Abstract of thesis entitled

Efficiency Enhancement for Nanoelectronic Transport Simulations

Submitted by

Jun Huang

for the degree of Doctor of Philosophy
at The University of Hong Kong
in July 2013

Continual technology innovations make it possible to fabricate electronic devices on the order of 10nm. In this nanoscale regime, quantum physics becomes critically important, like energy quantization effects of the narrow channel and the leakage currents due to tunneling. It has also been utilized to build novel devices, such as the band-to-band tunneling field-effect transistors (FETs). Therefore, it presages accurate quantum transport simulations, which not only allow quantitative understanding of the device performances but also provide physical insight and guidelines for device optimizations.

However, quantum transport simulations usually require solving repeatedly the Green's function or the wave function of the whole device region with open boundary treatment, which are computationally cumbersome. Moreover, to overcome the short-channel effects, modern devices usually employ multi-gate structures that are three-dimensional, making the computation very challenging. It is the major target of this thesis to enhance the simulation efficiency by proposing several fast numerical algorithms. The other target is to apply these algorithms to study the physics and performances of some emerging electronic devices.

First, an efficient method is implemented for real space simulations with the effective mass approximation. Based on the wave function approach, asymptotic waveform evaluation combined with a complex frequency hopping algorithm is successfully adopted to characterize electron conduction over a wide energy range. Good accuracy and efficiency are demonstrated by

simulating several n-type multi-gate silicon FETs. This technique is valid for arbitrary potential distribution and device geometry, making it a powerful tool for studying n-type silicon nanowire (SiNW) FETs in the presence of charged impurity and surface roughness scattering.

Second, a model order reduction (MOR) method is proposed for multi-band simulation of nanowire structures. Employing three- or six-band k.p Hamiltonian, the non-equilibrium Green's function (NEGF) equations are projected into a much smaller subspace constructed by sampling the Bloch modes of each cross-section layer. Together with special sampling schemes and Krylov subspace methods for solving the eigenmodes, large cross-section p-type SiNW FETs can be simulated. A novel device, junctionless FET, is then investigated. It is found that its doping density, channel orientation, and channel size need to be carefully optimized in order to outperform the classical inversion-mode FET.

With a spurious band elimination process, the MOR method is subsequently extended to the eight-band k.p model, allowing simulation of band-to-band tunneling devices. In particular, tunneling FETs with indium arsenide (InAs) nanowire channel are studied, considering different channel orientations and configurations with source pockets. Results suggest that source pocket has no significant impact on the performances of the nanowire device due to its good electrostatic integrity.

At last, improvements are made for open boundary treatment in atomistic simulations. The trick is to condense the Hamiltonian matrix of the periodic leads before calculating the surface Green's function. It is very useful for treating leads with long unit cells.

An abstract of exactly 487 words

EFFICIENCY ENHANCEMENT FOR NANOELECTRONIC TRANSPORT SIMULATIONS

BY

JUN HUANG

（黃 俊）

DISSERTATION

Submitted in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
at The University of Hong Kong.

July 2013

Hong Kong

To my wife and my parents

There is plenty of room at the bottom.
—Richard Feynman

DECLARATION

I declare that this thesis represents my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed _____

Jun Huang

ACKNOWLEDGMENTS

I am indebted to many people who have offered me great help during my PhD years. First of all, I would like to express my deepest gratitude and appreciation toward my principal advisor, Professor Weng Cho Chew. During the past four years, Prof. Chew reshaped my view about research. I was always reminded to "identify important problems, understand them deeply, and do science oriented engineering". He defines what a good researcher is like through his inspiring lectures and passion in research. I also would like to thank him for inviting me to study at University of Illinois, Urbana-Champaign (UIUC).

I would also like to express my great thankfulness to my advisor, Dr. Lijun Jiang for giving me comments on paper writing, suggestions on career development, and generous financial support.

I would like to thank my advisor, Prof. Guanhua Chen. I was privileged to attend his group seminars and participate in their software development meetings, from which I benefited a lot.

I am thankful to my former advisor, Prof. Wen Yan Yin, for his advice on my academic planning and giving me the opportunity to interact with his group .

I would like to thank my colleagues and labmates in HKU and UIUC for the ongoing friendship and fruitful discussions all these years. They are Dr. Wei Sha, Dr. Yumao Wu, Dr. Yongpin Chen, Dr. Sheng Sun, Dr. Min Tang, Dr. Yathei Lo, Dr. Yang Liu, Dr. Bo Zhu, Dr. Shaoying Huang, Dr. Peng Yang, Dr. Shulabh Gupta, Yan Li, Qi Dai, Zuhui Ma, Nick Huang, Ping Li, Yanlin Li, Xiaoyan Xiong, Lingling Meng, Qin Liu, Shanshan Gao, Zilong

Ma in HKU and Philip Atkins, Hui Gan, Tian Xia, Aditya Sarathy, Michael Wei, Michael Qiao, Palash Sarker, Aiyin Liu, Junwei Wu, Hanru Shao, Kai Zheng, Dr. Babak Kia, Joe Rutherford, Christopher Ryu from UIUC.

I was fortunate to get involved in a large AoE project, which is "Theory, modeling, and simulation of emerging electronics", where I had the opportunity to collaborate and learn from many great researchers in physics, chemistry, and engineering. Thanks go to Prof. Fuchun Zhang for his quantum mechanics lectures that helped me a lot, Prof. Jian Wang for his solid state physics lectures and many helpful discussions, and Prof. Hong Guo for his interesting talks on first-principle methods. I would like to mention with gratitude Dr. ChiYung Yam from whom I learnt about density functional tight binding (DFTB) method and atomistic simulations, Dr. Jian Sun who showed me how to access the computer clusters and parallelize the code, Dr. Stanislav Markov who offered detailed feedback on my work, and Dr. Ferdows Zahid who shared his viewpoint about quantum transport modeling. I appreciated many helpful discussions with Dr. Quan Chen, Dr. Peng Jie, Dr. Hang Xie, Dr. Yan Zhou, Dr. Yong Wang, Yu Zhang, Qing Zhang, SiuKong Koo, YanHo Kwok, Dr. Heng Tian, Lining Zhang, Zubair Ahmed and Raju Salahuddin.

Finally, I am grateful to my parents and my little sister who stood by me all the time, and to my wife Hanbing Gao for her love, understanding, and encouragement throughout my entire graduate studies.

ABSTRACT

Continual technology innovations make it possible to fabricate electronic devices on the order of 10nm. In this nanoscale regime, quantum physics becomes critically important, like energy quantization effects of the narrow channel and the leakage currents due to tunneling. It has also been utilized to build novel devices, such as the band-to-band tunneling field-effect transistors (FETs). Therefore, it presages accurate quantum transport simulations, which not only allow quantitative understanding of the device performances but also provide physical insight and guidelines for device optimizations.

However, quantum transport simulations usually require solving repeatedly the Green's function or the wave function of the whole device region with open boundary treatment, which are computationally cumbersome. Moreover, to overcome the short-channel effects, modern devices usually employ multi-gate structures that are three-dimensional, making the computation very challenging. It is the major target of this thesis to enhance the simulation efficiency by proposing several fast numerical algorithms. The other target is to apply these algorithms to study the physics and performances of some emerging electronic devices.

First, an efficient method is implemented for real space simulations with the effective mass approximation. Based on the wave function approach, asymptotic waveform evaluation combined with a complex frequency hopping algorithm is successfully adopted to characterize electron conduction over a wide energy range. Good accuracy and efficiency are demonstrated by simulating several n-type multi-gate silicon FETs. This technique is valid for arbitrary potential distribution and device geometry, making it a powerful tool for studying n-type silicon nanowire (SiNW) FETs in the presence of charged impurity and surface roughness scattering.

Second, a model order reduction (MOR) method is proposed for multi-band simulation of nanowire structures. Employing three- or six-band k.p Hamiltonian, the non-equilibrium Green's function (NEGF) equations are projected into a much smaller subspace constructed by sampling the Bloch modes of each cross-section layer. Together with special sampling schemes and Krylov subspace methods for solving the eigenmodes, large cross-section p-type SiNW FETs can be simulated. A novel device, junctionless FET, is then investigated. It is found that its doping density, channel orientation, and channel size need to be carefully optimized in order to outperform the classical inversion-mode FET.

With a spurious band elimination process, the MOR method is subsequently extended to the eight-band k.p model, allowing simulation of band-to-band tunneling devices. In particular, tunneling FETs with indium arsenide (InAs) nanowire channel are studied, considering different channel orientations and configurations with source pockets. Results suggest that source pocket has no significant impact on the performances of the nanowire device due to its good electrostatic integrity.

At last, improvements are made for open boundary treatment in atomistic simulations. The trick is to condense the Hamiltonian matrix of the periodic leads before calculating the surface Green's function. It is very useful for treating leads with long unit cells.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AlAs	Aluminium Arsenide
AWE	Asymptotic Waveform Evaluation
BTBT	Band-To-Band Tunneling
CBR	Contact Block Reduction
CFH	Complex Frequency Hopping
CG	Conjugate Gradient
CMOS	Complementary Metal-Oxide-Semiconductor
CNT	Carbon Nanotube
CPU	Central Processing Unit
DFT	Density Functional Theory
DFTB	Density Functional Tight Binding
DIBL	Drain Induced Barrier Lowering
DKK	Dresselhaus-Kip-Kittel
DOS	Density of States
EGL	Effective Gate Length
EMA	Effective Mass Approximation
EVP	Eigenvalue Problem
FDM	Finite Difference Method
FEM	Finite Element Method
FET	Field-Effect Transistor

FFT	Fast Fourier Transform
GAA	Gate-All-Around
GaAs	Gallium Arsenide
GEVP	Generalized Eigenvalue Problem
GNR	Graphene Nanoribbon
GPU	Graphics Processing Unit
IM	Inversion Mode
InAs	Indium Arsenide
InP	Indium Phosphide
ITRS	International Technology Roadmap for Semiconductors
JL	Junctionless
KPM	$k \cdot p$ Model
LDOS	Local Density of States
MoM	Method of Moment
MOR	Model Order Reduction
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MS	Mode Space
NEGF	Non-Equilibrium Green's Function
NEVP	Normal Eigenvalue Problem
RGF	Recursive Green's Function
RS	Real Space
SiNW	Silicon Nanowire
SS	Sub-threshold Slope (Swing)
SVD	Singular Value Decomposition
TBM	Tight Binding Model
TCAD	Technology Computer Aided Design
TFET	Tunneling Field-Effect Transistor

CHAPTER 1

INTRODUCTION

1.1 Background and Motivations

1.1.1 Emerging Electronics

The quest for high performance and low cost computers and consumer electronics like cell phones has been the driving force of the electronics industry. Besides the enlargement of the chip and the optimization of the circuit (or system), the major approach is to decrease the feature size as it allows more transistors and functional units to be integrated into the same chip. The scaling trend is summarized by Moore's law [1], which is often stated as doubling of transistor performance and quadrupling of the number of devices on a chip every three years. Remarkably, the semiconductor industry has followed this trend for almost half a century, and now the feature size of integrated circuits has already been scaled to only tens of nanometers (22nm in year 2011). In this nanoelectronics era, the traditional CMOS devices suffer from severe short channel effects leading to serious performance degradation, including increased power consumption and larger performance variations. It is generally accepted that novel device engineering by introducing new device structures or new materials is needed to extend Moore's law [2].

According to ITRS (International Technology Roadmap for Semiconductors) [3], alternative materials like carbon nanotube, graphene, and III-V compound semiconductors are expected to replace or complement silicon as channel materials; new structures like nanowires are also being utilized as candidates for replacing traditional planar technologies. Due to its two-dimensional monolayer structure, graphene exhibits outstanding transport properties [4], such as light effective mass and long mean free path. III-V compound materials such as gallium arsenide (GaAs), aluminium arsenide

(AlAs), indium arsenide (InAs), indium phosphide (InP) and their ternary and quaternary alloys draw much attention due to their much higher carrier mobilities than silicon [5]. On the other hand, nanowires allow multi-gate architectures which provide better electrostatic control over the channel [6].

1.1.2 Quantum Transport Simulation

Besides the fabrication issues, these emerging electronic devices also impose a challenging task for the transport modeling and simulation. Modeling and simulation methodologies are very helpful for understanding new device working mechanisms [7]. They also constitute the key part of technology computer aided design (TCAD) tools, which reduce the cost and expedite the product cycle by replacing direct experiments.

Various kinds of transport models exist under different levels of approximations [8]. In the past, the mainstream models were based on semi-classical Boltzmann transport equation and its simplified models, like hydrodynamic equations and drift-diffusion equations. These models can be adequate when the device size is large, but as the device size becomes smaller, quantum mechanical effects start to emerge. Some of the quantum effects that are important for MOSFET applications are the energy quantization due to the small channel cross section, source-to-drain tunneling in ultra-short channel devices, and gate leakage current [9]. To account for these quantum effects, corrections have been made to the classical models, leading to the quantum hydrodynamic equations and quantum Monte Carlo methods [8]. As the device size is further shrunk to be comparable to the electron wavelength, quantum physics dominates. Besides, there are also devices making use of quantum mechanics, such as the resonant tunneling diodes [10] and band-to-band tunneling devices [11]. This necessitates full quantum transport modeling. Boltzmann equation is not compatible with quantum mechanics as it specifies momentum and position at the same time. The Green's function method, in particular the non-equilibrium Green's function (NEGF) approach, is suitable for this purpose due to its ability to capture both the wave physics and the phase-breaking process [12]. It also has the potentiality to treat large devices since it is based on single electron approximation. In the coherent transport limit, it has been proven to be equivalent to the wave

function approach [13].

To model a device quantum mechanically, an appropriate representation must be chosen for the electron Hamiltonian. This is difficult since the crystal potential can be very complicated due to its non-periodicity and the presence of disorders. In addition, there are many electrons interacting with each other. Although *ab initio* methods [14] are preferred from the scientific point of view, empirical models are usually adopted for practical efficiency. The following are three of the most commonly used empirical electronic structure models, each featuring its distinct ability and validity.

Effective Mass Approximation (EMA): It is well known that the conduction band bottom of many common semiconductor materials, such as silicon, germanium, and GaAs, is approximately parabolic and can be described by single band effective masses [15]. Although this is a drastic approximation, it has been extensively used to evaluate the performance of traditional CMOS devices. It has also been widely adopted to study quantum transport in n-type devices due to its simplicity [16]. For ultra-small nanostructures, however, the effective masses vary from the bulk ones [17]. By adjusting the effective masses through mapping to more accurate band structure calculations, EMA still works well down to very small sizes [18].

Multi-Band $k{\cdot}p$ Models (KPMs): In contrast, the valence band top of above mentioned semiconductor materials is not parabolic any more. It actually involves three strongly coupled bands, namely the heavy hole, light hole, and split-off hole band [15]. Still, the band structure in the vicinity of the Γ point can be described by a relatively simple model, the multi-band effective mass $(k \cdot p)$ model [19]. In three-band model, the three bands are treated exactly whereas the remote bands are taken into account approximately. The number of bands doubles when spin-orbit coupling is considered. For narrow and direct band gap materials, like InAs, the two conduction bands are usually included to make an eight-band model. It should be mentioned that this approach is based on perturbation and thus is only valid for a certain energy range. But it suffices for nanoelectronic transport modeling and optoelectronic modeling, as only band structures near the band gap are of interest [20]. Note that by including more bands, full zone band structures can be accurately produced. For instance, 30-band model has been developed recently [21]. Moreover, strain effects can easily be included as a perturbation. The approach is flexible in the way that the number of

bands can be tailored for specific materials or applications.

Tight Binding Models (TBMs): For ultra-small nanostructures, one may question the validity of EMA due to the non-parabolicity nature of the conduction band [15]. Tight binding models are better band structure models and thus can provide more accurate quantization levels and masses [22]. Take the $sp^3d^5s^*$ nearest neighbor TBM for example [23], it can reproduce the band edges of the bulk silicon band structure over the entire Brillouin zone. Its atomistic treatment is another advantage, since the device under study is so small that it can hardly be treated as a continuum system. Furthermore, the dopants and interfaces could be modeled in an atomic resolution.

1.1.3 Numerical Bottlenecks

Once the Hamiltonian is written down, the NEGF (or wave function) equations [24] can be solved routinely. However, this is computationally prohibitive for realistic problems.

To identify the numerical bottlenecks, a typical simulation flow for coherent transport is illustrated in Fig. 1.1. At the beginning, an initial guess of the potential profile is needed to set up the initial Hamiltonian. Then, it requires solving the open boundary Hamiltonian equation repeatedly in an energy band to get the charge density (which is an integral over energy). This consists of two steps, one is calculating self energy matrices of the semi-infinite leads, and the other is solving for the Green's function or wave function of the device region. The charge density obtained is then fed into Poisson equation to get a new potential distribution, which in turn updates the Hamiltonian of the system. The iteration process continues until self consistency is achieved. For transistors, this flow has to be repeated for each gate (drain) bias to get the transfer (output) characteristics. The self energy and Green's function calculations constitute the most expensive parts of the whole simulation since they require inversion of the lead and device Hamiltonian (note that direct inversion is $O(N^3)$ in computational complexity and $O(N^2)$ in memory, where N is the dimension of the Hamiltonian). Therefore, many quantum transport simulations are performed on small model systems and for theoretical study only.

Unfortunately, emerging devices such as silicon nanowire transistors are

three-dimensional. Moreover, devices fabricated recently are usually of size $(10\text{nm})^3$ approximately [25]. Discretization of such big devices usually results in a huge Hamiltonian matrix, even when the simplest EMA is used, not to mention the more sophisticated KPMs and TBMs. In addition, the convergence of the charge density integral is poor when singularities are involved. This situation is commonly encountered when simulating low-dimensional devices due to Van Hove singularities at the sub-band edges and possible bound or resonant states [26].

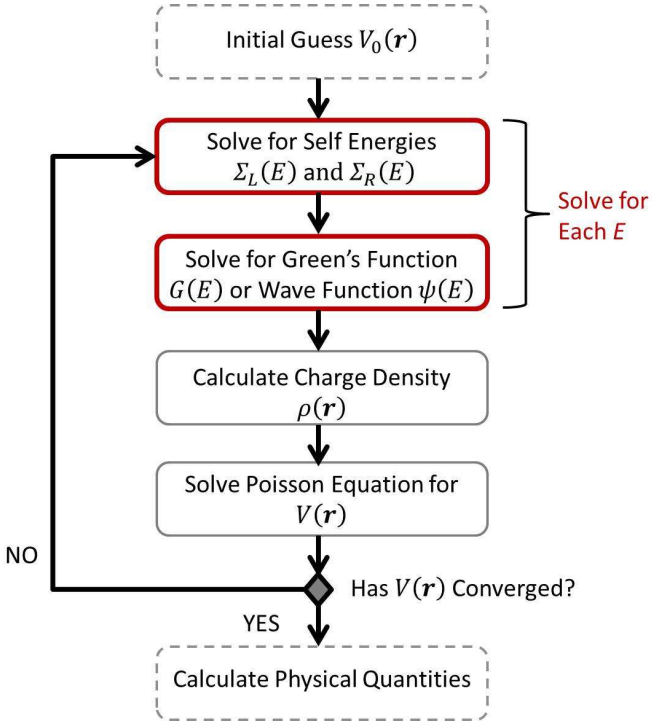


Figure 1.1: Flowchart of device simulation with the NEGF or wave function approach. The steps in the dash-line boxes are calculated once, while the bold line boxes define the numerical bottlenecks.

To break through the above numerical bottlenecks will bridge the gap between theory and experiment, thus to achieve the ultimate goal of interpreting and guiding experiments.

1.2 Outline of the Thesis

This thesis endeavors to break through the numerical bottlenecks by proposing several efficient numerical algorithms. The description of the Hamilto-

nian ranges from effective mass approximation, $k \cdot p$ models, to tight binding models, since each can model a wide range of devices. With these methods, several representative emerging devices are studied as well.

Chapter 2 introduces a method to speed up the simulations with the effective mass approximation. As the wave function needs to be calculated many times in an energy band, the method improves the efficiency by computing the wave function only at several sampled energy points and evaluates its neighboring energy points by an approximated asymptotic form. The sampling points are chosen automatically through a binary searching process. The accuracy and efficiency is demonstrated by the analysis of several n-type silicon transistors. By the implementation of this method, silicon nanowire transistors with the presence of several kinds of imperfection are studied.

Chapter 3 dedicates to accelerating the simulations with the three- and six-band $k \cdot p$ models. The proposed model order reduction (MOR) technique is aimed to reduce the dimension of the $k \cdot p$ Hamiltonian matrix by a transformation of basis. The construction of the reduced basis is discussed in detail. This method is particularly suitable for the simulation of p-type silicon nanowire FETs. The novel junctionless FETs are studied and compared to classical inversion mode FETs.

Chapter 4 extends the MOR technique proposed in Chapter 3 to deal with the eight-band $k \cdot p$ model. A spurious band elimination process is found to be crucial for obtaining a useful reduced model. The method allows simulation of band-to-band tunneling devices. In particular, tunneling FETs with InAs nanowire channel is studied by varying channel orientations and by introducing a source pocket.

Chapter 5 deals with the self-energy matrix calculation, a bottleneck in atomistic simulations. The key step is to condense the Hamiltonian matrix of the periodic atomic leads, after which the unit cell size becomes much smaller and the conventional self-energy calculation methods can be applied with some modifications. Nearest neighbor $sp^3d^5s^*$ tight binding scheme is given as an example to show the validity and efficiency of the methods.

Chapter 6 concludes the thesis by a brief review of the numerical methods proposed and the key device physics found. It also gives some viewpoints about the continual improvement of the nanodevice modeling.

CHAPTER 2

ASYMPTOTIC WAVEFORM EVALUATION FOR SIMULATIONS WITH THE EFFECTIVE MASS APPROXIMATION

Quantum mechanical modeling of ballistic transport in nanodevices usually requires solving Schrödinger equation at multiple energy points within an energy band. To speed up the simulation and analysis, the asymptotic waveform evaluation (AWE) is introduced in this chapter. Using this method, the wave function is only rigorously solved at several sampled energy points, while those at other energies are computed through Padé approximation. This allows us to obtain the physical quantities over the whole energy band with very little computational cost. In addition, the accuracy is controllable by a complex frequency hopping (CFH) algorithm. The validity and efficiency of the proposed method are demonstrated by detailed study of several multi-gate silicon nano-MOSFETs. The method is applied to study silicon nanowire MOSFETs in the presence of charged impurity and surface roughness scattering, it is found that these scattering events have significant impacts on the device variabilities.

2.1 Introduction

Since the dimensions of nanodevices have shrunk to be comparable to electron wavelength, quantum-mechanical modeling of electron transport through these nanodevices is indispensable to capture their wave-physics features. Several quantum transport models have been developed with different levels of approximations [8]. To calculate ballistic current through ultra-small nanostructures, a widely used scheme is to solve the coupled Schrödinger-Poisson system self-consistently, either directly [27–29] or by using non equilibrium Green's function (NEGF) approach [30–32].

Both methods essentially generate the same results. In terms of computational burden, the former approach usually requires less computer time than

the latter one, because the wavefunction is computed directly for each mode coming from the contact and the number of modes with energy below Fermi level is usually very small. However, NEGF approach is quite convenient since all the modes in the contacts are automatically taken into account in the Green's function and all the physical quantities are expressed in very compact forms.

Solving the self-consistent Schrödinger-Poisson system is computationally intensive, as it requires solving the open boundary Schrödinger equation for each energy point, each incoming mode, each iteration, and each bias. There are several efficient methods developed over the past years to reduce the complexity of the large matrix inversion, including the quantum transmitting boundary methods [27], the coupled (and uncoupled) mode space approach [28, 30], scattering matrix method [33], recursive Green's function (RGF) method [34], contact block reduction (CBR) method [35, 36], and R-matrix method [37–39]. In the mode space approaches and scattering matrix method, the wave function is expanded in a basis set which is obtained by solving an eigenvalue problem for each cross sectional layer. Due to the strong confinement, few modes are usually sufficient and the matrix dimension is drastically reduced. However, in cases where there are sharp potential or geometrical variations, a large number of modes are required which makes this kind of methods even slower than the real space approaches. In RGF approach, the Hamiltonian matrix is inverted layer by layer and the complexity is reduced from $O((N_x N)^3)$ to $O(N_x^3 N)$, where N_x is the dimension in one layer and N is the number of layers. This method is extremely useful when inelastic scattering is included. However, it is still very expensive when N_x is large. In CBR method, the matrix inversion is separated into a large eigenvalue problem for the isolated device which only needs to be solved once, and a small energy dependent part for the contact which needs to be solved repeatedly. The disadvantage is that the eigenvalue problem is very expensive, especially for devices with sharp variations of potential or boundaries, as a large number of eigenmodes needs to be solved. For the recently developed R-matrix method, the simulation domain is divided into several sub-regions, which allows us to use different basis sets for different sub-regions. This is a very powerful method and is flexible enough to treat various device imperfections.

However, solving the Schrödinger equation repeatedly at every energy

point is still time-consuming; and there is a need to find approximate solutions that can efficiently simulate the energy response over a wide band. AWE combined with CFH technique is a very popular frequency sweep method in high speed circuit analysis and computational electromagnetics [40], which has been verified to be able to reduce the computer time by over one order of magnitude. In this chapter, this technique is employed for efficient simulation and analysis of quantum electron transport in nanodevices. It will be shown that this method considerably speeds up the simulation while good accuracy can be maintained.

In Section 2.2, quantum ballistic transport equations will be reviewed first, and then the idea of AWE and CFH will be presented and incorporated into the simulation flow. In Section 2.3, the method is demonstrated by simulating several multi-gate silicon MOSFETs; the accuracy and simulation time are compared with traditional approach. In Section 2.4, silicon nanowire MOSFETs in the presence of charged impurity and surface roughness scattering are then analyzed by performing 3D real space simulations. Some conclusions are drawn at the end.

2.2 Method Description

2.2.1 Quantum Ballistic Transport Problem

A general 2D quantum device is illustrated in Fig. 2.1. The solution domain we are interested in can be divided into two parts, the device region Ω_D and the contact regions Ω_α (α=1, 2, and 3). Ω_D is a finite region with boundary denoted by Γ_D; Ω_α is a semi-infinite region with boundary denoted by Γ_α. Denote the intersection of Ω_D and Ω_α by $\Gamma_{D,\alpha}$, then the rest of Γ_D is $\Gamma_{D,0}$ and the rest of Γ_α is $\Gamma_{\alpha,0}$.

The device region is characterized by space varying potential $V_D(x,y)$ and effective mass $m_D^*(x,y)$. Since the potential and effective mass inside the contact region should be independent of the position along the contact although it may have complicated transverse structure [27], the potential $V_\alpha(\xi_\alpha,\eta_\alpha) = V_\alpha(\xi_\alpha)$ and $m_\alpha^*(\xi_\alpha,\eta_\alpha) = m_\alpha^*(\xi_\alpha)$, where ξ_α and η_α are respectively the transverse position and the longitude position inside each contact. Therefore, our major problem is [27]:

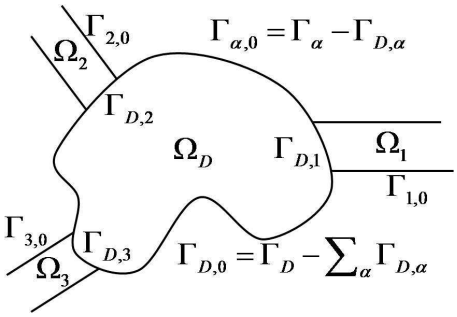


Figure 2.1: Geometry of a generalized 2-dimensional quantum device with three semi-infinite leads.

Given: (1) the potential and effective mass in every region: $V_D(x,y)$, $m_D^*(x,y)$, $V_\alpha(\xi_\alpha)$, $m_\alpha^*(\xi_\alpha)$, (2) amplitude for each wave incoming from contact α with mode n and energy E,

Find: $\psi_D^{\alpha,n}(x,y,E) \in C^2(\Omega_D)$, which satisfies the following Schrödinger equation,

$$\begin{aligned} &-\frac{\hbar^2}{2}\nabla\cdot\left[\frac{1}{m_D^*(x,y)}\nabla\right]\psi_D^{\alpha,n}(x,y)+V_D(x,y)\,\psi_D^{\alpha,n}(x,y)\\ &=E\psi_D^{\alpha,n}(x,y),\ (x,y)\in\Omega_D, \end{aligned} \tag{2.1}$$

and boundary conditions,

$$\psi_D^{\alpha,n} = \psi_\alpha \text{ on } \Gamma_{D,\alpha}, \tag{2.2}$$

$$\frac{1}{m_D^*}\nabla\psi_D^{\alpha,n}\cdot\mathbf{n}_{\Gamma_{D,\alpha}} = \frac{1}{m_\alpha^*}\nabla\psi_\alpha\cdot\mathbf{n}_{\Gamma_{D,\alpha}} \text{ on } \Gamma_{D,\alpha}, \tag{2.3}$$

$$\psi_D^{\alpha,n} = 0 \text{ on } \Gamma_{D,0}, \tag{2.4}$$

$$\psi_\alpha = 0 \text{ on } \Gamma_{\alpha,0}, \tag{2.5}$$

$$\psi_\alpha \text{ bounded as } \eta_\alpha \to \infty. \tag{2.6}$$

Once the wave function is obtained by solving (2.1)-(2.6), all the physical quantities can be obtained. For example, the electron density is given by

$$n(x,y) = 2\sum_\alpha\sum_n\int_0^{+\infty}|\psi_D^{\alpha,n}(x,y,E)|^2 f_{FD}(E-\mu^\alpha)\frac{dk}{dE}\frac{dE}{2\pi}, \tag{2.7}$$

where k is the wavenumber, f_{FD} is the Fermi-Dirac distribution function, μ_α is the Fermi level associated to contact α. Note that incoming wave of differ-

ent modes or contacts are uncorrelated, so they are calculated independently and added up [41]. The terminal current can be obtained through transmission function by Landauer-Büttiker formula (a derivation is provided in Appendix A),

$$I_\alpha = \frac{2q}{h} \sum_{\alpha \neq \alpha'} \int_0^{+\infty} T_{\alpha\alpha'}(E) \Big[f_{FD}(E - \mu^\alpha) - f_{FD}\left(E - \mu^{\alpha'}\right) \Big] dE, \tag{2.8}$$

$$T_{\alpha\alpha'}(E) = \sum_n \sum_m \frac{k_m^\alpha}{k_n^{\alpha'}} \left| \psi_D^{\alpha',n}(x, y, E)^\dagger \cdot \chi_m^\alpha(\xi_\alpha) \right|^2, \tag{2.9}$$

where $\chi_m^\alpha(\xi_\alpha)$ is the mth normalized eigenmode of the contact α, which will be defined later on.

It should be mentioned that the potential distribution for (2.1) is usually determined by a self-consistent procedure, which requires solving the Poisson equation with the charge density obtained from (2.7),

$$\nabla \cdot [\epsilon(x, y) \nabla V_D(x, y)] = q [n(x, y) - N_d(x, y)], \tag{2.10}$$

where ϵ is the dielectric constant, N_d is the doping density, q is the electron charge. The boundary conditions for Poisson equation will be specified later for the specific device.

2.2.2 Numerical Solution

The problem can be solved with different methods, like the finite element method (FEM) in Ref. [27, 31] and the finite difference method (FDM) in the following. It can also be transformed to integral equations and solved by method of moment (MoM), a formulation is provided in Appendix B.

To obtain the wave function in the device region, let us first write down the solution in the contact regions, for one incoming mode, as the summation of incident and scattered waves,

$$\psi_\alpha(\xi_\alpha, \eta_\alpha) = a_n^\alpha \chi_n^\alpha(\xi_\alpha) \exp(-ik_n^\alpha \eta_\alpha) + \sum_{m=1}^{N\alpha} b_m^\alpha \chi_m^\alpha(\xi_\alpha) \exp(ik_m^\alpha \eta_\alpha), \tag{2.11}$$

where a_n^α and b_m^α are the amplitudes of incident wave and scattered wave, respectively. Here, $\chi_m^\alpha(\xi_\alpha)$ is the mth normalized eigenmode of the contact

that satisfies following eigenvalue problem (suppose $m_\alpha^*(\xi_\alpha)$ is constant in the contacts) and boundary condition (2.5),

$$-\frac{\hbar^2}{2m_\alpha^*}\frac{\partial^2}{\partial\xi_\alpha^2}\chi_m^\alpha(\xi_\alpha)+V_\alpha(\xi_\alpha)\chi_m^\alpha(\xi_\alpha)=E_m^\alpha\chi_m^\alpha(\xi_\alpha), \tag{2.12}$$

which can be numerically solved. Here, k_m^α is the longitudinal wave number, and

$$k_m^\alpha=\sqrt{2m^*(E-E_m^\alpha)}\Big/\hbar. \tag{2.13}$$

It should be noted that k_m^α can be either real or imaginary, which corresponds to traveling wave or evanescent wave in the contact. Here, N_α should be truncated to include enough number of evanescent waves.

According to the orthogonality of the eigenmodes, b_m^α can be evaluated by

$$b_m^\alpha=\int\chi_m^\alpha(\xi_\alpha)\psi_\alpha(\xi_\alpha,\eta_\alpha=0)\,d\xi_\alpha-a_n^\alpha\delta_{mn}. \tag{2.14}$$

Substituting above expression back to (2.11) and using boundary condition (2.2) lead to

$$\begin{aligned}\psi_\alpha(\xi_\alpha,\eta_\alpha)=&-2ia_n^\alpha\chi_n^\alpha(\xi_\alpha)\sin(k_n^\alpha\eta_\alpha)\\&+\sum_{m=1}^{N_\alpha}\left(\int\chi_m^\alpha(\xi_\alpha)\psi_D^{\alpha,n}(\xi_\alpha,0)\,d\xi_\alpha\right)\chi_m^\alpha(\xi_\alpha)\exp(ik_m^\alpha\eta_\alpha).\end{aligned} \tag{2.15}$$

This is the solution in the contact region in terms of the unknowns at the interface; it is subsequently utilized to express the boundary conditions for the device region.

Next, applying FDM to discretize the 2D Schrödinger equation (2.1), we can obtain the following matrix equation (see Appendix C for details),

$$\left[E\overline{\mathbf{I}}-\overline{\mathbf{H}}-\sum_\alpha\overline{\mathbf{S}}^\alpha(E)\right]\boldsymbol{\psi}_D^{\alpha,n}(E)=\mathbf{v}_n^\alpha(E), \tag{2.16}$$

where $\overline{\mathbf{I}}$ is the identity matrix, $\overline{\mathbf{H}}$ is the isolated device Hamiltonian matrix, $\overline{\mathbf{S}}^\alpha(E)$ is the energy dependent self energy matrix that represents the boundary condition, $\mathbf{v}_n^\alpha(E)$ is the vector represents the incident wave from the contact. $\overline{\mathbf{S}}^\alpha(E)$ and $\mathbf{v}_n^\alpha(E)$ only have non-zero elements in the parts that have coupling to the contacts.

Equation (2.16) can then be solved by various matrix solvers. It must be pointed out that, for multiple incoming modes from multiple contacts which share the same energy, they have the same Hamiltonian matrix that only needs to be inverted once.

2.2.3 Asymptotic Waveform Evaluation

It is obvious from (2.7) and (2.8) that (2.16) need to be solved repeatedly within the energy range of interest so as to obtain the integral. In particular, when the wave function changes rapidly with energy, the energy grid must be very fine so as to achieve convergence. In addition, sometimes the spectrum of the electron density or the transmission coefficients over some energy range are needed to analyze the device physics and guide the design process. These can be very time consuming for large problems.

To obtain the solution of (2.16) over a wide energy band, following the steps in Ref. [40], let's rewrite (2.16) as

$$\overline{\mathbf{A}}(E)\boldsymbol{\psi}\left(E\right) = \mathbf{v}\left(E\right), \tag{2.17}$$

and expand $\boldsymbol{\psi}\left(E\right)$ in terms of Taylor series at E_0

$$\boldsymbol{\psi}\left(E\right) \approx \sum_{n=0}^{Q} \mathbf{m}_n (E - E_0)^n. \tag{2.18}$$

Similarly, $\overline{\mathbf{A}}\left(E\right)$ and $\mathbf{v}\left(E\right)$ can also be expanded in terms of Taylor series at E_0 with coefficients $\overline{\mathbf{A}}^{(n)}$ and $\mathbf{v}^{(n)}$. Matching the coefficients of equal powers on both sides of (2.17) leads to following recursive algorithm for $\mathbf{m}_n$

$$\mathbf{m}_0 = \overline{\mathbf{A}}^{-1}\left(E_0\right)\mathbf{v}\left(E_0\right), \tag{2.19}$$

$$\mathbf{m}_n = \overline{\mathbf{A}}^{-1}\left(E_0\right)\left[\frac{\mathbf{v}^{(n)}\left(E_0\right)}{n!} - \sum_{i=1}^{n}\frac{\overline{\mathbf{A}}^{(i)}\left(E_0\right)\mathbf{m}_{n-i}}{i!}\right],\ n \geq 1. \tag{2.20}$$

A wider bandwidth can be obtained by approximate $\boldsymbol{\psi}\left(E\right)$ with a rational Padé approximant of order $[L/M]$

$$\boldsymbol{\psi}\left(E\right) \approx \frac{\sum_{i=0}^{L}\mathbf{a}_i(E - E_0)^i}{\mathbf{1} + \sum_{j=1}^{M}\mathbf{b}_j(E - E_0)^j}, \tag{2.21}$$

where $L+M=Q$. The elements of unknown coefficient vectors $\mathbf{a}_i$ $(0 \leq i \leq L)$ and $\mathbf{b}_j$ $(1 \leq j \leq M)$ can be calculated by equating the right hand sides of (2.18) and (2.21), multiplying both sides with the denominator of the Padé approximant, and matching the coefficients of equal powers of $E-E_0$, which results in

$$\begin{bmatrix} m_L & m_{L-1} & m_{L-2} & \cdots & m_{L-M+1} \\ m_{L+1} & m_L & m_{L-1} & \cdots & m_{L-M+2} \\ m_{L+2} & m_{L+1} & m_L & \cdots & m_{L-M+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{L+M-1} & m_{L+M-2} & m_{L+M-3} & \cdots & m_L \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_M \end{bmatrix} = -\begin{bmatrix} m_{L+1} & m_{L+2} & m_{L+3} & \cdots & m_{L+M} \end{bmatrix}^T, \tag{2.22}$$

and

$$a_i = \sum_{j=0}^{i} b_j m_{i-j}, 0 \leq i \leq L. \tag{2.23}$$

(2.22) is first solved to obtain b_j, which is then substituted into (2.23) to calculate a_i.

Once the coefficient vectors $\mathbf{a}_i$ and $\mathbf{b}_j$ are evaluated, the wave function at any energy (within the bandwidth of accuracy) can be found by (2.21). Note that if LU decomposition of the sparse matrix $\overline{\mathbf{A}}$ is done, then (2.19) and (2.20) can be solved efficiently with forward and backward substitutions. The implementation is simple, since the derivatives of $\overline{\mathbf{A}}$ only have non-zero elements in the self-energy parts, which have very simple analytical forms for the effective mass approximation as can be derived from equations (C-5) and (2.13). Similarly, the derivatives of $\mathbf{v}$ only have non-zero elements in the layer that couples to the contact and they are also analytical.

In addition, for multiple incoming modes from multiple contacts, the asymptotic form (2.21) for each mode needs to be evaluated. Note that the LU decomposition of $\overline{\mathbf{A}}$ can be reused, and the computational cost is slightly increased since more forward and backwards substitutions are needed.

