©Copyright 2017
Yunqi Zhao

A Nested Dissection Approach to Modeling Transport in Nanodevices: Algorithms and Applications

Yunqi Zhao

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Manjeri P. Anantram, Chair

Ulrich L. Hetmaniuk

Arka Majumdar

Program Authorized to Offer Degree:
Department of Electrical Engineering

University of Washington

Abstract

A Nested Dissection Approach to Modeling Transport in Nanodevices: Algorithms and Applications

Yunqi Zhao

Chair of the Supervisory Committee:
Manjeri P. Anantram
Department of Electrical Engineering

Modeling nanoscale devices quantum mechanically is a computationally challenging problem where new methods to solve the underlying equations are in a dire need.

In this Ph.D. work, we design and implement an efficient and high quality numerical algorithm to solve Green's functions, within the framework of non-equilibrium Green's function (NEGF) calculation, which is the most accurate approach in electronic transport simulation. Our approach exploits and extends a recent advance in using an established graph partitioning method, namely the nested dissection. The developed method has the capability to handle open boundary conditions that are represented by full self-energy matrices required for realistic modeling of nanoscale devices. We demonstrate that our method has a reduced complexity and significant speedup compared to the state-of-the-art recursive Green's function (RGF) approach across a variety of two-dimensional systems and, more important, three-dimensional structures including the traditional silicon nanowire, emerging graphene based multilayer devices, and DNA molecules.

As a novel application of the proposed simulator, we investigate the tunneling transport properties of heterostructures consisting of a few atomic layers thick hexagonal Boron Nitride (hBN) film sandwiched between armchair edged graphene nanoribbon electrodes. By incorporating our efficient Green's function solver, the modeled device ranges from a small

system with 6,000 atoms to experimental feasible sizes up to 70,000 atoms. We show a gate-controllable vertical transistor exhibiting strong negative differential resistance (NDR) effect with multiple resonant peaks, which originate from two distinct mechanisms depending on the gate and applied bias in the same device. We perform a scaling analysis of the NDR feature as a function of the system size and gain instructive insights for future theoretical and experimental investigations.

To convey more experimentally realistic simulation, we incorporate (i) angular misorientation between multilayer heterostructure, which inducing a distinct resonant mechanism depending on both gate bias and twisting angle; (ii) electron-phonon scattering decoherence mechanism, which successfully captures the current NDR peaks degradation observed in room-temperature experiments. The NDR feature with multiple resonant peaks, combined with the ultrafast tunneling speed provides prospect for the graphene-hBN-graphene heterostructure in the high-performance electronics.

TABLE OF CONTENTS

Page

List of Tables iii

List of Figures iv

Chapter 1: Introduction 1
1.1 Background 1
1.2 NEGF Approach 4
1.3 Review of Exact Algorithms 12

Chapter 2: Mathematical Description of The Algorithm 23
2.1 Description for a Simple Case 23
2.2 Description for a Multilevel Case 25
2.3 Comments on the System Partition 30

Chapter 3: Numerical Experiments for 2D Structures 35
3.1 Cost Analysis 35
3.2 Results 36
3.3 Super-lattice Device 36
3.4 Graphene 39

Chapter 4: Numerical Experiments for 3D Structures 42
4.1 Motivation 42
4.2 Operation Count Analysis for 3D Brick-like Devices 43
4.3 Numerical Experiments 45
4.4 Summary 62

Chapter 5: Development of NEGF – Poisson Solver 64
5.1 NEGF – Poisson Simulator 64

5.2 Development of 2D NEGF – Poisson Solver 66
5.3 Poisson Solver for 3D Gate All-around MOSFET 75
5.4 Poisson Solver for Graphene FET 77

Chapter 6: Application: Transport Simulation in Emerging 2D Electronic Devices 84
6.1 Negative Differential Resistance in Boron Nitride Graphene Heterostructures: Physical Mechanisms and Size Scaling Analysis 85
6.2 Negative Differential Resistance in Graphene Boron Nitride Heterostructure Controlled by Twist and Phonon-Scattering 99

Chapter 7: Conclusion 109

Appendix A: Derivation of HSC-Extention (Simple Case) 111

Appendix B: Description of the Algorithm for a Three-Level Tree 114

Appendix C: Complexity Derivation of HSC-extension for 3D Cuboidal Structures . 123

Bibliography 125

LIST OF TABLES

Table Number Page

1.1 The range of validity for several device simulation approaches. Here, L is the feature size of the electronic device; l_{e-ph} and l_{e-e} are the scattering lengths between electron-phonon and electron-electron; λ denotes the particle transport wave length of charge carriers. 3

1.2 NEGF-Poisson simulation requirement of physical quantities. Here $\mathbf{G}^{r(<)}$ only represents the diagonal and required off-diagonal entries, rather than the full matrix. 13

1.3 Complexity of algorithms to compute diagonal blocks of $\mathbf{G}^r$. 21

1.4 Complexity of algorithms to compute diagonal blocks of $\mathbf{G}^<$. 22

2.1 Features summarized for three exact NEGF algorithms, RGF, FIND and HSC(-extension) 34

4.1 Operation counts of HSC-extension and RGF for various configurations of cuboid mesh. 45

4.2 Extrapolated CPU timings of transmission calculation for SiNW devices at one energy point with $L = 20$nm for various shapes. 59

4.3 CPU timings of transmission calculation for DNA molecules at one energy point. 60

4.4 CPU timings of transmission calculation for DNA molecules at one energy point. 62

6.1 Peak current and PVR values as a function of for both mechanisms. (I-V curves from Figure 6.5) 96

LIST OF FIGURES

Figure Number Page

1.1 Nano-device partitioned into N_y layers. Each layer contains N_x grid points. . . 6

1.2 $\mathbf{H}$ (left) matrix shape for device in Figure 1.1, non-zero entries are highlighted. $N_x = 5$ and $N_y = 9$ for the system. $\mathbf{H}$ corresponds to equation (1.11). Each $\mathbf{H}_{i,j}$ is a 5×5 sparse matrix and diagonal blocks number is 9. 6

1.3 Left: An example zigzag-edged graphene sheet. The rectangular box represents the unit cell layer repeating along y direction infinitely. Right: $\mathbf{H}$ matrix for the graphene sheet, non-zero entries are highlighted. $N_x = 6$ and $N_y = 6$ for the system. 9

1.4 $\mathbf{\Sigma}^r_L$ (left), $\mathbf{\Sigma}^r_R$ (middle) and $\mathbf{A}$ (right) matrix shape for device in Figure 1.1, non-zero entries are highlighted. $N_x = 5$ and $N_y = 9$ for the system. Each $\mathbf{\Sigma}^r_{L/R}$ contains a $N_x \times N_X$ dense block at its first/last diagonal block site. $\mathbf{A}$ is computed by equation (1.2) with left and right semi-infinite leads effect folded into $\mathbf{\Sigma}^r_L$ and $\mathbf{\Sigma}^r_R$, which brings two $N_x \times N_x$ dense blocks into the first and last diagonal blocks of $\mathbf{A}$ respectively. 10

1.5 Computational intensity varies for different self-energy included. The contact self-energy increases the complexity for Green's function solution at each energy point by introducing dense blocks, while the scattering self-energy brings additional iterative loop with lesser Green's function. 12

1.6 System is partitioned into two disjoint regions Z and Z', which has interaction with each other. 16

1.7 RGF partitions the systme into disjoint layers with interaction at specific direction. Solid arrows and dash arrows relate to the *fold* and *extract* passes respectively in computing Green's functions. 19

2.1 Nanodevice partitioned into two subregions (L, R) and a separator S (L stands for Left and R for Right). Solid arrows and dash arrows relate to the *fold* and *extract* passes respectively in computing Green's functions. 24

2.2 Example of a multilevel partition. 26

2.3 Flowchart for NEGF calculation incorporated with HSC-extension algorithm. 31

2.4 Partition generating the RGF algorithm. 32

2.5 First method to partition system with two dense layers and two ends. 33

2.6 Partition generated by METIS for system including dense layers at two ends. 33

3.1 Numerical count comparison for our algorithm (blue) and RGF (red). . . . 36

3.2 Barrier structure for a model super-lattice device. 36

3.3 Electron density profile and electron density in y direction for a model super-lattice device. 37

3.4 Superlattice device NEGF simulation computation time comparison for RGF and our methods, all systems grid spacing is 0.1nm. (a) Square system of with diagonal self-energy matrix; (b) Square system of with dense self-energy matrix; (c) For systems in this plot, the length in the x-direction is fixed at 25 nm while the length in the y-direction is increased. (d) For systems in this plot, the length in the y-direction is fixed at 10 nm while the length in the x-direction is increased. Dense self-energy matrices are used in (c) and (d) devices. 38

3.5 Graphene hexagonal structure decomposed by tight biding method. Dashed rectangular illustrates one repeating hexagon layer. Dashed lines represent inner four atom layers, showing the atoms ordering in tight binding Hamiltonian construction. 40

3.6 Graphene device NEGF simulation computation time comparison for RGF and the HSC extension, based on tight binding theory. (a) Square system of with diagonal self-energy matrix. (b) Square system of with dense self-energy matrix. (c) For systems in this plot, the number of atoms in the x-direction is set to $N_x = 250$, while the number of atoms in the y-direction is increased. (d) For systems in this plot, the number of atoms in the y-direction is set to $N_y = 100$, while the number of atoms in the x-direction is increased. Dense self-energy matrices are used in (c) and (d) devices. 41

4.1 A Cartesian 3D mesh with 7-point-stencil discretization of dimension $N_x \times N_z \times N_y$ (the y-direction is the transport direction.). The colored layers along the y-direction show the layered-structure organization of grid points for the RGF approach. 44

4.2 (a) The partition of grid points obtained from the multilevel nested dissection. The colored clusters show the separators defined at each level. (b) A binary tree representing the clustering. Each colored block matches a colored separator. 44

4.3 Non-zero pattern of $\mathbf{A}$ for a 3D cuboid system with $N_x = 3$, $N_z = 3$ and $N_y = 5$. Entries for the diagonal self-energy approximation are marked in red. The matrix exhibits a block-tridiagonal structure, where each block is of dimension $N_xN_z \times N_xN_z$. The arrow highlights the diagonal width in each block controlled by the ratio N_x/N_z. In all the matrix pattern graphs, `nz` specifies the number of non-zero entries. 46

4.4 CPU timing for cubic system versus the dimension N_x. For all plots in result section, the timing includes the $\mathbf{G}^r$ and $\mathbf{G}^<$ calculation at one energy point. The runtime of RGF, HSC-extension and LU-factorization for $\mathbf{A}$ with *SPARSE* self-energy are presented. For comparison, we also plot black dashed curves, reflecting the theoretically asymptotic slopes for HSC-extension: $\mathcal{O}(N^6)$, and for RGF: $\mathcal{O}(N^7)$. 47

4.5 (a) CPU timing for elongated mesh versus N_x with fixed $N_x = N_z = N \ll N_y$, $N_y = 200$. (b) CPU timing for elongated mesh versus N_y with fixed $N_x = N_z = 16$. The theoretically asymptotic slopes (black dashed curves) for HSC-extension correspond to Table 4.1, $\mathcal{O}(N^5N_y)$, and for RGF to $\mathcal{O}(N^6N_y)$. 48

4.6 CPU timing for flattened mesh versus N_x with $N_x = N_y = N \gg N_z$, $N_z = 4$. The theoretically asymptotic slope (black dashed curves) for flattened mesh is $\mathcal{O}(N^3)$ for HSC-extension, and $\mathcal{O}(N^4)$ for RGF. 49

4.7 Non-zero pattern of $\mathbf{A}$ for a 3D cuboid system with $N_x = 3$, $N_z = 3$ and $N_y = 5$. Entries for the diagonal self-energy approximation are marked in red. The matrix exhibits a block-tridiagonal structure, where each block is of dimension $N_xN_z \times N_xN_z$. 50

4.8 CPU timing for cubic system $N_x = N_y = N_z = N$ with both *DENSE* and *SPARSE* self-energies. The black dashed curve shows the asymptotic rates, namely $\mathcal{O}(N^6)$ for HSC-extension and $\mathcal{O}(N^7)$ for RGF. 51

4.9 CPU timing for different N_y and fixed $N_x = N_z = 16$. The black dashed curve shows the asymptotic rate $\mathcal{O}(N_y)$. 52

4.10 (a) Schematic view of graphene-hBN-graphene multilayer heterostructure. Two graphene layers are semi-infinitely long used as contacts. (b) The non-zero pattern of the $\mathbf{A}$ matrix with $N_x = 16$, $N_y = 8$ and $N_z = 3$. 53

