63], and also treats alloy effects on bandstructure [64].

Tight-binding modeling of nanostructures, using s, p and d type atomic orbital basis, has already been successfully employed for the predictive simulation of 1-D structures such as resonance tunneling diode (RTD) [65], and zero dimensional system such as quantum dots [66]. In silicon technology, its application is limited to calculation of MOS gate oxide tunneling current, an inherently 1-D problem [67–71]. Nanoscale UTB MOSFETs, however, require a 2-D atomistic description of the structure, along channel and along thickness, making them extremely challenging, computationally. As a result, no attempt has yet been made to address the carrier transport in a MOSFET using the tight-binding method. Since, it is computationally prohibitive to do a full 2-D atomistic treatment of MOSFETs, the second challenge in simulation of novel channel material devices is to develop the appropriate transport model and make the necessary simplifying assumptions that allow the use of full band atomistic tight-binding model to simulate such devices.

1.4 Outline of The Thesis

This thesis is divided into following chapters:

- Chapter 2 (page 11): This chapter generalizes the effective mass approach to treat nanoscale n-MOSFET fabricated on arbitrarily oriented wafers. It is shown that under certain simplifications, which are generally valid for UTB nanoscale devices, an arbitrarily oriented constant energy ellipsoid can be transformed into a regular ellipsoid which has its ellipsoidal axes aligned along the device axes. The three effective masses along channel, width and thickness, calculated from the transformed ellipsoid, can then be used in any quantum simulation tool developed for Si n-MOSFETs on (100) wafers (such as

nanoMOS) to treat quantum transport of electrons. The approach in this chapter has been published as a full length journal paper in Journal of Applied Physics [72].

- Chapter 3 (page 38): An application of the generalized effective mass approach, introduced in Ch. 2, is demonstrated in this chapter. A design study for the UTB dual-gate (DG) 10nm gate length germanium n-MOSFET is performed using the specifications given for end of the ITRS 2001 device. All simulation were performed by using the correct effective masses for Ge in nanoMOS 2.5. Contrary to popular belief, that devices fabricated on Ge (111) wafers will show best performance, it was observed that n-MOSFETs fabricated on (100) wafers will offer the best performance. The reason for this is, the density-of-states for electrons in Ge (111) is severely degraded, which in turn degrades the gate capacitance. As a result, although the electron injection velocity for Ge(111) devices is very high, fewer carriers in the channel reduces the current. NEGF scattering simulation using a simple Büttiker-probe treatment shows that the electron transport in the channel is near ballistic, but the performance will be limited by the low mobility, highly doped source and drain regions. The results of this chapter has been published in the Technical Digest of 2003 International Electron Device Meeting [73].

- Chapter 4 (page 51): A top-of-the-barrier semiclassical ballistic transport model, for arbitrary bandstructures, is introduced in this chapter. It was demonstrated that for a given set of *gate and drain control parameters, and source Fermi level position*, all extracted from the subthreshold region of device operation, this model can reproduce nanoMOS 2.5 results for a Si (100) UTB DG n-MOSFET. The treatment of *floating source potential*, a characteristic phenomena in ballistic FETs, can be easily incorporated in this model. Also, *quantum capacitance effects* due to finite density-of-states in a nanoscale MOSFET is autometically included in this model. This top-of-the-barrier ballistic

transport model will be the key vehicle for the assessments of novel channel material devices presented in subsequent chapters. The model in this chapter has been published as a full length journal paper in IEEE Transactions on Electron Devices [74].

- Chapter 5 (page 68): This chapter describes the semi-empirical tight-binding model and its application to nanoelectronic devices. The historical development and the assumptions involved are clearly described. Also the technique to apply the model to nanostructures to calculated 2-D bandstructure is explained.

- Chapter 6 (page 77): In this chapter the application of tight-binding approach to UTB DG MOSFETs is presented. First, the validity of using parabolic E-k effective mass approach to Ge n-MOSFETs was investigated and was observed that below a body thickness of 4nm, use of parabolic E-k can introduce serious error in the performance assessment of n-MOSFETs. These results has been published in the Technical Digest of 2004 International Electron Device Meeting [75]. In the second half of this chapter, performances of UTB DG nanoscale n- and p- MOSFETs with Si, Ge, GaAs and InAs were compared. It was observed that there is often *a tread-off between the high-injection velocity and the gate capacitance degradation due to low density-of-states*. As a result, for highly scaled MOSFETs, GaAs and InAs channel devices fail to out-perform Si and Ge channel devices and Ge channel becomes the clear winner. The results presented in this part will appear in the Technical Digest of IEDM 2005 [76].

- Chapter 7 (page 103): Application of tight-binding model with self-consistent electrostatics is presented in this chapter. Two type of devices, bulk strained MOSFETs and heterostructure on insulator (HOI) devices, were explored. It was demonstrated that the experimentally observed hole mobility vs. gate field behavior in single-gate bulk strained p-MOSFETs can be explained in terms of their ballistic performances. However, bandstructure alone cannot

explain similar behavior in the n-MOSFETs. Simulation of charge profile in HOI devices shows that the profile critically depends on the gate bias, and the band discontinuity alone does not control the peak of hole density profile along the thickness.

- Chapter 8 (page 119): Summary of the thesis is presented and future work suggested.

- Appendix A (page 132): Mathematical details and derivations relevant to Ch. 2 are presented.

- Appendix B (page 139): Mathematical details and derivations relevant to Ch. 4 are presented.

- Appendix C (page 146): Mathematical details and derivations relevant to Ch. 5 are presented.

2. A GENERALIZED EFFECTIVE MASS APPROACH FOR N-MOSFETS

In this chapter, the general theory for quantum simulation of cubic semiconductor n-type metal oxide semiconductor field effect transistors (MOSFETs) is presented within the effective mass equation approach. The full three-dimensional transport problem is described in terms of coupled transverse subband modes, which arise due to quantum confinement along the body thickness direction. Couplings among the subbands are generated for two reasons: due to spatial variations of the confinement potential along the transport direction, and due to non-alignment of the device coordinate system with the principal axes of the constant energy conduction band ellipsoids. The problem simplifies considerably if the electrostatic potential is separable along transport and confinement directions, and further, if the potential variations along the transport direction are slow enough to prevent dipolar coupling (Zener tunneling) between subbands. In this limit, the transport problem can be solved by employing two unitary operators to transform an arbitrarily oriented constant energy ellipsoid into a regular ellipsoid with principal axes along the transport, width and confinement directions of the device. The effective masses for several technologically important wafer orientations for silicon (Si) and germanium (Ge) are calculated in this chapter.

2.1 Introduction

Metal oxide semiconductor field effect transistors (MOSFETs) constitute the fundamental building block of present day complementary metal oxide semiconductor (CMOS) technology. Current research in this field is largely geared towards improv-

ing MOSFET performance and increasing device density through aggressive scaling of their feature sizes [6, 7]. The importance of quantum mechanical size effects in MOSFETs, where the inversion layers are just a few nanometers thick, was realized during the early period of their development [77,78]. Moreover, as MOSFET channel lengths approach few tens of nanometers, source-to-drain and gate tunneling in these near-ballistic devices become important issues [79]. Numerical simulations provide valuable insight into the physics of device operation at this scale, requiring an appropriate treatment of the device bandstructure as well as a rigorous formulation of quantum transport.

The effective mass equation (EME) provides an accurate, easy to implement model Hamiltonian that does justice to the device bandstructure including quantum confinement effects within the inversion layer, and describes the slowly varying envelope part of the underlying Bloch wavefunction. The Non-Equilibrium Green's Function (NEGF) method provides a rigorous formulation of quantum transport in nanoscale devices [80]. Together, the NEGF formalism and the EME have been used to describe transport in nanoscale MOSFETs both in the ballistic limit [49, 81–83], as well as including the effects of carrier scattering [51, 84]. In [49] and [85] the coupled- and decoupled-mode-space approaches were introduced, and in [52] the coupled-mode-space approach is used in order to assess the effects of channel access geometry and series resistance in nanoscale n-MOSFETs. In the presence of strong dephasing with band like transport, the NEGF equation reduces to the semiclassical Boltzmann transport equation (BTE). The BTE has also been used, along with related concepts such as density-of-states (DOS) and conduction band (CB) effective masses, in order to explore the upper limit of nanoscale MOSFET performance [48, 74, 86]. The two dimensional (2D) numerical simulator nanoMOS 2.5, a freeware, has been developed to simulate the ballistic and scattering characteristics of ultra-thin-body (UTB), double gate (DG) Si (100) n-MOSFETs using both semiclassical (BTE) and fully quantum (NEGF) methods [50, 87].

Silicon (100) wafers are almost universally used by the semiconductor industry for CMOS integrated circuit fabrication. Simulation of n-MOSFETs is generally performed for devices fabricated on (100) wafers, motivated by its technological importance. The quantum simulation of Si (100) devices is substantially simplified by the fact that the principal axes of the six fold degenerate conduction band ellipsoids are aligned along the device coordinate axes, effectively decoupling the kinetic energies along the device coordinate axes. In general, however, the principal axes of the conduction band ellipsoids are not aligned with the device axes, so that the associated kinetic energies become coupled and the effective mass equation becomes non-trivial. Such a situation arises for transistors that employ germanium as a high-mobility channel material [32, 36, 38], as well as for alternate wafer orientations of silicon [27]. To extend the application of EME to analyze these novel n-MOSFETs it is necessary to generalize the EME approach to arbitrary wafer orientations. In the past, Stern *et al.* proposed a method which, in such non-trivial cases, decouples the kinetic energy associated with the quantum confinement direction from that associated with the motion in the transport plane [77, 88]. In this chapter we introduce a technique that decouples the energy associated with all three device axes, i.e., transport, width and confinement directions, for devices with unvarying cross sections and slowly varying channel-directed potentials. This allows us to use all the EME based simulation tools developed so far for modeling novel channel material n-MOSFETs.

This chapter is organized as follows. In Sec. 2.2 we outline our general solution procedure, describing the full three dimensional (3D) problem, and the conditions under which it can be simplified. In Sec. 2.3 we discuss the conduction band structure in cubic semiconductors and derive the effective-mass-tensor (EMT) in an arbitrary, orthogonal device coordinate system. In Sec. 2.4 we present the technique to solve the resulting EME for n-MOSFETs. This general technique shows that under certain conditions one can employ two unitary transformations that map any arbitrarily oriented constant energy ellipsoid onto a regular ellipsoid having principal axes ori-

ented along the device axes. In Sec. 2.5, the effective masses are calculated for several technologically significant silicon and germanium wafer orientations. We follow by presenting the discussions in Sec. 2.6, and finally, we summarize and conclude this section in Sec. 2.7.

2.2 Summary of the Overall Solution Procedure

The complete problem involves the full 3D quantum transport and electrostatics of the system. In page 18, we start by writing down the non-diagonal EMT in the device coordinate system, and the corresponding dispersion relationship (2.12) for the arbitrary oriented conduction-band ellipsoids. We then perform a basis transformation which recasts the general effective-mass equation in a *fully equivalent* form in terms of a complete set of transverse subband eigenmodes. These transverse modes are obtained by considering the confinement potential along the principal-axis directions of the constant energy ellipsoids (2.19). The general 3D nature of the problem is manifested in (2.34) (page 24) through coupling among these subbands, so that the corresponding coupled-mode-space transport formulation is fully equivalent to the original effective-mass (2.14), *with no further assumptions*. Some of the coupling terms disappears if the confinement potential remains same as we move along the transport direction, making the overall electrostatic potential separable (2.35, 2.36); at this stage, however, there are still coupling among the different subbands representing subband-to-subband Zener tunneling caused by the nonalignment of the ellipsoidal principal axes with the device coordinate axes (2.39). These couplings are not present when the device axes align with ellipsoidal axes, as in Si (100) devices. The origin of these couplings can be traced back to the nonalignment of the device and ellipsoidal axes, which complicates the description of kinetic energy in the device coordinate system through the non-diagonal effective-mass-tensor (2.12). In effect the couplings arise because the channel potential itself is varying along the confinement direction, effectively coupling the two coordinates. If now

the transport potential varies slowly enough that the total variation in the channel potential between the confinement planes along the ellipsoidal axis is much smaller than the subband separation (2.40), the inter-subband coupling terms are further eliminated, leading to a simplified decoupled-mode-space description in terms of isolated ellipsoids with their principal axes oriented along the device axes (2.41), albeit with modified effective masses. The problem is further simplified computationally for ultra-thin-body MOSFETs with a large energy difference between the transverse subbands, so that only the lowest few modes that are thermally populated need to be considered.

2.3 Conduction-band Structure in cubic semiconductors

The conduction-band minima of cubic semiconductor materials appear either at a single point (for direct band-gap materials such as GaAs or InAs) or at multiple equivalent points (for indirect band-gap materials such as Si and Ge) within the first Brillouin zone (BZ). The constant energy surfaces become non-parabolic and warped for energies away from the band minima; close to the band edge, however, the relevant electronic states for transport calculations can be described by simple ellipsoidal surfaces. We can safely ignore the coupling with the valence band for semiconductors with moderately large band gaps. Under these circumstances, the constant energy surface for electron in the direct band-gap material is spherical, centered on the Γ point and described as

$$E = \frac{\hbar^2 k^2}{2m_{eff}}, \tag{2.1}$$

with a constant, isotropic effective mass, m_{eff}. For indirect semiconductor materials, the CB minima are located at multiple equivalent points: six points near X along the Δ ($\equiv< 100 >$) crystallographic directions for silicon, and eight equivalent points at L along Λ ($\equiv< 111 >$) for germanium. In indirect semiconductors, the constant energy surfaces are ellipsoids of revolution around Δ and Λ axes, respectively [89, 90], requiring two effective masses, longitudinal m_l and transverse m_t, for

description. In general, the nonalignment of the ellipsoidal principal axes with the device coordinate axes causes the effective mass to become a 3×3 tensor quantity in the device coordinate system [91, 92]. In this section, we will systematically derive this effective-mass-tensor in an arbitrary orthogonal coordinate system, and in Sec. 2.4 we will simplify the resulting EME for quantum transport simulation of n-MOSFETs. We formulate the generalized EME by defining three orthogonal coordinate systems, presented schematically in Fig. 2.1. They are called the device coordinate system (DCS), the crystal coordinate system (CCS), and the ellipsoid coordinate system (ECS). Three unit vectors, $\hat{k}_1$, $\hat{k}_2$, and $\hat{k}_3$, span the DCS and form its basis. We take $\hat{k}_3$ along the body thickness (i.e., quantum confinement of inversion carriers), $\hat{k}_1$ along the source-to-drain (i.e., transport) direction, and $\hat{k}_2 (\equiv \hat{k}_3 \times \hat{k}_1)$ along the device width direction. The second coordinate system, CCS, is spanned by three unit vectors, $\hat{k'}_1$, $\hat{k'}_2$, and $\hat{k'}_3$, oriented along the tree orthogonal $< 100 >$ crystallographic directions of the underlying channel material. Finally, the basis for the ECS consists of the unit vectors $\hat{k}_{||}$, $\hat{k}_{\perp 1}$, and $\hat{k}_{\perp 2}$, chosen along the principal axes of each constant energy ellipsoid. In summary, the CCS is unique for all our simulations, the DCS depends on the fabrication choice (that is, on the wafer orientation and the source-to-drain direction in the chip design layout), and the ECS depends on the specific channel material and is unique to each ellipsoid.

We now describe the key steps in determining the EMT for a general CB ellipsoid. In the ECS the constant energy ellipsoid can be expressed as:

$$E = \frac{\hbar^2 k_{||}^2}{2m_l} + \frac{\hbar^2 (k_{\perp 1}^2 + k_{\perp 2}^2)}{2m_t}. \tag{2.2}$$

In (2.2), the k-space origin is translated to the CB minima, which serves as the reference for the electronic energy. In compact vector notation, (2.2) can be written as

$$E = \frac{\hbar^2}{2} \vec{k}_E^T \left[M_E^{-1} \right] \vec{k}_E, \tag{2.3}$$

where $\vec{k}_E = (k_{||} k_{\perp 1} k_{\perp 2})^T$ consists of the components of an arbitrary wave vector in the ECS, and the inverse EMT, $\left[M_E^{-1} \right]$, is a 3×3 diagonal matrix with m_l^{-1}, m_t^{-1}, and

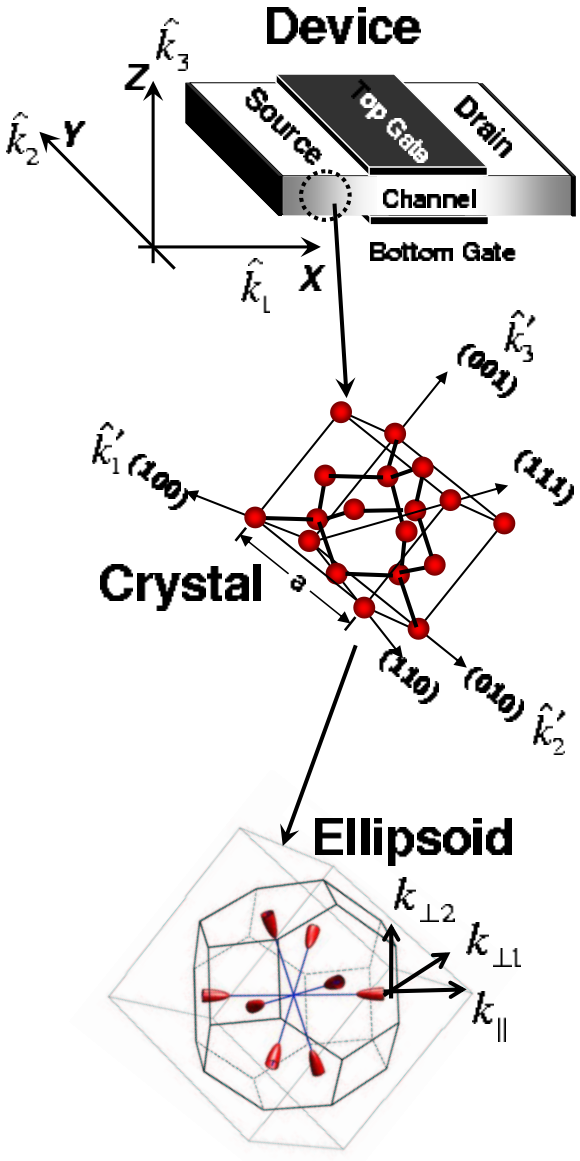

Fig. 2.1. Three orthogonal coordinate systems: Device coordinate system (DCS), Crystal coordinate system (CCS), and Ellipse coordinate system (ECS).

m_t^{-1} along the diagonal. For a given channel material and for a given CB ellipsoid, the directions of the unit basis vectors $\hat{k}_{||}$, $\hat{k}_{\perp 1}$, and $\hat{k}_{\perp 2}$ are known relative to the CCS, thus allowing us to write the $3x3$ rotation matrix $\Re_{E \leftarrow C}$, which transforms

the components of an arbitrary vector $\vec{k}_C \equiv (k'_1 k'_2 k'_3)^T$ defined in the CCS, to its components in the ECS,

$$\vec{k}_E = \Re_{E\leftarrow C}\vec{k}_C. \tag{2.4}$$

A similar rotation matrix $\Re_{C\leftarrow D}$ transforms a wavevector $\vec{k}_D \equiv (k_1 k_2 k_3)^T$ in the DCS to $\vec{k}_C$ in the CCS as

$$\vec{k}_C = \Re_{C\leftarrow D}\vec{k}_D. \tag{2.5}$$

Combing 2.4 and 2.5 we obtain

$$\vec{k}_E = \Re_{E\leftarrow D}\vec{k}_D, \tag{2.6}$$

where the rotation matrix is defined as

$$\Re_{E\leftarrow D} = \Re_{E\leftarrow C}\Re_{C\leftarrow D}. \tag{2.7}$$

Inserting 2.6 into 2.3 we obtain

$$E = \frac{\hbar^2}{2}\vec{k}_D^T \left[M_D^{-1}\right] \vec{k}_D, \tag{2.8}$$

where the inverse effective mass $\left[M_D^{-1}\right]$ in the DCS is

$$\left[M_D^{-1}\right] = \Re_{E\leftarrow D}^T \left[M_E^{-1}\right] \Re_{E\leftarrow D}. \tag{2.9}$$

In Sec. 2.5 we will evaluate $\Re_{E\leftarrow D}$ for various wafer orientations and for both Δ and Λ type CB valleys. From (2.9) we find that the general EMT, $\left[M_D^{-1}\right]$, is a full $3x3$ symmetric matrix whose elements $\left[M_D^{-1}\right]_{ij}$ are

$$\frac{1}{m_{ij}} = \frac{a_{1i}a_{1j}}{m_l} + \frac{a_{2i}a_{2j} + a_{3i}a_{3j}}{m_t}, \tag{2.10}$$

where $a_{ij} = [\Re_{E\leftarrow D}]_{ij}$. Equation 8 can now be written in compact form as

$$E(k_1, k_2, k_3) = \sum_{i,j=1}^{3} \frac{\hbar^2 k_i k_j}{2m_{ij}}. \tag{2.11}$$

From (2.10) we find $m_{ij} = m_{ji}$ and therefore the above expression can be rewritten as

$$E(k_1, k_2, k_3) = \sum_{i=1}^{3} \frac{\hbar^2 k_i^2}{2m_{ii}} + 2\sum_{i=1}^{3}\sum_{i<j\leq 3} \frac{\hbar^2 k_i k_j}{2m_{ij}}. \tag{2.12}$$

Comparing with (2.1), we see that the expression for the constant energy ellipsoidal surface in (2.12) contains additional cross terms $k_i k_j$ in kinetic energy. In Sec. 2.4 we will see that this makes the corresponding general EME non-trivial and there we will outline our treatment of the problem.

2.4 The Generalized Effective-Mass Equation and the Solution

The bulk band structure for any semiconductor is calculated by solving Schrödinegr equation using Bloch's theorem for a periodic lattice. Although this technique yields and accurate description of the $E(\vec{k})$ relationship for the electron over the entire BZ, it is unnecessarily complicated for treating transport problems in the MOSFET device structure. Since an accurate description of only the band-edge electronic states is sufficient for transport simulation, the effective-mass approximation scheme (also known as the envelop function approximation) becomes an attractive alternative. The effective-mass approximation uses an accurate description of the $E(\vec{k})$ relationship over only a limited range of energy near the valence- or conduction-band extrema. The $\vec{k} \cdot \vec{p}$ perturbation technique is employed in this regard, which describes the band structure over a limited range of energy near the band edge with sufficiently accuracy [93]. A Schrödinger like effective mass equation is obtained by replacing certain component of the wave vector, k_j, in the expression for $E(k)$ with their quantum-mechanical operators, $-i\partial/\partial x_j$, and the electronic states are obtained by solving this differential eigenvalues equation.

The EME scheme described above is universally used by the electronic device community for quantum-mechanical simulation of MOSFETs. For silicon devices fabricated on (100) wafers with the source-to-drain direction oriented along [010], the EMT continues to be diagonal and, therefore, the cross terms in 2.12 drop out. In this case, the EME can be solved without difficulty and the quantum-mechanical effects are accurately included in the simulation. Since the work of Stern *et al.* [78], the above-mentioned device orientation has been exclusively used for simulation.

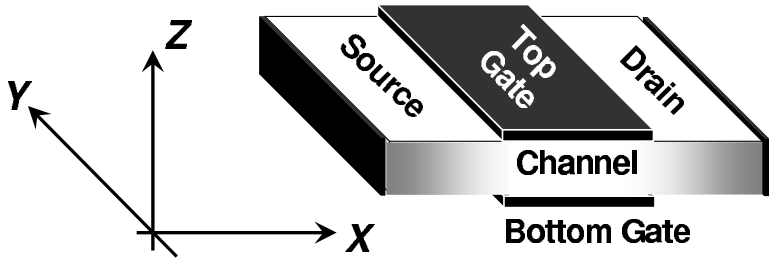

Fig. 2.2. The ultra-thin-body, double-gate device structure. The device coordinate system consists of orthogonal axes X, Y and Z along transport, width and thickness directions, respectively. The wavevectors along the thickness, k_3, and along the transport, k_1, are treated quantum mechanically while in the width direction, Y, plane waves were assumed.

The quantum-mechanical simulation of MOSFETs fabricated on germanium (100) wafers or silicon (111) or (110) wafers still remains a non-trivial problem, since their EMT in the device coordinate system are full 3×3 matrices. In this section we will introduce the general solution technique and will show that under certain simplifying conditions we can decouple the energies in (2.12), thereby, eliminating the limitations of EME. In Fig. 2.2 the UTB DG semiconductor-on-insulator (SOI) MOSFET device structure is shown as the model device; however, the general theory we are developing is valid for bulk MOSFETs as well. In this figure we have defined X, Y, and Z, as the real-space Cartesian axes along the previously mentioned $\hat{k}_1$, $\hat{k}_2$, and $\hat{k}_3$ unit vectors, respectively. Accordingly, we replace the k-subscripts in (2.12) from $\{123\}$ to $\{xyz\}$. For the MOSFET, X, Y, and Z represent the transport, quantum confinement and width directions, respectively.

Our general strategy will be to first solve the quantum problem along the confinement direction and then use the corresponding eigenvectors to construct a complete basis set for the full three-dimensional quantum transport problem. Using the new labels for the axes of the DCS, the general $E(\vec{k})$ in (2.12) becomes

$$E\left(k_x, k_y, k_z\right) = \frac{\hbar^2 k_x^2}{2m_{11}} + \frac{\hbar^2 k_y^2}{2m_{22}} + \frac{\hbar^2 k_z^2}{2m_{33}} + \frac{\hbar^2 k_x k_y}{m_{12}} + \frac{\hbar^2 k_y k_z}{m_{23}} + \frac{\hbar^2 k_z k_x}{m_{31}}. \qquad (2.13)$$

By substituting $k_x \rightarrow -i\partial/\partial x$ and $k_z \rightarrow -i\partial/\partial z$ we now find the corresponding 2D effective-mass equation,

$$\left[-\frac{\hbar^2}{2m_{11}}\frac{\partial^2}{\partial x^2} - i\frac{\hbar^2 k_y}{m_{12}}\frac{\partial}{\partial x} + \frac{\hbar^2 k_y^2}{2m_{22}} + \left\{-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z} + W(x,z)\right\}\right]\Psi_{k_y}(x,z) = E\Psi_{k_y}(x,z). \quad (2.14)$$

The potential along the width direction is unvarying and therefore, $W(x,z)$ is not a function of y. k_y is a constant arising from the commutation of $-i\partial/\partial y$ and the Hamiltonian in 2.14. The two independent degenerate solutions, $e^{\pm ik_y y}$, and of the y-dependent part of 2.14 carry no net current along Y, despite the non-diagonal character of the EMT. We now discuss the general mode-space formalism for solving this problem in general, without postulating any separability for the potential energy $W(x,z)$.

2.4.1 The Quantum Confinement Problem

The confinement modes diagonalize the part of the Hamiltonian associated with the confinement potential and serve as basis sets for evaluating the complete transport equation. From (2.14) we separate the terms dealing with the quantum confinement problem at a given x,

$$[H_z + W(x,z)]\,\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) = \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y : x\right)\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : x, z\right), \quad (2.15)$$

where the confinement Hamiltonian is

$$H_z = -\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z}. \quad (2.16)$$

We now perform a canonical transformation by substituting

$$\zeta\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) = e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(x,z), \quad (2.17)$$

in (2.15) and left multiplying the equation by $e^{i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}$. The exponential term in (2.17) is a unitary operator that basis transforms the wave function. The algebra is considerably simplified by employing the following well-known operator identity:

$$e^{-B}Ae^{B} = A + [A,B] + \frac{1}{2}[[A,B],B] + ... \tag{2.18}$$

The above operation on the kinetic energy H_z causes the linear term in $\partial/\partial z$ to drop out (the exponential term being just the translation operator in $\partial/\partial z$ space), while the corresponding unitary operation on the potential $W(x,z)$ transforms it into $W\left(x + \frac{m_{33}}{m_{31}}z, z\right)$. For a given x, this expression implies that the quantum confinement potential needs to be sampled along the principal axis of the constant energy ellipse at fixed k_y . At the end of this canonical transformation, the confinement problem becomes

$$\left[-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} + W\left(x + \frac{m_{33}}{m_{31}}z, z\right)\right]\phi_i(x,z) = \epsilon_i\left(x + \frac{m_{33}}{m_{31}}z\right)\phi_i(x,z), \tag{2.19}$$

which we have to solve in order to obtain the orthonormal eigenvectors, ϕ_i's, hereafter referred to as "modes" Using the inverse canonical transformation, the confinement problem can be rewritten as

$$\begin{aligned}[H_z + W(x,z)]\, e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(x,z) \\ = \left\{\epsilon_i(x) - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right)\right\} e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(x,z),\end{aligned} \tag{2.20}$$

where $\epsilon_i(x)$ is the i-th subband energy at x and

$$\epsilon\left(-i\frac{\partial}{\partial x}, k_y\right) = \frac{\hbar^2}{2}\left(-\frac{m_{33}}{m_{31}^2}\frac{\partial^2}{\partial x^2} + \frac{m_{33}}{m_{32}^2}k_y^2 - 2i\frac{m_{33}}{m_{31}m_{23}}k_y\frac{\partial}{\partial x}\right), \tag{2.21}$$

is the kinetic energy.

Equation (2.20) embodies two accomplishments when compared with (2.15). Firstly, we have identified the transverse modes ϕ_i which together with the exponential pre-factor selectively diagonalize the confinement part of the Hamiltonian; secondly, we see that the eigenenergy term is separated into subband energy and kinetic energy. Next we will exploit these two accomplishments.

2.4.2 The Transport Problem

We now return to the original 2D EME in (2.14), by

$$\left[-\frac{\hbar^2}{2m_{11}}\frac{\partial^2}{\partial x^2}-i\frac{\hbar^2 k_y}{m_{12}}\frac{\partial}{\partial x}+\frac{\hbar^2 k_y^2}{2m_{22}}+\overline{H_z+W\left(x,z\right)}\right]\Psi\left(-i\frac{\partial}{\partial x},k_y:x,z\right) = E\Psi\left(-i\frac{\partial}{\partial x},k_y:x,z\right). \quad (2.22)$$

At a given x, the eigenfunctions $e^{-i\left(\frac{m_{33}}{m_{23}}k_y-i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i\left(x,z\right)$ diagonalize the confinement Hamiltonian in (2.20) and form a complete set. This allows us to expand the wavefunction Ψ in (2.22) in this complete basis

$$\Psi\left(-i\frac{\partial}{\partial x},k_y:x,z\right)=\sum_m e^{-i\left(\frac{m_{33}}{m_{23}}k_y-i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_m\left(x,z\right)\chi_m\left(x,k_y\right), \quad (2.23)$$

where $\chi_m\left(x,k_y\right)$ are the corresponding expansion coefficients. We substitute (2.23) in (2.22) and left multiply it by $\phi_n^*\left(x,z\right)e^{i\left(\frac{m_{33}}{m_{23}}k_y-i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}$, which amounts to doing a unitary transformation for the transport Hamiltonian. Using the operator identity (2.18) and the confinement eigenvalues from (2.20), equation (2.22) boils down to

$$\sum_m \phi_n^*\left(x,z\right)\left[H_{trans}+\epsilon_m\left(x+\frac{m_{33}}{m_{31}}z\right)\right]\phi_m\left(x,k_y\right) = E\sum_m \phi_n^*\left(x,z\right)\phi_m\left(x,z\right)\chi_m\left(x,k_y\right), \quad (2.24)$$

where the Hamiltonian for this transport problem is now

$$H_{trans}=-\frac{\hbar^2}{2m_1'}\frac{\partial^2}{\partial x^2}-i\frac{\hbar^2 k_y}{m_{12}'}\frac{\partial}{\partial x}+\frac{\hbar^2 k_y^2}{2m_2'}. \quad (2.25)$$

In (2.24) the new effective masses are obtained by regrouping terms as

$$\frac{1}{m_1'} = \left(\frac{1}{m_{11}}-\frac{m_{33}}{m_{31}^2}\right) \quad (2.26)$$

$$\frac{1}{m_2'} = \left(\frac{1}{m_{22}}-\frac{m_{33}}{m_{23}^2}\right) \quad (2.27)$$

$$\frac{1}{m_{12}'} = \left(\frac{1}{m_{12}}-\frac{m_{33}}{m_{23}m_{31}}\right) \quad (2.28)$$

Equation (2.24) can be simplified further using another canonical transformation,

$$\chi_m(x,k_y)=e^{-i\frac{m_1'}{m_{12}'}k_y x}\psi_m(x), \quad (2.29)$$

which on substituting in (2.24) and left multiplying with $e^{i\frac{m_1'}{m_{12}'}k_y x}$ eliminates the linear terms in $\partial/\partial x$. We find

$$\sum_m \phi_m^*(x,z)\left[-\frac{\hbar^2}{2m_1'}\frac{\partial^2}{\partial x^2}+\frac{\hbar^2 k_y^2}{2m_2''}+\epsilon_m\left(x+\frac{m_{33}}{m_{31}}z\right)\right]\phi_m(x,z)\psi_m(x) = E\sum_m \phi_n^*(x,z)\phi_m(x,z)\psi_m(x), \tag{2.30}$$

where we have again made use of (2.18) and defined the mass

$$\frac{1}{m_2''}=\left(\frac{1}{m_2'}-\frac{m_1'}{m_{12}'^2}.\right) \tag{2.31}$$

Equation (2.30) is now integrated over z. Employing the orthogonality condition

$$\int \phi_n^*(x,z)\phi_m(x,z)dz=\delta_{n,m},$$

we find

$$\sum_m \int dz\left\{\phi_n^*(x,z)\left[-\frac{\hbar^2}{2m_1'}\frac{\partial^2}{\partial x^2}+\epsilon_m\left(x+\frac{m_{33}}{m_{31}}z\right)\right]\ phi_m(x,z)\right\}\psi_m(x) = \left(E-\frac{\hbar^2 k_y^2}{2m_2''}\right)\psi_n(x). \tag{2.32}$$

The potential term in (2.32) can be expanded by using the Taylor series as,

$$\begin{aligned}\int \phi_n^*(x,z)\epsilon_m\left(x+\frac{m_{33}}{m_{31}}z\right)\phi_m(x,z)dz &= \int \phi_n^*\left[\epsilon_m(x)+\frac{m_{33}}{m_{31}}z\frac{\partial\epsilon_m(x)}{\partial x}+\cdots\right]\phi_m dz\\ &= \epsilon_m(x)\delta_{nm}+\frac{m_{33}}{m_{31}}\frac{\partial\epsilon_m(x)}{\partial x}\int z\phi_n^*\phi_m dz+\cdots\\ &= \epsilon_m(x)\delta_{nm}+\frac{m_{33}}{m_{31}}\frac{\partial\epsilon_m(x)}{\partial x}\mu_{nm}+\cdots\\ &= \epsilon_m(x)\delta_{nm}+W_{nm}(x)\end{aligned} \tag{2.33}$$

while the kinetic-energy operator can be simplified using integration by parts, yielding our general coupled-mode-space equation in full 3D transport problem.

$$\begin{aligned}&\left[-\frac{\hbar^2}{2m_1'}\frac{\partial^2\psi_n(x)}{\partial x^2}+\epsilon_n(x)\psi_n(x)\right]+\sum_m W_{nm}(x)\psi_m(x)\\ &+\frac{\hbar^2}{2m_1'}\sum_m\left\{\psi_m(x)\int dz\left[\phi_m(x,z)\frac{\partial^2\phi_n^*(x,z)}{\partial x^2}\right]+2\int dz\frac{\partial\phi_n^*(x,z)}{\partial x}\frac{\partial[\phi_m(x,z)\psi_m(x)]}{\partial x}\right\}\\ &= \left(E-\frac{\hbar^2 k_y^2}{2m_2''}\right)\psi_n(x).\end{aligned} \tag{2.34}$$

Equation (2.34) serves as the generalization of a similar one derived from the restricted case of silicon (100) MOSFETs in [49]; however, in that treatment the summation terms involving coupling $W_{nm}(x)$ among the subbands, arising due to the nonalignment of device axes with ellipsoid axes, were absent. These terms represent Zener tunneling between the subbands. Ignoring higher-order corrections, they are proportional to the inter-subband dipole μ_{nm} and the local field generated by the variation in the subband eigenvalues along the transport direction, $\frac{\partial \epsilon_m(x)}{\partial x}$. Contributions to $W_{nm}(x)$ arise both from the cross terms in the potential energy, as in a channel with varying cross section, as well as from the cross terms in the kinetic energy due to the nonaligned device and ellipsoidal axes. While the latter terms never arise for silicon (100) devices, the effect of the former cross terms has been included in the coupled-mode-space approach of [52] to analyze the effects of channel access geometry in nanoscale silicon n-MOSFETs.

2.4.3 Simplifications in UTB Devices

In an UTB SOI MOSFET with uniform channel thickness (Fig. 2.2), the electrostatic potential is separable along the confinement and transport (channel) directions,

$$W(x,z) = U(z) + V(x). \tag{2.35}$$

The channel potential $V(x)$ simply shifts the confining potential $U(z)$, but does not alter the shape of the models as we move from one point to next along the transport direction. This implies

$$\frac{\phi_j(x,z)}{\partial x} = \frac{\partial^2 \phi_j(x,z)}{\partial x^2} = 0. \tag{2.36}$$

The x dependence of $\phi_j(x,z)$ is eliminated since it selectively diagonalizes the confinement Hamiltonian which now depends only on $U(z)$ and therefore x independent. In this special case (2.34) simplifies considerably and becomes

$$\left[-\frac{\hbar^2}{2m_1'}\frac{\partial^2 \psi_n(x)}{\partial x^2} + \epsilon_n(x)\,\psi_n(x)\right] + \sum_m W_{nm}(x)\,\psi_m(x) = \left(E - \frac{\hbar^2 k_y^2}{2m_2''}\right)\psi_n(x). \tag{2.37}$$

Additionally, (2.37) allows us to write

$$\epsilon_n\left(x+\frac{m_{33}}{m_{31}}z\right)=\epsilon_n+V\left(x+\frac{m_{33}}{m_{31}}z\right). \tag{2.38}$$

ϵ_n is the n-th subband energy arising from diagonalizing the x-independent confinement potential $U(z)$, and is unvarying along x. The only role of the channel potential $V(x)$ in this special separable case is to shift the bottom of the subbands:

$$\left[-\frac{\hbar^2}{2m_1'}\frac{\partial^2}{\partial x^2}+V(x)\right]\psi_n(x)+\sum_{m\neq n}V_{nm}(x)\psi_m(x)=\left(E-\epsilon_n-\frac{\hbar^2k_y^2}{2m''_2}\right)\psi_n(x). \tag{2.39}$$

In (2.39), the coupling between models is still present and represents the Zener tunneling between subbands due to the cross terms in the kinetic energy, arising from the arbitrary orientation of the ellipsoids. The tunneling is negligible if

$$|V_{nm}(x)|<<|\epsilon_n(x)-\epsilon_m(x)|,$$

which amounts to

$$\left|\frac{m_{33}}{m_{31}}\frac{\partial V(x)}{\partial x}\mu_{nm}\right|<<|\epsilon_n-\epsilon_m|. \tag{2.40}$$

It can be shown, in general, that $\left|\frac{m_{33}}{m_{31}}\right|<1$ and $|\mu_{nm}|<t_b$, so that ignoring the Zener tunneling amounts to

$$t_b\left|\frac{\partial V(x)}{\partial x}\right|<<|\epsilon_n-\epsilon_m|,$$

where t_b is the body thickness. This inequality can be restated as follows: *if at any point along the channel, the channel-directed potential drop between the confining planes and sampled along the ellipsoidal principal-axis direction is much smaller than the corresponding inter-subband separation, the coupling between modes can be safely ignored.* Indeed, the coupling arose precisely because the channel potential varies along the ellipsoid's principal axis, which determines the confinement potential generating the transverse subbands (2.19). For Si (100) the coupling does not exist since there is no drop in the channel-directed potential along the direction of the ellipsoidal axis.

A simple estimate tells us the conditions under which this Zener tunneling is negligible. The smallest inter-subband separation is given roughly by $(3^2-1^2)\frac{\hbar^2\pi^2}{2m_{33}t_b^2}$.

The channel potential drop along the ellipsoid between the confinement planes depends on the local field. For a linear channel-directed potential profile, this drop is given by $\frac{Vt_b}{L}$, where V is the applied bias and L is the channel length. Near the top of the barrier, the field is smaller and the drop is given roughly by $\frac{Vt_b^2}{2L^2}$. For a 10nm channel length with a 0.6V applied bias and for $m_{33} = 0.1$, the drop is the channel potential along the confinement direction is negligible compared to the subband separation, provided the body thickness is smaller than about 5nm (for the average potential, and 10nm for the top of the barrier potential).

In the absence of inter-subband Zener coupling, the corresponding decoupled-mode-space equation finally becomes

$$-\frac{\hbar^2}{2m'_1}\frac{\partial^2\psi_n(x)}{\partial x^2} + V(x)\psi_n(x) = \left(E - \epsilon_n - \frac{\hbar^2 k_y^2}{2m''_2}\right)\psi_n(x). \tag{2.41}$$

In this section we have demonstrated an exact mathematical transform that considerably simplifies the quantum-mechanical treatment of electronic transport in a nanoscale n-MOSFET with arbitrarily oriented transport, width, and thickness directions w.r.t. the CB ellipsoidal. The conditions under which this simplified equation is valid are:

1. an unvarying device cross section that allows us to separate the confinement and transport potentials and

2. a channel-directed potential that varies slowly enough that there is no Zener tunneling between transverse subbands.

A simplified mapping operation is performed by two unitary operations in (2.17), (2.29) that map each conduction-band ellipsoid into an equivalent regular ellipsoid whose principal axes are oriented along the device coordinate axes X, Y, and Z with corresponding effective masses m'_1, m''_2, and m_{33}, respectively. For separable potentials with coincident device and ellipsoidal axes (i.e., no cross terms in the kinetic energy), the normal mode $\phi_m(z)$ selectively diagonalizes the confinement Hamiltonian $-\partial^2/\partial z^2 + U(z)$. Non coincident device and ellipsoidal directions introduce additional

cross terms in the Hamiltonian which are diagonalized by $e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_m(z)$. Finally, in the general case of an arbitrary potential $W(x,z)$, the transverse modes $\phi_m(x,z)$ depend on both x and z. In Sec. 2.5, we present several applications of the transformation developed here.

2.5 Application to Ultra-thin-body Si and Ge MOSFETs

We now demonstrate the usefulness of our generalized treatment by applying it to several technologically important materials and wafer orientations. Until recently, nearly all quantum simulations of MOSFETs were performed for silicon (100) wafers with transport along $< 100 >$. We will discuss several non-trivial cases here–effective masses for UTB silicon and germanium MOSFETs fabricated on (100), (111), and (110) wafers.

2.5.1 Transformation Matrices for the DCS

First, we will evaluate the transformation matrix $\Re_{C\leftarrow D}$ for various wafer orientations.

(100) **Wafers**

This is the most common wafer orientation used for the fabrication and simulation of nanoscale silicon MOSFETs. In these devices the inversion layer electron are confined along the [001] direction which is the Z axis for the DCS. The transport and width directions (X and Y axes) are along [100] and [010], respectively, so that the basis vectors are $\hat{k}_1 = (100)$, $\hat{k}_2 = (010)$, and $\hat{k}_3 = (001)$. Since the columns of

the transformation matrix $\Re_{C\leftarrow D}$ are components of $\hat{k}_1$, $\hat{k}_2$, and $\hat{k}_3$, the matrix itself becomes and identity matrix,

$$\Re_{C\leftarrow D}^{(001)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.42}$$

(111) **Wafers**

For MOSFETs fabricated on (111) wafers, the gate electric field confines the inversion layer carriers along the [111] crystallographic orientation. We choose the transport direction along $[\bar{2}11]$ and the width direction along $[0\bar{1}1]$. These crystallographic orientations represent the Z, X and Y axes for the device. The basis vectors for the DCS are $\hat{k}_1 = \left(-\frac{2}{\sqrt{6}}\frac{1}{\sqrt{6}}\frac{1}{\sqrt{6}}\right)$, $\hat{k}_2 = \left(0 - \frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}\right)$, and $\hat{k}_3 = \left(\frac{1}{\sqrt{3}}\frac{1}{\sqrt{3}}\frac{1}{\sqrt{3}}\right)$, respectively. The rotation matrix $\Re_{C\leftarrow D}$ becomes

$$\Re_{C\leftarrow D}^{(111)} = \begin{bmatrix} -\frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix}. \tag{2.43}$$

(110) **Wafers**

In this case the inversion layer electrons are confined along the [110] crystallographic orientation, which is the Z axis. We choose the transport direction, X, along [001] and the width direction, Y, along $[1\bar{1}0]$, so that the corresponding unit vectors are $\hat{k}_1 = (001)$, $\hat{k}_2 = \left(\frac{1}{2} - \frac{1}{2}0\right)$, and $\hat{k}_3 = \left(\frac{1}{2}\frac{1}{2}0\right)$, and the rotation matrix $\Re_{C\leftarrow D}$:

$$\Re_{C\leftarrow D}^{(110)} = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 1 & 0 & 0 \end{bmatrix}. \tag{2.44}$$

2.5.2 Transformation matrices for the ECS

There are two types of valleys in indirect band-gap semiconductors–the sixfold degenerate Δ valleys and the eightfold degenerate Λ valleys. These are classified according the orientation of the major axes of the constant energy ellipsoids along the $< 100 >$ or $< 111 >$ directions. In bulk silicon the Δ valleys are energetically lower than the Λ valleys; consequently, the conduction-band electron populate the Δ valleys, while the Λ valleys can be ignored for transport simulations. The opposite is true for bulk germanium, where the Λ valleys are energetically lower than their Δ counterparts, and therefore, the states near the CB edge are of the former type. Interesting phenomena can be observed when quantum confinement is present, since there, in addition to the energy of the bulk band edge, inversion layer thickness and the confinement direction effective mass determine which valley forms the energetically lowest subband. In this subsection, we will evaluate the transformation matrix, $\Re_{E\leftarrow C}$, for these valleys.

Δ-Valleys

Figure 2.3 shows the three doubly degenerate constant energy Δ valley conduction-band ellipsoids. The basis vectors for the ellipsoid coordinate system are unique for each ellipsoid, with $\hat{k}_{||}$ along the major axis, and $\hat{k}_{\perp 1}$ and $\hat{k}_{\perp 2}$ along two orthogonal minor axes. For ellipsoid 1, for example, $\hat{k}_{||} = (100)$, $\hat{k}_{\perp 1} = (010)$, and $\hat{k}_{\perp 2} = (001)$. For each ellipsoid there is a unique transformation matrix $\Re_{E\leftarrow C}$, the rows of which are the components of $\hat{k}_{||}$, $\hat{k}_{\perp 1}$, and $\hat{k}_{\perp 2}$. For ellipsoid 1, $\Re_{E\leftarrow C}$ becomes

$$\Re_{E\leftarrow C}^{\Delta_1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.45}$$

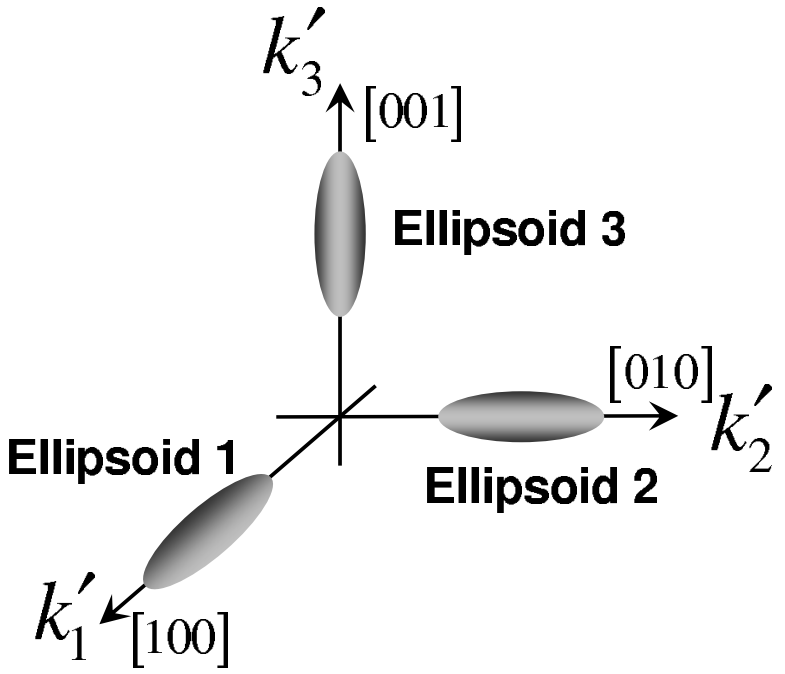

Fig. 2.3. Conduction band constant energy ellipsoids along Δ. Each of the three ellipsoids is doubly degenerate. In silicon, such valleys form the conduction band minima.