2.2.4 Complex Frequency (Energy) Hopping

Since the bandwidth of Padé approximation is limited, multiple points' expansion is necessary to obtain an accurate solution over the whole energy

band. The locations of the energy points can be selected by CFH technique as described in the following.

Given an energy band $[E_1, \mu + mk_BT]$ and a maximum error tolerance ε for the wave function $\boldsymbol{\psi}$, where E_1 is the first eigen energy of the incoming wave, μ is the Fermi level, k_B is the Boltzmann constant, T is the temperature in Kelvin, and m is a value large enough that the Fermi function is close to 0 at $\mu + mk_BT$,

1. Let $E_{min} = E_1$ and $E_{max} = \mu + mk_BT$;
2. Do AWE at E_{min} and E_{max}, obtain $\boldsymbol{\psi}_1(E)$ and $\boldsymbol{\psi}_2(E)$;
3. Calculate $\boldsymbol{\psi}_1(E_{mid})$ and $\boldsymbol{\psi}_2(E_{mid})$ respectively at middle energy $E_{mid} = (E_{min} + E_{max})/2$;
4. If $\max|\boldsymbol{\psi}_1(\mathrm{E_{mid}}) - \boldsymbol{\psi}_2(\mathrm{E_{mid}})| < \varepsilon$, stop. Otherwise, do AWE at E_{mid} and repeat the above steps for subregions $[E_{min}, E_{mid}]$ and $[E_{mid}, E_{max}]$.

However, a numerical difficulty arises in the above process because the derivatives reach singularities at subband edges $E = E_m^{\alpha}$ as can be seen from (2.13). In order to avoid the singularities, several small intervals should be skipped from the expansion region.

Suppose there are N subbands within region $[E_1, \mu + mk_BT]$, the subband edge energies of which are

$$E_1 < E_2 < \cdots < E_N < \mu + mk_BT. \tag{2.24}$$

Then we divide the whole region into several sub-regions:

$$(E_1 + \sigma, E_2 - \sigma), (E_2 + \sigma, E_3 - \sigma), \cdots, (E_N + \sigma, \mu + mk_BT), \tag{2.25}$$

where σ is a small value, e.g., 5×10^{-4}eV. In each sub-region, the CFH algorithm is employed as described above.

2.3 Numerical Examples and Discussion

The above proposed method is applied to simulate several 2D and 3D multi-gate silicon MOSFETs, as shown in Fig. 2.2 and Fig. 2.3. They are very promising candidates for the next generation nano-transistors [42].

For the Poisson equation, Dirichlet boundary condition is enforced at the gate region, while the floating boundary condition, *i.e.* the normal derivative

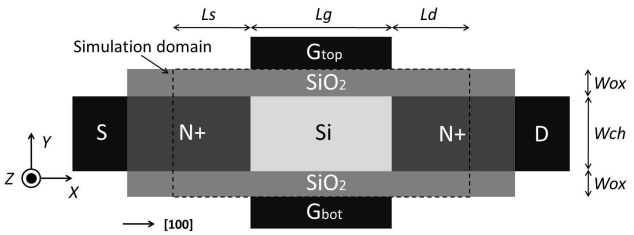


Figure 2.2: 2-D view of the N-type double gate silicon MOSFET. The structure is infinite in Z direction. Gate length is denoted by L_g, source and drain extension length is denoted by L_s and L_d, silicon channel thickness is W_{ch}, oxide thickness is W_{ox}.

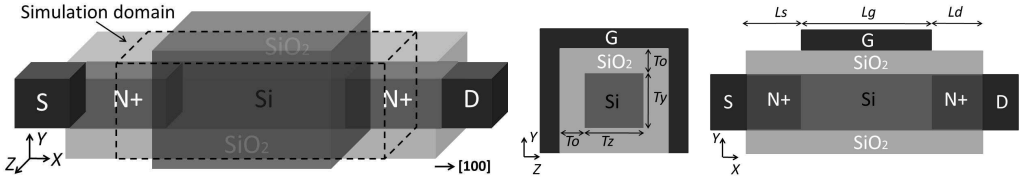


Figure 2.3: 3-D view (left), YZ cross section view (middle) and XY cross section view (right) of the N-type triple gate silicon MOSFET. Gate length is denoted by L_g, source and drain extension length is denoted by L_s and L_d, silicon channel thickness is T_x and T_y, oxide thickness is T_o.

is zero, is applied at the remaining boundary. This can be used to maintain the charge neutrality at the source and drain extensions [16]. In addition, Gummel iterative scheme is adopted to speed up the convergence of the coupled Schrödinger Poisson system [28].

2.3.1 2D Double Gate MOSFET

As shown in Fig. 2.2, the device parameters are: $L_g = 10$nm, $L_s = L_d = 4$nm, $W_{ch} = 5$nm, $W_{ox} = 1$nm, doping density $N^+ = 10^{26}/\mathrm{m}^3$, longitudinal and transverse effective mass are $m_l^* = 0.91m_e$, $m_t^* = 0.19m_e$, work function of gate metal is 4.25eV, affinity of silicon is 4.05eV, permittivity of silicon is 11.9, permittivity of SiO_2 is 3.8. Temperature $T = 300$K. The grid spacing is 0.1nm in both x and y directions. The source Fermi level is set to be 0eV. In the following simulation, Padé approximant of order [4/4] is used and the error tolerance of CFH is set to be $2 \times 10^{-2} \times \max|\boldsymbol{\chi}_{\mathrm{n}}^{\alpha}|$.

At first, the code is verified by comparing the I-V curve to that generated by NanoMOS tool [43], which is a program using mode space NEGF formal-

ism. It can be seen from Fig. 2.4 that good agreement is obtained at every bias point.

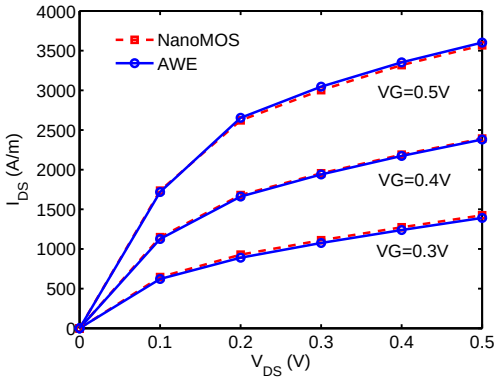
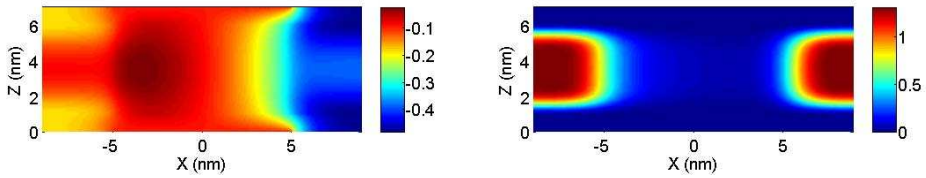


Figure 2.4: Currents of the double gated MOSFET as a function of drain bias for different gate voltages ($V_G = 0.3, 0.4, 0.5$V): comparison between results calcualted by AWE and the results of NanoMOS.



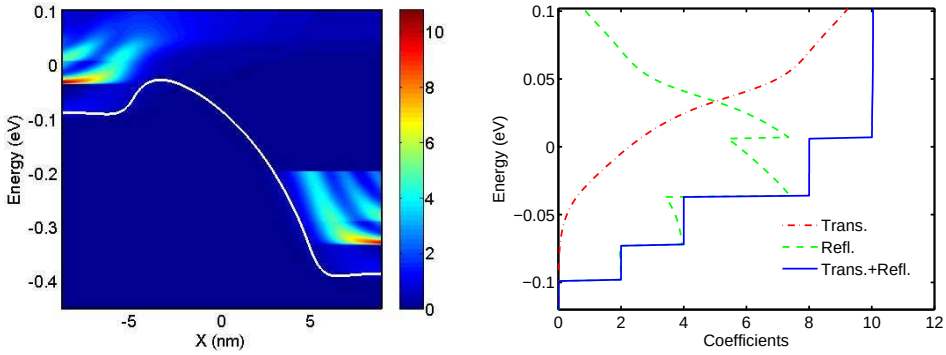
(a) (b)

Figure 2.5: 2D plot of the potential (a) and electron density (b) distributions at $V_G = V_D = 0.3$V.

The self-consistent potential and electron density distribution are plotted in Fig. 2.5. It can be seen that the electron density is very high at the source and drain extensions; and thus charge neutrality should be achieved because of the heavy positive doping density. Homogeneous Neumann boundary condition makes the potential constant along the transport direction at the source and drain ends.

The local density of states (LDOS) along the center of the silicon layer is depicted in Fig. 2.6(a). It is observed that the interference of incoming wave and reflected wave leads to standing-wave like phenomenon. In addition, some high energy electrons coming from the source side may go into the channel and eventually escape to the drain side, but the electrons coming

from the drain can hardly go to the source side due to large potential barrier. Therefore, current is formed with direction from right to left.

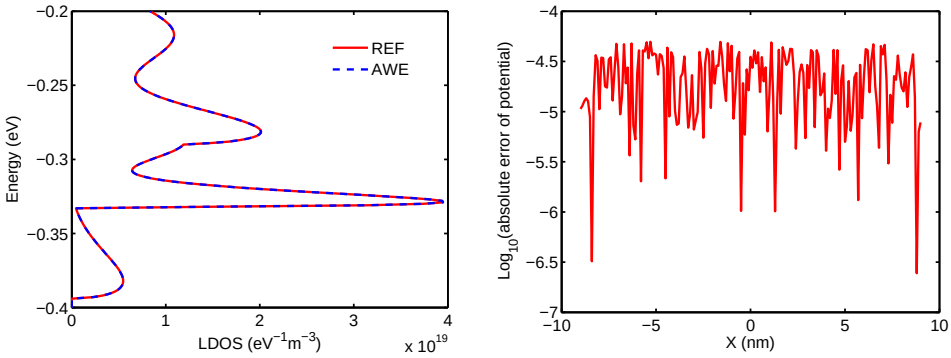


(a) (b)

Figure 2.6: (a) Local density of states along the center of the silicon layer. $V_G = V_D = 0.3$V. Conduction band edge is also shown in white line. (b) Transmission and reflection coefficients defined in Landauer formula for the electrons coming from the source. $V_G = V_D = 0.3$V.

Transmission and reflection coefficients are plotted in Fig. 2.6(b). it is noticed that the transmission coefficient is continuous whereas the reflection coefficient jumps when a new mode starts to propagate. This is because when an electron mode starts to propagate, most of it will be reflected back. However, the summation of transmission and reflection coefficients is always equal to the number of propagating modes of that energy, which is an integer. The jumps of coefficient from 0 to 2, 2 to 4, and 8 to 10 correspond to the increase of propagating modes of the electrons with heavy effective mass in the confinement direction; their valley degeneracy is 2. Conversely, the jump of coefficient from 4 to 8 corresponds to the propagating mode of electrons with light effective mass in the confinement direction; their valley degeneracy is 4.

To investigate the accuracy of the proposed method, the results of direct method (calculating the values at each energy point) with very fine energy grid (energy step: 0.001eV) is taken as the reference. Fig. 2.7(a) gives the LDOS in the middle of the drain end; it is shown that this method can produce almost the same results as the reference one. Fig. 2.7(b) plots the absolute error of the potential along the center of the silicon layer; it is shown that the error can be controlled below 10^{-4}V, which is very accurate.



(a) (b)

Figure 2.7: (a) Local density of states at the center of the drain end; both reference results and AWE results are plotted. $V_G = V_D = 0.3$V. (b) Absolute error of the potential energy along the center of silicon layer with $V_G = V_D = 0.3$V.

Table 2.1: List of The Energy Points, CPU Time, and Current (2D Case)

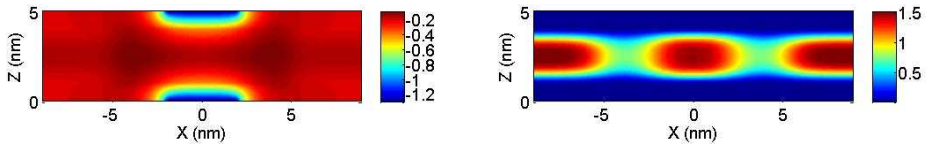
@$V_G = V_D = 0.3$V	Energy Points	Current (A/m)	CPU Time (seconds)	Speed Up
Direct method	951	1075.1	1413	1.0×
$\varepsilon = 2 \times 10^{-2}$, Order [3/3]	73	1075.0	196	7.2×
$\varepsilon = 2 \times 10^{-2}$, Order [4/4]	65	1075.1	200	7.1×
$\varepsilon = 2 \times 10^{-2}$, Order [5/5]	55	1075.0	204	6.9×
$\varepsilon = 4 \times 10^{-2}$, Order [3/3]	67	1074.9	185	7.6×
$\varepsilon = 4 \times 10^{-2}$, Order [4/4]	62	1075.1	196	7.2×
$\varepsilon = 4 \times 10^{-2}$, Order [5/5]	53	1075.0	203	7.0×

The performances of various orders of Padé approximant with different CFH tolerances are then investigated, in comparison with the direct method (energy grid: 0.001eV). The number of energy points (for AWE, it should be understood as the expansion points) of the last Poisson Schrödinger iteration, total CPU time and drain current for one bias point ($V_G = V_D = 0.3$V) are summarized in Table 2.1 (matrix solver: sparse LU decomposition with permutation matrices P and Q using UMFPACK routines). Compared with the direct method, the method reduces the inversion points by over one order of magnitude. More accurate results can be obtained by reducing the CFH tolerance but at the cost of more computer time since more expansion points

are needed. To achieve the same accuracy, higher order Padé approximant takes more computational cost although the number of expansion points decreases with increasing order. This is because higher order needs more forward and backward substitutions. When $\varepsilon = 4 \times 10^{-2}$, the drain currents obtained are still accurate enough and it is over 7 times faster.

2.3.2 2D Double Gate (Underlapped) MOSFET

To further demonstrate the advantage of AWE, a similar structure as Fig. 2.2 is analyzed, but this time the gate length is reduced to 4nm (note that the channel length is 10nm), which means it is underlapped. The other device parameters are the same as those in the last Section except that $W_{ch} = 3$nm.

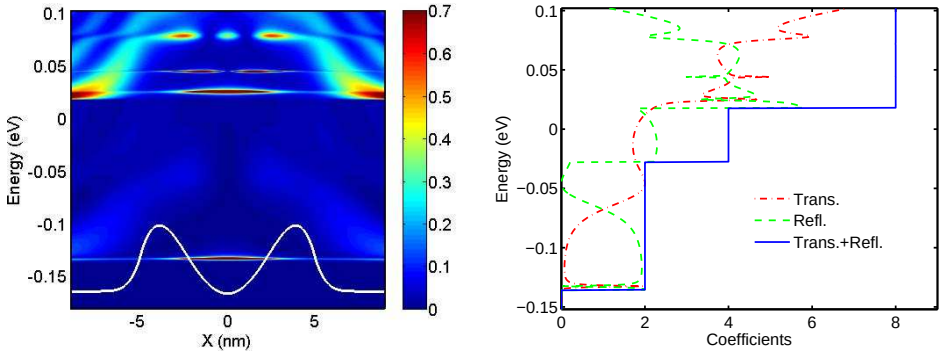


(a) (b)

Figure 2.8: 2D plot of the potential (a) and electron density (b) distributions at $V_G = 1.5$V and $V_D = 0$V.

The self-consistent potential and electron density distributions are plotted in Fig. 2.8. It can be seen that the potential in the middle part of the channel is significantly lowered by the large gate bias, whereas the potential at the end parts of the channel is less affected by the gate bias; and correspondingly, the electron density only concentrates at the middle part of the channel.

The LDOS and conduction band edge along the center of the silicon layer are further plotted in Fig. 2.9(a); it is obvious that the wave can penetrate through the potential barriers. In addition, because of the two potential barriers formed at the channel ends, there exist some resonant states inside the channel. The transmission and reflection coefficients are plotted in Fig. 2.9(b). It shows several sharp peaks corresponding to the resonant tunneling behavior. These sharp peaks are well captured by this method, which is hard to obtain by the direct method because it requires very fine energy grids.



(a) (b)

Figure 2.9: (a) Local density of states along the center of the silicon layer. $V_G = 1.5$V, $V_D = 0$V. Conduction band edge is also shown in white line. (b) Transmission and reflection coefficients defined in Landauer formula for the electrons coming from the source. $V_G = 1.5$V, $V_D = 0$V.

2.3.3 3D Triple Gate MOSFET

A triple gate silicon MOSFET is simulated in this example, as shown in Fig. 2.3. The device parameters are: $L_g = 10$nm, $L_s = L_d = 4$nm, $T_o = 1$nm, $T_y = T_z = 3$nm. Due to the small cross-section of the silicon nanowire, the effective masses are chosen as those in [44], which are extracted from $sp^3d^5s^*$ tight binding calculation of the E-k dispersion. The other parameters are the same as those in Section A. The grid spacing is 0.2nm in all x, y, and z directions.

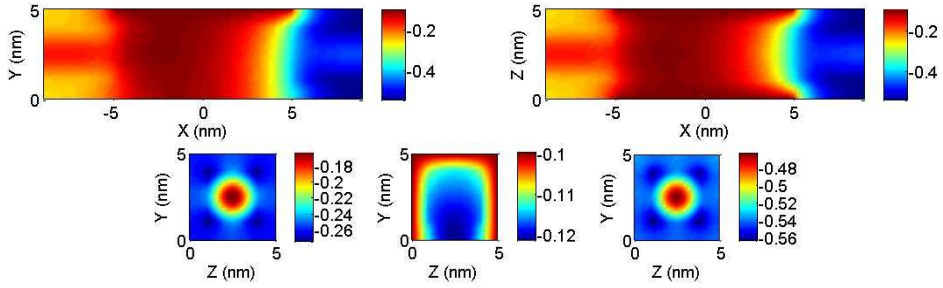


Figure 2.10: 2D plot of the potential distribution in XY plane (top left), XZ plane (top right) and YZ plane at the source end (bottom left), channel center (bottom middle) and drain end (bottom right). $V_G = V_D = 0.3$V.

The potential and electron density distributions are plotted in Fig. 2.10 and Fig. 2.11, respectively. It is shown that the potential in the XY plane is asymmetrical whereas the potential in the XZ plane is symmetrical due to

the tri-gate structure. It is also observed that the electron density is mainly confined in the center of the silicon channel due to the ultra small channel thickness.

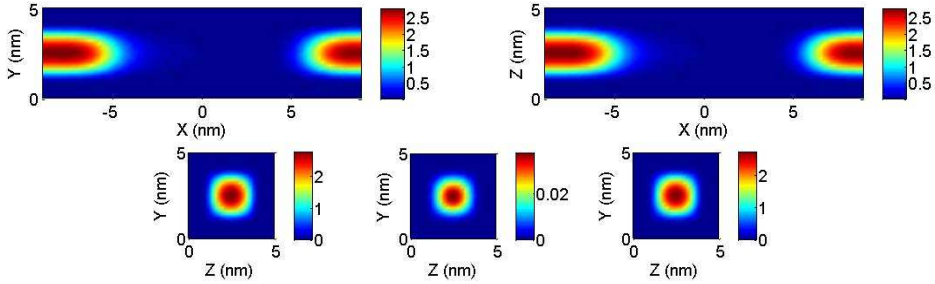
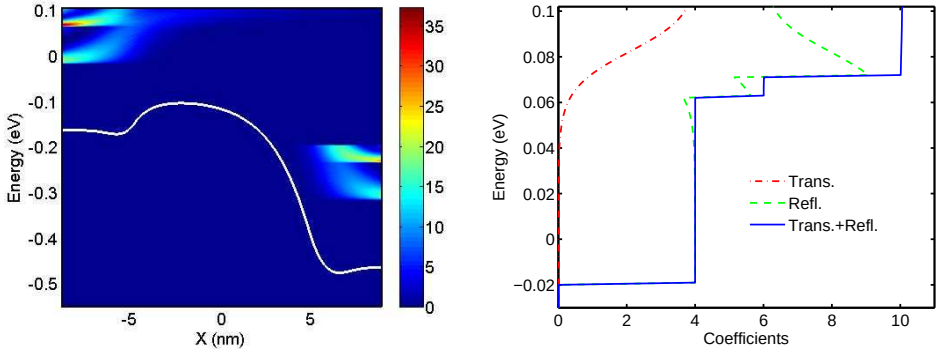


Figure 2.11: 2D plot of the electron distribution in XY plane (top left), XZ plane (top right) and YZ plane at the source end (bottom left), channel center (bottom middle) and drain end (bottom right). $V_G = V_D = 0.3$V.

The LDOS and conduction band edge along the center of the silicon layer are illustrated in Fig. 2.12(a). The wave phenomenon is evident. In addition, compared with Fig. 2.6(a), the conduction band edge in the channel part is relatively flat. This suggests that the potential in the channel is mainly modulated by the gates; and short channel effect due to drain-induced barrier lowering is effectively suppressed.



(a) (b)

Figure 2.12: (a) Local density of states along the center of the silicon layer. $V_G = V_D = 0.3$V. Conduction band edge is also shown in white line. (b) Transmission and reflection coefficients defined in Landauer formula for the electrons coming from the source. $V_G = V_D = 0.3$V.

Transmission and reflection coefficients are plotted in Fig. 2.12(b). Again, the approach produces very accurate results (note that the summation of

Table 2.2: List of The Energy Points, CPU Time, and Current (3D Case)

$@V_G = V_D = 0.3$V	Energy Points	Current (nA)	CPU Time (seconds)	Speed Up
Direct method	566	253.24	9122	1.0×
$\varepsilon = 2 \times 10^{-2}$, Order [3/3]	61	253.24	1277	7.1×
$\varepsilon = 2 \times 10^{-2}$, Order [4/4]	56	253.23	1270	7.2×
$\varepsilon = 2 \times 10^{-2}$, Order [5/5]	48	253.24	1163	7.8×
$\varepsilon = 4 \times 10^{-2}$, Order [3/3]	58	253.22	1166	7.8×
$\varepsilon = 4 \times 10^{-2}$, Order [4/4]	52	253.25	1141	8.0×
$\varepsilon = 4 \times 10^{-2}$, Order [5/5]	44	253.24	1076	8.5×

transmission and reflection is an integer) over the entire energy band interested.

The comparison of various orders of Padé approximant with different CFH tolerances for this 3D case is summarized in Table 2.2 (matrix solver is the same as the 2D case). The reference is the direct method with energy grid 0.001eV. The bias is $V_G = V_D = 0.3$V. It is observed that the number of inversion points is reduced by over one order of magnitude with our method. The accuracy is mainly determined by the CFH tolerance. More accurate results can be obtained by minimizing the CFH tolerance but at the cost of more computer time since more expansion points are needed. To achieve the same accuracy, higher order Padé approximant takes less computational cost as it requires less expansion points. When $\varepsilon = 4 \times 10^{-2}$, the drain currents obtained are still very accurate and it can be over 8 times faster.

2.4 Applications: Silicon Nanowire Transistors with Charged Impurity and Surface Roughness Scattering

Real world devices always have lots of imperfections that can significantly alter the device performances. These can be random impurities in the channel and in the source/drain extension, spatial fluctuations due to surface roughness at the silicon-oxide interfaces, and remote Coulomb scattering due

to trapped charges at silicon oxide/high-k material interfaces. Some quantum studied have been carried out based on the mode space approaches or real space approach [39, 45–48]. It is found that full 3D real space quantum simulations are needed to treat these imperfections accurately. The above developed AWE algorithm is suitable for this purpose since it is valid for arbitrary device geometry and potential profile.

2.4.1 Charged Impurity and Surface Roughness Modeling

The geometry of the silicon nanowire MOSFET considered is shown in Fig. 2.3. The parameters in the following simulation are $L_g = 10$nm, $L_s = L_d = 4$nm, $T_o = 1$nm, $T_y = T_z = 3$nm, doping density $N^+ = 10^{26}/\text{m}^3$. The order of the Padé approximant is chosen to be [5/5] and the error tolerance of CFH is set to be $2 \times 10^{-2} \times \max|\boldsymbol{\chi}_{\text{n}}^{\alpha}|$.

The charged impurity sitting in the silicon channel is modeled by a δ source in the Poisson equation. The screening effect is automatically included when self-consistency is achieved between Poisson and Schrödinger equations.

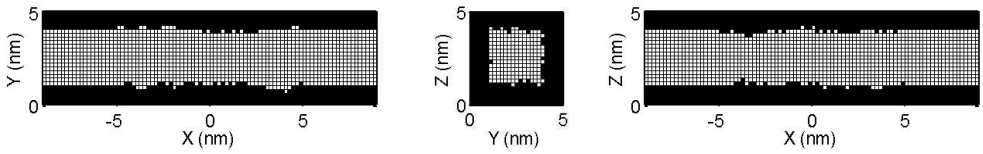


Figure 2.13: The generated random roughness pattern in one XY layer (left), YZ layer (middle), and XZ layer (right). The white color stands for silicon while the black stands for SiO_2. Only the portion in the channel is assumed to have roughness.

The roughness at the four Si/SiO_2 surfaces is generated by an exponential auto-covariance function [49],

$$C(\mathbf{r}) = \langle \Delta(\mathbf{r}') \Delta(\mathbf{r}' - \mathbf{r}) \rangle = \Delta_m^2 \cdot \exp\left(-\sqrt{2} \cdot |\mathbf{r}|/L_m\right), \tag{2.26}$$

where Δ_m is the root mean square of the fluctuation $\Delta(\mathbf{r})$, $|\mathbf{r}|$ is the distance between two positions at the Si/SiO_2 interface, and L_m is the correlation length. Typical values are used with $\Delta_m = 0.14$nm and $L_m = 0.7$nm. The generated random patterns are shown in Fig. 2.13.

2.4.2 Results and Discussion

Transistors with four situations are simulated, they are of the same dimensions but with a (a) perfect channel; (b) positively charged impurity sitting in the middle of the channel; (c) channel with rough surfaces; (d) channel with both a positively charged impurity and rough surfaces.

With bias $V_g = 0.5$V and $V_d = 0.3$V, the electron density distribution, local density of states (LDOS), and transmission characteristics are compared in Fig. 2.14, Fig. 2.15, and Fig. 2.16, respectively.

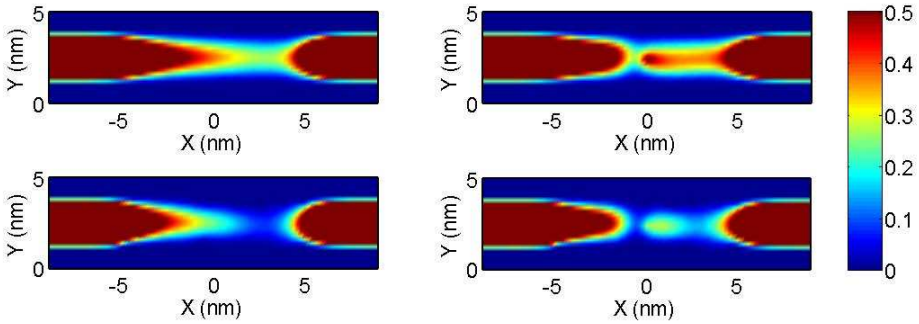


Figure 2.14: The electron density distributions in the XY plane for the four cases. Top left: (a), top right: (b), bottom left: (c), and bottom right: (d).

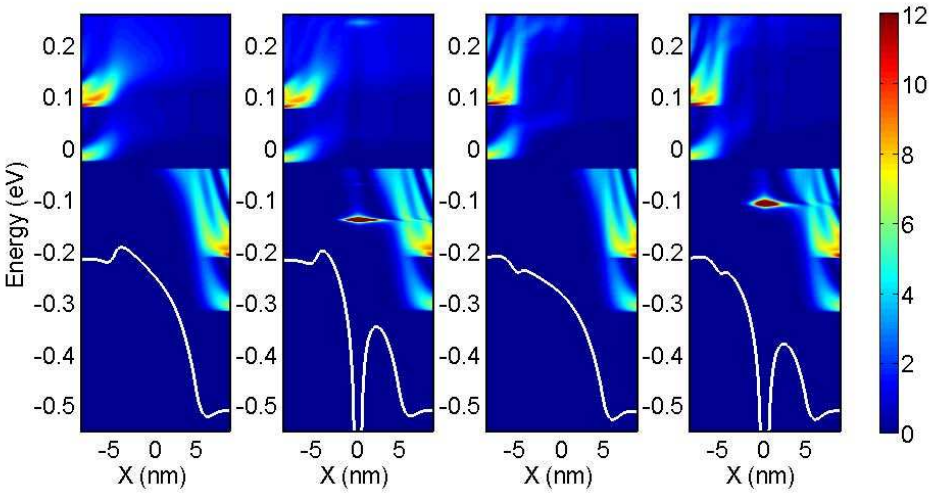


Figure 2.15: The averaged LDOS and conduction band bottom in the center of the channel (white line) for the four cases. From left to right: (a), (b), (c), and (d).

By comparing case (b) with case (a), it is observed that, the positive charge lowers the potential barrier in the channel, so that more electrons from the

source can jump to the channel and reach the drain. Also, the potential barrier becomes thinner, electrons from the source can tunnel to the drain more easily. As a result, the transmission is increased, although some weak scattering is induced at high energy. It also creates some resonant states inside the channel. In this case, there is one which can be filled by the drain lead, leading to strong distortion of the electron density distribution.

By comparing case (c) with case (a), it is seen that the surface roughness causes strong scattering, which blocks the transmission and decreases the electron density in the channel. The potential in the channel is also lowered due to the lower electron density.

Case (d) shows a mixing effect of case (b) and case (c). In addition, the position of the resonant state is shifted comparing with case (b).

As a check of the accuracy, the summation of transmission and reflection coefficient is an integer over the whole energy range for all four cases, as shown in Fig. 2.16.

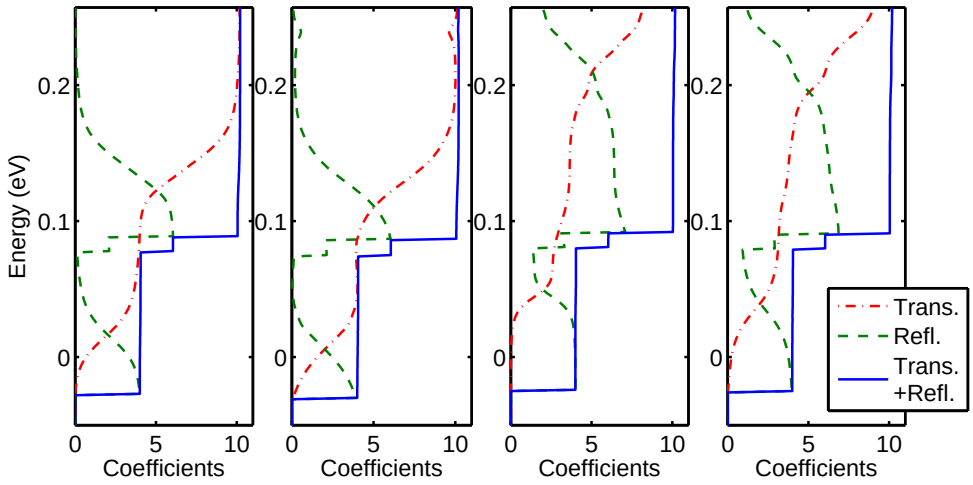


Figure 2.16: The transmission and reflection coefficients for the four cases. From left to right: (a), (b), (c), and (d).

Finally, Table 2.3 gives the currents of the four cases. It clearly indicates that surface roughness and charged impurity scattering cause drastic changes of the drain current. Such large effect is due to the small nanowire used, where a small change of the cross section size can drastically shift the quantization levels and single charged impurity can significantly alter the potential of a large portion. Note that only one configuration is analyzed

Table 2.3: Comparison of the Currents for the Four Cases

@Vg=0.5V and Vd=0.3V	Current	Changes
(a) Perfect Channel	3.75μA	NA
(b) Charged Impurity	5.38μA	43.5% $\Uparrow$
(c) Rough Surfaces	1.06μA	71.7% $\Downarrow$
(d) Charged Impurity & Rough Surfaces	1.98μA	47.2% $\Downarrow$

here for demonstration purpose. To get quantitative prediction of the device variabilities, it is required to perform statistical analysis by averaging the results of many configurations.

2.5 Summary

In this chapter, quantum ballistic transport of multi-terminal devices has been modeled by the self-consistent Schrödinger Poisson system. The AWE integrated with CFH technique is proposed to accelerate the solution of Schrödinger equation in a wide energy range. Numerical results show that this method can reduce by over 8 times the computer time; and the precision can be controlled to an acceptable level. The characteristic parameters of the device, such as the spectral density, LDOS and transmission (reflection) coefficients at any energy are readily accessed by this method.

With this method, full 3D real space simulations have been performed for silicon nanowire transistors in the presence of surface roughness and charged impurity scattering. The results suggest that these scattering events should be seriously taken into account considering the device variability.

As a general method for wide band simulation, this algorithm can be incorporated to FEM [27, 31] in the similar way. In addition, it can be combined with coupled (uncoupled) mode space approach [28, 30] to further improve the efficiency. To extend this work to include inelastic scattering, however, is non-trivial, since a full NEGF simulation is required. The implementation of AWE to the Green's function is obviously more expensive and the scattering self energy is not analytical in most cases. Extension to tight-binding model and first principle model can be realized if the derivatives of the contact self energy can be obtained.

CHAPTER 3

MODEL ORDER REDUCTION FOR MULTI-BAND SIMULATION OF NANOWIRE DEVICES

In this chapter, an efficient method is developed for multi-band simulation of quantum transport in nanowire electronic devices within non-equilibrium Green's function formalism. The efficiency relies on a model order reduction technique, which projects the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian into a much smaller subspace constructed by sampling the Bloch modes of each cross-section layer. Several sampling approaches are discussed to obtain a minimum and accurate basis with reduced computational overhead. The technique is verified by calculating the valence bands of silicon nanowires (SiNWs) and by solving I-V curves of p-type SiNW transistors. It is then applied to study for the first time the performances of large cross-section p-type junctionless (JL) transistors in the quantum ballistic transport limit. The influences of doping density, transport direction, channel length, and cross-section size are examined. It is found that, larger doping densities may lead to worse sub-threshold slopes due to the enhanced source-to-drain tunneling. Compared with their counterparts, i.e., classical inversion-mode (IM) transistors, they have better sub-threshold behaviors, but they do not necessarily provide better "on/off" ratio except when the channel is short or thin. In addition, unlike IM transistors, [110] and [111] channel directions in JL transistors are very robust against channel thickness scaling.

3.1 Introduction

One-dimensional nanowire structures, such as carbon nanotubes (CNTs), graphene nanoribbons (GNRs), and silicon (germanium or III-V material) nanowires (SiNWs), have attracted much attention during the past two decades. Due to their excellent physical properties, they are believed to have great potential in many applications, including the building blocks of

future electronic devices.

To understand the electrical properties of nanodevices built upon these small structures, a quantum-mechanical method, non-equilibrium Green's function (NEGF) approach [41], has been widely used to simulate their carrier transport. As the computational cost of the real space (RS) NEGF approach is huge, mode space (MS) approaches have been successfully developed for simulating nanostructures with strong confinement in the lateral directions such as the nanowires mentioned above [28,30,50]. These approaches expand the device Hamiltonian in the space spanned by the eigenmodes of the cross sections. Making use of the fact that usually a few modes participate in the transport process, the dimension of the Hamiltonian matrix in the MS can be greatly reduced and thus the Green's function in the MS can be easily solved. This is true for single-band effective mass approximation, since the eigenmodes for each cross section are wave vector $\mathbf{k}$-independent. For more accurate multi-band models, such as the tight-binding and $\mathbf{k}\cdot\mathbf{p}$ models, as pointed out in [50–52], the modes are generally $\mathbf{k}$-dependent and thus the transformation from the RS to MS does not exist. It is recently shown that the $\mathbf{k}$-dependent modes also make the contact block reduction (CBR) method troublesome when it is combined with tight-binding model [53].

The exception is the tight-binding model of gate-all-around (GAA) CNT transistors, which has a rigorous MS approach [54]. However, to simulate general CNT transistors which do not possess GAA feature [55] and GNR transistors [56], some crude approximations have been made so that the MS approach can still be applied. As a result, the accuracy is compromised. To improve the accuracy, a criterion of mode selection for GNR transistors has been suggested [52], which works pretty well near the conduction band minima (and valence band maxima). Quite recently, low-dimensional equivalent transport models have been constructed for tight-binding Hamiltonians of SiNWs [57], thanks to a spurious mode elimination process. Another low rank approximation method has been tried [58], but it involves large eigenvalue problems requiring much computational cost.

A MS approach has also been proposed for multi-band $\mathbf{k}\cdot\mathbf{p}$ models [51], which demonstrates great success to simulation of p-type SiNW transistors and InAs nanowire tunneling transistors. Unfortunately, the modes adopted can only accurately expand the wave function near the Γ point. Away from the Γ point, many modes are actually needed which limits its performance.

Generally, to capture the feature of **k**-dependent modes in multi-band simulations, as is commonly done in MOR methods in electromagnetics [59] and asymptotic waveform evaluation (AWE) in the previous Chapter [60], it is better to adopt multi-point expansion. This is adopted in this chapter. This approach will be demonstrated by simulation of hole transport in p-type SiNW field effect transistors using the three-band and six-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian.

Recently, JL transistors have been proposed [61] and experimentally demonstrated [62], which show extraordinarily promising performance but with simpler fabrication. To characterize the performance and illustrate the physics, much simulations have been carried out, either semi-classically [61–68] or quantum mechanically [69–72]. For quantum mechanical study, only n-type ones with large cross sections have been carried out [70–72]. This is because the single-band effective mass model is enough for the description of conduction band and it can be done in the mode space. Since the description of the valence band requires computationally more intensive multi-band model, simulation of p-type ones has been limited to only 1.15nm diameter [69] and performances of large cross-section ones remain unexamined. This gap can be filled with the MOR technique in this chapter. As a step towards more sophisticated full NEGF simulation, coherent transport will be assumed in this work.

In Section 3.2, the multi-band model is first described and the NEGF approach is outlined, then the MOR technique is presented in detail and its accuracy is checked with caution. In Section 3.3, the method is applied to simulate p-type JL transistors with different channel materials and geometries. Various device figures of merit are extracted and compared with those of classical IM transistors. Conclusions are drawn in Section 3.4.

3.2 Method Description

3.2.1 Multiband Effective Mass Equation

According to multi-band effective mass theory [20], the wavefunction inside the nanostructures can be found by solving the following coupled differential

equation for envelop function F_m $(m = 1, 2, \cdots, N)$,

$$\sum_{n=1}^{N} \left[\overline{\mathbf{H}}_{mn}^{kp} (-i\nabla) + V(\mathbf{r}) \delta_{mn} \right] F_n(\mathbf{r}) = E F_m(\mathbf{r}) \tag{3.1}$$

where N is the number of bands considered, $V(\mathbf{r})$ is the slowly varying perturbed potential distribution, and operator $\overline{\mathbf{H}}_{mn}^{kp}(-i\nabla)$ is the element of the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian with $\mathbf{k}$ replaced by differential operator $-i\nabla$.