4.11 CPU timing for G-BN-G system as a function of $N_x = N_y$ and fixed $N_z = 5$. Dashed curves illustrate asymptotic rates, namely $\mathcal{O}(N_x^3)$ for HSC-extension and $\mathcal{O}(N_x^4)$ for RGF. 54

4.12 (a) CPU timings for G-BN-G system as a function of N_x and fixed $N_y = 32$, $N_z = 5$. (b) CPU timings for different N_y and fixed $N_x = 64$, $N_z = 5$. (c) CPU timings as a function of N_z and fixed $N_x = N_y = 48$. The dashed curves indicates the asymptotic operation counts. For HSC-extension, they are $\mathcal{O}(N_x^{1.5})$ for (a), $\mathcal{O}(N_y^{1.5})$ for (b), and $\mathcal{O}(N_z^2)$ for (c). The operation counts for RGF are as follows: $\mathcal{O}(N_x^3)$ for (a), $\mathcal{O}(N_y)$ for (b), and $\mathcal{O}(N_z^3)$ for (c). . 55

4.13 (a) CPU timing for G-BN-G system with different N_x and fixed $N_y = 256$, $N_z = 5$. (b) CPU timing for different N_y and fixed $N_x = 48$, $N_z = 5$. The dashed curves indicates the asymptotic operation counts. For HSC-extension, they are $\mathcal{O}(N_x^2)$ for (a) and $\mathcal{O}(N_y)$ for (b). The operation counts for RGF are as follows: $\mathcal{O}(N_x^3)$ for (a) and $\mathcal{O}(N_y)$ for (b). 56

4.14 (a) Atomic view of a silicon nanowire example with 4 unit cells. Each unit cell has two atomic layers and hexagonal cross-section shape, with each atomic layer containing 108 Si atoms. Cross-section is along $x-z$ plane and transport direction is along y direction. This example corresponds to $N_{cs} = 108$ and $N_y = 8$. (b) The non-zero pattern of the **A** matrix with $N_{cs} = 108$ and $N_y = 8$. 57

4.15 CPU timing for SiNW system with (a) $L = 10D$ with largest $L = 20$nm, (b) $L = 6D$ with largest $L = 15$nm and (c) $L = 3D$ with largest $L = 9.3$nm. Note that $N_y \propto L$ and $N_{cs} \propto D^2$. The dashed curves represent asymptotes: $\mathcal{O}(N_y^6)$ for HSC-extension and $\mathcal{O}(N_y^7)$ for RGF. 58

4.16 (a) Sketch of the simulated DNA sequence with 7, 9, 11, 13 and 15 base pairs respectively. Cytosine (C) and guanine (G) are two types of bases in DNA. The left/right contacts are connected to the bases on one strand. (b) The corresponding non-zero pattern of the **A** matrix for the 9 base pairs DNA. All tri-diagonal blocks are fully dense. 61

4.17 Cluster definitions for HSC-extension, for RGF, and for two customizations. 61

5.1 Flowchart for NEGF-Poisson self-consistent simulation, including self-consistent solution of decoherence self-energy. 65

5.2 A 2D dual-gate SOI MOSFET toy model employed as an example in the development of Poisson solver. 68

5.3 The computational domain for the NEGF calculation part. Note that we only discretize the semiconductor region, since the oxide regions do not hold charge and thus have no contribution to the transport. 70

5.4 The computational domain and boundary conditions for the Poisson calculation. 71

5.5 Simulation results for a thick body 2D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.2V$. 73

5.6 Simulation results for a thin body 2D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.2V$. 74

5.7 Simulation results for a short channel 2D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.6V$. 75

5.8 A 3D gate all-around SOI MOSFET device, which is a straightforward extension from the 2D dual-gate MOSFET. 76

5.9 Simulation results for a gate all-around 3D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.6V$. 77

5.10 A 3D graphene FET device formed by a single graphene layer sandwiched by two thin oxide layers. Left graph shows the graphene monolayer and the right graph shows the lateral view of the device. 78

5.11 (a) Illustration of grid points for the graphene layer. Blue circles denote the atomic sites and red circles denote the interbond grid points. (b) A final tetrahedron FEM mesh generated for the graphene-FET device consisting of a graphene monolayer vertically sandwiched by two insulating layers. In the sample device, the graphene layer contains 160 carbon atoms and locates at $z = 0$. 80

5.12 Simulation results for a 3D graphene-FET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.3V$. 82

5.13 Simulation results for a 3D graphene-FET structure for $V_g = 0.3$V with different numbers of interbond grid points. 83

6.1 A schematic view of the heterostructure device. N_x and N_y represent the width and stacking length of the device respectively. N_z is the number of hBN layers sandwiched between the two AGNR ribbons. All the dimensions are in unit of atomic layers. (b) The average DOS versus Energy of hBN for device with $N_x = 200$, $N_y = 32$ and $N_z = 3$. This shows a 4.72eV bandgap of atomically thin hBN material and a 1.38eV valence band-offset between graphene and hBN stacking structure. 88

6.2 (a) Current versus drain voltage for a device with $N_x = 62$, $N_y = 32$ and $N_z = 3$. V_g varies from -45V to 0. The black solid arrows in the four plots mark the current resonant peaks due to mechanism 2, and the empty arrows marks the NDR peak due to mechanism 1. The inset explains the resonant tunneling induced from mechanism 2. The difference between Fermi energy and Dirac point in bottom graphene is induced by gate potential. When $V_b = -0.01V_g$, the electronic spectra of top and bottom electrodes are tuned into alignment, allowing the resonant tunneling. (b) Current versus drain voltage for large device with $N_x = 200$, $N_y = 32$ and $N_z = 1$. Here V_g varies from -45V to +45V. Inset shows an asymmetric PVR relationship with the applied vertical gate potential. 89

6.3 Transmission and DOS plot at various biases for I-V curve ($V_g = 0$) in Figure 6.2(a). (a)-(c) specifies the bias potential $V_b = 0.3$V, 0.46V and 0.66V respectively. In the transmission plots, black dashed curves are transmission coefficient when $V_b = 0$. In DOS$_g$ plots, blue and red curves represent DOS of bottom and top graphene sheets respectively. Vertical dash-dot lines give the chemical potentials at both graphene ends μ_B and μ_T, which determines the bias window. S mark the DOS peaks resulting from the zigzag shaped edges of graphene cut ends. P mark the transmission peaks that mainly contribute to the current. T represent the tunneling peaks due to the energy alignment of subbands in top and bottom graphene contacts; they do not contribute significantly to current. Units of DOS are number of states per atom per eV. 91

6.4 (a) Current-voltage curves for devices with different N_z, with fixed $N_x = 62$ and $N_y = 32$. Here the current value for cases when $N_z = 3$ and $N_z = 5$ are scaled by 1E2 and 1E4 respectively. The inset plots the low bias conductance of the three current-voltage curves. (b) Transmission relationship for devices with different N_z and fixed $N_x = 62$ and $N_y = 32$ at $V_b = 0.3V$, corresponding to the first current peaks shown in (a). Again, the transmission coefficient value for cases when $N_z = 3$ and $N_z = 5$ are scaled by 1E2 and 1E4 respectively. 94

6.5 (a) Current versus drain voltage at $V_g = 0$ for devices with various N_x, with fixed $N_y = 32$ and $N_z = 1$. (b) Current versus drain voltage at $V_g = -45V$ for devices with various N_x, with $N_y = 32$ and $N_z = 1$. Black arrows mark the NDR peaks due to mechanism 2. 95
6.6 (a) Current versus drain voltage at $V_g = 0$ for devices with various N_y , with fixed $N_x = 200$ and $N_z = 1$. (b) Current versus drain voltage at $V_g = -45V$ for devices with various N_y, with $N_x = 200$ and $N_z = 1$. Black arrows mark the NDR peaks due to mechanism 2. 97
6.7 (a) A schematic view of the twisted heterostructure device. An external gate electrode is applied on bottom graphene sheet. The top graphene layer is rotated with hBN insulator by an exaggerated angle θ. Inset: The Brillouin zones for bottom and top graphene layers in momentum space. The neutrality points of different graphene layers are displaced by ΔK. (b)-(e): The horizontal distance between neutrality points is determined by the rotation angle θ and the vertical distance between them are determined by the applied gate voltage. (b) depicts the situation of $V_b = V_b^R$. (c)-(e) correspond to situations of $V_b < V_b^P$, $V_b = V_b^P$ and $V_b > V_b^P$. The red and blue cones represent the energy dispersions of bottom and top graphene layers respectively. Occupied and unoccupied states are distinguished by different transparency. The transmissive states that can carry tunnel current is highlighted by yellow curves. 101
6.8 Calculated I-V curves for a family of twisting angle without external gate electrode. 102
6.9 V_b^R and V_b^P as a function of θ for both simulated results (markers) and analytical (dashed) estimations. Top inset: illustration of the situation at V_b^P, a 2D version of Figure 6.7(d). Bottom inset: illustration of the situation at V_b^R where tunneling current begins to increase rapidly from zero, a 2D version of Figure 6.7(b). The electrostatic model is defined in equations 6.1-6.2. 104
6.10 Calculated I-V curves for a fixed non-zero twisting angle $\theta = 4°$ with various gate voltages. 105
6.11 Calculated I-V curves at $V_g = -45V$ for devices ($\theta = 0$) with $N_z = 1$, $N_x = 62$ and $N_y = 64$ with consideration of electron-phonon scattering. 107
6.12 Calculated I-V curves for a family of θ at $V_g = 0$ with electron-phonon scattering (dashed curves). For comparison, the corresponding results of coherent transport (from Figure 6.8) are plotted as solid curves. 108
B.1 Partition for a three-level system. 114
B.2 Sparsity of matrix $\mathbf{A}$ and matrix $\mathbf{A}^{(1)}$. 116
B.3 Sparsity of matrix $\mathbf{G}^{(0)}$. 119

C.1 (a) The domain decomposition from cube of dimension $2a + 1$ to cubes of dimension a. Three levels of separators are colored by red, purple and green respectively. (b) The multilevel binary tree corresponding to the cuboid decomposition. The three levels of separators are depicted with matching colors. The blue blocks denote the corresponding blue clusters. 123

Chapter 1
INTRODUCTION

1.1 Background

Since 1960s, semiconductor industry has achieved tremendous development and grown to be a giant market with over $300 billion value today. The continuous semiconductor scaling drives the ever-decreasing costs of processing, transmission and storage capabilities, which makes semiconductor engineering research an essentially important part in the industrial development. Device physics, as one of the fundamentals in semiconductor research, studies how electronic devices operates. Based on such knowledge, operational behaviors of an electronic device can be accurately predicted by incorporating appropriate physical models, without any actual fabrication of the device. This leads to the emergence of computational electronics, which develops Technology Computer Aided Design (TCAD) tools to model semiconductor device operation. Computational simulation offers many advantages such as decreasing industrial design cycle time, providing problem diagnostics, gaining insights for future products, and shortening time to market.

The operation of semiconductor electronics is determined by the transport behavior of electrons and holes, two basic charge carriers in semiconductor devices. In traditional device physics, charge carriers are commonly treated as semi-classical particles with effective mass moving through the device driven by electric field. This process can be clearly expressed by drift-diffusion equations. The diffusive carriers transport scenario has served device engineers and researchers for over 50 years, and is usually clarified and adequate for most simple devices (e.g., field-effect-transistors). Traditional TCAD tools solves the drift-diffusion equation, current continuity equation, self-consistently with the Poisson's equation in the simulation process. This semi-classical approach has achieved great success for devices in micrometer

and sub-micrometer scale.

Things are changing today. Moore's law predicted the rapid scaling down of semiconductor feature size, now leading the whole device engineering world to nanoscale range. Traditional silicon-based devices is about to reach the scaling limitation of a few nanometers. People start to investigate new devices built from semiconductor nanowires, graphene, carbon nanotubes and organic molecules. Due to quantum confinement, the electrical properties of materials are sensitive to the structures in atomistic level. Treating carrier transport as diffusive particles has lost its validity, and traditional TCAD technique has failed in considering the wave nature of carrier transport. [4] On the other hand, doing experiments with semiconductors at atomistic scale is difficult and demanding. Seeking practical analyzing and simulation approaches based on more fundamental physics is imminent for nanomaterial engineers. This leads to the application of quantum mechanics in semiconductor physics. The electrons and holes shall be described quantum mechanically, yielding a new treatment of carriers in atomistic scale rather than continuous one.