Similar expressions can be obtained for ellipsoids 2 and 3:

$$\Re_{E\leftarrow C}^{\Delta_2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \tag{2.46}$$

and

$$\Re_{E\leftarrow C}^{\Delta_3} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{2.47}$$

Λ-Valleys

Figure 2.4a shows the eight-half Λ-valley ellipsoids with centers at the equivalent L points at the surface of the first Brillouin zone. Since the diagonally opposite L

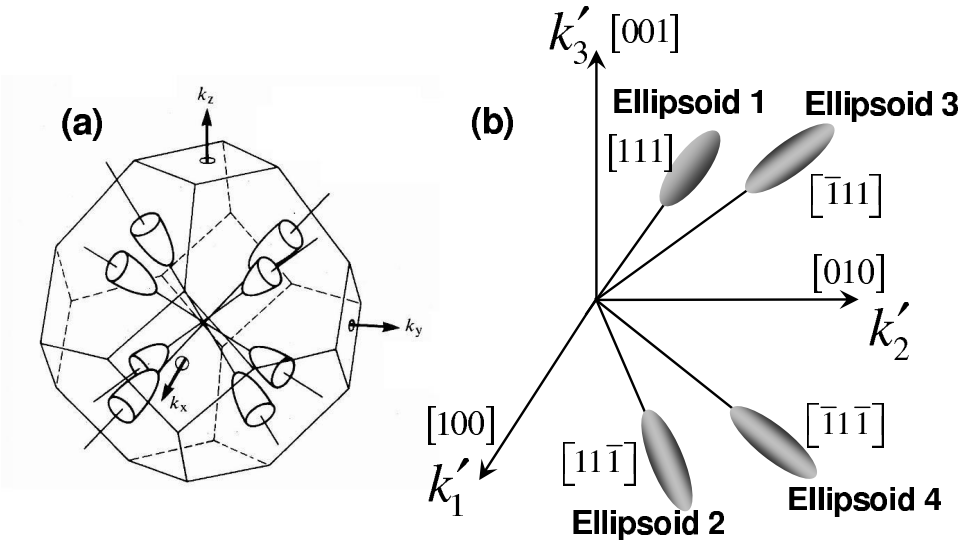

Fig. 2.4. Conduction band constant energy ellipsoids around the L points in the first BZ. (a) The major axis of the eight half ellipsoids are along Λ. (b) Since the centers of the diagonally opposite half ellipsoids are one wavevector apart, they can be combined into four equivalent full ellipsoids. In bulk Ge they form the conduction band edge.

points are one reciprocal-lattice vector apart, they can be combined into four full ellipsoids, as shown in Fig. 2.4b. The major axes of ellipsoids 1-4 are along [111], [11$\bar{1}$], [$\bar{1}$11], and [$\bar{1}$1$\bar{1}$], respectively. The rotation matrix $\Re_{E\leftarrow C}^{\Lambda_1}$ for ellipsoid 1 can be written from the components of the basis vectors $\hat{k}_{||}$, $\hat{k}_{\perp 1}$, and $\hat{k}_{\perp 2}$ in the CCS, and is

$$\Re_{E\leftarrow C}^{\Lambda_1} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \end{bmatrix}. \tag{2.48}$$

Similar matrices for ellipsoids 2-4 can be readily calculated, and are

$$\Re_{E\leftarrow C}^{\Lambda_2} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \end{bmatrix}. \tag{2.49}$$

$$\Re^{\Lambda_3}_{E\leftarrow C} = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix}. \tag{2.50}$$

and

$$\Re^{\Lambda_4}_{E\leftarrow C} = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix}. \tag{2.51}$$

2.5.3 Evaluating Effective Masses

Using the results presented above, we can now calculate the effective masses for both Δ and Λ valleys in the conduction band, and for different wafer orientations. In Table 2.1, the results are given in terms of bulk m_l and m_t. The following steps were used to obtain the results:

1. for the given wafer orientation, choose the appropriate $\Re_{C\leftarrow D}$ from (2.42–2.44).

2. for the given valley type (Δ or Λ) and for each of the conduction-band ellipsoids, choose the appropriate $\Re_{E\leftarrow C}$ from (2.45–2.47) and (2.48–2.51).

3. From (2.7), evaluate $\Re_{E\leftarrow D}$, and then using (2.10) find the effective mass tensor, $\left[M_D^{-1}\right]$, in the device coordinate system.

4. The confinement effective mass, $m_Z\,(= m_{33})$, is directly obtained from $\left[M_D^{-1}\right]$. The transport effective mass, $m_X\,(= m'_1)$, is calculated from (2.26). Finally, the effective mass along the width direction, $m_Y\,(= m''_2)$, is calculated from (2.31).

Table 2.2 shows the m_l and m_t values for the Δ and Λ valleys of bulk silicon and bulk germanium. The effective masses of the lowest valleys for each materials, i.e., Δ valleys for silicon and Λ valleys for germanium, were obtained from cyclotron experiments [89, 90, 94, 95], while those for the upper valleys are obtained from empirical

pseudopotential calculations [96,97]. Using the effective masses in Table 2.2 and the expressions for m_X, m_Y, and m_Z in Table 2.1, the effective masses for the Δ valleys and Λ valleys in silicon and germanium can be calculated for the corresponding wafer orientation.

2.6 Discussion

In this chapter we introduced a generalized effective mass equation framework for the quantum-mechanical simulation of cubic semiconductor n-MOSFETs. It is well known that when one or more principal axes of the conduction-band constant energy ellipsoid do not coincide with the device coordinate axes, X, Y, and Z, the solution of the effective-mass equation is non-trivial. Our treatment simplifies if the electrostatic potential is separable, valid if the cross section is unvarying along the transport direction. Further simplifications occur for thin-body MOSFETs, where the cross terms in the kinetic energy arising due to the arbitrarily oriented conduction-band ellipsoids do not couple the various subbands. The unitary operation in (2.17) decouples the energy along the confinement direction, Z, from the energy associated with the carrier's motion in the transport plane. A physical picture of the result of this operation is schematically present in Fig. 2.5, where we see that this transforms the arbitrarily oriented ellipsoid (top) in such a way, that the resultant ellipsoid (middle) becomes symmetric across the $X-Y$ plane, and therefore, one principal axis of the transformed ellipsoid becomes aligned with the Z axis. Consequently, in (2.20) we see that the $X-Y$ plane energy is decoupled from the quantum confinement problem. By substituting $-i\partial/\partial x \rightarrow k_x$ in (2.21), we find the constant energy contours for the electrons in the $X-Y$ plane as ellipses. In general, their principal axes are not aligned with X and Y, and the Hamiltonian in (2.24) remains complicated due to the presence of the first derivative. The second unitary operation in (2.29) transforms these ellipses in such a way that the transformed constant energy elliptical contours have their principal axis aligned along the transport direction. This is also schemat-

(Wafer) /[Transport]/ [Width]	Valley	m_X	m_Y	m_Z	Deg.
(001)/[100]/[010]	Δ	m_t	m_t	m_l	2
		m_l	m_t	m_t	2
		m_t	m_l		2
	Λ	$m_t\frac{2m_l+m_t}{m_l+2m_t}$	$\frac{m_l+2m_t}{3}$	$\frac{3m_lm_t}{2m_l+m_t}$	4
(111)/[$\bar{2}$11]/[0$\bar{1}$1]	Δ	$\frac{2m_l+m_t}{3}$	m_t	$\frac{3m_lm_t}{2m_l+m_t}$	2
		$\frac{2}{3}m_t\frac{2m_l+m_t}{m_l+m_t}$	$\frac{m_l+m_t}{2}$	$\frac{3m_lm_t}{2m_l+m_t}$	4
	Λ	m_t	m_t	m_l	1
		$\frac{8m_l+m_t}{9}$	m_t	$\frac{9m_lm_t}{8m_l+m_t}$	1
		$\frac{m_t}{3}\frac{8m_l+m_t}{2m_l+m_t}$	$\frac{2m_l+m_t}{3}$		2
(110)/[001]/[0$\bar{1}$0]	Δ	m_t	$\frac{m_l+m_t}{2}$	$\frac{2m_lm_t}{m_l+m_t}$	4
		m_l	m_t	m_t	2
	Λ	$\frac{m_l+2m_t}{3}$	m_t	$\frac{3m_lm_t}{m_l+2m_t}$	2
		$\frac{3m_lm_t}{2m_l+m_t}$	$\frac{2m_l+m_t}{3}$	m_t	2

Table 2.1
Transport, width and confinement effective masses and subband degeneracies for three different technologically important semiconductor wafer orientations.

Table 2.2
Transverse and longitudinal effective masses for the Δ- and Λ-type valleys in silicon and germanium.

Material	Valley	m_l	m_t
Silicon	Δ	0.91	0.19
	Λ	1.7	0.12
Germanium	Δ	0.95	0.2
	Λ	1.64	0.08

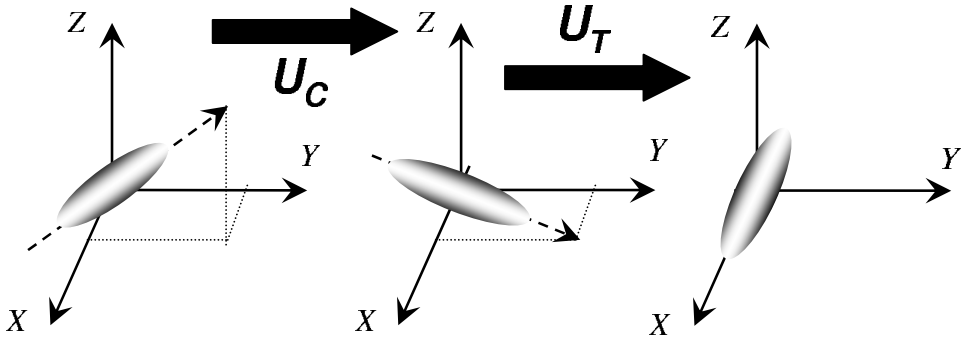

Fig. 2.5. The effects of unitary transformations of (2.17) and (2.29). The first operator transforms the arbitrarily oriented CB ellipsoid into an equivalent one, which is symmetric across $k_x - k_y$ plane. The second unitary operation transforms it into the regular ellipsoid with its principal axes along X, Y and Z. The density-of-states effective mass and group velocity of each k-state is conserved.

ically presented in Fig. 2.5, where we see that the bottom ellipsoid has its principal axes along the device axes. In summary, since the top and the bottom ellipsoids in Fig. 2.5 are exactly equivalent, the well-defined effective masses, m_X, m_Y, and m_Z, determined from the bottom ellipsoid describe the effective masses of the original ellipsoid. It can be observed from Table 2.1 that for each row $m_X m_Y m_Z = m_l m_t^2$, which ensures that the volume of the transformed ellipsoid is the same as that of the original one, and therefore, the density of states is conserved. Additionally, these unitary operations change only the phase velocity of the states and thus the group

velocity of the carriers, determined from the gradient of the $E(\vec{k})$, is also conserved. Conservation of the density of states and carrier group velocity ensures that the results obtained by performing quantum simulation of a MOSFET using m_X, m_Y, and m_Z gives the same result as by treating the nondiagonlal effective mass tensor, $\left[M_D^{-1}\right]$. The density of states effective mass per valley can be readily obtained from Table 2.1, using the expression, $m_d = \sqrt{m_X m_Y}$.

2.7 Summary and Conclusion

The simple technique for mapping arbitrarily oriented conduction-and constant energy ellipsoids into regular ellipsoids, where the principal axes are aligned along the device axes, allows us to perform quantum transport simulation in ultra thin body n-MOSFETs with any channel materials and arbitrary wafer orientations. The effective masses presented in Table 2.1 can be readily used in any quantum-mechanical simulator to determine I-V and C-V characteristics.

3. ASSESSMENT OF GERMANIUM N-MOSFETS BY NEGF SIMULATION

In this chapter, quantum simulations of ultra-thin-body (UTB), double-gate (DG), end of the ITRS-2001 roadmap germanium n-MOSFETs are performed using the Non-Equilibrium Green's Function (NEGF) formalism within the generalized effective mass equation framework presented in Ch. 2. Ballistic simulations show that Ge (111) n-MOSFETs suffer from high source-to-drain tunneling in the off-state, and low semiconductor capacitance in the on-state. However, devices fabricated on Ge (100) wafers perform better compared to their silicon counterparts. Design optimization studies show that a stiff tolerance for body thickness variations and a super-steep source-drain doping gradient are necessary to optimize the device performance. Finally, it was observed from quantum scattering simulations that the source-drain series resistance limits the otherwise near-ballistic intrinsic device operation.

3.1 Introduction

Germanium is an interesting candidate for ultimate CMOS because of its excellent transport properties. Compared to bulk silicon, bulk germanium has two times and four times higher mobilities for electrons and holes, respectively. Several experimental groups have recently reported devices based on germanium [32, 36, 37]. The quantum simulation of germanium MOSFETs is a challenge because for arbitrary surface orientation and transport direction, the effective mass tensor becomes a 3×3 non-diagonal tensor and the effective mass equation becomes complicated. This problem was partially addressed in [77], which treated quantum confinement, not transport. In this chapter, we use the general theoretical approach, introduced

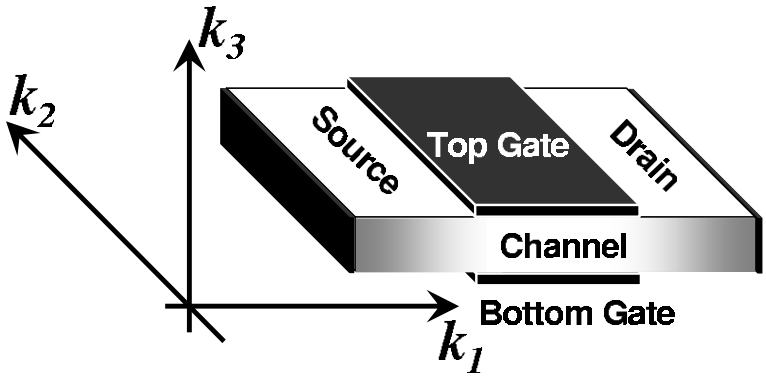

Fig. 3.1. The ultra-thin-body, double-gate device structure. The device coordinate system consists of orthogonal axes k_1, k_2 and k_3, along transport, width and thickness directions, respectively. In this work, confinement along k_3 and transport along k_1 are treated quantum mechanically. The width direction, k_2, is considered to be much longer, so plane waves were assumed. An equivalent gate insulator thickness (EOT) of 0.6 nm [32] and channel thickness of 2.5nm were assumed in this work.

in Chapter 2, for the quantum mechanical simulation of n-MOSFETs within the Non Equilibrium Green's Function (NEGF) formalism [72], and demonstrate its application by performing a scaling study for the end of the ITRS germanium MOSFETs [6, 7]. We show that ultra-thin-body germanium MOSFETs fabricated on the (111) wafers are not scalable due to source-to-drain tunneling problems in the off-state and low semiconductor capacitance in the on-state. Devices fabricated on germanium (100) wafers, however, perform better than their silicon counterparts. We perform both ballistic and scattering NEGF simulations of such devices to determine the ultimate and realistic performance limits, and to identify key design and fabrication issues. In this work, we have treated only the eight conduction-band half ellipsoids around the equivalent L points in the first Brillouin zone. The higher energy valleys, such as those along the delta directions, may have important effects in charge transport and subband occupation, which we will explore in Sec. 6.2 of chapter 6 (page 79).

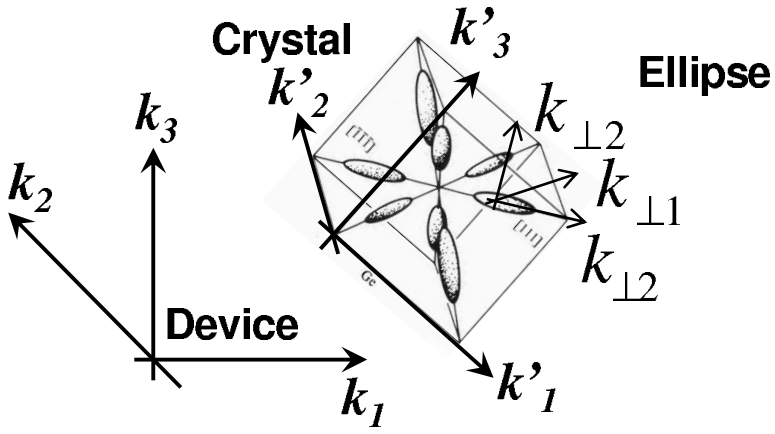

Fig. 3.2. Three orthogonal coordinate systems, device, crystal and ellipse, are shown. The effective mass tensor is diagonal only in the ellipse coordinate system, and becomes a full matrix in the device coordinates. The resulting effective mass equation becomes complicated. The technique presented in Ch. 2 diagonalizes the effective mass tensor in the device coordinate system, thereby decoupling energies along k_1, k_2 and k_3 [72].

3.2 Approach

The effective mass tensor is diagonal when the device coordinate axes (Fig. 3.1) are along the major and minor axes of each of the constant-energy ellipsoids. This condition is satisfied for n-MOSFETs fabricated on silicon (001) wafers, but not for devices on germanium (001) or silicon (111) wafers. The general scenario is presented in Fig. 3.2, where we see three different coordinate systems: device, crystal and ellipse. To solved this problem within the EME approximation, we:

1. identify the rotation matrix for the coordinate transformation between the ellipse and device coordinate systems,

2. use it to get the EMT and the EME in the device coordinate system, and

3. use two unitary transformations to decouple the energies in the EME along the device coordinate axes [72].

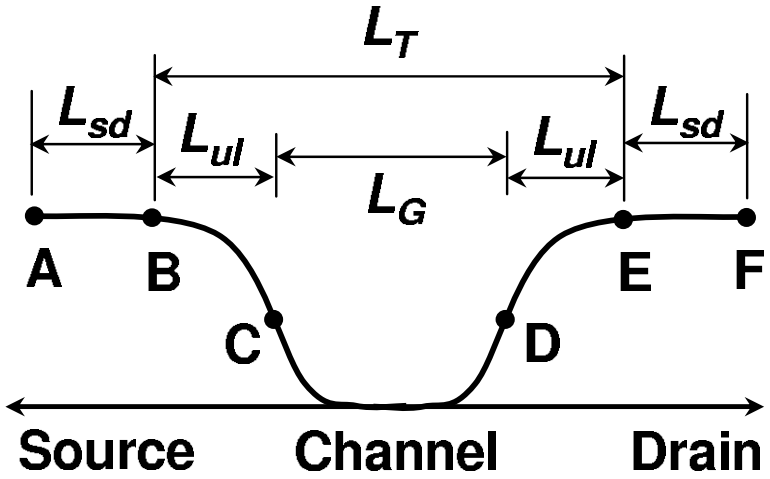

Fig. 3.3. Sketch showing the doping profile along the device and the key length parameters. The source and drain contacts are at A and F, respectively. The region between A and B (E and F) is the source (drain) and is doped at N_{sd} ($10^{20} cm^{-3}$). Between B and E, the doping falls off with a Gaussian profile in both directions. L_T is a measure of the channel length. Between C and D, the length L_G (printed gate length) is under the direct electrostatic control of the gate electrodes. The region between B and C (D and E) is the source/drain under-lap and improves short-channel effects.

As demonstrated in Ch. 2, this method maps the arbitrarily oriented constant-energy ellipsoid in to a regular ellipsoid with the principal axes aligned along the device coordinate system, and thus allows use of all the techniques developed so far for silicon (001) n-MOSFET simulation. We made necessary changes in the 2D quantum simulation nanoMOS 2.5 [50] and performed ballistic and scattering simulations using the NEGF formalism.

To ensure electrostatic integrity, we choose the UTB double gate (DG) MOSFET as our model device (Fig. 3.1), but similar results are expected for other device geometries. The device dimensions and the optimized parameters for the nominal device are given in Table 3.1 (page 49), and in Fig. 3.3 the doping profile and the key device lengths are shown. The source/drain (S/D) under-lap shown here plays an important role in device operation by suppressing source-to-drain tunneling and

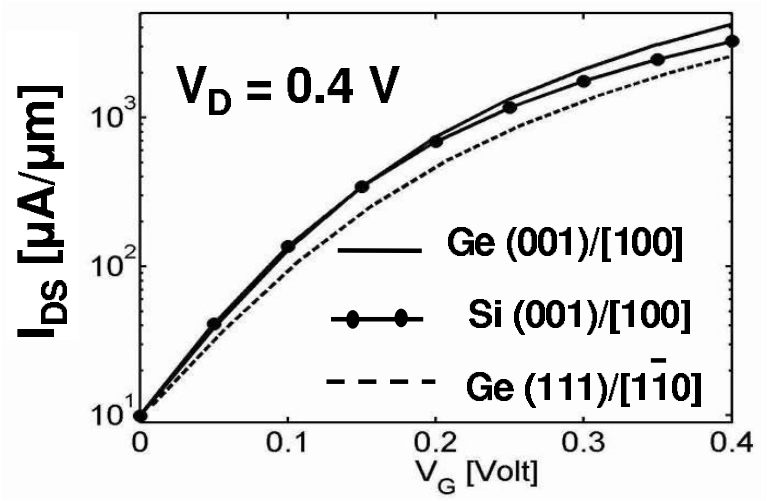

Fig. 3.4. Quantum ballistic (NEGF) $I_D - V_G$ simulation at $V_D = V_{DD}$, for different channel materials and surface orientations. The first indices specify wafer orientation and the second is the transport direction. (001) oriented Ge and Si have almost identical transport effective masses, 0.16 m_0 and 0.19 m_0, respectively, and hence they have identical sub-threshold slope. However, because of a very low transport effective mass (0.08 m_0), for Ge (111) tunneling current is dominant in the off-state, which degrades the sub-threshold slope. The off-currents are adjusted to the ITRS limit, 10 $\mu A/\mu m$.

improving short-channel effects. Effects of the source-drain doping gradient and the gate under-lap are explored in this work in order to optimize the device design.

3.3 Results

The ballistic performance of germanium (001)/[100] and germanium (111)/[1$\bar{1}$0] devices are first compared. Here the first and second sets of indices specify the wafer orientation (quantum confinement) and the transport direction, respectively. The $I_D - V_G$ plot (Fig. 3.4) at high drain bias shows degraded sub-threshold slope due to strong source-to-drain tunneling in the Ge (111) devices. The output characteristics (Fig. 3.5) show that the on-current for Ge (111) devices suffer. This happens because the low density-of-states (DOS) in their conduction band lowers the semiconductor

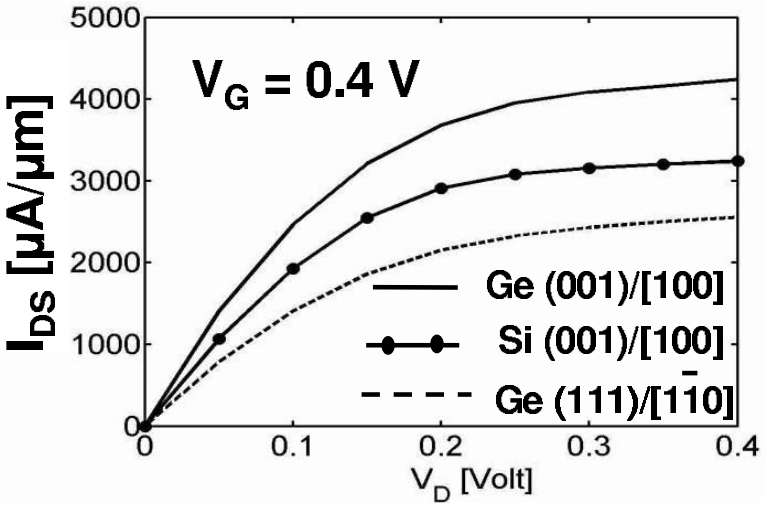

Fig. 3.5. Output characteristics for the devices in Fig. 3.4 at $V_G = V_{DD}$. The on-current degradation for the Ge (111) device is due to very low density-of-states ($m_{DOS} = 0.08m_0$ and valley degeneracy one). Maximum ballistic on-current is delivered by Ge (001), which has $m_{DOS} = 0.3m_0$ per valley, and a valley degeneracy of four. Performance of Si (001) lies in between.

capacitance, which lowers the overall gate capacitance. The silicon (001) device performs in between the germanium limits. These figures show that best performance can be obtained from germanium (001)/[100] devices.

The optimization procedure for choosing the device dimensions is summarized in Fig. 3.6. Using the approach of [56], we calculated the ballistic on-current for the nominal device with different S/D under-laps. For each under-lap, the off-current for the worst case device was adjusted to $10\mu A/\mu m$ by varying the gate work function, Φ_m, which is also plotted. The worst case device is assumed to have 5% thicker body and 10% shorter gate length due to process variation. Results for two different channel doping gradients, $G = 1$ and $G = 2$ nm/dec, are shown. Both gradients deliver almost the same maximum on-current, but, $G = 1$ nm/decade gives a shorter device and thereby less series resistance. In this figure, for each L_T , the I_{OFF} for the corresponding worst case device was adjusted to $10\mu A/\mu m$ by varying Φ_m. Therefore, when the Φ_m for the optimum L_T was used to simulate the nominal device,

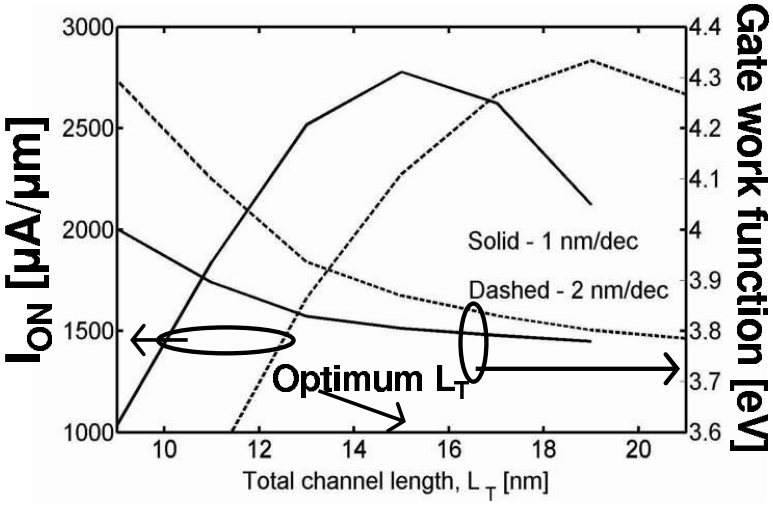

Fig. 3.6. Device design optimization technique. Left axis: The quantum ballistic I_{ON} vs. L_T is plotted for a nominal case Ge (001)/[100] device. The total channel length is the sum of gate length, L_G, and under-lap, L_{ul}, on both sides of the gate. Two different doping gradients, 1 nm/dec and 2 nm/dec are considered. Both offer almost the same optimum I_{ON}, but the optimum L_T for the steeper doping gradient is shorter. Therefore, the optimum device is chosen to have 1nm/dec S/D doping gradient and 3 nm under-lap. Right axis: Gate work function needed to adjust I_{OFF} to 10 $\mu A/\mu m$ for the worst case device.

it is naturally ensured that the worst case device operates within the ITRS specified limit for I_{Off}. The optimum values for device parameters are given in Table 3.1, and the intrinsic performance metrics for the ballistic nominal device are summarized in Table 3.2.

The quantum ballistic I-V characteristics are shown in Figs. 3.7 and 3.8. The off-current is lower than the ITRS specification (Fig. 3.7) since the Φ_m used here assures that if the body thickness increases 5% and gate length decreases 10% due to process variation, the I_{OFF} still remains below the ITRS limit. Figure 3.9 plots I_{OFF} as a function of gate length, L_G, with channel thickness, t_{body}, as a parameter. Here the ballistic I_{OFF} from NEGF simulation is plotted as a function of printed gate length, at three different channel thicknesses. All other simulation parameters

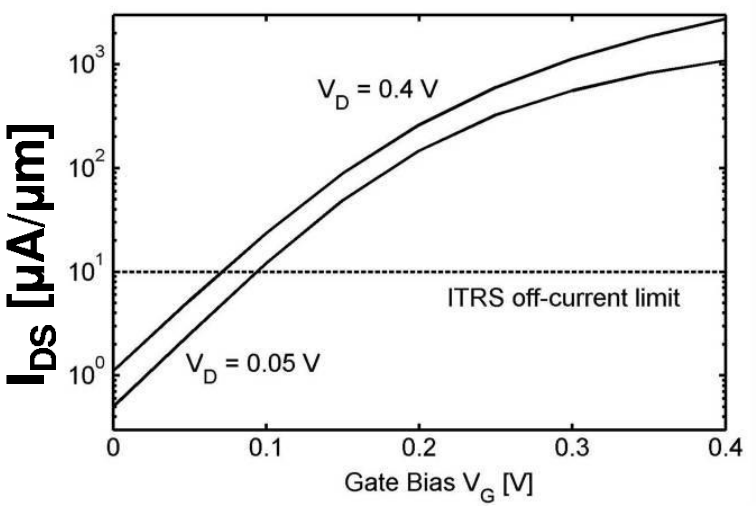

Fig. 3.7. Ballistic NEGF $I_D - V_G$ characteristics at high and low drain bias for the nominal device; the gate length is L_G = 9nm and the body thickness is t_{body} = 2.5nm. The under-lap, doping gradient and gate work-function were chosen from the device optimization technique shown in Fig. 3.6. The off-current is lower than the ITRS limit since the work function was adjusted to obtain $I_{OFF} = 10\mu A/\mu m$ for the worst case device.

are the same as in Table 3.1. Variation in I_{OFF} can be observed due to changes in both L_G and t_{body}, but, the effect of the latter is more serious. The plots are equivalent to the V_T roll-off plot for bulk MOSFETs. They represent the effects of process variations in fabricating such devices. Here the worst case device operates exactly at the ITRS limit for I_{OFF} and for the nominal device it remains within the limit for L_G as short as 6nm. Also, the thinner body results in higher V_T and therefore, the on-current is degraded.

Figure 3.10 shows results with scattering for the Ge (100) n-MOSFET. A doping dependent Caughy-Thomas mobility model was used to calculate the local mobility for the doping profile shown in Fig. 3.3. A simple phenomenological model based on the Büttiker-probe treatment was used for scattering simulation within the NEGF formalism. This model has been successful for silicon MOSFETs [51]. Since the channel of this $L_G = 9nm$ MOSFET is near ballistic, increasing the channel mobility

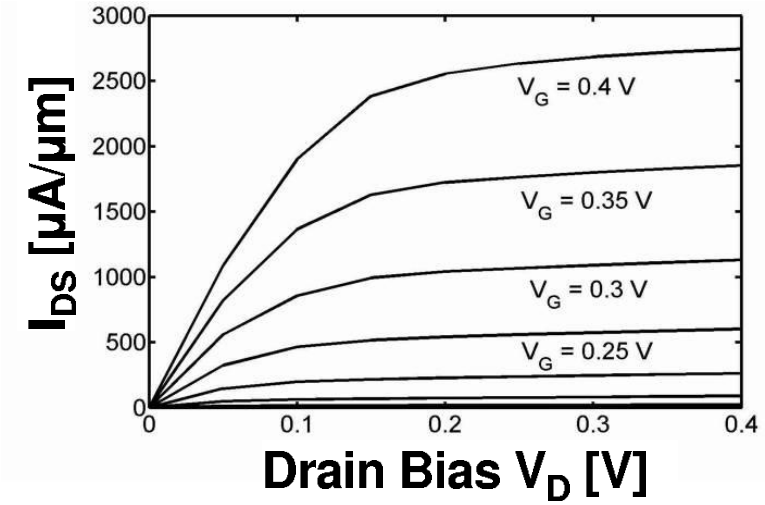

Fig. 3.8. Output characteristics from quantum ballistic (NEGF) simulation for the nominal device.

does not improve the performance significantly; however, the mobility in the highly doped S/D regions has an important effect on performance enhancement.

3.4 Discussion

Figures 3.4 and 3.5 present the general characteristics of bandstructure effects in ballistic nanoscale MOSFETs. Low effective mass along the channel direction is necessary for the channel material in order to obtain high injection velocity at the source end of the channel. However, high density-of-states (DOS) is also needed to induce sufficient charge at the top of the source-channel barrier. For the channel materials with low DOS effective mass, such as germanium (111) wafers, a greater fraction of applied gate voltage is consumed to shift the potential at the top of the barrier; therefore, less voltage is available to drop across the gate insulator. As a result, the gate cannot efficiently induce charge and the current suffers. The degradation of I_{ON} in low DOS channel materials has been discussed in detains by Solomon and Laux [98].

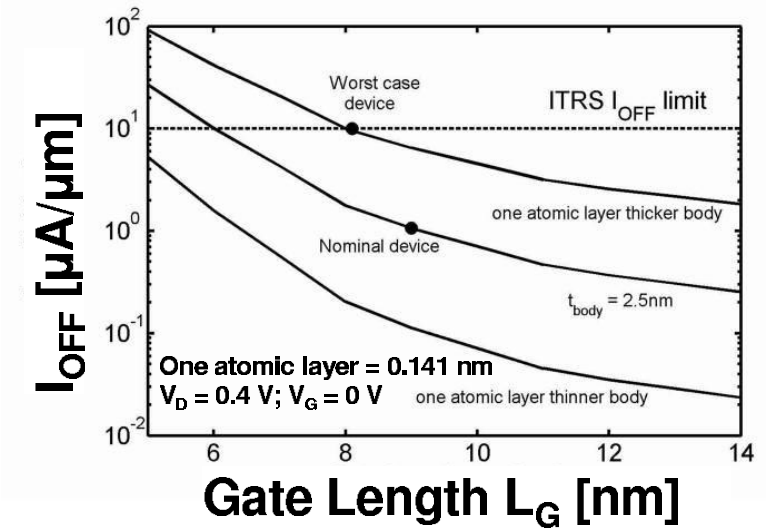

Fig. 3.9. Effects of process variation. It can be seen that in UTB MOSFETs, the off-current (or equivalently threshold voltage) is very sensitive to the body thickness variation.

The S/D under-lap plays an important role in device scaling in two ways. First, the source-channel barrier becomes thicker with under-lap and this suppresses tunneling. Second, the source and drain regions are moved farther apart, which improve short-channel effects. However, since the potential in this region has to be modulated by the fringing field of the gate, the modulation is not as efficient as the channel region directly under the gate; therefore, we can see in Fig. 3.6 that after attaining its peak value, I_{ON} decreases with under-lap.

In UTB devices, threshold voltage, V_T, is a strong function of the body thickness. In order to fabricate operational devices with a body thickness of 4nm or less, the process variation of thickness has to be within a single atomic layer. Thin body devices also suffer from the series resistance in the S/D regions. Scattering simulation has shown that the channel in the nanoscale device is near ballistic, and the degradation of carrier mobility in the S/D regions causes degradation of I_{ON}. The higher mobility in germanium is attractive in this regard, but, currently the key challenge of fabricate germanium n-MOSFETs is that the S/D regions cannot

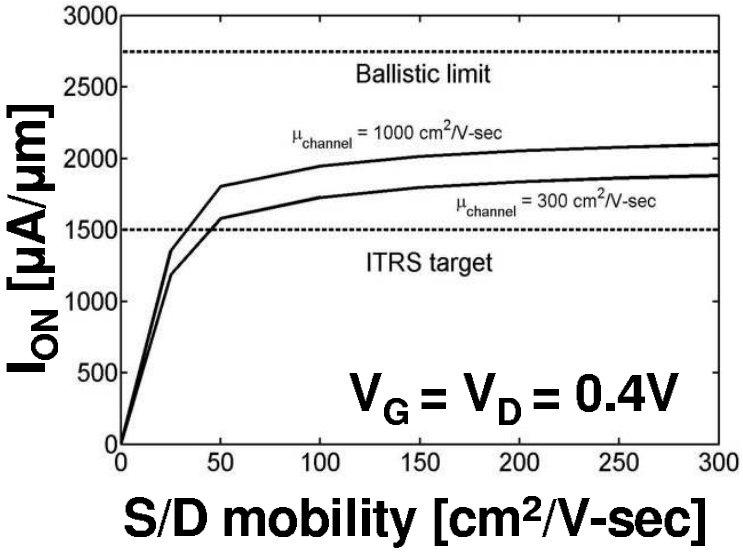

Fig. 3.10. Quantum scattering simulation for the nominal device. On-current vs. source/drain mobility is plotted for two different channel mobilities. The key observation is, increasing channel mobility by more than 300% does not enhance the I_{ON} accordingly, since, for this L_G = 9nm MOSFET, the channel is near ballistic. The current is limited by the series resistance in the highly doped, low mobility S/D regions.

be doped a sufficiently high density due to low solid-solubility for the n-type impurities. Technological advancement is necessary to overcome this challenge before germanium CMOS can be viable.

3.5 Summary and Conclusion

Ballistic NEGF Simulations show that UTB DG Germanium n-MOSFETs fabricated on (111) wafers suffer from high source-to-drain tunneling which makes it difficult to turn off the device. The on-current suffers due to a very low density-of-states effective mass. Germanium (001)/[100] MOSFETs deliver high on-currents as a result of their high density of states and performs better than their silicon counterparts. One key device design parameter is the body thickness, and I_{OFF} is very sensitive to body thickness variations. Successful fabrication of such devices depends

Table 3.1
Device parameters for the optimally designed device. The worst case device has 10% shorter channel length and one atomic monolayer (one quarter of the lattice constant) thicker body, and hence exhibits degraded short-channel behavior.

Parameter	Value
t_{ins} [nm]	0.6 (EOT)
t_{body} [nm]	2.5
L_G [nm]	9
L_T [nm]	15
N_b [cm^{-3}]	0
N_{sd} [cm^{-3}]	$2X10^{20}$
L_{sd} [nm]	7.5
G [nm/dec]	1
Φ_m [eV]	3.8
V_{dd} [V]	0.4

on technological advancement that will limit the process variation in body thickness to within one atomic layer. Quantum scattering simulations for germanium MOSFETs show that higher mobility in the highly doped S/D regions is desirable and is important to achieve target on-currents.

Table 3.2
Performance projection for a ballistic Ge n-MOSFET.

	ITRS Target	Nominal Ge (001)/ < 100 > device
$I_{ON}[\mu A/\mu m]$	1500	2750
$I_{OFF}[\mu A/\mu m]$	10	1
$V_{DD}/I_{ON}[\Omega - \mu m]$	267	145
DIBL [mV/V]	–	60
S [mV/dec]	75	75

4. A TOP OF THE BARRIER BALLISTIC MODEL

Numerical simulations are used to guide the development of a simple analytical theory for ballistic field-effect transistors. When two-dimensional (2-D) electrostatic effects are small (and when the insulator capacitance is much less than the semiconductor (quantum) capacitance), the model reduces to Natoris theory of the ballistic MOSFET. The model also treats 2-D electrostatics and the quantum capacitance limit where the semiconductor quantum capacitance is much less than the insulator capacitance. This new model provides insights into the performance of MOSFETs near the scaling limit and a unified framework for assessing and comparing a variety of novel transistors.

4.1 Introduction

MOSFET channel lengths continue to shrink rapidly toward the sub-10 nm dimensions called for by the International Technology Roadmap for Semiconductors [6, 7, 99]. Coupled with the use of high-mobility channel materials [32, 36, 100–104], nanoscale channel lengths open up the possibility of near-ballistic MOSFET operation. For these reasons, it is important to understand ballistic operationboth in conventional MOSFETs and in unconventional transistors. Our objectives in this paper are to present a simple analytical theory for ballistic transistors and to explore its application to MOSFETS and to unconventional field-effect transistors.

The operation of MOSFETs in the ballistic regime has recently been explored by simple, analytical models [105], [106], [107] as well as by detailed numerical simulations [50,54,55,82,108,109]. In Section 4.2, we review our understanding of the device physics of ballistic MOSFETs as developed in previous publications [79,110,111]. In Section 4.3, we present a simple, analytical model, and in Section 4.4, we show that

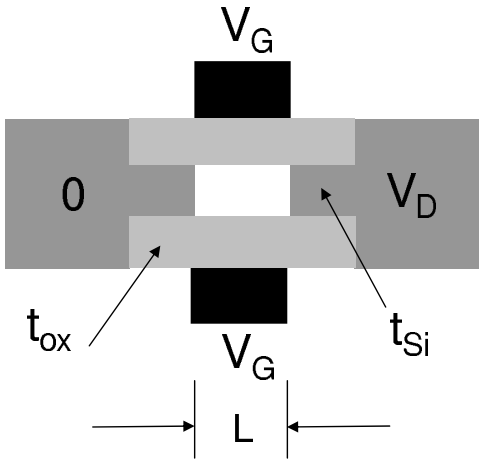

Fig. 4.1. Structure of the model device, a double-gate MOSFET. A body thickness of 1.5 nm and an oxide thickness of 1.5 nm were assumed. Both the source and drain regions were doped at $10^{20}/\text{cm}^3$. The gate workfunction was set to 4.25 eV which produced an off-current of 1.6 nA/μm.

it agrees with two-dimensional (2-D) numerical simulations of ballistic MOSFETs. Finally, in Section 4.5, we discuss why the model developed here does not describe devices like Schottky barrier FETs before concluding in Section 4.6.

4.2 Device Physics of Ballistic MOSFETs

Numerical simulations provide detailed information on the operation of nanoscale devices. Two transport models have proven to be especially useful in our work. The first is a numerical solution of the ballistic Boltzmann equation, and the second is the non-equilibrium Green's function (NEGF) formalism for quantum transport. Fig. 4.1 shows a model 10-nm MOSFET, and Fig. 4.2 shows the computed ballistic distribution function within the device under on-state conditions. The results show that two distinct carrier populations exist: one due to source injection and another due to drain injection (scattering would mix these two populations). Deep within the channel, the drain-injected population retains a near-equilibrium shape, but the source-injected population is strongly distorted. Fig. 4.3 is an NEGF simulation of

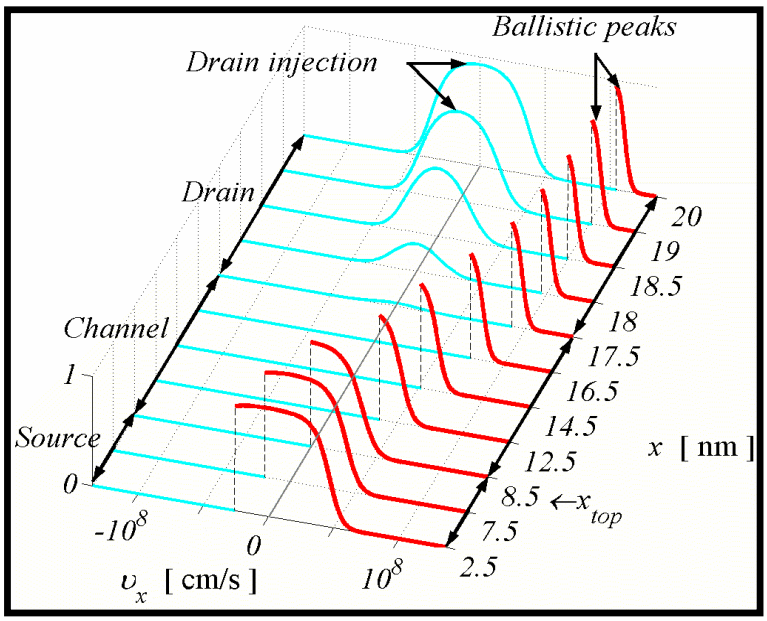

Fig. 4.2. The ballistic distribution function within the model device under on-state conditions as computed by solving the ballistic BTE (From [111])

the energy-resolved electron density under on-state conditions. Although quantum interference effects are seen as well as tunneling of carriers beneath the source-channel barrier, NEGF simulations of the terminal I-V characteristics of well-designed MOSFETs agree rather well with semi-classical simulations-even at the 10-nm scale (when the strong effects of quantum confinement are included in both simulations). Both the quantum and classical simulations show rich, complex phenomena within the device, but it turns out that a simple description of the current versus voltage characteristics is possible. Fig. 4.4 shows the computed self-consistent potentials within the model nanoscale MOSFET under low and high drain bias with gate voltage as a parameter (what is plotted is actually the bottom of the first subband versus position). At low gate voltages, the energy barrier between the source and drain is high, and the device is off. A high drain bias lowers the energy in the drain, and when a high gate voltage lowers the potential energy barrier, electrons flow from source to drain. This picture of the MOSFET is essentially that of the bipolar transistor; tran-

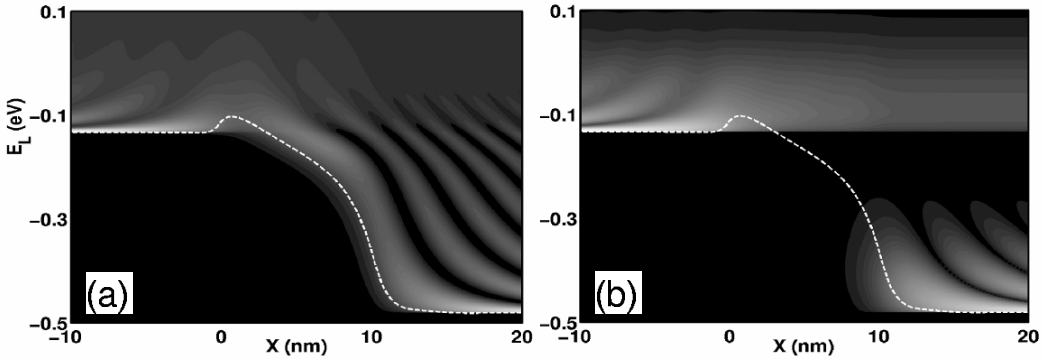

Fig. 4.3. The energy-resolved (a) density-of-states vs. position and (b) electron density vs. positions with the model device as computed using the ballistic NEGF formalism.

sistor action occurs by modulating the height of an energy barrier. It is more common to thin of MOSFETs in terms of the gate modulating the charge in the channel, but the charge in the channel is controlled by the height of the barrier. MOSFETs are bipolar transistors operate by similar principles (both below and above threshold); in the bipolar transistor, the height of the energy barrier is controlled directly by the base-emitter voltage, whereas in the MOSFET, it is controlled indirectly by the voltage on the gate. As will be discussed in Section 4.5, not all transistors operate by this charge (or barrier height) modulation principle.