The six-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian can be written as (if the basis is arranged in this order, three spin up p atomic orbital-like states and three spin down ones) [73]

$$\overline{\mathbf{H}}^{kp} = \left(E_{VB,0} + \frac{\hbar^2 k^2}{2m_0} \right) \overline{\mathbf{I}} + \begin{pmatrix} \overline{\mathbf{H}}^{dkk} & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{H}}^{dkk} \end{pmatrix} + \overline{\mathbf{H}}^{so}, \tag{3.2}$$

where $E_{VB,0}$ is the valence band edge, $\overline{\mathbf{I}}$ is the identity matrix. The DKK (Dresselhaus-Kip-Kittel) Hamiltonian $\overline{\mathbf{H}}^{dkk}$ is

$$\overline{\mathbf{H}}^{dkk} = \begin{pmatrix} L_M k_x^2 + M k^2 & N k_x k_y & N k_x k_z \\ N k_x k_y & L_M k_y^2 + M k^2 & N k_y k_z \\ N k_x k_z & N k_y k_z & L_M k_z^2 + M k^2 \end{pmatrix}, \tag{3.3}$$

where $L_M = L - M$, and the parameters L, M, N are related to the effective masses which can be found in [74]. The spin-orbit interaction $\overline{\mathbf{H}}^{so}$ can be written as

$$\overline{\mathbf{H}}^{so} = \begin{pmatrix} \overline{\mathbf{A}} & \overline{\mathbf{B}} \\ -\overline{\mathbf{B}}^* & \overline{\mathbf{A}}^* \end{pmatrix}, \tag{3.4}$$

with

$$\overline{\mathbf{A}} = \frac{\Delta}{3} \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \overline{\mathbf{B}} = \frac{\Delta}{3} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & -i \\ -1 & i & 0 \end{pmatrix}, \tag{3.5}$$

where Δ is the spin-orbit splitting parameter, which can be set to zero to reduce to three-band model. A derivation of the above Hamiltonian is provided in Appendix D.

To numerically solve equation (3.1), it is needed to discretize the differential operator, which can be done using finite difference method (FDM) provided in [75]. Note that for nanowire directions other than [100], coordi-

nate transformation for (3.2) should be performed before the discretization.

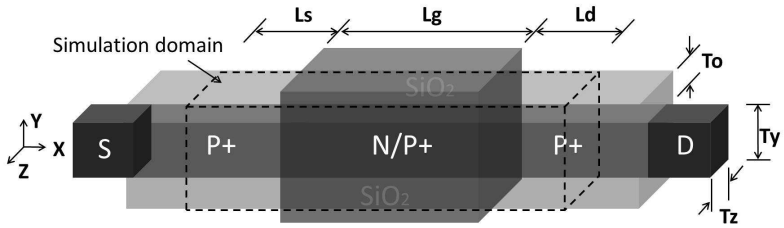


Figure 3.1: P-type GAA SiNW transistor. If the channel is N-type doping, the device is a classical IM transistor. If the channel has the same doping type and doping density with the source (or drain) extension region, it is a JL transistor.

3.2.2 NEGF Solution

For transport problems, it is required to solve (3.1) with open boundary conditions and then get the physical quantities of interest such as charge density and current. This can be nicely formulated with NEGF approach [41].

In this formalism, it is required to solve the retarded Green's function $\overline{\mathbf{G}}$ of the device region (in real space) defined as

$$\left[E\overline{\mathbf{I}} - \overline{\mathbf{H}}_0 - \overline{\mathbf{\Sigma}}(E)\right]\overline{\mathbf{G}}(E) = \overline{\mathbf{I}}, \tag{3.6}$$

where $\overline{\mathbf{H}}_0$ is the discretized $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian of the isolated device (with potential term included), and $\overline{\mathbf{\Sigma}}$ is the self-energy matrix due to the semi-infinite leads [76]. For nanowire structure like Fig. 3.1, $\overline{\mathbf{H}}_0$ can take a block tridiagonal form, the diagonal block $\overline{\mathbf{H}}_{i,i}$ (with size $N_t \times N_t$) is the on-site Hamiltonian for the ith layer; the off-diagonal block $\overline{\mathbf{H}}_{i,i\pm1}$ (with size $N_t \times N_t$) is the coupling Hamiltonian between the ith and the $(i \pm 1)$th layer. Thus, $\overline{\mathbf{H}}_0$ is of size $(N_t N_l) \times (N_t N_l)$ with N_l being the number of layers. In coherent transport limit, $\overline{\mathbf{\Sigma}}$ just has two non-zero blocks, which are the first one $\overline{\mathbf{\Sigma}}_{1,1}$ and the last one $\overline{\mathbf{\Sigma}}_{N_l,N_l}$.

The charge density $n(\mathbf{r})$ and the current $J_{i\to i+1}$ flowing between layer i and layer $i+1$, can both be expressed in terms of $\overline{\mathbf{G}}(E)$ [24, 41],

$$n(\mathbf{r}) = 2\int \frac{dE}{2\pi}\overline{\mathbf{G}}^n(\mathbf{r},\mathbf{r},E), \tag{3.7}$$

$$J_{i\to i+1} = 2\frac{ie}{\hbar}\int \frac{dE}{2\pi}\mathrm{Trace}[\overline{\mathbf{H}}_{i,i+1}\overline{\mathbf{G}}^{n}_{i+1,i}(E) - \overline{\mathbf{H}}_{i+1,i}\overline{\mathbf{G}}^{n}_{i,i+1}(E)], \tag{3.8}$$

where correlation function $\overline{\mathbf{G}}^{n} = \overline{\mathbf{G}}\overline{\mathbf{\Sigma}}^{in}\overline{\mathbf{G}}^{\dagger}$. Here, $\overline{\mathbf{\Sigma}}^{in}$ takes the same format as $\overline{\mathbf{\Sigma}}$, but with the first and last diagonal blocks replaced by $-2\Im m\left(\overline{\mathbf{\Sigma}}_{1,1}\right) f(E-\mu_L)$ and $-2\Im m\left(\overline{\mathbf{\Sigma}}_{N_l,N_l}\right) f(E-\mu_R)$, where $f(E)$ is the Fermi-Dirac distribution function, μ_L and μ_R are the chemical potentials of the left and right leads.

The problem is that solving (3.6) for realistic systems is very difficult in terms of both CPU and memory requirements. Furthermore, solving (3.6) for different E is required by (3.7) and (3.8), and it needs to recalculate (3.7) once a new potential is generated by Poisson equation until self-consistency is achieved. In spite of the popular recursive Green's function (RGF) algorithm [24], its CPU and memory cost is $O(N_t^3 N_l)$ and $O(N_t^2 N_l)$, respectively, and therefore is only feasible for small cross-section size.

3.2.3 Model Order Reduction

Similar to the mode space approach, the first step is to construct a unitary transformation matrix $\overline{\mathbf{U}}$ of size $(N_t N_l)\times(N_m N_l)$ in the following format,

$$\overline{\mathbf{U}} = \mathrm{diag}\left(\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2, \cdots, \overline{\mathbf{V}}_i, \cdots, \overline{\mathbf{V}}_{N_l}\right), \tag{3.9}$$

where $\overline{\mathbf{V}}_i$ $(i = 1, 2, \cdots, N_l)$ is a $N_t \times N_m$ $(N_m < N_t)$ sub-matrix that contains reduced basis for layer i.

Then, equation (3.6) can be transformed into this reduced basis,

$$\left[E\widetilde{\overline{\mathbf{I}}} - \widetilde{\overline{\mathbf{H}}}_0 - \widetilde{\overline{\mathbf{\Sigma}}}(E)\right]\widetilde{\overline{\mathbf{G}}}(E) = \widetilde{\overline{\mathbf{I}}}, \tag{3.10}$$

where,

$$\widetilde{\overline{\mathbf{H}}}_0 = \overline{\mathbf{U}}^{\dagger}\overline{\mathbf{H}}_0\overline{\mathbf{U}}, \quad \widetilde{\overline{\mathbf{\Sigma}}}(E) = \overline{\mathbf{U}}^{\dagger}\overline{\mathbf{\Sigma}}(E)\overline{\mathbf{U}}, \quad \widetilde{\overline{\mathbf{G}}}(E) = \overline{\mathbf{U}}^{\dagger}\overline{\mathbf{G}}(E)\overline{\mathbf{U}}. \tag{3.11}$$

Note that $\widetilde{\overline{\mathbf{H}}}_0$ is still block tridiagonal and $\widetilde{\overline{\mathbf{\Sigma}}}(E)$ can be directly calculated from lead Hamiltonian in the reduced space.

Solving (3.10) instead of (3.6) presents numerical advantages since the matrix involved is of reduced size $(N_m N_l)\times(N_m N_l)$, and this can be done

efficiently by the standard RGF algorithm with CPU cost $O\left(N_m^3 N_l\right)$ and memory cost $O\left(N_m^2 N_l\right)$. With $\widetilde{\overline{\mathbf{G}}}\left(E\right)$, the physical quantities in the reduced space are calculated with similar expressions as (3.7) and (3.8). After that, the physical quantities in the real space can be recovered with inverse transformation. The accuracy and efficiency of this method are very much dependent on how we the reduced basis set (3.9) is constructed.

3.2.4 Construction of the Reduced Basis

It is known that the E-k relation of layer i repeating along the transport direction x can be obtained by solving the following eigenvalue problem (EVP),

$$\left(\overline{\mathbf{H}}_{i,i} + \overline{\mathbf{H}}_{i,i+1} e^{ik_x \Delta x} + \overline{\mathbf{H}}_{i,i+1}^{\dagger} e^{-ik_x \Delta x}\right) \mathbf{\Psi}_i = E \mathbf{\Psi}_i, \tag{3.12}$$

where k_x is the wave number in the transport direction, Δx is the layer thickness, and $\mathbf{\Psi}_i$ is the eigenmode.

The criteria for constructing $\overline{\mathbf{V}}_i$ is that, while N_m is kept as small as possible, when $\overline{\mathbf{V}}_i$ is applied to (3.12), the reduced EVP should produce the original E-k relation as accurately as possible. The reduced EVP is,

$$\left(\widetilde{\overline{\mathbf{H}}}_{i,i} + \widetilde{\overline{\mathbf{H}}}_{i,i+1} e^{ik_x \Delta x} + \widetilde{\overline{\mathbf{H}}}_{i,i+1}^{\dagger} e^{-ik_x \Delta x}\right) \widetilde{\mathbf{\Psi}}_i = \widetilde{E} \widetilde{\mathbf{\Psi}}_i, \tag{3.13}$$

where,

$$\widetilde{\overline{\mathbf{H}}}_{i,i} = \overline{\mathbf{V}}_i^{\dagger} \overline{\mathbf{H}}_{i,i} \overline{\mathbf{V}}_i, \quad \widetilde{\overline{\mathbf{H}}}_{i,i+1} = \overline{\mathbf{V}}_i^{\dagger} \overline{\mathbf{H}}_{i,i+1} \overline{\mathbf{V}}_{i+1}, \quad \mathbf{\Psi}_i = \overline{\mathbf{V}}_i \widetilde{\mathbf{\Psi}}_i. \tag{3.14}$$

3.2.4.1 K space sampling

The reduced basis can be constructed by solving (3.12) for each layer i at n judiciously sampled k_x points (instead of solving (3.12) only at $k_x = 0$, as in [51]). Suppose m_j eigenmodes are obtained at $k_x = k_j$ $(1 \leq j \leq n)$ with eigenvalues inside the window $E_m \leq E \leq 0$, where E_m is the minimum energy of interest (usually several hundred meV below the top of the valence

band), from which a matrix $\overline{\mathbf{W}}_i$ is constructed,

$$\begin{aligned}\overline{\mathbf{W}}_i = \{&\boldsymbol{\Psi}_i^1(k_1), \cdots, \boldsymbol{\Psi}_i^{m_1}(k_1), \boldsymbol{\Psi}_i^1(k_2), \cdots, \boldsymbol{\Psi}_i^{m_2}(k_2), \cdots, \\ &\boldsymbol{\Psi}_i^1(k_j), \cdots, \boldsymbol{\Psi}_i^{m_j}(k_j), \cdots, \boldsymbol{\Psi}_i^1(k_n), \cdots, \boldsymbol{\Psi}_i^{m_n}(k_n)\},\end{aligned} \tag{3.15}$$

which is then $\overline{\mathbf{QR}}$ factorized. The unitary matrix $\overline{\mathbf{Q}}$ then serves as the submatrix $\overline{\mathbf{V}}_i$ in (3.9).

Solving (3.12) for several lowest eigenmodes can be done efficiently with iterative solvers since the matrix is highly sparse as a result of FDM. Moreover, it just needs to be solved at positive k_x (or negative k_x) saving half the cost. Suppose we already have $\boldsymbol{\Psi}_i$ and E as the eigenpairs of (3.12) at k_x. When spin-orbit coupling is not considered, $\overline{\mathbf{H}}_{i,i}$ and $\overline{\mathbf{H}}_{i,i+1}$ are both real matrices. In this case, it is easy to prove that $\boldsymbol{\Psi}_i' = (\boldsymbol{\Psi}_i)^*$ and E are the eigenpairs of (3.12) at $-k_x$. When spin-orbit coupling is taken into account, instead, the following transformation is performed to obtain those at $-k_x$ (which can be verified through (3.4)),

$$\boldsymbol{\Psi}_i = \begin{pmatrix} \boldsymbol{\Psi}_i \uparrow \\ \boldsymbol{\Psi}_i \downarrow \end{pmatrix} \Rightarrow \boldsymbol{\Psi}_i' = \begin{pmatrix} \boldsymbol{\Psi}_i' \uparrow \\ \boldsymbol{\Psi}_i' \downarrow \end{pmatrix}, \tag{3.16}$$

where $\boldsymbol{\Psi}_i \uparrow$ and $\boldsymbol{\Psi}_i \downarrow$ are the spin up and spin down components, respectively; $\boldsymbol{\Psi}_i' \uparrow = (\boldsymbol{\Psi}_i \downarrow)^*$ and $\boldsymbol{\Psi}_i' \downarrow = -(\boldsymbol{\Psi}_i \uparrow)^*$.

3.2.4.2 E space sampling

Alternatively, to obtain the eigenmode $\boldsymbol{\Psi}_i$ for each layer i, one can solve the following generalized eigenvalue problem (GEVP) (see Chapter 5) at n judiciously sampled E points,

$$\begin{pmatrix} \overline{\mathbf{0}} & \overline{\mathbf{I}} \\ \overline{\mathbf{H}}_{i,i+1}^\dagger & \overline{\mathbf{H}}_{i,i} - E\overline{\mathbf{I}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}_i \\ \boldsymbol{\Psi}_{i+1} \end{pmatrix} = \lambda \begin{pmatrix} \overline{\mathbf{I}} & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & -\overline{\mathbf{H}}_{i,i+1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}_i \\ \boldsymbol{\Psi}_{i+1} \end{pmatrix}, \tag{3.17}$$

where $\lambda = e^{ik_x \Delta x}$. It is well known that the eigenpairs with $|\lambda| = 1$ correspond to the propagating modes; whereas the eigenpairs with $|\lambda| < 1$ ($|\lambda| > 1$) correspond to the decaying (growing) modes. Here, only the propagating modes are used to construct $\overline{\mathbf{W}}_i$ as in (3.15), which is then orthonormalized to form $\overline{\mathbf{V}}_i$.

To selectively solve the eigenpairs with $|\lambda| = 1$, one can adopt the Krylov subspace method with some shift-and-invert strategies [77]. As these eigenvalues distribute in a circle in the complex plane and for low energy range they tend to cluster around 1 (since k_x is small), the shift $\sigma = e^{i\theta}$ with $\theta = 0$ is chosen. For the two cases mentioned before, the cost can be further reduced by choosing $\theta = \hat{\theta}$ (where $\hat{\theta}$ is a value slightly larger than 0) and solving the eigenmodes having eigenvalues in the upper half plane; then the corresponding transformation of these eigenmodes are those in the lower half plane.

3.2.4.3 Hybrid sampling

Both of the above sampling schemes work well but each has its advantages and disadvantages. Sampling in the k space is fast since we only need to solve EVP (3.12) several times, the drawback is that it is not easy to determine the sampling points. While sampling in the E space has the advantage that it is easy to establish the energy window to sample, it is slow as it is required to solve interior GEVPs (3.17) with matrix dimension twice the layer size. Therefore, it is better to hybridize the above two methods to construct $\overline{\mathbf{V}}_i$.

It turns out that, as will be shown later, sampling at one particular k_x point ($k_x = 0$) and at one E point ($E = E_m$) can approximate very well the E-k dispersion for energy range $E_m \leq E \leq 0$. The reason is that the modes at $k_x = 0$ can well produce the band structure near the Brillouin zone center, while the modes at $E = E_m$ are excellent for correcting the band structure far away from the center. In this simple scheme, for each layer it only needs to solve (3.12) once and solve (3.17) once. In addition, the only parameter needs to obtain is E_m. It is expected that a lower E_m may improve the accuracy since a wider energy range is approximated, but it also slows down the simulation as N_m becomes larger.

3.2.5 Validation of the Method

To validate this MOR method, which is essentially a new method of constructing the reduced basis, let's follow two steps: First, let's check if the reduced EVP can accurately capture the E-k diagram. SiNWs with cross-section size 5nm $\times$ 5nm in the [100], [110], and [111] directions are examined.

Six-band $\mathbf{k}\cdot\mathbf{p}$ model is used and the parameters are chosen to tuned to tight-binding model provided in [78]. $E_m = E_t - 0.3\text{eV}$ is chosen, where E_t is the top of the valence band, which results in N_m equal to 96, 102, and 126 respectively. Here, $\hat{\theta} = \pi/12$ is used as the eigenvalues fall within a very narrow region with $-\pi/6 < \theta < \pi/6$. As can be seen from Fig. 3.2, the E-k diagrams obtained by this MOR method are very close to the exact solutions in all three cases. The reduction is tremendous, compared with the original matrix size $N_t = 9126$ as a result of 0.125nm mesh size.

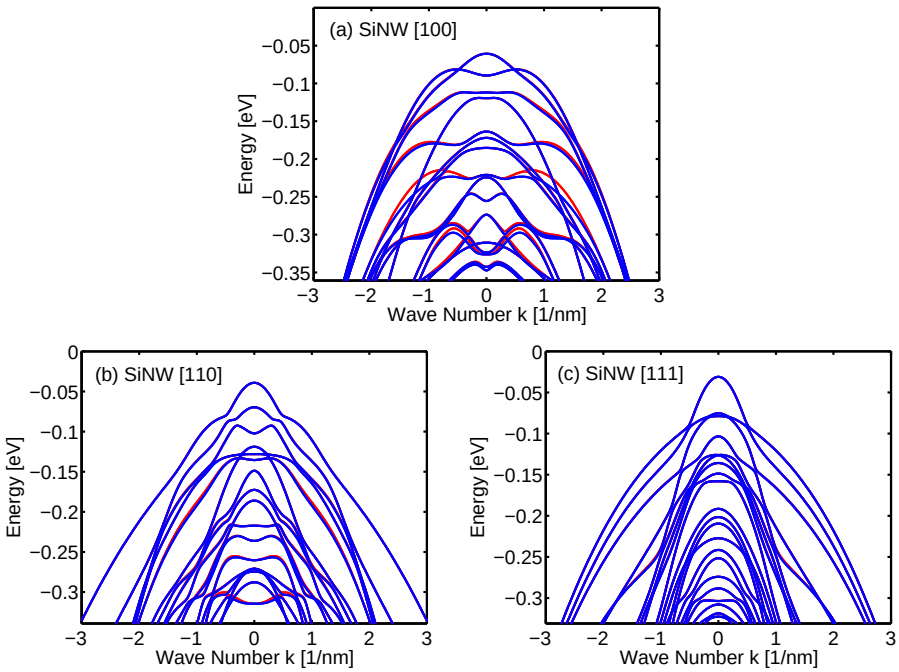


Figure 3.2: E-k relations for 5nm × 5nm SiNWs in the [100], [110], and [111] directions. The potential is assumed to be zero everywhere inside the nanowire. Red lines: exact solution, blue lines: MOR solution.

Then, one may wonder if the NEGF results are also correctly produced. A p-type IM SiNW transistor as shown in Fig. 3.1 is simulated, with NEGF and Poisson equations solved self-consistently. Again, $E_m = E_t - 0.3\text{eV}$ is chosen and the drain currents are compared to those obtained by setting $E_m = E_t - 0.5\text{eV}$, which can be regarded as a more accurate solution. Three-band $\mathbf{k}\cdot\mathbf{p}$ model is used as spin-orbit coupling plays negligible role [78], which leads to N_m values that are roughly half of the values in six-band model. For simplicity, hard-wall boundary condition at the silicon-oxide interface is implemented, which is valid for large cross-section nanowires. The

Poisson equation is solved with the same method and boundary conditions as in Chapter 2. The integration (3.7) is done by using adaptive Simpson's method. As can be seen in Fig. 3.3(a), the two solutions almost overlap with each other. The relative errors of the two solutions are further manifested in Fig. 3.3(b), which shows that errors of [110] and [111] directions for the whole bias range are within 0.5%, whereas the errors of [100] direction below threshold are larger, but still within 2.5%. Therefore, to be consistent with the E-k calculation, $E_m = E_t - 0.3\text{eV}$ is justified. The relatively larger errors of [100] direction below threshold are due to the larger errors of the band structure approximation, as can be observed from Fig. 3.2(a). In particular, the subband edges near the middle of the energy window are not aligned, which may induce a small threshold voltage shift. To improve the accuracy, one can sample one more energy point at $(E_t + E_m)/2$. Note that for SiNWs with large cross sections ($\geq 4\text{nm} \times 4\text{nm}$), 0.2nm mesh size has been adopted as this leads to negligible errors in the I-V curves (compared with 0.125nm mesh size).

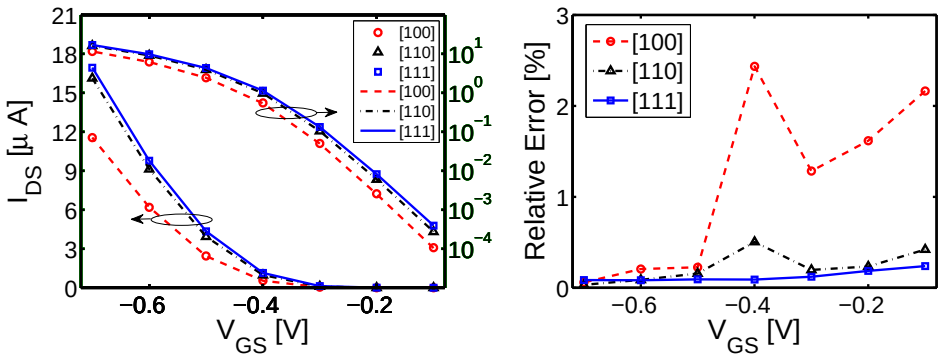


Figure 3.3: (a) I_{DS}-V_{GS} characteristics obtained by setting $E_m = E_t - 0.3\text{eV}$ (lines) and $E_m = E_t - 0.5\text{eV}$ (symbols). IM SiNW transistors in the [100], [110], and [111] directions are considered. Doping density N_d in the source and drain is $1 \times 10^{20}\text{cm}^{-3}$, while in the channel it is $1 \times 10^{15}\text{cm}^{-3}$. $T_y = T_z = 5\text{nm}$, $L_g = 10\text{nm}$, $T_o = 1\text{nm}$. Drain bias is set to be $V_{DS} = -0.5\text{V}$. (b) Relative error of the two sets of currents.

Although not shown here, the nanowires with different cross sections are also examined. It is found that for nanowires larger than $5\text{nm} \times 5\text{nm}$, choosing $E_m = E_t - 0.3\text{eV}$ is enough, while for cross sections smaller than $5\text{nm} \times 5\text{nm}$, a slightly lower E_m is recommended to ensure that enough basis functions

are included (for example, $E_m = E_t - 0.35$eV for 4nm $\times$ 4nm nanowires). It should be mentioned that, to achieve the same accuracy, N_m required in this MOR is expected to be less than that reported in [51] as the band structures here are better approximated. In addition, thanks to the sparse solvers, construction of the reduced basis and the basis transformations do not occupy much CPU time, and most of the CPU time is spent on solving the reduced transport problem. As a result, simulation of an I_{DS}-V_{GS} curve with 10 bias points for a SiNW transistor with 5nm $\times$ 5nm channel cross-section size and 30nm device length is within 10 hours using single PC (Intel i5-2400 CPU @ 3.10GHz).

3.3 Application to p-Type Junctionless Transistors

The MOR method above is applied to study hole transport in p-type JL SiNW FETs, which are then compared with similar IM ones. The impacts of various device structures and parameters on their performances will be examined. The gate oxide layer is assumed to be 1nm, the channel length is varied from 15nm to 5nm, and the channel cross-section size is varied from 6nm $\times$ 6nm to 4nm $\times$ 4nm. Channel orientations [100], [110], and [111] are considered. For IM FETs, the source and drain junctions are assumed to be abrupt, with N_d in the source/drain equal to 1×10^{20}cm^{-3} and in the channel it is 1×10^{15}cm^{-3}. The temperature is set to 300K.

To characterize the device, the device metrics such as sub-threshold slope (SS), doping density or geometry induced threshold voltage change (ΔV_{th}), and drain-induced-barrier-lowering (DIBL) will be extracted from the I-V curves. SS is expressed as millivolts of gate voltage needed for a decade change of drain current. Threshold voltage V_{th} is extracted using constant current method at 100nA. DIBL is expressed as millivolts of ΔV_{th} induced by one volt change of drain voltage.

3.3.1 The Role of Doping Densities

For JL transistors, the first thing needs to be ascertained is the doping density to use. Unlike the IM devices (which is lightly doped in the channel), the doping density actually affects the device performance greatly. In Fig. 3.4(a),

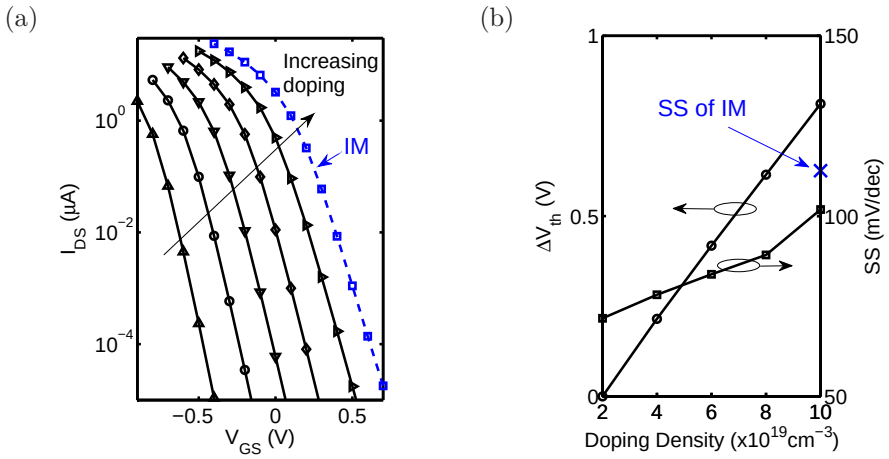


Figure 3.4: (a) I_{DS}-V_{GS} curves for JL transistors with different doping densities N_d. The curve for IM device is also plotted. The dimensions of the devices are all fixed to $T_y = T_z = L_g = 5$nm. The N_d for JL devices is linearly varied from 2×10^{19}cm^{-3} to 1×10^{20}cm^{-3}. Transport direction [100] is considered here and $V_{DS} = -0.5$V. (b) The extracted ΔV_{th} and SS with respect to doping density. Also shown is the SS for IM device.

the I-V curves of some short-channel JL devices is plotted, it is seen that SS increases as N_d increases. The SS values have been extracted and plotted in Fig. 3.4(b), which shows that for $N_d = 1 \times 10^{20}$cm^{-3}, the SS can exceed 100mV/dec. On the other hand, as seen from Fig. 3.4(a), the ON current for low N_d is very limited and it increases as N_d increases, as expected. Note that here we define the ON current as the current at flat band condition, which means no further increase of current is observed if the gate voltage is further increased. Therefore, one may suggest an N_d value based on a compromise between SS and ON current. It is also seen from Fig. 3.4(a) that, as N_d increases, the I-V curve shifts towards the positive direction, which means that more positive gate voltage is needed to turn the device off. In other words, V_{th} is shifted and ΔV_{th} is positive. As shown in Fig. 3.4(b), ΔV_{th} is almost linearly proportional to the change of N_d, say around 0.2V per 2×10^{19}cm^{-3} change of N_d, indicating that JL transistors are very sensitive to doping density variations.

To explain why SS degrades as doping density increases, Fig. 3.5 plots the potential distributions along the transport direction and their corresponding current spectra for three different doping densities. Those of IM transistor are also plotted in the same figure for comparison. The current spectrum is then roughly divided into two parts based on the peak of the potential

barrier, the part below it can be attributed to thermionic current, while the part above it is largely due to tunneling contribution. It is clearly seen that, as the doping density increases, the width of the potential barrier decreases, which results in larger source-to-drain tunneling current contribution that is known to degrade the SS. As this is due to electrostatics, similar behavior should be observed in n-type JL transistors as well. It is also obvious that the IM transistor has the thinnest potential barrier and the largest tunneling current, and thus the worst SS as shown in Fig. 3.4(b).

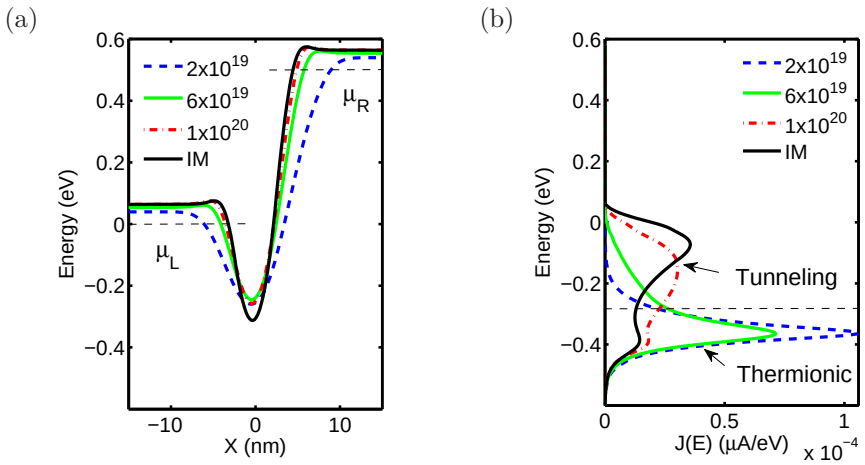


Figure 3.5: Potential distributions (a) and current spectra (b) of JL transistors with different doping density N_d. The case of IM transistor is also shown. The potentials are sampled at the center of the silicon layer. All device parameters are the same as those in Fig. 3.4, with the gate voltages of all cases tuned to achieve the same current at $I_{DS} = 1 \times 10^{-5}\mu A$.

3.3.2 Channel Orientations and Scaling

In order to study the impacts of channel orientations and channel sizes, N_d of JL transistors is fixed to be $8 \times 10^{19}\text{cm}^{-3}$. Note that different results may be obtained by choosing a different doping density, but the trends can be inferred based on the above analysis.

3.3.2.1 SS and DIBL

The SS as functions of channel aspect ratio and channel thickness are plotted in Fig. 3.6. From Fig. 3.6(a), it is observed that when channel length is

long, the SS is very close to the 60mV/dec limit. As the channel length is scaled down (with fixed channel thickness), the SS degrades as anticipated, for all nanowire directions and for both IM and JL cases. However, the JL devices degrade much slower than IM ones, which makes them very attractive for ultra-scaled applications. The excellent SSs of JL devices are very much related to the effective gate length (EGL) concept that has been used to explain n-type JL transistors [72]. In fact, the EGL of JL device in the sub-threshold range is longer than that of IM device, which results in wider potential barrier that greatly helps the suppression of source-to-drain tunneling current. The plots (for example, as shown in Fig. 3.5) confirm that this is also true for p-type devices. It is also observed that direction [100] has better SS although directions [110] and [111] do not differ too much, and this is more obvious when the channel becomes very short, indicating that [100] direction is more immune to short channel effect. The [100] direction's superior short channel performances can be explained by the fact that its larger effective mass of the first subband (as can be seen from Fig. 3.2) reduces the tunneling current contribution, as was first discovered in p-type double-gate transistors [79].

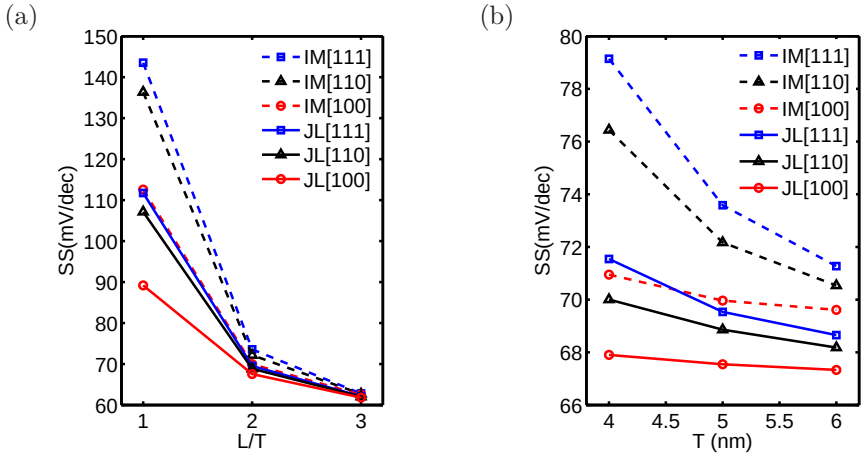


Figure 3.6: (a): SS for different channel aspect ratio, the channel thickness is fixed to T=5nm. (b): SS for different channel thickness, the channel aspect ratio is fixed to L/T=2.

From Fig. 3.6(b), it is seen that, for IM devices, the SSs of [110] and [111] directions degrade when the channel is narrowed (with fixed aspect ratio), although SS of [100] direction does not degrade much. This is because the light hole effective mass of [100] direction grows much faster as the cross section

becomes smaller, which suppresses the tunneling current and consequently compensates the otherwise faster increase of SS [78]. What should be emphasized and is observed from Fig. 3.6(b) is that, unlike IM devices, SSs of JL ones change very little when the channel thickness is reduced, regardless of channel direction. Again, this is due to the longer EGL of JL devices, which makes the tunneling current less important.

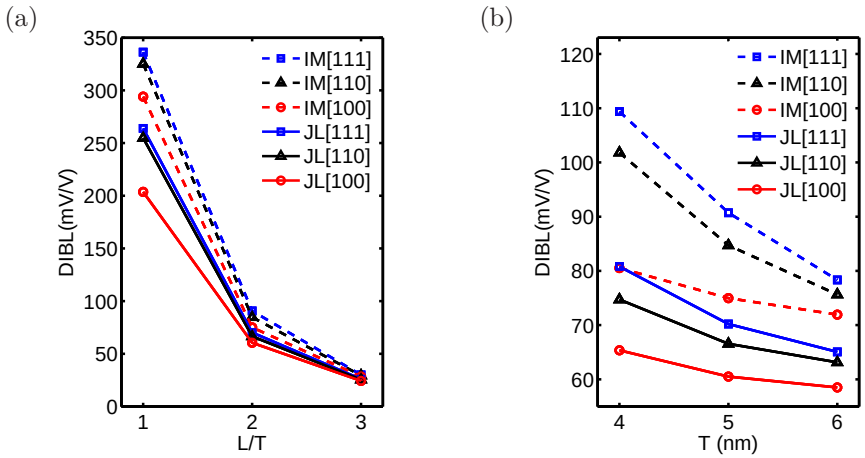


Figure 3.7: (a): DIBL for different channel aspect ratio, the channel thickness is fixed to T=5nm. (b): DIBL for different channel thickness, the channel aspect ratio is fixed to L/T=2.

The DIBL as functions of channel aspect ratio and channel thickness are plotted in Fig. 3.7. Similar trends as SS plotted in Fig. 3.6 are observed, as SS and DIBL are closely related quantities.

3.3.2.2 ΔV_{th}

The ΔV_{th} as functions of channel aspect ratio and channel thickness are plotted in Fig. 3.8. From Fig. 3.8(a), it is found that as the channel length is scaled down (with fixed channel thickness), more positive V_{th} is required to maintain the threshold current. The reason is twofold, one is that the negative drain voltage tends to raise the channel potential, the other is that tunneling current contribution becomes larger due to the narrower barrier. Direction [100] is again the most robust. In all cases, the JL devices outperform IM ones, particularly for ultra-short devices. These trends are similar to SSs in Fig. 3.6 and can be supported with similar arguments.

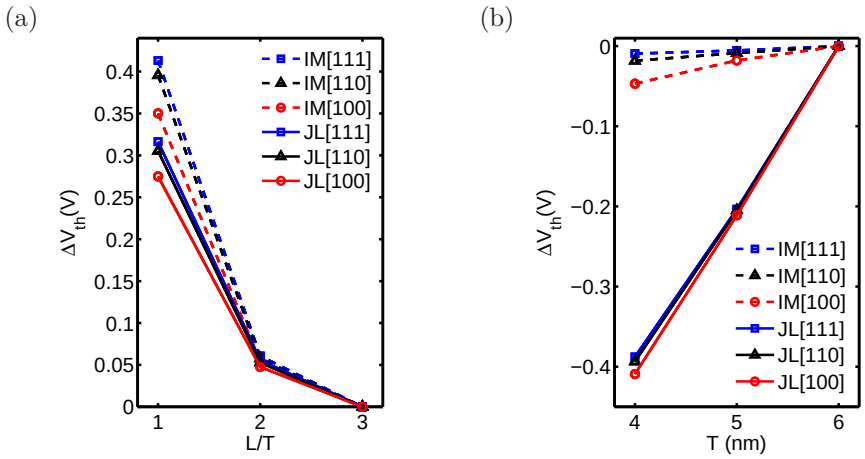


Figure 3.8: (a): ΔV_{th} with respect to channel aspect ratio, the channel thickness is fixed to T=5nm, ΔV_{th} is obtained with reference to L/T=3. (b): ΔV_{th} with respect to channel thickness, the channel aspect ratio is fixed to L/T=2, ΔV_{th} is obtained with reference to T=6nm.

From Fig. 3.8(b) it is seen that while ΔV_{th} is very small for IM devices as the channel thickness is narrowed (with fixed aspect ratio), it is significantly larger for JL ones. This is consistent with the semi-classical studies [64, 65]. Regarding channel orientations, on the contrary, [100] is more sensitive to the channel thickness scaling and [111] turns out to be the most robust for both kinds of devices, although the distinction is less pronounced for JL ones. This has been attributed to the larger subband modulation in the [100] direction [80].

3.3.2.3 I_{ON}

The I_{ON}/W (where W=4T is nanowire perimeter) as functions of channel aspect ratio and channel thickness are plotted in Fig. 3.9. The I_{ON} are obtained by setting $V_{GS} = V_{DS} = -0.5$V after adjusting the gate work functions such that the I_{OFF} are all equal to 10nA/μm [67]. From Fig. 3.9(a), it is found that JL devices have better I_{ON}/W only when the channel length is short, mainly due to their better short-channel SSs. However, when the channel length is long, the JL cases lose their advantages as the SSs become similar for both devices.

In Fig. 3.9(b), it is seen that the JL devices have better I_{ON}/W only when the channel thickness is small, especially in the [110] and [111] directions, as

a result of their better thin-channel SSs. Overall, [110] and [111] directions have similar I_{ON}/W and they are greater than those of [100] direction. The only exception is when $L/T = 1$ as shown in Fig. 3.9(a), where [100] direction provides the largest I_{ON}/W. This is due to [100] direction's excellent short channel SS mentioned before. Their performances in low standby power applications are also examined by lowering the I_{OFF}, similar trends have been observed that JL devices have better I_{ON}/W only when the channel is short or thin, although they do have more margin to perform better.