When the semiconductor feature size is below 10nm, naturally, device dimensions become comparable to the scattering length due to phonons, photons and other electrons. To investigate accurately the transport properties of charge carriers by quantum mechanical method, the modeling approach should capture mechanisms such as quantum tunneling, quantum confinement, and scattering mechanisms. Under such circumstance, simply replacing drift-diffusion equation with the basic Schrödinger's equation is not adequate. The non-equilibrium Green's function (NEGF) method [4, 13, 14] has emerged as a powerful modeling approach for nanodevices and nanomaterials. The NEGF approach is based on the self-consistent coupling of Schrödinger and Poisson equations and is designed to capture decoherence effects including electron-phonon scattering. The improvements of NEGF-based computational electronics contains:

- Correctly evaluate the size-quantization effect of nanoscale electronic devices and benchmark the quantum corrections for semi-classical modeling.

- Capture physics underlying atomistic scale effects such as quantum mechanical tunneling, and quantum interference effects.

- Accurately treat scattering and other phase-breaking mechanism. Formulation and implementation of scattering in NEGF is quite straightforward.

- Capable of dealing with correlations in both space and time domain.

Table 1.1 shows the validity range for major semiconductor simulation methods including both semi-classical and quantum mechanical approaches.

	Model	Range of validity
Semi-classical approaches	Drift-diffusion equations	$L \gg l_{e-ph}$
	Hydrodynamic equations	$L \gg l_{e-e}$
	Boltzmann transport equation, Monte Carlo methods	$L \geqslant l_{e-e}$, Accurate up to classical limits
Quantum mechanical approaches	Quantum hydrodynamics	$L < l_{e-e}$
	Quantum-Kinetic equations	$L < l_{e-e}$, Accurate up to single particle description
	(non-equilibrium) Green's function method	$L < \lambda$

Table 1.1: The range of validity for several device simulation approaches. Here, L is the feature size of the electronic device; l_{e-ph} and l_{e-e} are the scattering lengths between electron-phonon and electron-electron; λ denotes the particle transport wave length of charge carriers.

1.2 NEGF Approach

1.2.1 Governing Equations

A typical NEGF-based simulation solves three Green's function equations,

$$\begin{cases} \mathbf{A}(E)\,\mathbf{G}^r(E) &= \mathbf{I} \\ \mathbf{A}(E)\,\mathbf{G}^<(E) &= \mathbf{\Sigma}^<\left(\mathbf{G}^r(E)\right)^\dagger \\ \mathbf{A}(E)\,\mathbf{G}^>(E) &= \mathbf{\Sigma}^>\left(\mathbf{G}^r(E)\right)^\dagger \end{cases} \tag{1.1}$$

where the matrix $\mathbf{A}$ is defined by

$$\mathbf{A} = E\mathbf{I} - \mathbf{H} - \mathbf{\Sigma}_L^r - \mathbf{\Sigma}_R^r - \mathbf{\Sigma}_{\text{Phonon}}^r \tag{1.2}$$

$\mathbf{G}^r(E)$ is called the retarded Green's function, describing local density of states and the propagation of electrons injected in the device, and $(\mathbf{G}^r(E))^\dagger$ its Hermitian conjugate. $\mathbf{G}^<(E)$, the lesser Green's function, represents the electron correlation function for energy level E; the diagonal elements of $\mathbf{G}^<(E)$ represent the electron density per unit energy. $\mathbf{G}^>(E)$, the greater Green's function, represents the hole correlation function for energy level E, which is proportional to the density of unoccupied states. $\mathbf{I}$ is the identity matrix and $\mathbf{H}$ the system Hamiltonian, including both dynamical and potential parts. $\mathbf{\Sigma}_L^r$ and $\mathbf{\Sigma}_R^r$ represent the self-energies due to left and right contact coupling and $\mathbf{\Sigma}_{\text{Phonon}}^r$ corresponds to the self-energy governing electron-phonon scattering. The matrix $\mathbf{\Sigma}^<$ corresponds to the lesser self-energy and the matrix $\mathbf{\Sigma}^>$ to the greater self-energy. Physically, the lesser(greater) self-energies represent the in(out)-scattering of carriers, reflecting the charge occupancy in devices. They are determined by the following equations.

$$\mathbf{\Sigma}_L^< = -2i\,\text{Im}[\mathbf{\Sigma}_L^r] f_L(E) \tag{1.3}$$

$$\mathbf{\Sigma}_R^< = -2i\,\text{Im}[\mathbf{\Sigma}_R^r] f_R(E) \tag{1.4}$$

$$\mathbf{\Sigma}_{\text{Phonon}}^< = -2i\,\text{Im}[\mathbf{\Sigma}_{\text{Phonon}}^r] f_{\text{Phonon}}(E) \tag{1.5}$$

$$\mathbf{\Sigma}^< = \mathbf{\Sigma}_L^< + \mathbf{\Sigma}_R^< + \mathbf{\Sigma}_{\text{Phonon}}^< \tag{1.6}$$

and

$$\begin{aligned}
\mathbf{\Sigma}_L^> &= -2i\,\mathrm{Im}[\mathbf{\Sigma}_L^r](1 - f_L(E)) && (1.7)\\
\mathbf{\Sigma}_R^> &= -2i\,\mathrm{Im}[\mathbf{\Sigma}_R^r](1 - f_R(E)) && (1.8)\\
\mathbf{\Sigma}_{\mathrm{Phonon}}^> &= -2i\,\mathrm{Im}[\mathbf{\Sigma}_{\mathrm{Phonon}}^r](1 - f_{\mathrm{Phonon}}(E)) && (1.9)\\
\mathbf{\Sigma}^> &= \mathbf{\Sigma}_L^> + \mathbf{\Sigma}_R^> + \mathbf{\Sigma}_{\mathrm{Phonon}}^> && (1.10)
\end{aligned}$$

Where $f_L(E)$ and $f_R(E)$ are Fermi-Dirac distribution function for left and right contacts respectively, and $f_{\mathrm{Phonon}}(E)$ is distribution function related with phonon. Both $\mathbf{\Sigma}^<$ and $\mathbf{\Sigma}^>$ self-energy matrices consist of three parts, corresponding to two contacts and electron-phonon scattering. The charge carriers density can then be obtained from the diagonal elements of $\mathbf{G}^<$ and $\mathbf{G}^>$ (see future definitions). The Green's functions are then incorporated in the coupling between the Schrödinger and Poisson equations – see [4] for further details.

1.2.2 Hamiltonian Matrix

The Hamiltonian matrix (electronic structure) of the device can be commonly obtained in two ways, which will be briefly described in this section.

Effective Mass Approximation

We introduce a simple effective-mass 2D system Hamiltonian and describe how physical quantities can be computed from Green's function matrices. Consider a device that can be topologically broken down into layers as shown in Figure 1.1. For an effective mass Hamiltonian, these dots (both blue and black) represent grid points of the discretized Green's function equation in coordinate space.

The Hamiltonian operator $H(\mathbf{x}) = -(\hbar^2/2m)\nabla^2 + V(\mathbf{x})$ can then be decomposed by using first-order finite difference method, *i.e.* each dot is connected with four nearest neighbor dots. The resulting five-point-stencil Hamiltonian is a block tri-diagonal matrix as shown in Figure 1.2, where each diagonal block represents the Hamiltonian of a layer in Figure 1.1. The i-th

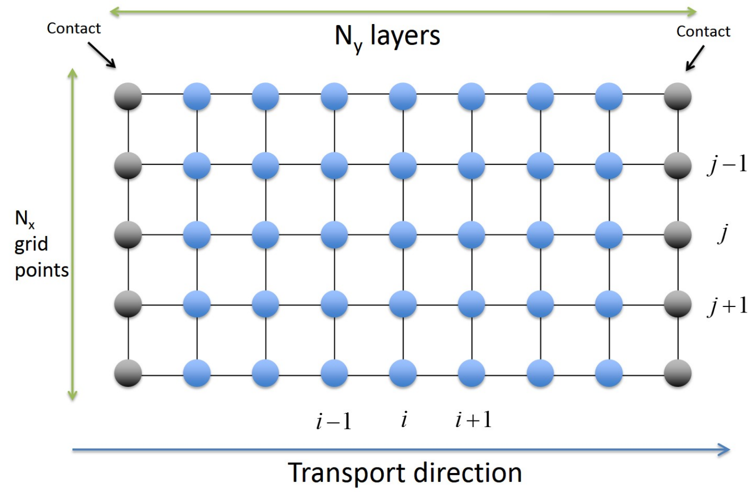

Figure 1.1: Nano-device partitioned into N_y layers. Each layer contains N_x grid points.

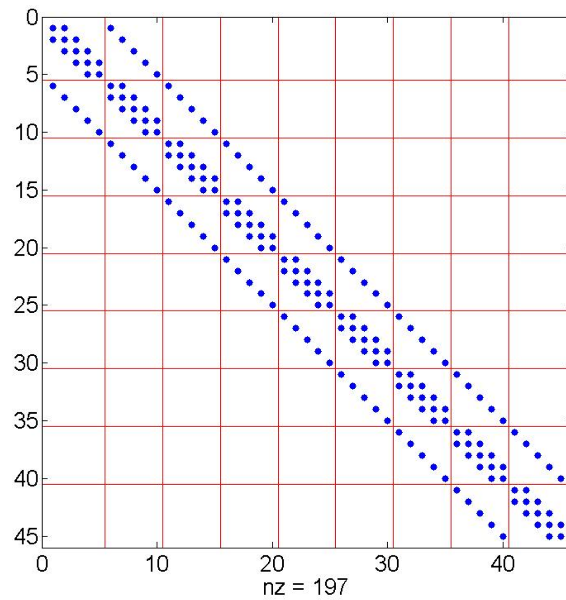

Figure 1.2: $\mathbf{H}$ (left) matrix shape for device in Figure 1.1, non-zero entries are highlighted. $N_x = 5$ and $N_y = 9$ for the system. $\mathbf{H}$ corresponds to equation (1.11). Each $\mathbf{H}_{i,j}$ is a 5×5 sparse matrix and diagonal blocks number is 9.

diagonal block of the Hamiltonian represents the coupling between grid points in a layer. The off-diagonal blocks to its left and right represents coupling to between layers i and $i-1$ and $i+1$ respectively. Both the diagonal and off-diagonal blocks of the Hamiltonian are sparse in many examples where an effective mass Hamiltonian represents the dynamics. The block representation of the Hamiltonian matrix is expressed in (1.11). Each block $\mathbf{H}_{i,j}$ is a $N_x \times N_x$ sparse matrix.

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} & & & \\ \mathbf{H}_{2,1} & \mathbf{H}_{2,2} & \mathbf{H}_{2,3} & & \\ & \mathbf{H}_{3,2} & \ddots & \ddots & \\ & & \ddots & \mathbf{H}_{N_y-1,N_y-1} & \mathbf{H}_{N_y-1,N_y} \\ & & & \mathbf{H}_{N_y,N_y-1} & \mathbf{H}_{N_y,N_y} \end{pmatrix} \tag{1.11}$$

The constructed Hamiltonian matrix can then be involved in Green's function calculation and the physical quantities can be computed by the following equations [4, 13]. Local density of states in grid point (i,j) can be calculated by

$$N_{i,j}(E) = -\frac{1}{\pi}\mathrm{Im}[\mathbf{G}^r(k,k,E)] \tag{1.12}$$

where $k = N_x \times (i-1) + j$ is the numbering of the grid point corresponding to the row number in the matrix. $\mathbf{G}^r(k,k,E)$ represents the (k,k) element in matrix $\mathbf{G}^r$ that is calculated at energy point E. Electron density is computed by

$$n(i,j) = 2\int \frac{\mathrm{d}E}{2\pi}\mathrm{Im}[\mathbf{G}^<(k,k,E)] \tag{1.13}$$

Then the current density flowing from layer i to layer $i+1$ can be computed.

$$J_{i\to i+1} = \frac{2e}{\hbar}\int \frac{\mathrm{d}E}{2\pi}\mathrm{Tr}[\mathbf{T}_{i,i+1}\mathbf{G}^<_{i+1,i}(E) - \mathbf{T}_{i+1,i}\mathbf{G}^<_{i,i+1}(E)] \tag{1.14}$$

Here, $\mathbf{T}$ is the kinetic part of the Hamiltonian matrix $-(\hbar^2/2m)\nabla^2$, and $\mathbf{T}_{i,i+1}$ represents the $(i,i+1)$ block in the $\mathbf{T}$ matrix — see equation (1.11).

Tight Binding Approximation

Tight-binding model decomposes the system by using an approximate set of wave functions based upon superposition of wave functions for isolated atoms located at each atomic site. According to tight-binding approximation, the electrons in this model should be tightly bound to the atom to which they belong and they should have limited interaction with states and potentials on surrounding atoms of the device. As a result the wave function of the electron will be rather similar to the atomic orbital of the free atom to which it belongs. The energy of the electron will also be rather close to the ionization energy of the electron in the free atom or ion because the interaction with potentials and states on neighboring atoms is limited. Though the tight-binding model is a one-electron model, it also provides a basis for more advanced calculations like the calculation of surface states and application to various kinds of many-body problem and quasiparticle calculations. The model gives good qualitative results in many cases and can be combined with other models, like Linear Combination of Atomic Orbitals (LCAO) method used in Density Functional Theory (DFT). Here we show an example of Hamiltonian matrix constructed from tight-binding approach.