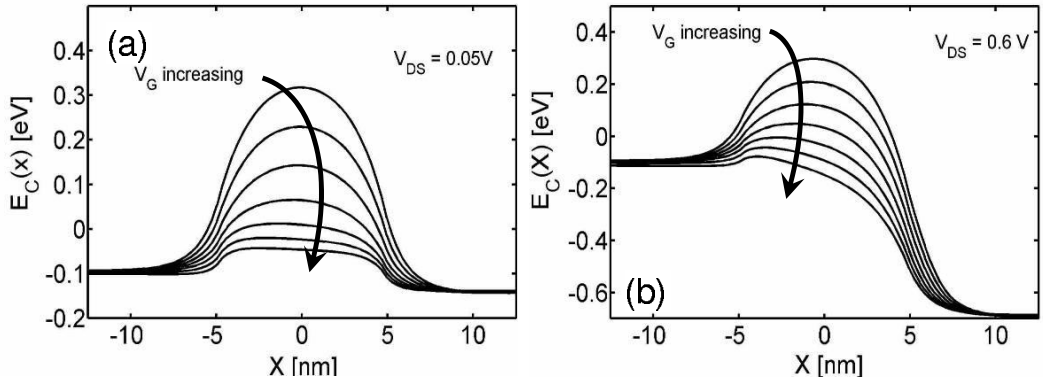

Fig. 4.4. The computed energy band diagrams under: (a) low drain bias and (b) high drain bias. The parameter is the gate voltage.

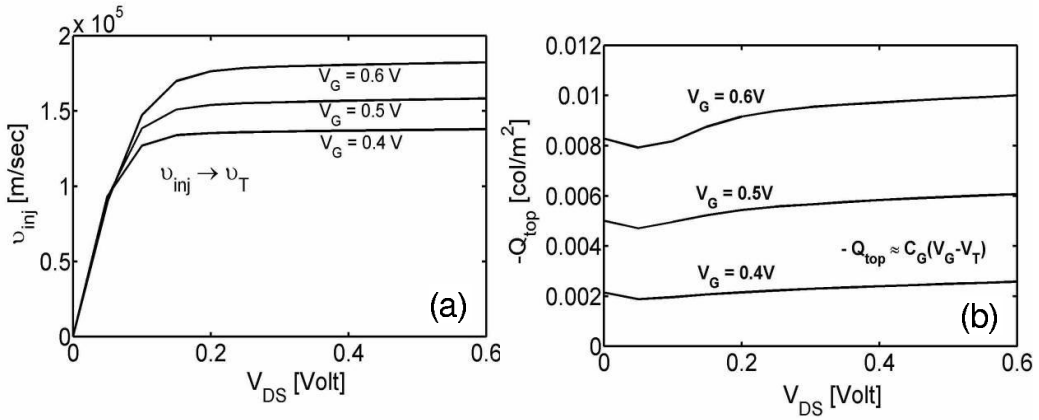

Fig. 4.5. Computed behavior at the top of the source to channel barrier. (a) The electron charge, $Q_n(0)$ at the top of the barrier vs. V_{DS}, (b) The average electron velocity at the top of the barrier vs. V_{DS}.

Current is the product of charge and velocity, which we plot separately in Fig. 4.5. In this figure, the gate voltage is high, and we plot the two quantities as a function of V_{DS}, fig. 4.5(a) shows that the charge at the top of the barrier is nearly independent of V_{DS} for a well-designed MOSFET, and for operation above threshold, it is given by MOS electrostatics as

$$Q_n(0) \approx C_{ox}\,(V_{GS} - V_T) \tag{4.1}$$

We will show in a later section that the initial dip in $Q_n(0)$ and the subsequent rise can be explained. Fig. 4.5(b) shows that the average electron velocity at the top of the barrier increases with V_{DS} and then saturates. The saturated velocity at the top of the barrier is simply the velocity of the thermal equilibrium hemi-Fermi-Dirac distribution shown in fig. 4.2. Note that above threshold, the electron gas is degenerate, and the magnitude of this injection velocity depends on the gate voltage. It is interesting to note that velocity saturation occurs in a ballistic MOSFET, but it occurs at the top of the barrier where the field is zero rather than at the drain end where the field is high.

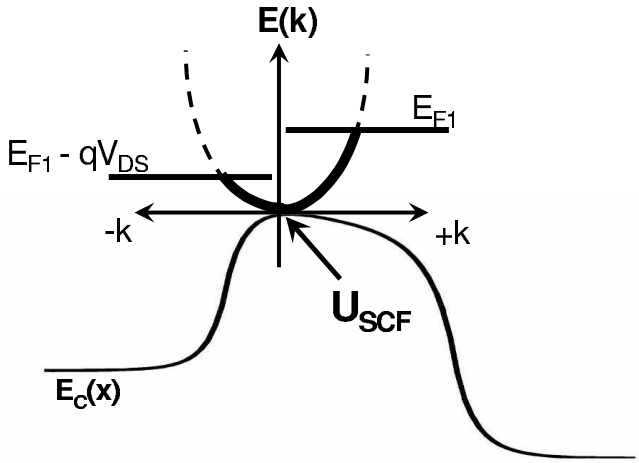

Fig. 4.6. Filling of the k-states at the top of the barrier by two Fermi levels E_{F1} and E_{F2}.

Because the top of the barrier has special significance, it is the starting point for our analytical model. For a ballistic transistor, the states at the top of the barrier are filled from either the source or the drain. For a quantum transport model, the local density of states fillable by the source and drain can be evaluated directly from the spectral function. In a semi-classical model, the local density of states is determined by the $E(\vec{k})$ relation for the semiconductor shifted by the self-consistent potential at the top of the barrier. Fig. 4.6 shows how the states at the top of the barrier are filled for a simple bandstructure. The positive velocity states are populated according to the Fermi level of the source and the negative velocity states by the Fermi level of the drain. Our key task is in developing an analytical model will be to devise a simple approach to determine the self-consistent potential at the top of the barrier. Finally, we mention ones subtle point. A careful examination of fig. 4.4b indicates that the conduction band in the source region actually floats down by about 10mV as the gate voltage increases. This unfamiliar behavior is a consequence of transport at the ballistic limit. The source Fermi level is fixed at 0eV and represents the Fermi level of the equilibrium source reservoir/contact. Under low gate bias, most of the positive velocity electrons injected from the contact

reflect from the energy barrier so that both positive and negative velocity states in the source extension are filled. When the gate voltage is high, however, the barrier decreases, and the fewer of the injected electrons reflect from the barrier so that it is mainly positive-velocity states in the source that are occupied. To achieve space-charge neutrality in the highly doped source extension, the conduction band must float down so that more electrons are injected from the source contact. When strong scattering is present inside the source extension, electrons occupy both positive and negative velocity states, and this effect is absent. For a more complete discussion of boundary conditions fro ballistic MOSFETs, see [112].

In the following section, we introduce a simple analytical model, and in section 4.4, we show that it accurately describes the physics of ballistic nanoscale MOSFETs.

4.3 Model

A simple 2D model for the ballistic MOSFET is shown in Fig. 4.7. It consists of three capacitors, which represent the effect of the three terminals on the potential at the top of the barrier. As also indicated by the shaded region in Fig. 4.7, mobile charge can be placed at the top of the barrier. The mobile charge is determined by the local density of states at the top of the barrier, the location of the source and drain Fermi levels E_{F1} and E_{F2}, and by the self-consistent potential at the top of the barrier U_{scf}. Because there is a relation between the local potential and the charge, this effect can be described by a nonlinear quantum capacitance [113]. In equilibrium

$$C_Q \equiv \frac{d(qN)}{d(-U_{scf}/q)} = q^2 \int_{-\infty}^{+\infty} D(E) \left(-\frac{\partial f(E - E_F)}{\partial E} \right) dE, \tag{4.2}$$

which, since $(-\partial f/\partial E)$ is sharply peaked about the Fermi energy, is q^2 times the density of states near the Fermi energy. Solomon *et al.* have pointed out in [98] that Natori's analytical ballistic model [105] does not include this non-geometric, quantum (or degeneracy) capacitance. Neglecting the quantum capacitance is justified for thick gate insulators (i.e., when $C_G << C_Q$); however, it fails to describe gate

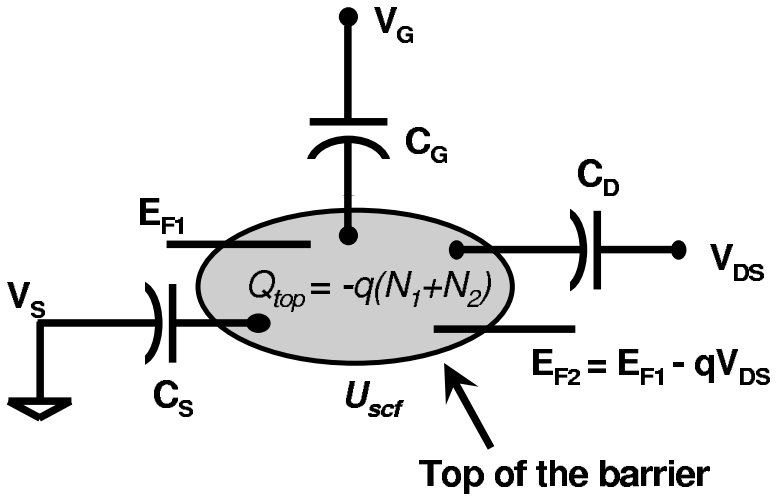

Fig. 4.7. The 2D circuit model for ballistic transistors. The potential at the top of the barrier, U_{scf}, is controlled by the gate, drain, and source potentials through the three capacitors shown. The mobile charge at the top of the barrier is determined by U_{scf} and by the location of the two Fermi levels. The nonlinear semiconductor (or quantum) capacitance is not shown explicitly but is implicit in the treatment of band filling.

electrostatics when the insulator capacitance is large compared with the quantum capacitance (i.e., when $C_G \geq C_Q$), which occurs when the electrical thickness is small or when the quantum capacitance is small, as in a one-dimensional conductor. Our model does not treat the quantum capacitance explicitly; however, it is included naturally through the treatment of self-consistent gate electrostatics.

When the terminal biases are zero, the equilibrium electron density at the top of the barrier is

$$N_0 = \int_{-\infty}^{+\infty} D(E) f(E - E_F) dE \tag{4.3}$$

where $D(E)$ is the local density of states at the top of the barrier, and $f(E - E_F)$ is the equilibrium Fermi function. The function $D(E)$ is nonzero for positive values of its argument only, which represents the minimum of the density of states and is specified as $E = 0$ in equilibrium. When a bias is applied to the gate and drain terminals (the source terminal is always grounded in the work), two things happen:

1. the self-consistent potential at the top of the barrier becomes U_{scf}, and
2. the states at the top of the barrier are now populated by two different Fermi levels.

The positive velocity states are filled by the source, according to

$$N_1 = \frac{1}{2}\int_{-\infty}^{+\infty} D(E - U_{scf}) f(E - E_{F1}) dE, \tag{4.4}$$

and the negative velocity states are filled by the drain according to

$$N_2 = \frac{1}{2}\int_{-\infty}^{+\infty} D(E - U_{scf}) f(E - E_{F2}) dE, \tag{4.5}$$

where $E_{F1} = E_F$, and $E_{F2} = E_F - qV_{DS}$. A change of variables can be used to re-express these equations as

$$N_1 = \frac{1}{2}\int_{-\infty}^{+\infty} D(E) f_1(E) dE, \tag{4.6}$$

$$N_2 = \frac{1}{2}\int_{-\infty}^{+\infty} D(E) f_2(E) dE, \tag{4.7}$$

where

$$f_1(E) \equiv f(E + U_{scf} - E_{F1}), \tag{4.8}$$

and

$$f_2(E) \equiv f(E + U_{scf} - E_{F2}). \tag{4.9}$$

Given an arbitrary density of states $D(E)$ and the location of the source and drain Fermi levels, we can evaluate the electron density at the top of the barrier $N = N_1 + N_2$ if the self-consistent potential U_{scf} is known.

Finding the self-consistent potential involves solving the two-dimensional Poisson equation as represented by the three capacitors in Fig. 4.7 with the bias induced charge $\Delta N = (N_1 + N_2) - N_0$ at their common terminal. We obtain the solution by superposition. First, ignoring the presence of the mobile charge in the channel, we calculate the Laplace potential at the top of the barrier due to terminal biases, which is

$$U_L = -q(\alpha_G V_G + \alpha_D V_D + \alpha_S V_S). \tag{4.10}$$

In this equation, the three αs describe how the gate, drain and source control the Laplace solution and are given by

$$\alpha_G = \frac{C_G}{C_\Sigma}; \quad \alpha_D = \frac{C_D}{C_\Sigma}; \quad \alpha_S = \frac{C_S}{C_\Sigma}, \tag{4.11}$$

where C_Σ is the parallel combination of the three capacitors in Fig. 4.7.

For a so-called, well-tempered MOSFET, the gate controls the potential, and $\alpha_G \approx 1$ and $\alpha_S, \alpha_D \approx 0$. The second part of the solution consists of grounding the three terminals and computing the potential due to the mobile charge, at the top of the barrier ΔN, from

$$U_P = \frac{q^2}{C_\Sigma}\Delta N. \tag{4.12}$$

Physically, a positive bias applied to the drain and gate terminals pushes down the potential energy at the top of the barrier as described by U_L, but because of the charge, the potential floats up, as described by U_P. The complete solution is obtained by adding the two contributions to obtain

$$U_{scf} = U_L + U_P = -q(\alpha_G V_G + \alpha_D V_D + \alpha_S V_S) + U_C \Delta N, \tag{4.13}$$

where

$$U_C = \frac{q^2}{C_\Sigma}, \tag{4.14}$$

is the charging energy.

Equations (4.6), (4.7), (4.13), and (4.14) represent the coupled nonlinear equations for the two unknowns N and U_{scf}. These equations can be solved iteratively to find the carrier density and self-consistent potential at the top of the barrier. Finally, the drain current is evaluated from

$$I_D = \int_{-\infty}^{+\infty} J(E)[f_1(E) - f_2(E)]dE, \tag{4.15}$$

where $J(E)$ is the "current-density-of-states" defined in the Appendix B.1.

In summary, the procedure for computing $I_D(V_G, V_{DS})$ consists of the following steps.

1. Specify the semiconductor carrier and current-density-of-states, $D(E)$ and $J(E)$ either analytically or by a numerical table.

2. Specify V_G, V_D, V_S and E_F.

3. Iteratively solve (4.6), (4.7), (4.13) and (4.14) for U_{scf} and N.

4. Evaluate the current from (4.15) for the assumed V_G and V_{DS}.

We have defined the model in terms of two density of states-one for the carrier density $D(E)$ and one for the current density $J(E)$ -which can be determined directly from the semiconductor bandstructure. In general, the integrals in 4.6, 4.7, and 4.15, must be done numerically, but for simple bandstructures, they can be done analytically. In the Appendix B, we evaluate these expressions for 2D carriers in a simple band and discuss how to use more general bandstructures.

4.4 Application to Ballistic MOSFETs

To illustrate the use of the model, we apply it to the double gate MOSFET presented in Fig. 4.1 and compare the results to 2-D numerical simulations with nanoMOS 2.0 [50]. Although the expressions for the α's given in (4.11) are exact, they are difficult to evaluate in practice because they depend on the 2D structure of the device. We will, therefore, treat them as fitting parameters and present a step-by-step procedure for determining the three parameters E_F, α_G, and α_D. The result show that this simple, three-parameter model does a good job of fitting the simulated I-V characteristics over the full range of operation.

4.4.1 Parameters for the Analytical Model

The first step is to set the Fermi level $E_F = E_{F1}$ for the correct threshold voltage, which is equivalent to setting the correct gate work function. Alternatively, setting the Fermi level is equivalent to setting the correct equilibrium carrier density at the

top of the barrier as given by (4.3). For a well-designed MOSFET at low gate and drain bias, α_D, U_L, and ΔN are all small so $U_{scf} \approx 0$, and (4.3) for N_0 depends on a single parameter E_F. In practice, we adjust the Fermi level in the analytical model so that the current matches that of the simulator for $V_G = 0$ and $V_{DS} = 50mV$.

Next, after setting E_{F1}, we adjust the gate control parameter α_G until the analytical model gives the same low V_{DS} sub-threshold swing as does the simulation. We do this for $V_G << V_T$ and for $V_{DS} = 50mV$. The induced charge at the top of the barrier ΔN is very small so that the gate controls the position of the top of the barrier U_{scf} through U_L. For complete gate control ($\alpha_G = 1$), the sub-threshold swing is ideal, i.e., $S = 60$ mV/dec at room temperature. For our model device, we obtained $\alpha_G = 0.87$.

Finally, having specified E_{F1} and α_G, the drain control parameter α_D was obtained by horizontal shift of the $\log(I_D)$ versus V_G characteristics in the sub-threshold regime (i.e., by matching the drain-induced barrier lowering (DIBL) of the simple model to the detailed numerical model). This parameter describes the additional change of the potential at the top of the barrier due to the drain bias. For our model device, we found $\alpha_D = 0.033$. Figures 4.8a and b compare the I-V characteristics from the analytical to those obtained by numerical simulation. From the $\log(I_{DS})$ versus V_G plot of Fig. 4.8a, we see that the sub-threshold characteristics match very well both for low and high V_{DS}. From the linear plot in the same figure, we also see that at low V_{DS} and high V_G, the characteristics match very well. However, when both V_{DS} and V_G are high, the match is poor, and the analytical model underestimates I_{DS}. This mismatch is also clear in the output characteristics presented in Fig. 4.8b, where we can see that for V_G above threshold, the drain current from the analytical model saturates at a lower value than numerical simulation. The reason for this mismatch under high V_G and V_{DS} and a way to treat it are discussed next.

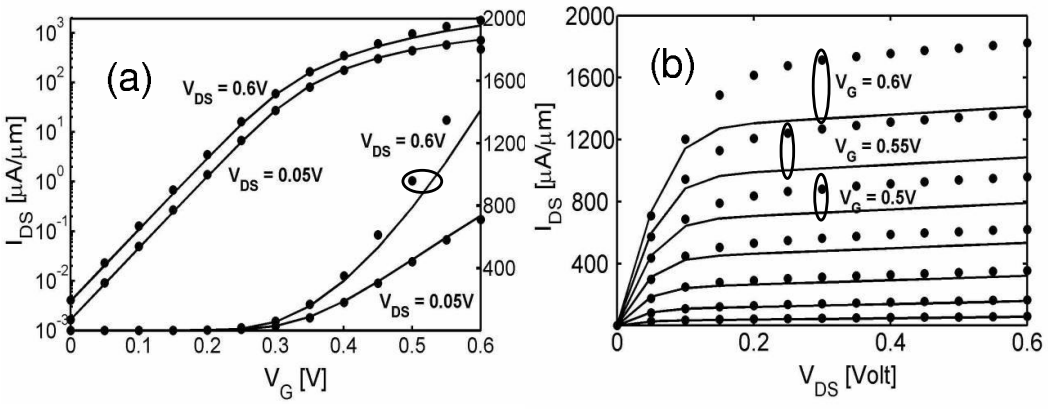

Fig. 4.8. Comparison of the analytical model to numerical simulations for the ballistic MOSFET of Fig. 4.1. (a) The transfer characteristics under both low and high drain biases. (b) The output characteristics. In both cases, the solid lines are from the analytical modes and the points are from nanoMOS simulations.

4.4.2 Treatment of the Floating Source Potential

The discrepancy between the analytical and numerical models under high gate and drain biases is related to the floating source potential, which was discussed briefly in Sec. 4.2. This phenomenon, which is important only in ballistic devices, is correctly implemented in the numerical simulator but has yet to be considered in our analytical model.

As discussed in Section 4.2 (and, at greater length, in [112]), for ballistic transport, a floating source potential is necessary to maintain charge neutrality in the highly doped source and the drain region under high bias conditions. As the gate voltage increases, fewer electrons are reflected from the barrier; the source potential must drop, so that enough additional electrons are injected to restore space-charge neutrality in the source. When the source potential decreases, so does U_{scf} at the top of the barrier. The result is that this floating source effect increases the carrier density at the top of the barrier, which explains the discrepancy observed in Fig. 4.8 under high gate and drain biases.

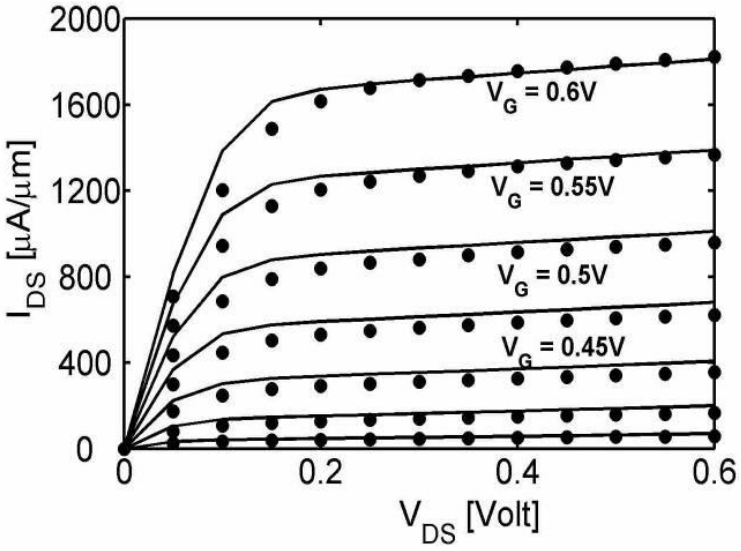

Fig. 4.9. Comparison of the analytical model to numerical simulations for the ballistic MOSFET of Fig. 4.1. In this case, the floating source potential was accounted for. In addition to good agreement at low gate and drain biases and as low gate and high drain biases, this plot show that the agreement at high gate and drain biases is also good.

With regard to the simulation procedure, the floating source potential means that the source Fermi level ($E_{F1} - E_C$) cannot be fixed at the beginning to produce a given V_T since it both gate and drain bias dependent. As discussed in the Appendix B.1, one can readily extend the procedure so that the Fermi level is iteratively adjusted to maintain space charge neutrality in the source under all bias conditions. Fig. 4.9 compares the $I_D - V_{DS}$ plots from the ballistic numerical simulation to the analytical model with the floating source treated, as discussed in the Appendix B.1. Figure 4.9 shows that when the floating source effect is included, the analytical model reproduces the full, numerical simulation quite well. The agreement is very good under high V_G and V_{DS} (where ignoring the floating source potential produced serious errors) but not quite as good under high V_G and low V_{DS}, where the model without floating source correction worked better.

4.4.3 Charge and Velocity at the Top of the Barrier

Finally, we examine the charge density

$$Q(0) = -q[N_1(V_G, V_{DS}) + N_2(V_G, V_{DS})], \tag{4.16}$$

and the carrier velocity

$$< v(0) > \equiv \frac{I_D(V_G, V_{DS})}{Q(V_G, V_{DS})}, \tag{4.17}$$

at the top of the barrier. Recall that the nanoMOS simulation of Fig. 4.5 shows that these quantities had a simple behavior at the top of the barrier. In Figs. 4.10a and b, we plot these two quantities from the analytical model. Figure 4.10a shows, in agreement with Fig. 4.5a, that the charge at the top of the barrier is nearly independent of the drain bias. The initial dip and subsequent rise are also seen, although not as pronounced as in the full, numerical mode. (The simpler model, which ignores the floating sources, actually does better in this regime.) the initial rise and subsequent saturation of the velocity at the top of the barrier is well-described by the simple model. These results show that Natori's assumption (and our own in

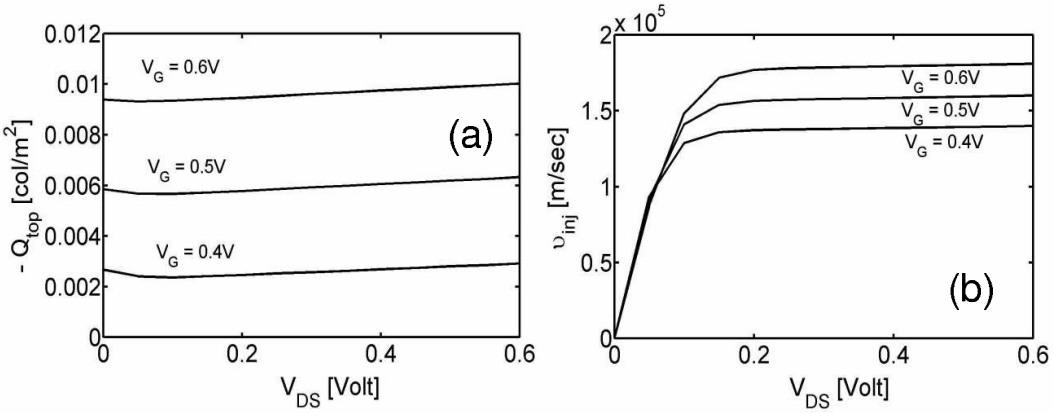

Fig. 4.10. Charge and injection velocity behavior at the top of the source to channel barrier as obtained from the analytical model. (a) The electron charge, $Q_n(0)$ at the top of the barrier vs. V_{DS}, (b) The average electron velocity at the top of the barrier vs. V_{DS}.

subsequent publications), where $Q_n(0)$ is independent of drain bias, is a good one for typical MOSFETs.

In practice, the model developed in this paper may be useful to compare the measured characteristics of nanoscale MOSFETs to their ballistic limits. From the measured electrical characteristics, the techniques presented in this section can be used to extract the parameters needed for the model. Another use for the model might be to compare the upper limit performance of devices that use novel channel materials to that of the conventional silicon MOSFET. (The model has been formulated to allow the use of numerically tabulated bandstructures) Finally, we note that the ballistic model is not entirely academic. Comparisons with experiments suggest that present-day MOSFETs operate at roughly 50% of the ballistic limit [48,114] and much of the research on new channel materials is motivated by a desire to approach the ballistic limit.

4.5 Discussion

The simple model we have developed does a rather good but not perfect job of reproduction more detained numerical simulations. The discrepancy under high gate and drain bias was resolved by forcing the potential at the top of the barrier U_{scf} to follow the floating source potential, but the high-gate, low-drain bias region is better described when U_{scf} is not allowed to follow the floating source potential. The reason for this behavior can be understood from Fig. 4.4. Under high gate and drain bias, the potential energy maximum is pushed up against the source; therefore, it seems reasonable that U_{scf} follows the floating potential in the source. Under low drain bias, it is not pushed as close to the source, and therefore, U_{scf} is not as tightly coupled to the source potential. Whether this physics can be captured in an analytically simple way is still under investigation.

4.6 Summary and Conclusion

In this chapter, we have developed a simple analytical model for ballistic nanotransistors that operate by modulating the charge in the device (as opposed to modulating the current at the contact). For conditions typical of silicon MOSFETs and when 2D effects are small, this surface potential model reduces to Natori's theory of ballistic MOSFET. When the insulator capacitance exceeds the quantum capacitance, however, some interesting new effects arise. This analytical model captures the essential physics of MOSFET-like ballistic nanotransistors and provides a convenient way to assess and compare transistors at the ballistic limit.

5. TIGHT-BINDING APPROACH TO SIMULATE NANOSCLAE MOSFETS

In this chapter we are going to relax one of the most widely employed assumptions in the simulation of nanoscale CMOS devices—the use of effective-mass bandstructure model with parabolic E-k relationship. Effective mass approach greatly reduces the computational effort required to calculate device electrical characteristics. This approach, however, comes under serious scrutiny when applied to the nanoelectronic devices. We will first list the key limitations of effective-mass approach when employed at nanoscale.

In deeply scaled devices, the quantum size effect along the thickness direction forms the discrete subbands with energies far from the bulk band edges. Since effective masses are calculated from the band curvatures near the bulk band edges, they do not provide a good description of these subband energies and their dispersions [115]. Subband dispersions, being formed at higher energies relative to the bulk band edges, often display anisotropy. An inherent limitation of the effective mass approach is its inability to treat such anisotropic dispersions. Valence bandstructure, especially the dispersion for heavy-hole band, cannot even be described by a simple parabolic dispersion due its anisotropic nature. Here one needs to use at least a six band $\vec{k} \cdot \vec{p}$ Hamiltonian to describe dispersions near bulk band edge, which is not accurate at higher energies where the subbands are formed. All these show the limitations of effective-mass approach in treating nanoelectronic devices.

Since, a correct description of the band structure is crucial for accurately calculating the subband energies and their dispersions at nanoscale, in this chapter we will introduce a semi-empirical atomistic tight-binding approach, a powerful tool to incorporate accurate full bandstructure in nanoelectronic devices, for this purpose.

This chapter is organized as follows. In Sec. 5.1 we present a very brief overview of tight-binding approach. In Sec. 5.2 we will highlight the semi-empirical nature of the approach, list the approximations and assumptions involved, and discuss on the choice of a basis set. In Sec. 5.3, we present tight-binding Hamiltonian in a $sp^3d^5s^*$-SO basis set and will show how to treat strained material. The application of the model to finite dimensional materials is presented in Sec. 5.5, and finally, the chapter is summarized in 5.6.

5.1 Introduction

In recent years, the semi-empirical atomistic tight-binding approach has received considerable attention, mainly because it offers an intuitively simple accurate description of the bandstructure in terms of chemical bonds, and its applicability to simulated nanostructures. Historically, the model was known as linear combination of atom(like) orbitals or LCAO and, Slater *et al.* first proposed it as a semi-empirical approach by treating the Hamiltonian matrix elements as disposable constants [60]. The Slater-Koster suggestion to treat the TB approach as an interpolation scheme was extensively used in a wide range of material systems–from transition metals to elemental and compound semiconductors. We will next highlight the approximations involved in tight-binding approach.

5.2 Semi-empirical Tight-binding Model

5.2.1 Assumptions Leading to A Semi-empirical Model

The mathematical details for the steps involved in deriving the tight-binding Hamiltonian is described in details in Appendix C.1 on page 146. Here, we will present the underlying assumptions and approximations involved.

Tight-binding or LCAO method begins with the assumption of an *one-electron Schrödinger equation* for the crystal, where the Hamiltonian is the sum of kinetic

energy and the potential energy operators. The potential energy operator itself is again a sum of spherically symmetrical nuclear potential at each atomic location in the crystal. We then expand the electronic wavefunction as a linear combination of a finite number of *Bloch sums* and attempt to write a secular equation in this Bloch sum basis. Use of Bloch sums as the basis for expanding the electronic wavefunction is necessary since in an infinite periodic potential, the solution of one-electron Schrödinger equation must obey Bloch's theorem, and Bloch sums always satisfy this condition.

Now we will see how these Bloch sums are constructed. In principal, a set of true atomic orbitals, such as those with s, p, d type symmetries, can be assigned at each location of the infinite periodic crystal and for each orbital in the set, Bloch sum can be performed. However, this immediately leads to problem. Two different true atomic orbitals sitting on different atomic sites are non-orthogonal to each other, and as a result, two Bloch sums that corresponds to these two orbitals are also non-orthogonal to each other. To avoid this non-orthogonality problem, instead of using true atomic orbitals, we use the Löwdin orbitals corresponding to them. These Löwdin orbitals are the symmetrically orthogonalized form of the original true atomic orbitals, constructed by subtracting all the overlaps with other orbitals on all other atomic locations in the crystal. As a result, Löwdin orbitals, whether located on different atomic sites or on same atomic site, are always orthogonal to each other and so are their Bloch sums [116]. Additionally, it can also be shown that Löwdin orbitals retain the symmetry of the original orbitals from which they are constructed, therefore, a Löwdin orbital constructed from a true atomic p_x orbital continues to possess all symmetry properties of the p_x orbital. For this property, from now we will call them as *atom-like orbitals*. The only disadvantage of employing the Löwdin orbital is, they are constructed by combining atomic orbitals located on different atomic sites, therefore, are more delocalized in space compared to the true atomic orbitals. However, since we will use the tight-binding method as a semi-empirical approach, we will not really attempt to calculate matrix elements corresponding to

these atom-like orbital, rather use them as fitting parameters. Thus the Bloch sums corresponding to these atom-like orbitals can be constructed and used as the basis for expanding the one electron wavefunction.

We then expand the crystal Hamiltonian in the Bloch sum basis and attempt to write the secular equation, the roots of which at a given $\vec{k}$ will give the bandstructure. The problem now is, the expression for Hamiltonian matrix elements between two Bloch functions is the sum of many terms, each of them is has the form of an spherical atomic potential sandwiched between two atomic orbitals with every possible combinations of their locations. Three major types of combinations can be identified:

1. **On-site** Both orbitals and the potential are located on the same atom.

2. **Two-center** Two orbitals located on different atoms and the potential is on one of these two atoms.

3. **Three-center** Each of the two orbitals and the potential are located at different atoms.

In addition to these three types of integrals, there is also intra-site overlaps, where two orbitals are located on same atom but the potential is on a different atom. Such integral is added to the on-site integral and becomes the diagonal, k-independent on-site energy. We now make an assumption that the *three-center integrals are much smaller compared to the on-site and two-center integrals* and therefore, will be ignored. This assumption is known as the *two-center integral assumption*, which leads to an important consequence, as shown in [60], that the orbital overlaps can be decomposed into a fewer two-center overlap energies with σ, π, or δ type bonding symmetries.

Two center integral assumption greatly reduces the number of integrals in the Hamiltonian matrix element, but still there are interactions present between orbitals located all over the crystal. To reduce the size of Hamiltonian, we now invoke

the *tight-binding* assumption, where we limit the maximum relative distance between atoms on which the orbitals are located. Most often, interactions between orbitals located on the nearest-neighboring atoms are only retained, leading to a nearest-neighbor tight-binding model, but occasionally, treatments of second even third neighbor interactions can be found in the published literature.

Finally, a *semi-empirical assumption* is employed, according to which we do not calculate the Hamiltonian matrix elements from the Löwdin orbitals and the spherical atomic potentials, rather we use a set of *fitting parameters for the two-center integrals and the on-site energies*, which finally leads to the correct band gap, position of band extrema, and band edge curvatures.

A few words on cubic semiconductor crystal structure is helpful to understand the structure of their tight-binding Hamiltonians. Semiconductors are tetrahedrally coordinated cubic materials. In general, they display zinc blende crystal structure, which consists of two inter-penetrating face-centered-cubic (FCC) lattices; each FCC may be considered as a sub-lattice. The two sub-latices are displaced by one quarter of a body diagonal and each consists entirely of one species of atoms. If the two sublattices are identical, we have a diamond structure. The two sublattices are called anionic and cationic, respectively. The nearest-neighbor atoms for an anion are four cations, and vice versa. The second-nearest neighbors of an anion are twelve anion. Similarly, other distant neighbors can be accounted for.

5.2.2 Choice of Basis

For cubic semiconductors, the minimal tight-binding basis is sp^3, and in the simplest case, interactions between the nearest-neighbor atoms are only treated. Such model satisfactorily describes the valence-band energy dispersion, as shown in [117], but fails to reproduce the indirect gaps of semiconductors at X and L correctly. The reason for this is pointed out in [118], where we see that at X and L

points, the electronic states contains substantial contributions from d-type orbitals, which are not included in the minimal basis set.

In [61], Vogl *et al.* attempts to mimic the influence of the excited d orbitals by adding an excited s orbital, namely s^*, to the minimal sp^3 basis. The resulting model is called sp^3s^*, which offers the correct positions of the lowest conduction minima at X and L. The limitations of the model are, the transverse masses for these indirect valleys and, the higher conduction bands show poor agreement with the experiment. The sp^3s^* model has limited use for the calculation of optical properties involving points at the surface of the Brillouin zone, and is not suitable for use in simulation of semiconductor devices. This basis set, sp^3s^*, with nearest neighbor interactions treated only, is indeed inherently incapable to describe correct transverse effective mass [119]. In order to correctly describe the effective masses of the indirect bands we have to either go beyond the nearest-neighbor interaction in sp^3s^* basis [119,120], or have to enlarge it by including higher d orbitals, resulting in a $sp^3d^5s^*$ model [62]. It is difficult to treat strained materials when interactions between distant neighbors considered, the only remaining choice is to include the d orbitals to the nearest-neighbor $sp^3d^5s^*$ basis.

In [62], a nearest-neighbor $sp^3d^5s^*$ tight-binding model is introduced. By including five d orbitals, the size of the atomic basis set has now grown to ten orbitals per atom per spin. With the inclusion of five d orbitals, the Bloch functions at the X and L points are considerably improved. In [62], the density-of-states, effective masses, and the deformation potentials are correctly reproduced. Finally, in [121], the usefulness of nearest-neighbor $sp^3d^5s^*$ tight-binding model is further extended by including treatment of local strain in nanoscale devices such as quantum well or heterostructure.

The tight-binding Hamiltonian can be written with or without treating spin-orbit (SO) coupling. When SO coupling is treated, the Hamiltonian size doubles. SO coupling introduces coupling elements between opposite spin orbitals on the same atomic site. Therefore, it does not include any matrix elements in the overlap energy bloc

of the Hamiltonian. The diagonal on-site energy blocs receives off-diagonal elements when SO coupling is treated. An excellent overview of the general treatment of SO coupling in cubic semiconoductors is presented by Dresselhaus in [122]. Treating SO coupling within the tight-binding context is elaborated in Appendix C.3 (page 160). SO coupling affects valence band only and when it is not accounted for, the heavy hole, light hole and split-off hole bands become degenerate at Γ point of the Brillouin zone. SO coupling removes the degeneracy of the split-off band. Such splitting off is large for Ge and GaAs, however, is small for Si valence bands.

5.3 The 3-D Tight-binding Hamiltonian in $spds^*$ Basis

The tigh-binding Hamiltonian in a $sp^3d^5s^*$ basis and without spin-orbit coupling can be written as,

$$H = \begin{bmatrix} [H_{aa}] & [H_{ac}] \\ [H_{ca}] & [H_{cc}] \end{bmatrix}. \tag{5.1}$$

In $\{s, p, d, s^*\}$ bais, each of the blocs are 10×10 matrices. In absence of SO coupling, the diagonal blocs, $[H_{aa}]$ and $[H_{cc}]$, contain only the diagonal elements, which are the on-site orbial energies for anion and cation respectively. The off-diagonal overlap energy blocs, $[H_{ac}]$ or $[H_{ca}]$, however, are full matrices. The structures of the matrices are discussed in details in Appendix C.1 on page 146.

5.4 The Treatment of Strain

Boykin *et al.* presented a highly accurate parametric model to treat local strain in semi-empirical tight-binding formalism in [121]. In this model, modifications of the Hamiltonian matrix elements due to the displacements of the atoms from their equilibrium unstrained positions, is described. Strain not only affects the overlap matrix elements of the orbitals, but also changes the on-site energies of the orbitals.

Treating the modulation of overlap matrix elements is rather straightforward. In strained materials, the relative positions of the neighboring atoms are different

from their equilibrium values where both the bond-angle and bond-length between nearest neighbors has changed. For a relaxed crystal, the directional cosines for the anion-to-cation bond orientation are equal and

$$l = \frac{1}{\sqrt{3}},\ m = \frac{1}{\sqrt{3}},\ n = \frac{1}{\sqrt{3}}.$$

For a strained material they are no longer equal and must be calculated from the deformed crystal structure. These new directional cosines automatically incorporate the modification of overlap matrix elements due to change in bond angle. The modification of the Slater-Koster two-center-integrals due to change in bond length is treated according to

$$V'_{i_a j_c \alpha} = \left(\frac{d_0}{d}\right)^{\eta} V_{i_a j_c \alpha}. \tag{5.2}$$

In (5.2), d and d_0 represent the strained and unstrained bond lengths and, $\alpha \in \{\sigma, \pi, \delta\}$ is the bond type. Primed and unprimed matrix elements, $V'_{i_a j_c \alpha}$ and $V_{i_a j_c \alpha}$, denote the strained and unstrained values, respectively. The parameter, η, is different for different types of bondings and the value is within the range of 1 to 4 [121]. Equation 5.2 is the generalization of Harrison's scaling law, which has similar expression but with $\eta = 2$ [123].

Treating the modulation in on-site energy due to strain is not straightforward. In [121], the expression for diagonal parameter shift is derived, and the diagonal parameter for strained material becomes

$$E'_{i,\alpha} = E_{i,\alpha} + \sum_{j \in N, N-i} \sum_{\beta} \left[C_{(j,\beta),(i,\alpha)} \frac{V^2_{(j,\beta),(i,\alpha)} - V'^2_{(j,\beta),(i,\alpha)}}{E_{j,\beta} + E_{i,\alpha}} \right]. \tag{5.3}$$

Equation (5.3) describes the relationship between the strained (primed) and unstrained (unprimed) on-site energies. The strained on-site energy for a given orbital, α, at the lattice point, i, now depends on its unstrained value, $E_{i,\alpha}$, the unstrained on-site energies, $E_{j,\beta}$, of all orbitals, β, of all of its nearest-neighbors, j, and the unstrained and strained overlap matrix elements, $V_{(j,\beta),(i,\alpha)}$ and $V'_{(j,\beta),(i,\alpha)}$, with all the neighboring orbitals.

5.5 Application to Finite Dimensional Systems

In Sec. 5.3, the tight-binding Hamiltonian for bulk material is presented. Here we will briefly review its application to the ultra-thin-body devices. In Refs. [120, 124], some numerical details of writing the tight-binding Hamiltonian in finite dimensional system is shown. In [124], Støvneng and Lipavský presented a inverse Fourier transform method to decompose the interaction matrices between the different atomic layers along the growth direction. For the (001) wafers, the UTB structure can be constructed by alternaively stacking anion and cation type atomic layers. The Hamiltonian for this becomes,

$$H_{TF}\left(k_x, k_y\right) = \begin{bmatrix} [H_{aa}] & [V_{ac}] & & & \\ [V_{ac}]^\dagger & [H_{cc}] & [U_{ac}]^\dagger & & \\ & [U_{ac}] & [H_{aa}] & [V_{ac}] & \\ & \ddots & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{5.4}$$

where the overlap blocs, $[V_{ac}]$ and $[U_{ac}]$, are functions of the in-plane wavevector $\vec{k}_{||}$.The detailed structure of the matrix is discussed in Appendix C.4 on page 164.

5.6 Summary

In this research, we have used a nearest-neighbor $sp^3d^5s^*$ tight-binding model, with spin-orbit (SO) coupling, for treating nanoscale CMOS devices with strained/unstrained Si, Ge, GaAs and InAs as novel channel material. The applications of this model are presented in Ch. 6 and Ch. 7.

6. APPLICATION OF TIGHT-BINDING APPROACH TO UTB DG MOSFETS

In Ch. 5, a semi-empirical tight-binding approach was introduced and its potential to offer an improved understanding of the nanoscale device operation was discussed. In this chapter we present specific applications of a 20 orbital $sp^3d^5s^*$-SO tight-binding approach to UTB DG nanoscale CMOS devices. We assume ballistic transport and use the top-of-the-barrier transport model introduced in Ch. 4.

This chapter is organized as follows. First, the calculation of the bandstructure in UTB MOSFETs is briefly described in Sec. 6.1. Then in Sec. 6.2, we discuss the validity of using the parabolic $E\left(\vec{k}\right)$ relationship in UTB DG Ge n-MOSFETs and will show that below 4 nm body thicknesses, the use of parabolic $E\left(\vec{k}\right)$ may introduce serious error. Also in this section, the effects of relative orientation of uniaxial strain and quantum confinement axes in UTB Si p-MOSFETs is discussed [75]. Finally, in Sec. 6.3, we explore novel channel materials for nanoscale MOSFETs by comparing the ballistic performances of UTB DG Si, Ge, GaAs and InAs MOSFETs. Here we will see that very high injection velocities of III-V materials does not always ensure high on-current, since such high velocity also degrades density-of-states which, in turn, lowers the gate capacitance [76].

6.1 Overall Simulation Approach: Tight-binding without Band Bending

This chapter presents results for UTB DG devices where the body is thin enough to ignore the electrostatic potential variation along the body thickness direction. We call this *non-self-consistent* application of tight-binding model. In Fig. 6.1 the bandstructure is schematically presented. For the ultra-thin-body, we first calculate the

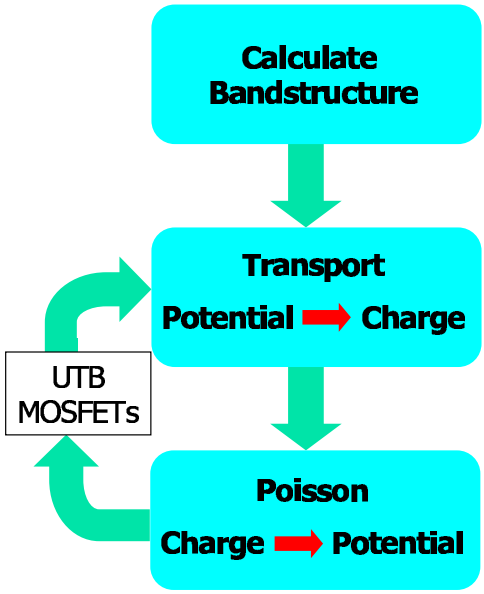

Fig. 6.1. Tight-binding simulation approach without treating self-consistent electrostatics. For an UTB MOSFET, the band bending along the thickness direction due to electrostatic potential is negligible. As a result, a constant potential can be assumed to calculate the 2-D bandstructure.

bandstructure over the entire 2-D Brillouin zone assuming zero externally applied electrostatic potential for all atomic layers. The gate electrostatics is then included by solving the Poisson equation at the top-of-the-barrier, as described in Ch. 4. The self-consistent potential at this point along the channel raises or lowers the bandstructure, but does not affect the band bending inside the body, hence bandstructure is not changed.

In Fig. 6.2 the band profile in UTB device along thickness is schematically presented. For body thickness less than 3–4nm, the subbands are formed high enough and the negligible band bending does not affect the bandstructure.

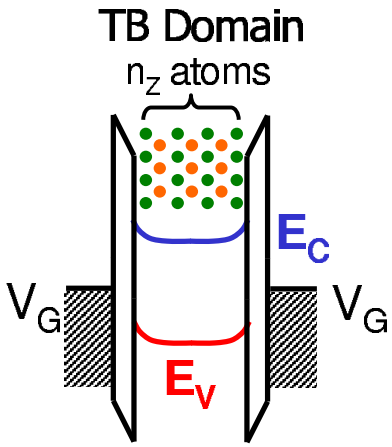

Fig. 6.2. The band profile in an UTB DG MOSFET. For thin bodies, the top and bottom gate insulators offer the carrier confinement in the channel. Electrostatic potential has negligible effect on bandstructure.

6.2 Application: On the Validity of Parabolic $E(\vec{k})$ in Nanoscale Ge n-FETs and Effects of Strain Orientation in Si p-FETs

New device geometries, ultra-thin bodies, new channel materials, and the use of strain engineering are being explored to address the device design challenges of the ITRS [6, 7]. Strained channels have been shown to be effective for improving the transport properties of bulk silicon MOSFETs [18–22]. New channel materials and device structures, such as pure germanium channels, ultra-thin-bodies (UTB), dual-gate, tri-gate, and FinFET structures, are also being studied. Assessing performances of such devices by physics-based quantitative simulation is a challenge as the use of the traditional effective mass (EM) approach becomes questionable. Although carrier transport depends on both bandstructure (effective mass) and scattering, in this work we neglect scattering because our objective is to understand the role that bandstructure plays. It is important to understand how atomistic effects, quantization, and strain influence the extremely scaled ballistic novel devices, and this section presents and atomistic approach to address the questions. The results show that bandstructure effects play an important role in nanoscale MOSFETs.