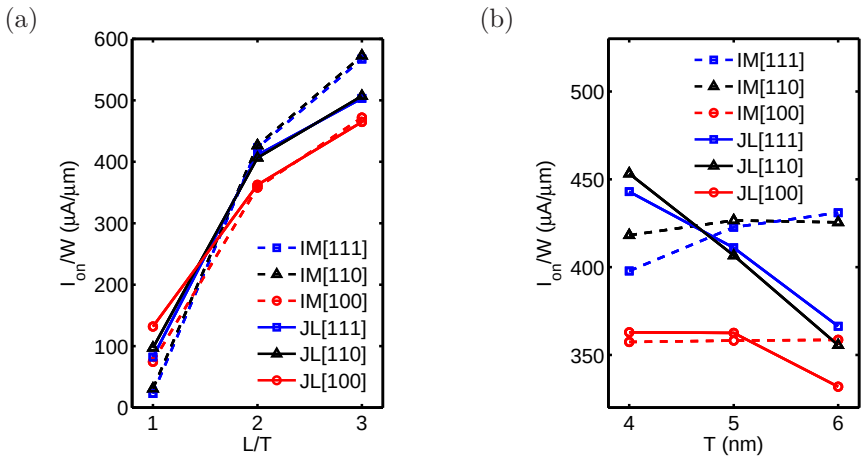


Figure 3.9: (a): I_{ON}/W for different channel aspect ratio, the channel thickness is fixed to T=5nm. (b): I_{ON}/W for different channel thickness, the channel aspect ratio is fixed to L/T=2.

3.3.3 Discussions

The code is based on continuum $\mathbf{k} \cdot \mathbf{p}$ method and the dopants are modeled through doping concentration in the Poisson equation, which is valid when the devices are large. As the devices are aggressively scaled, the atomistic effects become crucial, calling for atomistic simulator. For example, just a few discrete dopants in the JL devices can result in a large doping density and there will be a doping density limit. Besides, the positions of these dopants matter, which may induce large performance variabilities, as reported recently by studying n-type JL transistors [71]. There is also an issue related to the dopant de-activation and dielectric screening at very small nanowires [81]. Such kind of studies are out of the scope of this work and will be published elsewhere.

3.4 Summary

In summary, an MOR technique is presented for efficient simulations of nanowire transistors based on the multi-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian and NEGF method. Numerical results show that this method can correctly produce the band structures of SiNWs and I-V curves of p-type SiNW transistors, meanwhile significant reduction is achieved. With this method, the influences of various device parameters on the performances of p-type JL transistors are then studied and compared to IM devices for the first time.

The method not only applies to GAA structures, but also applies to trigate structures, such as FinFETs. Alternative channels using Germanium or III-V materials could also be simulated in this framework. Moreover, strain effects can easily be incorporated into the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian. In the next chapter, studies will be devoted to eight-band models which enable simulation of band-to-band tunneling devices.

CHAPTER 4

MODEL ORDER REDUCTION FOR SIMULATION OF BAND-TO-BAND TUNNELING DEVICES

In last chapter, a model order reduction method was developed for simulation of hole transport in silicon nanowires using three- and six-band $\mathbf{k} \cdot \mathbf{p}$ models. In this chapter, it is shown that, with a spurious band elimination process, the method can be readily extended to the eight-band case that enables us to simulate band-to-band tunneling devices. The method is demonstrated via constructing reduced models for indium arsenide (InAs) nanowires and simulation of I-V characteristics of InAs tunneling field-effect transistors (TFETs). It is shown that significant model reduction is achieved, meanwhile good accuracy can be retained. The method is then applied to investigate InAs TFETs with different channel orientations and source-pocket TFETs with n-p-i-p doping profiles.

4.1 Introduction

Band-to-band tunneling (BTBT) is a very interesting quantum phenomenon in electronic device applications. It accounts for a portion of the leakage current in the subthreshold region of carbon nanotube (CNT) field-effect transistors (FETs) [82]. It has also been utilized to build novel devices, like tunneling diodes [83] and tunneling FETs (TFETs) [84]. TFETs are energy efficient switches since their subthreshold slope can be less than 60mV/dec at room temperature [85]. This is impossible for conventional FETs which are based on thermal injection. Therefore, TFET has been selected by ITRS as a very attractive candidate device for future low-power applications [3].

Non-equilibrium Green's function (NEGF) is among the most popular approaches for quantum transport calculations [41]. Combined with tight binding [86] or eight-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian [51], which describes both the conduction and valence bands, the BTBT process can be rigorously simulated.

Unfortunately, these multi-band NEGF studies require huge computational resource. Several efforts have been devoted to improve their efficiency. In Ref. [57], equivalent but greatly reduced tight-binding models are constructed for silicon nanowires (SiNWs), which greatly speed up the simulation of p-type SiNW FETs even in the presence of inelastic scattering. For multi-band $\mathbf{k} \cdot \mathbf{p}$ models, a mode space approach is proposed to simulate p-type SiNW FETs and indium arsenide (InAs) TFETs [51]. However, it selects the modes only at the Γ point, i.e., at $k = 0$, which is insufficient since the modes are generally k-dependent.

For three- and six-band $\mathbf{k} \cdot \mathbf{p}$ models, as is shown in the last section, by sampling the Bloch modes at multiple points in the k space and (or) E space, a significantly reduced Hamiltonian can be constructed that describes very well the valence band top, based on which p-type SiNW FETs are simulated with good accuracy and efficiency [87]. The purpose here is to extend the method to eight-band $\mathbf{k} \cdot \mathbf{p}$ model to simulate BTBT devices. However, as will be shown later, the direct extension fails. The problem is that the reduced model constructed by multi-point expansion generally leads to some spurious bands, in addition to the normal bands, a situation similar to constructing the equivalent tight binding models [57], rendering the reduced model useless. Therefore, it is essential to eliminate these spurious bands, meanwhile retaining the accuracy of the normal bands.

In Section 4.2, the eight-band $\mathbf{k} \cdot \mathbf{p}$ approach will be given first. Then the scheme of model order reduction will be outlined, followed by some discussion of the discretization. A procedure to eliminate the induced spurious bands will be discussed in detail. The accuracy will be checked by comparing the band structures as well as the I-V curves. In Section 4.3, the method developed will be applied to simulate TFETs with different crystalline orientations and with source pockets. Some conclusions will be drawn in the end.

4.2 Theory and Method

The gate-all-around (GAA) InAs nanowire TFET to be simulated is illustrated in Fig. 4.1. The InAs nanowire is n-type (p-type) doped in the source and p-type (n-type) doped in the drain, while it is intrinsic in the channel. The InAs nanowire is surrounded by the oxide layer, through which the gate

controls the channel portion. The working principle of this device is based on BTBT as described in Ref. [84,85]. InAs is chosen as the channel material because high "on" current is possible due to its small direct band gap and light effective masses [86].

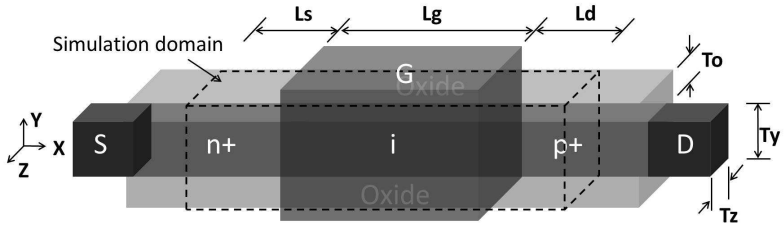


Figure 4.1: GAA InAs TFET with n-i-p (or p-i-n) type doping profile. The transport direction is x. The source, channel, and drain lengths are L_s, L_g, and L_d respectively. Nanowire thickness is T_y and T_z. Oxide layer thickness is denoted by T_o, dielectric constant is ε_{ox}.

4.2.1 Eight-Band $k \cdot p$ Approach

To describe the band structure involving both the conduction and valence bands of III-V compound semiconductor materials, a widely used approach is the eight-band $\mathbf{k} \cdot \mathbf{p}$ model. When the eight basis functions are chosen to be spin-up and spin-down s and p atomic orbital-like states, the Hamiltonian can be written as [74, 88] (for its derivation, please refer to Appendix D),

$$\overline{\mathbf{H}}^{8}(\mathbf{k}) = \begin{bmatrix} \overline{\mathbf{G}}(\mathbf{k}) & \overline{\mathbf{\Gamma}} \\ -\overline{\mathbf{\Gamma}}^{*} & \overline{\mathbf{G}}^{*}(\mathbf{k}) \end{bmatrix}. \tag{4.1}$$

The matrix $\overline{\mathbf{G}}(\mathbf{k})$ is defined as,

$$\overline{\mathbf{G}}(\mathbf{k}) = \overline{\mathbf{G}}_1(\mathbf{k}) + \overline{\mathbf{G}}_2(\mathbf{k}) + \overline{\mathbf{G}}_{so}, \tag{4.2}$$

where

$$\overline{\mathbf{G}}_1(\mathbf{k}) = \begin{pmatrix} E_g & ik_xP & ik_yP & ik_zP \\ -ik_xP & -\Delta/3 & 0 & 0 \\ -ik_yP & 0 & -\Delta/3 & 0 \\ -ik_zP & 0 & 0 & -\Delta/3 \end{pmatrix}, \tag{4.3}$$

$$\overline{\mathbf{G}}_2(\mathbf{k}) = \begin{pmatrix} Ak^2 & Bk_yk_z & Bk_xk_z & Bk_xk_y \\ Bk_yk_z & Mk^2 + L_Mk_x^2 & Nk_xk_y & Nk_xk_z \\ Bk_zk_x & Nk_xk_y & Mk^2 + L_Mk_y^2 & Nk_yk_z \\ Bk_xk_y & Nk_xk_z & Nk_yk_z & Mk^2 + L_Mk_z^2 \end{pmatrix}, \quad (4.4)$$

and

$$\overline{\mathbf{G}}_{so} = \frac{\Delta}{3} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (4.5)$$

The matrix $\overline{\mathbf{\Gamma}}$ is

$$\overline{\mathbf{\Gamma}} = \frac{\Delta}{3} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -i \\ 0 & -1 & i & 0 \end{pmatrix}. \quad (4.6)$$

Here the parameter E_g is the band gap, Δ is the spin-orbit split-off energy, P is proportional to the momentum matrix element and can be evaluated by its equivalent energy E_p. The parameter A is determined from the conduction band effective mass and B is set to be 0. $L_M = L - M$ has been used to shorten the expression. The parameters L, M, and N are related to the Luttinger parameters. For more discussion of the parameters, please refer to Ref. [74, 88].

For nanostructures like InAs nanowires that are considered here, the periodicity is broken by the finite sizes and the external potentials. The wave function can be found by solving the following coupled differential equation for envelop function F_m $(m = 1, 2, \cdots, 8)$,

$$\sum_{n=1}^{8} [\overline{\mathbf{H}}_{mn}^{8}(-i\nabla) + V(\mathbf{r})\,\delta_{mn}] F_n(\mathbf{r}) = EF_m(\mathbf{r}) \quad (4.7)$$

where $V(\mathbf{r})$ is the slowly varying perturbed potential distribution, and operator $\overline{\mathbf{H}}_{mn}^{8}(-i\nabla)$ is the element of $\overline{\mathbf{H}}^{8}(\mathbf{k})$ with $\mathbf{k}$ replaced by the differential operator $-i\nabla$.

The parameters of bulk InAs material [74] are used in this Chapter, except that the parameter E_p is reduced to 18eV according to Ref. [89]. Note that adjustment of the bulk parameters may be needed in order to match other band structure models, such as in Ref. [78].

4.2.2 Model Order Reduction

In order to solve (4.7) numerically with the NEGF approach, the operator needs to be discretized first. For reasons which will be stated later, finite difference method (FDM) is adopted in the transport direction while k-space discretization is employed in the transverse directions [51]. For simplicity, hard wall boundary condition is assumed at the interfaces between the oxide layer and the InAs channel. The resultant matrix equation for Green's function $\overline{\mathbf{G}}(E)$ can be written as,

$$\left[E\overline{\mathbf{I}} - \overline{\mathbf{H}}_0 - \overline{\mathbf{\Sigma}}(E)\right]\overline{\mathbf{G}}(E) = \overline{\mathbf{I}}, \tag{4.8}$$

where $\overline{\mathbf{H}}_0$ is the discretized $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian of the isolated device (including the potential term), and $\overline{\mathbf{\Sigma}}$ is the self-energy matrix due to the semi-infinite leads. For nanowire structures like Fig. 4.1, the Hamiltonian can be written down layer by layer and $\overline{\mathbf{H}}_0$ takes block tridiagonal form (each block is of size N_t).

As this equation can be large, to solve it efficiently for many different energy E, a reduced-order matrix equation can be constructed,

$$\left[E\widetilde{\overline{\mathbf{I}}} - \widetilde{\overline{\mathbf{H}}}_0 - \widetilde{\overline{\mathbf{\Sigma}}}(E)\right]\widetilde{\overline{\mathbf{G}}}(E) = \widetilde{\overline{\mathbf{I}}}, \tag{4.9}$$

and the reduced-order Green's function $\widetilde{\overline{\mathbf{G}}}(E)$ is to be solved. Here, the reduced Hamiltonian, self energy, and Green's function are

$$\widetilde{\overline{\mathbf{H}}}_0 = \overline{\mathbf{U}}^\dagger\overline{\mathbf{H}}_0\overline{\mathbf{U}}, \quad \widetilde{\overline{\mathbf{\Sigma}}}(E) = \overline{\mathbf{U}}^\dagger\overline{\mathbf{\Sigma}}(E)\,\overline{\mathbf{U}}, \quad \widetilde{\overline{\mathbf{G}}}(E) = \overline{\mathbf{U}}^\dagger\overline{\mathbf{G}}(E)\,\overline{\mathbf{U}}, \tag{4.10}$$

where $\overline{\mathbf{U}}$ is a block-diagonal unitary matrix containing the reduced basis $\overline{\mathbf{V}}_i$ (with dimension $N_t \times N_m$, where N_m is the number of reduced basis) for each layer i. Then the problem is on how to construct this transformation matrix $\overline{\mathbf{U}}$ so that the reduced system is as small as possible, and yet it still accurately describes the original system.

To construct the reduced basis $\overline{\mathbf{V}}_i$ for layer i, the Hamiltonian of layer i is repeated to form an infinite periodic nanowire. The reduction comes from the fact that only the electrons near the conduction band bottom and valence band top are important in the transport process. To approximate the band structure over that small region, $\overline{\mathbf{V}}_i$ then consists of the Bloch modes

with energy lying in that region. Multiple-point construction based on k space sampling and (or) E space sampling can be employed, as described in Chapter 3. Here k space sampling is adopted since E space sampling is more costly and that the eight-band matrix is much larger than the six- or three-band case. Before going to the examples, there is a need to discuss the discretization because solving the Bloch modes for each layer is itself very costly when N_t is large.

4.2.3 The Discretization

In Chapter 3, FDM is adopted and it results in extremely sparse matrices. Therefore, the Bloch modes can be obtained efficiently with sparse matrix solvers. In fact, with the shift-and-invert strategy implemented, the Krylov subspace based eigenvalue solver converges very quickly, as the interested eigenvalues (close to the valence band top) distribute in a very small area. However, it is found that the Krylov subspace method is less efficient in the eight-band case. The reason is that the interested eigenvalues distribute over a larger area, as both conduction and valence bands are of interest and between them there is a band gap.

Therefore, the method used in Ref. [51] is adopted. In that method, the transport direction is still discretized by FDM while the transverse directions are discretized by spectral method. Spectral method has high spectral accuracy (i.e., the error decreases exponentially with the increase of discretization points N) if the potential distribution is smooth [90]. This is true for devices that do not have any explicit impurities or surface roughness. So, the Hamiltonian matrix size of a layer, i.e., N_t, can be kept very small (although it is less sparse or even dense), making direct solution of the eigenvalue problem possible. The discretized form valid for arbitrary nanowire orientation is provided in Appendix E.

4.2.4 Spurious Band Elimination

As an example, Fig. 4.2(a) plots the E-k dispersion for an ideal InAs nanowire orientated in the [100] direction. Fig. 4.2(b) is the result using the reduced Hamiltonian $\widetilde{\overline{\mathbf{H}}}_0$. The reduced basis $\overline{\mathbf{V}}_i$ (i is arbitrary here) is con-

structed by sampling the Bloch modes evenly in the Brillouin zone (at $k = 0$, $\pm\pi/4$, $\pm 2\pi/4$, and $\pm 3\pi/4$ [1/nm]), with the energy $E \in [-0.5\text{eV}, 1.7\text{eV}]$, which results in $N_m = 262$ modes. Note that the modes at negative k can be obtained by a transformation of those at positive k. Clearly, the reduced Hamiltonian reproduces quite well the dispersion bands in that energy window (except at the very bottom, which can be improved by sampling a slightly larger energy window or more k points), demonstrating that the k space sampling is effective. However, there are also some spurious bands appearing in the conduction and valence bands, and even in the band gap, making the reduced model useless. This situation is not encountered in the three- or six-band model involving only the valence bands, or in the one-band effective mass model involving the conduction band only. It should be caused by the coupling between the conduction and valence bands. The coupling is important for materials with narrow band gap.

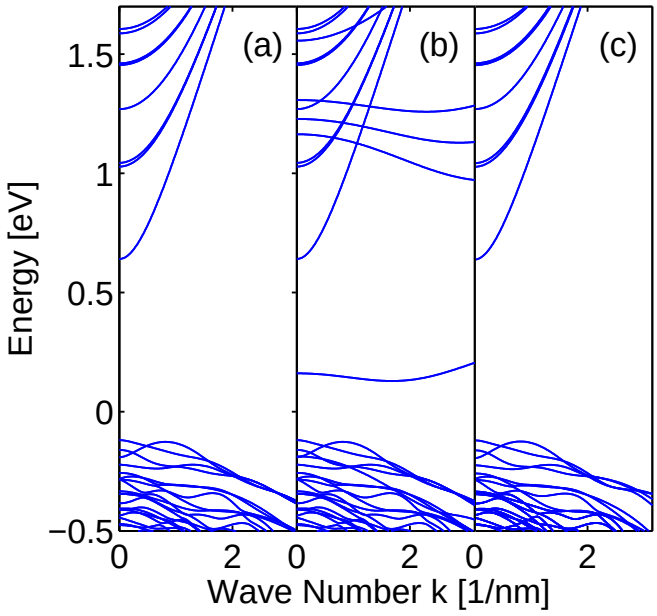
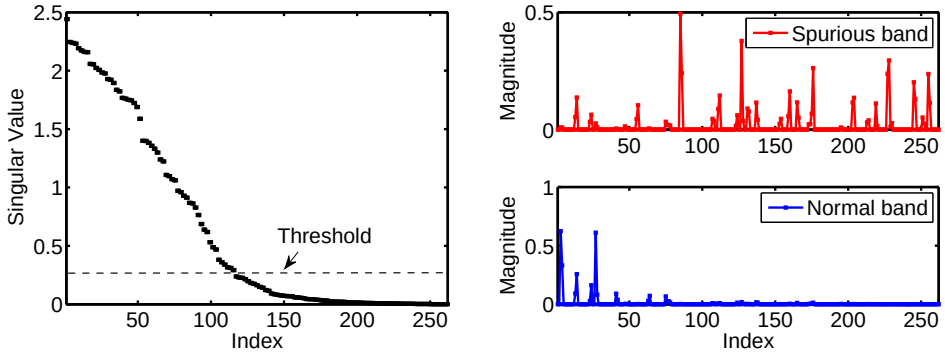


Figure 4.2: E-k diagram of a 5nm × 5nm InAs nanowire in [100] direction. The potential is assumed to be zero everywhere inside the nanowire. (a): exact solution, (b): reduced model solution, (c): reduced model solution with spurious bands eliminated. Only $+k$ is shown as the band structure is symmetric with respect to $k = 0$.

To make the reduced model useful for TFET simulation, the spurious bands must be suppressed. To this end, a singular value decomposition (SVD) is

applied to the matrix $\overline{\mathbf{V}}_i$. As plotted in Fig. 4.3(a), the singular values spread from a large value down to zero. It is further found that the normal bands are mainly contributed by singular vectors having large singular values, in contrast to the spurious bands where singular vectors with small singular values have large contribution. An example of this is shown in Fig. 4.3(b). By removing the vectors with small singular values, i.e., vectors with $v \leq v_{th}$ where $v_{th} = 0.25$ is the threshold, a new reduced basis $\widetilde{\overline{\mathbf{V}}}_i$ is generated with $\widetilde{N}_m = 116$. Using this new reduced basis, a new reduced Hamiltonian is constructed, the E-k diagram of which is given in Fig. 4.2(c). It is observed that all the spurious bands have been eliminated but at the cost of a slightly compromised accuracy.



(a) (b)

Figure 4.3: (a) Distribution of singular values of matrix $\overline{\mathbf{V}}_i$. (b) Contribution of the singular vectors to the two modes at $k = 0$: one is the spurious mode with $E = 0.16$eV, the other is the normal mode with $E = 0.64$eV (please refer to Fig. 4.2(b)).

For the BTBT process, the evanescent dispersion inside the band gap is particularly important, and it is thus plotted in Fig. 4.4. Only the smallest imaginary k is plotted, since evanescent waves decay exponentially and thus higher modes' contribution to the tunneling can be neglected. As can be seen, the MOR solution (after the spurious band elimination) agrees well with the exact solution.

The choice of v_{th} is found to be crucial. A small v_{th} might be insufficient to remove all the spurious bands while a large v_{th} may degrade the accuracy severely. Moreover, adjustment of v_{th} may be required when different sampling points or sampling energy windows are chosen. To determine the value of v_{th} automatically, we propose a search process as follows:

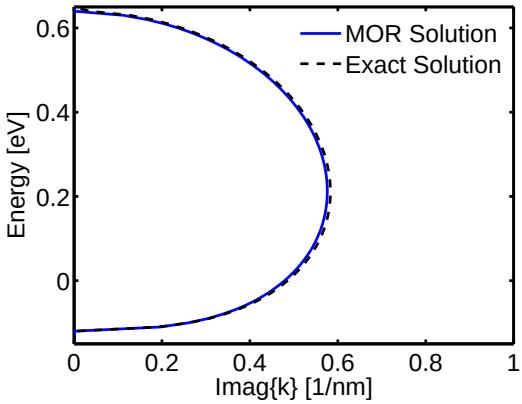


Figure 4.4: Evanescent E-k diagram in the band gap of a 5nm $\times$ 5nm InAs nanowire in the [100] direction. Only the band with the smallest imag$\{k\}$ is shown.

1. Sample enough Bloch modes and store them in matrix $\overline{\mathbf{B}}$. Suppose I points are sampled in the k space, and m_i modes with energy $E \in [E_1, E_2]$ are obtained at the ith point k_i $(1 \leq i \leq I)$, then the size of matrix $\overline{\mathbf{B}}$ is $N_t \times N_m$, where $N_m = \sum_{i=1}^{I} m_i$.

2. Do SVD of $\overline{\mathbf{B}}$, i.e., $\overline{\mathbf{B}} = \overline{\mathbf{U}}\overline{\mathbf{\Sigma}}\overline{\mathbf{V}}^{\dagger}$.

3. Set an initial value for v_{th}. Let us use $v_{th} = 0$ here.

4. Use v_{th} to construct a reduced basis $\widetilde{\overline{\mathbf{U}}}$ by removing the singular vectors with $v < v_{th}$ in $\overline{\mathbf{U}}$. The size of $\widetilde{\overline{\mathbf{U}}}$ will be $N_t \times \widetilde{N}_m$.

5. Use $\widetilde{\overline{\mathbf{U}}}$ to build a reduced Hamiltonian $\widetilde{\overline{\mathbf{H}}}$. For each layer of $\widetilde{\overline{\mathbf{H}}}$, the size will be $\widetilde{N}_m \times \widetilde{N}_m$.

6. Solve the E-k relation of $\widetilde{\overline{\mathbf{H}}}$ for certain k_i, obtaining $\widetilde{m}_i$ modes with $E \in [E_1, E_2]$. It is found that $k_i = 0$ is a good choice.

7. If $\widetilde{m}_i > m_i$ (which means that there are still some spurious bands), increase v_{th} appropriately and go back to step 4. Otherwise, stop.

The above searching process is fast, since step 5 and step 6 are much cheaper than step 1 although they have to be repeated many times. In fact, the complexity of step 1 is $I \times O\left(N_t^3\right)$, step 2 is $O\left(N_t N_m^2\right)$, step 5 is $O\left(\widetilde{N}_m N_t^2\right)$, and step 6 is $O\left(\widetilde{N}_m^3\right)$. Note that $\widetilde{N}_m < N_m < N_t$.

The $v_{th} = 0.25$ used earlier is the result of the above searching process. It has also been tested for energy windows $[-0.4\text{eV}, 1.5\text{eV}]$ and $[-0.6\text{eV}, 1.9\text{eV}]$ (see Fig. 4.5), and for [110] and [111] directions (not shown here), with good results obtained. It should be mentioned that this process results in a smaller basis set, which is different from the method for tight binding

models in Ref. [57], where the basis is enlarged to eliminate those spurious modes.

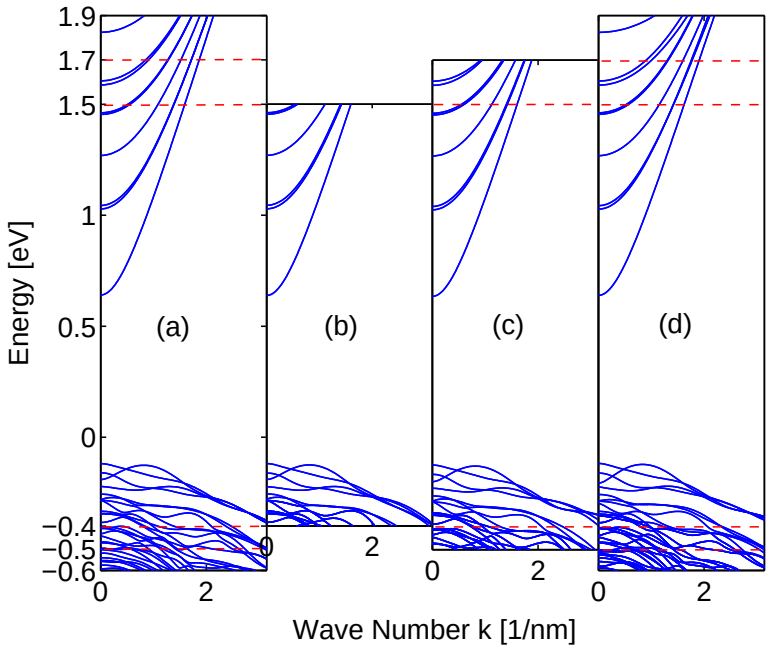
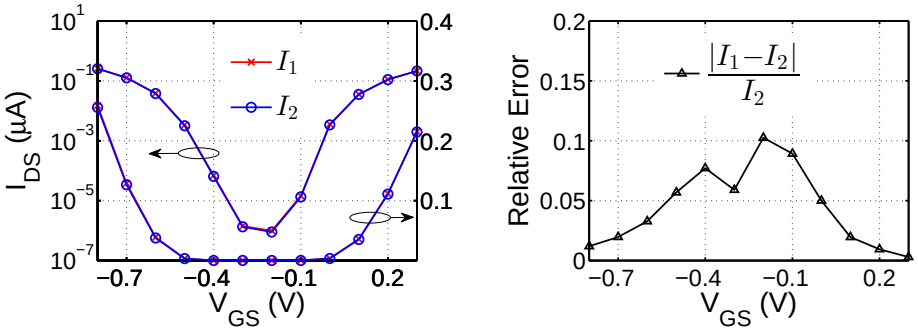


Figure 4.5: E-k diagrams of a 5nm × 5nm InAs nanowire in the [100] direction. (a): exact solution; (b), (c), and (d): reduced model solution for $E \in [-0.4\text{eV}, 1.5\text{eV}]$, $E \in [-0.5\text{eV}, 1.7\text{eV}]$ (exactly the same as Fig. 4.2(c)), and $E \in [-0.6\text{eV}, 1.9\text{eV}]$, respectively. Note that for case (d) two more sampling points ($k = \pm\pi$) are included. The reduced basis is equal to 90, 116, and 144, respectively.

4.2.5 Error Analysis

Now this reduced model can be applied to simulate a TFET as shown in Fig. 4.1. NEGF equations and Poisson equation are solved self-consistently. Phonon scattering has a negligible effect on the I-V curve [91] and thus is excluded in this work. The charge density involving both the electrons and the holes is calculated by the method in Ref. [54]. To improve the efficiency, the reduced basis is constructed only once for one layer and is assumed to be the same for each layer of the nanowire. Moreover, it remains unchanged during the self-consistent iterations. This is a fairly good approximation for GAA nanowire devices like Fig. 4.1, where the potential does not vary

drastically. This approximation has also been adopted in Ref. [57] with good accuracy demonstrated.



(a) (b)

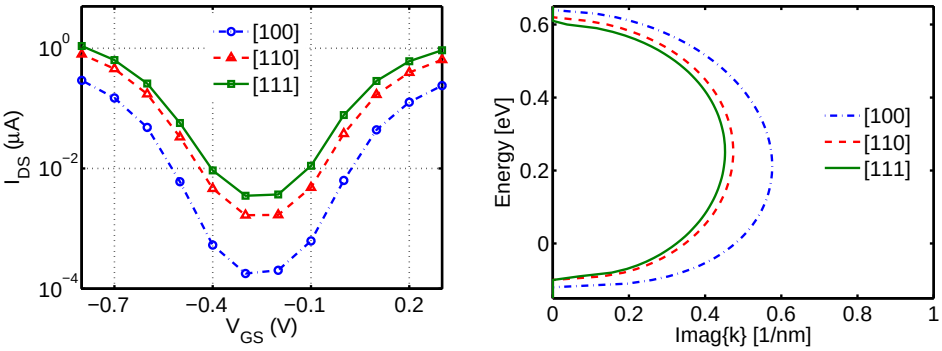
Figure 4.6: (a) I_{DS}-V_{GS} transfer characteristics of a n-i-p TFET as shown in Fig. 4.1. The nanowire is oriented in the [100] direction. T_{ox} = 1nm, $T_y = T_z$ = 5nm, $L_g = L_s = L_d$ = 15nm, ε_{ox} = 12.7. The doping density is equal to $5 \times 10^{19}\text{cm}^{-3}$ at both the source and the drain. The drain bias is fixed to $V_{DS} = -0.3$V. I_1 and I_2 are obtained by sampling energy windows [−0.4eV, 1.5eV] and [−0.5eV, 1.7eV] (as in Fig. 4.5(b) and Fig. 4.5(c)). (b) Relative errors of the two sets of currents.

The I_{DS}-V_{GS} transfer characteristics is obtained and plotted in Fig. 4.6(a), in both linear and logarithm scales. To check how large the sampling energy window is sufficient to produce the correct I-V curve, [−0.4eV, 1.5eV] and [−0.5eV, 1.7eV] are tried, which result in I_1 and I_2. It is observed that I_1 is very close to I_2, indicating that the results converge and [−0.4eV, 1.5eV] should be sufficient. The relative errors of the two sets of currents are calculated in Fig. 4.6(b). It is seen that the relative errors are very small for the region near "on" state, but are larger (up to 10%) for the subthreshold region. The reason is that, in the subthreshold region, the tunneling path is longer and thus the tunneling current is more sensitive to the (evanescent) band structures errors. In the following of this Chapter, we use energy window [−0.5eV, 1.7eV].

4.3 Applications

4.3.1 Different Channel Orientations

Fig. 4.7(a) compares the I_{DS}-V_{GS} characteristics of InAs nanowire TFETs oriented in the [100], [110], and [111] directions. It is found that [100] has the best subthreshold slope (SS) but the smallest "on" current; [111] has the worst SS but the largest "on" current. In addition, SS less than 60mV/dec is observed in Fig. 4.6(a), but it is not observed here due to a shorter channel used.



(a) (b)

Figure 4.7: (a) I_{DS}-V_{GS} curves of n-i-p InAs TFETs oriented in the [100], [110], and [111] directions. $L_g = 10$nm, the other device parameters are the same as those in Fig. 4.6. (b) Evanescent E-k relation in the band gap, only the band with the smallest imag$\{k\}$ is shown.

To explain it, Fig. 4.7(b) compares the lowest evanescent E-k relation in the band gap for the three cases. It is found that [100] has the largest $\Im m\{k\}$ leading to the smallest tunneling probability, while [110] has slightly larger $\Im m\{k\}$ than [111] leading to a modest tunneling probability. Denoting the tunneling length as L_n (L_f) for the "on" ("off") state, the "on/off" ratio can be estimated by WKB method with uniform electric field approximation as $\exp\left(k_I\left(L_f - L_n\right)\right)$, where $k_I = \Im m\{k\}$. This means that (i) large k_I has large "on/off" ratio when $L_f - L_n$ is fixed, which is the case for the [100] direction; (ii) the "on/off" ratio increases with $L_f - L_n$ and the increasing speed is larger for larger k_I, so in practice long channel (with long L_f) is employed to increase the "on/off" ratio and [100] is expected to have much better "on/off" ratio than the other directions when the channel is very long. The analysis

also suggests that the evanescent E-k relation should be accurately modeled for correctly predicting the tunneling current, in particular the "off" state current.

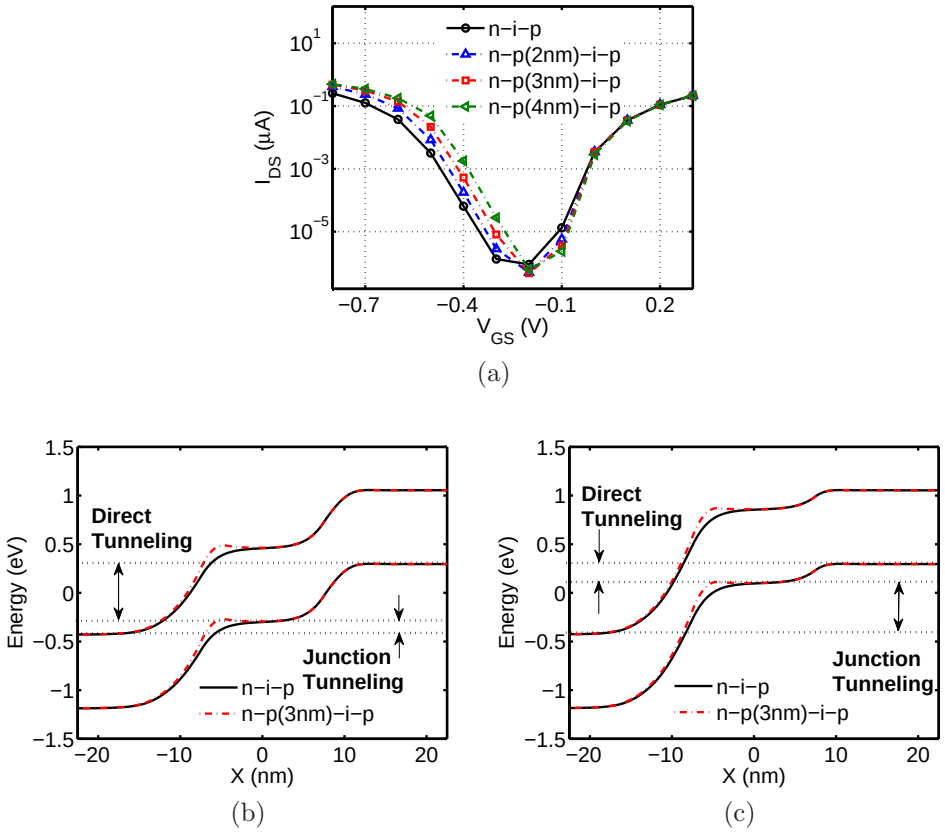


Figure 4.8: Comparison of source-pocket n-p-i-p TFETs with different pocket lengths and n-i-p one with no pocket. The device settings are the same as those in Fig. 4.6. (a) I_{DS}-V_{GS} curves. (b) Band diagrams at "off" state ($V_{GS} = -0.2$V). (c) Band diagrams at "on" state ($V_{GS} = -0.6$V). The source Fermi level is 0eV, while the drain Fermi level is -0.3eV.

4.3.2 The Source-Pocket TFETs

Many TFETs suffer from low "on" current [85]. It is theoretically predicted and experimentally demonstrated that the source-pocket TFETs have significantly improved "on" current and steeper subthreshold swing (SS) over the classical TFETs [92, 93]. In addition, the significantly degraded linear-region I_{DS}-V_{DS} characteristics of classical TFETs [94] can be improved by incorporating the source pocket [95]. The source-pocket TFET is formed by

inserting a thin layer of p-type (n-type) doping between the n-type (p-type) source and the intrinsic channel, which results in an n-p-i-p (p-n-i-n) doping profile. Most of the simulations were based on semiclassical methods [92,93]. Recently, 2-D quantum simulations were performed for all-Si and all-Ge double gate structures [96], confirming the semiclassical simulation results. Here, we investigate InAs nanowire based source-pocket TFETs for the first time, using 3-D quantum simulations.

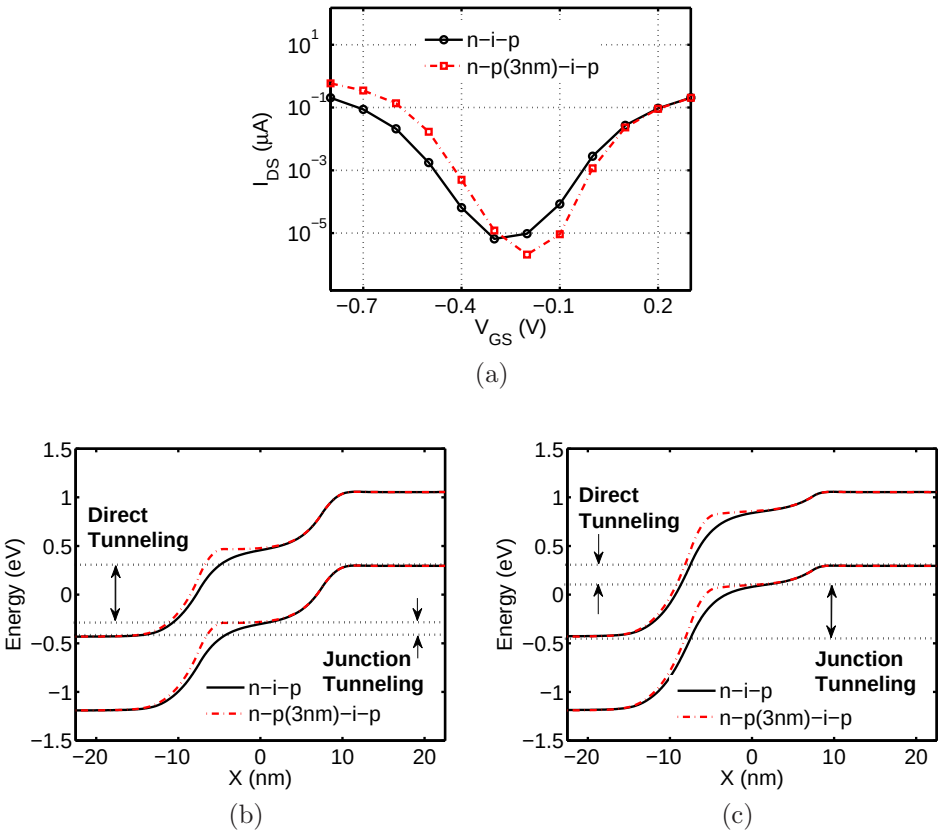


Figure 4.9: Same plots as Fig. 4.8, except that ε_{ox} is reduced from 12.7 to 3.8.