Figure 1.3 (left) shows the atomistic structure of a zig-zag edged graphene sheet. For each carbon atom, only p_z orbital is considered and the resulting Hamiltonian is shown in Figure 1.3. The diagonal entries in the matrix correspond to the on-site potential energy for each p_z orbital and the off diagonal ones correspond to the interaction parameters between p_z orbitals from the nearest neighbor carbon atoms. Usually, the tight-binding parameters are obtained by fitting electronic structure (dispersion relationship) between tight-binding and *ab initio* calculations. We notice that the size of the Hamiltonian matrix is determined by the number of atoms and the number of orbitals considered for each atom (which is one in the graphene example). Increasing the number of orbitals will reduce the computational efficiency while improving the calculation accuracy.

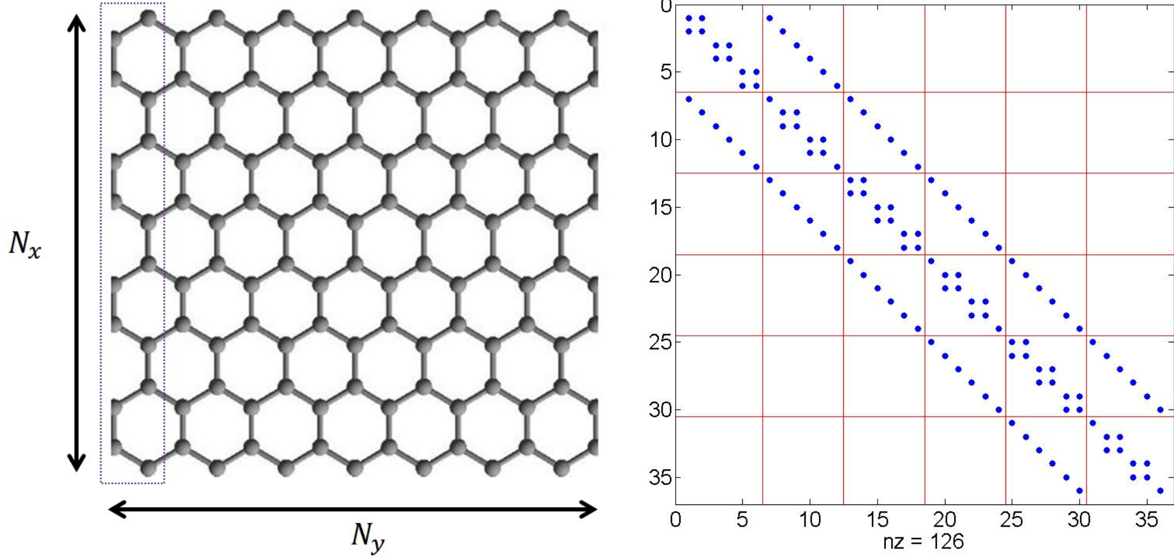

Figure 1.3: Left: An example zigzag-edged graphene sheet. The rectangular box represents the unit cell layer repeating along y direction infinitely. Right: $\mathbf{H}$ matrix for the graphene sheet, non-zero entries are highlighted. $N_x = 6$ and $N_y = 6$ for the system.

1.2.3 Self-energy Matrix

Self-energy matrix can be expressed by equation $\mathbf{\Sigma}^{r,<,>} = \mathbf{\Sigma}_L^{r,<,>} + \mathbf{\Sigma}_R^{r,<,>} + \mathbf{\Sigma}_{\text{Phonon}}^{r,<,>}$, which consists of three parts: self-energy due to left contact, right contact and electron-phonon scattering. We will show the self-energy matrices for the device described in section 1.2.2, which can be easily extended to other systems. For open boundary system, the left and right coupling contacts (indicated in Figure 1.1) are two semi-infinite leads connected to the device and should be represented by infinitely large Hamiltonian matrices. Their respective effect can be folded into layer 1 and layer N_y, resulting in dense blocks for the first and the N_y-th diagonal blocks of the self-energy. The matrix structure of the left and right self energy matrices are shown in Figure 1.4.

The dense contact self-energy matrices, namely the surface Green's function (SGF), need to be computed repeatedly for each energy point which makes algorithm speedup desirable. Recently, the layered structure SGF tends to be computed by an efficient iterative algorithm

[59]. This method will include 2^i+1 layers in the ith iteration, until convergence is achieved. This approach is generic in handling the semi-infinite leading contacts and has an exponential convergence speed, compared to some other methods with linear convergence speed [60]. We also adopt this approach in the current work.

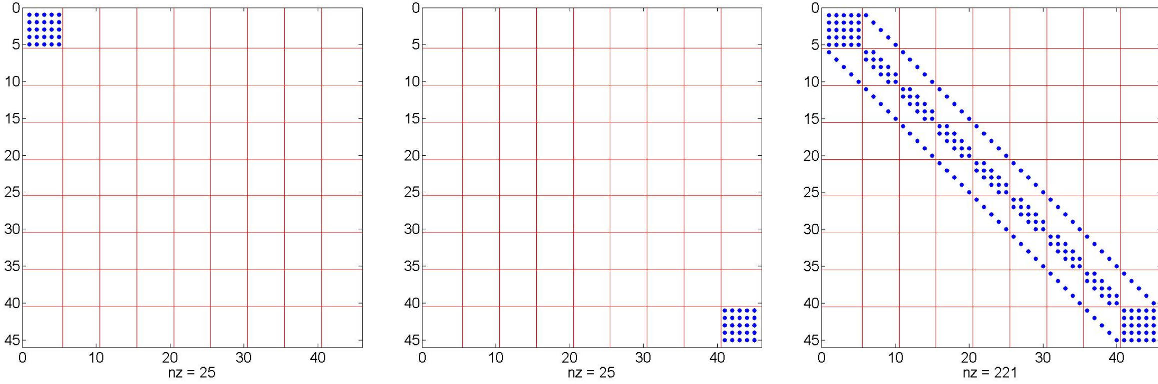

Figure 1.4: $\mathbf{\Sigma}_L^r$ (left), $\mathbf{\Sigma}_R^r$ (middle) and $\mathbf{A}$ (right) matrix shape for device in Figure 1.1, non-zero entries are highlighted. $N_x = 5$ and $N_y = 9$ for the system. Each $\mathbf{\Sigma}_{L/R}^r$ contains a $N_x \times N_X$ dense block at its first/last diagonal block site. $\mathbf{A}$ is computed by equation (1.2) with left and right semi-infinite leads effect folded into $\mathbf{\Sigma}_L^r$ and $\mathbf{\Sigma}_R^r$, which brings two $N_x \times N_x$ dense blocks into the first and last diagonal blocks of $\mathbf{A}$ respectively.

For realistic system whose size is comparable with electron mean free path, the phase-breaking scattering is involved by corresponding self-energy matrices ($\mathbf{\Sigma}_{\text{Phonon}}^{r,<,>}$) such that the transport is between the ballistic and diffusive limit and energy is relaxed. In the decoherence transport regime, NEGF approach has a distinct advantage over self-consistently solving Schrödinger's and Poisson's equation. In numerical simulation, the scattering self-energy matrices $\mathbf{\Sigma}_{\text{Phonon}}^{r,<,>}$ are set to be diagonal at each interior grid point, which may arise due to electron-phonon interaction or other decoherence interactions. Such assumption does not affect the sparsity of the system matrix and the computational complexity. Relaxing the requirement of diagonal self-energies to include more realistic models of scattering and solving equations (1.1) remains a challenge. The resulting matrix $\mathbf{A}$, calculated by (1.2), is shown in Figure 1.4.

The inclusion of decoherence mechanisms introduces an inner self-consistent process in NEGF calculation, which brings huge extra numerical cost in spite of the assumption of diagonal self-energy matrices. Due to the challenge in computation, the phase-breaking scattering is often considered by quasi-1D approximation in mode-space [4, 66], which definitely loses accuracy in simulation results. The NEGF – Poisson modeling infrastructure constructed in the current work contains a full-2D (or 3D) self-consistent solution of electron-phonon scattering self-energies, rather than reduced-dimensional mode space approximations. The simulation can be performed in real space by assuming isotropic scattering and neglecting interband scattering due to electrostatic potential. The governing equations to form a self-consistent computation within NEGF framework will be described in section 6.2.5, yielding a straightforward implementation of the electron-phonon scattering calculation.

1.2.4 Computational Intensity

A complete NEGF-Poisson TCAD simulation requires solving Green's function and Poisson's equation self-consistently. In order to include the phase decoherence mechanism, self-energy matrix should involve nonzero scattering term $\mathbf{\Sigma}^{r,<,>}_{\text{Phonon}}$ which can be obtained from $\mathbf{G}^{<,>}$. Therefore, an extra self-consistent calculation between Green's function and self-energy matrices (due to electron-phonon scattering) is performed at each energy point E. A complete solution of Green's function requires to achieve this self-consistency among a family of energy points $E_1, E_2, \ldots E_n$. Consequently, there are two self-consistent calculations in NEGF-Poisson simulation which are illustrated in Figure 1.5. Table 1.2 summarizes the computational requirement of several physical quantities.

In NEGF simulation, both of the self-consistent calculations (see Figure 1.5) requires solving (1.1) many times until consistency is achieved. It is well appreciated that the computationally intensive part of this calculation is solving (1.1) for the diagonal element of $\mathbf{G}^<$ (electron density) and $\mathbf{G}^>$ at all energies E_i. Although it has been demonstrated that there exist other time-consuming parts, such as calculating contacts self-energy matrices ($\mathbf{\Sigma}^r_L$, $\mathbf{\Sigma}^r_R$) and physical quantities (like current density or transmission coefficient), in practice, comput-

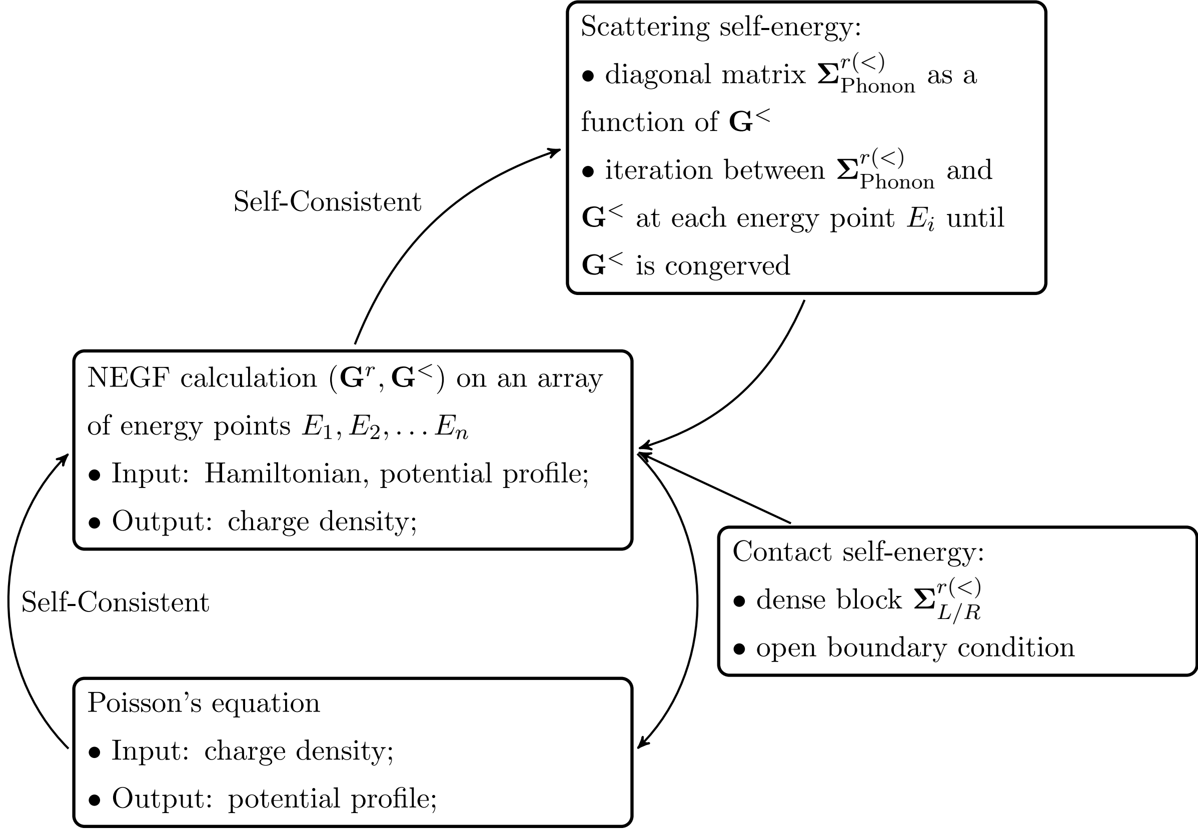

Figure 1.5: Computational intensity varies for different self-energy included. The contact self-energy increases the complexity for Green's function solution at each energy point by introducing dense blocks, while the scattering self-energy brings additional iterative loop with lesser Green's function.

ing the diagonal and required off-diagonal elements of Green's functions remains the most computationally challenging part.