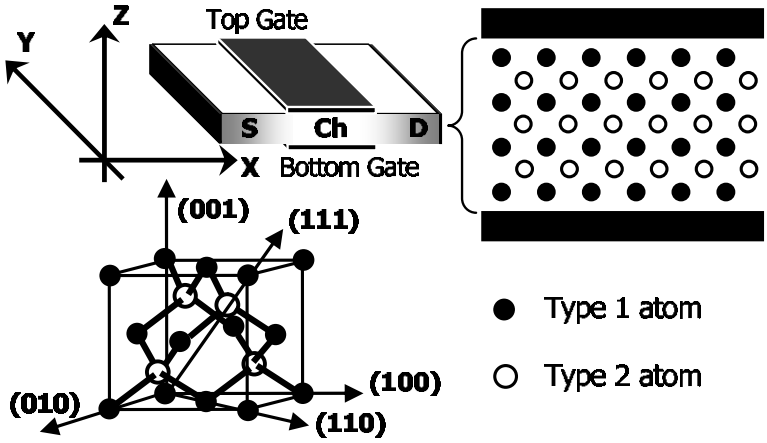

Fig. 6.3. The UTB DG MOSFET structure fabricated on (100) wafer and the underlying atomic arrangement. The body is organized as atomic layers (right). The TB modeling of cubic semiconductors as two inter penetrating FCC lattice classifies the atoms according to their positions-lattice atom (black) and basis atom (white). For elemental semiconductors, both types are the same atom.

6.2.1 Approach

Figure 6.3 sketches the UTB-DG device structure and the underlying atomic arrangement. The translational symmetry of the crystal is broken along the channel thickness direction and therefore, the bulk TB Hamiltonian is discretized along this direction (Appendix C.4 on page 164). The channel is organized as N_z layers of atoms, each atom is modeled using 10 orbitals per atom per spin – totaling 20 orbitals per atom [125]. These atom-like localized orbitals have s, p, and d type symmetries, respectively, and the model is known as $sp^3d^5s^*$. Each layer contributes a 20×20 block in the block tri-diagonal thin film Hamiltonian. The resulting $20N_z \times 20N_z$ Hamiltonian is solved by sparse matrix technique to find the bandstructure. This atomistic approach allows device simulation with acceptable computational burden. A zero boundary condition for the wavefunction is used at the top and bottom semiconductor-oxide interface. The dangling bonds at these interfaces are pacified using a hydrogen termination model of the sp^3 hybridized interface atoms. This

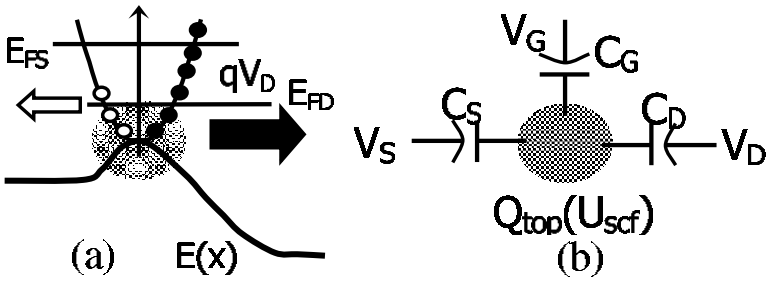

Fig. 6.4. The semi-numerical ballistic model for arbitrary bandstructures. (a) The states at the top-of-the-barrier are classified in two groups according to the direction of their velocity–towards drain or towards source. The velocity is calculated from tabulated 2D bandstructure, and the carrier groups are in equilibrium with the source or drain Fermi level, respectively. (b) The equivalent circuit for modeling 2D electrostatics such as DIBL and output conductance.

technique, introduced in [126] and elaborated in Appendix C.5 on page 169 of this thesis, successfully removes all the interface states from the bands gap.

Figure 6.4 summarizes the top-of-the-barrier ballistic transport model used in here to assess the ballistic performance of the nanoscale devices. This model is a generalization of the analytical model introduced in Chapter 4 [74]. The states at the top of the barrier are divided into two groups according to their group velocities and two fluxes of carriers are found, source injected and drain injected. For each gate and drain bias, the Poisson's equation is solved at the top of the source-channel barrier to get a self-consistent potential. The equivalent circuit model to treat 2D electrostatics is also shown.

The bulk conduction band (CB) of germanium, near and away from conduction band edge, is shown in Fig. 6.5. The near CB edge constant energy surface is shown in Fig. 6.5a, where it can be observed that the lowest valleys of Ge conduction band are L type-ellipsoids of revolution around eight equivalent $\Lambda \equiv [111]$ directions. Half of teach ellipsoid is within the first Brillouin zone. In addition to this, in the Ge CB there are spherical Γ valley and six silicon like X valleys along Δ, all within 250 meV from the L-valley edge. In Fig 6.5b, 300 meV above the CB edge, all valleys

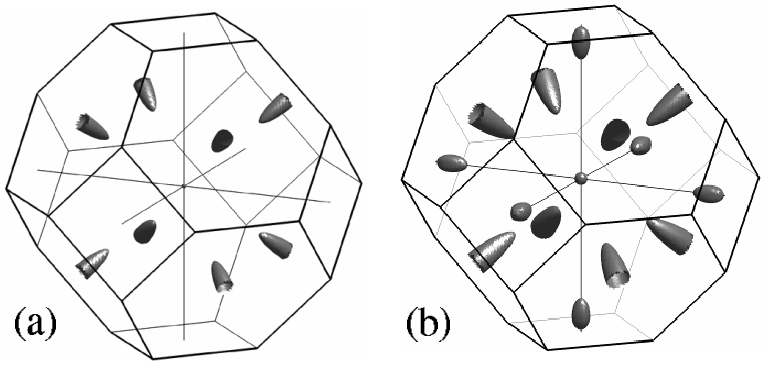

Fig. 6.5. Tight binding simulations of the constant energy surfaces of bulk Ge: (a) near CB edge, ΔE_C=100meV and (b) away ΔE_C=300meV.

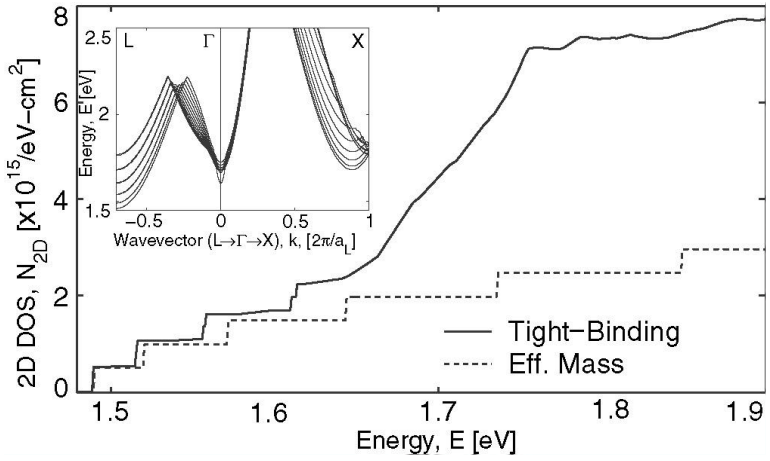

Fig. 6.6. Tight binding (solid) density-of-states of a 16 nm (113 atomic layers) thick Ge (100) channel compared with the results from EM approach (L-valley). At this thick body limit, Ge bandstructure near the band edge is well described by parabolic $E - k$. Inset: Conduction subband $E(\vec{k})$. Here, the L valley subbands form the lowest ladder of subbands at the band edge (as in bulk Ge).

are visible. Due to close proximity of these band edges and the difference in their curvatures (effective mass) along the quantum confinement directions, the valley energies re-order when quantization forms subbands in ultra-thin-body devices. Thus,

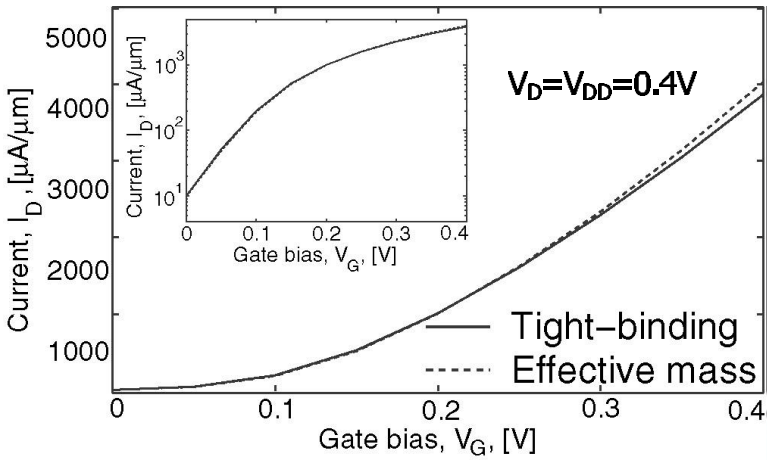

Fig. 6.7. Ballistic $I_D - V_G$ for tight-binding and effective mass bandstructures of Fig. 6.6 at $V_D = V_{DD} = 0.4$V. Transport along [100] is considered. Neither threshold voltage shift, nor density-of-states is an issue at this body thickness and the two curves practically lie on top of each other.

below a certain body thickness, X-valleys become lower than L valleys [127]. This dramatically changes the transport properties of the thin channel devices.

6.2.2 Results and Discussions

Using the tight-binding bandstructure and ballistic transport models described above, the performance of unstrained germanium n-MOSFETs are first compared with the effective-mass (parabolic E-k) results. Then the effect of strain orientation with respect to channel thickness direction is explored in an UTB strained silicon p-MOSFET.

Unstrained Ge n-MOSFETs

Figures 6.6 and 6.7 explore bandstructure and ballistic characteristics of a (100) Ge film in the thick body limit, 113 atomic layers (approximately 16nm). In Fig. 6.6, as a result of the finite body thickness, the CB conduction band splits into subbands with the L valleys forming the lowest subband ladder (inset). The L-point in the 2D Brillouin one is actually the projection of the 3D L-point on the $k_x - k_y$ plane.

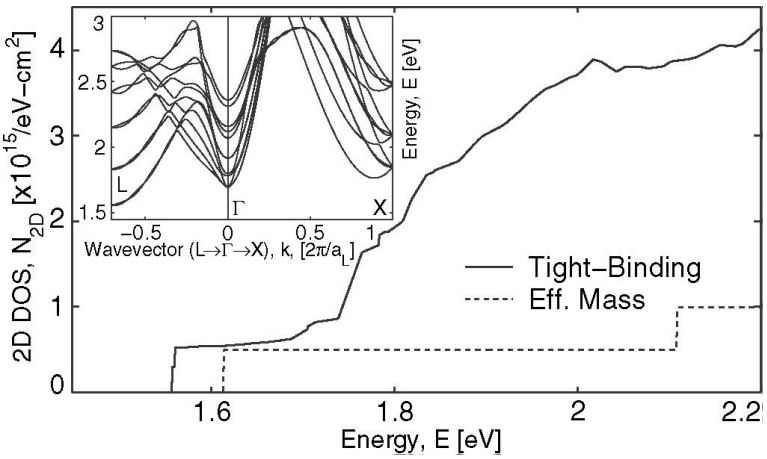

Fig. 6.8. The 2D density of states and subband $E - \vec{k}$ from TB and EM calculations for a 4nm (30 atomic layers) body. The non-parabolicity in L valley subbands is apparent here, manifested as a disagreement in the band edge position between the two approaches. The lowest subbands are still from L type valleys but the projection of two of the X valleys at Γ forms nearby subbands.

Similarly, at the Γ point, both the 3D Γ point and the projection of the Δ line along k_z on the $k_x - k_y$ plane are located. As a result, the subbands at Γ can be either from bulk Γ valleys or from the X valleys with major axis along k_z. The 2D density-of-states (DOS) plot shows that the tight-binding and effective mass (always treating only lowest L valleys) approaches agree well in predicting both the DOS at band edge and the lowest subband energies. At this thick body limit, Ge bandstructure is approximately described by the EM approach in an energy range of 120 meV above the band edge. For energies below 1.62eV, both the position of the band edge and the 2-D DOS agree. Although, the TB DOS includes the effects of higher subbands from higher valleys at Γ and X, the ballistic $I_D - V_G$ at $V_D = V_{DD}$ in Fig. 6.7 shows that they don's participate in transport, and the I-Vs from effective mass and tight-binding are essentially identical. At this body thickness limit conduction subbands are closely separated in energy and six subbands were used in these I-V calculation.

In Figs. 6.8 and 6.9, similar characteristics were investigated for a 4nm (30 atomic layers) body. In Fig. 6.8, the effects of non-parabolicity in L valleys, which still form

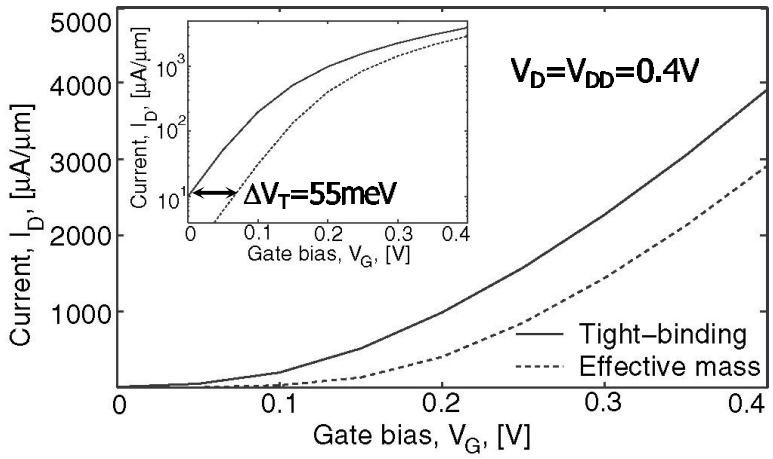

Fig. 6.9. The ballistic $I_D - V_G$ for a 4nm Ge (100) MOSFET at $V_D = V_{DD}$=0.4V. For this moderately thin body, the shift of V_T becomes the prime issue. The V_T from the EM approach is higher and the corresponding I_{ON} is underestimated by about 25%.

the lowest subband ladder, are visible. At this body thickness, the effective mass and tight-binding subband energies do not agree, which is manifested in the $I_D - V_G$ at $V_D = V_{DD}$ plots (Fig. 6.9) as a threshold voltage, V_T, shift. The result is a 25% underestimation of I_{ON} from the effective mass approach. The key limitation for the effective mass approach at this body thickness range is the non-parabolicity of the L valleys. The 2D DOS, calculated from these two approaches agree at their respective subband band edges, therefore, if V_T is adjusted by correct amount, their I-Vs will coincide.

The case of extreme scaling of body thickness is presented in Fig. 6.10. Here, for a 2nm (12 atomic layers) thick body, the scenario changes completely from the previous cases. The inset $E - k$ from tight-binding approach reveals that the two X type conduction band valleys, with major axes along [001] forms the 2-fold degenerate lowest subband at Γ. This is similar to the unprimed subband ladder in Si (100) wafers. Additionally, due to strong non-parabolicity in L valleys, their four-fold degenerate subbands remain close to the lowest X_2 valley subbands, and the lowest subband practically become 6-fold degenerate. The effective mass approach

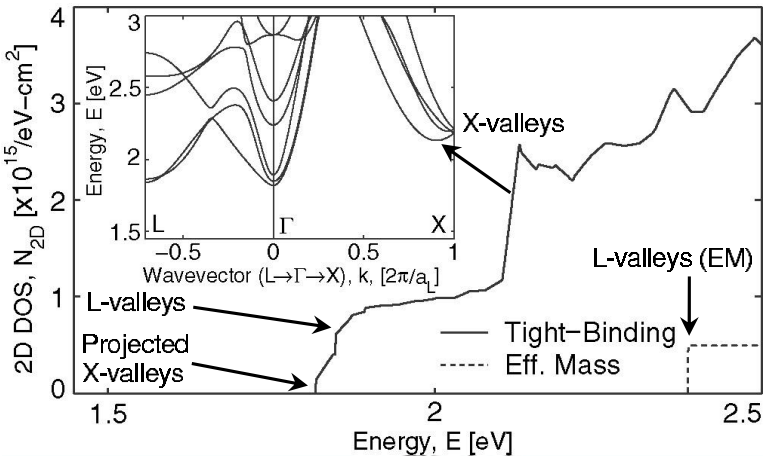

Fig. 6.10. Bandstructure and 2D DOS for a 2nm (12 atomic layer) Ge (100) body. L valleys are strongly affected by non-parabolicity. The lowest subband at Γ stems from the two longitudinal X valleys along Z. The fourfold degenerate L subbands are very close to the X subband edge-all six subbands effectively participate in transport. The EM approach does not capture such effects.

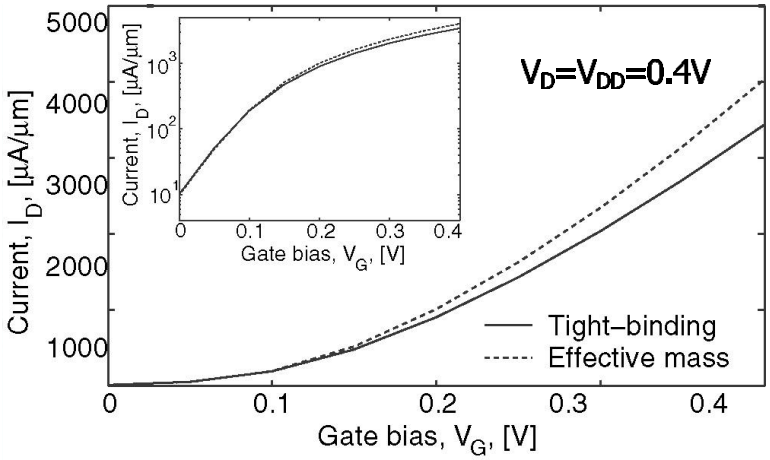

Fig. 6.11. The ballistic I_D-V_G for a 2nm Ge n-MOSFET compared at V_D =0.4V. The large mismatch of the EM subband energies from the TB results is compensated by adjusting the gate work function, and both have the same V_T (inset). Compared to TB, the EM approach overestimates I_{ON} by about 15%.

is unable to treat the L-valley non-parabolicity and predicts very high subbands

from them. A large band edge disagreement (0.57 V) between these two approaches is visible, as well as a disagreement between the values of 2D DOS at the respective band edges. The huge band edge disagreement between the effective mass and tight-binding approaches necessitates V_T adjustment, otherwise comparison of their ballistic characteristics is meaningless. In Fig. 6.11, the $I_D - V_G$s are compared after adjusting the gate work function to account for this V_T difference, and the characteristics show a 15% overestimation of I_{ON} is predicted by the effective mass approach.

Strained Si p-MOSFETs

Strain alters the position of the atoms in a crystal and in a tight-binding model, both the *overlap energies* between orbitals sitting on neighboring atoms and the *on-site orbital energies* are altered because of this. The relevant theory to treat strain in semi-empirical tight-binding approach, and its application to III-V semiconductors are presented in [121] and [66]. In this work we have used the same formalism to treat strain in ultra-thin-body Si p-MOSFETs.

It has already been observed experimentally that relative orientation of compressive strain with respective to the gate electric field has a profound effect on the hole transport properties in planar Si MOSFETs, especially at high gate bias [18–22]. It has been pointed out in [21] that hole mobility enhancement in *substrate induced* biaxially strained Si MOSFETs is present only at low gate field and disappears completely at high gate bias. However, if *process induced* compressive strain is applied uniaxially along channel direction, then the hole mobility enhancement, over unstrained devices, continues to be present both at high and low gate biases. The main difference between these two kind of strained devices is – in the first case, compressive strain axis and the gate field are aligned, while in the second case they are normal to each other. Gate field causes quantum confinement of carriers along the thickness direction and in these two device geometries, relative orientations of strain and quantum confinement are different. Since both strain and quantum con-

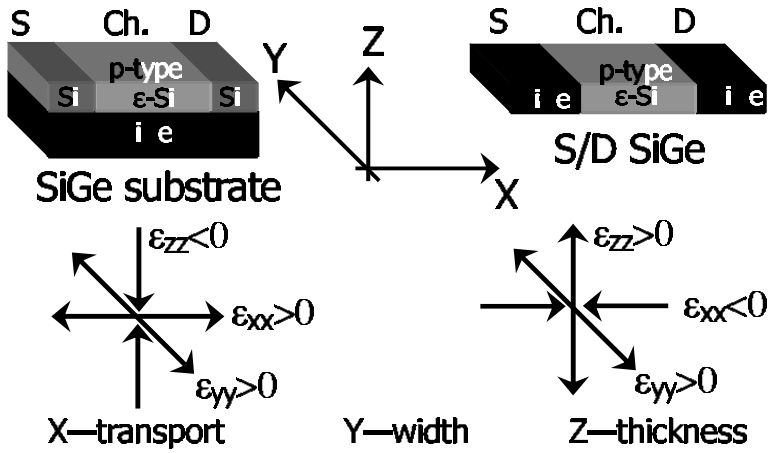

Fig. 6.12. Two strained Si p-MOSFET geometries compared in this work-SiGe substrate and SiGe source/drain devices. For the SiGe substrate device, compressive strain is along *quantum confinement* direction of inversion layer holes, while in the SiGe S/D device it is along *transport*.

finement substantially change the bandstructure, their combined effect on valence bandstructure must be treated to explain such mobility behavior in Si p-MOSFETs.

Figure 6.12 sketches the two different strained p-MOSFET geometries compared in this work. They mimic the two strained devices described above. In both cases, relaxed SiGe induces strain in the pseudomorphic Si channel. Lattice mismatch between SiGe and Si causes biaxial tensile strain ($\epsilon > 0$) in Si channel, parallel to the SiGe-Si interface, and compressive strain ($\epsilon < 0$) normal to the interface. The holes are quantum mechanically confined along thickness direction, Z, and we find that in these two geometries, the relative orientation of the uniaxial compressive strain and confinement direction are different.

The combined effect of strain and quantum confinement on the Si valence subbands is presented in Fig. 6.13. In bulk Si, quantum confinement is absent and biaxial tensile strain removes the degeneracy of heavy hole (HH)-light hole (LH) band at Γ. As a result, the LH band with better transport properties rise in energy and is primarily populated by holes. In Fig. 6.13 we observe that such strain ad-

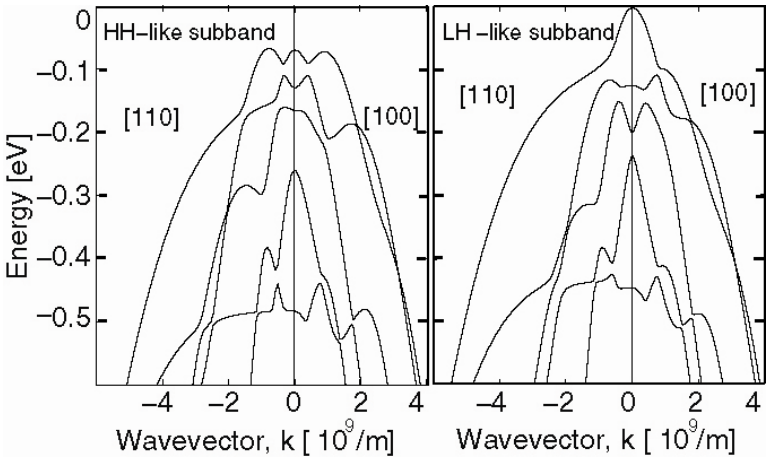

Fig. 6.13. TB simulations of strain effects in quantized valence subbands of Si. Thickness of Si-(001) channel is 24 atomic layers (3.5 nm). Strain inducing SiGe regions contain 25% Ge. Left: In SiGe substrate device, top hole subband in Si is HH type (small curvature). Right: Top hole subbands of Si in SiGe S/D device is LH type (large curvature). In presence of strong quantum confinement, a clear bandstructure advantage of SiGe S/D device is visible.

vantage is nullified by quantum effects in silicon-on-SiGe devices (left), and the top subband again becomes HH like, resulting in poor transport properties. However, for the strained Si channel in SiGe-S/D devices (right), the top subband remains LH like (large curvature) even at strong quantum confinement. In Fig. 6.14, the ballistic I-V of such devices reflects this bandstructure advantage, and we find the SiGe S/D devices offers 25% more I_{ON} compared to its counterpart. In order to perform a valid comparison, the same I_{OFF} is obtained in both devices by adjusting their respective gate work-functions (inset). The results clearly highlight the importance of bandstructure effects in ballistic p-MOSFETs.

6.2.3 Summary and Conclusion

In this subsection, the key bandstructure related issues in the operation of ballistic nanoscale CMOS have been investigated. We have gone beyond the effective mass

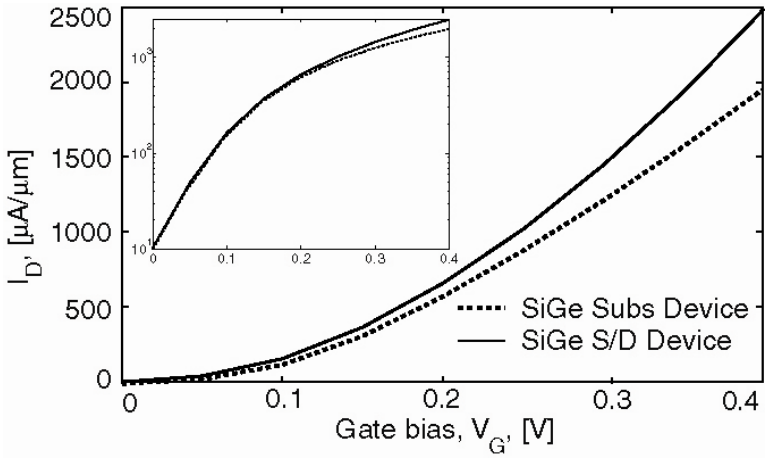

Fig. 6.14. Ballistic $I_D - V_G$ for strained Si p-MOSFETs: SiGe-substrate vs. SiGe-S/D devices. A bandstructure advantages in the SiGe S/D device is clearly visible-a 25% higher I_{ON} than its counterpart at the ballistic limit.

approach (parabolic $E - k$) by using an atomistic model, the $sp^3d^5s^*$ semi-empirical tight-binding model with spin-orbit coupling. The electronic subband structure, 2-D DOS, and ballistic device operation of UTB-DG Ge (100) n-MOSFETs have been studied using these two approaches with 16, 4, and 2 nm body thickness, respectively. The results show that below 4nm body thickness, the tight-binding approach captures non-parabolicity and multi-band transport phenomenon in Ge correctly. These effects are, however, missed by the effective mass approach. For the extremely thin bodies (2nm), the lowest subbands in Ge are of a very different nature than the results from the results of parabolic $E - k$ approach; therefore, the scattering dominated transport properties are going to be completely different from those in bulk materials. Finally, a study of the effects of strain vs. quantization orientations in p-MOSFETs show that when strong quantization is present, there is a clear bandstructure advantage in UTB devices where uniaxial compressive strain is applied along the transport direction, compared to the device under biaxial strain.

6.3 Application: Comparing Si, Ge, GaAs and InAs as Nanoscale CMOS Channel Materials

Novel materials such as pure Ge or III-V semiconductors (e.g. GaAs or InAs) has raised considerable interest for their possible use in nanoscale CMOS devices as alternative channel material. Their improved transport properties, along with the use of new device designs are expected to enhance the performance of end of ITRS CMOS devices. Novel process techniques, such as ALD, high-κ dielectrics, and metal gates are now being used explore III-V MOSFETs [43]. New materials in the channel promise reduced series resistance and higher on-currents. The theoretical assessment of such devices, however, is a challenge because bandstructure, quantum effects and electrostatics must all be treated. In this subsection, a systematic analysis of the bandstructure effects in deeply scaled n- and p- MOSFETs with Si, Ge, GaAs and InAs channel is performed. A $sp^3d^5s^*$-SO tight-binding (TB) model and a top-of-the-barrier ballistic transport model have revealed important trends in deeply scaled novel channel material devices.

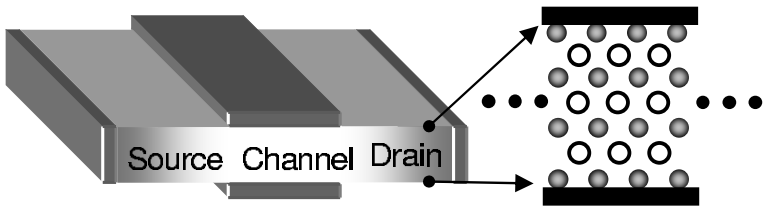

Fig. 6.15. UTB DG MOSFET as model device and the underlying atomic structure. A 20 orbital $sp^3d^5s^*$-SO TB model is used to calculate the 2D subband dispersions for Si, Ge, GaAs and InAs. Bandstructure is calculated assuming translation symmetry along transport and width directions; however, along thickness direction the body is resolved atomistically.

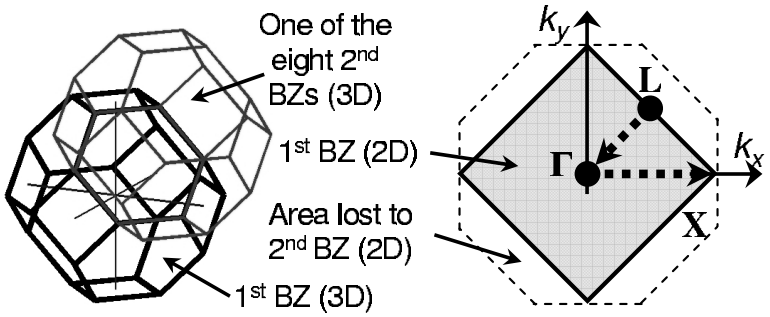

Fig. 6.16. Formation of 2D Brillouin zone. The shaded square in right is the 2D BZ. The bandstructure is calculated over $L\Gamma X$ triangle and using symmetry, the entire 2D BZ is filled [67].

6.3.1 Approach

The model device, an ultra-thin-body (UTB), symmetrical, dual gate (DG) MOSFET, is shown in Fig. 6.15. Here the shaded and open spheres represent atoms from two (anion or cation) inter-penetrating FCC lattices of the diamond or zinc-blend crystals, respectively. Two dimensional bandstructures for the channel materials were calculated using a 20 orbital $sp^3d^5s^*$-SO semi-empirical tight-binding model. The tight-binding input parameters are the orbital on-site energies and the overlap energies of nearest neighbors. The parameter sets for each material, Si, Ge, GaAs, and InAs, is optimized to give accurate bandgaps and effective masses simultaneously for the conduction (CB) and valence bands (VB) [121, 125].

While calculating 2-D bandstructure, crystalline translational symmetry is assumed along transport, [100], and width, [010], directions. Due to ultra-thin-body nature of the device, along the thickness direction, [001], periodicity is broken and the body is treated as discrete atomic layers. As elaborated in Appendix C.4 in page 164, the tight-binding Hamiltonian of a N_Z mono atomic layers (AL) thick body is represented by a $20N_Z \times 20N_Z$ hermitian matrix for each k point in the 2D Brillouin zone (BZ) (Fig. 6.16). The size of this sparse matrix may be huge, from which few eigenvalues near the bandgap are obtained using a sparse matrix technique. Hard-wall boundary condition for wavefunction is assumed at the top and

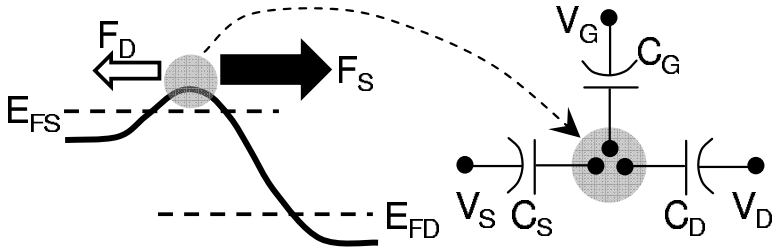

Fig. 6.17. The semi-numerical quasi-2D top-of-the-barrier ballistic transport model.

bottom interfaces, which results in discrete subbands due to energy quantization along thickness direction.

Figure 6.16 describes the formation and shape of 2-D Brillouin zone in the k-space. We know that reciprocal lattice for real space FCC lattice has BCC type symmetry. In order to obtain the 3-D BZ we need to draw the Wigner-Seitz cell for this BCC reciprocal lattice. Since diamond or zinc-blend crystal structure consists of two inter-penetrating real space FCC lattices, formation of their 3-D BZ follows the same process, and in Fig. 6.16 shape of such zones in k-space is shown (left). Now to find 2-D BZ in presence of quantum confinement of 3D crystal along [001], we first, project the reciprocal BCC lattice on (001) plane, and then draw the Wigner-Seitz cell on this 2-D plane.

We thus obtain the shaded square on right side, which is the 2-D BZ for all cubic semiconductors. Here, The L point is the projection of 3D L point on $k_x - k_y$ plane. The Δ line along k_z is projected at the center, Γ point. As a result, the X_2 conduction band valleys form the unprimed subbands at 2D Γ point. Due to zone folding, the 2D BZ is smaller in area than the octagonal shape (dotted) obtained by slicing 3D BZ at $k_z = 0$. The bandstructure is calculated over $L\Gamma X$ triangle and symmetry is used to fill the entire 2D BZ.

Abrupt termination of crystal symmetry at top and bottom semiconductor-insulator interfaces leaves unsatisfied dangling bonds. These bonds generates lo-

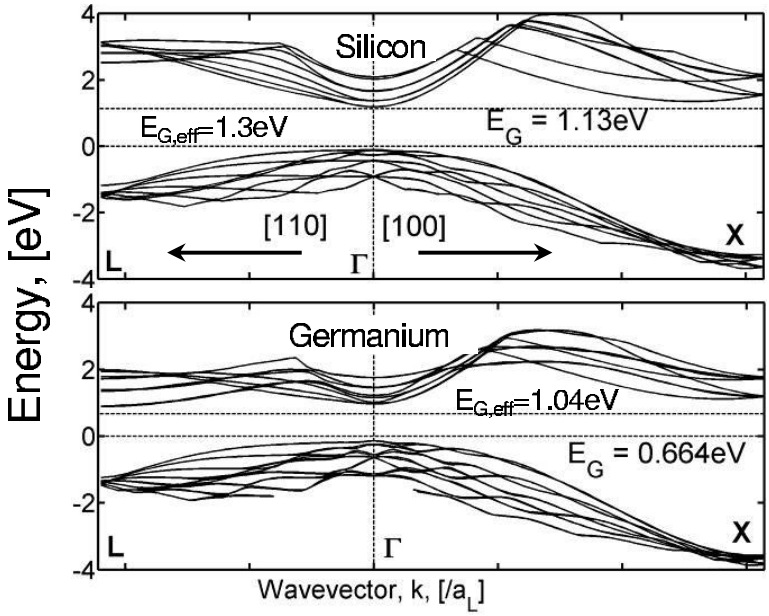

Fig. 6.18. UTB (19 AL, $\approx$ 2.75nm) (001) Si (top) and Ge (bottom) VB and CB subband dispersions.

calized interface states causing charge trapping. These interface states are removed using a technique similar to hydrogen passivation of Si-SiO_2 interfaces [126].

The 2-D bandstructure thus obtained were used to calculate the ballistic I-V using the semi-numerical, top-of-the-barrier transport model shown in Fig. 6.17 [74]. This model treats ballistic transport semiclassically by filling the k-states at top of the source-channel barrier. Carriers are grouped in two fluxes, F_S and F_D, which are in equilibrium with source and drain Fermi levels, E_{FS} and E_{FD}, respectively. F_S and F_D are calculated by populating the positive and negative velocity states in 2D k-space according to E_{FS} and E_{FD} and weighting them with velocity at each point. The model can treat arbitrary bandstructures. Their difference is the net current. Self-consistent potential at beginning of channel obtained by solving Poisson equation and the 2D electrostatics treated using the circuit model shown right.

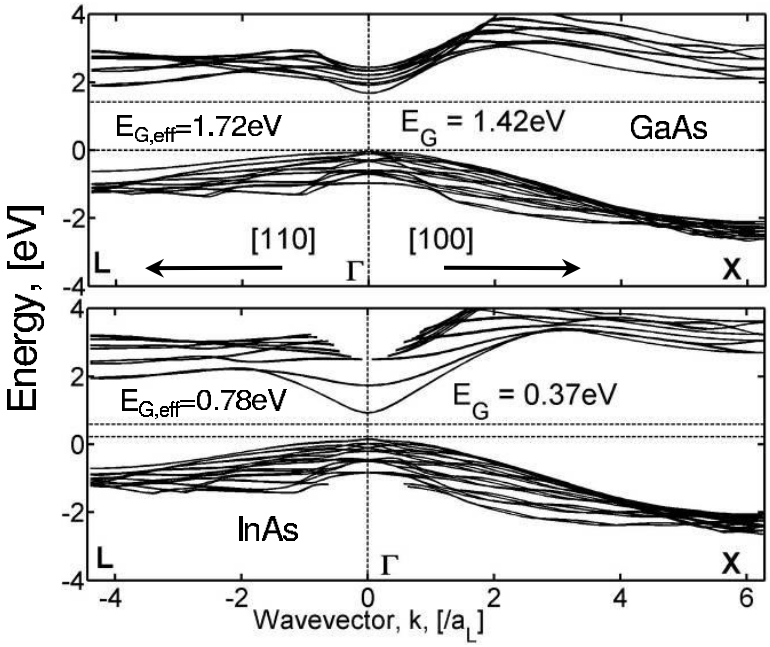

Fig. 6.19. UTB (19 atomic monolayers, $\approx$ 2.9nm) (001) subband $E - \vec{k}$ for III-V materials. Top: GaAs, Bottom: InAs.

6.3.2 Results and Discussion

The ITRS 2004 specifications for the 2016 MOSFET are: physical gate length, $L_G = 9nm$, equivalent oxide thickness (EOT) $t_{OX} = 0.5nm$, off-current, $I_{OFF} = 0.5\mu A/\mu m$ and supply voltage, $V_{DD} = 0.8V$ (0.6V used here considering series resistance drop). Using the full 2D ballistic simulator nanoMOS 2.5, a $L_G = 9nm$ UTB DG silicon device is first simulated and the optimum body thickness of 3nm (19 AL) is determined. From the full 2D Si I-Vs, the circuit model parameter are then extracted and used in the quasi-2D model of Fig. 6.17 to treat the 2D electrostatics.

The 2-D conduction and valence subband dispersions for 19 atomic layer thick Si and Ge bodies are presented in Fig. 6.18. Here, strong quantum confinement increases their effective bandgaps (energy difference between lowest conduction subband edge and highest valence subband edge) to 1.3 and 1.04eV, respectively. For Si conduction band, the 2-fold degenerate X_2 valleys form the unprimed ladder at Γ.

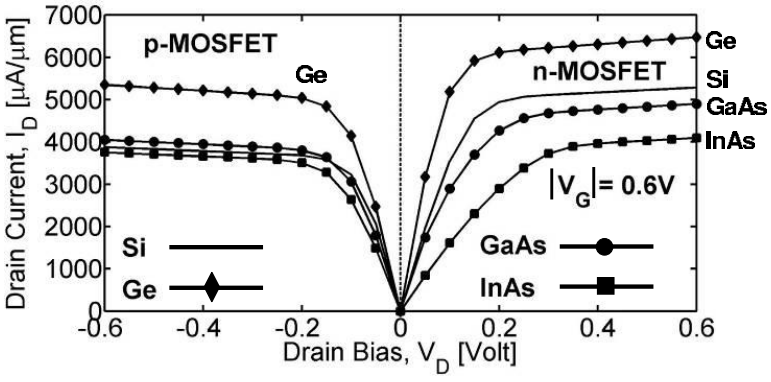

Fig. 6.20. Ballistic $I_D - V_D$ at $V_G = V_{DD}$ =0.6V for 19 AL DG p- and n-MOSFETs; calculated using the 2D TB bandstructures for various channel materials (Figs. 6.18, 6.19). The reason for the poor performance of III-V MOSFETs is their very low CB DOS, which the degrades their gate capacitance C_G and lowers Q_{top}.

In Ge conduction band, the lowest subbands are the projected 4-fold degenerate L valleys. In addition to these L valleys, 2-fold degenerate X_2 valleys at Γ are only 75 meV higher and therefore, Ge conduction band at this body thickness range is effectively 6-fold degenerate. High density-of-states in Ge conduction band is attractive since it increases the semiconductor capacitance.

Similar dispersions for 19 AL GaAs and InAs are presented in Fig. 6.19. Here, quantum confinement increases the effective bandgaps for GaAs (top) and InAs (bottom) to 1.72 and 0.78eV, respectively. Conduction subband for both are from singly degenerate Γ valleys. Additionally, for GaAs the 4-fold degenerate higher L valleys are just 200 meV away. Although, the small transport m^* in III-V materials is desirable for high v_{inj}, their low DOS m^* has the adverse effect of degraded quantum capacitance effect in deeply scaled MOSFETs.

Figure 6.20 compares the ballistic $I_D - V_D$ for p-and n-MOSFETs at $|V_G|$ = 0.6V using the 2-D tight-binding bandstructures for various channel materials in Figs. 6.18 and 6.19. The EOT is 0.5nm and for all devices, and I_{OFF} is adjusted to 0.5 μA/μ m by varying gate work functions. For both n- and p- devices Ge offers maximum ballistic I_{ON}. While GaAs and InAs n-channels offers very high room

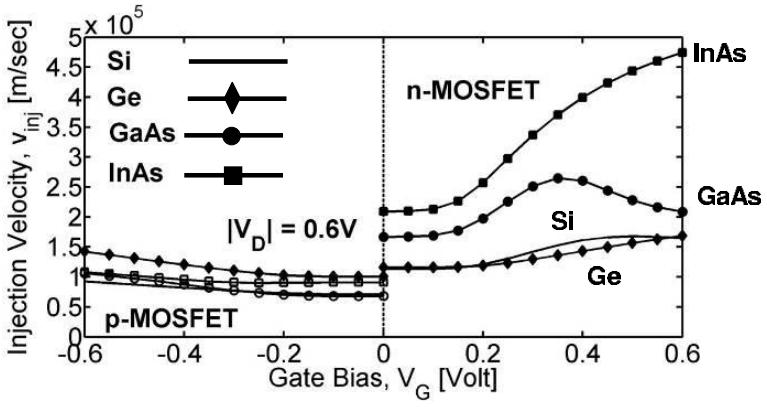

Fig. 6.21. Injection velocity, v_{inj}, vs. V_G plots for UTB (19 AL) n- and p- MOSFETs at $V_D = V_{DD} = 0.6$V. Holes have nearly same v_{inj}, but electron v_{inj} vary widely.

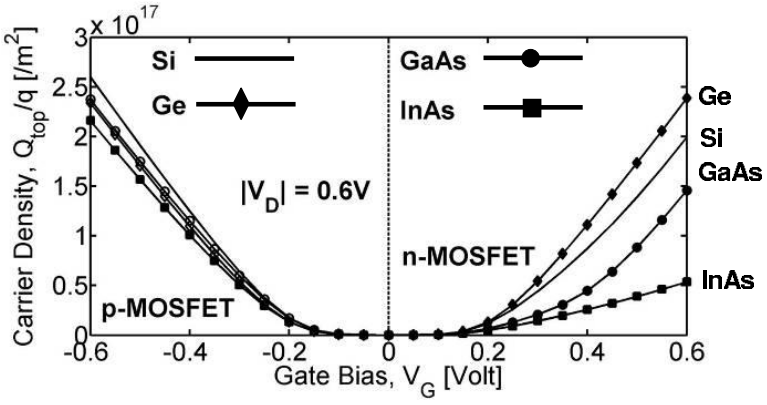

Fig. 6.22. Carrier density, Q_{top}, vs. V_G plots for UTB (19 AL) n- and p- MOSFETs at $V_D = V_{DD}$ =0.6V.

temperature mobility, neither of them offer improved ballistic performance over Si or Ge. The explanation is offered in next paragraph using the injection velocity, $vinj$, and carrier density, Q_{top}, vs. V_G plots in Figs. 6.21 and 6.22. Because p- type devices behave nearly the same for all materials, we will focus on the n- devices.

First recall that current is the product of Q_{top} and v_{inj}. In Fig. 6.21, at low V_G the carriers are non-degenerate and v_{inj} is flat. At high V_G, v_{inj} increases due to degenerate carrier statistics. Electron v_{inj} is very high for III-V materials due to

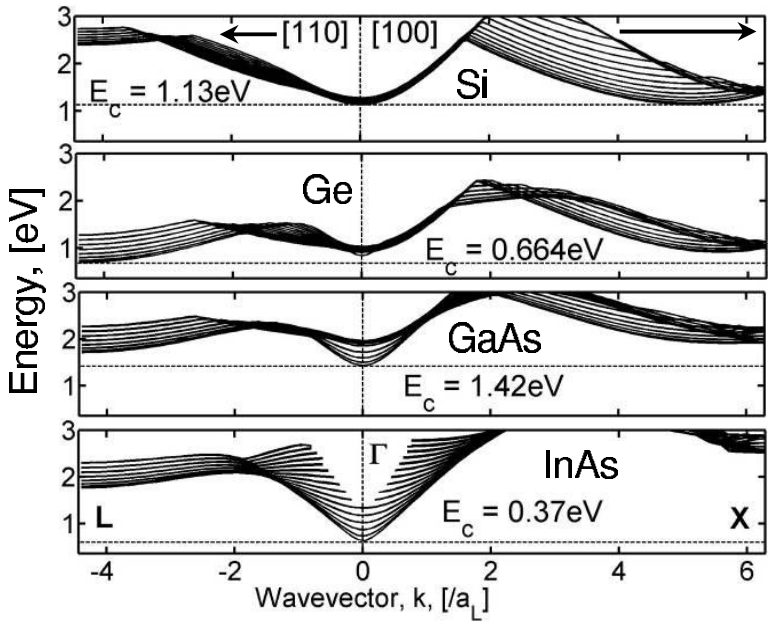

Fig. 6.23. CB subband $E - \vec{k}$'s for 100 AL bodies on (001) wafers. The lowest subband edges coincide with the bulk band edges since the quantum confinement is negligible.

low CB m^*. In this figure, InAs displays a very high v_{inj}, as expected, however, its Q_{top} in Fig. 6.22 is drastically reduced. This happens because the very small CB DOS lowers the overall gate capacitance C_G. Similarly, Q_{top} for GaAs displays some reduction due to the low DOS; however, additional degradation of v_{inj} at high V_G, due to carrier filling higher slow velocity L valley subbands, is the main reason for its overall poor ballistic performance. GaAs v_{inj} starts to fall from V_G 0.4V due to carriers populating higher energy, yet low velocity L valleys. Similarly, Si v_{inj} starts to flatten due to populating higher X_4 valleys. Finally, although Ge has the smallest v_{inj}, it is the overall winner because of very high Q_{top}.

The importance of maintaining high density-of-states for deeply scaled EOT (0.5nm) is apparent in Fig. 6.22. Valence band DOS are nearly same for all materials, so are Q-V for holes. However, CB DOS can vary widely and the associated C_G ($\equiv$ slope of Q-V plots) also varies over a wide range. For III-V n-channels, CB DOS are very small and as a result, they display strongly degraded C_G. Higher L

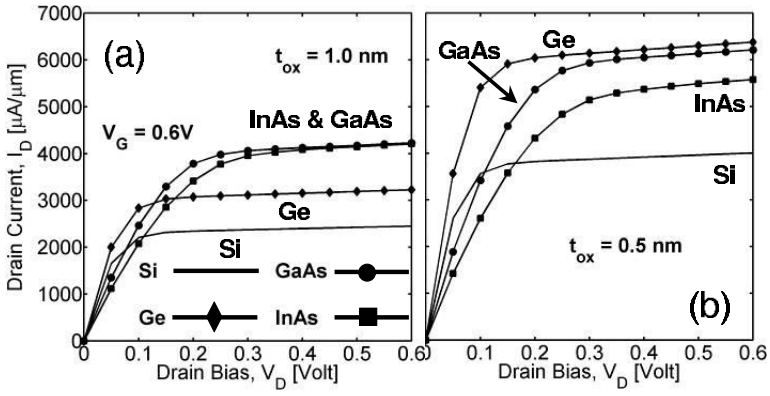

Fig. 6.24. $I_D - V_D$ at $V_G = 0.6$V for thick body (100 AL) (001)/[100] DG n-MOSFETs for two different EOTs, (a) 1nm and (b) 0.5nm. For thicker oxide (a), the degradation of C_G in III-V devices is minimal and they offer the highest I_{ON}. For thin EOT (b), C_G for III-V degrade and the Ge channel offers maximum I_{ON}.

valleys in GaAs begin to populate at V_G 0.4V and C_G increases due to this increase in DOS. Finally, Ge, although have lowest v_{inj}, offers maximum ballistic I_{ON} due to high C_G.