Fig. 4.8(a) shows the I_{DS}-V_{GS} of the n-p-i-p TFETs with three different pocket lengths, in comparison with the n-i-p one without source pocket. For the right part of the curve (due to the ambipolar nature of TFETs), as expected, these pockets have negligible influence on the turn-on property, since the conduction there is via tunneling through the drain junction. However, a better SS is observed due to the lower "off" current at $V_{GS} = -0.2$V. For the left part of the I-V curve, which is of interest, the pockets merely shift

the threshold voltage and less negative gate voltage can now turn the device on. However, the SS remains almost unchanged, in contrast to previous studies [92, 93, 96].

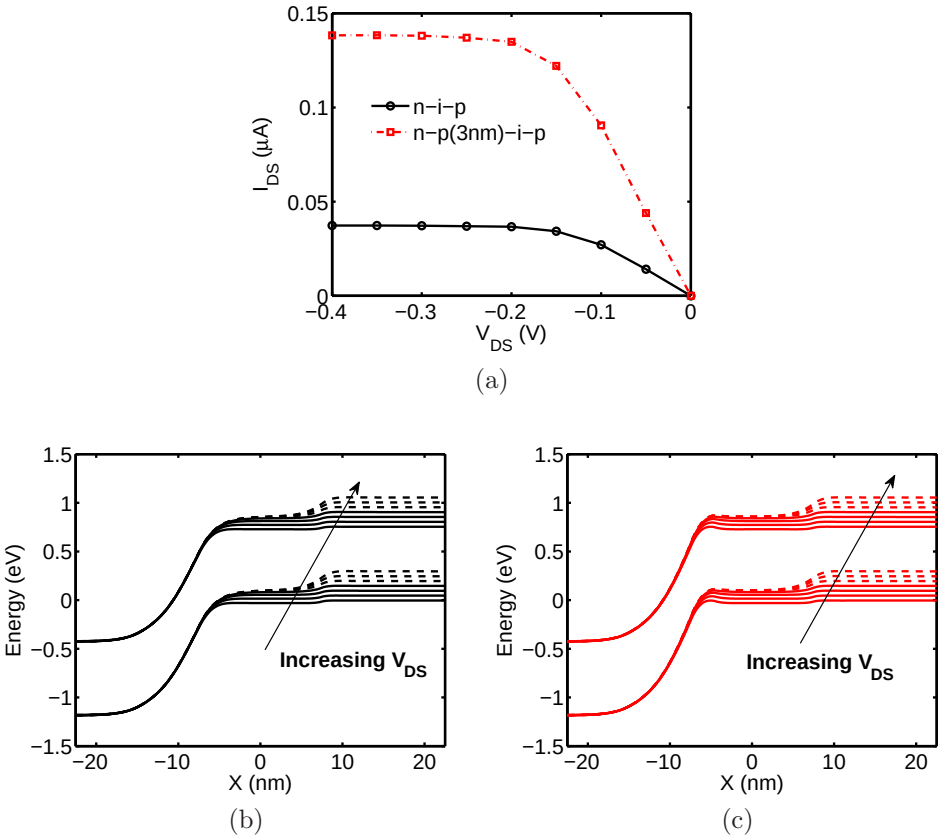


Figure 4.10: (a): I_{DS}-V_{DS} curve of source-pocket n-p-i-p TFET with 3nm pocket length, comparing with n-i-p one with no pocket. The device settings are the same as those in Fig. 4.6. (b) and (c): Band diagrams of n-i-p TFET (b) and source-pocket n-p-i-p TFET with 3nm pocket length (c), for V_{DS} varying from 0V to −0.3V with −0.05V step. The gate bias is fixed to $V_{GS} = -0.6$V. The solid lines are for linear region while the dashed lines are for saturation region.

Fig. 4.8(b) and Fig. 4.8(c) compare the band diagrams of the TFETs with and without the source pocket, at "off" and "on" state, respectively. It is seen that the source pocket enhances the band bending at the source junction, which leads to longer source-to-drain direct tunneling path but shorter source-to-channel junction tunneling path. As the "off" ("on") state current is dominated by direct (junction) tunneling, the source pocket will decrease (increase) the "off" ("on") state current. But the band bending of

the n-i-p structure at the source-channel junction is already very sharp; the insertion of source pockets does not improve it much. The reason is that, for the n-i-p structure here, the channel is fully controlled by the gate due to the small cross-section nanowire, high-k gate oxide, and GAA geometry used, making the lateral electric field between the source and the channel very strong.

Fig. 4.9(a) plots the case when the gate oxide is reduced to $\varepsilon_{ox} = 3.8$. Now the source pocket has a larger impact on the I-V curve. The band diagrams shown in Fig. 4.9(b) and Fig. 4.9(c) confirm that the band bending improvement is more significant than Fig. 4.8(b) and Fig. 4.8(c).

Fig. 4.10(a) shows the I_{DS}-V_{DS} of the n-p-i-p TFET with 3nm pocket length, in comparison with the n-i-p one. The linear dependence of I_{DS} on small V_{DS} is observed here for both cases. This is in agreement with Ref. [95] for the TFETs with source pockets, but in contrast to the exponential dependence of I_{DS} on small V_{DS} observed for the p-i-n TFETs [94].

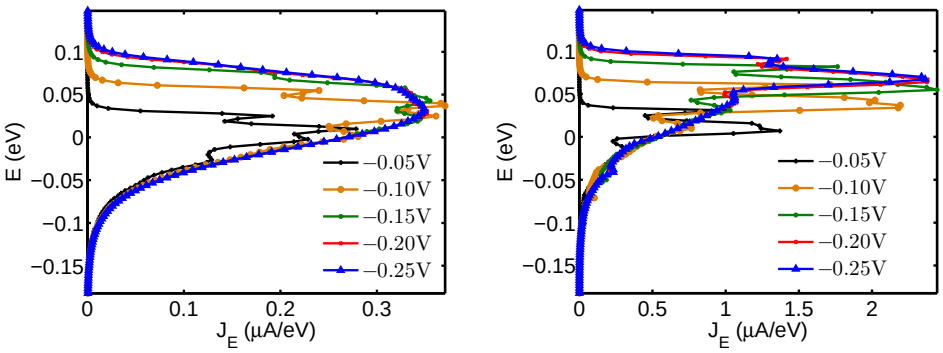


Figure 4.11: Current Spectra of n-i-p TFET (a) and n-p-i-p TFET with 3nm pocket length (b), for V_{DS} varying from -0.05V to -0.25V with -0.05V step. Please refer to Fig. 4.10 for the device setting.

To explain it, Fig. 4.10(b) and Fig. 4.10(c) compare the band diagrams by varying V_{DS}, for the TFETs with and without the source pocket, respectively. When V_{DS} is small, as can be seen, the potential in the channel changes with V_{DS}. This should modulate the source-to-channel tunneling width leading to exponential change of the tunneling probability (and current). But here, the tunneling width is almost unchanged, which is again due to the already very sharp source-channel band bending. The change of V_{DS} merely changes

the tunneling window leading to the linear behavior. When V_{DS} is large, saturation occurs since the potential in the channel (and thus the source-channel tunneling junction) no longer changes with V_{DS}. The current spectra plotted in Fig. 4.11 confirm this observation.

4.4 Summary

A MOR method is developed for efficient simulation of BTBT devices based on solving self-consistently the Poisson equation and the NEGF equations employing the eight-band $\mathbf{k} \cdot \mathbf{p}$ model. By introducing a spurious band elimination process, reduced models can be constructed for reproducing the band structures in any energy window near the band gap. The reduced models can also capture the I-V characteristics of TFETs with acceptable accuracy. InAs TFETs with different channel orientations are compared, it is found that [100] direction has better subthreshold swing but smaller "on" current than [110] and [111] directions, due to its larger imaginary wave vectors. Source-pocket TFETs with an n-p-i-p doping profile are also studied, it is observed that band bending at the source-to-channel junction is enhanced by the source pocket, which results in better "on"/"off" ratio, SS, and output behaviors. But such effects tend to be diminished when the electrostatic integrity of the device is improved. These studies are beneficial for high-performance TFET design in the future.

CHAPTER 5

FAST EVALUATION OF SELF-ENERGY MATRICES IN ATOMISTIC SIMULATIONS

Besides the heavy computational cost of obtaining the wave function or the Green's function of the device region, another major part of the computational burden is the calculation of self-energy matrices. The calculation in atomistic schemes usually requires dealing with matrices of the size of a unit cell in the leads. Since a unit cell always consists of several planes (for example, in silicon nanowire, four atomic planes for [100] crystal orientation and six for [111] and [112]), it is shown in this chapter that a condensed Hamiltonian matrix can be constructed with reduced dimension ($\sim 1/4$ of the original size for [100] and $\sim 1/6$ for [111] and [112] in the nearest neighbor interaction) and thus greatly speeding up the calculation. Examples of silicon nanowires with $sp^3d^5s^*$ basis set and the nearest neighbor interaction are given to show the accuracy and efficiency of the proposed methods.

5.1 Introduction

Non-equilibrium Green's function (NEGF) approach [41,97] has been widely adopted to simulate quantum transport in nanoscale devices. However, the large computational cost of this method limits its application to small problems. One major part of computational cost is the inversion of the large Hamiltonian matrix so as to obtain the Green's function of the device. Considerable effort has been made to reduce the complexity, such as recursive Green's function algorithm [34], mode space approaches [28,30], contact block reduction method [35,36], and the recent R-matrix method [37,38]. Another major source of the cost is the open boundary treatment, which is expressed explicitly through the self-energy matrices. In the effective mass approximation [30, 37], the self-energy matrices can be obtained for the whole energy band once the eigenmodes of the leads are solved [60]. Beyond the effective

mass approximation, such as the *ab initio* methods [26] and the empirical tight-binding approaches [38], however, the self-energy matrices must be evaluated for each energy point individually, further increasing the computational burden. The tight binding models will be the focus of this work, as they are well-suited for nanodevice modeling due to limited-range interactions and reasonably-sized basis sets [98].

Traditionally, there are roughly two kinds of approaches for self-energy evaluation [99], one is through iterative evaluation of the surface Green's function [100], the other is by solving the Bloch modes of the leads [101–103]. The underlining assumption of both approaches is that the leads are characterized by a periodic potential and thus a principle layer [104, 105] (usually a unit cell in tight-binding schemes) can be defined with translational invariance along the leads. The former approach usually requires many inversions of a Hamiltonian matrix of the size of the unit cell. The latter one, instead, requires solving a generalized eigenvalue problem (GEVP) for a matrix of the size twice that of the unit cell.

Several improvements that speed up the calculation have been developed over the past years. The widely used decimation algorithm [104] greatly improves the convergence of the iterative method by reducing the iteration steps from N to $\log(N)$. The shift-and-invert method transforms the GEVP to a normal eigenvalue problem (NEVP) [106]. The Krylov subspace method reduces the cost of the GEVP approach by computing only a portion of the eigenmodes that have contribution to the transmission [77]. However, the calculation is still very slow when the size of the unit cell matrix becomes very large. By imposing absorbing boundary conditions into the leads, the open system is transformed to a closed system and the surface Green's function (and then the self energy) can be constructed for any energy by spectral representation [107]. But this should be designed very carefully in order to eliminate possible reflections (less reflection with more absorbing layers, but with more computational cost).

However, by taking a closer look at the structure of the unit cell, it is easy to find that there are some redundancies when these traditional methods are applied to tight-binding schemes. Take silicon (or germanium) for example, the [100] crystal direction nanowire consists of four atomic planes in the unit cell and the [111] (or [112]) direction consists of six planes, as shown in Fig. 5.1. Moreover, take the nearest neighbor tight binding scheme [23]

for example, the surface Green's function of the size of the unit cell is not needed, but actually the size of an atomic plane is needed. Despite the method in [108], which transforms the GEVP to a NEVP of reduced size, it calculates the whole surface Green's function of the size of the unit cell and at the same time involves inverting a matrix of the size of the unit cell that incurs additional cost. In fact, due to the short-range interactions, it is possible to compress the Hamiltonian matrix of a unit cell to that of an atomic plane. Then, after some slight modifications, the decimation method and the eigenvalue methods can be employed to calculate the surface Green's function (and then the self energy). The gain is obvious, as we are now dealing with a much smaller matrix (approximately by a factor of 1/4 for [100] and 1/6 for [111] and [112]).

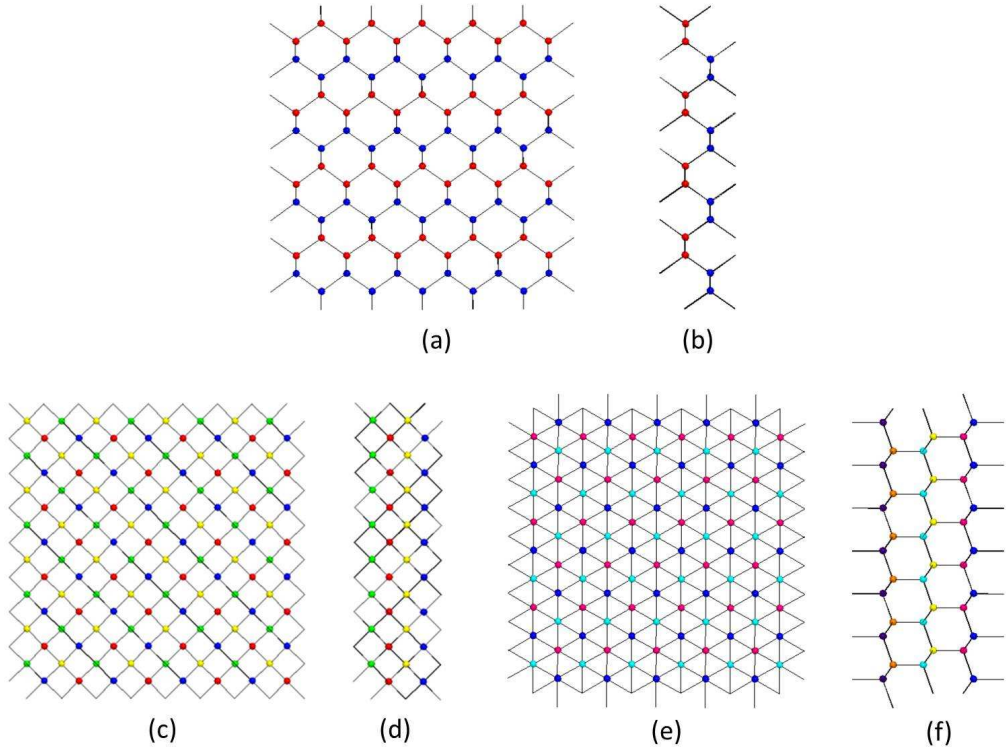


Figure 5.1: Cross section and profile of a unit cell for silicon nanowires along the [110] direction (a and b), [100] direction (c and d), and the [111] direction (e and f). The unit cell consists of two, four, and six atomic planes, respectively. Different planes are denoted with different colors.

In Section 5.2, the condensation of the Hamiltonian matrix (in the nearest neighbor tight-binding schemes) for the semi-infinite leads is derived in detail, followed by the applications of the decimation approach and the eigenvalue

approach, respectively. Some numerical examples are provided in Section 5.3 to show the accuracy and the efficiency. In Section 5.4, the methods in this chapter are generalized to second-near (and third-near) neighbor interaction schemes. Section 5.5 gives a brief summary and also some possible extensions.

5.2 Description of the Methods

5.2.1 Condensation of the Hamiltonian Matrix

A typical two-probe system as illustrated in Fig. 5.2 is considered here, where the system Hamiltonian is divided into $\overline{\mathbf{H}}_L$, $\overline{\mathbf{H}}_D$, and $\overline{\mathbf{H}}_R$. The self energy calculation for the right lead will be the focus as the left lead can be done similarly. The Green's function matrix $\overline{\mathbf{g}}_R$ for the right lead at energy point E is defined as

$$\left(E\overline{\mathbf{I}} - \overline{\mathbf{H}}_R\right)\overline{\mathbf{g}}_R = \overline{\mathbf{I}}, \tag{5.1}$$

where $\overline{\mathbf{I}}$ is the identity matrix. Orthogonal basis has been assumed here; non-orthogonal basis case can be done by replacing $E\overline{\mathbf{I}}$ with overlap matrix $E\overline{\mathbf{S}}$.

Take the nearest neighbor interaction scheme for example (the generalization to second-near or third-near neighbor interaction schemes is discussed in Section 5.4), the matrix $\overline{\mathbf{H}}_R$ can be written in a block tridiagonal form and $\overline{\mathbf{g}}_R$ is usually a full matrix,

$$\overline{\mathbf{H}}_R = \begin{pmatrix} \overline{\mathbf{H}}_{1,1} & \overline{\mathbf{H}}_{1,2} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{H}}_{1,2}^{\dagger} & \overline{\mathbf{H}}_{2,2} & \overline{\mathbf{H}}_{2,3} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{H}}_{2,3}^{\dagger} & \overline{\mathbf{H}}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \ \overline{\mathbf{g}}_R = \begin{pmatrix} \overline{\mathbf{g}}_{1,1} & \overline{\mathbf{g}}_{1,2} & \overline{\mathbf{g}}_{1,3} & \cdots \\ \overline{\mathbf{g}}_{2,1} & \overline{\mathbf{g}}_{2,2} & \overline{\mathbf{g}}_{2,3} & \cdots \\ \overline{\mathbf{g}}_{3,1} & \overline{\mathbf{g}}_{3,2} & \overline{\mathbf{g}}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5.2}$$

where $\overline{\mathbf{H}}_{p,q}$ with $p = q$ denotes the on-site Hamiltonian for atomic plane p and $\overline{\mathbf{H}}_{p,q}$ with $p \neq q$ denotes the coupling Hamiltonian between atomic plane p and q, $\overline{\mathbf{H}}_{p,q}^{\dagger}$ is the Hermitian conjugate of $\overline{\mathbf{H}}_{p,q}$. $\overline{\mathbf{H}}_{1,0} = 0$ has been used since the semi-infinite lead terminates at plane 1.

According to (5.1) and (5.2), the Green's function $\overline{\mathbf{g}}_{p,q}$ for $q = 1$ should

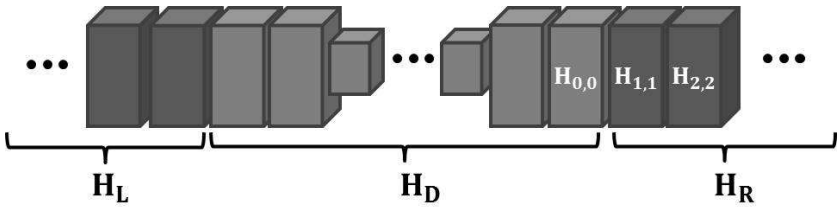


Figure 5.2: Schematic representation of a two-probe system. The system consists of a device part with Hamiltonian $\overline{\mathbf{H}}_D$ and two semi-infinite leads with Hamiltonian $\overline{\mathbf{H}}_L$ and $\overline{\mathbf{H}}_R$. The system is divided into many atomic planes and the right lead is described with atomic plane Hamiltonian $\overline{\mathbf{H}}_{p,p}$ $(p = 1, 2, \cdots)$ as illustrated.

satisfy the following equation,

$$\begin{pmatrix} E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{H}}_{1,1} & -\overline{\mathbf{H}}_{1,2} & \overline{\mathbf{0}} & \cdots \\ -\overline{\mathbf{H}}_{1,2}^{\dagger} & E\overline{\mathbf{I}}_{2,2} - \overline{\mathbf{H}}_{2,2} & -\overline{\mathbf{H}}_{2,3} & \cdots \\ \overline{\mathbf{0}} & -\overline{\mathbf{H}}_{2,3}^{\dagger} & E\overline{\mathbf{I}}_{3,3} - \overline{\mathbf{H}}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \overline{\mathbf{g}}_{1,1} \\ \overline{\mathbf{g}}_{2,1} \\ \overline{\mathbf{g}}_{3,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{I}}_{1,1} \\ \overline{\mathbf{0}} \\ \overline{\mathbf{0}} \\ \vdots \end{pmatrix}. \tag{5.3}$$

Assuming that a unit cell in the lead consists of P atomic planes, the Hamiltonian then repeats every P atomic planes, i.e.,

$$\overline{\mathbf{H}}_{nP+p,nP+q} = \overline{\mathbf{H}}_{p,q}, \;\; (p = 1, 2, \cdots, P;\; q = p, p+1;\; n = 1, 2, \cdots). \tag{5.4}$$

Utilizing this fact, equation (5.3) can be rewritten in the following format with matrix partitioning,

$$\begin{pmatrix} E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{H}}_{1,1} & \overline{\mathbf{B}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{B}}^{\dagger} & \overline{\mathbf{C}} & \overline{\mathbf{D}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{D}}^{\dagger} & E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{H}}_{1,1} & \overline{\mathbf{B}} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{B}}^{\dagger} & \overline{\mathbf{C}} & \overline{\mathbf{D}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \bullet$$

$$\begin{pmatrix} \overline{\mathbf{g}}_{1,1} \\ \overline{\mathbf{g}}_{2\sim P,1} \\ \overline{\mathbf{g}}_{P+1,1} \\ \overline{\mathbf{g}}_{(P+2)\sim 2P,1} \\ \overline{\mathbf{g}}_{2P+1,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{I}}_{1,1} \\ \overline{\mathbf{0}} \\ \overline{\mathbf{0}} \\ \overline{\mathbf{0}} \\ \vdots \end{pmatrix}, \tag{5.5}$$

where several new blocks have been defined,

$$\overline{\mathbf{B}} = \left(-\overline{\mathbf{H}}_{1,2},\ \overline{\mathbf{0}},\ \cdots,\ \overline{\mathbf{0}}\right), \quad \overline{\mathbf{D}} = \begin{pmatrix} \overline{\mathbf{0}} \\ \vdots \\ \overline{\mathbf{0}} \\ -\overline{\mathbf{H}}_{P,P+1} \end{pmatrix}, \tag{5.6}$$

$$\overline{\mathbf{C}} = \begin{pmatrix} E\overline{\mathbf{I}}_{2,2} - \overline{\mathbf{H}}_{2,2} & -\overline{\mathbf{H}}_{2,3} & \cdots & \overline{\mathbf{0}} \\ -\overline{\mathbf{H}}_{2,3}^{\dagger} & E\overline{\mathbf{I}}_{3,3} - \overline{\mathbf{H}}_{3,3} & \cdots & \overline{\mathbf{0}} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} & \cdots & E\overline{\mathbf{I}}_{P,P} - \overline{\mathbf{H}}_{P,P} \end{pmatrix}, \tag{5.7}$$

$$\overline{\mathbf{g}}_{2\sim P,1} = \begin{pmatrix} \overline{\mathbf{g}}_{2,1} \\ \overline{\mathbf{g}}_{3,1} \\ \vdots \\ \overline{\mathbf{g}}_{P,1} \end{pmatrix}, \quad \overline{\mathbf{g}}_{(P+2)\sim 2P,1} = \begin{pmatrix} \overline{\mathbf{g}}_{P+2,1} \\ \overline{\mathbf{g}}_{P+3,1} \\ \vdots \\ \overline{\mathbf{g}}_{2P,1} \end{pmatrix}. \tag{5.8}$$

Eliminating $\overline{\mathbf{g}}_{2\sim P,1}$, $\overline{\mathbf{g}}_{(P+2)\sim 2P,1}$, $\cdots$, in equation (5.5) results in,

$$\begin{pmatrix} E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}_s & -\overline{\mathbf{\Pi}} & \overline{\mathbf{0}} & \cdots \\ -\overline{\mathbf{\Pi}}^{\dagger} & E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}} & -\overline{\mathbf{\Pi}} & \cdots \\ \overline{\mathbf{0}} & -\overline{\mathbf{\Pi}}^{\dagger} & E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \overline{\mathbf{g}}_{1,1} \\ \overline{\mathbf{g}}_{P+1,1} \\ \overline{\mathbf{g}}_{2P+1,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{I}}_{1,1} \\ \overline{\mathbf{0}} \\ \overline{\mathbf{0}} \\ \vdots \end{pmatrix}, \tag{5.9}$$

where,

$$\overline{\mathbf{\Xi}}_s = \overline{\mathbf{H}}_{1,1} + \overline{\mathbf{BC}}^{-1}\overline{\mathbf{B}}^{\dagger}, \tag{5.10}$$

$$\overline{\mathbf{\Xi}} = \overline{\mathbf{H}}_{1,1} + \overline{\mathbf{BC}}^{-1}\overline{\mathbf{B}}^{\dagger} + \overline{\mathbf{D}}^{\dagger}\overline{\mathbf{C}}^{-1}\overline{\mathbf{D}}, \tag{5.11}$$

$$\overline{\mathbf{\Pi}} = \overline{\mathbf{BC}}^{-1}\overline{\mathbf{D}}. \tag{5.12}$$

From equation (5.9), a condensed Hamiltonian can be identified that only consists of planes $p = nP + 1$ $(n = 0, 1, \cdots)$, i.e.,

$$\overline{\mathbf{H}}_{cnd} = \begin{pmatrix} \overline{\mathbf{\Xi}}_s & \overline{\mathbf{\Pi}} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{\Pi}}^{\dagger} & \overline{\mathbf{\Xi}} & \overline{\mathbf{\Pi}} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{\Pi}}^{\dagger} & \overline{\mathbf{\Xi}} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5.13}$$

where the blocks are of the size $\sim (N/P) \times (N/P)$ with N being the matrix dimension of a unit cell. Note that the condensed on-site Hamiltonian $\overline{\overline{\mathbf{\Xi}}}_s$ of plane 1 differs from condensed on-site Hamiltonian $\overline{\overline{\mathbf{\Xi}}}$ of plane $p = nP + 1$ $(n = 1, 2, \cdots)$, as $\overline{\overline{\mathbf{\Xi}}}_s$ only includes the influences of right side planes (plane 2 to P) while $\overline{\overline{\mathbf{\Xi}}}$ includes the influences of both sides (plane $(n-1)P + 2$ to nP and plane $nP + 2$ to $(n+1)P$). The condensed coupling Hamiltonian $\overline{\overline{\mathbf{\Pi}}}$ connects plane $p = nP + 1$ to plane $p = (n+1)P + 1$ directly.

The problem now is to evaluate the expressions of $\overline{\overline{\mathbf{\Xi}}}_s$, $\overline{\overline{\mathbf{\Xi}}}$, and $\overline{\overline{\mathbf{\Pi}}}$ as shown in (5.10)-(5.12). This requires inversion of matrix $\overline{\mathbf{C}}$ of the size $\sim \left(\frac{P-1}{P}N\right) \times \left(\frac{P-1}{P}N\right)$, which can be done efficiently since it is highly sparse. Alternatively, the full inversion can be avoided by noticing that the matrix $\overline{\mathbf{B}}$ or $\overline{\mathbf{D}}$ consists of only one non-zero block and thus several blocks in $\overline{\mathbf{C}}^{-1}$ are actually needed. In fact, by denoting $\overline{\mathbf{C}}^{-1}$ as

$$\overline{\mathbf{C}}^{-1} = \begin{pmatrix} \widetilde{\overline{\mathbf{C}}}_{2,2} & \widetilde{\overline{\mathbf{C}}}_{2,3} & \cdots & \widetilde{\overline{\mathbf{C}}}_{2,P} \\ \widetilde{\overline{\mathbf{C}}}_{3,2} & \widetilde{\overline{\mathbf{C}}}_{3,3} & \cdots & \widetilde{\overline{\mathbf{C}}}_{3,P} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\overline{\mathbf{C}}}_{P,2} & \widetilde{\overline{\mathbf{C}}}_{P,3} & \cdots & \widetilde{\overline{\mathbf{C}}}_{P,P} \end{pmatrix}, \tag{5.14}$$

due to (5.6), it is found that only $\widetilde{\overline{\mathbf{C}}}_{2,2}$, $\widetilde{\overline{\mathbf{C}}}_{2,P}$, and $\widetilde{\overline{\mathbf{C}}}_{P,P}$ are relevant. Furthermore, these blocks can be calculated efficiently with forward and backward recursions since $\overline{\mathbf{C}}$ is block tridiagonal. The details are as follows:

ALGORITHM 0 (Recursive Condensation of the Hamiltonian Matrix):

1. $\widetilde{\overline{\mathbf{H}}}_{P,P} = \left(E\overline{\mathbf{I}}_{P,P} - \overline{\mathbf{H}}_{P,P}\right)^{-1}$
2. For $p = P-1, P-2, \cdots, 2$ (in this order), do $\{$
3. $\widetilde{\overline{\mathbf{H}}}_{p,p} = \left(E\overline{\mathbf{I}}_{p,p} - \overline{\mathbf{H}}_{p,p} - \overline{\mathbf{H}}_{p,p+1}\widetilde{\overline{\mathbf{H}}}_{p+1,p+1}\overline{\mathbf{H}}^{\dagger}_{p,p+1}\right)^{-1}$
4. $\widetilde{\overline{\mathbf{H}}}_{p,P} = \widetilde{\overline{\mathbf{H}}}_{p,p}\overline{\mathbf{H}}_{p,p+1}\widetilde{\overline{\mathbf{H}}}_{p+1,P}$ $\}$
5. $\widetilde{\overline{\mathbf{C}}}_{2,2} = \widetilde{\overline{\mathbf{H}}}_{2,2}$, $\widetilde{\overline{\mathbf{C}}}_{2,P} = \widetilde{\overline{\mathbf{H}}}_{2,P}$
6. For $p = 3, \cdots, P$ (in this order), do $\{$
7. $\widetilde{\overline{\mathbf{C}}}_{p,p} = \widetilde{\overline{\mathbf{H}}}_{p,p} + \widetilde{\overline{\mathbf{H}}}_{p,p}\left(\overline{\mathbf{H}}_{p,p-1}\widetilde{\overline{\mathbf{C}}}_{p-1,p-1}\overline{\mathbf{H}}^{\dagger}_{p,p-1}\right)\widetilde{\overline{\mathbf{H}}}_{p,p}$ $\}$
8. Obtain $\overline{\overline{\mathbf{\Xi}}}_s = \overline{\mathbf{H}}_{1,1} + \overline{\mathbf{H}}_{1,2}\widetilde{\overline{\mathbf{C}}}_{2,2}\overline{\mathbf{H}}^{\dagger}_{1,2}$
9. Obtain $\overline{\overline{\mathbf{\Xi}}} = \overline{\overline{\mathbf{\Xi}}}_s + \overline{\mathbf{H}}^{\dagger}_{P,P+1}\widetilde{\overline{\mathbf{C}}}_{P,P}\overline{\mathbf{H}}_{P,P+1}$
10. Obtain $\overline{\overline{\mathbf{\Pi}}} = \overline{\mathbf{H}}_{1,2}\widetilde{\overline{\mathbf{C}}}_{2,P}\overline{\mathbf{H}}_{P,P+1}$

With this condensed Hamiltonian (5.13) of reduced size, it is now ready to calculate the self energy either by the iterative approach or the eigenvalue

approach as described separately in the following.

5.2.2 Iterative Approach

As seen from matrix (5.13), the translational invariance is broken by the first block. Nevertheless, the decimation method [104] can still be applied to the chain in (5.9).

The main idea is to continue eliminating $\overline{\mathbf{g}}_{P+1,1}$, $\overline{\mathbf{g}}_{3P+1,1}$, $\cdots$, in equation (5.9), which results in,

$$\begin{pmatrix} E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}_s^1 & -\overline{\mathbf{\Pi}}^1 & \overline{\mathbf{0}} & \cdots \\ -\overline{\mathbf{\Pi}}^{1\dagger} & E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}^1 & -\overline{\mathbf{\Pi}}^1 & \cdots \\ \overline{\mathbf{0}} & -\overline{\mathbf{\Pi}}^{1\dagger} & E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}^1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \overline{\mathbf{g}}_{1,1} \\ \overline{\mathbf{g}}_{2P+1,1} \\ \overline{\mathbf{g}}_{4P+1,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{I}}_{1,1} \\ \overline{\mathbf{0}} \\ \overline{\mathbf{0}} \\ \vdots \end{pmatrix}, \tag{5.15}$$

where,

$$\overline{\mathbf{\Xi}}_s^1 = \overline{\mathbf{\Xi}}_s + \overline{\mathbf{\Pi}}\left(E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}\right)^{-1}\overline{\mathbf{\Pi}}^{\dagger}, \tag{5.16}$$

$$\overline{\mathbf{\Xi}}^1 = \overline{\mathbf{\Xi}} + \overline{\mathbf{\Pi}}^{\dagger}\left(E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}\right)^{-1}\overline{\mathbf{\Pi}} + \overline{\mathbf{\Pi}}\left(E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}\right)^{-1}\overline{\mathbf{\Pi}}^{\dagger}, \tag{5.17}$$

$$\overline{\mathbf{\Pi}}^1 = \overline{\mathbf{\Pi}}\left(E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}\right)^{-1}\overline{\mathbf{\Pi}}, \tag{5.18}$$

$$\overline{\mathbf{\Pi}}^{1\dagger} = \overline{\mathbf{\Pi}}^{\dagger}\left(E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}\right)^{-1}\overline{\mathbf{\Pi}}^{\dagger}. \tag{5.19}$$

Now equations (5.15) defines a new chain with planes $p = 2nP + 1$ ($n = 0, 1, 2, \cdots$).

The above process is repeated by eliminating $\overline{\mathbf{g}}_{2P+1,1}$, $\overline{\mathbf{g}}_{6P+1,1}$, $\cdots$, in (5.15), which results in a yet new chain with planes $p = 4nP + 1$ ($n = 0, 1, 2, \cdots$), and the process is repeated continuously. After i repetitions, a chain with planes $2^i nP + 1$ ($n = 0, 1, 2, \cdots$) will be obtained, where the on-site Hamiltonian are $\overline{\mathbf{\Xi}}_s^i$ and $\overline{\mathbf{\Xi}}^i$, the coupling Hamiltonian are $\overline{\mathbf{\Pi}}^i$ and $\overline{\mathbf{\Pi}}^{i\dagger}$. It should be noted that $\overline{\mathbf{\Pi}}^i$ and $\overline{\mathbf{\Pi}}^{i\dagger}$ will become weaker and weaker. The process can be terminated when the coupling is small enough ($\overline{\mathbf{\Pi}}^i \approx 0$ and $\overline{\mathbf{\Pi}}^{i\dagger} \approx 0$), and then the surface Green's function $\overline{\mathbf{g}}_{1,1}$ can be computed by,

$$\overline{\mathbf{g}}_{1,1} = \left(E\bar{\mathbf{I}}_{1,1}-\overline{\mathbf{\Xi}}_s^i\right)^{-1}. \tag{5.20}$$

Finally, the self energy is constructed through $\overline{\mathbf{g}}_{1,1}$ by

$$\overline{\mathbf{\Sigma}} = \overline{\mathbf{H}}_{0,1}\overline{\mathbf{g}}_{1,1}\overline{\mathbf{H}}_{0,1}^{\dagger}. \tag{5.21}$$

The above approach is implemented in the following way,

ALGORITHM I (Iterative method):

0. Do ALGORITHM 0.

1. Let $E^* = E + i\eta$, where η is an infinitesimal positive quantity. Let $\overline{\mathbf{A}} = \left(E^*\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}\right)$, $\overline{\mathbf{A}}_s = \left(E^*\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}_s\right)$. Note that $\overline{\mathbf{A}} \neq \overline{\mathbf{A}}_s$.

2. While $(\max\left(\left|\overline{\mathbf{\Pi}}\right|, \left|\overline{\mathbf{\Pi}}^{\dagger}\right|\right) > \delta)$, do {

3. Solve $\overline{\mathbf{A}}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}} = \overline{\mathbf{\Pi}}$ for $\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}}$

4. Solve $\overline{\mathbf{A}}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}^{\dagger}} = \overline{\mathbf{\Pi}}^{\dagger}$ for $\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}^{\dagger}}$

5. Update $\overline{\mathbf{A}} = \overline{\mathbf{A}} - \overline{\mathbf{\Pi}}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}^{\dagger}} - \overline{\mathbf{\Pi}}^{\dagger}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}}$

6. Update $\overline{\mathbf{A}}_s = \overline{\mathbf{A}}_s - \overline{\mathbf{\Pi}}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}^{\dagger}}$

7. Update $\overline{\mathbf{\Pi}} = \overline{\mathbf{\Pi}}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}}$

8. Update $\overline{\mathbf{\Pi}}^{\dagger} = \overline{\mathbf{\Pi}}^{\dagger}\overline{\mathbf{X}}_{\overline{\mathbf{\Pi}}^{\dagger}}$ }

9. Solve $\overline{\mathbf{A}}_s\overline{\mathbf{Y}} = \overline{\mathbf{H}}_{0,1}^{\dagger}$ for $\overline{\mathbf{Y}}$

10. Obtain the self energy $\overline{\mathbf{\Sigma}} = \overline{\mathbf{H}}_{0,1}\overline{\mathbf{Y}}$

It should be emphasized that, although the decimation may be directly applied to the original chain in (5.3), the implementation here is systematic and much simpler, as now all the layers (except the first one) in (5.13) are made identical no matter how many different atomic planes there are in a unit cell.

5.2.3 Eigenvalue Approach

The eigenvalue approach [102, 103], however, cannot be directly applied to the chain in (5.13). Fortunately, it can still be applied to the chain starting from layer 2, and the extra treatment of layer 1 can be done without too much effort.