1.3 Review of Exact Algorithms

The most common approach to compute blocks of $\mathbf{G}^r$ and $\mathbf{G}^<$ is the recursive Green's function method (RGF) [4, 29, 30, 31, 40, 63, 66]. RGF is an effective method, often used in practice, to compute $\mathbf{G}^<$ and $\mathbf{G}^>$. For elongated devices, this approach remains the most efficient. Recently, the Hierarchical Schur Complement (HSC) method [46] and the Fast

Physical quantities	$\mathbf{G}^r$	$\mathbf{G}^<$	self-consistent with Poisson solution	self-consistent between $\mathbf{\Sigma}^{r(<)}_{\text{Phonon}}$ and $\mathbf{G}^<$
DOS/ electron density at equilibrium (coherent)	√		√	
DOS/ electron density at non-equilibrium (coherent)	√	√	√	
Transmission coefficient/ current density (coherent)	√	√	√	
Phase de-coherence calculation	√	√	√	√

Table 1.2: NEGF-Poisson simulation requirement of physical quantities. Here $\mathbf{G}^{r(<)}$ only represents the diagonal and required off-diagonal entries, rather than the full matrix.

Inverse using Nested Dissection (FIND) method [42] exploit the nested dissection method [25] to exhibit a significant speedup. Nested dissection recursively partitions the unstructured graph (representing the discretized system) into subgraphs using separators, small subsets of vertices the removal of which allows the graph to be partitioned into disconnected subgraphs with minimized number of vertices. The key ideas behind FIND and HSC algorithms are to partition the whole matrix $\mathbf{A}$ via nested dissection and then to perform efficiently a block LDL^T-factorization. This factorization is then re-used to fill in all diagonal blocks of the Green's function and some off-diagonal blocks in a specific order. These two algorithms are more efficient than RGF and have reduced the operation count down to a multiple of the cost for a block LDL^T-factorization of a sparse matrix.

For calculating the lesser Green's function $\mathbf{G}^<$, advanced algorithms for an arbitrary sparse matrix are still in their infancy. The RGF method remains an effective method, especially for elongated devices. The extension of FIND [43] for $\mathbf{G}^<$ yields a reduced asymptotic complexity but the constant in front of the asymptotic term hinders the reduction in run-

time. Li et al. [43] have proposed a modification of FIND for a significant speedup but their partitioning of the matrix $\mathbf{A}$ requires some pre-processing.

> *We aim to develop an exact method that works in the presence of scattering, with self-energy matrices at least block diagonal. The major contribution of our work is to present an extension of the HSC method for calculating diagonal blocks for* $\mathbf{G}^<$ *with partitions from existing graph partitioning libraries (like, for example, the package METIS [36]).*

This section aims to provide a comprehensive understanding of different algorithms for Green's function calculation. We will provide a mathematical description of block LDL^T-factorization, which is employed to illustrate RGF approach. The computational complexity of the other exact algorithms (FIND and HSC) will be compared with that of the RGF method.

1.3.1 Solving Green's Function by Block LDL^T-Factorization

Before the mathematical description of block LDL^T-factorization, it is worth to give a brief discussion about the symmetry properties utilized in Green's function calculation.

- The resulting retarded Green's function $\mathbf{G}^r$ is a symmetric matrix due to the symmetry of Hamiltonian and self-energy matrices.

$$\mathbf{G}^r = (\mathbf{G}^r)^T$$

- As defined in equations (1.3) to (1.10), the lesser (greater) self-energy matrices are imaginary, therefore, the resulting lesser (greater) Green's function $\mathbf{G}^{<(>)}$ is skew-Hermitian, which satisfies

$$\mathbf{G}^{<(>)\dagger} = -\mathbf{G}^{<(>)}$$

The calculation of the whole matrix $\mathbf{G}^r$ requires the solution of n ($n = N_x N_y$) linear systems, $\mathbf{A}\mathbf{x} = \mathbf{b}$. After computing $\mathbf{\Sigma}^{<(>)}\mathbf{G}^{r\dagger}$, the whole matrix $\mathbf{G}^{<(>)}$ is also obtained after

solving n linear systems. For large-scale applications, computing the whole matrices $\mathbf{G}^r$ and $\mathbf{G}^{<(>)}$ becomes quickly computationally infeasible.

Luckily, for NEGF method, only the diagonal and nearest neighbor off-diagonal elements of $\mathbf{G}^r$ and $\mathbf{G}^{<(>)}$ are computed repeatedly, given that they correspond to physical quantities such as the density of states, electron density, and current (see equations (1.12) to (1.14)). The block LDL^T-factorization of $\mathbf{A}(E)$ provide a systematic approach to calculate the diagonal and the nearest neighbor off-diagonal elements of $\mathbf{G}^r$ and $\mathbf{G}^{<(>)}$ without full inversion of the matrix $\mathbf{A}$. Exploiting the resulting algebraic relations yields an algorithm with a cost significantly smaller than the solution of $2n$ linear systems.

First we review the concept of matrix LDL^T-factorization. As a linear algebra manipulation, the LDU-factorization factors a matrix $\mathbf{A}$ as the product of a lower triangular matrix $\mathbf{L}$, a diagonal matrix $\mathbf{D}$ and a upper triangular matrix $\mathbf{U}$, where diagonal entries for $\mathbf{L}$ and $\mathbf{U}$ are 1. For symmetric matrix $\mathbf{A}$, the factors satisfy $\mathbf{U} = \mathbf{L}^T$, yielding the LDL^T-factorization. The decomposition can be viewed as the Gaussian elimination in matrix format, which is a fundamental matrix inversion method. The LDL^T-factorization of $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T \tag{1.15}$$

Incorporating this relation into the definition of $\mathbf{G}^r$ gives the formula

$$\mathbf{G}^r = (\mathbf{I} - \mathbf{L}^T)\mathbf{G}^r + \mathbf{D}^{-1}\mathbf{L}^{-1}, \tag{1.16}$$

which is the governing equation, used by RGF [4] and HSC [46]. Similarly, the LDL^T-factorization of $\mathbf{A}$ is also generalized in calculating $\mathbf{G}^<$, which is defined as $\mathbf{A}\mathbf{G}^< = \mathbf{\Sigma}^<\mathbf{G}^{r\dagger}$. Petersen et al. [56] generalized Takahashi's method by writing

$$\mathbf{L}^T\mathbf{G}^< \left(\mathbf{L}^T\right)^\dagger = \mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^<\mathbf{L}^{-\dagger}\mathbf{D}^{-\dagger} \tag{1.17}$$

The resulting governing equation used in RGF is [56]

$$\mathbf{G}^< = (\mathbf{I} - \mathbf{L}^T)\mathbf{G}^< + \mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^<\mathbf{L}^{-\dagger}\mathbf{D}^{-\dagger}(\mathbf{L}^T)^{-\dagger} \tag{1.18}$$

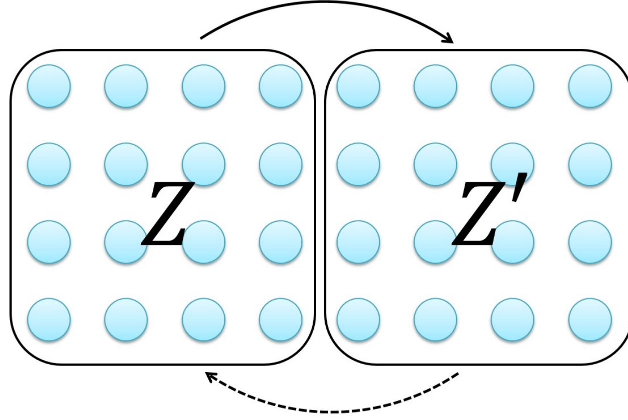

Figure 1.6: System is partitioned into two disjoint regions Z and Z', which has interaction with each other.

Next we illustrate the block LDL^T-factorization by partitioning the discretized system into two disjoint regions Z and Z', shown in Figure 1.6. The matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{ZZ} & \mathbf{A}_{ZZ'} \\ \mathbf{A}_{Z'Z} & \mathbf{A}_{Z'Z'} \end{bmatrix} \tag{1.19}$$

The block LDL^T-factorization of $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{Z'Z}\mathbf{A}_{ZZ}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{ZZ} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{Z'Z'} - \mathbf{A}_{Z'Z}\mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{1.20}$$

Incorporating this relation into the definition (1.1) of $\mathbf{G}^r$ gives

$$\begin{bmatrix} \mathbf{I} & \mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{G}^r = \begin{bmatrix} (\mathbf{A}_{ZZ})^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{A}_{Z'Z'} - \mathbf{A}_{Z'Z}\mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'}\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{Z'Z}\mathbf{A}_{ZZ}^{-1} & \mathbf{I} \end{bmatrix}$$

or

$$\mathbf{G}^r = -\begin{bmatrix} \mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'}\mathbf{G}^r_{Z'Z} & \mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'}\mathbf{G}^r_{Z'Z'} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} (\mathbf{A}_{ZZ})^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{A}_{Z'Z'} - \mathbf{A}_{Z'Z}\mathbf{A}_{ZZ}^{-1}\mathbf{A}_{ZZ'}\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{Z'Z}\mathbf{A}_{ZZ}^{-1} & \mathbf{I} \end{bmatrix}. \tag{1.21}$$

The blocks of $\mathbf{G}^r$ satisfies

$$\begin{aligned}
\mathbf{G}^r_{Z'Z'} &= \left(\mathbf{A}_{Z'Z'} - \mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\right)^{-1} \\
\mathbf{G}^r_{Z'Z} &= -\mathbf{G}^r_{Z'Z'}\mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ} \\
\mathbf{G}^r_{ZZ} &= \mathbf{A}^{-1}_{ZZ} - \mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\mathbf{G}^r_{Z'Z}.
\end{aligned}$$

Once the block LDL^T-factorization of $\mathbf{A}$ is known, the block $\mathbf{G}^r_{Z'Z'}$ is available. Then two different approaches are possible to compute the blocks of $\mathbf{G}^r$:

1. define a family of regions (Z, Z') and calculate, repeatedly, block LDL^T-factorizations for all this regions; this approach is sometimes referred to as a *one-way* method [42];

2. use only one block LDL^T-factorization and back-substitute the blocks to calculate new entries; this approach is sometimes referred to as a *two-way* method [46, 56, 66].

Similarly, the block LDL^T-factorization of $\mathbf{A}$ may be incorporated into the definition of $\mathbf{G}^<$,

$$\mathbf{A}\mathbf{G}^< = \Sigma^<\mathbf{G}^{r\dagger},$$

as follows

$$\begin{bmatrix} \mathbf{I} & \mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{G}^< = \begin{bmatrix} (\mathbf{A}_{ZZ})^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{A}_{Z'Z'} - \mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ} & \mathbf{I} \end{bmatrix} \Sigma^<\mathbf{G}^{r\dagger} \quad (1.22)$$

or

$$\mathbf{G}^< = -\begin{bmatrix} \mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\mathbf{G}^<_{Z'Z} & \mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\mathbf{G}^<_{Z'Z'} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} (\mathbf{A}_{ZZ})^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{A}_{Z'Z'} - \mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ} & \mathbf{I} \end{bmatrix} \Sigma^<\mathbf{G}^{r\dagger}. \quad (1.23)$$

The blocks of $\mathbf{G}^<$ are defined as follows

$$\begin{aligned}
\mathbf{G}^<_{Z'Z'} &= \mathbf{G}_{Z'Z'}\left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{Z'Z'} \\
\mathbf{G}^<_{Z'Z} &= -\mathbf{G}_{Z'Z'}\mathbf{A}_{Z'Z}\mathbf{A}^{-1}_{ZZ}\left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{ZZ} + \mathbf{G}_{Z'Z'}\left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{Z'Z} \\
\mathbf{G}^<_{ZZ} &= \mathbf{A}^{-1}_{ZZ}\left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{ZZ} - \mathbf{A}^{-1}_{ZZ}\mathbf{A}_{ZZ'}\mathbf{G}^<_{Z'Z}.
\end{aligned}$$

The computational cost of evaluating these relations depends, of course, on the expression for the self-energy $\Sigma^<$. In the particular case where $\Sigma^<$ is block diagonal, the blocks of $\Sigma^<\mathbf{G}^{r\dagger}$ satisfy

$$\left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{Z'Z'} = \Sigma^<_{Z'Z'}\mathbf{G}^{r\dagger}_{Z'Z'} \quad \text{and} \quad \left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{ZZ} = \Sigma^<_{ZZ}\mathbf{G}^{r\dagger}_{ZZ}$$

and

$$\left(\Sigma^<\mathbf{G}^{r\dagger}\right)_{Z'Z} = \Sigma^<_{Z'Z'}\mathbf{G}^{r\dagger}_{ZZ'}.$$

Once the block LDL^T-factorization of $\mathbf{A}$, the self-energy $\Sigma^<$, and the diagonal elements of $\mathbf{G}^r$ are known, the block $\mathbf{G}^<_{Z'Z'}$ is available. Again two different approaches are possible to compute the blocks of $\mathbf{G}^<$:

1. define a family of regions (Z, Z') and calculate, repeatedly, block LDL^T-factorizations for all this regions; this approach is still referred to as a *one-way* method [43];

2. use only one block LDL^T-factorization and back-substitute the blocks to calculate new entries of $\mathbf{G}^<$; this approach is still referred to as a *two-way* method [46, 56, 66].