In chapter 4, the concept of semiconductor capacitance, C_S, (also known as quantum capacitance, C_Q) was introduced. This capacitance is a function of the density-of-states of the carriers in a device. Since this capacitance appears in series with the gate insulator capacitance, C_{OX}, the overall gate capacitance, C_G, which is the series combination of C_{OX} and C_S, is always smaller than C_{OX}. In order to further explore the relationship between C_{OX}, C_S, and ballistic I_{ON}, we simulate n-MOSFETs with different combinations of C_{OX} and C_S. Smaller C_{OX} can be obtained by increasing the EOT. In order to increase C_S, we use thicker bodies, where due to negligible quantum confinement effects the subbands are closely spaced in energy and therefore, higher C_S results. The calculated 2-D conduction subband E-k's for 100 AL thick Si, Ge, GaAs and InAs bodies are presented in Fig. 6.23. Body thicknesses in nm are 13.7 (Si), 14.3 (Ge and GaAs) and, 15.3 (InAs). Quantum confinement effects are minimal here, therefore, subbands are closely spaced in energy and multiple subbands can be populated at high V_G. Thus, degradation of C_G due

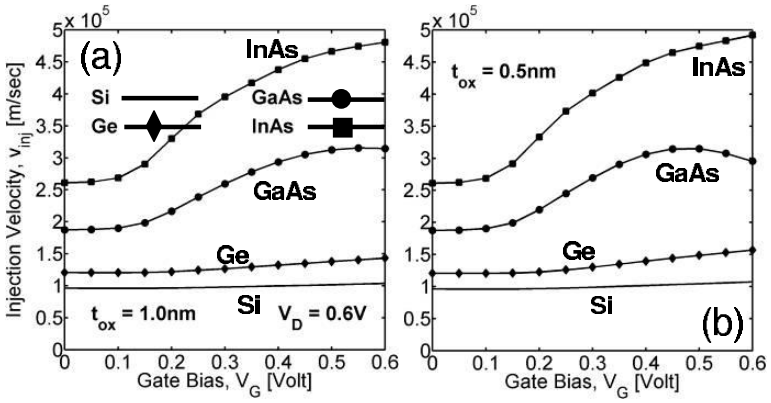

Fig. 6.25. Injection velocity, υ_{inj}, vs. V_G characteristics at V_D = 0.6V for thick body (100 AL) DG n-MOSFETs with two EOTs, (a) 1nm and (b) 0.5nm. III-V channel materials are clear winner due to their low CB m^* along transport. In (a) such enhancement in υ_{inj} offsets the degradation of C_G and in Fig. 6.24(a) III-V materials offer maximum I_{ON}. In (b), however, enhanced υ_{inj} cannot offset degradation of C_G and Ge offers maximum I_{ON}. In (a) and (b), at low V_G, υ_{inj} is for non degenerate carriers, which increases at high V_G due to degenerate carriers following Fermi-Dirac distribution. In (b), υ_{inj} for GaAs drops at high V_G due to occupation of higher L subbands.

to smaller DOS becomes less severe. Si band edge is 6-fold degenerate X type and Ge band edge is 4-fold degenerate L type. For both III-Vs they are singly degenerate Γ type.

The ballistic I-V calculations for these thick body devices were computed for two different EOTs, 1.0nm and 0.5 nm. In contrast to the I-V results for the 3nm (19 AL) bodies (Fig. 6.20), now Fig. 6.24a shows that among 14nm thick devices, ballistic I_{ON} for III-V devices (GaAs and InAs) exceeds Si and Ge performances. However, in Fig. 6.24b, when a thinner EOT (0.5nm) was used for the same thick body, III-V devices fall back, and Ge device offered maximum I_{ON}. For fair comparison, I_{OFF} = 0.5 μ A/ μ m obtained in all devices by tuning gate work function. Thick Si always compares worst due to carriers occupying the low velocity 4-fold degenerate primed (X_4) subband ladder. For thick bodies, due to multiple subband occupation, the

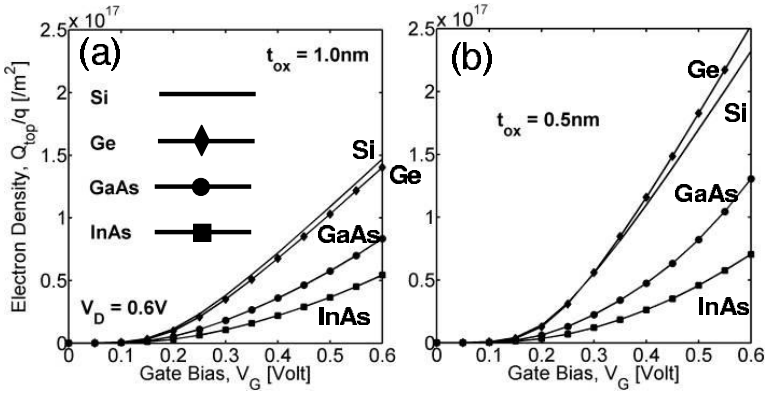

Fig. 6.26. Carrier density Q_{top} vs. V_G at $V_D = 0.6$V for thick body (100 AL) DG n-MOSFETs with two EOTs, (a) 1nm and (b) 0.5 nm. Current, I_D, is the product of Q_{top} and υ_{inj}, and here, Q_{top} is presented. Although threshold voltage, V_T, is adjusted for all devices to obtain same I_{OFF}, C_G is different for different channel materials with same EOT. C_G is the series combination of C_{OX} and C_S, where C_S depends on DOS at band edge. Low CB DOS in III-V materials always results in lower C_S and so have degraded C_G. Doubling C_{OX} does not double Q_{top} due to series C_S.

semiconductor capacitance, C_S, is higher than that of thin bodies. This C_S is in series with C_{OX}, therefore, although C_{OX} for 6.24b is double of that in 6.24a, I_{ON} is less than twice. These results show that maintaining a high value of the ratio C_S/C_{OX} is crucial to maximize I_ON.

Figures 6.25 and 6.26 present the υ_{inj} and Q_{top} vs. V_G plots for 100 AL devices. For 1nm EOT, the degradation of Q_{top} in III-V devices is offset by the enhancement of υ_{inj}. For 0.5nm EOT, however, the degradation of Q_{top} is too severe and the III-V material currents do not increase as expected. These results can be explained as follows: the semiconductor capacitance, C_S (also known as quantum capacitance, C_Q) is a function of the 2D DOS. The 2D DOS depends on number of subbands occupied, number of valleys participating, and the DOS of each such subband. The gate capacitance C_G, which determines Q_{top}, is the series combination of insulator capacitance, C_{OX} (function of EOT), and C_S. When $C_{OX} << C_S$ (thick EOT and thick body) effects of C_S are minimal on C_G, therefore, υ_{inj} determines the ballistic

I_{ON}. However, if $C_{OX} \geq C_S$ (thin EOT and thin body), the degradation of C_G due to C_S is severe and consequently I_{ON} suffers. Finally, thick body Si always compares poorly because of carriers populating the X_4 valleys, two of which have very low v_{inj}.

6.3.3 Summary and Conclusion

Bandstructure plays a crucial role in the ballistic performance of novel channel material MOSFETs. The 2D DOS and v_{inj} are the most important quantities determined by bandstructure. A low DOS degrades C_S, which lowers C_G and hence, reduces Q_{top} for the same gate overdrive. The degradation due to DOS becomes more severe as EOT becomes thinner. Such degradation is often associated with an enhancement in v_{inj}, however, only when both the body and EOT are thick enough, can an enhanced v_{inj} offsets the reduction of C_G and increases ballistic I_{ON}. As a result, III-V devices offer maximum ballistic I_{ON} for a thick body and thick EOT. For deeply scaled DG FETs, however, a Ge channel is found to be the best choice in the ballistic limit.

7. APPLICATION OF TIGHT-BINDING APPROACH WITH SELF-CONSISTENT ELECTROSTATICS

In chapter 6, band bending along the thickness direction due to electrostatic potential was ignored while treating the ultra-thin-body (UTB) dual-gate (DG) MOSFETs. Such band bending is negligible when the body thickness is scaled below 3-4 nm, and can be safely ignored. For thicker bodies, however, ignoring band bending may introduce serious error, especially, when strong electric field limits the spatial extent of the wavefunction within a small part of the device. Treatment of electrostatic potential is necessary to treat gate electrostatics in planar, single-gate, bulk MOSFETs. Additionally, band bending due to applied electrostatic potential plays an important role in the operation of heterostructure-on-insulator (HOI) devices [128, 129]. HOI devices rely on the band discontinuity of various material regions to preferentially populate the high mobility regions and therefore, improving I_{ON}. For HOI devices operating at high gate bias, band bending may override band discontinuity and so accurate treatment of band bending is important.

The objective of this chapter is to use a self-consistent tight-binding approach, where Poisson's equation is iteratively solved with the bandstructure calculation. we will see that this approach explains some experimentally observed phenomena in deeply scaled CMOS devices, and offers an improved understanding of their operation. In Sec. 7.1, the method to incorporate self-consistent electrostatics within the tight-binding formalism is described. We next apply this method to treat two different experimental device structures: single-gate, planar CMOS devices and HOI p-devices. In Sec. 7.2, the experimentally observed mobility vs. gate field behavior in uniaxially and biaxially strained bulk Si p-MOSFETs is explained in terms of their

ballistic behavior. Then in Sec. 7.3, the charge profile inside HOI p-type devices is explored and their C-V characteristics calculated.

7.1 Overall Simulation Approach: Self-consistent Electrostatics

Treatment of self-consistent electrostatics demands heavy computational burden since, 2-D bandstructure in the transport-width plane now depends on the electrostatic potential at every atomic layer. The process is schematically presented in Fig. 7.1. Unlike UTB bandstructure calculation, where bandstructure needs to be calculated only once, bandstructures now must be calculated until it converges with the 1-D Poisson equation along thickness direction. In Fig. 7.2, the domain for atomistic bandstructure calculation and that for Poisson equation are shown. The Poisson domain is much larger compared to tight-binding domain. Using entire Poisson domain for tight-binding calculation is computationally prohibitib and also unnecessary, since at reasonable gate bias, band bending due to external electrostatic potential limits the spatial extent of the wavefunction and tight-binding domain does not need to include the region where the wavefunction practically vanishes.

The approach is as follows. For a given gate bias, we first assume an electrostatic potential profile for the 1-D Poisson equation. From this we then interpolate the potential within the tight-binding domain and get the potential energy at each atomic layer. Adding the potential at each atomic layer to its on-site energy, we then calculate the 2-D bandstructure over the entire $k_x - k_y$ plane. Each point in the k-space is then filled with carriers, according to either source or drain Fermi levels, grouped according to velocity component along channel, and the charge is multiplied by the modulus square of the tight-binding wavefunction to get the spatial charge profile. Sum of all charge profiles over the entire 2-D Brillouin zon gives the charge profile along thickness. This charge is then used as an input to Poisson equation, which calculates the correction needed in the potential profile. The corrected potential profile is then used again to calculate bandstructure, and the steps are repeated

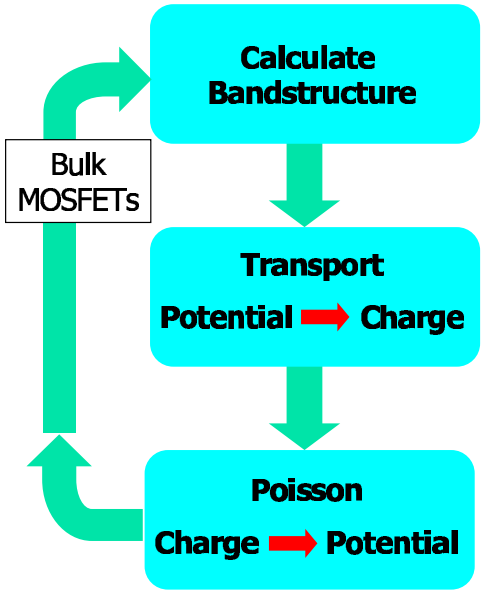

Fig. 7.1. Bandsturctures must be calculated many times before it is self-consister

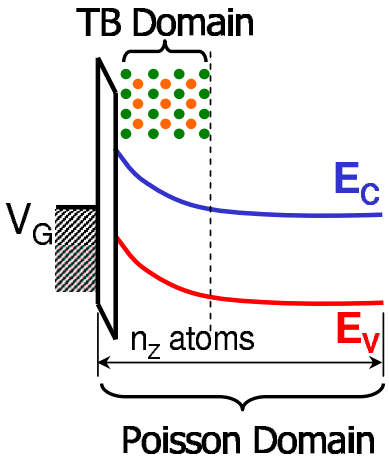

Fig. 7.2. The Poisson equation domain and the tight-binding band-structure calculation domain.

until the charge profile becomes self-consistent with the potential profile. After self-consistence is achieved, we calculate the ballistic current from the self-consistent bandstructure and move to next gate and drain bias.

7.2 Application: Mobility vs. Ballisticity in Strained Bulk single-gate COMS Devices

7.2.1 Background

Improvement of transport properties of carriers in CMOS devices, by the use of strained materials, has been known for long time. It is experimentally observed that correct type of strain enhances both electron and holes mobility.

Over three decades, use of biaxial strain remained the prime technique to incorporate strain in the channel of the device. Biaxialy strained Si is achieved by growing silicon layers epitaxially on relaxed $Si_{1-x}Ge_x$ substrate. On the plane of epitaxial growth, silicon layers achieve the same in-plane lattice constant as $Si_{1-x}Ge_x$, which is determined by the linear interpolation of those from Si and Ge as,

$$a_{SiGe} = (1 - x)\, a_{Si} + x a_{Ge}.$$

Lattice mismatch between relaxed Si and Ge is about 4%, therefore, as x is changed from 0 to 1, the strain in the epitaxially grown silicon layer changes linearly from 0 to 4%. The grown Si layer is under biaxial tensile strain since for [001] growth direction, silicon crystal along both [100] and [010] are under tension. Due to this biaxial strain, the lattice constant along [001] becomes less that that of relaxed Si and the strained silicon thus grown is called pseudomorphically grown silicon. Since the SiGe substrate acts as strain inducing region, biaxial strain is often called *substrate-induced* strain. Due to matured growth techniqe, this has been the most common way of introducing strain in the MOSFET channel. As the thickness of the pseudomorphic layer increases, it stores more and more elastic energy, and beyond a critical thickness (depending on x in substrate), stacking fault or dislocation appear, which relaxes the strain and the grown layer ceases to be a single crystal. A MOSFET with biaxial strain in channel is shown in the middle of Fig. 7.3.

Strain can also be induced in the channel during device processing and packaging. In the bottom of Fig. 7.3, we find the demonstration of *process-induced* strain. Here,

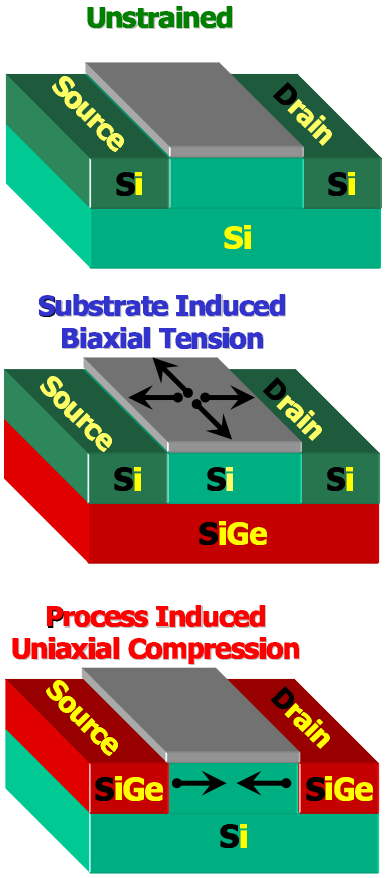

Fig. 7.3. MOSFETs with unstrained and strained channel. Top: unstrained, middle: biaxially tensile strain, bottom: uniaxialy compressive strain

SiGe is grown in the source and drain regions by selective epitaxy. This SiGe in source and drain, with $a_{SiGe} > a_{Si}$, tries to expand and therefore, causes compression along the channel direction. Unlike the biaxial case, the strain in channel is now compressive and acts *uniaxialy*, along the channel. We next will show that mobility behavior of different types strains are also very different.

In Fig. 7.4, the hole mobility vs. gate effective field characteristics for devices with substrate induced and process induced strain are presented and compared with the unstrained behavior. The biaxial strain results by Rim *et al.* show that [130,131], hole mobility is enhanced over relaxed (unstrained) universal mobility at low gate field only. As gate field increases, such improvement diminishes and finally totally

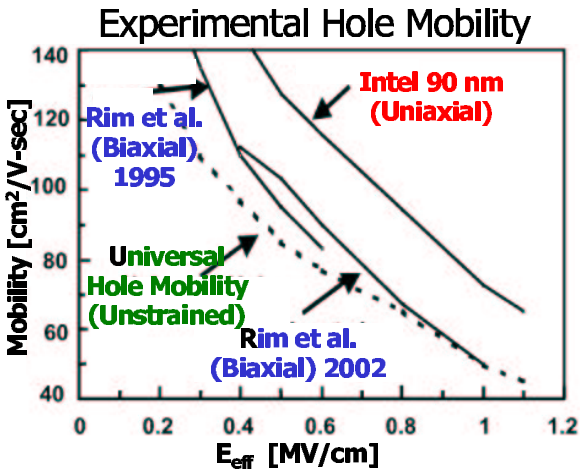

Fig. 7.4. Experimentally observed hole mobility vs. effective field characteristics for strained and unstrained p-MOSFETs

lost at high field. The uniaxialy strained device, however, consistently offers improved hole mobility over the entire gate field range. Using self-consistent tight-binding approach, where strain is treated atomistically within $sp^3d^5s^*$-SO formalism, we will next show that such behavior is directly related to the bandstructure modulation of silicon under combined effecs of strain and quantum confinement—a phenomena also directly related to their ballistic characteristics.

7.2.2 Approach

When applied gate bias depletes and eventually inverts the channel, it creates a potential well along the thickness direction in which carriers are quantum mechanically confined. Their energy is quantized along this direction and discrete subbands are formed. The degree of quantum mechanical confinement depends on the gate bias—higher is the bias, stronger is the confinement. The direction of confinement is always along the thickness, which is true for both the strained devices presented in Fig. 7.3. The relative orientation of the compressive strain direction, with respect to quantum confinement direction, is however different in the two strained devices. For the substrate induced biaxialy strained device, quantum confinement axis is aligned

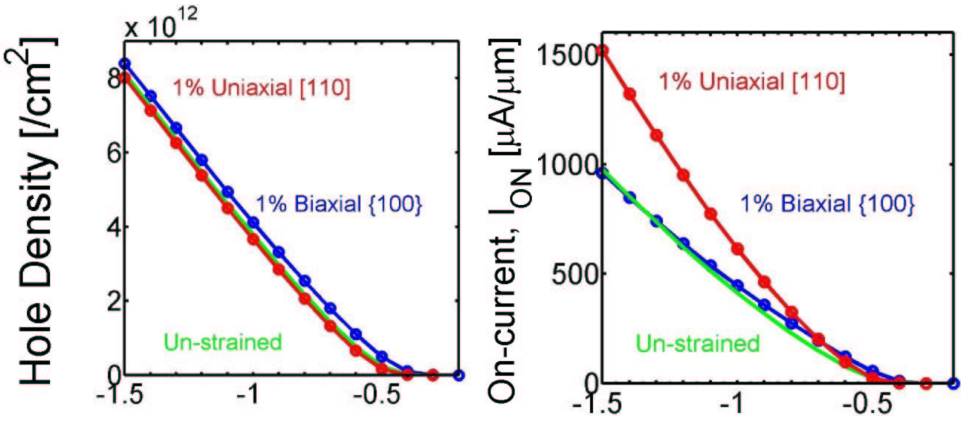

Fig. 7.5. Ballistic Q-V (left) and I-V (right) plot for three different p-MOSFETs. Green: unstrained, blue: biaxialy strained, red: uniaxialy strained.

along the compressive strain axis, while for the process-induced strained device, they are perpendicular to each other. Since in each device strain is a fixed in direction and magnitude and quantum confinement is what changing with gate bias, we explore this orientation dependence by simulating these two device structures for the gate bias range, V_G =0- -1.5V, at V_D =-1.5V. The result is presented in next section.

7.2.3 Results and Discussion

First, we use the self-consistent tight-binding approach, described in Sec. 7.1, to calculate the Q-V characteristics for the strained and unstrained devices in Fig. 7.3 at high drain bias, V_D =1.5V. This result is presented in Fig. 7.5 (left). Here we observe that, above threshold voltage, V_T, all devices, strained and unstrained, have almost same charge at the beginning of the channel. This happens because in p-MOSFETs, the amount of charge is controlled by gate electrostatics, not the bandstructure. Recall the results in Fig. 6.22 on page 97, where similar results were observed. However, the I_D-V_G plots on right of Fig. 7.5, show strong dependence of currents on the strain. As it is for the experimental case, the current is calculated along [100] for both unstrained and biaxialy strained p-MOSFET, but it is calculated

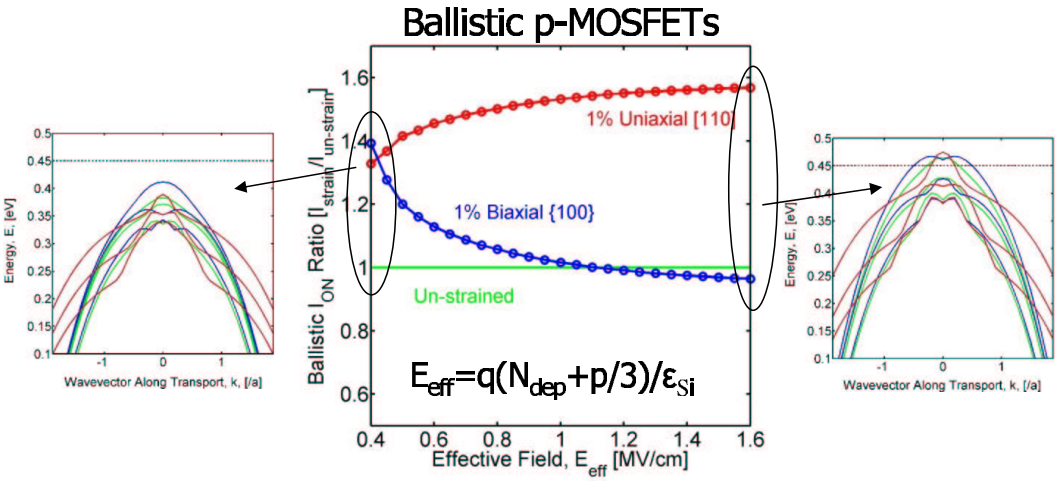

Fig. 7.6. Ballistic current ratio showing performance enhancement for the two strained p-MOSFET with respect to the unstrained device. The left and right plots show bandstructure at low and high gate field

along [110] for uniaxialy strained device. The reason for such difference in currents is in their bandstructures, the curvature of which determines the group velocity of the carriers, and hence current.

We now compare the I-V results in Fig. 7.5 in the same way mobility were compared in Fig. 7.4. First the gate bias is changed to effective gate field, using the well known empirical formula,

$$E_{eff} = \frac{q}{\epsilon_{Si}} \left(N_{dep} + p/3 \right) .$$

Then the ratio of the ballistic currents, $I_{Strained}/I_{NoStrain}$, is plotted against effective field, E_{eff}, in Fig. 7.6. This ratio shows a unique characteristics, and it is equivalent to the experimental results presented in Fig. 7.4. Here we observe that, at low gate bias the biaxialy strained device offers enhanced performance over unstrained device, but as the gate field increases, the currents goes down and finally all strain advantage in ballistic current is diminished at high gate field. Uniaxially strained device, however, consistently offers enhanced ballistic performance over the entire range of gate field. The explanation for this behavior can be found in the bandstructure plots

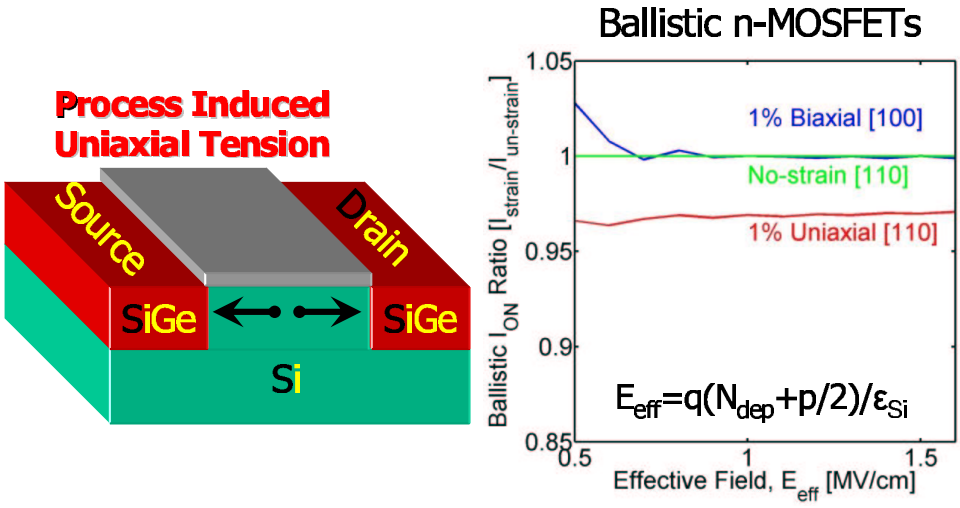

Fig. 7.7. Left: n-MOSFET under uniaxial tensile strain. Right: ballistic current ration of strained and unstrained current ratio is almost unity. Thus ballistic simulation of strained n-MOSFETs cannot explain why strain improves electron mobility.

on the same figure. On the left, at low gate bias and therefore, negligible quantum confinement, we find that strain has lifted the heavy and light hole band degeneracy in both biaxially and uniaxialy strained devices. Here, the low effective mass, high velocity, light-hole band forms the top subband, thereby improving strained mobility. On the right plot, where bandstructure at high gate field is presented, the scenario is very different. Here we observe that, quantum confinement has nullified all strain advantage in the biaxially strained device, where the top subband for this device is now heavy-hole like (with a dip in the middle). This explains why the performance enhancement for biaxially strained device diminished at high gate field. Bandstructure for uniaxially strained device at high field, however, is unaffected by quantum confinement, and the top subband is still light-hole like and thereby enhanced performance is retained.

After successfully explaining the mobility vs. effective field behavior in strained p-MOSFETs, we repeat the same simulation for n-MOSFETs. Experimentally, in n-MOSFETs, mobility enhancement is also observed for both biaxial and uniaxial

tensile strain. In Fig. 7.7, we see that bandstructure effects cannot explain such effect in n-FETs since there is no enhancement of ballistic I_{ON} under strain. Originally, Takagi *et al.*, proposed that strain lifts the degeneracy between the X_2 and X_4 conduction band valleys, which causes improvement of electron mobility [132]. Later, Fischetti *et al.* [133], has shown by detailed simulation that bandstructure alone cannot explain the mobility enhancement in n-MOSFETs under strain. Strain induced lifting of degeneracy of the X_2 and X_4 valleys has marginal effect since, in the channel of an n-MOSFETs, as a result of quantum confinement, this degeneracy has already been sufficiently lifted during formation of unprimed and primed ladder. Our results on strained n-MOSFETs are consistent with the conclusion in [133], and we see *no-effect* of strain on conduction band, which can explain electron mobility enhancement.

7.2.4 Summary and Conclusion

In this section, a 20 orbital $sp^3d^5s^*$-SO strained tight-binding model is solved self-consistently with Poisson equation to incorporate self-consistent electrostatics in planar single gate strained CMOS devices. The results show that experimentally observed mobility vs. gate field behavior in strained p-MOSFETs can be explained by the unique bandstructure modulation under strain and quantum confinement. However, for bulk n-MOSFETs, the origin of mobility enhancement is still an open question and ballistic behavior, therefore, bandstructure effects, alone cannot explain .

7.3 Application: C-V Characteristics and Hole Density Profile in Strained HOI p-MOSFETs

In this section we will demonstrate the application of self-consistent tight-binding approach by simulating the gate bias dependence of the hole density profile and the C-V characteristics of a p-type heterostructure-on-insulator (HOI) device structure.

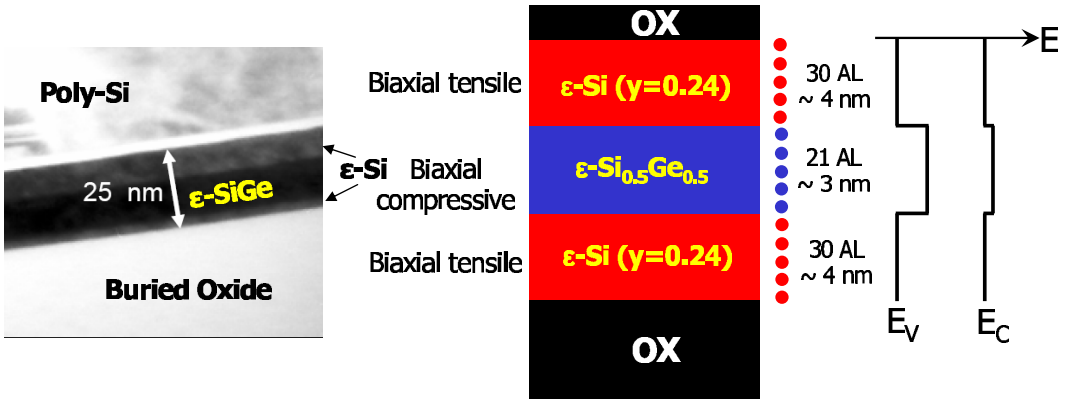

Fig. 7.8. Left: Fabricated HOI structure Prof. Hoyt's group, MIT. Center: The model device for simulation. Top and bottom Si layers are under biaxial tension while the central SiGe layer is under biaxial compression. Right: The conduction and valence band profile along thinkness.

It has already been experimentally demonstratrated that such devices can offer improved transport properties for both holes and electrons [128, 129].

The HOI structure shown in Fig. 7.8 operates by exploiting the band discontinuity between strained Si and SiGe layer to preferentially populate the carriers in high-mobility strained regions. We will primarily focus on the gate electrostatics of p-type HOI devices here. The structure is pseudomorphically grown along thickness direction and the lattice constant in the plane normal to the growth direction is same for all three layers: Si, SiGe and Si. The whole structure sits on a thick bottom oxide layer, below which bottom gate is located. All three layers are strained, top layer was originally grown on a relaxed $Si_{1-y}Ge_y$ layer with y =0.24 and hence is under biaxially tensile strain. The Ge content, y, fixes the in-plane lattice constant for all three layers. The Ge content of the central $Si_{1-z}Ge_z$ layer is z =0.5, and since $z > y$, this central layer is under biaxialy compressive strain. Finally, the bottom Si layer is identical to the top Si layer, except its thickness may be different.

As the Ge content in SiGe changes from 0% to 100%, the band gap of SiGe also changes from that of Si to that of Ge. For any intermidiate value of Ge mole fraction,

the band gap is in-between of that of Si and Ge. Strain modulate the bandgaps of both Si and SiGe in a comlicated way. Band discontinuity between Si and SiGe appears mostly in the valence band and in Fig. 7.8 we find a quantum well for holes in the central SiGe layer.

7.3.1 Approach

We have simualted a HOI structure where both top and bottom Si layers are 30 atomic layers (≈4nm) thick, and the central SiGe layer is 21 atomic layers (≈3nm) thick. For the central SiGe layer, since z =0.5, there are equal numbers of Si and Ge atoms. For this layer we considered an ordered structure consisting of alternatively stacked Si type and Ge type atomic layers. The tight-binding Hamiltonaian is written in $sp^3d^5s^*$-SO basis including all three layers and is a 1620×1620 matrix for each $\{k_x - k_y\}$ point in the first 2-D Brillouin zone. The Si-SiGe hetero-interfaces are treated in a symmetric way to ensure Hermicity of the Hamiltonian. The top and bottom insulators layers are 2 and 10 nm, respectively, with bottom gate always considered grounded. For a given bias at the top gate, Poisson equation is solved self-consistently with the bandstructure calculation and the charge profile calculated. The Poisson simulation domain consists of both insulators as well as the hetero-channel body, however, bandstructure is calculated only in the body, with zero boundary condition for the wavefunction at the top and bottom insulator interfaces.

7.3.2 Results and Discussion

In Fig. 7.9, valence band profile and hole density profile for the HOI structure is presented for the gate bias range, V_G = 0 to -0.4V. The Q-V characteristics on left shows that the device is operating in the subthreshold region. The applied bias drops within both top and bottom insulators, as well as, in the body itself. The hole density profile in the bottom-right plot of Fig. 7.9 shows that the holes are preferentially occupying the central SiGe region, where, due to band discontinuity

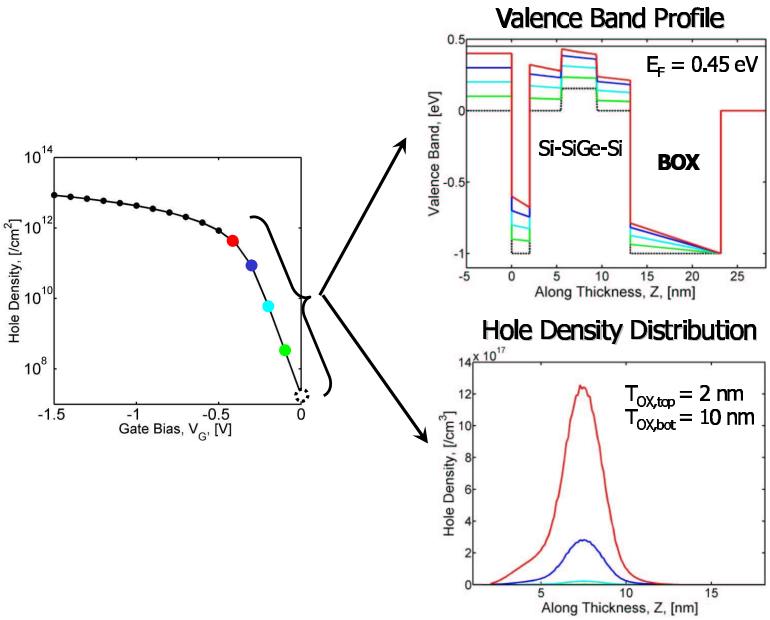

Fig. 7.9. V_G varied from 0 to -0.4V. Left: The device is in subthreshold. Top: The potential drops both in top and bottom insulator, as well as in the body. Bottom: Holes occupy only the central SiGe region

a quantum well is present for holes. It is also observed that, not only holes stay in the central region, but also for any incremental bias, the additional holes also accumulate in this central region.

In Fig. 7.10, valence band profile and hole density profile for the HOI structure is presented for the gate bias range, V_G = -0.5 to -0.9V. Here the carrier concentration vs. gate bias plot on left shows that the device has come out of the subthreshold region and operates just above threshold voltage. The valence band profile in the top-right shows that the bottom insulator, as well as, the bottom Si layer and the central SiGe layers are almost electrostatically isolated from the top insulator and the top Si layer. We observe that the applied bias drops across these two regions only. The hole density profile in the bottom-right plot of Fig. 7.10 is different from what observed in Fig. 7.9. Now we see that the hole density profile is splitted between the top Si layer and the central SiGe layer. The electrostatic band bending has created

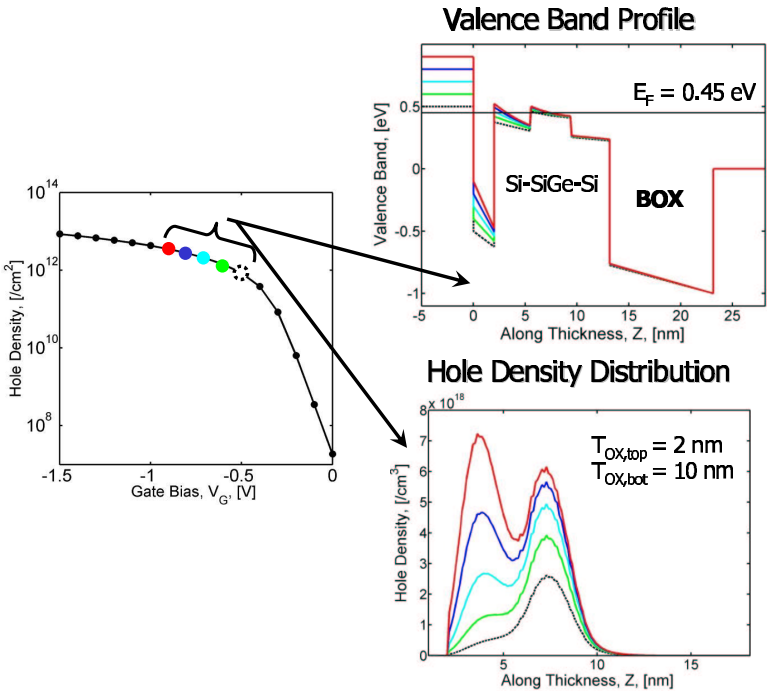

Fig. 7.10. V_G varied from -0.5 to -0.9V. Left: The device is at threshold. Top: The gate potential drops in top insulator, top Si layer and the SiGe layer. Bottom: Induced holes are shared between to Si layer and central SiGe layer

an additional quantum well near the surface where the holes now move. It can also be seen that, at the end of the gate bias range all incremental carrier density with gate bias accumulates near the top surface, which offers the lowest energy well for holes.

Finally, in Fig. 7.11, the same plots are shown at high gate bias range, V_G = -1.0 to -1.5V. The left plot shows that the device is well into threshold. The valence band profile profile shows that nearly all the incremental bias in the top gate now drops across the top gate insulator. Most of the hole concentration in the valence band is now located in the top Si layer and with increasing gate bias, the hole are added to this layer.

The hole density profiles for the HOI structure, paresented in Figs. 7.9–7.11, are integrated along the thickness direction, Z, and the hole density vs. gate bias

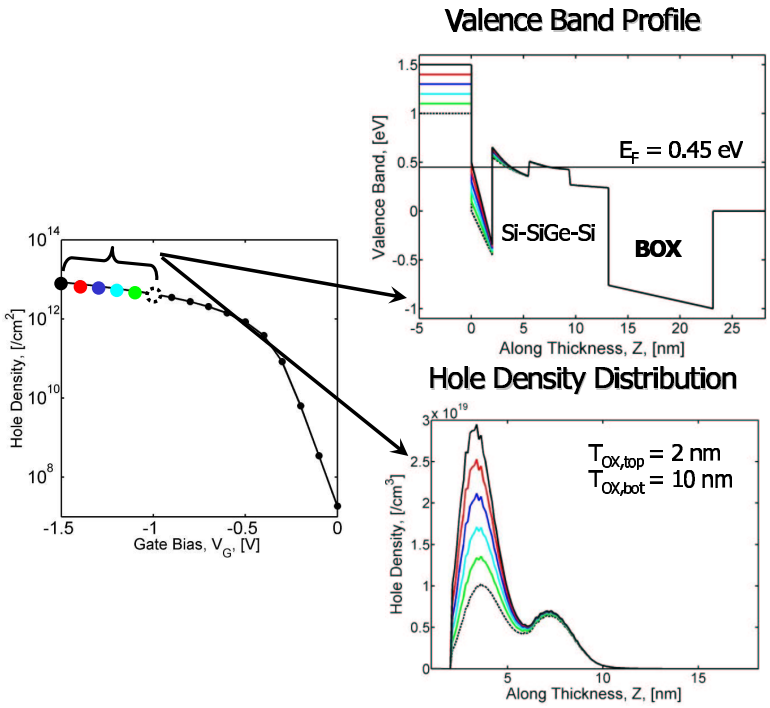

Fig. 7.11. Top gate bias, V_G, varied from -1.0 to -1.5V. Left: The device is in above threshold. Top: The potential drops only in top insulator. Bottom: Holes occupy only the top Si layer

is presented in Fig. 7.12 (left). In the same plot, hole density for an equivalent Strained-Silicon-On-Insulator (SSOI) structure is also shown. The SSOI structure is similar to the HOI except the central SiGe region is replaced by strained Si. In the left plot, two devices behave similarly above threshold voltage, however, the inset shows that the SSOI has slightly higher threshold voltage. This happens because the small bandgap SiGe region is not present in the SSOI device, therefore, the subbands form at higher hole energy. The C-V plot for these two device also show interesting behavior. At low gate bias, capacitance for SSOI is smaller than the HOI structure since the higher V_T SSOI has pushed it in deep subthreshold. However, around or just above V_T, the SSOI structue shows higher capacitance. This happens because, for SSOI all charges are added to the surface channel, while for HOI the charge is added to both central and surface channel. As a result, the equivalent oxide thickness

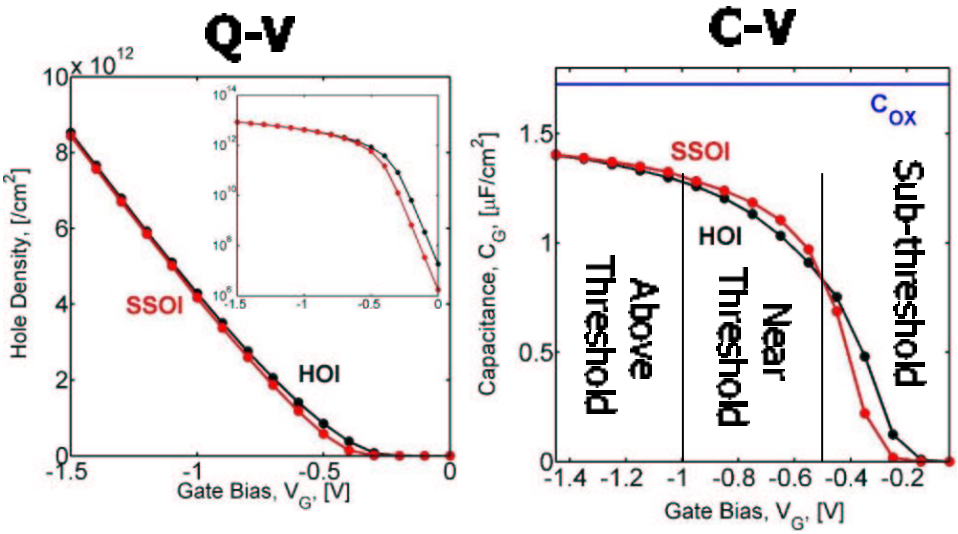

Fig. 7.12. Left: Q-V compared, HOI vs. SSOI. Right: C-V compared, HOI vs. SSOI.

for HOI become larger. Finally, at high V_G, when in both devices, charge is only added at the surface channel, the capacitance for two devices converge. Finally, we observe that the inversion capacitance for both SSOI and HOI are smaller than the gate insulator capacitance, C_{OX}, due to increase of oxide thickness with quantum confinement.

7.3.3 Summary and Conclusion

In this section we have demonstrated application of self-consistent tight-binding approach for HOI device structures. For the given device structure, we found that in subthreshold and around threshold the hole density profile remains close to the central SiGe channel. At high gate bias, however, the device behaves like a SSOI structure and the central SiGe channel becomes irrelevent. Simulation of devices with varying thicknesses of three different regions can find an optimized structure.

8. SUMMARY AND FUTURE WORK

8.1 Summary

This work has provided an improved understanding of the operation and the scaling characteristics of novel channel material nanoscale CMOS devices. The necessity of incorporating an accurate treatment of the full bandstructure of the channel material is highlighted. It has been observed that at the ballistic limit, the carrier density-of-states play a crucial role in determining performance of novel-channel material MOSFETs. Ballistic performances of low density-of-states material MOSFETs degrades drastically when gate insulator capacitance increases in a deeply scaled device. As a result, some of the density-of-states effects, which are not observed at 1-2nm EOT, may play crucial role when EOT is scaled down to 0.5nm. Ballistic performance comparison reveals that the germanium CMOS devices are the most promising candidates for highly-scaled novel channel material CMOS technology. However, when the role of parasitic series resistances in the source and drain become the limiting factor, high mobility materials such as III-V material MOSFETs show promise to outperform Si or Ge MOSFETs.

8.2 Future Work

A list of possible future works, directly related to this research, is presented below:

- **Explanation of experimental mobility behavior**: Recently, a great deal of experimental results has been published in IEDM, TED and EDL on the strained mobility behavior in deeply scaled planar and SOI CMOS devices. Such mobility data incorporates both bandstructure effects (i.e. effective mass)

and effects of carrier scattering. The top-of-the-barrier model in chapter 4 and the tight-binding approach in chapters 5–7 correctly incorporates the bandstructure effects displayed in the mobility data; however, completely ignores the scattering. A very useful study of the carrier transport in nanoscale devices can be performed by identifying a few key experiments from the published literature, and then using the theoretical approaches of this thesis to separate how much effect is due to bandstructure and how much effect is due to scattering.

- **Alternative wafer orientations**: In this thesis, only UTB devices with (100) wafer orientations have been treated. However, the top-down approach to fabricate UTB devices and FinFETs can result in an arbitrary orientation for the thickness direction. For a complete study of the orientation effects, it is necessary to generalize the approach and discretize the tight-binding Hamiltonian for wafer orientations other than (100); then calculate the associated bandstructures and device performances. A careful bookkeeping of the bond orientations between different layers along thickness is crucial for treating a general orientation.

- **Tight-binding Hamiltonian in mode-space**: A full 2-D description of the nanoscale MOSFET using tight-binding approach is prohibitive, computationally. However, if this 2-D Hamiltonian can be expanded in terms of a few coupled modes along confinement, a reduced Hamiltonian can be obtained which is suitable for the NEGF quantum transport treatment. Such technique can prove enormously useful for a better understanding of the quantum transport in nanoscale CMOS.

- **Application to HEMTs and QWFETs**: The self-consistent tight-binding approaches presented in this thesis can be readily applied to III-V devices, such as high-electron-mobility transistors (HEMT) and quantum-well field-effect-transistors (QWFETs)—provided that the tight-binding parameters for the materials are available. Such treatment can be very useful and relevant,

since it is widely believed that compared to Si MOSFETs, these devices operate closer to their ballistic limits due to the exceptionally high carrier mobilities in III-V.

- **Zone unfolding method to treat SRS and alloy**: Recently, zone-unfolding method has been developed within the context of tight-binding approach [64]. A direct consequence of this is, surface-roughness-scattering (SRS) and random alloy fluctuations can now be treated withing the tight-binding approach. Inclusion of these effects will make tight-binding a more complete approach for predictive simulation of nanoscale CMOS devices.

LIST OF REFERENCES

LIST OF REFERENCES

[1] D. Kahng and M. M. Atalla. Silicon-silicon dioxide field induced surface devices. *Proc. IRE-AIEE Solid-State Device Res. Conf., Pittsburgh, PA*, 1960.

[2] P. Bai *et al.* A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 cu interconnect layers, low-k ILD and $0.57\mu m^2$ SRAM cell. *IEEE International Electron Devices Meeting, 2004. IEDM Technical Digest.*, pages 657–660, Dec. 2004.