First, let's define a new semi-infinite chain that starts from layer 2 of (5.13). By using Bloch wave condition

$$\mathbf{\Psi}_{p+P} = \lambda\mathbf{\Psi}_p, \tag{5.22}$$

where $\lambda = e^{ikd}$ with k real (complex) for propagating (evanescent) modes,

we have the following equation for Bloch waves,

$$-\lambda^{-1}\overline{\mathbf{\Pi}}^{\dagger}\mathbf{\Psi}_p + \left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}\right)\mathbf{\Psi}_p - \lambda\overline{\mathbf{\Pi}}\mathbf{\Psi}_p = \overline{\mathbf{0}}. \tag{5.23}$$

This equation can be solved by transforming to a GEVP of size $2N_1$ (N_1 is the size of $\overline{\mathbf{\Xi}}$), i.e.,

$$\begin{pmatrix} \overline{\mathbf{0}} & \overline{\mathbf{I}}_{1,1} \\ -\overline{\mathbf{T}}^{\dagger} & -\overline{\mathbf{D}} \end{pmatrix} \begin{pmatrix} \mathbf{\Psi}_p \\ \mathbf{\Psi}_{p+P} \end{pmatrix} = \lambda \begin{pmatrix} \overline{\mathbf{I}}_{1,1} & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{\Psi}_p \\ \mathbf{\Psi}_{p+P} \end{pmatrix}, \tag{5.24}$$

where the blocks are

$$\overline{\mathbf{D}} = E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}, \quad \overline{\mathbf{T}} = -\overline{\mathbf{\Pi}}. \tag{5.25}$$

Second, let's define a new Green's function $\overline{\mathbf{g}}'$ for this new semi-infinite chain, the blocks $\overline{\mathbf{g}}'_{p,q}$ for $q = 1$ should satisfy the following,

$$\left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}\right)\overline{\mathbf{g}}'_{1,1} = \overline{\mathbf{I}}_{1,1} + \overline{\mathbf{\Pi}}\overline{\mathbf{g}}'_{P+1,1}, \tag{5.26}$$

$$\left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}\right)\overline{\mathbf{g}}'_{P+1,1} = \overline{\mathbf{\Pi}}^{\dagger}\overline{\mathbf{g}}'_{1,1} + \overline{\mathbf{\Pi}}\overline{\mathbf{g}}'_{2P+1,1}, \tag{5.27}$$

$$\vdots$$

Then $\overline{\mathbf{g}}'_{1,1}$ can be expanded through Bloch modes of the chain,

$$\overline{\mathbf{g}}'_{1,1} = \overline{\mathbf{U}}^{+}\overline{\mathbf{C}}^{+}, \tag{5.28}$$

where matrix $\overline{\mathbf{U}}^{+}$ (of size $N_1 \times M$) consists of M right-going normalized Bloch vectors constructed from the first N_1 elements of the solution of (5.24), and matrix $\overline{\mathbf{C}}^{+}$ (of size $M \times N_1$) consists of N_1 vectors of corresponding expansion coefficients, i.e.,

$$\overline{\mathbf{U}}^{+} = \left(\mathbf{u}_1^{+}, \mathbf{u}_2^{+}, \cdots, \mathbf{u}_M^{+}\right), \tag{5.29}$$

$$\overline{\mathbf{C}}^{+} = \left(\mathbf{c}_1^{+}, \mathbf{c}_2^{+}, \cdots, \mathbf{c}_{N_1}^{+}\right). \tag{5.30}$$

Since the waves go outward from the δ source, $\overline{\mathbf{g}}'_{P+1,1}$ can be expressed as,

$$\overline{\mathbf{g}}'_{P+1,1} = \overline{\mathbf{U}}^{+}\overline{\mathbf{\Lambda}}^{+}\overline{\mathbf{C}}^{+}, \tag{5.31}$$

where the propagator $\overline{\mathbf{\Lambda}}^{+}$ is a $M \times M$ diagonal matrix with elements

$$\overline{\mathbf{\Lambda}}^{+}_{mm} = \lambda^{+}_{m}. \tag{5.32}$$

By defining pseudo-inverse $\widetilde{\overline{\mathbf{U}}^{+}}$ of $\overline{\mathbf{U}}^{+}$, i.e.,

$$\widetilde{\overline{\mathbf{U}}^{+}}\overline{\mathbf{U}}^{+} = \overline{\mathbf{I}}, \tag{5.33}$$

and using (5.31), $\overline{\mathbf{g}}'_{P+1,1}$ can be related to $\overline{\mathbf{g}}'_{1,1}$ through the following way

$$\overline{\mathbf{g}}'_{P+1,1} = \overline{\mathbf{U}}^{+}\overline{\mathbf{\Lambda}}^{+}\widetilde{\overline{\mathbf{U}}^{+}}\overline{\mathbf{U}}^{+}\overline{\mathbf{C}}^{+} = \overline{\mathbf{U}}^{+}\overline{\mathbf{\Lambda}}^{+}\widetilde{\overline{\mathbf{U}}^{+}}\overline{\mathbf{g}}'_{1,1} = \overline{\mathbf{F}}\overline{\mathbf{g}}'_{1,1}, \tag{5.34}$$

where a new propagator was defined

$$\overline{\mathbf{F}} = \overline{\mathbf{U}}^{+}\overline{\mathbf{\Lambda}}^{+}\widetilde{\overline{\mathbf{U}}^{+}}. \tag{5.35}$$

Similarly, the following holds

$$\overline{\mathbf{g}}'_{2P+1,1} = \overline{\mathbf{F}}\overline{\mathbf{g}}'_{P+1,1}. \tag{5.36}$$

Putting (5.34) and (5.36) into (5.26) and (5.27), we have

$$\left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}} - \overline{\mathbf{\Pi}\mathbf{F}}\right)\overline{\mathbf{g}}'_{1,1} = \overline{\mathbf{I}}_{1,1}, \tag{5.37}$$

$$\left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}} - \overline{\mathbf{\Pi}\mathbf{F}}\right)\overline{\mathbf{F}}\overline{\mathbf{g}}'_{1,1} = \overline{\mathbf{\Pi}}^{\dagger}\overline{\mathbf{g}}'_{1,1}. \tag{5.38}$$

From above two we can solve for the surface Green's function $\overline{\mathbf{g}}'_{1,1}$, which is

$$\overline{\mathbf{g}}'_{1,1} = \overline{\mathbf{F}\mathbf{\Pi}}^{\dagger -1}. \tag{5.39}$$

In the case when $\overline{\mathbf{\Pi}}^{\dagger}$ is not invertable, we solve for self energy directly, i.e.,

$$\overline{\mathbf{\Sigma}}' = \overline{\mathbf{\Pi}}\overline{\mathbf{g}}'_{1,1}\overline{\mathbf{\Pi}}^{\dagger} = \overline{\mathbf{\Pi}\mathbf{F}}. \tag{5.40}$$

Finally, the surface Green's function for the original chain (including layer 1 of (5.13)) is obtained as,

$$\overline{\mathbf{g}}_{1,1} = \left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}_{s} - \overline{\mathbf{\Sigma}}'\right)^{-1}, \tag{5.41}$$

and the self energy is constructed using,

$$\overline{\mathbf{\Sigma}} = \overline{\mathbf{H}}_{0,1}\overline{\mathbf{g}}_{1,1}\overline{\mathbf{H}}_{0,1}^{\dagger}. \tag{5.42}$$

The above approach (the approach is self-consistent since we can verify that Eqs. (5.28), (5.31), and (5.36) satisfy Eqs. (5.26) and (5.27) by direct substitution) is implemented in the following way,

ALGORITHM II (Eigenvalue method):

0. Do ALGORITHM 0.

1. Let $\overline{\mathbf{A}} = \begin{pmatrix} \overline{\mathbf{0}} & \overline{\mathbf{I}}_{1,1} \\ -\overline{\mathbf{T}}^{\dagger} & -\overline{\mathbf{D}} \end{pmatrix}$, and $\overline{\mathbf{B}} = \begin{pmatrix} \overline{\mathbf{I}}_{1,1} & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{T}} \end{pmatrix}$.

2. Instead of solving a generalized eigenvalue problem $\overline{\mathbf{A}}\mathbf{\Psi} = \lambda\overline{\mathbf{B}}\mathbf{\Psi}$, we resort to a normal eigenvalue problem by constructing $\widetilde{\overline{\mathbf{A}}} = \left(\overline{\mathbf{A}} - \sigma\overline{\mathbf{B}}\right)^{-1}\overline{\mathbf{B}}$, where σ is a shift. Note that the 2×2 block matrix $\left(\overline{\mathbf{A}} - \sigma\overline{\mathbf{B}}\right)$ can be inverted efficiently by using the Schur complement block [106].

3. Solve the normal eigenvalue problem $\widetilde{\overline{\mathbf{A}}}\mathbf{\Psi} = \widetilde{\lambda}\mathbf{\Psi}$, obtain the eigenpairs $\left(\widetilde{\lambda}, \mathbf{\Psi}\right)$.

4. Obtain the eigenpairs of the original problem: $\left(\lambda = \widetilde{\lambda}^{-1} + \sigma, \mathbf{\Psi}\right)$.

5. Retrieve all the eigenpairs corresponding to the right-going propagating modes with $|\lambda| = 1$; Retrieve a part of the eigenpairs corresponding to the right-going evanescent modes with $\epsilon < |\lambda| < 1$, where ϵ can be truncated to include only slowly decaying evanescent modes. Construct a $N_1 \times M$ matrix $\overline{\mathbf{U}}^{+}$ and a $M \times M$ diagonal matrix $\overline{\mathbf{\Lambda}}^{+}$ from these eigenpairs.

6. Obtain pseudo-inverse $\widetilde{\overline{\mathbf{U}}^{+}}$ of $\overline{\mathbf{U}}^{+}$ by factorizing $\overline{\mathbf{U}}^{+} = \overline{\mathbf{QR}}$ and solving $\widetilde{\overline{\mathbf{RU}}^{+}} = \overline{\mathbf{Q}}^{\dagger}$.

7. Construct $\overline{\mathbf{F}}$ according to (5.35). Solve $\left(E\overline{\mathbf{I}}_{1,1} - \overline{\mathbf{\Xi}}_s - \overline{\mathbf{\Pi F}}\right)\overline{\mathbf{Y}} = \overline{\mathbf{H}}_{0,1}^{\dagger}$ for $\overline{\mathbf{Y}}$. Note that this is the only step where layer 1 ($\overline{\mathbf{\Xi}}_s$) comes in.

8. Obtain the self energy $\overline{\mathbf{\Sigma}} = \overline{\mathbf{H}}_{0,1}\overline{\mathbf{Y}}$.

5.2.4 Computational Cost

To reduce the Hamiltonian to (5.13), as shown in ALGORITHM 0, it requires $P-1$ inversions of the small matrices of the size $\sim (N/P)$. The cost is $(P-1)\times O\left((N/P)^3\right)$, which is very cheap. Once (5.13) is obtained, the computational cost of ALGORITHM I is $(M+1)\times O\left((N/P)^3\right)$ if the process converges in M steps (usually 20 to 50 steps, depending on the value of

η). This is a tremendous reduction compared with the original decimation method [104], where the complexity is $(M+1) \times O(N^3)$ (here it is assumed that the inversions are carried out for matrices of the size of a unit cell). The computational cost of ALGORITHM II is $O((2N/P)^3) + O((N/P)^3)$, where the first term is due to step 3, and the second term due to steps 2, 6, and 7. This is also a significant improvement over the original eigenvalue approach [101–103], the cost of which is about $O((2N)^3) + O(N^3)$. Note that $P = 4$ for [100] orientation and $P = 6$ for [111] and [112].

5.3 Results and Discussion

The testing examples are rectangular silicon nanowires. The representation of the Hamiltonian matrix is through $sp^3d^5s^*$ tight binding scheme with nearest neighbor interaction (10 orbits per atom without spin-orbit coupling, and 20 orbits per atom with spin-orbit coupling) [23]. The dangling sp^3 hybridized bonds at the surfaces are passivated using hydrogen-like atoms [109]. This tight binding scheme has been widely employed to study nanowire transistors. Please refer to Ref. [22, 110] for the details of the Hamiltonian construction.

5.3.1 Validation of the Method

First, to validate these methods, the transmission spectrum of an unbiased perfect silicon nanowire was calculated with Green's function approach [41, 97]. The self energies involved were obtained by ALGORITHM I and II respectively. The results are shown in Fig. 5.3, also shown are the E-k dispersion and density of states (DOS) calculated for an infinite periodic nanowire.

It is clearly seen that the transmission is an integer over the whole band and it steps up or down when a transmission channel is opened or closed. The transition points of the transmission match perfectly with the positions of the one dimensional DOS peaks (van Hove singularities), indicating that our transmission calculation is reliable, and in turn, validating our self energy calculations. Note that to explain the transmission in valance band, it is better to trace through the E-k diagram since there are additional DOS peaks which do not correspond to the van Hove singularities and the number of

transmission channel remains unchanged when one goes through these peaks. Similar phenomena can be observed for a [111] oriented silicon nanowire (see Fig. 5.4).

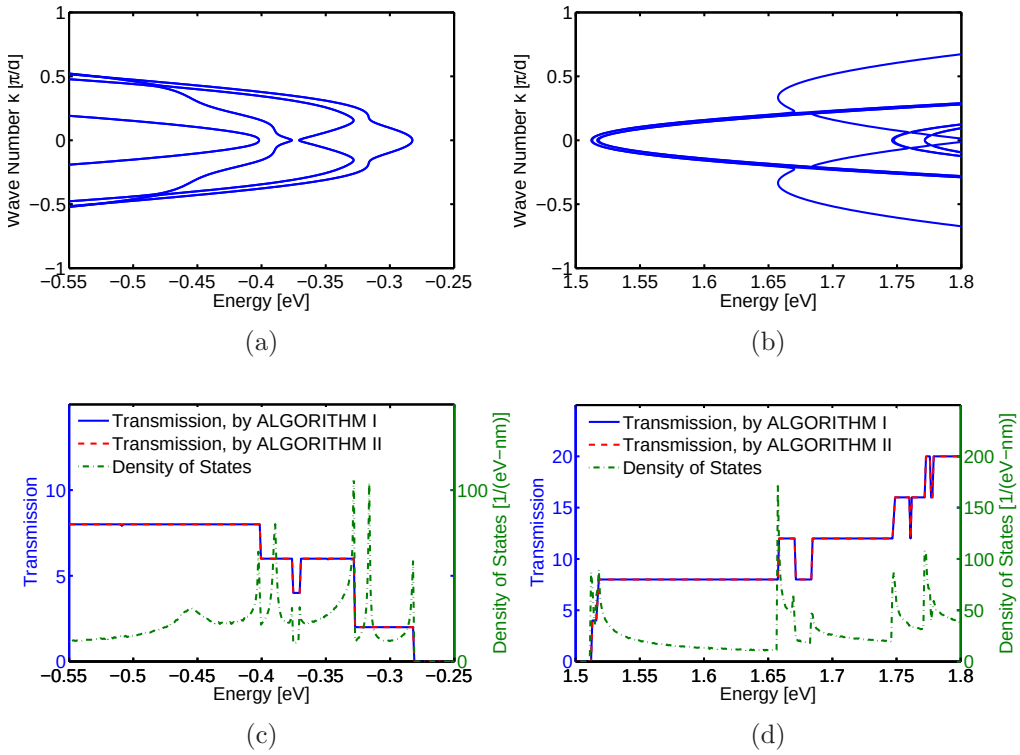


Figure 5.3: (a) and (b): E-k relation, (c) and (d): transmission spectrum and DOS, for an ideal [100] oriented silicon nanowire with cross-section $\sim$ 2nm $\times$ 2nm. (a) and (c): for valance band, and spin-orbit coupling is included in the calculation, (b) and (d): for conduction band, and spin-orbit coupling is not included in the calculation. No external bias is applied. The transmissions calculated by the two methods in this chapter lie almost on top of each other.

5.3.2 Comparison with Other Methods

Next, to show the efficiency, the run times of these algorithms along with those of the existing methods are list in Table 5.1.

For the iterative methods (methods 1, 2, and 3), $\eta = 10^{-9}$eV is chosen that the iterative processes converge in a certain number of steps. It is seen that ALGORITHM I can greatly speed up the simulation compared with the fastest iterative one, i.e., method 2. It should be mentioned that in this

work, method 2 is implemented by inverting the matrices of the unit cell. In particular, for [100] and [111] directions, an acceleration factor of about 40 to 80 is gained. Note that for these two cases, sparse matrix operations have been implemented in method 2 as the matrices involved have many zero blocks.

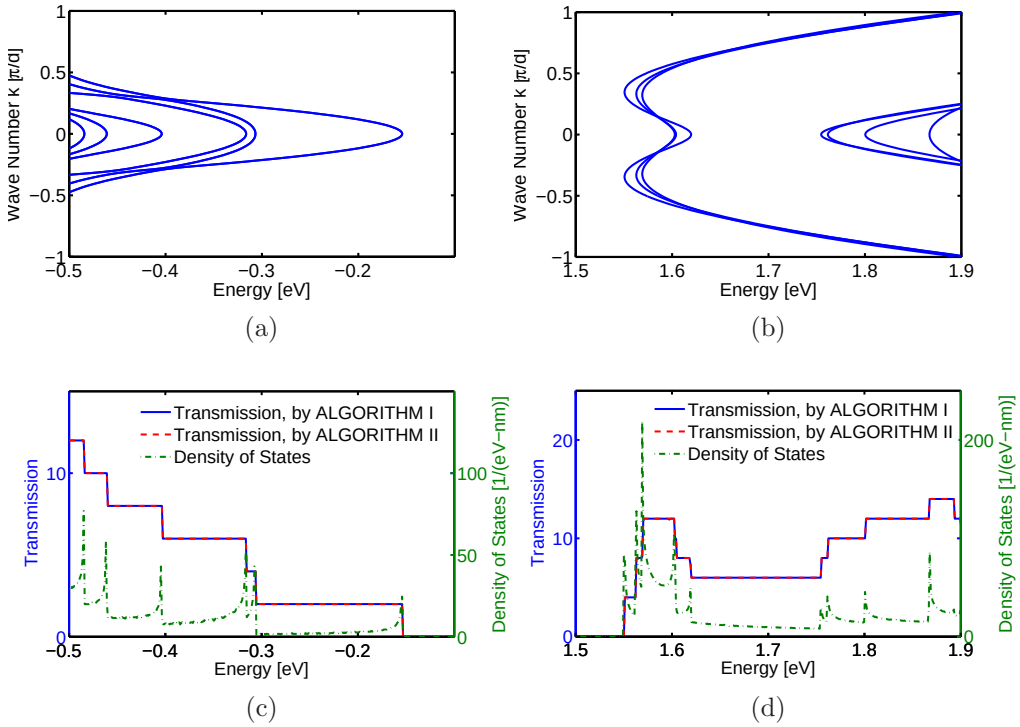


Figure 5.4: Same plots as Fig. 5.3, except that the silicon nanowire is [111] oriented.

While among the eigenvalue approaches (methods 4, 5, and 6), ALGORITHM II is the best and it slightly outperforms the fastest existing one, i.e., method 5. Note that sparse matrix operations have been implemented in method 5 so that the matrix inversion involved is very efficient.

To include spin-orbit interaction, which is important for hole transport, the computational cost is significantly increased. The reason is two fold, one is that the number of orbits doubles, the other is the introduction of complex operations (in the eigenvalue approaches) as a result of complex Hamiltonian elements. Generally speaking, ALGORITHM I and II are comparable in terms of speed when spin-orbit coupling is included; ALGORITHM II shows advantage when spin-orbit coupling is not included due to the real arithmetic,

Table 5.1: List of run times (in seconds) for self energy evaluation in one energy point for silicon nanowires with cross section $\sim 2\text{nm} \times 2\text{nm}$. Calculations are carried out for three crystal directions ([110], [100], and [111]) and for two basis sets (without and with spin-orbital coupling). Six methods are implemented (in MATLAB). The quantities in the brackets are the speed degradation factors compared with the fastest method. The simulations are performed on an Intel Xeon processor (restricted to four cores, 2.66 GHz).

Orientation	[110]	[100]	[111]
Number of planes p.u.c	2	4	6
Number of atoms p.u.c	88	128	208
Matrix size p.u.c	880	1280	2080
1. Iterative method[a]	1816 (386×)	819.4 (394×)	239.9 (79.2×)
2. Decimation [104]	34.3 (7.3×)	86.6 (41.6×)	169.2 (55.8×)
3. ALGORITHM I	4.71	2.08	3.03
4. NEVP method [106]	5.04 (5.5×)	11.1 (21.8×)	38.9 (45.2×)
5. Advanced NEVP [108]	1.52 (1.7×)	1.77 (3.5×)	3.43 (4.0×)
6. ALGORITHM II	0.92	0.51	0.86
Matrix size p.u.c	1760	2560	4160
1. Iterative method[b]	13475 (409×)	5468 (390×)	1590 (83.4×)
2. Decimation [104]	262.6 (8.0×)	722.0 (51.5×)	1473 (77.2×)
3. ALGORITHM I	32.91	14.02	19.07
4. NEVP method [106]	108.6 (7.1×)	314.2 (40.9×)	1302 (92.4×)
5. Advanced NEVP [108]	22.36 (1.5×)	18.63 (2.4×)	36.47 (2.6×)
6. ALGORITHM II	15.26	7.69	14.09

[a] This is done by repetitive use of relations,
$\overline{\mathbf{g}}^{(n)}_{p,p} = \left(E^{*}\overline{\mathbf{I}}_{p,p} - \overline{\mathbf{H}}_{p,p} - \overline{\mathbf{H}}_{p,p+1}\overline{\mathbf{g}}^{(n)}_{p+1,p+1}\overline{\mathbf{H}}^{\dagger}_{p,p+1}\right)^{-1}$, for $p = P, P-1, \cdots, 1$, and
$\overline{\mathbf{g}}^{(n)}_{P+1,P+1} = \overline{\mathbf{g}}^{(n-1)}_{1,1}$.

[b] As described in footnote (a) above.

which is not the case in ALGORITHM I since a small imaginary part is introduced to ensure convergence.

5.4 Generalization to the Second- and Third-Near Neighbor Interaction Schemes

The methods proposed can be generalized to the second- and the third-near neighbor (2NN and 3NN) interaction schemes. Take 2NN interaction for example (3NN can be done in the same spirit), the Hamiltonian matrix in terms of atomic planes takes the form,

$$\overline{\mathbf{H}}_R = \begin{pmatrix} \overline{\mathbf{H}}_{1,1} & \overline{\mathbf{H}}_{1,2} & \overline{\mathbf{H}}_{1,3} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{H}}_{1,2}^{\dagger} & \overline{\mathbf{H}}_{2,2} & \overline{\mathbf{H}}_{2,3} & \overline{\mathbf{H}}_{2,4} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{H}}_{1,3}^{\dagger} & \overline{\mathbf{H}}_{2,3}^{\dagger} & \overline{\mathbf{H}}_{3,3} & \overline{\mathbf{H}}_{3,4} & \overline{\mathbf{H}}_{3,5} & \overline{\mathbf{0}} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{H}}_{2,4}^{\dagger} & \overline{\mathbf{H}}_{3,4}^{\dagger} & \overline{\mathbf{H}}_{4,4} & \overline{\mathbf{H}}_{4,5} & \overline{\mathbf{H}}_{4,6} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{H}}_{3,5}^{\dagger} & \overline{\mathbf{H}}_{4,5}^{\dagger} & \overline{\mathbf{H}}_{5,5} & \overline{\mathbf{H}}_{5,6} & \cdots \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{H}}_{4,6}^{\dagger} & \overline{\mathbf{H}}_{5,6}^{\dagger} & \overline{\mathbf{H}}_{6,6} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5.43}$$

which can be rewritten in a block tridiagonal form,

$$\overline{\mathbf{H}}_R = \begin{pmatrix} \widetilde{\overline{\mathbf{H}}}_{1,1} & \widetilde{\overline{\mathbf{H}}}_{1,2} & \overline{\mathbf{0}} & \cdots \\ \widetilde{\overline{\mathbf{H}}}_{1,2}^{\dagger} & \widetilde{\overline{\mathbf{H}}}_{2,2} & \widetilde{\overline{\mathbf{H}}}_{2,3} & \cdots \\ \overline{\mathbf{0}} & \widetilde{\overline{\mathbf{H}}}_{2,3}^{\dagger} & \widetilde{\overline{\mathbf{H}}}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5.44}$$

where the blocks are

$$\widetilde{\overline{\mathbf{H}}}_{1,1} = \begin{pmatrix} \overline{\mathbf{H}}_{1,1} & \overline{\mathbf{H}}_{1,2} \\ \overline{\mathbf{H}}_{1,2}^{\dagger} & \overline{\mathbf{H}}_{2,2} \end{pmatrix}, \ \widetilde{\overline{\mathbf{H}}}_{22} = \begin{pmatrix} \overline{\mathbf{H}}_{3,3} & \overline{\mathbf{H}}_{3,4} \\ \overline{\mathbf{H}}_{3,4}^{\dagger} & \overline{\mathbf{H}}_{4,4} \end{pmatrix}, \ \widetilde{\overline{\mathbf{H}}}_{33} = \begin{pmatrix} \overline{\mathbf{H}}_{5,5} & \overline{\mathbf{H}}_{5,6} \\ \overline{\mathbf{H}}_{5,6}^{\dagger} & \overline{\mathbf{H}}_{6,6} \end{pmatrix},$$
$$\widetilde{\overline{\mathbf{H}}}_{12} = \begin{pmatrix} \overline{\mathbf{H}}_{1,3} & \overline{\mathbf{0}} \\ \overline{\mathbf{H}}_{2,3} & \overline{\mathbf{H}}_{2,4} \end{pmatrix}, \ \widetilde{\overline{\mathbf{H}}}_{23} = \begin{pmatrix} \overline{\mathbf{H}}_{3,5} & \overline{\mathbf{0}} \\ \overline{\mathbf{H}}_{4,5} & \overline{\mathbf{H}}_{4,6} \end{pmatrix}. \tag{5.45}$$

Now, the method in Section 5.2.1 can be applied to equation (5.44) to condense the Hamiltonian matrix into a small one which consists only the planes $p = nP + 1$ and $p = nP + 2$, where $n = 0, 1, 2, \cdots$. Thus, a size reduction factor of $\sim 1/2$ is gained for [100] orientation and $\sim 1/3$ for [111] and [112]. With the condensed Hamiltonian matrix, the self energy matrix can be evaluated with the methods described in Sections 5.2.2 and 5.2.3.

5.5 Summary

In order to efficiently simulate quantum transport in nanodevices within NEGF or wave function formalism, two algorithms are proposed for the fast evaluation of self-energy matrices in atomistic simulations. The efficiency of the algorithms is based on constructing a condensed Hamiltonian with reduced size for the semi-infinite leads. The condensation successfully takes advantage of the crystal structures together with the short-range interactions of tight binding schemes. The reliability of the methods has been demonstrated by studying the transmission of an ideal silicon nanowire in the nearest neighbor interaction scheme. Extensive numerical examples and comparisons have shown that the methods can speed up the decimation approach by 7 to 80 times and can also out-perform the advanced eigenvalue approach by several times.

The methods are particularly useful when the unit cell in the leads is made very long due to the presence of doping atoms. This situation is very common in nano-electronics nowadays as the doping density (per nanometer) in the leads is usually very low as a result of the ultra-small cross sections. Furthermore, the methods can be applied to *ab initio* models as long as the interaction range is short compared with the unit cell length.

CHAPTER 6

CONCLUSION AND OUTLOOK

6.1 Numerical Methods

In this thesis, several numerical methods are developed to improve the efficiency of nanoelectronic transport simulations, with effective mass, $k \cdot p$, and tight binding models adopted as the Hamiltonian models. It is found that the numerical properties of the transport problems are strongly related to the Hamiltonian model employed.

For the effective mass model, AWE combined with CFH algorithm in Chapter 2 can reduce by over $10\times$ the number of energy points needed for getting the wide band results. The algorithm is valid for any number of leads, arbitrary device geometry, and various potential profile. This makes a very powerful tool for studying n-type devices with scattering induced by geometrical variations (surface roughness for instance) and potential variations (ionized impurities for example). In the case of multiband $k \cdot p$ simulations, the MOR methods in Chapters 3 and 4 construct reduced models which are nearly $100\times$ smaller, making it possible to simulate hole transport and band-to-band tunneling in large cross-section nanowire devices. The algorithms in Chapter 5 improve the self energy evaluation by condensing the atomistic Hamiltonian of the leads. Acceleration up to $80\times$ is demonstrated for silicon nanowires, and the acceleration will be larger if the unit cell in the leads is longer. Moreover, the condensation is exact with no approximation made and can be applied to any atomistic schemes.

Although the results are promising, there is still some room for improvements.

The $\overline{\mathbf{A}}\mathbf{x} = \mathbf{b}$ problem in Chapter 2 is based on sparse LU decomposition, the computational cost of which scales as $O\left(NM^2\right)$ where M is the bandwidth of $\overline{\mathbf{A}}$. The bandwidth M is usually proportional to the nanowire

cross section. Therefore, iterative solvers with lower complexity, like conjugate gradient (CG) based solvers, should be used for large nanowires. As there are multiple right hand sides, block methods can be used to generate block Krylov subspace, which is cheaper than solving each right hand side individually.

The eigenvalue solution in Chapter 3 is still very costly, especially when the nanowire size is large. In fact, the shift-and-invert Krylov subspace solvers scale as $O\left(N^2\right)$ where N is the matrix size of a layer. As the eigenvalue problems need to be solved for each layer, the cost may be reduced by reusing the Krylov subspace information generated in the neighboring layers, instead of generating the Krylov subspace individually. The other way is to parallelize the calculation of each layer, since they are independent of each other. The recursive Green's function algorithm for each energy point can be easily parallelized as well. In Chapter 4, the discretization is in Fourier space, which results in a dense matrix (due to the potential terms, although the kinetic terms are still relatively sparse). If Krylov subspace based iterative solvers are used to solve the eigenvalue problem, the matrix-vector product step will cost $O(N^2)$. However, the matrix-vector product can be separated into two parts, the sparse matrix-vector product part (for the kinetic terms) can be done with $O(N)$ complexity, and the dense matrix-vector product part (for the potential term) is a convolution that can be accelerated by Fast Fourier Transform (FFT) algorithm which is $O\left(N\log\left(N\right)\right)$ in complexity. This might be useful if the matrix in Fourier space is still too large to be solved directly.

Self energy matrix calculation in Chapter 5 could also be improved. Note that once the self energy matrix is obtained for a certain bias, it can be reused in the simulation of other biases. Therefore, it is better to calculate the self energy once and then store it in the memory or even in the hard disk for later.

6.2 Device Physics

By applying the developed methods, several emerging electronic devices are studied. In particular, p-type junctionless transistors and source-pocket InAs tunneling FETs operating in the quantum ballistic transport limit are studied for the first time. It is found that tunneling and band structure effects have

significant influences on the performances of junctionless transistors, and thus the doping density, channel orientation, and channel size should be carefully optimized in order to outperform classical inversion mode transistors. Source pocket is a good performance booster for tunneling FETs as it decreases (increases) the tunneling length at the "on" ("off") state and thus improves the subthreshold swing, but such effect tends to be diminished when the electrostatic integrity is improved by using GAA structure together with high-k gate oxide.

There are also some physical issues in need of further investigation.

First, inelastic scattering is not included in this work. It plays a very important role in classical devices, and still exists when devices shrink to nanoscale. Recent studies show that phonon-scattering has a significant impact on the I-V curves of nanoscale transistors [86, 111]. Electron-phonon scattering plays an important role in junctionless transistors as well [62] and thus should be taken into account in the future work.

Second, for ultra small devices and new materials (like MoS_2 [112]) with unknown parameters, first-principles calculations are necessary. First-principles method combining density functional theory (DFT) and NEGF approach [26] has found applications in simulation of exotic molecular devices. It is a parameter-free approach and treats exchange-correlation effects rigorously. But the computational cost is so heavy that most codes can only deal with thousands of atoms or less. It is worth mentioning that significant improvements have been made recently by employing parallelization and GPUs (graphics processing unit) [113]. The other choice is the density functional tight binding (DFTB) method [114,115], which is a compromise between the accuracy of DFT and the efficiency of tight binding method.

Third, multi-scale method combining the efficiency of classical models and the accuracy of quantum models is another feasible approach to efficient electronic device modeling [116, 117]. Fourth, electronic transport is inherently coupled with phonon transport, so, multi-physics solution could be important in understanding some electronic and thermal problems. With regard to optoelectronic applications, especially devices incorporating quantum-well or quantum dot structures [118], there is a call for solving quantum transport equations and Maxwell's equations together.

APPENDIX A

DERIVATION OF THE LANDAUER-BÜTTIKER FORMULA

Because the electron density of a single k state in a large conductor of length L is $1/L$, the current I^+ carried by the $+k$ states can be calculated by [97],

$$I^+ = \frac{e}{L}\sum_k v f^+(E) = \frac{e}{L}\sum_k \frac{1}{\hbar}\frac{\partial E}{\partial k} f^+(E), \tag{A-1}$$

where the velocity $v = (1/\hbar)(\partial E/\partial k)$ is substituted and $f^+(E)$ specifies the electron occupation probability.

Since density of states in k space is $L/2\pi$, the above summation can be converted into integral by the following prescription

$$\sum_k \to 2 \times \frac{L}{2\pi}\int dk, \tag{A-2}$$

where the prefactor 2 accounts for the spin, which results in

$$I^+ = \frac{2e}{h}\int \frac{\partial E}{\partial k} f^+(E)\, dk. \tag{A-3}$$

Now, if an electron with energy E comes in from contact α' with mode n (denoting as $\chi^{\alpha',n}$), it will induce wave function in the device (denoting as $\psi_D^{\alpha',n}(E)$). The probability of the electron going out to another contact α with mode m (denoting as $\chi^{\alpha,m}$) is then,

$$P_{\alpha\alpha'}^{mn}(E) = |\psi_D^{\alpha',n\dagger}(E)\cdot\chi^{\alpha,m}|^2, \tag{A-4}$$

which should carry current according to (A-3) as

$$I_{\alpha\alpha'}^{mn} = \frac{2e}{h}\int P_{\alpha\alpha'}^{mn}(E)\frac{\partial E^{\alpha,m}}{\partial k} f^{\alpha'}(E)\, dk. \tag{A-5}$$

Note that the velocity in contact α is used, and the Fermi function in contact

α' is used. The above equation can be modified to

$$\begin{aligned} I_{\alpha\alpha'}^{mn} &= \frac{2e}{h}\int P_{\alpha\alpha'}^{mn}(E)\,\frac{\partial E^{\alpha,m}/\partial k}{\partial E^{\alpha',n}/\partial k}\left(\partial E^{\alpha',n}/\partial k\right) f^{\alpha'}(E)\,dk \\ &= \frac{2e}{h}\int P_{\alpha\alpha'}^{mn}(E)\,\frac{\partial E^{\alpha,m}/\partial k}{\partial E^{\alpha',n}/\partial k} f^{\alpha'}(E)\,dE \\ &= \frac{2e}{h}\int T_{\alpha\alpha'}^{mn}(E)\, f^{\alpha'}(E)\,dE, \end{aligned} \tag{A-6}$$

where

$$T_{\alpha\alpha'}^{mn}(E) = P_{\alpha\alpha'}^{mn}(E)\,\frac{\partial E^{\alpha,m}/\partial k}{\partial E^{\alpha',n}/\partial k} = P_{\alpha\alpha'}^{mn}(E)\,\frac{k^{\alpha,m}}{k^{\alpha',n}}, \tag{A-7}$$

is the transmission from contact α' with mode n to contact α with mode m. Note that the second equation in (A-7) is only valid for parabolic band structure, i.e., effective mass approximation.

When there are multiple modes in the contacts, we have to take into account all of them to get the current from α' to α, i.e.,

$$I_{\alpha\alpha'} = \frac{2e}{h}\int T_{\alpha\alpha'}(E)\, f^{\alpha'}(E)\,dE, \tag{A-8}$$

where

$$T_{\alpha\alpha'}(E) = \sum_{m,n} T_{\alpha\alpha'}^{mn}(E), \tag{A-9}$$

Similarly, we have the current from α to α', i.e.,

$$I_{\alpha'\alpha} = \frac{2e}{h}\int T_{\alpha'\alpha}(E)\, f^{\alpha}(E)\,dE. \tag{A-10}$$

It can be argued that $T_{\alpha'\alpha}(E) = T_{\alpha\alpha'}(E)$, so the net current from α' to α is,

$$\hat{I}_{\alpha\alpha'} = I_{\alpha\alpha'} - I_{\alpha'\alpha} = \frac{2e}{h}\int T_{\alpha\alpha'}(E)\left[f^{\alpha'}(E) - f^{\alpha}(E)\right]dE. \tag{A-11}$$

Finally, if there are more than two contacts, the above Landauer formula should be generalized to Landauer-Büttiker formula to get the net current of contact α,

$$\hat{I}_{\alpha} = \frac{2e}{h}\sum_{\alpha'\neq\alpha}\int T_{\alpha\alpha'}(E)\left[f^{\alpha'}(E) - f^{\alpha}(E)\right]dE. \tag{A-12}$$

APPENDIX B

INTEGRAL EQUATION FORMULATION

Let us consider an arbitrary quantum device as shown in Fig. B.1.

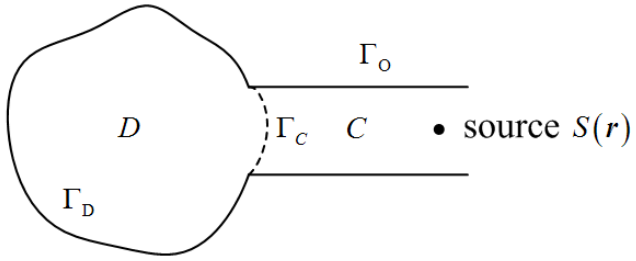


Figure B.1: An arbitrary quantum device and notations.

We let $M^{-1}(\mathbf{r}) = \frac{\hbar^2}{2m(\mathbf{r})}$. Then, the equation in the device region is

$$-\nabla \cdot M^{-1}(\mathbf{r}) \nabla \psi_D(\mathbf{r}) + [V(\mathbf{r}) - E] \psi_D(\mathbf{r}) = 0. \tag{B-1}$$

In the contact region, the governing equation is

$$-\nabla \cdot M^{-1}(\mathbf{r}) \nabla \psi_C(\mathbf{r}) + [V(\mathbf{r}) - E] \psi_C(\mathbf{r}) = S(\mathbf{r}), \tag{B-2}$$

with the boundary conditions that

$$\psi_D = \psi_C \qquad \text{on } \Gamma_C, \tag{B-3}$$

$$M_D^{-1}\mathbf{n} \cdot \nabla \psi_D = M_C^{-1}\mathbf{n} \cdot \nabla \psi_C \qquad \text{on } \Gamma_C, \tag{B-4}$$

$$\psi_D = 0 \qquad \text{on } \Gamma_D, \tag{B-5}$$

$$\psi_C = 0 \qquad \text{on } \Gamma_O, \tag{B-6}$$

and ψ_C satisfies the outgoing wave radiation condition at the far end of the contact.