1.3.2 RGF Method

In section 1.3.1, we describe the resolution of Green's function by exploiting block LDL^T-factorization and corresponding recurrence formulas on the system decomposed into two disjoint clusters. RGF approach extends this scheme to more general cases. The main idea is to partition the nanodevice into N_y disjoint layers, interacting only with their nearest neighbors (N_y layers horizontally aligned with the transport direction in Figure 1.1, with each

layer containing N_x grid points). Given the system decomposition into layered-structure, one can calculate Green's functions by using the similar block LDL^T-factorization and recurrence strategy.

State-of-the-art *two-way* RGF algorithms were developed in [40, 66] for cases where the matrix $\mathbf{A}$ is block tridiagonal. RGF is an algorithm composed of two passes to compute $\mathbf{G}^r$ and two passes to compute $\mathbf{G}^<$. In both cases, the passes are interpreted as follows:

1. the first pass marches one layer at a time from *left to right* along the y-direction and, recursively, *folds* the effect of left layers into the current layer;

2. the second pass marches one layer at a time from *right to left* along the y-direction and, recursively, *extracts* the diagonal blocks and the nearest neighbor off-diagonal blocks for the final result.

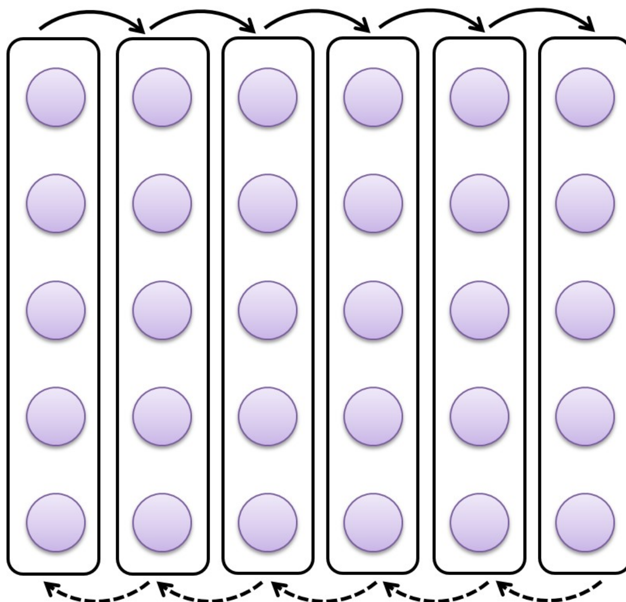

Figure 1.7: RGF partitions the systme into disjoint layers with interaction at specific direction. Solid arrows and dash arrows relate to the *fold* and *extract* passes respectively in computing Green's functions.

The block LDL^T-factorization provides a systematic description of RGF method process, by performing block LDL^T-factorization between two disjoint layers *forward* at one direction (solid arrow in Figure 1.7) and *backward* again (dashed arrow in Figure 1.7). Then the algorithm performs the following steps:

1. compute one block LDL^T-factorization of $\mathbf{A}$, marching one layer at a time from left to right in the transport direction, often labelled the y-direction (a.k.a. the *forward* pass for $\mathbf{G}^r$);

2. apply (1.16) recursively, marching one layer at a time from right to left in the transport direction (a.k.a. the *backward* pass for $\mathbf{G}^r$);

3. calculate the right hand side of (1.17), marching one layer at a time from left to right in the transport direction (a.k.a. the *forward* pass for $\mathbf{G}^<$);

4. apply (1.18) recursively, marching one layer at a time from right to left in the transport direction (a.k.a. the *backward* pass for $\mathbf{G}^<$).

Recall that the full inversion of an $N \times N$ matrix takes $\mathcal{O}(N^3)$ operations, indicating that the full inversion of $\mathbf{A}$ would require $\mathcal{O}\left(N_x^3 N_y^3\right)$ operations. Numerically, it is essential to notice that the RGF method exploits the matrix sparsity *only* at the block level, which means that it separates the whole problem into sub-problems of *full* matrix operations. The operation count for the algorithms [40, 66] scales like $\mathcal{O}\left(N_x^3 N_y\right)$. The dependence on N_x^3 arises because only blocks of dimension $N_x \times N_x$ are inverted and multiplied. These inversions and multiplications are repeated recursively for N_y layers. The reduction from $\mathcal{O}\left(N_x^3 N_y^3\right)$ to $\mathcal{O}\left(N_x^3 N_y\right)$ is dramatic but is limited to the case where the matrix $\mathbf{A}$ is block tridiagonal. The complexity of this method is, at most, $10 N_x^3 N_y$ (when $N_x \leqslant N_y$). Further details about RGF, including mathematical proof and code, are presented in [4, 66].

1.3.3 Comments on Exact Algorithms

To compute block entries of $\mathbf{G}^r$, two recent advances, namely FIND [42, 43] and HSC [46, 47], utilize the nested dissection method [25] to exhibit a significant speedup. These methods explicitly exploit the sparsity of $\mathbf{A}$ via a sparse block LDL^T-factorization of the whole matrix and re-use this factorization to fill in all diagonal blocks of the Green's function and some off-diagonal blocks in a specific order. FIND and HSC have a strong mathematical component and their physical interpretation is less obvious. The main difference between RGF and these methods is the replacement of *layers* of grid points organized along a specific direction with *arbitrarily-shaped clusters* of grid points organized in a binary tree. Such choice allows to *fold* and to *extract* in any physical direction when following the vertical hierarchy of the binary tree. Further details about FIND and HSC can be found in their respective references. Table 1.3 summarizes the complexity of these three state-of-the-art algorithms when computing entries in $\mathbf{G}^r$.

Algorithm	Complexity when $N_x = N_y$	Complexity when $N_x < N_y$
RGF [66]	$\mathcal{O}(N_x^4)$	$\mathcal{O}(N_x^3 N_y)$
FIND [42, 56]	$\mathcal{O}(N_x^3)$	$\mathcal{O}(N_x^2 N_y)$
HSC [46, 56]	$\mathcal{O}(N_x^3)$	$\mathcal{O}(N_x^2 N_y)$

Table 1.3: Complexity of algorithms to compute diagonal blocks of $\mathbf{G}^r$.

To compute diagonal entries for $\mathbf{G}^<$, only the RGF [4] and FIND [43] methods have been extended. Table 1.4 displays the runtime complexity for computing diagonal blocks of $\mathbf{G}^<$. We aim to present an extension of the HSC method [46] for the calculation of diagonal blocks for $\mathbf{G}^<$, which is compatible with existing partitioning libraries — namely, the extension is combined with the graph partitioning package METIS [36].

Algorithm	Complexity when $N_x = N_y$	Complexity when $N_x < N_y$
RGF [66]	$\mathcal{O}\left(N_x^4\right)$	$\mathcal{O}\left(N_x^3 N_y\right)$
FIND [43]	$\mathcal{O}\left(N_x^3\right)$	$\mathcal{O}\left(N_x^2 N_y\right)$

Table 1.4: Complexity of algorithms to compute diagonal blocks of $\mathbf{G}^<$.

Chapter 2

MATHEMATICAL DESCRIPTION OF THE ALGORITHM

In this chapter, a detailed mathematical description for the extension of HSC to compute blocks of $\mathbf{G}^<$ is given. The key ingredients are:

- an efficient sparse block LDL^T-factorization of $\mathbf{A}$; The block sparse factorization will gather grid points into arbitrarily-shaped clusters (instead of layers, like in RGF). Such choice allows to fold and to extract in any physical direction when eliminating entries in $\mathbf{A}$. The factorization yields formulas to calculate the diagonal blocks and off-diagonal blocks for $\mathbf{G}^r$ and $\mathbf{G}^<$. Exploiting the resulting algebraic relations results in an algorithm with a cost significantly smaller than the full inversion of matrix $\mathbf{A}$.

- an appropriate order of operations. The cost of a matrix multiplication $\mathbf{BCD}$ depends on the order of operations. When $\mathbf{B}$ is $m \times p$, $\mathbf{C}$ is $p \times k$, and $\mathbf{D}$ is $k \times n$, $(\mathbf{BC})\,\mathbf{D}$ costs $2mk(n+p)$ operations and $\mathbf{B}\,(\mathbf{CD})$ costs $2np(m+k)$. The order of operations can have a large effect when multiplying series of matrices together (which is the case for computing entries of $\mathbf{G}^r$ and $\mathbf{G}^<$). Furthermore, when working with sparse matrices, one order of operations may preserve sparsity, while another may not.

First, a simple description with three clusters is given. Then the approach is extended to an arbitrary number of levels and a multilevel binary tree.

2.1 *Description for a Simple Case*

The basic idea is to partition the nano-device into three disjoint regions (L, R, S) — see Figure 2.1

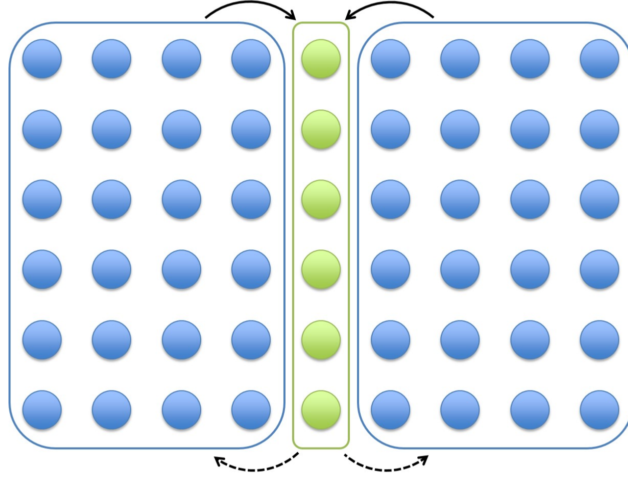

Figure 2.1: Nanodevice partitioned into two subregions (L, R) and a separator S (L stands for Left and R for Right). Solid arrows and dash arrows relate to the *fold* and *extract* passes respectively in computing Green's functions.

Such a partition is easily obtained via the nested dissection, introduced by George [25]. Nested dissection divides the system into two disconnected sets and an interface, called the separator S. With this partition, the matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{LL} & \mathbf{0} & \mathbf{A}_{LS} \\ \mathbf{0} & \mathbf{A}_{RR} & \mathbf{A}_{RS} \\ \mathbf{A}_{LS}^T & \mathbf{A}_{RS}^T & \mathbf{A}_{SS} \end{bmatrix}$$

Similar to equation (1.19), here L + R $\Rightarrow Z$ and S $\Rightarrow Z'$. According to the block LDL^T-factorization (1.15), the calculation of $\mathbf{G}^r$ can be performed by the same *fold* and *extract* scheme, which is also used by RGF.