[3] S. E. Thompson, R. S. Chau, T. Ghani, K. Mistry, S. Tyagi, and M. T. Bohr. In search of "forever," continued transistor scaling one new material at a time. *IEEE Transactions on Semiconductor Manufacturing*, 18(1):26–36, Feb. 2005.

[4] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, pages 114–117, April 19, 1965.

[5] G. E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, Jan. 1998.

[6] International technology roadmap for semiconductors, 2001 edition. www.itrs.net, 2001.

[7] International technology roadmap for semiconductors, 2003 edition. www.itrs.net, 2003.

[8] H. Wakabayashi, T. Ezaki, M. Hane, S. Yamagami, N. Ikarashi, K. Takeuchi, T. Yamamoto, T. Mogami, T. Ikezawa, T. Sakamoto, and H. Kawaura. Transport properties of sub-10-nm planar-bulk-CMOS devices. *IEEE International Electron Devices Meeting, 2004. IEDM Technical Digest.*, pages 429–432, Dec. 2004.

[9] Robert Chau, Boyan Boyanov, Brian Doyle, Mark Doczy, Suman Datta, Scott Hareland, Ben Jin, Jack Kavalieros, and Matthew Metz. Silicon nanotransistors for logic applications. *Physica E*, 19(1/2):1–5, July 2003.

[10] B. Doris, Meikei Ieong, T. Kanarsky, Ying Zhang, R. A. Roy, O. Dokumaci, Zhibin Ren, Fen-Fen Jamin, Leathen Shi, W. Natzle, Hsiang-Jen Huang, J. Mezzapelle, A. Mocuta, S. Womack, M. Gribelyuk, E. C. Jones, R. J. Miller, H. S. P. Wong, and W. Haensch. Extreme scaling with ultra-thin Si channel MOSFETs. *International Electron Devices Meeting, 2002. IEDM '02. Digest.*, pages 267–270, Dec. 2002.

[11] K. Uchida, Junji Koga, and Shin ichi Takagi. Experimental study on carrier transport mechanisms in double- and single-gate ultrathin-body MOSFETs - coulomb scattering, volume inversion, and δT_{SOI}-induced scattering. *IEEE International Electron Devices Meeting, 2003. IEDM '03 Technical Digest.*, pages 33.5.1–33.5.4, Dec. 2003.

[12] R. S. Shenoy and K. C. Saraswat. Optimization of extrinsic source/drain resistance in ultrathin body double-gate FETs. *IEEE Transactions on Nanotechnology*, 2(4):265–270, Dec. 2003.

[13] R. S. Shenoy and K. C. Saraswat. Novel process for fully self-aligned planar ultrathin body double gate FET. *Proceedings of IEEE International SOI Conference, 2004*, pages 190–191, 2004.

[14] B. S. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau. High performance fully-depleted tri-gate CMOS transistors. *IEEE Electron Device Letters*, 24(4):263–265, April 2003.

[15] B. Doyle, B. Boyanov, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, R. Rios, and R. Chau. Tri-gate fully-depleted CMOS transistors: fabrication, design and layout. *Digest of Technical Papers. 2003 Symposium on VLSI Technology*, pages 133–134, June 2003.

[16] R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros, and M. Metz. High-κ metal-gate stack and its MOSFET characteristics. *IEEE Electron Device Letters*, 25(6):408–410, June 2004.

[17] R. Chau, S. Datta, M. Doczy, B. Doyle, B. Jin, J. Kavalieros, A. Majumdar, M. Metz, and M. Radosavljevic. Benchmarking nanotechnology for high-performance and low-power logic transistor applications. *IEEE Transactions on Nanotechnology*, 4(2):153–158, March 2005.

[18] T. Ghani *et al.* A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors. *IEDM 2003 Technical Digest.*, pages 11.6.1–11.6.3, 8-10 Dec. 2003.

[19] K. Mistry *et al.* Delaying forever: Uniaxial strained silicon transistors in a 90nm CMOS technology. *Digest of Technical Papers, 2004 Symposium on VLSI Technology*, pages 50–51, 15-17 June 2004.

[20] S. Thompson *et al.* A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 μm^2 SRAM cell. *IEDM 2002. Digest.*, pages 61–64, 8-11 Dec. 2002.

[21] S. E. Thompson *et al.* A 90-nm logic technology featuring strained-silicon. *IEEE Transactions on Electron Devices*, 51(11):1790–1797, Nov. 2004.

[22] S. E. Thompson *et al.* A logic nanotechnology featuring strained-silicon. *IEEE Electron Device Letters*, 25(4):191–193, April 2004.

[23] K. Rim *et al.* Fabrication and mobility characteristics of ultra-thin strained si directly on insulator (SSDOI) MOSFETs. *IEDM '03 Technical Digest. IEEE InternationalElectron Devices Meeting*, pages 3.1.1–3.1.4, Dec. 2003.

[24] I. Aberg, O. O. Olubuyide, C. Ni Chleirigh, I. Lauer, D. A. Antoniadis, J. Li, R. Hull, and J. L. Hoyt. Electron and hole mobility enhancements in sub-10 nm-thick strained silicon directly on insulator fabricated by a bond and etch-back technique. *Digest of Technical Papers–Symposium on VLSI Technology, 2004*, pages 52–53, 2004.

[25] T. Mizuno, N. Sugiyama, T. Tezuka, T. Numata, T. Maeda, and S. Takagi. Thin-film strained-SOI CMOS devices-physical mechanisms for reduction of carrier mobility. *IEEE Transactions on Electron Devices*, 51(7):1114–1121, July 2004.

[26] K. Uchida, R. Zednik, Ching-Huang Lu, H. Jagannathan, J. McVittie, P. C. McIntyre, and Y. Nishi. Experimental study of biaxial and uniaxial strain effects on carrier mobility in bulk and ultrathin-body SOI MOSFETs. *IEEE International Electron Devices Meeting, 2004. IEDM Technical Digest.*, pages 229–232, Dec. 2004.

[27] T., N. Sugiyama, T. Tezuka, and S. Takagi. (110)strained-SOI n-MOSFETs with higher electron mobility. *IEEE Electron Device Letters*, 24(4):266–268, 2003.

[28] M. Yang *et al.* High performance CMOS fabricated on hybrid substrate with different crystal orientations. *IEEE International Electron Devices Meeting 2003*, pages 18.7.1–18.7.4, 2003.

[29] Min Yang *et al.* Performance dependence of CMOS on silicon substrate orientation for ultrathin oxynitride and hfo_2 gate dielectrics. *IEEE Electron Device Letters*, 24(5):339–341, May 2003.

[30] T. Mizuno, N. Sugiyama, T. Tezuka, Y. Moriyama, S. Nakaharai, and S. Takagi. (110)-surface strained-SOI CMOS devices. *IEEE Transactions on Electron Devices*, 52(3):367–734, March 2005.

[31] M. Myronov, T. Irisawa, S. Koh, O. A. Mironov, T. E. Whall, E. H. C. Parker, and Y. Shiraki. Temperature dependence of transport properties of high mobility holes in Ge quantum wells. *Journal of Applied Physics*, 97(8):1–6, 2005.

[32] C. O. Chui, S. Ramanathan, B. B. Triplett, P. C. McIntyre, and K. C. Saraswat. Germanium MOS capacitors incorporating ultrathin high-κ gate dielectric. *IEEE Electron Device Letters*, 23(8):473–475, Aug. 2002.

[33] Chi On Chui, Hyoungsub Kim, P. C. McIntyre, and K. C. Saraswat. A germanium nMOSFET process integrating metal gate and improved hi-κ dielectrics. *IEEE International Electron Devices Meeting 2003*, pages 18.3.1–18.3.4, 2003.

[34] D. Chi, Chi On Chui, K. C. Saraswat, B. B. Triplett, and P. C. McIntyre. Zirconia grown by ultraviolet ozone oxidation on germanium (100) substrates. *Journal of Applied Physics*, 96(1):813–819, July 2004.

[35] Chi On Chui, H. Kim, P. C. McIntyre, and K. C. Saraswat. Atomic layer deposition of high-κ dielectric for germanium MOS applications - substrate. *IEEE Electron Device Letters*, 25(5):274–276, May 2004.

[36] M. L. Lee, C. W. Leitz, Z. Cheng, A. J. Pitera, T. Langdo, M. T. Currie, G. Taraschi, E. A. Fitzgerald, and D. A. Antoniadis. Strained Ge channel p-type metal-oxide-semiconductor field-effect transistors grown on $Si_{1-x}Ge_x$/Si virtual substrates. *Applied Physics Letters*, 79(20):3344–3346, 2001.

[37] H. Shang *et al.* Electrical characterization of germanium p-channel MOSFETs. *IEDM Tech. Digest*, pages 441–444, 2002.

[38] H. Shang, H. Okorn-Schimdt, J. Ott, P. Kozlowski, S. Steen, E. C. Jones, H. S. P. Wong, and W. W. Hanesch. Electrical characterization of germanium p-channel MOSFETs. *IEEE Electron Device Letters*, 24(4):242–244, 2003.

[39] Chi On Chui and K. C. Saraswat. Advanced germanium MOSFET technologies with high-κ gate dielectrics and shallow junctions. *2004 International Conference on Integrated Circuit Design and Technology*, pages 245–252, 2004.

[40] Chi On Chui, F. Ito, and K. C. Saraswat. Scalability and electrical properties of germanium oxynitride MOS dielectrics. *IEEE Electron Device Letters*, 25(9):613–615, Sept. 2004.

[41] A. Nayfeh, Chi On Chui, T. Yonehara, and K. C. Saraswat. Fabrication of high-quality p-MOSFET in Ge grown heteroepitaxially on Si. *IEEE Electron Device Letters*, 26(5):311–313, May 2005.

[42] P. D. Ye, B. Yang, K. K. Ng, J. Bude, G. D. Wilk, S. Halder, M. Marbell, T. Hierl, and J. C. M. Hwang. RF power performance of GaAs MOSFETs with Al_2O_3 gate dielectric grown by atomic layer deposition. *Conference Proceedings. 34th European Microwave Conference*, 2:901–903, 2004.

[43] P. D. Ye *et al.* GaAs metaloxidesemiconductor field-effect transistor with nanometer-thin dielectric grown by atomic layer deposition. *Applied Physics Letters*, 83(1):180–182, 2003.

[44] P. D. Ye, G. D. Wilk, J. Kwo, B. Yang, H. J. L. Gossmann, M. Frei, S. N. G. Chu, J. P. Mannaerts, M. Sergent, M. Hong, K. K. Ng, and J. Bude. GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition. *IEEE Electron Device Letters*, 24(4):209–211, April, 2003.

[45] X. Li, Y. Cao, D. C. Hall, P. Fay, B. Han, A. Wibowo, and N. Pan. GaAs MOSFET using InAlP native oxide as gate dielectric. *IEEE Electron Device Letters*, 25(12):772–774, Dec. 2004.

[46] T. Ashley *et al.* Novel InSb-based quantum well transistors for ultra-high speed, low power logic applications. *7th International Conference on Solid-State and Integrated Circuits Technology Proceedings*, 3:2253–2256, 2005.

[47] M. V. Fischetti and S. E. Laux. Monte Carlo simulation of transport in technologically significant semiconductors of the diamond and zinc-blende structures. ii. submicrometer MOSFET's. *IEEE Transactions on Electron Devices*, 38(3):650–660, March 1991.

[48] F. Assad, Z. B. Ren, D. Vasileska, S. Datta, and M. Lundstrom. On the performance limits for Si MOSFET's: A theoretical study. *IEEE Transactions on Electron Devices*, 47(1):232–240, 2000.

[49] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic. Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches. *Journal of Applied Physics*, 92(7):3730–3739, 2002.

[50] Z. B. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom. nanoMOS 2.5: A two-dimensional simulator for quantum transport in double-gate MOSFETs. *IEEE Transactions on Electron Devices*, 50(9):1914–1925, Sept. 2003.

[51] R. Venugopal, M. Paulsson, S. Goasguen, S. Datta, and M. S. Lundstrom. A simple quantum mechanical treatment of scattering in nanoscale transistors. *Journal of Applied Physics*, 93(9):5613–5625, 2003.

[52] R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom. Quantum mechanical analysis of channel access geometry and series resistance in nanoscale transistors. *Journal of Applied Physics*, 95(1):292–305, 2004.

[53] 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. *Journal of Applied Physics*, 95(10):5545–5582, 2004.

[54] Z. Ren, R. Venugopal, S. Datta, M. S. Lundstrom, D. Jovanovic, and J. G. Fossum. The ballistic nanotransistor: A simulation study. *IEDM Tech. Dig.*, pages 715–718, 2000.

[55] Z. Ren, R. Venugopal, S. Datta, and M. S. Lundstrom. Examination of design and manufacturing issues in a 10 nm double gate MOSFET using nonequilibrium Greens function simulation. *IEDM Tech. Dig.*, pages 5.4.1–5.4.4, 2001.

[56] Sayed Hasan, Jing Wang, and Mark Lundstrom. Device design and manufacturing issues for 10 nm-scale MOSFETs: a computational study. *Solid-State Electronics*, 48(6):867–875, June 2004.

[57] A. Kumar, J. Kedzierski, and S. E. Laux. Quantum-based simulation analysis of scaling in ultrathin body device structures. *IEEE Transactions on Electron Devices*, 52(4):614–617, April 2005.

[58] S. E. Laux. Simulation study of Ge n-channel 7.5 nm DGFETs of arbitrary crystallographic alignment. *IEEE International Electron Devices Meeting, 2004. IEDM Technical Digest.*, pages 135–138, Dec. 2004.

[59] S. E. Laux. Optimum crystallographic alignment for Si n-channel ballistic DGFETs. *IEEE Electron Device Letters*, Accepted for publication, 2005.

[60] J. C. Slater and G. F. Koster. Simplified LCAO method for the periodic potential problem. *Phys. Rev.*, 94(6):1498–1524, 1954.

[61] P. Vogl, Harold P. Hjalmarson, and John D. Dow. A semi-empirical tight-binding theory of the electronic structure of semiconductors. *Journal of Physics and Chemistry of Solids*, 44(5):365–378, 1983.

[62] Jean-Marc Jancu, Reinhard Scholz, Fabio Beltram, and Franco Bassani. Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters. *Phys. Rev. B*, 57(11):6493–6507, 1998.

[63] T. B. Boykin, G. Klimeck, P. von Allmen, Seungwon Lee, and F. Oyafuso. Valley splitting in V-shaped quantum wells. *Journal of Applied Physics*, 97(11):113702(1–6), 2005.

[64] T. B. Boykin and G. Klimeck. Practical application of zone-folding concepts in tight-binding calculations. *Physical Review B*, 71(11):115215(1–6), 2005.

[65] G. Klimeck. NEMO 1-D: the first NEGF-based TCAD tool. *Simulation of Semiconductor Processes and Devices - SISPAD 2004*, pages 9–16, 2004.

[66] Gerhard Klimeck, Fabiano Oyafuso, Timothy B. Boykin, R. Chris Bowen, and Paul von Allmen. Development of a nanoelectronic 3-d (NEMO 3-d) simulator for multimillion atom simulations and its application to alloyed quantum dots. *Computer Modeling in Engineering & Sciences (CMES)*, 3(5):601–642, 2002.

[67] M. Städele, B. R. Tuttle, and K. Hess. Tunneling through ultrathin SiO_2 gate oxides from microscopic models. *Journal of Applied Physics*, 89(1):348–363, January 2001.

[68] Fabio Sacconi, Michael Povolotskyi, Aldo Di Carlo, Paolo Lugli, Martin Städele, Christian G. Strahberger, and Peter Vogl. Full-band approaches for the quantum treatment of nanometer-scale MOS structures. *Physica B*, 314(1–4):345–349, March 2002.

[69] M. Städele, A. Di Carlo, P. Lugli, F. Sacconi, and B. Tuttle. Atomistic tight-binding calculations for the investigation of transport in extremely scaled SOI transistors. *IEEE International Electron Devices Meeting, 2003. IEDM '03 Technical Digest*, pages 9.2.1–9.2.4, Dec. 2003.

[70] F. Sacconi, A. Di Carlo, P. Lugli, M. Stadele, and J.-M. Jancu. Full band approach to tunneling in MOS structures. *IEEE Transactions on Electron Devices*, 51(5):741–748, May 2004.

[71] Fabio Sacconi, Michael Povolotskyi, Aldo Di Carlo, Paolo Lugli, and Martin Städele. Full-band approaches to the electronic properties of nanometer-scale MOS structures. *Solid-State Electronics*, 48(4):575–580, April 2004.

[72] A. Rahman, M. S. Lundstrom, and A. W. Ghosh. Generalized effective-mass approach for n-type metal-oxide-semiconductor field-effect transistors on arbitrarily oriented wafers. *Journal of Applied Physics*, 97(5):xx–xx, 2005.

[73] A. Rahman, A. Ghosh, and M. Lundstrom. Assessment of Ge n-MOSFETs by quantum simulation. *IEDM Technical Digest.*, pages 19.4.1–19.4.4, Dec. 2003.

[74] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom. Theory of ballistic nanotransistors. *IEEE Transactions on Electron Devices*, 50(9):1853–1864, 2003.

[75] A. Rahman, G. Klimeck, T. B. Boykin, and M. Lundstrom. Bandstructure effects in ballistic nanoscale MOSFETs. *IEDM Technical Digest.*, pages 139–142, Dec. 2004.

[76] Anisur Rahman, Gerhard Klimeck, and Mark Lundstrom. Novel channel materials for ballistic nanoscale MOSFETs-bandstructure effects. *To appear in Technical Digest of IEDM 2005*, 2005.

[77] Frank Stern and W. E. Howard. Properties of semiconductor surface inversion layers in electric quantum limit. *Physical Review*, 163(1):816–835, 1967.

[78] Frank Stern. Self-consistent results for n-type Si inversion layers. *Physical Review B*, 5(12):4891–4899, 1972.

[79] M. Lundstrom and Z. B. Ren. Essential physics of carrier transport in nanoscale MOSFETs. *IEEE Transactions on Electron Devices*, 49(1):133–141, 2002.

[80] S. Datta. *Electronic Transport in Mesoscopic Systems*. Cambridge Univ. Press, Cambridge, U.K., 1997.

[81] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal. Two-dimensional quantum mechanical modeling of nanotransistors. *Journal of Applied Physics*, 91(4):2343–2354, 2002.

[82] J. Knoch, B. Lengeler, and J. Appenzeller. Quantum simulations of an ultrashort channel single-gated n-MOSFET on SOI. *IEEE Transactions on Electron Devices*, 49(7):1212–1218, 2002.

[83] J. Guo and M. S. Lundstrom. A computational study of thin-body, double-gate, Schottky barrier MOSFETs. *IEEE Transactions on Electron Devices*, 49(11):1897–1902, 2002.

[84] A. Svizhenko and M. P. Anantram. Role of scattering in nanotransistors. *IEEE Transactions on Electron Devices*, 50(6):1459–1466, 2003.

[85] P. Damle, A. Ghosh, and S. Datta. *Molecular Nanoelectronics*. American Scientific, Stevenson Ranch, CA, 2003. pp. 115-135.

[86] J. Wang and M. Lundstrom. Ballistic transport in high electron mobility transistors. *IEEE Transactions on Electron Devices*, 50(7):1604–1609, 2003.

[87] Available at www.nanohub.purdue.edu.

[88] T. Ando, A. B. Fowler, and F. Stern. Electronic-properties of two-dimensional systems. *Reviews of Modern Physics*, 54(2):437–672, 1982.

[89] G. Dresselhaus, A. F. Kip, and C. Kittel. Cyclotron resonance of electrons and holes in silicon and germanium crystals. *Physical Review*, 98(2):368–384, 1955.

[90] R. N. Dexter, B. Lax, A. F. Kip, and G. Dresselhaus. Effective masses of electrons in silicon. *Physical Review*, 96(1):222–223, 1954.

[91] Robert F. Pierret. *Advanced Semiconductor Fundamentals, 2nd ed.* Prentice Hall, 2002.

[92] S. M. Sze. *Physics of Semiconductor Devices, 2nd ed.* John Willey & Sons, 1981.

[93] J. M. Luttinger and W. Kohn. Motion of electrons and holes in perturbed periodic fields. *Physical Review*, 97(4):869–883, 1955.

[94] R. N. Dexter, H. J. Zeiger, and B. Lax. Cyclotron resonance experiments in silicon and germanium. *Physical Review*, 104(3):637–644, 1956.

[95] J. C. Hensel, H. Hasegawa, and M. Nakayama. Cyclotron resonance in uniaxially stressed silicon. 2. nature of covalent bond. *Physical Review*, 138(1A):A225–xxx, 1965.

[96] M. V. Fischetti and S. E. Laux. Band structure, deformation potentals, and carrier mobility in strained Si, Ge, and SiGe alloys. *Journal of Applied Physics*, 80(4):2234–2252, 1996.

[97] M. M. Rieger and P. Vogl. Electronic-band parameters in strained $si_{1-X}ge_x$ alloys on $si_{1-Y}ge_y$ substrates. *Physical Review B*, 48(19):14276–14287, 1993.

[98] P. M. Solomon and S. E. Laux. The ballistic FET: Design, capacitance and speed limit. *IEDM Tech. Digest*, pages 95–98, Dec. 2001.

[99] G. Timp *et al.* The ballistic nano-transistor. *IEDM Tech. Digest*, pages 55–58, Dec. 1999.

[100] C. W. Leitz, M. T. Currie, M. L. Lee, Z.-Y. Cheng, D. A. Antoniadis, and E. A.Fitzgerald. Hole mobility enhancements in strained $Si/Si_{1-y}Ge_y$ p-type metal-oxide-semiconductor field-effect transistors grown on relaxed $Si_{1-x}Ge_x(x < y)$ virtual substrates. *Applied Physics Letters*, 79(25):4246–4248, Dec. 2001.

[101] Z. Y. Cheng, M. T. Currie, C. W. Leitz, G. Taraschi, E. A. Fitzgerald, J. L.Hoyt, and D. A.Antoniadas. Electron mobility enhancement in strained-Si n-MOSFETs fabricated on SiGe-on-insulator (SGOI) substrates. *IEEE Electron Device Letters*, 22(7):321–323, July 2001.

[102] K. Rim, J. L. Hoyt, and J. F. Gibbons. Fabrication and analysis of deep submicron strained-Si n-MOSFETs. *IEEE Trans. Electron Devices*, 47(7):1406–1415, July 2000.

[103] K. Rim, J. L. Hoyt, and J. F.Gibbons. Transconductance enhancement in deep submicron strained Si n-MOSFETs. *IEDM Tech. Dig.*, pages 707–710, 1998.

[104] Y.-C. Yeo, V. Subramanian, J. Kedzierski, P. Xuan, T.-J. King, J. Bokor, , and H. Chenming. Design and fabrication of 50-nm thin-body p-MOSFETs with a SiGe heterostructure channel. *IEEE Transactions on Electron Devices*, 49(2):279–286, Feb. 2002.

[105] K. Natori. Ballistic metal-oxide-semiconductor field effect transistor. *J. Appl. Phys.*, 76(xx):4879–4890, 1994.

[106] K. Natori. Scaling limit of the MOS transistora ballistic MOSFET. *IEICE Trans. Electron.*, E84-C(xxx):1029–1036, 2001.

[107] S. Datta, F. Assad, and M. S. Lundstrom. The Si MOSFET from a transmission viewpoint. *Superlatt. Microstruct.*, 23(xx):771–780, 1998.

[108] Y. Naveh and K. K. Likharev. Modeling of 10-nm-scale ballistic MOSFETs. *IEEE Electron Device Lett.*, 21(5):242–244, May 2000.

[109] Z. Ren. Nanoscale MOSFETs: Physics, simulation, and design. Ph.D. dissertation, Dec. 2001.

[110] M. S. Lundstrom. Elementary scattering theory of the MOSFET. *IEEE Electron Device Lett.*, 18(8):361–363, Aug. 1997.

[111] Jung-Hoon Rhew, Zhibin Ren, and Mark S. Lundstrom. A numerical study of ballistic transport in a nanoscale MOSFET. *Solid-State Electronics*, 46(11):1899–1906, 2002.

[112] R. Venugopal, Zhibin Ren, and M. S. Lundstrom. Simulating quantum transport in nanoscale MOSFETs: ballistic hole transport, subband engineering and boundary conditions. *IEEE Transactions on Nanotechnology*, 2(3):135–143, Sept. 2003.

[113] S. Luryi. Quantum capacitance devices. *Appl. Phys. Lett.*, 52(xx):501–503, Feb. 1988.

[114] A. Lochtefeld and D. A. Antoniadis. "on experimental determination of carrier velocity in deeply scaled NMOS: How close to the thermal limit?". *IEEE Electron Device Lett.,*, 22(2):95–97,, Feb. 2001.

[115] Jing Wang. Xxx. Ph.D. dissertation, Aug. 2005.

[116] Per-Olov Löwdin. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *The Journal of Chemical Physics*, 18(3):365–375, March 1950.

[117] D. J. Chadi and M. L. Cohen. Tight-binding calculations of the valence bands of diamond and zincblende crystals. *Physica Status Solidi B*, 68(1):405–419, 1975.

[118] Steven L. Richardson, Marvin L. Cohen, Steven G. Louie, and James R. Chelikowsky. Electron charge densities at conduction-band edges of semiconductors. *Phys. Rev. B*, 33(2):1177–1182, 1986.

[119] Gerhard Klimeck, R. Chris Bowen, Timothy B. Boykin, Carlos Salazar-Lazaro, Thomas A. Cwik, and Adrian Stoica. Si tight-binding parameters from genetic algorithm fitting. *Superlattices and Microstructures*, 27(2/3):77–88, Feb. 2000.

[120] Timothy B. Boykin. Generalized eigenproblem method for surface and interface states: The complex bands of gaas and alas. *Phys. Rev. B*, 54(11):8107–8115, 1996.

[121] Timothy B. Boykin, Gerhard Klimeck, R. Chris Bowen, and Fabiano Oyafuso. Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory. *Phys. Rev. B*, 66:125207, 2002.

[122] G. Dresselhaus. Spin-orbit coupling effects in zinc blende structures. *Phys. Rev.*, 100(2):580–586, 1955.

[123] Walter A. Harrison. *Elementary Electronic Structure*. World Scientific Publishing Company, revised edition edition.

[124] J. A. Støvneng and P. Lipavský. Multiband tight-binding approach to tunneling in semiconductor heterostructures: Application to γ X transfer in GaAs. *Phys. Rev. B*, 49(23):16494–16504, 1994.

[125] Timothy B. Boykin, Gerhard Klimeck, and Fabiano Oyafuso. Valence band effective-mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization. *Phys. Rev. B*, 69:115201, 2004.

[126] Seungwon Lee, Fabiano Oyafuso, Paul von Allmen, and Gerhard Klimeck. Boundary conditions for the electronic structure of finite-extent embedded semiconductor nanostructures. *Phys. Rev. B*, 69:045316, 2004.

[127] T. Low *et al.* Investigation of performance limits of germanium double-gated MOSFETs. *IEDM 2003 Technical Digest*, pages 29.4.1–29.4.4, 8-10 Dec. 2003.

[128] I. Aberg, O.O. Olubuyide, J. Li, R. Hull, and J. L. Hoyt. Fabrication of strained si/strained sige/strained si heterostructures on insulator by a bond and etch-back technique. *Proceedings of IEEE International SOI Conference, 2004.*, pages 35–36, Oct. 2004.

[129] I. Aberg, C. Ni Chleirigh, O. O. Olubuyide, X. Duan, and J. L. Hoyt. High electron and hole mobility enhancements in thin-body strained Si/strained SiGe/strained Si heterostructures on insulator. *IEDM Technical Digest.*, pages 173–176, Dec. 2004.

[130] K. Rim, J. Welser, J. L. Hoyt, and J. F. Gibbons. Enhanced hole mobilities in surface-channel strained-Si p-MOSFETs. *International Electron Devices Meeting*, pages 517–520, Dec. 1995.

[131] K. Rim *et al.* Characteristics and device design of sub-100 nm strained Si n- and pMOSFETs. *Digest of Technical Papers:2002 Symposium on VLSI Technology*, pages 98–99, June 2002.

[132] Shin ichi Takagi, Judy L. Hoyt, Jeffrey J. Welser, and James F. Gibbons. Comparative study of phonon-limited mobility of two-dimensional electrons in strained and unstrained Si metaloxidesemiconductor field-effect transistors. *Journal of Applied Physics*, 80(3):1567–1577, Aug. 1996.

[133] M. V. Fischetti, F. Gámiz, and W. Hänsch. On the enhanced electron mobility in strained-silicon inversion layers. *Journal of Applied Physics*, 92(12):7320–7324, Dec. 2002.

[134] N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. Saunders, Philadelphia, PA, 1976.

[135] Shang Yuan Ren, Xia Chen, and John D. Dow. Tight-binding sp^3d^5 hamiltonian for Si. *Journal of Physics and Chemistry of Solids*, 59(3):403–410, March 1998.

[136] Walter A. Harrison. *Electronic Structure and the Properties of Solids : The Physics of the Chemical Bond*. Dover Publications, dover ed edition edition.

[137] Walter Ashley Harrison. *Solid State Theory*.

APPENDIX

APPENDIX A
SOLVING GENERALIZED EFFECTIVE MASS EQUATION

A.1 Separable Potential

Here the steps to solve (2.14) in page 21:

$$\left[-\frac{\hbar^2}{2m_{11}}\frac{\partial^2}{\partial x^2} - i\frac{\hbar^2 k_y}{m_{12}}\frac{\partial}{\partial x} + \frac{\hbar^2 k_y^2}{2m_{22}} + \left\{-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z} + W(x,z)\right\}\right] \times \Psi_{k_y}(x,z) = E\Psi_{k_y}(x,z),$$

for the separable potential given in (2.35)

$$W(x,z) = U(z) + V(x),$$

in page 25 are presented.

A.1.1 Quantum Confinement Problem

In the special case of separable potential, the part of (2.14) that deals with quantum confinement is

$$\left[H_z + U(z)\right]\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : z\right) = \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right)\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : z\right) \tag{A.1}$$

where H_z is given in (2.16)

$$H_z = -\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z}.$$

Substituting the canonical transformation of (2.17) (page 21)

$$\zeta\left(-i\frac{\partial}{\partial x}, k_y : z\right) = U_C\phi_i(z),$$

where,

$$U_C = e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z},$$

in (A.1) and left multiplying by $U_C^\dagger$ we find

$$\left[\bar{H}_z + U(z)\right]\phi_i(z) = \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right)\phi_i(z), \tag{A.2}$$

where,

$$\begin{aligned}\bar{H}_z &= U_C^\dagger H_z U_C \\ &= e^{i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\left\{-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z}\right\}e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}.\end{aligned} \tag{A.3}$$

We now use the identity given in (2.18) (page 22),

$$e^{-B}Ae^{B} = A + [A,B] + \frac{1}{2}[[A,B],B] + \cdots$$

where,

$$A = \left\{-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z}\right\}$$

$$B = -i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z.$$

The second term in the RHS of identity (2.18) can be evaluated as,

$$\begin{aligned}[A,B] &= \left[-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z}, -i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z\right] \\ &= \frac{\hbar^2}{2m_{33}}i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)\left[\frac{\partial^2}{\partial z^2}, z\right] + i^2\hbar^2 m_{33}\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)^2\left[\frac{\partial}{\partial z}, z\right] \\ &= i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z} - \frac{\hbar^2}{m_{33}}\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)^2.\end{aligned} \tag{A.4}$$

Similarly, the third term in the RHS of the identity can be evaluated as,

$$\begin{aligned}[[A,B],B] = \Bigg[&i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z} - \frac{\hbar^2}{m_{33}}\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial z}\right)^2 \\ &, -i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z\Bigg]\end{aligned}$$

$$
\begin{aligned}
&= -i^2\hbar^2 m_{33}\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)^2\left[\frac{\partial}{\partial x}, z\right] \\
&= \hbar^2 m_{33}\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)^2. \qquad \text{(A.5)}
\end{aligned}
$$

To evaluate the commutators the following relationships are useful,

$$
\left[\frac{\partial}{\partial z}, z\right] = 1; \quad \left[\frac{\partial^2}{\partial z^2}, z\right] = 2\frac{\partial}{\partial z}; \quad \left[\frac{\partial}{\partial z}, c\right] = 0. \qquad \text{(A.6)}
$$

Since the commutator $[[A, B], B]$ is indepent of z, all the higher orde terms in (2.18) is zero, and we have

$$
\begin{aligned}
\bar{H}_z &= e^{-B} A e^{B} \\
&= A + [A, B] + \frac{1}{2}[[A, B], B] \\
&= -\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right) \qquad \text{(A.7)}
\end{aligned}
$$

where

$$
\begin{aligned}
\epsilon\left(-i\frac{\partial}{\partial x}, k_y\right) &= \frac{\hbar^2}{2m_{33}}\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)^2 \\
&= \frac{\hbar^2}{2}\left(-\frac{m_{33}}{m_{31}^2}\frac{\partial^2}{\partial x^2} + \frac{m_{33}}{m_{32}^2}k_y^2 - 2i\frac{m_{33}}{m_{31}m_{23}}k_y\frac{\partial}{\partial x}\right). \qquad \text{(A.8)}
\end{aligned}
$$

Since, $\bar{H}_z$ is sum of z dependent and z independent parts, the eigenvalue of (A.2) can also be separated as

$$
\epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right) = \epsilon_i - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right), \qquad \text{(A.9)}
$$

and, we can rewrite (A.2) as

$$
\left[\bar{H}_z + U(z)\right]\phi_i(z) = \left[\epsilon_i - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right)\right]\phi_i(z). \qquad \text{(A.10)}
$$

here, the eigenvalues ϵ_i and the eigenfunctions $\phi_i(z)$ are found by solving the Z parts of the potential only, i.e.,

$$
\left[-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} + U(z)\right]\phi_i(z) = \epsilon_i\phi_i(z). \qquad \text{(A.11)}
$$

Finally, using

$$\bar{H}_z = e^{i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z} H_z e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z},$$

in (A.10) we find

$$\left[H_z + U(z)\right] e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(z) = \left\{\epsilon_i - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right)\right\} e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(z) \tag{A.12}$$

A.1.2 The Transport Problem

We now return to the original 2D effective mass equation (2.22) in page 23, which is

$$\left[-\frac{\hbar^2}{2m_{11}}\frac{\partial^2}{\partial x^2} - i\frac{\hbar^2 k_y}{m_{12}}\frac{\partial}{\partial x} + \frac{\hbar^2 k_y^2}{2m_{22}} + \overline{H_z + U(z)} + V(x)\right]\Psi\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) = E\Psi\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) \tag{A.13}$$

Since the eigenfunctions of (A.12), $e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(z)$, forms a complete set, at a given x, we expand the wavefunction in (A.13) as

$$\Psi\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) = \sum_m e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_m(x, z)\,\chi_m(x, k_y)\,. \tag{A.14}$$

Substituting (A.14) in (A.13), then left multiplying by $\phi_n^*(z)e^{i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}$ and integrating along Z, we find

$$\left[-\frac{\hbar^2}{2m_{11}}\frac{\partial^2}{\partial x^2} - i\frac{\hbar^2 k_y}{m_{12}}\frac{\partial}{\partial x} + \frac{\hbar^2 k_y^2}{2m_{22}} + \left(\epsilon_n - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right)\right)\right]\chi_n(x, ky) + \sum_m V_{nm}(x)\chi_m(x, k_y) = E\chi_m(x, k_y). \tag{A.15}$$

where we have made use of (A.12) and the orthogonality condition for $\phi_i(z)$

$$\int \phi_n^*(z)\phi_m(z)dz = \delta_{n,m}.$$

Additionally, in (A.15), we have defined the matrix elements for $V(x)$ as

$$V_{nm}(x) = \int \left[e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z} V(x) e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\right]\phi_n^*(z)\phi_m(z)dz \tag{A.16}$$

which can be evaluated as

$$\begin{aligned} V_{nm}(x) &= \int \left[e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z} V(x) e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z} \right] \phi_n^*(z)\phi_m(z)dz \\ &= \int \left\{ V(x) + \left[V(x), -\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z \right] + \cdots \right\} \phi_n^*(z)\phi_m(z)dz \\ &= \int \left\{ V(x) + \left(\frac{m_{33}}{m_{31}}z\right)\frac{\partial V(x)}{\partial x} + \frac{1}{2}\left(\frac{m_{33}}{m_{31}}z\right)^2 \frac{\partial^2 V(x)}{\partial x^2} + \cdots \right\} \phi_n^*(z)\phi_m(z)dz \\ &= \int V\left(x + \frac{m_{33}}{m_{31}}z\right) \phi_n^*(z)\phi_m(z)dz \end{aligned} \tag{A.17}$$

From (A.17) it can be seen that at a given x, say $x = x_c$, in order to evaluate $V_{nm}(x = x_c)$, inside the integration the potential should be samped along the line:

$$x = x_c + \frac{m_{33}}{m_{31}}z,$$

which is also along the principal axis of the constant energy CB ellipsoid at fixed k_y.

Although in (A.17), $V_{nm}(x)$ is writen in a compact form, its physical significance can be understood from the Taylor series expansion

$$\begin{aligned} V_{nm}(x) &= \int \left\{ V(x) + \left(\frac{m_{33}}{m_{31}}z\right)\frac{\partial V(x)}{\partial x} + \frac{1}{2}\left(\frac{m_{33}}{m_{31}}z\right)^2 \frac{\partial^2 V(x)}{\partial x^2} + \cdots \right\} \phi_n^*(z)\phi_m(z)dz \\ &= V(x)\delta_{nm} + \frac{m_{33}}{m_{31}}\left[\frac{\partial}{\partial x}, V(x)\right] \int z\phi_n^*(z)\phi_m(z)dz + \cdots \\ &= V(x)\delta_{nm} + \frac{m_{33}}{m_{31}}\frac{\partial V(x)}{\partial x}\mu_{nm} + \cdots \end{aligned} \tag{A.18}$$

The first term in (A.18) $V_{nn}(x) = V(x)$ respresents the on-sie potential and the second term, which depends on the electric field, represents Zener tunneling between subbands. this term and all higher order terms represent coupling between differet subbands.

A.2 General Potential

In this section we will show the steps to solve (2.14) in page 21 without assuming the potential to be separable.

A.2.1 Quantum Confinement Problem

From (2.14), the qunatum confinement problem at a given x is

$$\left[H_z + U\left(x,z\right)\right]\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) = \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y : x\right)\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : x, z\right), \tag{A.19}$$

where the Hamitonian is

$$H_z = -\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} - i\hbar^2\left(\frac{k_y}{m_{23}} - i\frac{1}{m_{31}}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial z}. \tag{A.20}$$

We now write

$$\zeta_i\left(-i\frac{\partial}{\partial x}, k_y : x, z\right) = e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(x,z). \tag{A.21}$$

Substituting (IV) in (III) and left multiplying by $e^{i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}$ we find

$$\left[\bar{H}_z + \left\{e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}U(x,z)e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\right\}\right]\phi_i(x,z) = e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\epsilon_i\left(-i\frac{\partial}{\partial x}, k_y : x\right)e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\phi_i(x,z). \tag{A.22}$$

In Sec. A.1.1 it has already been shown that

$$\bar{H}_z = -\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} + \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right) \tag{A.23}$$

$$\epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right) = -\frac{\hbar^2}{2m_{33}}\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)^2 \tag{A.24}$$

Using the identity in (2.18), the potential term in A.22 can be reduced as follows,

$$\begin{aligned}
e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}U(x,z)e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z} &= U(x,z) + \left[U(x,z), -\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z\right] + \cdots \\
&= U(x,z) - \frac{m_{33}}{m_{31}}\left[U(x,z), \frac{\partial}{\partial x}\right]z + \cdots \\
&= U(x,z) + \frac{m_{33}}{m_{31}}z\frac{\partial U(x,z)}{\partial x} + \left(\frac{m_{33}}{m_{31}}z\right)^2\frac{\partial^2 U(x,z)}{\partial x^2} + \cdots \\
&= U\left(x + \frac{m_{33}}{m_{31}}z, z\right).
\end{aligned} \tag{A.25}$$

For a given x, say $x = x_c$, (A.25) shows that the relevant potential in (A.22) must be taken along the line

$$x = x_c + \frac{m_{33}}{m_{31}}z,$$

which is also along the principal axis of the constant energy ellipse at $k_y = 0$. Substituting $\bar{H}_z$ and the potential term in (A.22) and after minor manipulation, we find

$$\left[-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} + U\left(x + \frac{m_{33}}{m_{31}}z, z\right)\right]\phi_i(x,z) = e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\left\{\epsilon_i\left(-i\frac{\partial}{\partial x}, k_y : x\right) - \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right)\right\}e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\phi_i(x,z). \quad \text{(A.26)}$$

Clearly, for a fixed value of x, the eigenvalues of (A.26) are $\epsilon_i\left(x + \frac{m_{33}}{m_{31}}z\right)$, and therefore, (A.26) can be written as

$$\left[-\frac{\hbar^2}{2m_{33}}\frac{\partial^2}{\partial z^2} + U\left(x + \frac{m_{33}}{m_{31}}z, z\right)\right]\phi_i(x,z) = \epsilon_i\left(x + \frac{m_{33}}{m_{31}}z\right)\phi_i(x,z), \quad \text{(A.27)}$$

which must be solved at each x to obtain modes $\phi_i(x,z)$. Additionally, since

$$\epsilon_i\left(x + \frac{m_{33}}{m_{31}}z\right) = e^{\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}\epsilon_i(x)e^{-\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}z}, \quad \text{(A.28)}$$

from (A.26) and (A.27) we can write

$$\epsilon_i\left(-i\frac{\partial}{\partial x}, k_y : x\right) = \epsilon_i(x) + \epsilon_i\left(-i\frac{\partial}{\partial x}, k_y\right) \quad \text{(A.29)}$$

This very important relationship shows that the total energy is the sum of the confinement energy and the in-plane kinetic energy and in (2.21) this is expressed as

$$\epsilon\left(-i\frac{\partial}{\partial x}, k_y\right) = \frac{\hbar^2}{2}\left(-\frac{m_{33}}{m_{31}^2}\frac{\partial^2}{\partial x^2} + \frac{m_{33}}{m_{32}^2}k_y^2 - 2i\frac{m_{33}}{m_{31}m_{23}}k_y\frac{\partial}{\partial x}\right).$$

Finally, using the inverse canonical transformation

$$\bar{H}_z = e^{i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}H_z e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}.$$

in (A.22) we find (2.20) in page 22, which is

$$\left[H_z + W(x,z)\right]e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(x,z) = \left\{\epsilon_i(x) - \epsilon\left(-i\frac{\partial}{\partial x}, k_y\right)\right\}e^{-i\left(\frac{m_{33}}{m_{23}}k_y - i\frac{m_{33}}{m_{31}}\frac{\partial}{\partial x}\right)z}\phi_i(x,z), \quad \text{(A.30)}$$

APPENDIX B
THE TOP OF THE BARRIER BALLISTIC MODEL

B.1 Treating Floating Boundary Condition

Figure 4.6 on page 56 shows how the states at the top of the barrier are occupied for a simple $E\left(\vec{k}\right)$ relationship. As mentioned in Sec. 4.3, the energy reference is the top of the barrier at zero terminal bias. We express the source Fermi level, E_{F1}, drain Fermi level, E_{F2}, and potential at the top of the barrier for first subband, U_{scf}, with respect to this reference. The positive k-states are then occupied according to the source Fermi level according to,

$$\begin{aligned} N_1 &= \frac{1}{A}\sum_{k_x>0,k_y} f\left(E-E_{F1}\right) \\ &= \int_{k_x>0}\int_{k_y} 2\frac{d^2k}{(2\pi)^2} f\left(E-E_{F1}\right) \\ &= \int_{-\infty}^{+\infty} dE f\left(E-E_{F1}\right)\int_{S(E)} \frac{dS}{2\pi^2}\frac{1}{\left|\vec{\nabla}E(k)\right|}, \end{aligned}$$

where $S(E)$ is a constant energy surface in k-space, dS is an elemental area on this surface, and $dE/\left|\vec{\nabla}E(k)\right|$ is the distance between the surfaces $S(E+dE)$ and $S(E)$ [134]. Defining the density-of-states as

$$D\left(E-U_{scf}\right) = \int_{S\left(E-U_{scf}\right)}]\frac{dS}{2\pi^2}\frac{1}{\left|\vec{\nabla}E(k)\right|},$$

we finally have

$$N_1 = \frac{1}{2}\int_{-\infty}^{+\infty} D\left(E-U_{scf}\right) f\left(E-E_{F1}\right) dE. \tag{B.1}$$

The last expression is valid for general bandstructure in 1, 2 or 3D. The density-of-states function is either analytically expressed or is numerically tabulated. For a 2D electron gas with isotropic and parabolic $E\left(\vec{k}\right)$ relationship, we have,

$$D\left(E-U_{scf}\right) = g_{2D}\theta\left(E-U_{scf}\right),$$

where $g_{2D} = 2m^*/\pi\hbar^2$ is the 2D density-of-states, when spin degeneracy and a valley degeneracy of two for the unprimed subband in silicon are considered. In this case the integral for N_1 can be analytically evaluated as

$$N_1 = \frac{N_{2D}}{2} \log\left(1 + e^{(E_{F1} - U_{scf})/k_B T}\right) = \frac{N_{2D}}{2} \Im_0 \left(\eta_{F1}\right), \tag{B.2}$$

where $N_{2D} = k_B T g_{2D}$ is the effective 2D density-of-states, $\Im_0$ is the Fermi Dirac integral of order 0, and $\eta_{F1} = \left(E_{F1} - U_{scf}\right)/k_B T$. A similar expression exists for N_2 with η_{F1} replaced by $\eta_{F2} = \eta_{F1} - qV_{DS}/k_B T$.

In addition to the carrier density, we can also evaluate current for the positive k population from

$$\begin{aligned} I_1 &= \frac{q}{A} \sum_{k_x > 0, k_y} \upsilon_x f\left(E - E_{F1}\right) \\ &= \int_{k_x > 0} \int_{k_y} \frac{d^2 k}{2\pi^2} q \upsilon_x f\left(E - E_{F1}\right) \\ &= \int_{-\infty}^{+\infty} dE f\left(E - E_{F1}\right) \frac{q}{2} \int_{S(E)} \frac{dS}{2\pi^2} \left|\upsilon_x\right| \frac{1}{\left|\vec{\nabla} E(k)\right|} \\ &= \int_{-\infty}^{+\infty} dE f\left(E - E_{F1}\right) \frac{q}{2} \bar{\upsilon_x}\left(E - U_{scf}\right) D\left(E - U_{scf}\right), \end{aligned} \tag{B.3}$$

where $\bar{\upsilon_x}(E)$ is the average value of $|\upsilon_x|$ over the constant energy surface, $S(E)$, expressed as

$$\bar{\upsilon_x}\left(E - U_{scf}\right) = \frac{\int_{S\left(E - U_{scf}\right)} \frac{dS}{2\pi^2} \left|\upsilon_x\right| \frac{1}{\left|\vec{\nabla} E(k)\right|}}{\int_{S\left(E - U_{scf}\right)} \frac{dS}{2\pi^2} \frac{1}{\left|\vec{\nabla} E(k)\right|}}.$$

Now defining the current-density-of-states as,

$$J\left(E - U_{scf}\right) = \frac{q}{2} \bar{\upsilon_x}\left(E - U_{scf}\right) D\left(E - U_{scf}\right), \tag{B.4}$$

we have

$$I_1 = \int_{-\infty}^{+\infty} J\left(E - U_{scf}\right) f\left(E - E_{F1}\right) dE. \tag{B.5}$$

In general, this expression can be evaluated for either numerically tabulated or analytically calculated bandstructures. For the 2D electron density considered here, we can analytically evaluate $\bar{v_x}(E)$ to obtain

$$J\left(E-U_{scf}\right)=\frac{1}{2}q\left(\frac{2}{\pi}\sqrt{\frac{2\left(E-U_{scf}\right)}{m^*}}\right)D\left(E-U_{scf}\right), \tag{B.6}$$

where the factor $2/\pi$ appears because of averaging v_x over all possible k_y values at energy $E-U_{scf}$. With this expression for $J(E-U_{scf})$, we can analytically integrate (B.5) to find

$$I_1=\frac{1}{2}qN_{2D}\sqrt{\frac{2k_BT}{\pi m^*}}\Im_{1/2}\left(\eta_{F1}\right). \tag{B.7}$$

Similar expression can be obtained for negative going carriers, with η_{F1} replaced by η_{F2}.