We define a Green's function for the device region

$$-\nabla \cdot M^{-1}(\mathbf{r}) \nabla g_D(\mathbf{r}, \mathbf{r}') + [V(\mathbf{r}) - E] g_D(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'). \quad \text{(B-7)}$$

Multiplying (B-1) by $g_D(\mathbf{r}, \mathbf{r}')$ from the left, and (B-7) by $\psi_D(\mathbf{r})$, and subtracting to get

$$\begin{aligned} &- g_D(\mathbf{r}, \mathbf{r}') \nabla \cdot M^{-1}(\mathbf{r}) \nabla \psi_D(\mathbf{r}) + \psi_D(\mathbf{r}) \nabla \cdot M^{-1}(\mathbf{r}) \nabla g_D(\mathbf{r}, \mathbf{r}') \\ &= \psi_D(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}'). \end{aligned} \quad \text{(B-8)}$$

Integration (B-8) over V_D, and using Green's theorem, we have

$$\begin{aligned} &- \int_{\Gamma_D + \Gamma_C} dS \mathbf{n} \cdot \left[g_D(\mathbf{r}, \mathbf{r}') M^{-1}(\mathbf{r}) \nabla \psi_D(\mathbf{r}) - \psi_D(\mathbf{r}) M^{-1}(\mathbf{r}) \nabla g_D(\mathbf{r}, \mathbf{r}') \right] \\ &= \begin{cases} \psi_D(\mathbf{r}'), & \mathbf{r}' \in V_D \\ 0, & \mathbf{r}' \notin V_D. \end{cases} \end{aligned} \quad \text{(B-9)}$$

Swapping $\mathbf{r}$ and $\mathbf{r}'$, we have

$$\begin{aligned} &\psi_D(\mathbf{r}) = \int_{\Gamma_D + \Gamma_C} dS' \mathbf{n}' \cdot \\ &\left[-g_D(\mathbf{r}, \mathbf{r}') M^{-1}(\mathbf{r}') \nabla' \psi_D(\mathbf{r}') + \psi_D(\mathbf{r}') M^{-1}(\mathbf{r}') \nabla' g_D(\mathbf{r}, \mathbf{r}') \right], \mathbf{r} \in V_D. \end{aligned} \quad \text{(B-10)}$$

We can pick $g_D(\mathbf{r}, \mathbf{r}')$ such that

$$g_D(\mathbf{r}, \mathbf{r}') = 0, \qquad \mathbf{r} \in \Gamma_D + \Gamma_C. \quad \text{(B-11)}$$

Making use of (B-5) and (B-11), we have

$$\psi_D(\mathbf{r}) = \int_{\Gamma_C} dS' \psi_D(\mathbf{r}') M^{-1}(\mathbf{r}') \mathbf{n}' \cdot \nabla' g_D(\mathbf{r}, \mathbf{r}'). \quad \text{(B-12)}$$

In the contact region, we define a Green's function such that

$$-\nabla \cdot M^{-1}(\mathbf{r}) \nabla g_C(\mathbf{r}, \mathbf{r}') + [V(\mathbf{r}) - E] g_C(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'). \quad \text{(B-13)}$$

By the same token as before, we have

$$-\int\limits_{\Gamma_O+\Gamma_C} dS'\mathbf{n}'\cdot\left[g_C\left(\mathbf{r},\mathbf{r}'\right)M^{-1}\left(\mathbf{r}'\right)\nabla'\psi_C\left(\mathbf{r}'\right)-\psi_C\left(\mathbf{r}'\right)M^{-1}\left(\mathbf{r}'\right)\nabla'g_C\left(\mathbf{r},\mathbf{r}'\right)\right]$$
$$=\int\limits_{V_C} g_C\left(\mathbf{r},\mathbf{r}'\right)S\left(\mathbf{r}'\right)d\mathbf{r}'+\begin{cases}\psi_C\left(\mathbf{r}\right), & \mathbf{r}\in V_C\\ 0, & \mathbf{r}\notin V_C.\end{cases} \tag{B-14}$$

We pick $g_C\left(\mathbf{r},\mathbf{r}'\right)$ such that

$$g_C\left(\mathbf{r},\mathbf{r}'\right)=0 \qquad \mathbf{r}\in\Gamma_{\mathrm{O}}, \tag{B-15}$$

$$\mathbf{n}'\cdot\nabla'g_C\left(\mathbf{r},\mathbf{r}'\right)=0 \qquad \mathbf{r}\in\Gamma_C. \tag{B-16}$$

Making use of (B-6), (B-15) and (B-16), we have

$$-\int\limits_{\Gamma_C} dS'\mathbf{n}'\cdot\nabla'\psi_C\left(\mathbf{r}'\right)g_C\left(\mathbf{r},\mathbf{r}'\right)M^{-1}\left(\mathbf{r}'\right)$$
$$=\int\limits_{V_C} g_C\left(\mathbf{r},\mathbf{r}'\right)S\left(\mathbf{r}'\right)d\mathbf{r}'+\psi_C\left(\mathbf{r}\right),\mathbf{r}\in V_C, \tag{B-17}$$

or

$$\psi_C\left(\mathbf{r}\right)=-\int\limits_{V_C} g_C\left(\mathbf{r},\mathbf{r}'\right)S\left(\mathbf{r}'\right)d\mathbf{r}'-\int\limits_{\Gamma_C} dS'\mathbf{n}'\cdot\nabla'\psi_C\left(\mathbf{r}'\right)M^{-1}\left(\mathbf{r}'\right)g_C\left(\mathbf{r},\mathbf{r}'\right), \tag{B-18}$$

or

$$\psi_C\left(\mathbf{r}\right)=\psi_C^{inc}\left(\mathbf{r}\right)+\psi_C^{ref}\left(\mathbf{r}\right). \tag{B-19}$$

We identify the first term in (B-18) to be the incident wave upon Γ_{C}, while the second term is the reflection from the device region.

Using (B-18) and (B-19) into (B-12), we have the integral equation

$$\psi_D\left(\mathbf{r}\right)=\int\limits_{\Gamma_C} dS'\psi_C^{inc}\left(\mathbf{r}'\right)M^{-1}\left(\mathbf{r}'\right)\mathbf{n}'\cdot\nabla'g_D\left(\mathbf{r},\mathbf{r}'\right)-$$
$$\int\limits_{\Gamma_C} dS'\left[\int\limits_{\Gamma_C} dS''\overbrace{\mathbf{n}''\cdot\nabla''\psi_C\left(\mathbf{r}''\right)}^{-\mathbf{n}''\cdot\nabla''\psi_D(\mathbf{r}'')}M^{-1}\left(\mathbf{r}''\right)g_C\left(\mathbf{r}',\mathbf{r}''\right)\right]M^{-1}\left(\mathbf{r}'\right)\mathbf{n}'\cdot\nabla'g_D\left(\mathbf{r},\mathbf{r}'\right), \tag{B-20}$$

which can be solved by method of moment (MoM).

To get a form similar to (2.16), we can write the above in operator notation

$$\boldsymbol{\psi}_D = \left(\overline{\boldsymbol{\tau}} \cdot \overline{\mathbf{g}}_D^t\right)^t \cdot \overline{\mathbf{M}}^{-1} \cdot \boldsymbol{\psi}_C^{inc} + \left(\overline{\boldsymbol{\tau}} \cdot \overline{\mathbf{g}}_D^t\right)^t \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\mathbf{g}}_C \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\boldsymbol{\tau}} \cdot \boldsymbol{\psi}_D, \quad \text{(B-21)}$$

where

$$\overline{\boldsymbol{\tau}} \Leftrightarrow \mathbf{n}' \cdot \nabla', \quad \text{(B-22)}$$

$$\boldsymbol{\psi}_D \Leftrightarrow \psi_D(\mathbf{r}), \quad \text{(B-23)}$$

$$\overline{\mathbf{g}}_D \Leftrightarrow g_D(\mathbf{r}, \mathbf{r}'), \quad \text{(B-24)}$$

$$\overline{\mathbf{g}}_C \Leftrightarrow g_C(\mathbf{r}, \mathbf{r}'), \quad \text{(B-25)}$$

$$\overline{\mathbf{M}}^{-1} \Leftrightarrow M^{-1}(\mathbf{r}). \quad \text{(B-26)}$$

Eq. (B-21) can be rewritten as

$$\boldsymbol{\psi}_D = \overline{\mathbf{g}}_D \cdot \overline{\boldsymbol{\tau}}^t \cdot \overline{\mathbf{M}}^{-1} \cdot \boldsymbol{\psi}_C^{inc} + \overline{\mathbf{g}}_D \cdot \overline{\boldsymbol{\tau}}^t \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\mathbf{g}}_C \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\boldsymbol{\tau}} \cdot \boldsymbol{\psi}_D, \quad \text{(B-27)}$$

or

$$\overline{\mathbf{g}}_D^{-1} \cdot \boldsymbol{\psi}_D = \overline{\boldsymbol{\tau}}^t \cdot \overline{\mathbf{M}}^{-1} \cdot \boldsymbol{\psi}_C^{inc} + \overline{\boldsymbol{\tau}}^t \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\mathbf{g}}_C \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\boldsymbol{\tau}} \cdot \boldsymbol{\psi}_D, \quad \text{(B-28)}$$

or

$$\left(\overline{\mathbf{g}}_D^{-1} - \overline{\boldsymbol{\tau}}^t \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\mathbf{g}}_C \cdot \overline{\mathbf{M}}^{-1} \cdot \overline{\boldsymbol{\tau}}\right) \boldsymbol{\psi}_D = \overline{\boldsymbol{\tau}}^t \cdot \overline{\mathbf{M}}^{-1} \cdot \boldsymbol{\psi}_C^{inc}. \quad \text{(B-29)}$$

Note that $\overline{\mathbf{g}}_D^{-1} = E\overline{\mathbf{I}} - \overline{\mathbf{H}}$.

APPENDIX C

THE MATRICES IN THE WAVE FUNCTION APPROACH

Consider a typical case where there are two contacts (contact 1 and contact 2) and the directions ξ_α and η_α ($\alpha = 1, 2$) are the same as x and y. A second order central finite difference method (FDM) with the following formulas is applied to discretize the 2D Schrödinger equation (2.1),

$$\begin{aligned}\nabla \cdot \left[\frac{1}{m^*(x,y)}\nabla\psi\right]_{x=x_i,y=y_j} \approx & \frac{1}{\Delta x^2}\left(\frac{\psi_{i+1,j}-\psi_{i,j}}{m^*_{i+1/2,j}} - \frac{\psi_{i,j}-\psi_{i-1,j}}{m^*_{i-1/2,j}}\right) \\ & + \frac{1}{\Delta y^2}\left(\frac{\psi_{i,j+1}-\psi_{i,j}}{m^*_{i,j+1/2}} - \frac{\psi_{i,j}-\psi_{i,j-1}}{m^*_{i,j-1/2}}\right), \quad \text{(C-1)}\end{aligned}$$

where $\psi_{i,j} = \psi(x_i, y_j)$ and $m^*_{i\pm1/2,j\pm1/2} = m^*\left(\left(x_i + x_{i\pm1}\right)/2, \left(y_i + y_{i\pm1}\right)/2\right)$, $i = 1, 2, \cdots, N_x$, $j = 1, 2, \cdots, N_y$. Δx and Δy are the uniform grid spacing in the x and y directions.

It is apparent that when $i = 1$ ($i = N_x$ is similar), the values $\psi_{0,j}$ for $j = 1, 2, \cdots, N_y$ need to be specified. These values are directly obtained from the solution in contact 1 (equation (2.15)) as

$$\begin{aligned}\psi_{0,j} = & -2ia^1_n\chi^1_n(y_j)\sin(k^1_n\Delta x) \\ & + \sum_{m=1}^{N_1}\left(\int \chi^1_m(y)\psi(x_1,y)\,dy\right)\chi^1_m(y_j)\exp\left(ik^1_m\Delta x\right), \\ = & -2ia^1_n\chi^1_n(y_j)\sin(k^1_n\Delta x) \\ & + \sum_{m=1}^{N_1}\left(\sum_{j'=1}^{N_y}\chi^1_m(y_{j'})\,\psi_{1,j'}\Delta y\right)\chi^1_m(y_j)\exp\left(ik^1_m\Delta x\right), \quad \text{(C-2)}\end{aligned}$$

where the integration is replaced by a summation using trapezoid rule. Similarly, the values $\psi_{N_x+1,j}$ for $j = 1, 2, \cdots, N_y$ can be obtained using the solution in contact 2.

Writing the discretized equations with matrix form results in equation

(2.16). The matrix $\overline{\mathbf{H}}$ for the isolated device is

$$\overline{\mathbf{H}} = \begin{pmatrix} \overline{\mathbf{H}}_1 & \overline{\mathbf{T}}_{12}^{\dagger} & \mathbf{0} & \cdots & \mathbf{0} \\ \overline{\mathbf{T}}_{12} & \overline{\mathbf{H}}_2 & \overline{\mathbf{T}}_{23}^{\dagger} & \ddots & \vdots \\ \overline{\mathbf{0}} & \ddots & \ddots & \ddots & \overline{\mathbf{0}} \\ \vdots & \ddots & \overline{\mathbf{T}}_{N-2,N-1} & \overline{\mathbf{H}}_{N-1} & \overline{\mathbf{T}}_{N-1,N}^{\dagger} \\ \mathbf{0} & \cdots & \overline{\mathbf{0}} & \overline{\mathbf{T}}_{N-1,N} & \overline{\mathbf{H}}_N \end{pmatrix}, \tag{C-3}$$

where $\overline{\mathbf{H}}_i$ is the tri-diagonal Hamiltonian matrix for layer i ($i = 1, 2, \cdots, N_x$) and $\overline{\mathbf{T}}_{ij}$ is the diagonal matrix represents the coupling between adjacent layers. The matrix $\overline{\mathbf{S}}$ for self energy is

$$\overline{\mathbf{S}} = \begin{pmatrix} \overline{\mathbf{S}}^1 & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \cdots & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \ddots & \vdots \\ \overline{\mathbf{0}} & \ddots & \ddots & \ddots & \overline{\mathbf{0}} \\ \vdots & \ddots & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \cdots & \overline{\mathbf{0}} & \overline{\mathbf{0}} & \overline{\mathbf{S}}^2 \end{pmatrix}, \tag{C-4}$$

where the non-zero elements are

$$\begin{aligned} \overline{\mathbf{S}}_{j,j'}^{\alpha} = & -\frac{\hbar^2 \Delta y}{2m_{1/2,j}^{*} \Delta x^2} \sum_{m=1}^{N_\alpha} \chi_m^{\alpha}(y_{j'}) \chi_m^{\alpha}(y_j) \exp(ik_m^{\alpha} \Delta x), \\ & \alpha \in \{1,2\}, \text{ and } (j,j') \in \{1, \cdots, N_y\}. \end{aligned} \tag{C-5}$$

The vector $\mathbf{v}$ is

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}^1 \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \text{ or } \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{v}^2 \end{pmatrix}, \tag{C-6}$$

for wave from the left or the right lead, where the non-zero elements are

$$\begin{aligned} \mathbf{v}_j^{\alpha} = & \frac{\hbar^2}{2m_{1/2,j}^{*} \Delta x^2} 2ia_n^{\alpha} \chi_n^{\alpha}(y_j) \sin(k_n^{\alpha} \Delta x), \\ & \alpha \in \{1,2\}, \text{ and } j \in \{1, \cdots, N_y\}. \end{aligned} \tag{C-7}$$

APPENDIX D

DERIVATION OF THE K · P HAMILTONIAN

D.1 Overview

The electronic structures of the semiconductors can be obtained by solving the single-electron Schrödinger equation, i.e.,

$$\hat{H}\psi\left(\mathbf{r}\right)=\left[\frac{\mathbf{p}^{2}}{2m_{0}}+V\left(\mathbf{r}\right)+\frac{\hbar}{4m_{0}^{2}c^{2}}\left(\boldsymbol{\sigma}\times\nabla V\right)\cdot\mathbf{p}\right]\psi\left(\mathbf{r}\right)=E\psi\left(\mathbf{r}\right), \tag{D-1}$$

where $\hat{H}$ is the Hamiltonian, ψ is the electronic wave function, E is the total energy, $\mathbf{p}$ is the momentum operator, $\mathbf{r}$ is the position vector, m_0 is the free electron mass, V is the periodic potential produced by the atoms of the crystal, the third term of $\hat{H}$ is due to spin-orbit coupling where $\boldsymbol{\sigma}$ is a vector with three components consisting of the Pauli spin matrices.

Since the Hamiltonian is translationally invariant in the crystal, the solution of the Schrödinger equation (D-1) is of the following form according to the Bloch theorem,

$$\psi_{n,\mathbf{k}}\left(\mathbf{r}\right)=e^{i\mathbf{k}\cdot\mathbf{r}}u_{n,\mathbf{k}}\left(\mathbf{r}\right). \tag{D-2}$$

where $\mathbf{k}$ is the wave vector, n is the band index, and $u_{n,\mathbf{k}}\left(\mathbf{r}\right)$ is the Bloch lattice function with periodicity equal to lattice vector $\mathbf{R}$. Substituting (D-2) into (D-1), we have the equation for $u_{n,\mathbf{k}}\left(\mathbf{r}\right)$ only,

$$\left(\hat{H}_{0}+\frac{\hbar^{2}k^{2}}{2m_{0}}+\frac{2\hbar\mathbf{k}\cdot\mathbf{p}}{2m_{0}}+\hat{H}_{so}\right)u_{n,\mathbf{k}}\left(\mathbf{r}\right)=E_{n,\mathbf{k}}u_{n,\mathbf{k}}\left(\mathbf{r}\right), \tag{D-3}$$

where,

$$\hat{H}_{0}=\frac{\mathbf{p}^{2}}{2m_{0}}+V\left(\mathbf{r}\right), \tag{D-4}$$

$$\hat{H}_{so}=\frac{\hbar}{4m_{0}^{2}c^{2}}\left(\boldsymbol{\sigma}\times\nabla V\right)\cdot\left(\hbar\mathbf{k}+\mathbf{p}\right). \tag{D-5}$$

Equation (D-3) can be solved by many methods with different kinds of approximations to obtain the band structure $E_{n,\mathbf{k}}$. Here, we use the $\mathbf{k}\cdot\mathbf{p}$ method, which is a useful technique for analyzing the band structure near a particular point $\mathbf{k}_0$.

The basic idea of $\mathbf{k}\cdot\mathbf{p}$ method is to solve (D-3) at $\mathbf{k}_0$, usually the high symmetry Γ point with $\mathbf{k}_0 = 0$, without taking into account spin-orbit coupling,

$$\hat{H}_0 u_{n,0}(\mathbf{r}) = E_{n,0} u_{n,0}(\mathbf{r}). \tag{D-6}$$

As $u_{n,0}(\mathbf{r})$ form a complete basis set, one can expand $u_{n,\mathbf{k}}(\mathbf{r})$ in this basis,

$$u_{n,\mathbf{k}}(\mathbf{r}) = \sum_j c_{n,\mathbf{k},j} u_{j,0}(\mathbf{r}). \tag{D-7}$$

To obtain the coefficient $c_{n,\mathbf{k},j}$, we substitute (D-7) into (D-3) and test both sides with $u^*_{i,0}(\mathbf{r})$, which results in a matrix equation,

$$\sum_j \overline{\mathbf{H}}_{i,j} c_{n,\mathbf{k},j} = E_{n,\mathbf{k}} c_{n,\mathbf{k},i}, \tag{D-8}$$

where the element of the matrix is

$$\overline{\mathbf{H}}_{i,j} = \left(E_{i,0} + \frac{\hbar^2 k^2}{2m_0}\right)\delta(i,j) + \int d\mathbf{r}\, u^*_{i,0}(\mathbf{r}) \left(\frac{2\hbar\mathbf{k}\cdot\mathbf{p}}{2m_0} + \hat{H}_{so}\right) u_{j,0}(\mathbf{r}). \tag{D-9}$$

Equation (D-8) can then be diagonalized to find $E_{n,\mathbf{k}}$. In practice, however, we only use a subset of the basis functions to approximate $u_{n,\mathbf{k}}(\mathbf{r})$ and it results in a matrix of small dimension. Usually, those corresponding to the lowest conduction and highest valence bands are included in this subset since the band structure near the band gap is of great interest. The rest of the bands can be taken into account approximately, which is necessary to produce the correct heavy-hole effective mass. This method is best described by Löwdin's perturbation theory, as will be derived later.

Note that to write down Hamiltonian matrix element (D-9), it is quite common to utilize the symmetry properties of $u_{n,0}(\mathbf{r})$, which will result in a small set of non-zero elements and many of them are equal, greatly reducing the number of parameters required in this model. For most semiconductors of interest, the top of the valence band can be described by three degenerate p-type states with energy $E_{VB,0}$: $u_{p_x,0} = \rho_v(r)\, x$, $u_{p_y,0} = \rho_v(r)\, y$, and $u_{p_z,0} =$

$\rho_v(r)\,z$, where $r=\sqrt{x^2+y^2+z^2}$. This means that they are antisymmetric with respect to a coordinate and symmetric with respect to the others. The bottom of the conduction band (of direct band gap material), however, is described by a non-degenerate symmetric s-type state with energy $E_{CB,0}$: $u_{s,0}=\rho_c(r)$.

D.2 Löwdin's Perturbation Theory

In Löwdin's method, the set of basis functions are grouped into class A that are treated exactly and class B that are treated approximately. With this method, we only need to solve eigenvalue problem restricted to A space instead of the full A+B space, i.e.,

$$\sum_{\alpha\in A}\overline{\mathbf{U}}_{m\alpha}c_\alpha = Ec_m,\quad m\in A, \tag{D-10}$$

since the renormalized matrix $\overline{\mathbf{U}}$ has taken into account the effect of B space,

$$\overline{\mathbf{U}}_{m\alpha}=\overline{\mathbf{H}}_{m\alpha}+\sum_{\beta\in B}\frac{\overline{\mathbf{H}}_{m\beta}\overline{\mathbf{H}}_{\beta\alpha}}{E-\overline{\mathbf{H}}_{\beta\beta}}+\cdots,\quad m,\alpha\in A. \tag{D-11}$$

To derive (D-10) and (D-11), the original eigenvalue problem in the full A+B space, i.e., equation (D-8), can be rewritten as

$$\sum_{n\in A\cup B}\overline{\mathbf{H}}_{mn}c_n = Ec_m,\quad m\in A\cup B, \tag{D-12}$$

which can be slightly reformulated to

$$\sum_{n\in A\cup B,n\neq m}\overline{\mathbf{H}}_{mn}c_n=\left(E-\overline{\mathbf{H}}_{mm}\right)c_m,\quad m\in A\cup B. \tag{D-13}$$

From (D-13), we can have a expression of c_m as

$$c_m=\sum_{\alpha\in A,\alpha\neq m}\frac{\overline{\mathbf{H}}_{m\alpha}}{E-\overline{\mathbf{H}}_{mm}}c_\alpha+\sum_{\beta\in B,\beta\neq m}\frac{\overline{\mathbf{H}}_{m\beta}}{E-\overline{\mathbf{H}}_{mm}}c_\beta,\quad m\in A\cup B, \tag{D-14}$$

from which we can write c_β as,

$$c_\beta = \sum_{\alpha\in A} \frac{\overline{\mathbf{H}}_{\beta\alpha}}{E-\overline{\mathbf{H}}_{\beta\beta}} c_\alpha + \sum_{\gamma\in B,\gamma\neq\beta} \frac{\overline{\mathbf{H}}_{\beta\gamma}}{E-\overline{\mathbf{H}}_{\beta\beta}} c_\gamma, \quad \beta\in B. \tag{D-15}$$

Inserting (D-15) back into (D-14) and repeating the process with the ultimate goal of eliminating all the coefficient of set B, we obtain a chain as follows,

$$c_m = \sum_{\alpha\in A,\alpha\neq m} \frac{\overline{\mathbf{H}}_{m\alpha}}{E-\overline{\mathbf{H}}_{mm}} c_\alpha + \sum_{\beta\in B,\beta\neq m} \frac{\overline{\mathbf{H}}_{m\beta}}{E-\overline{\mathbf{H}}_{mm}} \sum_{\alpha\in A} \frac{\overline{\mathbf{H}}_{\beta\alpha}}{E-\overline{\mathbf{H}}_{\beta\beta}} c_\alpha + \cdots. \tag{D-16}$$

Moving $E-\overline{\mathbf{H}}_{mm}$ to the left hand side, we have

$$\left(E-\overline{\mathbf{H}}_{mm}\right) c_m = \sum_{\alpha\in A,\alpha\neq m} \overline{\mathbf{H}}_{m\alpha} c_\alpha + \sum_{\beta\in B,\beta\neq m} \overline{\mathbf{H}}_{m\beta} \sum_{\alpha\in A} \frac{\overline{\mathbf{H}}_{\beta\alpha}}{E-\overline{\mathbf{H}}_{\beta\beta}} c_\alpha + \cdots. \tag{D-17}$$

Choosing $m\in A$ and moving $\overline{\mathbf{H}}_{mm}c_m$ to the right, we have

$$Ec_m = \sum_{\alpha\in A} \left(\overline{\mathbf{H}}_{m\alpha} + \sum_{\beta\in B} \frac{\overline{\mathbf{H}}_{m\beta}\overline{\mathbf{H}}_{\beta\alpha}}{E-\overline{\mathbf{H}}_{\beta\beta}} + \cdots \right) c_\alpha, \ m\in A, \tag{D-18}$$

which can be written as (D-10) and (D-11).

Note that equations (D-10) and (D-11) need to be solved iteratively to obtain self-consistent value of E. Several approximations are commonly used, (i) the series of (D-11) is truncated after the second term, (ii) the energy E in (D-11) is replaced with an approximated value so that no self-consistency is actually performed.

D.3 One-Band Model

By keeping only one single band in class A, one obtains the single band effective mass dispersion. Take the conduction band s for example, we have only one matrix element according to (D-3),

$$E_{s,\mathbf{k}} = E_{s,0} + \frac{\hbar^2k^2}{2m_0} + \frac{\hbar\mathbf{k}}{m_0}\cdot\langle u_{s,0}|\hat{\pi}|u_{s,0}\rangle + \frac{\hbar^2}{m_0^2}\sum_{i\neq s}\frac{|\mathbf{k}\cdot\langle u_{s,0}|\hat{\pi}|u_{i,0}\rangle|^2}{E_{s,0}-E_{i,0}}, \tag{D-19}$$

where

$$\hat{\pi} = \mathbf{p} + \frac{\hbar}{4m_0c^2}\left(\boldsymbol{\sigma} \times \nabla V\right). \tag{D-20}$$

At the band minima, the 3rd term in (D-19) reduces to zero, since it is linear in $\mathbf{k}$. The 4th term can be decomposed into x, y, z components, thus we have,

$$E_{s,\mathbf{k}} = E_{s,0} + \sum_{\alpha,\beta} \frac{\hbar^2}{2} k_\alpha k_\beta \left(\frac{1}{m_s^*}\right)_{\alpha,\beta}, \quad \alpha, \beta \in \{x, y, z\} \tag{D-21}$$

where the effective mass tensor is,

$$\left(\frac{1}{m_s^*}\right)_{\alpha,\beta} = \left(\frac{1}{m_0}\right)\delta_{\alpha,\beta} + \frac{2}{m_0^2}\sum_{i\neq s} \frac{\langle u_{s,0}|\hat{\pi}_\alpha|u_{i,0}\rangle\langle u_{i,0}|\hat{\pi}_\beta|u_{s,0}\rangle}{E_{s,0} - E_{i,0}}. \tag{D-22}$$

If the three valence bands are included in class B, and spin-orbit coupling is neglected in (D-20), we find that

$$\langle u_{s,0}|\mathbf{p}_\alpha|u_{p_\beta,0}\rangle = -\langle u_{p_\beta,0}|\mathbf{p}_\alpha|u_{s,0}\rangle = \frac{im_0P}{\hbar}\delta_{\alpha,\beta}, \quad \alpha, \beta \in \{x, y, z\}, \tag{D-23}$$

due to symmetry properties of the basis functions as mentioned before and that the momentum operator $\mathbf{p}_\alpha$ is odd under inversion of α. Consequently, the effective mass for the conduction band is isotropic and equal to

$$\left(\frac{1}{m_s^*}\right)_{\alpha,\beta} = \left(\frac{1}{m_0} + \frac{2P^2}{\hbar^2 E_g}\right)\delta_{\alpha,\beta}. \tag{D-24}$$

where $E_g = E_{CB,0} - E_{VB,0}$ is the band gap.

D.4 Three-Band Model

In the three-band model we have three degenerate states in class A: $u_{p_x,0}$, $u_{p_y,0}$, and $u_{p_z,0}$; the rest of the bands are included in class B. We also set $\hat{H}_{so} = 0$ in (D-3), which means that spin-orbit coupling is not considered.

The first order contribution to the matrix element $U_{m\alpha}$ in (D-11) is proportional to,

$$\langle u_{p_i,0}|\mathbf{p}_\alpha|u_{p_j,0}\rangle, \quad \alpha, i, j \in \{x, y, z\}, \tag{D-25}$$

which are evaluated to zero since the momentum operator is odd under in-

version.

The second order contribution to the matrix element is

$$I_{ij} = \sum_{\alpha,\beta\in\{x,y,z\}} \frac{\hbar^2}{m_0^2} k_\alpha k_\beta \sum_{n\neq u_{px,y,z,0}} \frac{\langle u_{p_i,0}|\mathbf{p}_\alpha|n\rangle\langle n|\mathbf{p}_\beta|u_{p_j,0}\rangle}{E_{VB,0}-E_n},\ i,j\in\{x,y,z\}. \tag{D-26}$$

To simplify these summations, we will make use of the symmetry properties,

$$\langle u_{p_x,0}|\mathbf{p}_x|n\rangle = \langle u_{p_y,0}|\mathbf{p}_y|n\rangle = \langle u_{p_z,0}|\mathbf{p}_z|n\rangle, \tag{D-27}$$

and

$$\begin{aligned}\langle u_{p_z,0}|\mathbf{p}_x|n\rangle = \langle u_{p_x,0}|\mathbf{p}_y|n\rangle = \langle u_{p_y,0}|\mathbf{p}_z|n\rangle = \\ \langle u_{p_z,0}|\mathbf{p}_y|n\rangle = \langle u_{p_x,0}|\mathbf{p}_z|n\rangle = \langle u_{p_y,0}|\mathbf{p}_x|n\rangle,\end{aligned} \tag{D-28}$$

It can be shown that the integral $\langle u_{p_i,0}|\mathbf{p}_\alpha|n\rangle\langle n|\mathbf{p}_\beta|u_{p_j,0}\rangle$ is non-zero only when all cartesian coordinates appear in pairs.

For the case when $i=j$, the integral $\langle u_{p_i,0}|\mathbf{p}_\alpha|n\rangle\langle n|\mathbf{p}_\beta|u_{p_i,0}\rangle$ is non-zero only when $\alpha=\beta$, and then I_{ii} reduces to

$$I_{xx} = Lk_x^2 + M\left(k_y^2+k_z^2\right), \tag{D-29}$$

$$I_{yy} = Lk_y^2 + M\left(k_x^2+k_z^2\right), \tag{D-30}$$

$$I_{zz} = Lk_z^2 + M\left(k_x^2+k_y^2\right), \tag{D-31}$$

where we only have two different coefficients,

$$L = \frac{\hbar^2}{m_0^2} \sum_{n\neq u_{px,y,z,0}} \frac{\langle u_{p_x,0}|\mathbf{p}_x|n\rangle\langle n|\mathbf{p}_x|u_{p_x,0}\rangle}{E_{VB,0}-E_n}, \tag{D-32}$$

$$M = \frac{\hbar^2}{m_0^2} \sum_{n\neq u_{px,y,z,0}} \frac{\langle u_{p_x,0}|\mathbf{p}_y|n\rangle\langle n|\mathbf{p}_y|u_{p_x,0}\rangle}{E_{VB,0}-E_n}, \tag{D-33}$$

Similarly, for the case when $i\neq j$, $\langle u_{p_i,0}|\mathbf{p}_\alpha|n\rangle\langle n|\mathbf{p}_\beta|u_{p_j,0}\rangle$ is non-zero when α and β are chosen to pair all the Cartesian coordinates. It turns out that α and β are not determined but up to a permutation. We have, making use of (D-27) and (D-28),

$$I_{xy} = I_{yx} = Nk_xk_y, \tag{D-34}$$

$$I_{xz} = I_{zx} = Nk_xk_z, \tag{D-35}$$

$$I_{yz} = I_{zy} = Nk_yk_z, \tag{D-36}$$

where

$$N = \frac{\hbar^2}{m_0^2} \sum_{n \neq u_{p_{x,y,z},0}} \frac{\langle u_{p_z,0}|\mathbf{p}_z|n\rangle\langle n|\mathbf{p}_x|u_{p_x,0}\rangle + \langle u_{p_z,0}|\mathbf{p}_x|n\rangle\langle n|\mathbf{p}_z|u_{p_x,0}\rangle}{E_{VB,0} - E_n}, \tag{D-37}$$

Therefore the three-band Hamiltonian can be written as,

$$\overline{\mathbf{U}}_{3\times3} = \left(E_{VB,0} + \frac{\hbar^2k^2}{2m_0}\right)\overline{\mathbf{I}}_{3\times3} + \overline{\mathbf{H}}_{DKK}, \tag{D-38}$$

where $\overline{\mathbf{I}}_{3\times3}$ is 3×3 identity matrix, $\overline{\mathbf{H}}_{DKK}$ is the DKK (Dresselhaus-Kip-Kittel) Hamiltonian,

$$\overline{\mathbf{H}}_{DKK} = \tag{D-39}$$
$$\begin{pmatrix} Lk_x^2 + M\left(k_y^2 + k_z^2\right) & Nk_xk_y & Nk_xk_z \\ Nk_yk_x & Lk_y^2 + M\left(k_x^2 + k_z^2\right) & Nk_yk_z \\ Nk_zk_x & Nk_zk_y & Lk_z^2 + M\left(k_x^2 + k_y^2\right) \end{pmatrix}.$$

D.5 Six-Band Model

To generalize the three-band model to six-band model, we include the spin-orbit coupling $\hat{H}_{so}$ in (D-3). The class A is extended to include spin up ones $u_{p_x,0}\uparrow$, $u_{p_y,0}\uparrow$, $u_{p_z,0}\uparrow$, and spin down ones $u_{p_x,0}\downarrow$, $u_{p_y,0}\downarrow$, $u_{p_z,0}\downarrow$. Again, the rest of the bands are put in class B.

First, the spin-orbit Hamiltonian defined in (D-5) can be simplified to

$$\hat{H}_{so} = \frac{\hbar}{4m_0^2c^2}(\boldsymbol{\sigma} \times \nabla V) \cdot \mathbf{p}, \tag{D-40}$$

since the momentum of the electron in its atomic orbit is very much greater than Bloch wave momentum.

It is then decomposed into Cartesian components,

$$\hat{H}_{so} = \frac{\hbar}{4m_0^2c^2}(\nabla V \times \mathbf{p}) \cdot \boldsymbol{\sigma} = \frac{\hbar}{4m_0^2c^2} \times \tag{D-41}$$

$$\left[\left(\frac{\partial V}{\partial y}\mathbf{p}_z - \frac{\partial V}{\partial z}\mathbf{p}_y\right)\sigma_x + \left(\frac{\partial V}{\partial z}\mathbf{p}_x - \frac{\partial V}{\partial x}\mathbf{p}_z\right)\sigma_y + \left(\frac{\partial V}{\partial x}\mathbf{p}_y - \frac{\partial V}{\partial y}\mathbf{p}_x\right)\sigma_z\right].$$

Third, we evaluate the matrix element of (D-41) up to the first order. Again, to have non-zero integral, the coordinates needs to be paired. It turns out that the non-zero integrals are equal, which are

$$\begin{aligned}&\langle u_{p_x,0}|\frac{\partial V}{\partial x}\mathbf{p}_y - \frac{\partial V}{\partial y}\mathbf{p}_x|u_{p_y,0}\rangle = \langle u_{p_y,0}|\frac{\partial V}{\partial y}\mathbf{p}_z - \frac{\partial V}{\partial z}\mathbf{p}_y|u_{p_z,0}\rangle = \\ &\langle u_{p_z,0}|\frac{\partial V}{\partial z}\mathbf{p}_x - \frac{\partial V}{\partial x}\mathbf{p}_z|u_{p_x,0}\rangle = \Delta\frac{4m_0^2c^2}{3i\hbar},\end{aligned} \tag{D-42}$$

and the transposed elements have opposite signs due to the $\mathbf{p}$ operators.

Therefore, the matrix form is

$$\overline{\mathbf{H}}_{so} = \begin{pmatrix} 0 & \sigma_z & -\sigma_y \\ -\sigma_z & 0 & \sigma_x \\ \sigma_y & -\sigma_x & 0 \end{pmatrix}\frac{\Delta}{3i}. \tag{D-43}$$

Inserting into the above the Pauli spin matrix,

$$\sigma_x \to \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y \to \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z \to \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{D-44}$$

we obtain $\overline{\mathbf{H}}_{so}$ in the basis arranged in the order $u_{p_x,0}\uparrow$, $u_{p_y,0}\uparrow$, $u_{p_z,0}\uparrow$, $u_{p_x,0}\downarrow$, $u_{p_y,0}\downarrow$, $u_{p_z,0}\downarrow$,

$$\overline{\mathbf{H}}_{so} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & i \\ -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -i & -1 & 0 \\ 0 & 0 & -i & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ i & -1 & 0 & 0 & 0 & 0 \end{pmatrix}\frac{\Delta}{3i}. \tag{D-45}$$

Finally, the total 6×6 Hamiltonian can be written as,

$$\overline{\mathbf{U}}_{6\times 6} = \left(E_{VB,0} + \frac{\hbar^2k^2}{2m_0}\right)\overline{\mathbf{I}}_{6\times 6} + \begin{pmatrix} \overline{\mathbf{H}}_{DKK} & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{H}}_{DKK} \end{pmatrix} + \overline{\mathbf{H}}_{so}, \tag{D-46}$$

where $\overline{\mathbf{I}}_{6\times 6}$ is 6×6 identity matrix and $\overline{\mathbf{H}}_{DKK}$ is the DKK Hamiltonian

derived in the three-band model.

D.6 Eight-Band Model

The eight-band model can be regarded as a generalization of the six-band model, since now we add in the two conduction bands. So class A consists of spin up ones $u_{s,0}\uparrow$, $u_{p_x,0}\uparrow$, $u_{p_y,0}\uparrow$, $u_{p_z,0}\uparrow$, and spin down ones $u_{s,0}\downarrow$, $u_{p_x,0}\downarrow$, $u_{p_y,0}\downarrow$, $u_{p_z,0}\downarrow$.

From the discussion of six-band model, it is easy to infer that the renormalized matrix $\overline{\mathbf{U}}_{8\times 8}$ takes the following form,

$$\overline{\mathbf{U}}_{8\times 8} = \begin{pmatrix} \overline{\mathbf{H}}_4 & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{H}}_4 \end{pmatrix} + \begin{pmatrix} \overline{\mathbf{H}}_R & \overline{\mathbf{0}} \\ \overline{\mathbf{0}} & \overline{\mathbf{H}}_R \end{pmatrix} + \overline{\mathbf{H}}'_{so}, \tag{D-47}$$

where $\overline{\mathbf{H}}_4$ is the direct $\mathbf{k}\cdot\mathbf{p}$ interaction, $\overline{\mathbf{H}}_R$ is the renormalized part of the $\mathbf{k}\cdot\mathbf{p}$ interaction, and $\overline{\mathbf{H}}'_{so}$ is the spin-orbit interaction part.

Due to (D-23) and (D-25),

$$\overline{\mathbf{H}}_4 = \tag{D-48}$$
$$\begin{pmatrix} E_{CB,0}+\frac{\hbar^2k^2}{2m_0} & ik_xP & ik_yP & ik_zP \\ -ik_xP & E_{VB,0}+\frac{\hbar^2k^2}{2m_0} & 0 & 0 \\ -ik_yP & 0 & E_{VB,0}+\frac{\hbar^2k^2}{2m_0} & 0 \\ -ik_zP & 0 & 0 & E_{VB,0}+\frac{\hbar^2k^2}{2m_0} \end{pmatrix},$$

where $P = -i\frac{\hbar}{m_0}\langle u_{s,0}|\mathbf{p}_x|u_{p_x,0}\rangle$ is the optical matrix element.

Similar to $\overline{\mathbf{H}}_{DKK}$ Hamiltonian, we have

$$\overline{\mathbf{H}}_R = \tag{D-49}$$
$$\begin{pmatrix} Ak^2 & Bk_yk_z & Bk_xk_z & Bk_xk_y \\ Bk_yk_z & Lk_x^2+M\left(k_y^2+k_z^2\right) & Nk_xk_y & Nk_xk_z \\ Bk_zk_x & Nk_xk_y & Lk_y^2+M\left(k_x^2+k_z^2\right) & Nk_yk_z \\ Bk_xk_y & Nk_xk_z & Nk_yk_z & Lk_z^2+M\left(k_x^2+k_y^2\right) \end{pmatrix},$$

where the parameters A, B, L, M, and N are expressed in terms of $\mathbf{k}\cdot\mathbf{p}$ perturbation sums over all bands other than the eight we consider. Note that L, M, and N are closely related to those in the DKK Hamiltonian

(D-39), the only difference is that the summation here does not include the two conduction bands, which are now treated exactly.