For calculating entries in the correlation matrix $\mathbf{G}^<$, our approach utilizes a variation of (1.17) and (1.18), namely

$$\mathbf{L}^T\mathbf{G}^< = \mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger \tag{2.1}$$

and

$$\mathbf{G}^< = \left(\mathbf{I} - \mathbf{L}^T\right)\mathbf{G}^< + \mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger \tag{2.2}$$

and exploits the symmetry of $\mathbf{G}^r$ and the skew-Hermitian property of $\mathbf{\Sigma}^<$ and $\mathbf{G}^<$,

$$(\mathbf{G}^<)^\dagger = -\mathbf{G}^< \quad \text{and} \quad (\mathbf{\Sigma}^<)^\dagger = -\mathbf{\Sigma}^< \tag{2.3}$$

A complete derivation for HSC-extention can be found at Appendix A. In our HSC-extension algorithm, the calculation of $\mathbf{G}^r$ matches exactly the HSC method [46]. However, the computational process for calculating the correlation matrix $\mathbf{G}^<$ differs from RGF and FIND methods [43, 44], in several aspects. It uses thin boundaries obtained directly from the nested dissection. This extension requires one sparse factorization and one back-substitution (*two-way*), while FIND utilizes only sparse factorizations but applied many times with different orderings (*one-way*). The order of operations to obtain the diagonal blocks of $\mathbf{G}^<$ is also different from the recurrence in Petersen et al. [56], which uses the sequence

$$\begin{aligned}\mathbf{\Sigma}^< &\to \mathbf{L}^{-1}\mathbf{\Sigma}^< \left(\mathbf{L}^{-1}\right)^\dagger \to \mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^< \left(\mathbf{L}^{-1}\right)^\dagger \left(\mathbf{D}^{-1}\right)^\dagger \\ &\to \mathbf{L}^{-T}\mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^< \left(\mathbf{L}^{-1}\right)^\dagger \left(\mathbf{D}^{-1}\right)^\dagger \left(\mathbf{L}^{-T}\right)^\dagger,\end{aligned}$$

while our HSC extension uses

$$\mathbf{\Sigma}^< \to \mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger \to \mathbf{L}^{-1}\mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger \to \mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger \to \mathbf{L}^{-T}\mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger$$

When working with sparse matrices, specific order of operations may result in fewer operations. In numerical experiments, the latter ordering was more efficient.

2.2 Description for a Multilevel Case

In this section, the description is extended to an arbitrary number of clusters.

Even though computing the diagonal for the inverse of a matrix is not equivalent to a sparse factorization, both problems benefit from matrix reordering. The multilevel nested dissection, introduced by George [25], lends itself naturally to the creation of grid points clusters. Typically, nested dissection divides the system into two disconnected sets and an interface, called the separator. Then the process is repeated recursively on each set to create a multilevel binary tree as shown in Figure 2.2. The system is thus re-organized into a

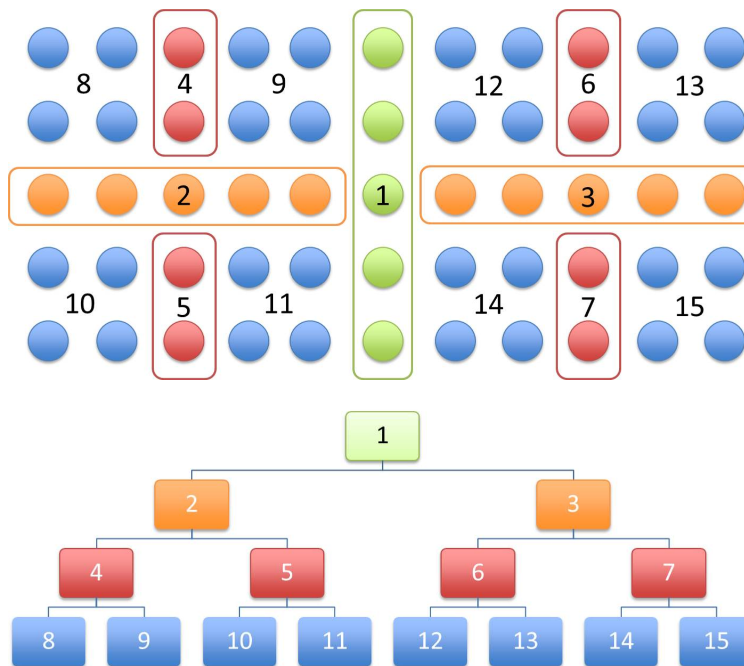

Figure 2.2: Example of a multilevel partition.

hierarchical structure, starting from the top level *(root)* to the lowest level *(leaf)*. From the mathematical derivation expressed in previous section, the *two-way* HSC-extension algorithm can be extended to a general multilevel partition case by performing the following steps:

1. compute one block LDL^T-factorization of $\mathbf{A}$ by 1.15, marching one level at a time from leaves to root (a.k.a. the *forward* pass for $\mathbf{G}^r$);

2. apply (1.16) recursively, marching one level at a time from root to leaves (a.k.a. the *backward* pass for $\mathbf{G}^r$);

3. calculate the right hand side of (2.1), marching one level at a time from leaves to root (a.k.a. the *forward* pass for $\mathbf{G}^<$);

4. apply (2.2) recursively, marching one level at a time from root to leaves (a.k.a. the *backward* pass for $\mathbf{G}^<$).

For the sake of clarification, we provide an example of computational steps in matrix notations of HSC-extension algorithm for a three-level system in Appendix B. The remaining of the section describes the computational process of HSC-extenion algorithm.

Based on the hierarchical structure of the tree shown in Figure 2.2, let P_i denote the set of all cluster indices j such that cluster j is an ancestor of cluster i. For example for Figure 2.2, P_5 is equal to $\{1, 2\}$ and $P_{15} = \{1, 3, 7\}$. Let C_i denote the set of all cluster indices j such that cluster j is a descendant of cluster i. For the partition on Figure 2.2, C_4 is equal to $\{8, 9\}$ and $C_3 = \{6, 7, 12, 13, 14, 15\}$. Note that a cluster may or may not have a direct coupling in the matrix $\mathbf{A}$ to any of its ancestors or descendants.

Once the partition is set, the algorithm may be separated into two distinct parts: computation of $\mathbf{G}^r$ and computation of $\mathbf{G}^<$.

2.2.1 Computation of Blocks for $\mathbf{G}^r$

In the binary tree, the levels are labeled from bottom to top, where level 1 contains all clusters at the end of the tree and level L contains only the original separator. For simplicity of presentation, let $\mathbf{A}^{(l)}$ denote the matrix transformed from $\mathbf{A}$ after folding all the clusters up to level l. Note that $\mathbf{A}^{(0)}$ is set to $\mathbf{A}$ and $\mathbf{A}^{(L-1)}$ is block diagonal.

The computation of blocks for $\mathbf{G}^r$ involve three steps: folding the lower level clusters unto the higher ones, inversion of the matrix for the main separator, and extracting of the diagonal blocks for the current level from blocks on higher level.

The algorithm for the first step goes as follows:

- For $l = 1$ up to $L - 1$,
  - $\mathbf{A}^{(l)} = \mathbf{A}^{(l-1)}$
  - For all the clusters i on level l,
    - $\mathbf{\Psi}_{i,j} = -\left(\mathbf{A}^{(l)}_{i,i}\right)^{-1} \mathbf{A}^{(l)}_{i,j}$ for all j in P_i
    - $\mathbf{A}^{(l)}_{j,k} = \mathbf{A}^{(l)}_{j,k} + \mathbf{\Psi}^T_{i,j} \mathbf{A}^{(l)}_{i,k}$ for all j and k in P_i

    - $\mathbf{A}_{k,j}^{(l)} = \left(\mathbf{A}_{j,k}^{(l)}\right)^T$ for all j and k in P_i
    - $\mathbf{A}_{i,j}^{(l)} = \mathbf{0}$ and $\mathbf{A}_{j,i}^{(l)} = \mathbf{0}$ for all j in P_i
  - end
- end

The next step is written as the inversion of $\mathbf{A}^{(L-1)}$, which is symmetric and block diagonal.

- $\mathbf{G}^{(L-1)} = \left(\mathbf{A}^{(L-1)}\right)^{-1}$

In practice, the operation requires only the inversion of the block for the top separator. All the other blocks have already been inverted during the folding steps.

Finally, all the diagonal blocks of $\mathbf{G}^r$ are extracted one level at a time. The algorithm goes as follows:

- For $l = L-2$ down to 0,
  - $\mathbf{G}^{(l)} = \mathbf{G}^{(l+1)}$
  - For all the clusters i on level l,
    - $\mathbf{G}_{i,j}^{(l)} = \mathbf{G}_{i,j}^{(l)} + \sum_{k \in P_i} \mathbf{\Psi}_{i,k} \mathbf{G}_{k,j}^{(l)}$ for all cluster indices j in P_i
    - $\mathbf{G}_{j,i}^{(l)} = \left(\mathbf{G}_{i,j}^{(l)}\right)^T$ for all cluster indices j in P_i
    - $\mathbf{G}_{i,i}^{(l)} = \mathbf{G}_{i,i}^{(l)} + \sum_{j \in P_i} \mathbf{\Psi}_{i,j} \mathbf{G}_{j,i}^{(l)}$
  - end
- end

The resulting algorithm to compute block entries in $\mathbf{G}^r$ matches exactly the HSC method [46].

Note that matrix $\mathbf{G}^{(0)}$ is not equal to $\mathbf{G}^r$ because $\mathbf{G}^{(0)}$ is incomplete (see an example in the Appendix). However, all the entries in $\mathbf{G}^{(0)}$, in particular the diagonal entries, match the corresponding entries in $\mathbf{G}^r$.

2.2.2 Computation of Blocks for $\mathbf{G}^<$

The algorithm consists of four steps. The first step uses the matrix $\mathbf{G}^{(0)}$ computed previously.

- $\mathbf{N} = \mathbf{\Sigma}^< \left(\mathbf{G}^{(0)}\right)^\dagger$

All the entries in $\mathbf{G}^{(0)}$ match the corresponding entries in $\mathbf{G}^r$ and are sufficient to compute diagonal blocks of $\mathbf{G}^<$. The matrix $\mathbf{\Sigma}^<$ is typically block diagonal. The matrix $\mathbf{N}$ will have the same structure and shape as $\mathbf{G}^{(0)}$. The matrix multiplication is done block by block.

Next the lower level clusters are folded into the higher ones. This step is critical and the most time consuming. Let $\mathbf{N}^{(l)}$ denote the matrix transformed from $\mathbf{N}$ after folding all the clusters up to level l. $\mathbf{N}^{(0)}$ is set to $\mathbf{N}$.

- For $l = 1$ up to $L - 1$,
  - $\mathbf{N}^{(l)} = \mathbf{N}^{(l-1)}$
  - For all the clusters i on level l,
    - $\mathbf{N}^{(l)}_{j,k} = \mathbf{N}^{(l)}_{j,k} + \mathbf{\Psi}^T_{i,j}\mathbf{N}^{(l)}_{i,k}$ for all j and k in P_i
  - end
- end

Similarly to Step 1, Step 3 is a block diagonal multiplication.

- $\mathbf{P}^{(L-1)} = \mathbf{G}^{(L-1)}\mathbf{N}^{(L-1)}$

Finally, Step 4 extracts all the diagonal blocks one level at a time. This step is similar to the extraction in the $\mathbf{G}^r$ algorithm. The operations are the following:

- For $l = L - 2$ down to 0,
  - $\mathbf{P}^{(l)} = \mathbf{P}^{(l+1)}$

- For all the clusters i on level l,
  - $\mathbf{P}_{i,j}^{(l)} = \mathbf{P}_{i,j}^{(l)} + \sum_{k \in P_i} \mathbf{\Psi}_{i,k} \mathbf{P}_{k,j}^{(l)}$ for all cluster indices j in P_i
  - $\mathbf{P}_{j,i}^{(l)} = -\left(\mathbf{P}_{i,j}^{(l)}\right)^\dagger$ for all cluster indices j in P_i
  - $\mathbf{P}_{i,i}^{(l)} = \mathbf{P}_{i,i}^{(l)} + \sum_{j \in P_i} \mathbf{\Psi}_{i,j} \mathbf{P}_{j,i}^{(l)}$
- end

• end

At the end, matrix $\mathbf{P}^{(0)}$ is not equal to $\mathbf{G}^<$ because $\mathbf{P}^{(0)}$ is incomplete (see an example in the Appendix). However, all the entries in $\mathbf{P}^{(0)}$, in particular the diagonal entries, match the corresponding entries in $\mathbf{G}^<$.

To analyze the speedup for HSC-extension, we emphasize that RGF exploits the matrix sparsity only at the block level, *i.e.* it merely separates the system in y direction and take the clusters in x direction as a dense block. Compared to RGF, HSC-extension exploits the system sparsity in both x and y direction. The arbitrarily shaped clusters organized in a hierarchical binary tree significantly reduces the computational cost. On the other hand, in HSC-extension, we do $\mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger$ instead of the common factorization $\mathbf{\Sigma}^< \mathbf{L}^{-\dagger} \mathbf{D}^{-\dagger} (\mathbf{L}^T)^{-\dagger}$ (see (2.1) and (2.2)), since the latter expression is less efficient in numerical computation. However, from the derivation in Appendix B, it is observed that this variation of recurrence formula requires computing extra $\mathbf{G}^r$ entries compared to the HSC approach [46]. As a result, the speedup of $\mathbf{G}^<$ computation (benchmarked against RGF) is smaller than the speedup of $\mathbf{G}^r$ computation for HSC-extension, although the asymptotic complexity of $\mathbf{G}^<$ remains to be $\mathcal{O}(N_x^2 N_y)$. With the HSC-extension algorithm built and incorporated, a complete NEGF calculation process is given in Figure 2.3.