When the drain bias, V_{DS}, is large, only the $+k_x$ states are occupied, and we can evaluate the maximum velocity at the top of the barrier as

$$\left\langle v(0)\right\rangle_{max}=v_{inj}\equiv\frac{I_1}{N_1}=\sqrt{\frac{2k_BT}{\pi m^*}}\frac{\Im_{1/2}\left(\eta_{F1}\right)}{\Im_0\left(\eta_{F1}\right)}. \tag{B.8}$$

The presence of the Fermi-Dirac integrals in this expression explains why the saturation injection velocities in Figs. 4.5b (page 55) and 4.10b (page 65) are gate bias dependent. Below threshold voltage, the ratio of the Fermi-Dirac integrals is one and the injection velocity becomes constant. The injection velocity at the highest gate bias determines the maximum on-current that a transistor can deliver.

Finally, we will discuss the treatment of the floating boundary condition in the analytical model. In Fig. B.1a, we see that at low gate and high drain bias the barrier height is large, i.e. $U_{scf} >> E_{CS}$, and inside the source, both positive and negative going states are at equilibrium with the source Fermi level. The charge neutrality condition demands

$$N_{SD}=\int_{-\infty}^{+\infty}D\left(E-E_{CS}\right)f\left(E-E_{F1}\right)dE. \tag{B.9}$$

where N_{SD} is the doping density in the source extension.

When high gate bias is applied, we can see in Fig. B.1b that the barrier height becomes small, and there are three distinct groups of carriers: i) carriers with energy lower than the barrier height and are reflected by the barrier ii) carriers with energy higher than the barrier and having positive velocity and iii) carriers having energy above barrier and going in the negative direction. Population groups i) and ii) are at equilibrium with E_{F1} and group iii) is in equilibrium with E_{F2}. Because the sum of the three populations in Fig B.1b is smaller than equilibrium carrier density in source, to maintain charge neutrality we have to increase $(E_{F1} - E_{CS})$. Physically, E_{F1} is fixed and E_{CS} floats down. Equivalently, as shown in Fig. B.1c, we can keep E_{CS} fixed and float E_{F1} up. In our analytical model we have treated the floating boundary condition by fixing E_{CS} and floating up E_{F1} to E'_{F1}. Therefore, the charge neutrality condition in the source is

$$\begin{aligned} N_{SD} &= \int_{-\infty}^{U_{scf}} D\left(E - E_{CS}\right) f\left(E - E'_{F1}\right) dE \\ &+ \frac{1}{2}\int_{U_{scf}}^{\infty} D\left(E - E_{CS}\right) \left\{f\left(E - E'_{F1}\right) + f\left(E - E'_{F1} + qV_{DS}\right)\right\} dE. \end{aligned} \tag{B.10}$$

Equation (B.10) is solved self-consistently with eqs. (4.3–4.14), i.e., for each E'_{F1}, barrier height is computed to distinguish three carrier populations and charge neutrality in the source is ensured.

B.2 Treating Arbitrary Bandstructure

In an ultra thin body MOSFET, the energy along the channel thickness direction becomes quantized and sub-bands are formed. The $E\left(k_x, k_y\right)$ relationship for carriers in a sub-band is on a two dimensional (k_x, k_y)-space. To generalize the analytical model for arbitrary bandstructure, we first calculate the "group velocity" for the states on this 2D k-space using

$$\vec{v}\left(k_x, k_y\right) = \upsilon_x \hat{k}_x + \upsilon_y \hat{k}_y = \frac{1}{\hbar}\vec{\nabla}_{\vec{k}} E, \tag{B.11}$$

and group them in source injected population or drain injected population according to the sign of the velocity component along the transport direction. Assuming the carrier transport direction (the direction from the source to the drain) makes an angle θ with the k_x direction, we define an unit vector, $\hat{n}$, along this direction and write it as

$$\hat{n} = \cos\theta\hat{k}_x + \sin\theta\hat{k}_y, \tag{B.12}$$

where, $\hat{k}_x$ and $\hat{k}_y$ are the unit vectors.

For a spherical, parabolic $E\left(\vec{k}\right)$ relationship, group velocity, $\vec{v}$, is radial. However, when the bandstructure is warped, as for the heavy hole band, or is not isotropic (e.g. Germanium with [100] quantized), the velocity field in the 2-D k-space does not show radial symmetry.

For an arbitrary 2D $E\left(\vec{k}\right)$ relationship, the equilibrium carrier density, N_0 , can be numerically calculated from

$$N_0 = \sum \sum_{\{k_x,k_y\}} 2g_v \frac{\Delta k_x \Delta k_y}{(2\pi)^2} f\left(E_i\left(k_x,k_y\right) - E_F\right), \tag{B.13}$$

where the factor 2 is for spin degeneracy, g_v is the valley degeneracy, Δk_x, Δk_y are the grid spacing in k-space and E_F is the equilibrium Fermi level. The sum is performed over the entire k-space.

Now, when the drain bias is applied, two distinct carrier populations at the top of the barrier exist. The states in the k-space which have positive velocity component in the transport direction, $\hat{n}$, are at equilibrium with the source Fermi level, μ_1, and the corresponding carrier concentration can be calculated from

$$N_1 = \sum \sum_{\{k_x,k_y\}|\vec{v}\cdot\hat{n}>0} 2g_v \frac{\Delta k_x \Delta k_y}{(2\pi)^2} f\left(E_i\left(k_x,k_y\right) + U_{scf} - \mu_1\right), \tag{B.14}$$

This expression can be simplified, using the fact

$$\sum \sum_{\{k_x,k_y\}} = 0, \tag{B.15}$$

implying, for each state in k-space with a certain positive velocity component along $\hat{n}$, there is a state with exactly equal negative velocity. Therefore, we find

$$\sum \sum_{\{k_x,k_y\}|\vec{v}\cdot\hat{n}>0} = \frac{1}{2}\sum \sum_{\{k_x,k_y\}} . \tag{B.16}$$

This simplifies (B.14), which can now be expressed as

$$N_1 = \frac{1}{2}\sum \sum_{\{k_x,k_y\}} 2g_v \frac{\Delta k_x \Delta k_y}{(2\pi)^2} f\left(E_i\left(k_x,k_y\right)+U_{scf}-\mu_1\right). \tag{B.17}$$

where the sum is over the whole k-space. Similarly, the expression for the drain injected carrier population, N_2, can be obtained by replacing μ_1 by $\mu_2 = \mu_1 - qV_{DS}$.

One can now solve for the self-consistent potential, U_{scf} , from (4.10–4.14), using the numerically evaluated carrier concentration in (B.14) and (B.15). After calculating U_{scf}, the ballistic current can also be calculated. The ballistic current carried by the source injected population can be expressed as

$$I_1 = \sum \sum_{\{k_x,k_y\}|\vec{v}\cdot\hat{n}>0} 2g_v \frac{\Delta k_x \Delta k_y}{(2\pi)^2} \left(\vec{v}\cdot\hat{n}\right) f\left(E_i\left(k_x,k_y\right)+U_{scf}-\mu_1\right),$$

where the sum is only over the states having positive "group velocity" component in the transport direction. However, from (15), it follows

$$\sum \sum_{\{k_x,k_y\}|\vec{v}\cdot\hat{n}>0} \left(\vec{v}\cdot\hat{n}\right) = \frac{1}{2}\sum \sum_{\{k_x,k_y\}} \left|\vec{v}\cdot\hat{n}\right|, \tag{B.18}$$

and therefore, the expression for I_1 is simplified to

$$I_1 = \frac{1}{2}\sum \sum_{\{k_x,k_y\}} 2g_v \frac{\Delta k_x \Delta k_y}{(2\pi)^2} \left|\vec{v}\cdot\hat{n}\right| f\left(E_i\left(k_x,k_y\right)+U_{scf}-\mu_1\right), \tag{B.19}$$

where the sum is now over the entire k-space. Similar expression for the current carried by the drain injected population, I_2, can be obtained by replacing μ_1 by $\mu_1 - qV_{DS}$. Finally, the net current is

$$I = I_1 - I_1. \tag{B.20}$$

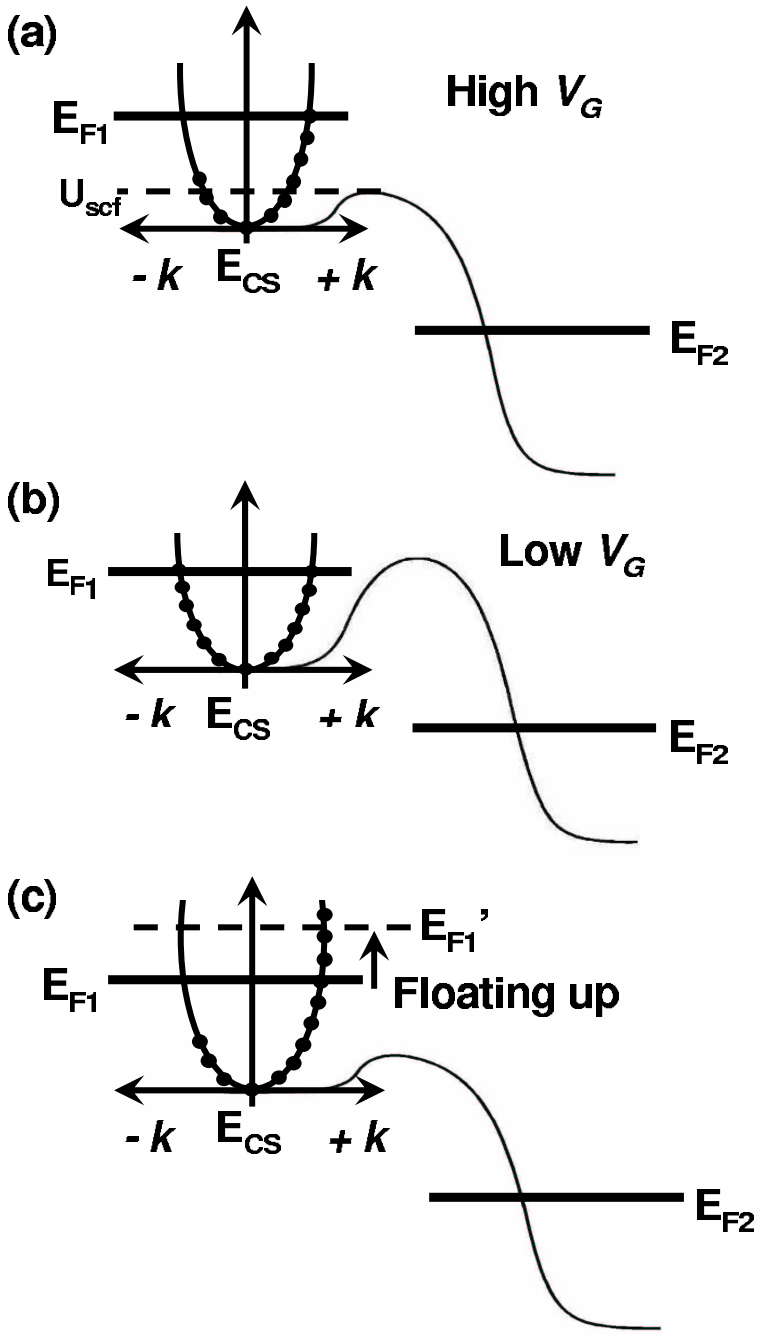

Fig. B.1. Treating floating boundary condition. (a) Under low V_G charge neutrality in source extension is maintained by only E_{F1}, (b) when V_G is increased barrier lowers and charge neutrality is not maintainted, (c) raising E_{F1} to E'_{F1} restores charge neutrality in source entension.

APPENDIX C
THE TIGHT-BINDING HAMILTONIAN WITH STRAIN

C.1 Derivation of Tight-binding Hamiltonian for Bulk Materials

To derive the tight-binding Hamiltonian, the following conditions are assumed:

1. *atom-like orbitals*–localized basis functions have atomic orbitals symmetry,
2. *tightly bound*–overlap of two orbitals on different atomic sites is zero,
3. *orthogonality*–overlap of two different orbials located on same atomic site is zero,
4. *nearest neighbor interaction*–nonzero matrix elements for Hamiltonian possible only between orbitals located on nearest neighboring sites. Nearest neighbors of a cation are four anions and vice versa, therefore, matrix elements between cations and anions are only possibility.
5. *two center integrals*–nonzero matrix element for Hamiltonian possible only when the potential is on one of the two atoms on which orbitals are located.

A set of such localized atomlike orbitals: $\left|nb\vec{R}_i\right\rangle$ is assigned at each atomic points. Here b is the atom type, cation or anion, $b \in \{a, c\}$, $\vec{R}_i$ is the lattice point coordinate with respect to some origin ($\vec{R}_{ji} = \vec{R}_i - \vec{R}_j$ relative position of i point with respect to j), and n is the orbital type, $n \in \{s, p_x, p_y, p_z, d_{xy}, d_{yz}, d_{zx}, d_{x^-y^2}, d_{3z^2-r^2}, s^*\}$. The atomic orbitals, $\left|nb\vec{R}_i\right\rangle$, are the eigenfunctions of the atomic Hamiltonian,

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + U_{i,b}\right]\left|nb\vec{R}_i\right\rangle = \epsilon_{n,b}\left|nb\vec{R}_i\right\rangle, \tag{C.1}$$

where, $U_{i,b}$ is the spherically symmetric nuclear potential of b type atom at i-th lattice point. The basis set for Hamiltonian consists of the Bloch sums of the localized orbitals, performed as

$$\left|nb\vec{k}\right\rangle = \frac{1}{\sqrt{N}} \sum_{\vec{R}_i} e^{i\vec{k}\bullet\left(\vec{R}_i+\vec{\nu}_b\right)} \left|nb\vec{R}_i\right\rangle . \tag{C.2}$$

There are two atoms, anion and cation, associated with each lattice point of FCC lattice, the position of atom b relative to the lattice point is defined as,

$$\vec{\nu}_b = \left[\frac{a_L}{4} \ \frac{a_L}{4} \ \frac{a_L}{4}\right] \delta_{b,c}.$$

The anion atoms occupy the lattice positions, therefore, $\vec{\nu}_a = 0$, and cations are one quarter of body diagonal displaced from the lattice points,

$$\vec{\nu}_c = \left[\frac{a_L}{4} \ \frac{a_L}{4} \ \frac{a_L}{4}\right].$$

We now assume the variational wavefunction as,

$$\begin{aligned} \left|\vec{k}\lambda\right\rangle &= \sum_{n,b} \left|nb\vec{k}\right\rangle \left\langle nb\vec{k}\right| \left|nb\vec{k}\right\rangle \\ &= \sum_{n,b} c_{n,b,\lambda} \left|nb\vec{k}\right\rangle , \end{aligned} \tag{C.3}$$

which is the linear combination of Bloch sums. In (C.3), λ is the band index and $c_{n,b,\lambda}$'s are the unknown expansion coefficients. The crystal Hamiltonian is the sum of kinetic and potential energy operators,

$$\begin{aligned} H &= -\frac{\hbar^2}{2m}\nabla^2 + \sum_{i,b} U_{i,b} \\ &= -\frac{\hbar^2}{2m}\nabla^2 + \sum_{i} \left[U_{i,a} + U_{i,c}\right] . \end{aligned} \tag{C.4}$$

The potential term is the sum of spherically symmetric nuclear potentials at all atomic locations.

$$\left[H - \epsilon\left(\vec{k}\lambda\right)\right] \left|\vec{k}\lambda\right\rangle = 0. \tag{C.5}$$

Substituting the variational wavefunction (C.3) in the Schrödinger equation (C.5), we have

$$\left[H - \epsilon\left(\vec{k}\lambda\right)\right] \sum_{n,b} c_{n,b,\lambda} \left|nb\vec{k}\right\rangle = \sum_{n,b} c_{n,b,\lambda} \left[H - \epsilon\left(\vec{k}\lambda\right)\right] \left|nb\vec{k}\right\rangle . \tag{C.6}$$

Now after left-multiplying by $\left\langle md\vec{k}\right|$, (C.6) can be expressed as,

$$\sum_{n,b} c_{n,b,\lambda} \left\langle md\vec{k}\right| \left[H - \epsilon\left(\vec{k}\lambda\right)\right] \left|nb\vec{k}\right\rangle = \sum_{n} c_{n,a,\lambda} \left[\left\langle md\vec{k}\right| H \left|na\vec{k}\right\rangle - \epsilon\left(\vec{k}\lambda\right) \left\langle md\vec{k}\right| na\vec{k}\right\rangle\right] + \sum_{n} c_{n,c,\lambda} \left[\left\langle md\vec{k}\right| H \left|nc\vec{k}\right\rangle - \epsilon\left(\vec{k}\lambda\right) \left\langle md\vec{k}\right| nc\vec{k}\right\rangle\right]. \tag{C.7}$$

Orthogonality of atomic orbitals becomes useful to evaluate terms in (C.7), such as

$$\begin{aligned} \left\langle md\vec{k}\right| na\vec{k}\right\rangle &= \frac{1}{N}\sum_{i,j} e^{i\vec{k}\cdot\left(\vec{R}_i+\vec{\nu}_a\right)-i\vec{k}\cdot\left(\vec{R}_j+\vec{\nu}_d\right)} \left\langle md\vec{R}_j\right| na\vec{R}_i\right\rangle \\ &= e^{-i\vec{k}\cdot\vec{\nu}_d}\sum_{i} e^{i\vec{k}\cdot\vec{R}_{ij}} \left\langle md\vec{R}_j\right| na\vec{R}_i\right\rangle \\ &= e^{-i\vec{k}\cdot\vec{\nu}_d}\sum_{i} e^{i\vec{k}\cdot\vec{R}_{ij}} \delta_{m,n}\delta_{d,a}\delta_{i,j} \\ &= \delta_{m,n}\delta_{d,a} \end{aligned} \tag{C.8}$$

where, δ_{ij} follows from the *tight-binding* assumption. Similarly, it follows

$$\left\langle md\vec{k}\right| nc\vec{k}\right\rangle = \delta_{m,n}\delta_{d,c}. \tag{C.9}$$

In order to calculate $\left\langle md\vec{k}\right| H \left|na\vec{k}\right\rangle$, interaction between orbitals sitting on anion atoms, i.e. $d = a$, is considered first,

$$\begin{aligned} \left\langle ma\vec{k}\right| H \left|na\vec{k}\right\rangle &= \frac{1}{N}\sum_{i,j} e^{i\vec{k}\cdot\left(\vec{R}_i+\vec{\nu}_a\right)-i\vec{k}\cdot\left(\vec{R}_j+\vec{\nu}_a\right)} \left\langle ma\vec{R}_j\right| H \left|na\vec{R}_i\right\rangle \\ &= \sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j\right| H \left|na\vec{R}_i\right\rangle \\ &= \sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j\right| \left[-\frac{\hbar^2}{2m}\nabla^2 + \sum_{l}\left[U_{l,a}+U_{l,c}\right]\right] \left|na\vec{R}_i\right\rangle \\ &= \sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j\right| \left[-\frac{\hbar^2}{2m}\nabla^2 + U_{i,a}\right] \left|na\vec{R}_i\right\rangle \\ &\quad + \sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j\right| \sum_{l\neq i} U_{l,a} \left|na\vec{R}_i\right\rangle \\ &\quad + \sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j\right| \sum_{l} U_{l,c} \left|na\vec{R}_i\right\rangle \end{aligned} \tag{C.10}$$

Using atomic Schrödinger equation (C.1) and the orthogonality condition of (C.8), the first term in RHS of (C.10) reduces as follows:

$$\sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j\right| \left[-\frac{\hbar^2}{2m}\nabla^2 + U_{i,a}\right] \left|na\vec{R}_i\right\rangle = \sum_{i} e^{i\vec{k}\cdot\vec{R}_{ji}} \epsilon_{n,a} \left\langle ma\vec{R}_j\right| na\vec{R}_i\right\rangle$$

$$
\begin{aligned}
&= \epsilon_{n,a} \sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \delta_{i,j}\delta_{m,n} \\
&= \epsilon_{n,a}\delta_{m,n}. \qquad \text{(C.11)}
\end{aligned}
$$

The second term in RHS of (C.10) reduces to zero since, if $i \neq j$ then two orbitals are no longer nearest neighbor (recall, nearest neighbors of anion are cations), and if $i = j$ then it become a three center integral since $l \neq i$, which are also assumed to be zero. As a result,

$$\sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle ma\vec{R}_j \right| \sum_{l\neq i} U_{l,a} \left| na\vec{R}_i \right\rangle = 0. \qquad \text{(C.12)}$$

Finally, the third term of (C.10) also reduces to zero, since the nearest neighbor condition satisfied only if $i = j$. However, the potential now is located on a cation atom, which makes it three center integral,

$$\sum_i e^{i\vec{k}\cdot\vec{R}_{ij}} \left\langle ma\vec{R}_j \right| \sum_{l} U_{l,c} \left| na\vec{R}_i \right\rangle = 0 \qquad \text{(C.13)}$$

Substituting (C.11–C.13) into (C.10), it follows

$$\left\langle ma\vec{k} \right| H \left| na\vec{k} \right\rangle = \epsilon_{n,a}\delta_{m,n}. \qquad \text{(C.14)}$$

Similarly, the Hamiltonian matrix elements between two cationic orbitals becomes,

$$\left\langle mc\vec{k} \right| H \left| nc\vec{k} \right\rangle = \epsilon_{n,c}\delta_{m,n}. \qquad \text{(C.15)}$$

Now the matrix elements between orbitals sitting on cation and anion atoms, respectively, are calculated

$$
\begin{aligned}
\left\langle mc\vec{k} \right| H \left| na\vec{k} \right\rangle &= \frac{1}{N} \sum_{i,j} e^{i\vec{k}\cdot\left(\vec{R}_i+\vec{\nu}_a\right)-i\vec{k}\cdot\left(\vec{R}_j+\vec{\nu}_c\right)} \left\langle mc\vec{R}_j \right| H \left| na\vec{R}_i \right\rangle \\
&= e^{-i\vec{k}\cdot\vec{\nu}_c} \sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle mc\vec{R}_j \right| \left[-\frac{\hbar^2}{2m}\nabla^2 + \sum_l \left[U_{l,a} + U_{l,c}\right] \right] \left| na\vec{R}_i \right\rangle \\
&= e^{-i\vec{k}\cdot\vec{\nu}_c} \sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle mc\vec{R}_j \right| \left[-\frac{\hbar^2}{2m}\nabla^2 + \frac{1}{2}\left(U_{i,a} + U_{j,c}\right) \right] \left| na\vec{R}_i \right\rangle \\
&\quad + e^{-i\vec{k}\cdot\vec{\nu}_c} \sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle mc\vec{R}_j \right| \frac{1}{2}\left(U_{i,a} + U_{j,c}\right) \left| na\vec{R}_i \right\rangle \\
&\quad + e^{-i\vec{k}\cdot\vec{\nu}_c} \sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle mc\vec{R}_j \right| \sum_{l\neq i} U_{l,a} \left| na\vec{R}_i \right\rangle \\
&\quad + e^{-i\vec{k}\cdot\vec{\nu}_c} \sum_i e^{i\vec{k}\cdot\vec{R}_{ji}} \left\langle mc\vec{R}_j \right| \sum_{l\neq j} U_{l,c} \left| na\vec{R}_i \right\rangle . \qquad \text{(C.16)}
\end{aligned}
$$

Using (C.1) and the orthogonality condition of (C.8), the first term in the RHS of (C.16) reduces to zero,

$$\begin{aligned} & e^{-i\vec{k}\cdot\vec{\nu}_c}\sum_i e^{i\vec{k}\cdot\vec{R}_{ji}}\left\langle mc\vec{R}_j\right|\left[-\frac{\hbar^2}{2m}\nabla^2+\frac{1}{2}\left(U_{i,a}+U_{j,c}\right)\right]\left|na\vec{R}_i\right\rangle \\ & =\frac{1}{2}e^{-i\vec{k}\cdot\vec{\nu}_c}\sum_i e^{i\vec{k}\cdot\vec{R}_{ji}}\left\langle mc\vec{R}_j\right|\left[\left\{-\frac{\hbar^2}{2m}\nabla^2+U_{i,a}\right\}+\left\{-\frac{\hbar^2}{2m}\nabla^2+U_{j,c}\right\}\right]\left|na\vec{R}_i\right\rangle \\ & =0. \end{aligned} \tag{C.17}$$

Additionally, both the third and fourth terms in RHS of (C.16) are zero since they are not two center integrals. For these two terms, the potential is not on either of the atoms on which orbitals are located, therefore,

$$e^{-i\vec{k}\cdot\vec{\nu}_c}\sum_i e^{i\vec{k}\cdot\vec{R}_{ji}}\left\langle mc\vec{R}_j\right|\sum_{l\neq i}U_{l,a}\left|na\vec{R}_i\right\rangle=0, \tag{C.18}$$

$$e^{-i\vec{k}\cdot\vec{\nu}_c}\sum_i e^{i\vec{k}\cdot\vec{R}_{ji}}\left\langle mc\vec{R}_j\right|\sum_{l\neq j}U_{l,c}\left|na\vec{R}_i\right\rangle=0. \tag{C.19}$$

The second term in RHS of (C.16), however, clearly fulfils two center integral requirment, therefore, definding,

$$U_{ij}^{ac}=\frac{1}{2}\left(U_{i,a}+U_{j,c}\right),$$

and treating nearest neighbor interaction, only four terms survives,

$$\begin{aligned} e^{-i\vec{k}\cdot\vec{\nu}_c}\sum_i e^{i\vec{k}\cdot\vec{R}_{ji}}\left\langle mc\vec{R}_j\right|U_{ij}^{ac}\left|na\vec{R}_i\right\rangle &= e^{i\vec{k}\cdot\left(\vec{R}_{ji}-\vec{\nu}_c\right)}\left\langle mc\vec{R}_j\right|U_{ij}^{ac}\left|na\vec{R}_i\right\rangle\delta_{\vec{R}_{ji},0} \\ &+ e^{i\vec{k}\cdot\left(\vec{R}_{ji}-\vec{\nu}_c\right)}\left\langle mc\vec{R}_j\right|U_{ij}^{ac}\left|na\vec{R}_i\right\rangle\delta_{\vec{R}_{ji},\left[\frac{a_L}{2}0\frac{a_L}{2}\right]} \\ &+ e^{i\vec{k}\cdot\left(\vec{R}_{ji}-\vec{\nu}_c\right)}\left\langle mc\vec{R}_j\right|U_{ij}^{ac}\left|na\vec{R}_i\right\rangle\delta_{\vec{R}_{ji},\left[0\frac{a_L}{2}\frac{a_L}{2}\right]} \\ &+ e^{i\vec{k}\cdot\left(\vec{R}_{ji}-\vec{\nu}_c\right)}\left\langle mc\vec{R}_j\right|U_{ij}^{ac}\left|na\vec{R}_i\right\rangle\delta_{\vec{R}_{ji},\left[\frac{a_L}{2}\frac{a_L}{2}0\right]}. \end{aligned} \tag{C.20}$$

The four terms in (C.20) correspond to anions at following lattice points,

$$\begin{aligned} \vec{R}_{ji} &= [0\ 0\ 0]; \\ \vec{R}_{ji} &= \left[\frac{a_L}{2}\ 0\ \frac{a_L}{2}\right]; \\ \vec{R}_{ji} &= \left[0\ \frac{a_L}{2}\ \frac{a_L}{2}\right]; \\ \vec{R}_{ji} &= \left[\frac{a_L}{2}\ \frac{a_L}{2}\ 0\right]. \end{aligned}$$

and are the nearest-neighbors of the cation associated to j-th lattice point. The relative position of these four anions relative to the cation are,

$$\begin{aligned}
\vec{x}_0 &= \left(+\frac{a_x}{4}, +\frac{a_y}{4}, +\frac{a_z}{4}\right) \\
\vec{x}_1 &= \left(+\frac{a_x}{4}, -\frac{a_y}{4}, -\frac{a_z}{4}\right) \\
\vec{x}_2 &= \left(-\frac{a_x}{4}, +\frac{a_y}{4}, -\frac{a_z}{4}\right) \\
\vec{x}_3 &= \left(-\frac{a_x}{4}, -\frac{a_y}{4}, +\frac{a_z}{4}\right)
\end{aligned} \tag{C.21}$$

Equation (C.20) can now be expressed as,

$$\begin{aligned}
\left\langle mc\vec{k}\right| H \left|na\vec{k}\right\rangle &= e^{-i\vec{k}\cdot\vec{x}_0} \left\langle mc\vec{R}_j\right| U_{i,j}^{ac} \left|na\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},0} \\
&+ e^{-i\vec{k}\cdot\vec{x}_1} \left\langle mc\vec{R}_j\right| U_{i,j}^{ac} \left|na\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},\left[0\frac{a_L}{2}\frac{a_L}{2}\right]} \\
&+ e^{-i\vec{k}\cdot\vec{x}_2} \left\langle mc\vec{R}_j\right| U_{i,j}^{ac} \left|na\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},\left[\frac{a_L}{2}0\frac{a_L}{2}\right]} \\
&+ e^{-i\vec{k}\cdot\vec{x}_3} \left\langle mc\vec{R}_j\right| U_{i,j}^{ac} \left|na\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},\left[\frac{a_L}{2}\frac{a_L}{2}0\right]}.
\end{aligned} \tag{C.22}$$

Similarly, it follows

$$\begin{aligned}
\left\langle ma\vec{k}\right| H \left|ma\vec{k}\right\rangle &= e^{i\vec{k}\cdot\vec{x}_0} \left\langle ma\vec{R}_j\right| U_{i,j}^{ac} \left|nc\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},0} \\
&+ e^{i\vec{k}\cdot\vec{x}_1} \left\langle ma\vec{R}_j\right| U_{i,j}^{ac} \left|nc\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},\left[0-\frac{a_L}{2}-\frac{a_L}{2}\right]} \\
&+ e^{i\vec{k}\cdot\vec{x}_2} \left\langle ma\vec{R}_j\right| U_{i,j}^{ac} \left|nc\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},\left[-\frac{a_L}{2}0-\frac{a_L}{2}\right]} \\
&+ e^{i\vec{k}\cdot\vec{x}_3} \left\langle ma\vec{R}_j\right| U_{i,j}^{ac} \left|nc\vec{R}_i\right\rangle \delta_{\vec{R}_{ji},\left[-\frac{a_L}{2}-\frac{a_L}{2}0\right]}.
\end{aligned} \tag{C.23}$$

Now, due to crystal symmetry, the four matrix elements in RHS of (C.22) or (C.23) are equal in magnitude, but may differ in signs due to the relative position of neighboring cations with respect to the anion. They can therefore, be combined and written as,

$$H_{mn}^{ac}\left(\vec{k}\right) = \left\langle ma\vec{k}\right| H \left|nc\vec{k}\right\rangle = g_i\left(\vec{k}\right) V_{mn}^{ac}, \tag{C.24}$$

where,

$$V_{mn}^{ac} = \left\langle ma\vec{R}_i\right| U_{i,a} \left|nc\vec{R}_i\right\rangle, \tag{C.25}$$

are the overlap energies, and $g_i\left(\vec{k}\right)$ takes one of the followin expressions,

$$4g_0 = e^{i\vec{k}\bullet\vec{x}_0} + e^{i\vec{k}\bullet\vec{x}_1} + e^{i\vec{k}\bullet\vec{x}_2} + e^{i\vec{k}\bullet\vec{x}_3} \quad \text{(C.26)}$$

$$4g_1 = e^{i\vec{k}\bullet\vec{x}_0} + e^{i\vec{k}\bullet\vec{x}_1} - e^{i\vec{k}\bullet\vec{x}_2} - e^{i\vec{k}\bullet\vec{x}_3} \quad \text{(C.27)}$$

$$4g_2 = e^{i\vec{k}\bullet\vec{x}_0} - e^{i\vec{k}\bullet\vec{x}_1} + e^{i\vec{k}\bullet\vec{x}_2} - e^{i\vec{k}\bullet\vec{x}_3} \quad \text{(C.28)}$$

$$4g_3 = e^{i\vec{k}\bullet\vec{x}_0} - e^{i\vec{k}\bullet\vec{x}_1} - e^{i\vec{k}\bullet\vec{x}_2} + e^{i\vec{k}\bullet\vec{x}_3} \quad \text{(C.29)}$$

The Hamiltonian in (C.7) can now be expressed in the Bloch sum basis of (C.2) as,

$$H = \begin{bmatrix} [H_{aa}] & [H_{ac}] \\ [H_{ca}] & [H_{cc}] \end{bmatrix}. \quad \text{(C.30)}$$

Since spin-orbit interaction, which couples orbitals on same atomic site, has not been treated so far, the diagonal blocs of (C.30) are diagonal matrices with the on-site energies of the orbials on the diagonals.

$$[H_{aa}] = \begin{bmatrix} E_s^a & & & & & & & & & \\ & E_p^a & & & & & & & & \\ & & E_p^a & & & & & & & \\ & & & E_p^a & & & & & & \\ & & & & E_{s^*}^a & & & & & \\ & & & & & E_d^a & & & & \\ & & & & & & E_d^a & & & \\ & & & & & & & E_d^a & & \\ & & & & & & & & E_d^a & \\ & & & & & & & & & E_d^a \end{bmatrix}. \quad \text{(C.31)}$$

$[H_{cc}]$ has a similar structure, except the diagonal elements are on-site energies of cationic orbitals, which are different from those of anion in III-V semiconductors, but are same for elemental semiconductors such as silicon or germanium. The off-diagonal bloc, $[H_{ac}]$, is however a dense 10×10 matrix, written in terms of 5×5 blocs

$$[H_{ac}] = \begin{bmatrix} \left[H^{ac}_{sp^3s^*-sp^3s^*}\right] & \left[H^{ac}_{sp^3s^*-d^5}\right] \\ \left[\ H^{ac}_{d^5-sp^3s^*}\ \right] & \left[\ H^{ac}_{d^5-d^5}\ \right] \end{bmatrix}. \quad \text{(C.32)}$$

where

$$\left[H^{ac}_{sp^3s^*-sp^3s^*}\right] = \begin{bmatrix} V^{ac}_{s,s}g_0 & V^{ac}_{s,x}g_1 & V^{ac}_{s,y}g_2 & V^{ac}_{s,z}g_3 & V^{ac}_{s,s^*}g_0 \\ V^{ac}_{x,s}g_1 & V^{ac}_{x,x}g_0 & V^{ac}_{x,y}g_3 & V^{ac}_{x,z}g_2 & V^{ac}_{x,s^*}g_1 \\ V^{ac}_{y,s}g_2 & V^{ac}_{y,x}g_3 & V^{ac}_{y,y}g_0 & V^{ac}_{y,z}g_1 & V^{ac}_{y,s^*}g_2 \\ V^{ac}_{z,s}g_3 & V^{ac}_{z,x}g_2 & V^{ac}_{z,y}g_1 & V^{ac}_{z,z}g_0 & V^{ac}_{z,s^*}g_3 \\ V^{ac}_{s^*,s}g_0 & V^{ac}_{s^*,x}g_1 & V^{ac}_{s^*,y}g_2 & V^{ac}_{s^*,z}g_3 & V^{ac}_{s^*,s^*}g_0 \end{bmatrix}, \tag{C.33}$$

$$\left[H^{ac}_{sp^3s^*-d^5}\right] = \begin{bmatrix} V^{ac}_{s,xy}g_0 & V^{ac}_{s,yz}g_3 & V^{ac}_{s,zx}g_1 & V^{ac}_{s,x^2-y^2}g_2 & V^{ac}_{s,3z^2-r^2}g_0 \\ V^{ac}_{x,xy}g_1 & V^{ac}_{x,yz}g_2 & V^{ac}_{x,zx}g_0 & V^{ac}_{x,x^2-y^2}g_3 & V^{ac}_{x,3z^2-r^2}g_1 \\ V^{ac}_{y,xy}g_2 & V^{ac}_{y,yz}g_1 & V^{ac}_{y,zx}g_3 & V^{ac}_{y,x^2-y^2}g_0 & V^{ac}_{y,3z^2-r^2}g_2 \\ V^{ac}_{z,xy}g_3 & V^{ac}_{z,yz}g_0 & V^{ac}_{z,zx}g_2 & V^{ac}_{z,x^2-y^2}g_1 & V^{ac}_{z,3z^2-r^2}g_3 \\ V^{ac}_{s^*,xy}g_0 & V^{ac}_{s^*,yz}g_3 & V^{ac}_{s^*,zx}g_1 & V^{ac}_{s^*,x^2-y^2}g_2 & V^{ac}_{s^*,3z^2-r^2}g_0 \end{bmatrix}, \tag{C.34}$$

$$\left[H^{ac}_{d^5-sp^3s^*}\right] = \begin{bmatrix} V^{ac}_{xy,s}g_0 & V^{ac}_{xy,x}g_1 & V^{ac}_{xy,y}g_2 & V^{ac}_{xy,z}g_3 & V^{ac}_{xy,s^*}g_0 \\ V^{ac}_{yz,s}g_3 & V^{ac}_{yz,x}g_2 & V^{ac}_{yz,y}g_1 & V^{ac}_{yz,z}g_0 & V^{ac}_{yz,s^*}g_3 \\ V^{ac}_{zx,s}g_1 & V^{ac}_{zx,x}g_0 & V^{ac}_{zx,y}g_3 & V^{ac}_{zx,z}g_2 & V^{ac}_{zx,s^*}g_1 \\ V^{ac}_{x^2-y^2,s}g_2 & V^{ac}_{x^2-y^2,x}g_3 & V^{ac}_{x^2-y^2,y}g_0 & V^{ac}_{x^2-y^2,z}g_1 & V^{ac}_{x^2-y^2,s^*}g_2 \\ V^{ac}_{3z^2-r^2,s}g_0 & V^{ac}_{3z^2-r^2,x}g_1 & V^{ac}_{3z^2-r^2,y}g_2 & V^{ac}_{3z^2-r^2,z}g_3 & V^{ac}_{3z^2-r^2,s^*}g_0 \end{bmatrix}, \tag{C.35}$$

and

$$\left[H^{ac}_{d^5-d^5}\right] = \begin{bmatrix} V^{ac}_{xy,xy}g_0 & V^{ac}_{xy,yz}g_3 & V^{ac}_{xy,zx}g_1 & V^{ac}_{xy,x^2-y^2}g_2 & V^{ac}_{xy,3z^2-r^2}g_0 \\ V^{ac}_{yz,xy}g_3 & V^{ac}_{yz,yz}g_0 & V^{ac}_{yz,zx}g_2 & V^{ac}_{yz,x^2-y^2}g_1 & V^{ac}_{yz,3z^2-r^2}g_3 \\ V^{ac}_{zx,xy}g_1 & V^{ac}_{zx,yz}g_2 & V^{ac}_{zx,zx}g_0 & V^{ac}_{zx,x^2-y^2}g_3 & V^{ac}_{zx,3z^2-r^2}g_1 \\ V^{ac}_{x^2-y^2,xy}g_2 & V^{ac}_{x^2-y^2,yz}g_1 & V^{ac}_{x^2-y^2,zx}g_3 & V^{ac}_{x^2-y^2,x^2-y^2}g_0 & V^{ac}_{x^2-y^2,3z^2-r^2}g_2 \\ V^{ac}_{3z^2-r^2,xy}g_0 & V^{ac}_{3z^2-r^2,yz}g_3 & V^{ac}_{3z^2-r^2,zx}g_1 & V^{ac}_{3z^2-r^2,x^2-y^2}g_2 & V^{ac}_{3z^2-r^2,3z^2-r^2}g_0 \end{bmatrix}. \tag{C.36}$$

The for $i \in \{0, 1, 2, 3\}$, the g_i factors with each terms obtained from a similar Hamiltonian in [135]. Finally, since tight-binding Hamiltonian, H, in (C.30) must be Harmitian, it follows

$$[H_{ca}] = [H_{ac}]\,.$$

C.2 Overlap Mtrix Elements from Slater-Koster Two-Center-Integrals

In this section, the overlap matrix elements, $V_{i,j}^{ac}$, are expressed in terms of the Slater-Koster two-center energy integrals [60]. Notice that, in [60], many of the matrix elements shown here, are absent. Those expressions are obtained by exploiting the symmetry of the orbitals with respect to the directional cosines l, m, and n. The overlap matrix elements between anion and cation possess additional symmetry when applied to elemental semiconductors such as Si or Ge. In this special case, both anionic and cationic sites are occupied by the same atom and switching the order of the orbitals in the overlap matrix element does not change their magnitude, $\left|V_{i,j}^{ac}\right| = \left|V_{j,i}^{ac}\right|$. Their signs, however, may change depending on the parity of the directional cosines, l, m, and n.