For parameter A, the non-zero integrals are

$$\langle u_{s,0}|\mathbf{p}_x|n\rangle\langle n|\mathbf{p}_x|u_{s,0}\rangle = \langle u_{s,0}|\mathbf{p}_y|n\rangle\langle n|\mathbf{p}_y|u_{s,0}\rangle = \langle u_{s,0}|\mathbf{p}_z|n\rangle\langle n|\mathbf{p}_z|u_{s,0}\rangle, \tag{D-50}$$

therefore

$$A = \frac{\hbar^2}{m_0^2} \sum_{n \neq u_{s,0}, u_{px,y,z,0}} \frac{\langle u_{s,0}|\mathbf{p}_x|n\rangle\langle n|\mathbf{p}_x|u_{s,0}\rangle}{E_{CB,0} - E_n}. \tag{D-51}$$

For parameter B, the non-zero integrals are

$$\langle u_{s,0}|\mathbf{p}_x|n\rangle\langle n|\mathbf{p}_y|u_{p_z,0}\rangle = \langle u_{s,0}|\mathbf{p}_y|n\rangle\langle n|\mathbf{p}_z|u_{p_x,0}\rangle = \langle u_{s,0}|\mathbf{p}_z|n\rangle\langle n|\mathbf{p}_x|u_{p_z,0}\rangle, \tag{D-52}$$

therefore, with E replaced by $\left(E_{CB,0} + E_{VB,0}\right)/2$, we have

$$B = \frac{2\hbar^2}{m_0^2} \sum_{n \neq u_{s,0}, u_{px,y,z,0}} \frac{\langle u_{s,0}|\mathbf{p}_x|n\rangle\langle n|\mathbf{p}_y|u_{p_z,0}\rangle}{\left(E_{CB,0} + E_{VB,0}\right)/2 - E_n}. \tag{D-53}$$

$\overline{\mathbf{H}}'_{so}$ is $\overline{\mathbf{H}}_{so}$ plus two more rows and columns for the conduction bands,

$$\overline{\mathbf{H}}'_{so} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & i \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -i & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -i & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & i & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \frac{\Delta}{3i}, \tag{D-54}$$

all added elements are zero due to symmetry.

For more details please refer to Ref. [19,20,73,119], where the above derivation is extracted.

APPENDIX E

DISCRETIZATION OF THE K · P HAMILTONIAN IN THE FOURIER SPACE

Discretization of the $\mathbf{k}\cdot\mathbf{p}$ operator usually results in very complicated forms, in particular when the number of bands is large, since it involves various differential operators. Therefore, it is very useful to have a simplified discretization form that is valid for arbitrary nanowire orientation.

Take the eight-band $\mathbf{k}\cdot\mathbf{p}$ operator in (4.7) as an example, it can be rewritten as [75],

$$\overline{\mathbf{H}}^8\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)=\overline{\mathbf{H}}_0+\overline{\mathbf{H}}_{so}+\overline{\mathbf{H}}_x\frac{\partial}{\partial x}+\overline{\mathbf{H}}_y\frac{\partial}{\partial y}+\overline{\mathbf{H}}_z\frac{\partial}{\partial z}-\overline{\mathbf{H}}_{xx}\frac{\partial^2}{\partial x^2}$$
$$-\overline{\mathbf{H}}_{yy}\frac{\partial^2}{\partial y^2}-\overline{\mathbf{H}}_{zz}\frac{\partial^2}{\partial z^2}-\overline{\mathbf{H}}_{xy}\frac{\partial^2}{\partial x\partial y}-\overline{\mathbf{H}}_{yz}\frac{\partial^2}{\partial y\partial z}-\overline{\mathbf{H}}_{zx}\frac{\partial^2}{\partial z\partial x}, \tag{E-1}$$

where the matrices $\overline{\mathbf{H}}_0$, $\overline{\mathbf{H}}_{so}$, $\overline{\mathbf{H}}_x$, $\overline{\mathbf{H}}_y$, $\overline{\mathbf{H}}_z$, $\overline{\mathbf{H}}_{xx}$, $\overline{\mathbf{H}}_{yy}$, $\overline{\mathbf{H}}_{zz}$, $\overline{\mathbf{H}}_{xy}$, $\overline{\mathbf{H}}_{yz}$, and $\overline{\mathbf{H}}_{zx}$ are the coefficients containing contributions from Löwdin's renormalization and spin-orbit interactions.

For nanowire directions other than [100], coordinate transformations have to be performed. For example, to obtain the coordinate for [110] direction, we rotate the coordinate of [100] direction in xy plane by $\phi=\pi/4$, therefore the components of $\mathbf{k}$ of the two systems are related by,

$$\begin{pmatrix}k_x\\k_y\\k_z\end{pmatrix}=\begin{pmatrix}\cos\phi&-\sin\phi&0\\\sin\phi&\cos\phi&0\\0&0&1\end{pmatrix}\begin{pmatrix}k'_x\\k'_y\\k'_z\end{pmatrix}. \tag{E-2}$$

Similarly, for [111] direction, we continue rotating the coordinate of [110] in yz plane by θ with $\sin\theta=1/\sqrt{3}$, therefore,

$$\begin{pmatrix}k_x\\k_y\\k_z\end{pmatrix}=\begin{pmatrix}\cos\phi&-\sin\phi&0\\\sin\phi&\cos\phi&0\\0&0&1\end{pmatrix}\begin{pmatrix}\cos\theta&0&-\sin\theta\\0&1&0\\\sin\theta&0&\cos\theta\end{pmatrix}\begin{pmatrix}k'_x\\k'_y\\k'_z\end{pmatrix}. \tag{E-3}$$

Once we have $\mathbf{k}$ in terms of $\mathbf{k}'$, we plug them into the Hamiltonian expression (4.1) to obtain the new Hamiltonian (now in $\mathbf{k}'$). It turns out that the new Hamiltonian operator can still be written in the form of (E-1), the only difference is that now we have different coefficient matrices $\overline{\mathbf{H}}_i$ and $\overline{\mathbf{H}}_{ij}$ $(i, j = x, y, z)$.

To discretize the above operator (E-1), the transversal components are expanded using Fourier series [51], i.e.,

$$\phi_{p,q}(y_m, z_n) = \frac{2}{\sqrt{N_y N_z}} \sin(k_p y_m) \sin(k_q z_n), \tag{E-4}$$

where N_y and N_z are the number of grid points in the y and z directions respectively, m and n $(1 \leq m \leq N_y, 1 \leq n \leq N_z)$ are the coordinates of the Rth grid point in real space, p and q $(1 \leq p \leq N_y, 1 \leq q \leq N_z)$ are the coordinates of the Sth grid point in the Fourier space,

$$k_p = \frac{p\pi}{L_y}, \quad k_q = \frac{q\pi}{L_z}, \tag{E-5}$$

where L_y and L_z are the nanowire length in the y and z directions.

The discretization is done by operating (E-1) on (E-4), multiplying the result with (E-4), and performing integrations. While the longitudinal component of the unknown envelope function is discretized with second-order central finite difference method. The discretized Hamiltonian for the nanowires will have the following format,

$$\overline{\mathbf{H}} = \begin{pmatrix} \overline{\mathbf{D}}_1 & \overline{\mathbf{T}} & & & \\ \overline{\mathbf{T}}^\dagger & \overline{\mathbf{D}}_2 & \overline{\mathbf{T}} & & \\ & \ddots & \ddots & \ddots & \\ & & \overline{\mathbf{T}}^\dagger & \overline{\mathbf{D}}_{N_x-1} & \overline{\mathbf{T}} \\ & & & \overline{\mathbf{T}}^\dagger & \overline{\mathbf{D}}_{N_x} \end{pmatrix}, \tag{E-6}$$

where $\overline{\mathbf{D}}_i$ is the on-site Hamiltonian for layer i $(1 \leq i \leq N_x$, N_x is the number of grid points in the longitudinal direction $x)$ and $\overline{\mathbf{T}}$ is the coupling Hamiltonian between adjacent layers.

It is found that the (S, S') block of $\overline{\mathbf{D}}_i$ (excluding the potential term) can

be written down using very simple prescription,

$$\begin{aligned}\overline{\mathbf{D}}_{S,S'} = &\left(\overline{\mathbf{H}}_0 + \overline{\mathbf{H}}_{xx}\frac{2}{(\Delta x)^2} + \overline{\mathbf{H}}_{yy}k_p^2 + \overline{\mathbf{H}}_{zz}k_q^2 + \overline{\mathbf{H}}_{so}\right)\delta_{p,p'}\delta_{q,q'} \\ &+ \left(\overline{\mathbf{H}}_y\frac{4k'_p}{\pi}\frac{p}{p^2-p'^2}\right)\delta_{p+p',odd}\delta_{q,q'} \\ &+ \left(\overline{\mathbf{H}}_z\frac{4k'_q}{\pi}\frac{q}{q^2-q'^2}\right)\delta_{q+q',odd}\delta_{p,p'} \\ &- \left(\overline{\mathbf{H}}_{yz}\frac{4k'_p}{\pi}\frac{p}{p^2-p'^2}\frac{4k'_q}{\pi}\frac{q}{q^2-q'^2}\right)\delta_{p+p',odd}\delta_{q+q',odd}, \end{aligned} \tag{E-7}$$

where Δx is the grid spacing in FDM, (p, q) and (p', q') are the coordinates of the Sth and S'th grid points respectively, and δ is Kronecker delta function. For instance, $\delta_{q+q',odd}$ is equal to 1 (0) if $q + q'$ is an odd (even) number.

Similarly, the (S, S') block of $\overline{\mathbf{T}}$ can be written as,

$$\begin{aligned}\overline{\mathbf{T}}_{S,S'} = &\left(-\overline{\mathbf{H}}_{xx}\frac{1}{(\Delta x)^2} + \overline{\mathbf{H}}_x\frac{1}{2\Delta x}\right)\delta_{p,p'}\delta_{q,q'} \\ &- \left(\overline{\mathbf{H}}_{xy}\frac{1}{2\Delta x}\frac{4k'_p}{\pi}\frac{p}{p^2-p'^2}\right)\delta_{p+p',odd}\delta_{q,q'} \\ &- \left(\overline{\mathbf{H}}_{xz}\frac{1}{2\Delta x}\frac{4k'_q}{\pi}\frac{q}{q^2-q'^2}\right)\delta_{q+q',odd}\delta_{p,p'}. \end{aligned} \tag{E-8}$$

Finally, the scheme in Ref. [51] can be adopted to index the grid points in the Fourier space, and the size of matrices $\overline{\mathbf{D}}_i$ and $\overline{\mathbf{T}}$ can be greatly reduced. In Chapter 4, $\Delta x = 0.125$nm is used and $1 \leq S \leq 183$.

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, p. 114, 1965.

[2] H.-S. Wong, "Beyond the conventional transistor," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 133–168, 2002.

[3] "ITRS." [Online]. Available: http://www.itrs.net/

[4] P. Avouris, "Graphene: Electronic and photonic properties and devices," *Nano Letters*, vol. 10, no. 11, pp. 4285–4294, 2010.

[5] J. A. Del Alamo, "Nanometre-scale electronics with III-V compound semiconductors," *Nature*, vol. 479, no. 7373, pp. 317–323, 2011.

[6] I. Ferain, C. A. Colinge, and J.-P. Colinge, "Multigate transistors as the future of classical metal-oxide-semiconductor field-effect transistors," *Nature*, vol. 479, no. 7373, pp. 310–316, 2011.

[7] M. Lundstrom and J. Guo, *Nanoscale Transistors: Device Physics, Modeling and Simulation*. Springer New York, 2006.

[8] D. Vasileska and S. M. Goodnick, *Computational Electronics*. San Rafael, CA: Morgan & Claypool, 2006.

[9] D. Vasileska, D. Mamaluy, I. Knezevic, H. Khan, S. Goodnick, and H. Nalwa, "Quantum transport in nanoscale devices," *Encyclopedia of Nanoscience and Nanotechnology*, American Scientific Publishers, Syracuse, 2010.

[10] T. Sollner, W. Goodhue, P. Tannenwald, C. Parker, and D. Peck, "Resonant tunneling through quantum wells at frequencies up to 2.5 THz," *Applied Physics Letters*, vol. 43, p. 588, 1983.

[11] H. Kisaki, "Tunnel transistor," *Proceedings of the IEEE*, vol. 61, no. 7, pp. 1053–1054, 1973.

[12] S. Datta, "Nanoscale device modeling: The Green's function method," *Superlattices and Microstructures*, vol. 28, no. 4, pp. 253–278, 2000.

[13] J. Wang and H. Guo, "Relation between nonequilibrium Green's function and Lippmann-Schwinger formalism in the first-principles quantum transport theory," *Physical Review B*, vol. 79, no. 4, p. 045119, Jan 2009.

[14] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, 1989.

[15] M. Lundstrom, *Fundamentals of Carrier Transport*. Cambridge University Press, 2000.

[16] Z. Ren, "Nanoscale MOSFETs: Physics, simulation and design," Ph.D. dissertation, Purdue University, 2001.

[17] K. Nehari, N. Cavassilas, J. Autran, M. Bescond, D. Munteanu, and M. Lannoo, "Influence of band structure on electron ballistic transport in silicon nanowire MOSFETs: An atomistic study," *Solid-State Electronics*, vol. 50, no. 4, pp. 716–721, 2006.

[18] J. Wang, A. Rahman, A. Ghosh, G. Klimeck, and M. Lundstrom, "On the validity of the parabolic effective-mass approximation for the I-V calculation of silicon nanowire transistors," *Electron Devices, IEEE Transactions on*, vol. 52, no. 7, pp. 1589–1595, 2005.

[19] L. C. L. Y. Voon and M. Willatzen, *The $k \cdot p$ Method: Electronic Properties of Semiconductors*. Springer, 2009.

[20] S. Chuang, *Physics of Optoelectronic Devices*, New York, NY, USA: Wiley, 1995.

[21] S. Richard, F. Aniel, and G. Fishman, "Energy-band structure of Ge, Si, and GaAs: A thirty-band $k \cdot p$ method," *Physical Review B*, vol. 70, no. 23, p. 235204, 2004.

[22] N. Neophytou, "Quantum and atomistic effects in nanoelectronic transport devices," Ph.D. dissertation, Purdue University, 2008.

[23] T. B. Boykin, G. Klimeck, and F. Oyafuso, "Valence band effective-mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization," *Physical Review B*, vol. 69, no. 11, p. 115201, 2004.

[24] M. Anantram, M. S. Lundstrom, and D. E. Nikonov, "Modeling of nanoscale devices," *Proceedings of the IEEE*, vol. 96, no. 9, pp. 1511–1550, 2008.

[25] S. Barraud, M. Berthome, R. Coquand, M. Cassé, T. Ernst, M.-P. Samson, P. Perreau, K. Bourdelle, O. Faynot, and T. Poiroux, "Scaling of trigate junctionless nanowire MOSFET with gate length down to 13 nm," *Electron Device Letters, IEEE*, vol. 33, no. 9, pp. 1225–1227, 2012.

[26] J. Taylor, H. Guo, and J. Wang, "*Ab initio* modeling of quantum transport properties of molecular electronic devices," *Physical Review B*, vol. 63, no. 24, p. 245407, 2001.

[27] C. S. Lent and D. J. Kirkner, "The quantum transmitting boundary method," *Journal of Applied Physics*, vol. 67, no. 10, pp. 6353–6359, 1990.

[28] E. Polizzi and N. Ben Abdallah, "Subband decomposition approach for the simulation of quantum electron transport in nanostructures," *Journal of Computational Physics*, vol. 202, no. 1, pp. 150–180, 2005.

[29] C. Cheng, J.-H. Lee, K. H. Lim, H. Z. Massoud, and Q. H. Liu, "3D quantum transport solver based on the perfectly matched layer and spectral element methods for the simulation of semiconductor nanodevices," *Journal of Computational Physics*, vol. 227, no. 1, pp. 455–471, 2007.

[30] J. Wang, E. Polizzi, and M. Lundstrom, "A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation," *Journal of Applied Physics*, vol. 96, no. 4, pp. 2192–2203, 2004.

[31] P. Havu, V. Havu, M. Puska, and R. Nieminen, "Nonequilibrium electron transport in two-dimensional nanostructures modeled using Green's functions and the finite-element method," *Physical Review B*, vol. 69, no. 11, p. 115325, 2004.

[32] H. Jiang, S. Shao, W. Cai, and P. Zhang, "Boundary treatments in non-equilibrium greens function (NEGF) methods for quantum transport in nano-MOSFETs," *Journal of Computational Physics*, vol. 227, no. 13, pp. 6553–6573, 2008.

[33] J. Torres and J. Sáenz, "Improved generalized scattering matrix method: Conduction through ballistic nanowires," *Journal of the Physics Society Japan*, vol. 73, no. 8, pp. 2182–2193, 2004.

[34] A. Svizhenko, M. Anantram, T. Govindan, B. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors," *Journal of Applied Physics*, vol. 91, no. 4, pp. 2343–2354, 2002.

[35] D. Mamaluy, M. Sabathil, and P. Vogl, "Efficient method for the calculation of ballistic quantum transport," *Journal of Applied Physics*, vol. 93, no. 8, pp. 4628–4633, 2003.

[36] D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl, "Contact block reduction method for ballistic transport and carrier densities of open nanostructures," *Physical Review B*, vol. 71, no. 24, p. 245321, 2005.

[37] G. Mil'nikov, N. Mori, Y. Kamakura, and T. Ezaki, "R-matrix theory of quantum transport and recursive propagation method for device simulations," *Journal of Applied Physics*, vol. 104, no. 4, p. 044506, 2008.

[38] G. Mil'Nikov, N. Mori, and Y. Kamakura, "R-matrix method for quantum transport simulations in discrete systems," *Physical Review B*, vol. 79, no. 23, p. 235337, 2009.

[39] G. Mil'nikov, N. Mori, and Y. Kamakura, "Application of the R-matrix method in quantum transport simulations," *Journal of Computational Electronics*, vol. 9, no. 3-4, pp. 256–261, 2010.

[40] W. C. Chew, E. Michielssen, J. Song, and J. Jin, *Fast and Efficient Algorithms in Computational Electromagnetics*. Norwood, MA: Artech House, 2001.

[41] S. Datta, *Quantum Transport: Atom to Transistor*. Cambridge University Press, 2005.

[42] J.-P. Colinge, *FinFETs and Other Multi-Gate Transistors*. Springer, 2008.

[43] Z. Ren, S. Goasguen, A. Matsudaira, S. S. Ahmed, K. Cantley, Y. Liu, M. Lundstrom, and X. Wang, "NanoMOS," May 2006. [Online]. Available: https://nanohub.org/resources/1305

[44] N. Neophytou, A. Paul, M. S. Lundstrom, and G. Klimeck, "Simulations of nanowire transistors: Atomistic vs. effective mass models," *Journal of Computational Electronics*, vol. 7, no. 3, pp. 363–366, 2008.

[45] J. Wang, E. Polizzi, A. Ghosh, S. Datta, and M. Lundstrom, "Theoretical investigation of surface roughness scattering in silicon nanowire transistors," *Applied Physics Letters*, vol. 87, no. 4, p. 043101, 2005.

[46] A. Martinez, M. Bescond, J. R. Barker, A. Svizhenko, M. Anantram, C. Millar, and A. Asenov, "A self-consistent full 3-D real-space NEGF simulator for studying nonperturbative effects in nano-MOSFETs," *Electron Devices, IEEE Transactions on*, vol. 54, no. 9, pp. 2213–2222, 2007.

[47] N. Seoane, A. Martinez, A. R. Brown, J. R. Barker, and A. Asenov, "Current variability in Si nanowire MOSFETs due to random dopants in the source/drain regions: A fully 3-D NEGF simulation study," *Electron Devices, IEEE Transactions on*, vol. 56, no. 7, pp. 1388–1395, 2009.

[48] S. Poli, M. G. Pala, and T. Poiroux, "Full quantum treatment of remote Coulomb scattering in silicon nanowire FETs," *Electron Devices, IEEE Transactions on*, vol. 56, no. 6, pp. 1191–1198, 2009.

[49] S. Goodnick, D. Ferry, C. Wilmsen, Z. Liliental, D. Fathy, and O. Krivanek, "Surface roughness at the Si (100)-SiO_2 interface," *Physical Review B*, vol. 32, no. 12, p. 8171, 1985.

[50] M. Luisier, A. Schenk, and W. Fichtner, "Quantum transport in two-and three-dimensional nanoscale transistors: Coupled mode effects in the nonequilibrium Green's function formalism," *Journal of Applied Physics*, vol. 100, no. 4, p. 043713, 2006.

[51] M. Shin, "Full-quantum simulation of hole transport and band-to-band tunneling in nanowires using the $k \cdot p$ method," *Journal of Applied Physics*, vol. 106, no. 5, p. 054505, 2009.

[52] R. Grassi, A. Gnudi, E. Gnani, S. Reggiani, and G. Baccarani, "Mode space approach for tight binding transport simulation in graphene nanoribbon FETs," *Nanotechnology, IEEE Transactions on*, vol. 10, no. 3, pp. 371–378, 2011.

[53] H. Ryu, H.-H. Park, M. Shin, D. Vasileska, and G. Klimeck, "Feasibility, accuracy, and performance of contact block reduction method for multi-band simulations of ballistic quantum transport," *Journal of Applied Physics*, vol. 111, no. 6, p. 063705, 2012.

[54] J. Guo, S. Datta, M. Lundstrom, and M. Anantam, "Toward multiscale modeling of carbon nanotube transistors," *International Journal for Multiscale Computational Engineering*, vol. 2, no. 2, pp. 257–276, 2004.

[55] G. Fiori, G. Iannaccone, and G. Klimeck, "Coupled mode space approach for the simulation of realistic carbon nanotube field-effect transistors," *Nanotechnology, IEEE Transactions on*, vol. 6, no. 4, pp. 475–480, 2007.

[56] P. Zhao and J. Guo, "Modeling edge effects in graphene nanoribbon field-effect transistors with real and mode space methods," *Journal of Applied Physics*, vol. 105, no. 3, p. 034503, 2009.

[57] G. Mil'nikov, N. Mori, and Y. Kamakura, "Equivalent transport models in atomistic quantum wires," *Physical Review B*, vol. 85, no. 3, p. 035317, 2012.

[58] Y. He, L. Zeng, T. Kubis, M. Povolotskyi, and G. Klimeck, "Efficient solution algorithm of non-equilibrium Green's functions in atomistic tight binding representation," in *Proc. 15th Int. Workshop Comput. Electron*, 2012, pp. 1–3.

[59] C. Scheiber, A. Schultschik, O. Biro, and R. Dyczij-Edlinger, "A model order reduction method for efficient band structure calculations of photonic crystals," *Magnetics, IEEE Transactions on*, vol. 47, no. 5, pp. 1534–1537, 2011.

[60] J. Z. Huang, W. C. Chew, M. Tang, and L. Jiang, "Efficient simulation and analysis of quantum ballistic transport in nanodevices with AWE," *Electron Devices, IEEE Transactions on*, vol. 59, no. 2, pp. 468–476, 2012.

[61] C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, and J.-P. Colinge, "Junctionless multigate field-effect transistor," *Applied Physics Letters*, vol. 94, no. 5, p. 053511, 2009.

[62] J.-P. Colinge, C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White et al., "Nanowire transistors without junctions," *Nature Nanotechnology*, vol. 5, no. 3, pp. 225–229, 2010.

[63] C.-W. Lee, I. Ferain, A. Afzalian, R. Yan, N. D. Akhavan, P. Razavi, and J.-P. Colinge, "Performance estimation of junctionless multigate transistors," *Solid-State Electronics*, vol. 54, no. 2, pp. 97–103, 2010.

[64] S.-J. Choi, D.-I. Moon, S. Kim, J. P. Duarte, and Y.-K. Choi, "Sensitivity of threshold voltage to nanowire width variation in junctionless transistors," *Electron Device Letters, IEEE*, vol. 32, no. 2, pp. 125–127, 2011.

[65] J. Colinge, A. Kranti, R. Yan, C. Lee, I. Ferain, R. Yu, N. Dehdashti Akhavan, and P. Razavi, "Junctionless nanowire transistor (JNT): Properties and design guidelines," *Solid-State Electronics*, vol. 65, pp. 33–37, 2011.

[66] R. T. Doria, M. A. Pavanello, R. D. Trevisoli, M. de Souza, C.-W. Lee, I. Ferain, N. D. Akhavan, R. Yan, P. Razavi, R. Yu et al., "Junctionless multiple-gate transistors for analog applications," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2511–2519, 2011.

[67] G. Leung and C. O. Chui, "Variability of inversion-mode and junctionless FinFETs due to line edge roughness," *Electron Device Letters, IEEE*, vol. 32, no. 11, pp. 1489–1491, 2011.

[68] G. Leung and C. O. Chui, "Variability impact of random dopant fluctuation on nanoscale junctionless FinFETs," *Electron Device Letters, IEEE*, vol. 33, no. 6, pp. 767–769, 2012.

[69] L. Ansari, B. Feldman, G. Fagas, J.-P. Colinge, and J. C. Greer, "Simulation of junctionless Si nanowire transistors with 3 nm gate length," *Applied Physics Letters*, vol. 97, no. 6, p. 062105, 2010.

[70] N. Dehdashti Akhavan, I. Ferain, P. Razavi, R. Yu, and J.-P. Colinge, "Improvement of carrier ballisticity in junctionless nanowire transistors," *Applied Physics Letters*, vol. 98, no. 10, p. 103510, 2011.

[71] M. Aldegunde, A. Martinez, and J. R. Barker, "Study of discrete doping-induced variability in junctionless nanowire MOSFETs using dissipative quantum transport simulations," *Electron Device Letters, IEEE*, vol. 33, no. 2, pp. 194–196, 2012.

[72] P. Razavi, G. Fagas, I. Ferain, R. Yu, S. Das, and J.-P. Colinge, "Influence of channel material properties on performance of nanowire transistors," *Journal of Applied Physics*, vol. 111, no. 12, p. 124509, 2012.

[73] C. Galeriu, "$k \cdot p$ theory of semiconductor nanostructures," Ph.D. dissertation, Worcester Polytechnic Institute, 2005.

[74] D. Gershoni, C. Henry, and G. Baraff, "Calculating the optical properties of multidimensional heterostructures: Application to the modeling of quaternary quantum well lasers," *Quantum Electronics, IEEE Journal of*, vol. 29, no. 9, pp. 2433–2450, 1993.

[75] M. A. Khayer and R. K. Lake, "Modeling and performance analysis of GaN nanowire field-effect transistors and band-to-band tunneling field-effect transistors," *Journal of Applied Physics*, vol. 108, no. 10, p. 104503, 2010.

[76] J. Z. Huang, W. C. Chew, Y. Wu, and L. J. Jiang, "Methods for fast evaluation of self-energy matrices in tight-binding modeling of electron transport systems," *Journal of Applied Physics*, vol. 112, no. 1, p. 013711, 2012.

[77] H. H. B. Sørensen, P. C. Hansen, D. E. Petersen, S. Skelboe, and K. Stokbro, "Krylov subspace method for evaluating the self-energy matrices in electron transport calculations," *Physical Review B*, vol. 77, no. 15, p. 155301, 2008.

[78] M. Shin, S. Lee, and G. Klimeck, "Computational study on the performance of Si nanowire pMOSFETs based on the $k \cdot p$ method," *Electron Devices, IEEE Transactions on*, vol. 57, no. 9, pp. 2274–2283, 2010.

[79] N. Cavassilas, N. Pons, F. Michelini, and M. Bescond, "Multiband quantum transport simulations of ultimate p-type double-gate transistors: Influence of the channel orientation," *Applied Physics Letters*, vol. 96, no. 10, p. 102102, 2010.

[80] M. Kobayashi and T. Hiramoto, "Experimental study on quantum confinement effects in silicon nanowire metal-oxide-semiconductor field-effect transistors and single-electron transistors," *Journal of Applied Physics*, vol. 103, no. 5, p. 053709, 2008.

[81] M. T. Björk, H. Schmid, J. Knoch, H. Riel, and W. Riess, "Donor deactivation in silicon nanostructures," *Nature Nanotechnology*, vol. 4, no. 2, pp. 103–107, 2009.

[82] J. Deng and H.-S. Wong, "A compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application–Part I: Model of the intrinsic channel region," *Electron Devices, IEEE Transactions on*, vol. 54, no. 12, pp. 3186–3194, 2007.

[83] D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*. McGraw-Hill New York, 2003.

[84] J. Appenzeller, Y.-M. Lin, J. Knoch, Z. Chen, and P. Avouris, "Comparing carbon nanotube transistors–The ideal choice: A novel tunneling device design," *Electron Devices, IEEE Transactions on*, vol. 52, no. 12, pp. 2568–2576, 2005.

[85] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, no. 7373, pp. 329–337, 2011.

[86] M. Luisier and G. Klimeck, "Atomistic full-band design study of InAs band-to-band tunneling field-effect transistors," *Electron Device Letters, IEEE*, vol. 30, no. 6, pp. 602–604, 2009.

[87] J. Z. Huang, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen, "Model order reduction for multiband quantum transport simulations and its application to p-type junctionless transistors," *Electron Devices, IEEE Transactions on*, vol. 60, no. 7, pp. 2111–2119, 2013.

[88] P. Enders and M. Woerner, "Exact block diagonalization of the eight-band Hamiltonian matrix for tetrahedral semiconductors and its application to strained quantum wells," *Semiconductor Science and Technology*, vol. 11, no. 7, p. 983, 1996.

[89] R. G. Veprek, S. Steiger, and B. Witzigmann, "Ellipticity and the spurious solution problem of $k \cdot p$ envelope equations," *Physical Review B*, vol. 76, no. 16, p. 165320, 2007.

[90] A. Paussa, F. Conzatti, D. Breda, R. Vermiglio, D. Esseni, and P. Palestri, "Pseudospectral methods for the efficient simulation of quantization effects in nanoscale MOS transistors," *Electron Devices, IEEE Transactions on*, vol. 57, no. 12, pp. 3239–3249, 2010.

[91] F. Conzatti, M. Pala, D. Esseni, E. Bano, and L. Selmi, "A simulation study of strain induced performance enhancements in InAs nanowire tunnel-FETs," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011, pp. 5.2.1–5.2.4.

[92] R. Jhaveri, V. Nagavarapu, and J. C. Woo, "Effect of pocket doping and annealing schemes on the source-pocket tunnel field-effect transistor," *Electron Devices, IEEE Transactions on*, vol. 58, no. 1, pp. 80–86, 2011.

[93] V. Nagavarapu, R. Jhaveri, and J. C. Woo, "The tunnel source (PNPN) n-MOSFET: A novel high performance transistor," *Electron Devices, IEEE Transactions on*, vol. 55, no. 4, pp. 1013–1019, 2008.

[94] Z. Chen, H. Yu, N. Singh, N. Shen, R. Sayanthan, G. Lo, and D.-L. Kwong, "Demonstration of tunneling FETs based on highly scalable vertical silicon nanowires," *Electron Device Letters, IEEE*, vol. 30, no. 7, pp. 754–756, 2009.

[95] H.-Y. Chang, B. Adams, P.-Y. Chien, J. Li, and J. C. Woo, "Improved subthreshold and output characteristics of source-pocket Si tunnel FET by the application of laser annealing," *Electron Devices, IEEE Transactions on*, vol. 60, no. 1, pp. 92–96, 2013.

[96] D. Verreck, A. Verhulst, K.-H. Kao, W. Vandenberghe, K. De Meyer, and G. Groeseneken, "Quantum mechanical performance predictions of p-n-i-n versus pocketed line tunnel field-effect transistors," *Electron Devices, IEEE Transactions on*, vol. 60, no. 7, pp. 2128–2134, 2013.

[97] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, 1995.

[98] T. B. Boykin, "Recent developments in tight-binding approaches for nanowires," *Journal of Computational Electronics*, vol. 8, no. 2, pp. 142–152, 2009.

[99] J. Velev and W. Butler, "On the equivalence of different techniques for evaluating the Green function for a semi-infinite system using a localized basis," *Journal of Physics: Condensed Matter*, vol. 16, no. 21, pp. R637–R657, 2004.

[100] L. Falicov and F. Yndurain, "Model calculation of the electronic structure of a (111) surface in a diamond-structure solid," *Journal of Physics C: Solid State Physics*, vol. 8, no. 2, p. 147, 1975.

[101] C. Rivas and R. Lake, "Non-equilibrium Green function implementation of boundary conditions for full band simulations of substrate-nanowire structures," *Physica Status Solidi (B)*, vol. 239, no. 1, pp. 94–102, 2003.

[102] P. A. Khomyakov and G. Brocks, "Real-space finite-difference method for conductance calculations," *Physical Review B*, vol. 70, no. 19, p. 195402, 2004.

[103] P. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, and P. Kelly, "Conductance calculations for quantum wires and interfaces: Mode matching and Green's functions," *Physical Review B*, vol. 72, no. 3, p. 035450, 2005.

[104] M. L. Sancho, J. L. Sancho, J. L. Sancho, and J. Rubio, "Highly convergent schemes for the calculation of bulk and surface Green functions," *Journal of Physics F: Metal Physics*, vol. 15, no. 4, p. 851, 1985.

[105] A. Di Carlo, M. Gheorghe, P. Lugli, M. Sternberg, G. Seifert, and T. Frauenheim, "Theoretical tools for transport in molecular nanostructures," *Physica B: Condensed Matter*, vol. 314, no. 1, pp. 86–90, 2002.

[106] H. H. B. Sørensen, "Computational aspects of electronic transport in nanoscale devices," Ph.D. dissertation, Technical University of Denmark, 2008.

[107] J. Driscoll and K. Varga, "Calculation of self-energy matrices using complex absorbing potentials in electron transport calculations," *Physical Review B*, vol. 78, no. 24, p. 245118, 2008.

[108] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations," *Physical Review B*, vol. 74, no. 20, p. 205323, 2006.

[109] S. Lee, F. Oyafuso, P. von Allmen, and G. Klimeck, "Boundary conditions for the electronic structure of finite-extent embedded semiconductor nanostructures," *Physical Review B*, vol. 69, no. 4, p. 045316, 2004.

[110] A. Rahman, "Exploring new channel materials for nanoscale CMOS devices: A simulation approach," Ph.D. dissertation, Purdue University, 2005.

[111] S. Jin, Y. J. Park, and H. S. Min, "A three-dimensional simulation of quantum transport in silicon nanowire transistor in the presence of electron-phonon interactions," *Journal of Applied Physics*, vol. 99, no. 12, p. 123719, 2006.

[112] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer MoS_2 transistors," *Nature Nanotechnology*, vol. 6, no. 3, pp. 147–150, 2011.

[113] J. Maassen, M. Harb, V. Michaud-Rioux, Y. Zhu, and H. Guo, "Quantum transport modeling from first principles," *Proceedings of the IEEE*, vol. 101, no. 2, pp. 518–530, 2013.

[114] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties," *Physical Review B*, vol. 58, no. 11, pp. 7260–7268, 1998.

[115] A. Pecchia, G. Penazzi, L. Salvucci, and A. Di Carlo, "Non-equilibrium Green's functions in density functional tight binding: Method and applications," *New Journal of Physics*, vol. 10, no. 6, p. 065022, 2008.

[116] M. Auf der Maur, G. Penazzi, G. Romano, F. Sacconi, A. Pecchia, and A. Di Carlo, "The multiscale paradigm in electronic device simulation," *Electron Devices, IEEE Transactions on*, vol. 58, no. 5, pp. 1425–1432, 2011.

[117] C. Yam, L. Meng, G. Chen, Q. Chen, and N. Wong, "Multi-scale quantum mechanics/electromagnetics simulation for electronic devices," *Physical Chemistry Chemical Physics*, vol. 13, no. 32, pp. 14 365–14 369, 2011.

[118] T. Kubis, C. Yeh, P. Vogl, A. Benz, G. Fasching, and C. Deutsch, "Theory of nonequilibrium quantum transport and energy dissipation in terahertz quantum cascade lasers," *Physical Review B*, vol. 79, no. 19, p. 195323, 2009.

[119] P. Marconcini and M. Macucci, "The $k \cdot p$ method and its application to graphene, carbon nanotubes and graphene nanoribbons: The Dirac equation," *Rivista Del Nuovo Cimento*, vol. 34, no. 8-9, pp. 489–584, 2011.

LIST OF PUBLICATIONS

Journal Publications

1. **J. Z. Huang**, L. Zhang, W. C. Chew, C.-Y. Yam, L. J. Jiang, G.-H. Chen, and M. Chan, "Model order reduction for quantum transport simulation of band-to-band tunneling devices," (under review).

2. C.-Y. Yam, J. Peng, Q. Chen, S. Markov, **J. Z. Huang**, N. Wong, W. C. Chew, and G.-H. Chen, "A multi-scale modeling of junctionless field-effect transistors," *Applied Physics Letters*, vol. 103, no. 6, p. 062109, 2013.

3. **J. Z. Huang**, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen, "Model order reduction for multiband quantum transport simulations and its application to p-type junctionless transistors," *Electron Devices, IEEE Transactions on*, vol. 60, no. 7, pp. 2111–2119, 2013.

4. **J. Z. Huang**, W. C. Chew, Y. Wu, and L. J. Jiang, "Methods for fast evaluation of self-energy matrices in tight-binding modeling of electron transport systems," *Journal of Applied Physics*, vol. 112, no. 1, p. 013711, 2012.

5. **J. Z. Huang**, W. C. Chew, M. Tang, and L. Jiang, "Efficient simulation and analysis of quantum ballistic transport in nanodevices with AWE," *Electron Devices, IEEE Transactions on*, vol. 59, no. 2, pp. 468–476, 2012.

6. **J. Z. Huang**, P. H. Yang, W. C. Chew, and T. T. Ye, "A novel broadband patch antenna for universal UHF RFID tags," *Microwave and Optical Technology Letters*, vol. 52, no. 12, pp. 2653–2657, 2010.

Conference Publications and Presentations

1. **J. Z. Huang**, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen, "Model order reduction methods for efficient quantum transport simulation of nanoelectronic devices," *The 34th Progress In Electromagnetics Research Symposium*, Stockholm, Sweden, Aug. 12–15, 2013.

2. **J. Z. Huang**, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen, "Full-quantum simulation of p-type junctionless transistors with multi-band k.p model," *2013 IEEE International Conference of Electron Devices and Solid-State Circuits*, Hong Kong, Jun. 3–5, 2013.

3. **J. Z. Huang**, W. C. Chew, Y. Wu, and L. J. Jiang, "Fast evaluation of self-energy matrices in atomistic modeling of electron transport systems," *Workshop on Computational Methods for Complex Systems*, Hong Kong, Dec. 9–12, 2012.

4. **J. Z. Huang**, W. C. Chew, M. Tang, L. J. Jiang, and W.-Y. Yin, "Fast three dimensional simulation of silicon nanowire transistors with asymptotic waveform evaluation," *28th International Review of Progress in Applied Computational Electromagnetics Conference*, Columbus, Ohio, USA, Apr. 10–14, 2012.

5. **J. Z. Huang**, W. C. Chew, M. Tang, and L. J. Jiang, "Efficient simulation and analysis of quantum transport in nanodevices with asymptotic waveform evaluation," *CECAM workshop on Simulation and Modeling of Emerging Electronics*, Hong Kong, Dec. 12–16, 2011.

CURRICULUM VITAE

Jun Huang was born in Shangrao, Jiangxi, China, in October 1982. He received his B.S. degree from Nankai University, Tianjin, China, in June 2004, and his M.S. degree from Shanghai Jiao Tong University, Shanghai, China, in March 2010, both in electrical engineering. Since December 2008, he has been with Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, first as a Research Assistant and then as a Ph.D. student. Since September 2012, he has been a visiting scholar in Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, USA.