$\langle s_a|$ with all cation orbitals

$$\begin{aligned}
V_{s,s}^{ac} &= V_{ss\sigma}. \\
V_{s,x}^{ac} &= lV_{s_a p_c \sigma}. \\
V_{s,y}^{ac} &= mV_{s_a p_c \sigma}. \\
V_{s,z}^{ac} &= nV_{s_a p_c \sigma}. \\
V_{s,xy}^{ac} &= \sqrt{3}lmV_{s_a d_c \sigma}. \\
V_{s,yz}^{ac} &= \sqrt{3}mnV_{s_a d_c \sigma}. \\
V_{s,zx}^{ac} &= \sqrt{3}nlV_{s_a d_c \sigma}. \\
V_{s,x^2-y^2}^{ac} &= \frac{1}{2}\sqrt{3}\left(l^2 - m^2\right)V_{s_a d_c \sigma}. \\
V_{s,3z^2-r^2}^{ac} &= \left[n^2 - \frac{1}{2}\left(l^2 + m^2\right)\right]V_{s_a d_c \sigma}. \\
V_{s,s^*}^{ac} &= V_{s_a s_c^* \sigma}.
\end{aligned}$$

$\langle p_{x_a}|$ with all cation orbitals

$$\begin{aligned}
V_{x,s}^{ac} &= -lV_{s_c p_a \sigma}. \\
V_{x,x}^{ac} &= l^2 V_{pp\sigma} + \left(1 - l^2\right)V_{pp\pi}.
\end{aligned}$$

$$
\begin{aligned}
V^{ac}_{x,y} &= lm\left(V_{pp\sigma}-V_{pp\pi}\right).\\
V^{ac}_{x,z} &= nl\left(V_{pp\sigma}-V_{pp\pi}\right).\\
V^{ac}_{x,xy} &= \sqrt{3}l^2mV_{p_ad_c\sigma}+m\left(1-2l^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{x,yz} &= lmn\left(\sqrt{3}V_{p_ad_c\sigma}-2V_{p_ad_c\pi}\right).\\
V^{ac}_{x,zx} &= \sqrt{3}nl^2V_{p_ad_c\sigma}+n\left(1-2l^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{x,x^2-y^2} &= \frac{1}{2}\sqrt{3}l\left(l^2-m^2\right)V_{p_ad_c\sigma}+l\left[1-\left(l^2-m^2\right)\right]V_{p_ad_c\pi}.\\
V^{ac}_{x,3z^2-r^2} &= l\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{p_ad_c\sigma}-\sqrt{3}ln^2V_{p_ad_c\pi}.\\
V^{ac}_{x,s^*} &= -lV_{s^*_cp_a\sigma}.
\end{aligned}
$$

$\langle p_{y_a}|$ with all cation orbitals $\langle p_{y_a}|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{y,s} &= -mV_{s_cp_a\sigma}.\\
V^{ac}_{y,x} &= ml\left(V_{pp\sigma}-V_{pp\pi}\right).\\
V^{ac}_{y,y} &= m^2V_{pp\sigma}+\left(1-m^2\right)V_{pp\pi}.\\
V^{ac}_{y,z} &= mn\left(V_{pp\sigma}-V_{pp\pi}\right).\\
V^{ac}_{y,xy} &= \sqrt{3}lm^2V_{p_ad_c\sigma}+l\left(1-2m^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{y,yz} &= \sqrt{3}nm^2V_{p_ad_c\sigma}+n\left(1-2m^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{y,zx} &= lmn\left(\sqrt{3}V_{p_ad_c\sigma}-2V_{p_ad_c\pi}\right).\\
V^{ac}_{y,x^2-y^2} &= \frac{1}{2}\sqrt{3}m\left(l^2-m^2\right)V_{p_ad_c\sigma}-m\left[1+\left(l^2-m^2\right)\right]V_{p_ad_c\pi}.\\
V^{ac}_{y,3z^2-r^2} &= m\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{p_ad_c\sigma}-\sqrt{3}mn^2V_{p_ad_c\pi}.\\
V^{ac}_{y,s^*} &= -mV_{s^*_cp_a\sigma}.
\end{aligned}
$$

$\langle p_{z_a}|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{z,s} &= -nV_{s_cp_a\sigma}.\\
V^{ac}_{z,x} &= nl\left(V_{pp\sigma}-V_{pp\pi}\right).
\end{aligned}
$$

$$
\begin{aligned}
V^{ac}_{z,y} &= nm\left(V_{pp\sigma}-V_{pp\pi}\right).\\
V^{ac}_{z,z} &= n^2V_{pp\sigma}+\left(1-n^2\right)V_{pp\pi}.\\
V^{ac}_{z,xy} &= lmn\left(\sqrt{3}V_{p_ad_c\sigma}-2V_{p_ad_c\pi}\right).\\
V^{ac}_{z,yz} &= \sqrt{3}mn^2V_{p_ad_c\sigma}+m\left(1-2n^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{z,zx} &= \sqrt{3}ln^2V_{p_ad_c\sigma}+l\left(1-2n^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{z,x^2-y^2} &= \frac{1}{2}\sqrt{3}n\left(l^2-m^2\right)V_{p_ad_c\sigma}-n\left(l^2-m^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{z,3z^2-r^2} &= n\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{p_ad_c\sigma}+\sqrt{3}n\left(l^2+m^2\right)V_{p_ad_c\pi}.\\
V^{ac}_{z,s^*} &= -nV_{s^*_cp_a\sigma}.
\end{aligned}
$$

$\langle d_{xy_a}|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{xy,s} &= \sqrt{3}lmV_{s_cd_a\sigma}.\\
V^{ac}_{xy,x} &= -\sqrt{3}l^2mV_{p_cd_a\sigma}-m\left(1-2l^2\right)V_{p_cd_a\pi}.\\
V^{ac}_{xy,y} &= -\sqrt{3}lm^2V_{p_cd_a\sigma}-l\left(1-2m^2\right)V_{p_cd_a\pi}.\\
V^{ac}_{xy,z} &= -lmn\left(\sqrt{3}V_{p_cd_a\sigma}-2V_{p_cd_a\pi}\right).\\
V^{ac}_{xy,xy} &= 3l^2m^2V_{dd\sigma}+\left(l^2+m^2-4l^2m^2\right)V_{dd\pi}+\left(n^2+l^2m^2\right)V_{dd\delta}.\\
V^{ac}_{xy,yz} &= 3lm^2nV_{dd\sigma}+nl\left(1-4m^2\right)V_{dd\pi}+nl\left(m^2-1\right)V_{dd\delta}.\\
V^{ac}_{xy,zx} &= 3l^2mnV_{dd\sigma}+mn\left(1-4l^2\right)V_{dd\pi}+mn\left(l^2-1\right)V_{dd\delta}.\\
V^{ac}_{xy,x^2-y^2} &= \frac{3}{2}lm\left(l^2-m^2\right)V_{dd\sigma}+2lm\left(m^2-l^2\right)V_{dd\pi}+\frac{1}{2}lm\left(l^2-m^2\right)V_{dd\delta}.\\
V^{ac}_{xy,3z^2-r^2} &= \sqrt{3}lm\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma}-2\sqrt{3}lmn^2V_{dd\pi}+\frac{1}{2}\sqrt{3}lm\left(1+n^2\right)V_{dd\delta}.\\
V^{ac}_{xy,s^*} &= \sqrt{3}lmV_{s^*_cd_a\sigma}.
\end{aligned}
$$

$\langle d_{(yz)_a}\Big|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{yz,s} &= \sqrt{3}mnV_{s_cd_a\sigma}.\\
V^{ac}_{yz,x} &= -lmn\left(\sqrt{3}V_{p_cd_a\sigma}-2V_{p_cd_a\pi}\right).
\end{aligned}
$$

$$
\begin{aligned}
V^{ac}_{yz,y} &= -\sqrt{3}nm^2V_{p_cd_a\sigma} - n\left(1-2m^2\right)V_{p_cd_a\pi}.\\
V^{ac}_{yz,z} &= -\sqrt{3}mn^2V_{p_cd_a\sigma} - m\left(1-2n^2\right)V_{p_cd_a\pi}.\\
V^{ac}_{yz,xy} &= 3lm^2nV_{dd\sigma} + nl\left(1-4m^2\right)V_{dd\pi} + nl\left(m^2-1\right)V_{dd\delta}.\\
V^{ac}_{yz,yz} &= 3m^2n^2V_{dd\sigma} + \left(m^2+n^2-4m^2n^2\right)V_{dd\pi} + \left(l^2+m^2n^2\right)V_{dd\delta}.\\
V^{ac}_{yz,zx} &= 3lmn^2V_{dd\sigma} + lm\left(1-4n^2\right)V_{dd\pi} + lm\left(n^2-1\right)V_{dd\delta}.\\
V^{ac}_{yz,x^2-y^2} &= \frac{3}{2}mn\left(l^2-m^2\right)V_{dd\sigma} - mn\left[1+2\left(l^2-m^2\right)\right]V_{dd\pi}\\
&\qquad +mn\left[1+\frac{1}{2}\left(l^2-m^2\right)\right]V_{dd\delta}.\\
V^{ac}_{yz,3z^2-r^2} &= \sqrt{3}mn\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma} - \sqrt{3}mn\left(l^2+m^2-n^2\right)V_{dd\pi}\\
&\qquad -\frac{1}{2}\sqrt{3}mn\left(l^2+m^2\right)V_{dd\delta}.\\
V^{ac}_{yz,s^*} &= \sqrt{3}mnV_{s^*_cd_a\sigma}.
\end{aligned}
$$

$\left\langle d_{(zx)_a}\right|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{zx,s} &= \sqrt{3}nlV_{s_cd_a\sigma}.\\
V^{ac}_{zx,x} &= -\sqrt{3}l^2nV_{p_cd_a\sigma} - n\left(1-2l^2\right)V_{p_cd_a\pi}.\\
V^{ac}_{zx,y} &= -lmn\left(\sqrt{3}V_{p_cd_a\sigma} - 2V_{p_cd_a\pi}\right).\\
V^{ac}_{zx,z} &= -\sqrt{3}ln^2V_{p_cd_a\sigma} - l\left(1-2n^2\right)V_{p_cd_a\pi}.\\
V_{(zx)_a,(xy)_c} &= 3l^2mnV_{dd\sigma} + mn\left(1-4l^2\right)V_{dd\pi} + mn\left(l^2-1\right)V_{dd\delta}.\\
V^{ac}_{zx,yz} &= 3lmn^2V_{dd\sigma} + lm\left(1-4n^2\right)V_{dd\pi} + lm\left(n^2-1\right)V_{dd\delta}.\\
V^{ac}_{zx,zx} &= 3l^2n^2V_{dd\sigma} + \left(l^2+n^2-4l^2n^2\right)V_{dd\pi} + \left(m^2+l^2n^2\right)V_{dd\delta}.\\
V^{ac}_{zx,x^2-y^2} &= \frac{3}{2}nl\left(l^2-m^2\right)V_{dd\sigma} + nl\left[1-2\left(l^2-m^2\right)\right]V_{dd\pi}\\
&\qquad -nl\left[1-\frac{1}{2}\left(l^2-m^2\right)\right]V_{dd\delta}.\\
V^{ac}_{zx,3z^2-r^2} &= \sqrt{3}nl\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma} + \sqrt{3}nl\left(l^2+m^2-n^2\right)V_{dd\pi}\\
&\qquad -\frac{1}{2}\sqrt{3}nl\left(l^2+m^2\right)V_{dd\delta}.\\
V^{ac}_{zx,s^*} &= \sqrt{3}nlV_{s^*_cd_a\sigma}.
\end{aligned}
$$

$\langle d_{(x^2-y^2)_a}|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{x^2-y^2,s} &= \frac{1}{2}\sqrt{3}\left(l^2-m^2\right)V_{s_c d_a\sigma}.\\
V^{ac}_{x^2-y^2,x} &= -\frac{1}{2}\sqrt{3}l\left(l^2-m^2\right)V_{p_c d_a\sigma} - l\left[1-\left(l^2-m^2\right)\right]V_{p_c d_a\pi}.\\
V^{ac}_{x^2-y^2,y} &= -\frac{1}{2}\sqrt{3}m\left(l^2-m^2\right)V_{p_c d_a\sigma} + m\left[1+\left(l^2-m^2\right)\right]V_{p_c d_a\pi}.\\
V^{ac}_{x^2-y^2,z} &= -\frac{1}{2}\sqrt{3}n\left(l^2-m^2\right)V_{p_c d_a\sigma} + n\left(l^2-m^2\right)V_{p_c d_a\pi}.\\
V^{ac}_{x^2-y^2,xy} &= \frac{3}{2}lm\left(l^2-m^2\right)V_{dd\sigma} + 2lm\left(m^2-l^2\right)V_{dd\pi} + \frac{1}{2}lm\left(l^2-m^2\right)V_{dd\delta}.\\
V^{ac}_{x^2-y^2,yz} &= \frac{3}{2}mn\left(l^2-m^2\right)V_{dd\sigma} - mn\left[1+2\left(l^2-m^2\right)\right]V_{dd\pi}\\
&\qquad + mn\left[1+\frac{1}{2}\left(l^2-m^2\right)\right]V_{dd\delta}.\\
V^{ac}_{x^2-y^2,zx} &= \frac{3}{2}nl\left(l^2-m^2\right)V_{dd\sigma} + nl\left[1-2\left(l^2-m^2\right)\right]V_{dd\pi}\\
&\qquad - nl\left[1-\frac{1}{2}\left(l^2-m^2\right)\right]V_{dd\delta}.\\
V^{ac}_{x^2-y^2,x^2-y^2} &= \frac{3}{4}\left(l^2-m^2\right)^2 V_{dd\sigma} + \left[l^2+m^2-\left(l^2-m^2\right)^2\right]V_{dd\pi}\\
&\qquad + \left[n^2+\frac{1}{4}\left(l^2-m^2\right)^2\right]V_{dd\delta}.\\
V^{ac}_{x^2-y^2,3z^2-r^2} &= \frac{1}{2}\sqrt{3}\left(l^2-m^2\right)\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma} + \sqrt{3}n^2\left(m^2-l^2\right)V_{dd\pi}\\
&\qquad + \frac{1}{4}\sqrt{3}\left(1+n^2\right)\left(l^2-m^2\right)V_{dd\delta}.\\
V^{ac}_{x^2-y^2,s^*} &= \frac{1}{2}\sqrt{3}\left(l^2-m^2\right)V_{s_c^* d_a\sigma}.
\end{aligned}
$$

$\langle d_{3z^2-r^2}|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{3z^2-r^2,s} &= \left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{s_c d_a\sigma}.\\
V^{ac}_{3z^2-r^2,x} &= -l\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{p_c d_a\sigma} + \sqrt{3}ln^2 V_{p_c d_a\pi}.\\
V^{ac}_{3z^2-r^2,y} &= -m\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{p_c d_a\sigma} + \sqrt{3}mn^2 V_{p_c d_a\pi}.\\
V^{ac}_{3z^2-r^2,z} &= -n\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{p_c d_a\sigma} - \sqrt{3}\left(l^2+m^2\right)V_{p_c d_a\pi}.\\
V^{ac}_{3z^2-r^2,xy} &= \sqrt{3}lm\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma} - 2\sqrt{3}lmn^2 V_{dd\pi}\\
&\qquad + \frac{1}{2}\sqrt{3}lm\left(1+n^2\right)V_{dd\delta}.
\end{aligned}
$$

$$
\begin{aligned}
V^{ac}_{3z^2-r^2,yz} &= \sqrt{3}mn\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma}+\sqrt{3}mn\left(l^2+m^2-n^2\right)V_{dd\pi} \\
&\quad -\frac{1}{2}\sqrt{3}mn\left(l^2+m^2\right)V_{dd\delta}. \\
V^{ac}_{3z^2-r^2,zx} &= \sqrt{3}nl\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma}+\sqrt{3}nl\left(l^2+m^2-n^2\right)V_{dd\pi} \\
&\quad -\frac{1}{2}\sqrt{3}nl\left(l^2+m^2\right)V_{dd\delta}. \\
V^{ac}_{3z^2-r^2,x^2-y^2} &= \frac{1}{2}\sqrt{3}\left(l^2-m^2\right)\left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{dd\sigma}+\sqrt{3}n^2\left(m^2-l^2\right)V_{dd\pi} \\
&\quad +\frac{1}{4}\sqrt{3}\left(1+n^2\right)\left(l^2-m^2\right)V_{dd\delta}. \\
V^{ac}_{3z^2-r^2,3z^2-r^2} &= \left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]^2V_{dd\sigma}+3n^2\left(l^2+m^2\right)V_{dd\pi} \\
&\quad +\frac{3}{4}\left(l^2+m^2\right)^2V_{dd\delta}. \\
V^{ac}_{3z^2-r^2,s^*} &= \frac{1}{2}\sqrt{3}\left(l^2-m^2\right)V_{s^*_c d_a\sigma}.
\end{aligned}
$$

$\langle s^*_a|$ with all cation orbitals

$$
\begin{aligned}
V^{ac}_{s^*,s} &= V_{s^*_a s_c\sigma}. \\
V^{ac}_{s^*,x} &= lV_{s^*_a p_c\sigma}. \\
V^{ac}_{s^*,y} &= mV_{s^*_a p_c\sigma}. \\
V^{ac}_{s^*,z} &= nV_{s^*_a p_c\sigma}. \\
V^{ac}_{s^*,xy} &= \sqrt{3}lmV_{s^*_a d_c\sigma}. \\
V^{ac}_{s^*,yz} &= \sqrt{3}mnV_{s^*_a d_c\sigma}. \\
V^{ac}_{s^*,zx} &= \sqrt{3}nlV_{s^*_a d_c\sigma}. \\
V^{ac}_{s^*,x^2-y^2} &= \frac{1}{2}\sqrt{3}\left(l^2-m^2\right)V_{s^*_a d_c\sigma}. \\
V^{ac}_{s^*,3z^2-r^2} &= \left[n^2-\frac{1}{2}\left(l^2+m^2\right)\right]V_{s^*_a d_c\sigma}. \\
V^{ac}_{s^*,s^*} &= V_{s^*_a s^*_c\sigma}.
\end{aligned}
$$

Using these overlap matrix elements, the overlap energy bloc becomes

$$
[H_{ac}] = \left[\bar{U}_1, \bar{U}_2\right], \tag{C.37}
$$

where

$$\bar{U}_1 = \begin{bmatrix} V^{ac}_{s,s}g_0 & V^{ac}_{s,x}g_1 & V^{ac}_{s,y}g_2 & V^{ac}_{s,z}g_3 & V^{ac}_{s,s^*}g_0 \\ V^{ac}_{x,s}g_1 & V^{ac}_{x,x}g_0 & V^{ac}_{x,y}g_3 & V^{ac}_{x,z}g_2 & V^{ac}_{x,s^*}g_1 \\ V^{ac}_{y,s}g_2 & V^{ac}_{y,x}g_3 & V^{ac}_{y,y}g_0 & V^{ac}_{y,z}g_1 & V^{ac}_{y,s^*}g_2 \\ V^{ac}_{z,s}g_3 & V^{ac}_{z,x}g_2 & V^{ac}_{z,y}g_1 & V^{ac}_{z,z}g_0 & V^{ac}_{z,s^*}g_3 \\ V^{ac}_{s^*,s}g_0 & V^{ac}_{s^*,x}g_1 & V^{ac}_{s^*,y}g_2 & V^{ac}_{s^*,z}g_3 & V^{ac}_{s^*,s^*}g_0 \\ V^{ac}_{xy,s}g_0 & V^{ac}_{xy,x}g_1 & V^{ac}_{xy,y}g_2 & V^{ac}_{xy,z}g_3 & V^{ac}_{xy,s^*}g_0 \\ V^{ac}_{yz,s}g_3 & V^{ac}_{yz,x}g_2 & V^{ac}_{yz,y}g_1 & V^{ac}_{yz,z}g_0 & V^{ac}_{yz,s^*}g_3 \\ V^{ac}_{zx,s}g_1 & V^{ac}_{zx,x}g_0 & V^{ac}_{zx,y}g_3 & V^{ac}_{zx,z}g_2 & V^{ac}_{zx,s^*}g_1 \\ V^{ac}_{x^2-y^2,s}g_2 & V^{ac}_{x^2-y^2,x}g_3 & V^{ac}_{x^2-y^2,y}g_0 & V^{ac}_{x^2-y^2,z}g_1 & V^{ac}_{x^2-y^2,s^*}g_2 \\ V^{ac}_{3z^2-r^2,s}g_0 & V^{ac}_{3z^2-r^2,x}g_1 & V^{ac}_{3z^2-r^2,y}g_2 & V^{ac}_{3z^2-r^2,z}g_3 & V^{ac}_{3z^2-r^2,s^*}g_0 \end{bmatrix} \tag{C.38}$$

and,

$$\bar{U}_1 = \begin{bmatrix} V^{ac}_{s,xy}g_0 & V^{ac}_{s,yz}g_3 & V^{ac}_{s,zx}g_1 & V^{ac}_{s,x^2-y^2}g_2 & V^{ac}_{s,3z^2-r^2}g_0 \\ V^{ac}_{x,xy}g_1 & V^{ac}_{x,yz}g_2 & V^{ac}_{x,zx}g_0 & V^{ac}_{x,x^2-y^2}g_3 & V^{ac}_{x,3z^2-r^2}g_1 \\ V^{ac}_{y,xy}g_2 & V^{ac}_{y,yz}g_1 & V^{ac}_{y,zx}g_3 & V^{ac}_{y,x^2-y^2}g_0 & V^{ac}_{y,3z^2-r^2}g_2 \\ V^{ac}_{z,xy}g_3 & V^{ac}_{z,yz}g_0 & V^{ac}_{z,zx}g_2 & V^{ac}_{z,x^2-y^2}g_1 & V^{ac}_{z,3z^2-r^2}g_3 \\ V^{ac}_{s^*,xy}g_0 & V^{ac}_{s^*,yz}g_3 & V^{ac}_{s^*,zx}g_1 & V^{ac}_{s^*,x^2-y^2}g_2 & V^{ac}_{s^*,3z^2-r^2}g_0 \\ V^{ac}_{xy,xy}g_0 & V^{ac}_{xy,yz}g_3 & V^{ac}_{xy,zx}g_1 & V^{ac}_{xy,x^2-y^2}g_2 & V^{ac}_{xy,3z^2-r^2}g_0 \\ V^{ac}_{yz,xy}g_3 & V^{ac}_{yz,yz}g_0 & V^{ac}_{yz,zx}g_2 & V^{ac}_{yz,x^2-y^2}g_1 & V^{ac}_{yz,3z^2-r^2}g_3 \\ V^{ac}_{zx,xy}g_1 & V^{ac}_{zx,yz}g_2 & V^{ac}_{zx,zx}g_0 & V^{ac}_{zx,x^2-y^2}g_3 & V^{ac}_{zx,3z^2-r^2}g_1 \\ V^{ac}_{x^2-y^2,xy}g_2 & V^{ac}_{x^2-y^2,yz}g_1 & V^{ac}_{x^2-y^2,zx}g_3 & V^{ac}_{x^2-y^2,x^2-y^2}g_0 & V^{ac}_{x^2-y^2,3z^2-r^2}g_2 \\ V^{ac}_{3z^2-r^2,xy}g_0 & V^{ac}_{3z^2-r^2,yz}g_3 & V^{ac}_{3z^2-r^2,zx}g_1 & V^{ac}_{3z^2-r^2,x^2-y^2}g_2 & V^{ac}_{3z^2-r^2,3z^2-r^2}g_0 \end{bmatrix}. \tag{C.39}$$

C.3 Treating Spin-orbit Coupling in Tight-binding Formalism

In Sec. C.1, tight-binding Hamiltonian has been derived without treating the spin-orbit (SO) coupling. In this section, spin-orbit Hamiltonian matrix elements is be calculated in $sp^3d^5s^*$ atomic orbital basis. In practice, the spin-orbit interaction

of d type orbitals are usually ignored for semiconductors, since they don't have any effect on the bandstructure near the band gap, and therefore, are not important. Spin-orbit interaction happens between orbitals located on the same atom, and not between orbitals sitting on neighboring atoms. Therefore, SO coupling not only adds off-diagonal elements to the otherwise diagonal same spin blocs, $[H_{aa}]$ and $[H_{cc}]$, but also adds matrix elements between opposit spin orbitals located on the same atom. The tight-binding Hamiltonian, without SO coupling, can be written as,

$$H_{No\ Spin} = \begin{bmatrix} \begin{bmatrix} [H_{a\uparrow a\uparrow}] & [H_{a\uparrow c\uparrow}] \\ [H_{c\uparrow a\uparrow}] & [H_{c\uparrow c\uparrow}] \end{bmatrix} & \\ & \begin{bmatrix} [H_{a\downarrow a\downarrow}] & [H_{a\downarrow c\downarrow}] \\ [H_{a\downarrow a\downarrow}] & [H_{a\downarrow c\downarrow}] \end{bmatrix} \end{bmatrix}. \tag{C.40}$$

wherer no matrix elements are present between the up and down spin orbitals. For clarity, the blocs of (C.40) can be rearranged to the following form,

$$H_{No\ Spin} = \begin{bmatrix} \begin{bmatrix} [H_{a\uparrow a\uparrow}] & \\ & [H_{a\downarrow a\downarrow}] \end{bmatrix} & \begin{bmatrix} [H_{a\uparrow c\uparrow}] & \\ & [H_{a\downarrow c\downarrow}] \end{bmatrix} \\ \begin{bmatrix} [H_{c\uparrow a\uparrow}] & \\ & [H_{c\downarrow a\downarrow}] \end{bmatrix} & \begin{bmatrix} [H_{c\uparrow c\uparrow}] & \\ & [H_{c\downarrow c\downarrow}] \end{bmatrix} \end{bmatrix} \tag{C.41}$$

where the large diagonal blocs contain Hamiltonian blocs from same atom, but with opposit spins. Since SO coupling present only between orbitals located on the same atom, the SO Hamiltonian, H_{SO}, has the following structure,

$$H_{SO} = \begin{bmatrix} \begin{bmatrix} [H_{SO}\uparrow\uparrow] & [H_{SO}\uparrow\downarrow] \\ [H_{SO}\downarrow\uparrow] & [H_{SO}\downarrow\downarrow] \end{bmatrix} & \\ & \begin{bmatrix} [H_{SO}\uparrow\uparrow] & [H_{SO}\uparrow\downarrow] \\ [H_{SO}\downarrow\uparrow] & [H_{SO}\downarrow\downarrow] \end{bmatrix} \end{bmatrix} \tag{C.42}$$

The various blocs of (C.41) are now calculated. With the nuclear electric field $\vec{E}$ and momentum operator $\vec{p}$, spin-orbit Hamiltonian, H_{SO}, is expressed as follows,

$$H_{SO} = \frac{q\hbar}{4m^2c^2}\vec{\sigma}\cdot\left(\vec{E}\times\vec{p}\right)$$

$$
\begin{aligned}
&= \frac{q\hbar}{4m^2c^2}\begin{bmatrix} (E_x p_y E_y p_x) & (E_y p_z - E_z p_y) - i\,(E_z p_x - E_x p_z) \\ (E_y p_z - E_z p_y) + i\,(E_z p_x - E_x p_z) & -(E_x p_y - E_y p_x) \end{bmatrix} \\
&= \begin{bmatrix} SO\uparrow\uparrow & SO\uparrow\downarrow \\ SO\downarrow\uparrow & SO\downarrow\downarrow \end{bmatrix} \qquad \text{(C.43)}
\end{aligned}
$$

where $p_x = -i\hbar\partial/\partial x$, $p_y = -i\hbar\partial/\partial y$ and $p_z = -i\hbar\partial/\partial z$. This H_{SO} is then be expressed in atomistic orbital basis.

$$
\begin{aligned}
\langle p_x\uparrow|\, H_{SO}\, |p_y\uparrow\rangle &= \begin{bmatrix} \langle p_x\uparrow| & 0 \end{bmatrix}\begin{bmatrix} SO\uparrow\uparrow & SO\uparrow\downarrow \\ SO\downarrow\uparrow & SO\downarrow\downarrow \end{bmatrix}\begin{bmatrix} |p_y\uparrow\rangle \\ 0 \end{bmatrix} \\
&= \frac{q\hbar}{4m^2c^2}\langle p_x\uparrow|\,(E_x p_y - E_y p_x)\,|p_y\uparrow\rangle \\
&= -i\frac{q\hbar^2}{4m^2c^2}\langle p_x\uparrow|\left(E_x\frac{\partial}{\partial y} - E_y\frac{\partial}{\partial x}\right)|p_y\uparrow\rangle \\
&= -i\delta
\end{aligned}
$$

$\langle p_y\uparrow|\, H_{SO}\, |p_x\uparrow\rangle$ directly follows from the Hermitian nature of H_{SO}

$$
\begin{aligned}
\langle p_y\uparrow|\, H_{SO}\, |p_x\uparrow\rangle &= \langle p_x\uparrow|\, H_{SO}\, |p_y\uparrow\rangle^{\dagger} \\
&= i\delta
\end{aligned}
$$

In order to evaluate other SO matrix elements, the following information becomes useful:

1. the electric field, E, and the p-orbitals are odd spatial functions along their respective axes,
2. product of even number of odd functions is even function, odd number of odd function is odd function, and any number of even function is a even function, and finally,
3. derivative of an even function is an odd function and vice versa.

Only those matrix elements, where even functions along all spatial axes results, produce nonzero values. Using these conditions other SO matrix elements can also be evaulated where only the following SO matrix elements survive:

$$
\langle p_x\downarrow|\, H_{SO}\, |p_y\downarrow\rangle = i\delta
$$

$$
\begin{aligned}
\langle p_y \downarrow| H_{SO} |p_x \downarrow\rangle &= -i\delta \\
\langle p_x \uparrow| H_{SO} |p_z \downarrow\rangle &= \delta \\
\langle p_z \downarrow| H_{SO} |p_x \uparrow\rangle &= \delta \\
\langle p_y \uparrow| H_{SO} |p_z \downarrow\rangle &= -i\delta \\
\langle p_z \downarrow| H_{SO} |p_y \uparrow\rangle &= i\delta \\
\langle p_z \uparrow| H_{SO} |p_x \downarrow\rangle &= -\delta \\
\langle p_x \downarrow| H_{SO} |p_z \uparrow\rangle &= -\delta \\
\langle p_z \uparrow| H_{SO} |p_y \downarrow\rangle &= i\delta \\
\langle p_y \downarrow| H_{SO} |p_z \uparrow\rangle &= -i\delta.
\end{aligned}
$$

Remaining matrix elements are zero by symmetry consideration. Now, from (C.42), the spin-orbit Hamiltonian blocs between the sp^3s^* orbitals can be written as,

$$
[H_{SO} \uparrow\uparrow]_{sp^3s^*,sp^3s^*} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -i\delta & 0 & 0 \\ 0 & +i\delta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{C.44}
$$

$$
[H_{SO} \downarrow\downarrow]_{sp^3s^*,sp^3s^*} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +i\delta & 0 & 0 \\ 0 & -i\delta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{C.45}
$$

$$
[H_{SO} \uparrow\downarrow]_{sp^3s^*,sp^3s^*} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & +\delta & 0 \\ 0 & 0 & 0 & -i\delta & 0 \\ 0 & -\delta & +i\delta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{C.46}
$$

$$[H_{SO}\downarrow\uparrow]_{sp^3s^*,sp^3s^*} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\delta & 0 \\ 0 & 0 & 0 & -i\delta & 0 \\ 0 & +\delta & +i\delta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{C.47}$$

The SO blocs in (C.42) now becomes,

$$[H_{SO}\uparrow\uparrow] = \begin{bmatrix} [H_{SO}\uparrow\uparrow]_{sp^3s^*,sp^3s^*} & [H_{SO}\uparrow\uparrow]_{sp^3s^*,d^5} \\ [H_{SO}\uparrow\uparrow]_{d^5,sp^3s^*} & [H_{SO}\uparrow\uparrow]_{d^5,d^5} \end{bmatrix}, \tag{C.48}$$

$$[H_{SO}\uparrow\downarrow] = \begin{bmatrix} [H_{SO}\uparrow\downarrow]_{sp^3s^*,sp^3s^*} & [H_{SO}\uparrow\downarrow]_{sp^3s^*,d^5} \\ [H_{SO}\uparrow\downarrow]_{d^5,sp^3s^*} & [H_{SO}\uparrow\downarrow]_{d^5,d^5} \end{bmatrix}, \tag{C.49}$$

$$[H_{SO}\downarrow\uparrow] = \begin{bmatrix} [H_{SO}\downarrow\uparrow]_{sp^3s^*,sp^3s^*} & [H_{SO}\downarrow\uparrow]_{sp^3s^*,d^5} \\ [H_{SO}\downarrow\uparrow]_{d^5,sp^3s^*} & [H_{SO}\downarrow\uparrow]_{d^5,d^5} \end{bmatrix}, \tag{C.50}$$

$$[H_{SO}\downarrow\downarrow] = \begin{bmatrix} [H_{SO}\downarrow\downarrow]_{sp^3s^*,sp^3s^*} & [H_{SO}\downarrow\downarrow]_{sp^3s^*,d^5} \\ [H_{SO}\downarrow\downarrow]_{d^5,sp^3s^*} & [H_{SO}\downarrow\downarrow]_{d^5,d^5} \end{bmatrix}, \tag{C.51}$$

where, except $sp^3s^* - sp^3s^*$ blocs, all other blocs are 5×5 zero matrices. Finally, the Hamiltonian with SO coupling are the sum of (C.41) and (C.42),

$$H_{Total} = H_{NoSpin} + H_{SO}. \tag{C.52}$$

C.4 Application to Finite Dimensional Structures: UTB MOSFETs

In Sec. C.1, bulk tight-binding Hamiltonian was derived under the assumption of infinite lattice periodicity along all three orthogonal spatial axes. However, real devices are finite dimensional where along one or more directions infinite periodicity is absent. The wave vector, k, does not have any physical significance along these finite dimensional directions. In order to apply tight-binding approach in nanostructures, in this section we will describe how to discretize the bulk Hamiltonian in one or more special directions.

To illustrate the technique of discretization of tight-binding Hamiltonian, we consider an ultra-thin-body device with the thickness along [001] direction, which is the Z-axis of the device coordinate system. Other two axes are are X, transport adn Y width directions, which are along [100] and [010] crystallographic orientations, respectively. Compared to thickness, along both X and Y the device is long enough so that infinite periodicity along these axes are assumed. We will show next how to write the tight-binding Hamiltonian for this finite dimensional system where along Z we need to represent the device atomistically, however, along X and Y plane wave states remains a good assumption.

We first recall that the four terms in each of the expressions in (C.26–C.29) represent the interaction between an anion with its four neighboring cations. Thus, $e^{i\vec{k}\bullet\vec{x}_0}$ represents the interaction between an anion and a cation where the cation is located $\vec{x}_0$ with respect to the anion. Similarly, the other three terms represent similar interactions of the anion with cations located at $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$, respectively. This bond orientation dependence of the terms become clear if we write (C.32) (or equivalently (C.37)) in the following form,

$$[H_{ac}] = X_V e^{i\vec{k}\cdot\vec{x}_0} + X_U e^{i\vec{k}\cdot\vec{x}_1} + Y_U e^{i\vec{k}\cdot\vec{x}_2} + Y_V e^{i\vec{k}\cdot\vec{x}_3}. \tag{C.53}$$

In (C.53), X_V is obtaine by replacing g_0, g_1, g_2 and g_3 by $+1$, and therefore,

$$X_V = \begin{bmatrix} +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \\ +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet & +\bullet \end{bmatrix}. \tag{C.54}$$

Each $\bullet$ in (C.54) represents the corresponding V_{ij}^{ac} in (C.37). Similarly, X_U is obtained by replacing g_0 and g_1 by +1, and g_2 and g_3 by -1 in (C.37). Thus we get,

$$X_U = \begin{bmatrix} +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & +\bullet & -\bullet & +\bullet \\ +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & +\bullet & -\bullet & +\bullet \\ -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & -\bullet & +\bullet & -\bullet \\ -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & -\bullet & +\bullet & -\bullet \\ +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & +\bullet & -\bullet & +\bullet \\ +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & +\bullet & -\bullet & +\bullet \\ -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & -\bullet & +\bullet & -\bullet \\ +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & +\bullet & -\bullet & +\bullet \\ -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & -\bullet & +\bullet & -\bullet \\ +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & +\bullet & -\bullet & +\bullet \end{bmatrix}. \tag{C.55}$$

The third matrix, Y_U is obtained by replacing g_0 and g_2 by +1, and g_1 and g_3 by -1 in (C.37). Thus we get,

$$Y_U = \begin{bmatrix} +\bullet & -\bullet & +\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet \\ -\bullet & +\bullet & -\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet \\ +\bullet & -\bullet & +\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet \\ -\bullet & +\bullet & -\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet \\ +\bullet & -\bullet & +\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet \\ +\bullet & -\bullet & +\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet \\ -\bullet & +\bullet & -\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet \\ -\bullet & +\bullet & -\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet \\ +\bullet & -\bullet & +\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet \\ +\bullet & -\bullet & +\bullet & -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet & +\bullet \end{bmatrix}. \tag{C.56}$$

Finally, the fourth matrix, Y_V, is obtained by replacing g_0 and g_3 by +1, and g_1 and g_2 by -1 in (C.37). Therefore,

$$Y_V = \begin{bmatrix} +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet \\ -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet \\ -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet \\ +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet \\ +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet \\ +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet \\ +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet \\ -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet \\ -\bullet & +\bullet & +\bullet & -\bullet & -\bullet & -\bullet & -\bullet & +\bullet & +\bullet & -\bullet \\ +\bullet & -\bullet & -\bullet & +\bullet & +\bullet & +\bullet & +\bullet & -\bullet & -\bullet & +\bullet \end{bmatrix}. \tag{C.57}$$

Now for a thin film, where the thickness is along [001]($\equiv$ Z), the translation symmetry is broken along Z. In order to treat this, we express (C.53) as,

$$\begin{aligned}
[H_{ac}] &= X_V e^{i\vec{k}\cdot\vec{x}_0} + X_U e^{i\vec{k}\cdot\vec{x}_1} + Y_U e^{i\vec{k}\cdot\vec{x}_2} + Y_V e^{i\vec{k}\cdot\vec{x}_3} \\
&= X_V e^{ik_x a_x/4} e^{ik_y a_y/4} e^{ik_z a_z/4} + X_U e^{ik_x a_x/4} e^{-ik_y a_y/4} e^{-ik_z a_z/4} \\
&\quad + Y_U e^{-ik_x a_x/4} e^{ik_y a_y/4} e^{-ik_z a_z/4} + Y_V e^{-ik_x a_x/4} e^{-ik_y a_y/4} e^{ik_z a_z/4} \\
&= \left[X_V e^{ik_x a_x/4} e^{ik_y a_y/4} + Y_V e^{-ik_x a_x/4} e^{-ik_y a_y/4}\right] e^{ik_z a_z/4} \\
&\quad + \left[X_U e^{ik_x a_x/4} e^{-ik_y a_y/4} + Y_U e^{-ik_x a_x/4} e^{ik_y a_y/4}\right] e^{-ik_z a_z/4} \\
&= [V_{ac}]\, e^{ik_z a_z/4} + [U_{ac}]\, e^{-ik_z a_z/4}.
\end{aligned} \tag{C.58}$$

For unstrained material, $a_x = a_y = a_z = a_L$, but for strained material they can be different. In (C.58), we have defined

$$[V_{ac}] = X_V e^{ik_x a_x/4} e^{ik_y a_y/4} + Y_V e^{-ik_x a_x/4} e^{-ik_y a_y/4}, \tag{C.59}$$

$$[U_{ac}] = X_U e^{ik_x a_x/4} e^{-ik_y a_y/4} + Y_U e^{-ik_x a_x/4} e^{ik_y a_y/4}, \tag{C.60}$$

both of which are function of in-plane wavevector k_x and k_y. In thin-bodies with thickness along Z, k_z does not have any physical significance, and after an inverse

Fourier transformation along Z, as shown in [124], the expontial term in k_z drops out in (C.58). As a result, $[V_{ac}]$ becomes the coupling matrix of an anion layer with the neighboring cation layer along [001] or $+Z$, while, $[U_{ac}]$ becomes the coupling of an anion layer with the neighboring cation layer along $[00\bar{1}]$ or $-Z$ direction. Knowing the coupling matrices between successive atomic layers, it is straightforward to write the thin-film Hamiltonian discretized along [001] orientation as

$$H_{TF}\left(k_x, k_y\right) = \begin{bmatrix} [H_{aa}] & [V_{ac}] & & & \\ [V_{ac}]^\dagger & [H_{cc}] & [U_{ac}]^\dagger & & \\ & [U_{ac}] & [H_{aa}] & [V_{ac}] & \\ & \ddots & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{C.61}$$

The tridiagonal nature of the thin-film Hamiltonian implies that each atomic layer, cation or anion, is connected with two neighboring atomic layer of opposit type, one above and one below. Size of each bloc is $N_b \times N_b$ where N_b is the size of atomic orbital basis set. Number of diagonal blocs in (C.61) is same as number of atomic layers, N_z, along body thickness. The layers are arranged $a - c - a - c - a - \cdots$ in this order. As a result, for each $\{k_x, k_y\}$ pair, size of the Hamiltonian in (C.61) is $(N_b N_z) \times (N_b N_z)$. Abrupt termination of Hamiltonian at the first and last elements or the diagonal blocs cause dangling bonds and the associated surface states. Removal of these states is discussed in Sec. C.5. Finally, since the first bloc diagonal element in (C.61) (equivalently the bottom surface of the thin body) is assumed to be an anionic layer, if the body consists of odd number of layers along thickness, then the last bloc diagonal element element is also anionic. However, if the body thickness consists of even number of layers then the last layer (top surface of thin body) is a cationic layer.

C.5 Passivation of Surface States in Tight-Binding Approach

C.5.1 Anion Atoms at Top and Bottom Surface

The orientations of four bonds from an anion to its neighboring cations are along [111], $[\bar{1}\bar{1}1]$, $[1\bar{1}\bar{1}]$ and $[\bar{1}1\bar{1}]$ crystalographic orientations. These bonds are primarily formed by sp^3-hybridization (linear combination) of $|s\rangle$, $|p_x\rangle$, $|p_y\rangle$ and $|p_z\rangle$ orbitals on the anion atom as [123, 136, 137],

$$\begin{aligned}
\left|sp^3\right\rangle_{[111]} &= \frac{1}{2}\left(|s\rangle + |p_x\rangle + |p_y\rangle + |p_z\rangle\right) \\
\left|sp^3\right\rangle_{[\bar{1}\bar{1}1]} &= \frac{1}{2}\left(|s\rangle - |p_x\rangle - |p_y\rangle + |p_z\rangle\right) \\
\left|sp^3\right\rangle_{[1\bar{1}\bar{1}]} &= \frac{1}{2}\left(|s\rangle + |p_x\rangle - |p_y\rangle - |p_z\rangle\right) \\
\left|sp^3\right\rangle_{[\bar{1}1\bar{1}]} &= \frac{1}{2}\left(|s\rangle - |p_x\rangle + |p_y\rangle - |p_z\rangle\right),
\end{aligned} \tag{C.62}$$

where, the first two bonds are above the (001) crystal plane and the last two are below the (001) plane. These anion-to-cation bonds of (C.62) can be written in a matrix notation as

$$\begin{bmatrix} |sp^3\rangle_{[111]} \\ |sp^3\rangle_{[\bar{1}\bar{1}1]} \\ |sp^3\rangle_{[1\bar{1}\bar{1}]} \\ |sp^3\rangle_{[\bar{1}1\bar{1}]} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}\begin{bmatrix} |s\ \rangle \\ |p_x\rangle \\ |p_y\rangle \\ |p_z\rangle \end{bmatrix} \tag{C.63}$$

From (C.63), the unitary operator that transforms from atomic $\{sp_xp_yp_z\}$ to the hybridized $|sp^3\rangle$ basis can be extracted as

$$[V]_{A\to H} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \tag{C.64}$$

Now, the sp^3 part of the tight-binding Hamiltonian's diagonal bloc dealing with onsite energy of anion is,

$$[H]_{Atom} = \begin{bmatrix} E_s^{aa} & & & \\ & E_p^{aa} & & \\ & & E_p^{aa} & \\ & & & E_p^{aa} \end{bmatrix} \tag{C.65}$$

It is straightforward to transform this on-site energy bloc, which is in the atomic orbital basis, to the hybridized bond-orbital basis,

$$\begin{aligned} [H]_{Hybrid} &= [V]_{A\to H}\,[H]_{Atom}\,[V]^{\dagger}_{A\to H} \\ &= \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{bmatrix}, \end{aligned} \tag{C.66}$$

where,

$$a = \frac{1}{4}\left(E_s^{aa} + 3E_p^{aa}\right) \text{ and } \; b = \frac{1}{4}\left(E_s^{aa} - E_p^{aa}\right).$$

Bottom Interface

For an ultra-thin-body of semiconductor with film growth direction along [001], if the bottom layer is anion type, then the two dangling bonds are $|sp^3\rangle_{[1\bar{1}\bar{1}]}$ and $|sp^3\rangle_{[\bar{1}1\bar{1}]}$, whose energies are the third and fourth diagonal elements of (C.66). In order to remove surface states due to these dangling bonds, their energies are increases by δ_{sp^3}, which is

$$\begin{aligned} [H]_{Hybrid} &= [V]_{A\to H}\,[H]_{Atom}\,[V]^{\dagger}_{A\to H} \\ &= \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a+\delta_{sp^3} & b \\ b & b & b & a+\delta_{sp^3} \end{bmatrix}, \end{aligned} \tag{C.67}$$

This Hamiltonian, although surface states are removed, is still in hybrid bond basis. Therefore, using

$$[H]_{Atom} = [V]^{\dagger}_{A\to H}\,[H]_{Hybrid}\,[V]_{A\to H}\,,$$

the corresponding bloc in atomic basis becomes

$$[H]^{Bottom}_{Atom} = \begin{bmatrix} E_s^{aa}+\frac{\delta_{sp^3}}{2} & 0 & 0 & -\frac{\delta_{sp^3}}{2} \\ 0 & E_p^{aa}+\frac{\delta_{sp^3}}{2} & -\frac{\delta_{sp^3}}{2} & 0 \\ 0 & -\frac{\delta_{sp^3}}{2} & E_p+\frac{\delta_{sp^3}}{2} & 0 \\ -\frac{\delta_{sp^3}}{2} & 0 & 0 & E_p^{aa}+\frac{\delta_{sp^3}}{2} \end{bmatrix}, \tag{C.68}$$

Top Interface

When the top surface is also anion type, the two dangling bonds there are $|sp^3\rangle_{[111]}$ and $|sp^3\rangle_{[\bar{1}\bar{1}\bar{1}]}$. They correspond to the first and second diagonal elements of (C.66), which should be raised by δ_{sp^3} to remove the associated surface states. Following the same treatment above, the new sp^3 Hamiltonian bloc for this top surface become

$$[H]^{Top}_{Atom} = \begin{bmatrix} E_s^{aa}+\frac{\delta_{sp^3}}{2} & 0 & 0 & +\frac{\delta_{sp^3}}{2} \\ 0 & E_p^{aa}+\frac{\delta_{sp^3}}{2} & +\frac{\delta_{sp^3}}{2} & 0 \\ 0 & +\frac{\delta_{sp^3}}{2} & E_p^{aa}+\frac{\delta_{sp^3}}{2} & 0 \\ +\frac{\delta_{sp^3}}{2} & 0 & 0 & E_p^{aa}+\frac{\delta_{sp^3}}{2} \end{bmatrix}, \tag{C.69}$$

C.5.2 Cation Atoms at Top and Bottom Surface

In Sec. C.5.1, the technique to remove surface states due to anionic dangling bonds at top and bottom interfaces is described. The technique can be readily extended for cationic dangling bonds at top and bottom interfaces. The orientations of bonds from a cation atom to its four neighboring anions are along $[\bar{1}11]$, $[1\bar{1}\bar{1}]$, $[\bar{1}\bar{1}\bar{1}]$ and $[11\bar{1}]$ crystallographic orientations, respectively. These bonds are also formed by hybridization (linear combination) of s, p_x, p_y and p_z orbitals as

$$\left|sp^3\right\rangle_{[\bar{1}11]} \;=\; \frac{1}{2}\left(|s\rangle - |p_x\rangle + |p_y\rangle + |p_z\rangle\right)$$

$$\begin{aligned}
\left|sp^3\right\rangle_{[1\bar{1}1]} &= \frac{1}{2}\left(|s\rangle + |p_x\rangle - |p_y\rangle + |p_z\rangle\right) \\
\left|sp^3\right\rangle_{[\bar{1}\bar{1}\bar{1}]} &= \frac{1}{2}\left(|s\rangle - |p_x\rangle - |p_y\rangle - |p_z\rangle\right) \\
\left|sp^3\right\rangle_{[11\bar{1}]} &= \frac{1}{2}\left(|s\rangle + |p_x\rangle + |p_y\rangle - |p_z\rangle\right)
\end{aligned} \tag{C.70}$$

Equations in (C.70) can be written in matrix notation as

$$\begin{bmatrix} |sp^3\rangle_{[\bar{1}11]} \\ |sp^3\rangle_{[1\bar{1}1]} \\ |sp^3\rangle_{[\bar{1}\bar{1}\bar{1}]} \\ |sp^3\rangle_{[11\bar{1}]} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 \end{bmatrix}\begin{bmatrix} |s\ \rangle \\ |p_x\rangle \\ |p_y\rangle \\ |p_z\rangle \end{bmatrix} \tag{C.71}$$

From (C.71), the unitary operator that transforms from atomic $\{sp_xp_yp_z\}$ to hybridized $|sp^3\rangle$ basis is

$$[V]_{A\to H} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 \end{bmatrix} \tag{C.72}$$

Now, the sp^3 part of the tight-binding Hamiltonian dealing with onsite energy is,

$$[H]_{Atom} = \begin{bmatrix} E_s^{cc} & & & \\ & E_p^{cc} & & \\ & & E_p^{cc} & \\ & & & E_p^{cc} \end{bmatrix} \tag{C.73}$$

This on-site energy bloc in atomic orbital basis, can easily be transformed in the hybridized bond-orbital basis as,

$$\begin{aligned}
[H]_{Hybrid} &= [V]_{A\to H}\,[H]_{Atom}\,[V]^{\dagger}_{A\to H} \\
&= \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{bmatrix},
\end{aligned} \tag{C.74}$$

where,

$$a = \frac{1}{4}\left(E_s^{cc} + 3E_p^{cc}\right) \text{and} b = \frac{1}{4}\left(E_s^{cc} - E_p^{cc}\right).$$

Bottom Interface

For an ultra-thin-body of semiconductor with film growth direction along [001], if the bottom layer is cation type, then the two dangling bonds are $|sp^3\rangle_{[\bar{1}\bar{1}\bar{1}]}$ and $|sp^3\rangle_{[11\bar{1}]}$, whose energies are the third and fourth diagonal elements of (C.74). In order to remove surface states due to these dangling bonds, their energies are increases by δ_{sp^3}, which is

$$\begin{aligned} [H]_{Hybrid} &= [V]_{A\to H}\,[H]_{Atom}\,[V]^{\dagger}_{A\to H} \\ &= \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a+\delta_{sp^3} & b \\ b & b & b & a+\delta_{sp^3} \end{bmatrix}, \end{aligned} \tag{C.75}$$

This Hamiltonian, although surface states removed, is in hybrid bond basis, therefore, using

$$[H]_{Atom} = [V]^{\dagger}_{A\to H}\,[H]_{Hybrid}\,[V]_{A\to H},$$

the corresponding bloc in atomic basis becomes

$$[H]^{Bottom}_{Atom} = \begin{bmatrix} E_s^{cc}+\frac{\delta_{sp^3}}{2} & 0 & 0 & -\frac{\delta_{sp^3}}{2} \\ 0 & E_p^{cc}+\frac{\delta_{sp^3}}{2} & -\frac{\delta_{sp^3}}{2} & 0 \\ 0 & -\frac{\delta_{sp^3}}{2} & E_p^{cc}+\frac{\delta_{sp^3}}{2} & 0 \\ -\frac{\delta_{sp^3}}{2} & 0 & 0 & E_p^{cc}+\frac{\delta_{sp^3}}{2} \end{bmatrix}, \tag{C.76}$$

Top Interface

Similarly, for the cation type top surface, the two dangling bonds are $|sp^3\rangle_{[\bar{1}11]}$ and $|sp^3\rangle_{[1\bar{1}1]}$, and therefore, the first and second diagonal elements of (C.74) should be

raised by δ_{sp^3} to remove the associated surface states. Following the same treatment above, the new sp^3 Hamiltonian bloc for this top surface become

$$[H]^{Top}_{Atom} = \begin{bmatrix} E_s^{cc} + \frac{\delta_{sp^3}}{2} & 0 & 0 & +\frac{\delta_{sp^3}}{2} \\ 0 & E_p^{cc} + \frac{\delta_{sp^3}}{2} & +\frac{\delta_{sp^3}}{2} & 0 \\ 0 & +\frac{\delta_{sp^3}}{2} & E_p^{cc} + \frac{\delta_{sp^3}}{2} & 0 \\ +\frac{\delta_{sp^3}}{2} & 0 & 0 & E_p^{cc} + \frac{\delta_{sp^3}}{2} \end{bmatrix}, \tag{C.77}$$

VITA

VITA

Anisur Rahman was born on June 1, 1972 in Khulna, Bangladesh. He received his BSc in Electrical Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 1997 and received the University Gold Medal for ranking top among all the graduating engineers. After that, he joined the Department of EEE at BUET and worked there for two years as a lecturer. Since fall 1999, he pursued a direct PhD program in Electrical Engineering at Purdue University, West Lafayette, IN, which he finished in 2005. Since September 2005, he has been working in the Quality and Reliability group at Intel Corp., Hillsboro, OR.