2.3 Comments on the System Partition

The partitioning of the system (or the clustering of points) is the key step for the efficiency of this algorithm. The partition should follow two rules:

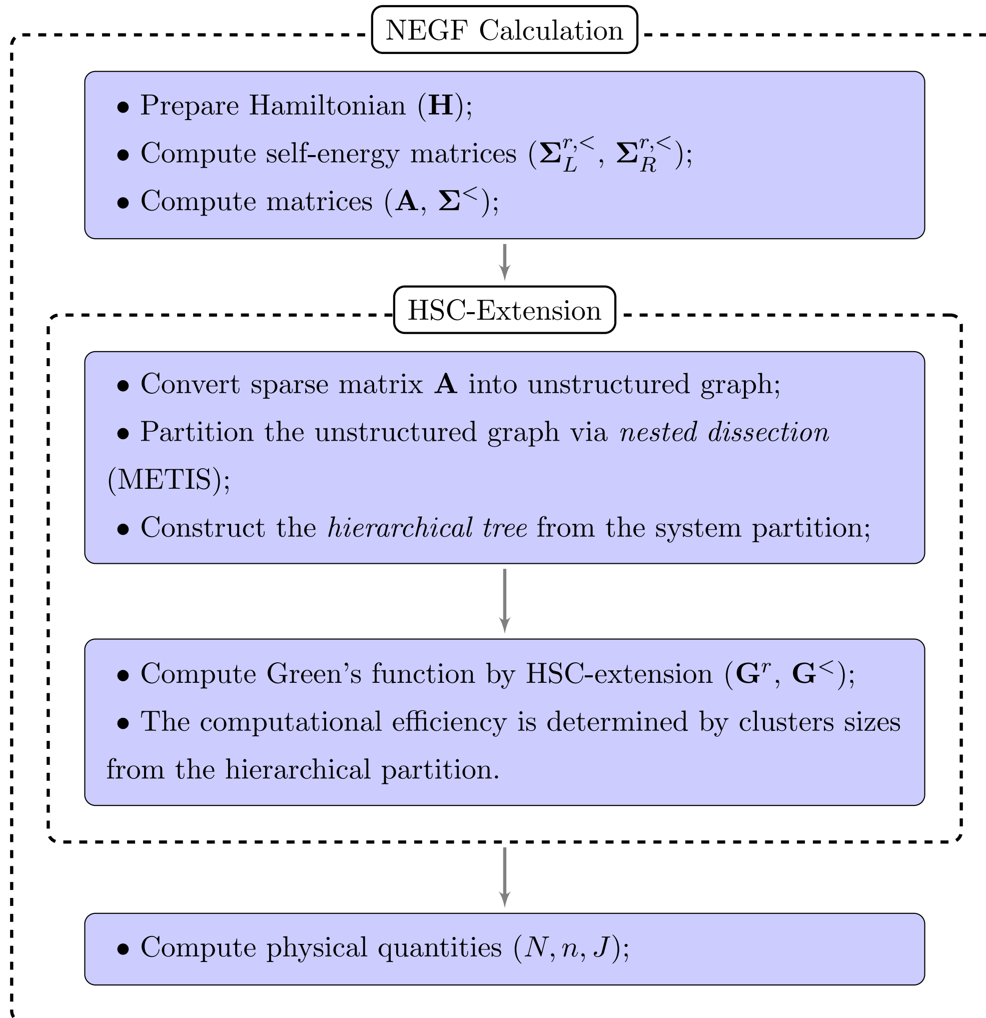

Figure 2.3: Flowchart for NEGF calculation incorporated with HSC-extension algorithm.

- for clusters within the same level on the binary tree, no interaction is allowed. Operations on blocks at the same level are performed independently.

- the partition should minimize the size of separators and reduce the clusters down to a size manageable for an inversion of the corresponding block matrix.

The multilevel nested dissection generates a partition that satisfies those rules. It is worthwhile to note that, as long as the rules stated above are followed, systems with non-uniform distribution of points or with a different stencil could be treated correctly.

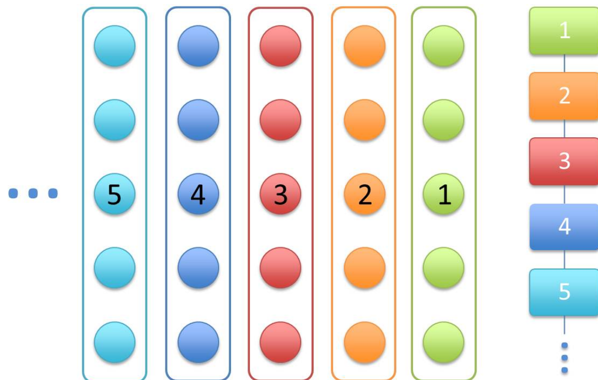

Figure 2.4: Partition generating the RGF algorithm.

The RGF algorithm is included in the previous description with a very specific partition. It re-organizes the nanodevice into N_y disjoint layers. Therefore the system can be taken as partitioned into N_y layers and linear hierarchy tree is performed, illustrated in Figure 2.4.

In many cases, self-energy functions add two $N_x \times N_x$ dense blocks into the input sparse matrix $\mathbf{A}$ in first and last block (see Figure 1.4). One possible partition combines the two contacts together with the middle separator 1, as shown in Figure 2.5.

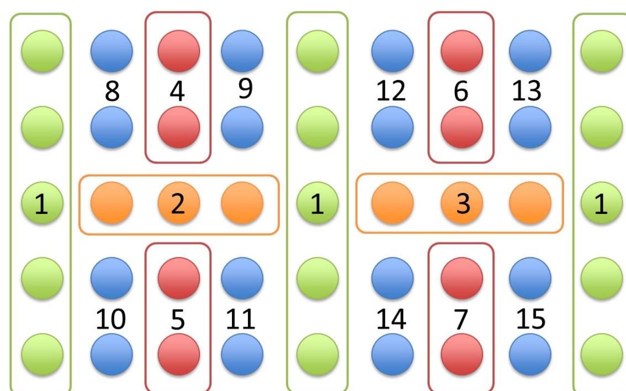

Figure 2.5: First method to partition system with two dense layers and two ends.

The weakness of such clustering is the size of the first separator (or root region). A larger separator increases the computational cost spent on this level. More descendant blocks are coupled with this separator and the total number of operations will increase dramatically. Another partition that satisfies the previous two rules is plotted in Figure 2.6.

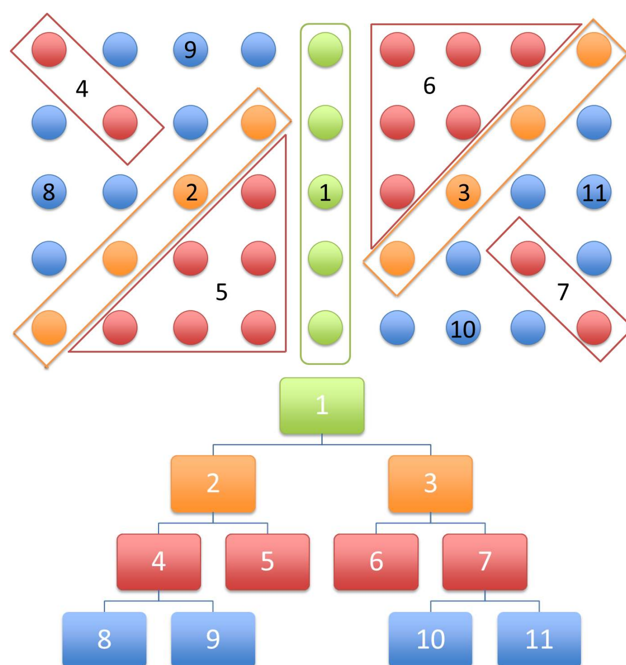

Figure 2.6: Partition generated by METIS for system including dense layers at two ends.

Finally, we summarize the features of RGF, FIND and HSC(-extension) approaches in table 2.1.

RGF[66]	• *two-way* method; • Linear hierarchy: layers organized in transport direction; • Perform inversion or multiplication of $N_x \times N_x$ blocks by number of $\mathcal{O}(N_y)$ times;
FIND[42, 56]	• *one-way* method; • Binary-tree hierarchy: arbitrarily-shaped clusters with thick separators; • Perform inversion or multiplication of block with size much smaller than N_x, but with operation number larger than $\mathcal{O}(N_y)$ times;
HSC (extension)[46, 56]	• *two-way* method; • Binary-tree hierarchy: arbitrarily-shaped clusters with thin separators; • Perform inversion or multiplication of block with size much smaller than N_x, but with operation number larger than $\mathcal{O}(N_y)$ times; • $\mathbf{G}^{<}$ requires calculation of extra entries in $\mathbf{G}^{r}$;

Table 2.1: Features summarized for three exact NEGF algorithms, RGF, FIND and HSC(-extension)

Chapter 3

NUMERICAL EXPERIMENTS FOR 2D STRUCTURES

This section describes numerical experiments on two simple models: a super-lattice structure and a graphene nanotube. Both algorithms (RGF and HSC-extension) are implemented as C codes (HSC-extension is interfaced with METIS [36]). All the runtime data corresponds to the total CPU time for the evaluation of the diagonal and desired off-diagonal entries of $\mathbf{G}^r$ and $\mathbf{G}^<$ at a single energy point. All numerical experiments are performed with one thread on a machine with Intel i7-2600 3.40GHz CPU and 12GB memory.

3.1 Cost Analysis

First the complexity of the HSC extension is compared numerically to the complexity of RGF. A model device is considered where the system Hamiltonian is discretized with a five-point stencil. The left and right contact self-energies are neglected for this section only. A typical partition is plotted in Figure 2.2

The numerical estimate tracks the operation counts step by step for all the matrix multiplications and matrix inversions throughout the code. For a multiplication of two matrices with dimensions $i \times j$ and $j \times k$, a total of ijk operations is added. For inversion of a matrix of dimension $i \times i$, i^3 operations are counted.

Figure 3.1 shows the cost comparison between the HSC extension and RGF for two-dimensional square systems with the same number of grid points (or atoms) per direction, *i.e.* $N_x = N_y = N$. The plot in logarithmic scale indicates that RGF exhibits a complexity of $\mathcal{O}(N^4)$, while the HSC-extension shows a $\mathcal{O}(N^3)$ complexity.

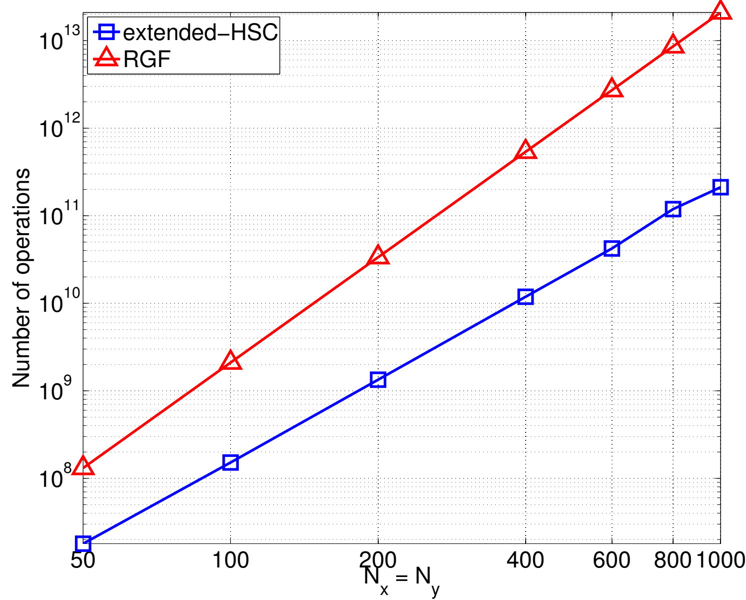

Figure 3.1: Numerical count comparison for our algorithm (blue) and RGF (red).

3.2 Results

3.3 Super-lattice Device

A super-lattice device is typically a multi-layered energy barriers system. The device is composed of repeating junctions of energy barriers and wells. To verify the simulation results, a two-dimensional system of lengths $l_x = 25$nm and $l_y = 20$nm is considered and plotted in Figure 3.2. Here the structure has eight barriers, each of width 1nm and of height 400meV. The wells have a width of 1nm. The length of the left flat band region is 2nm and the right flat band region is 3nm long.

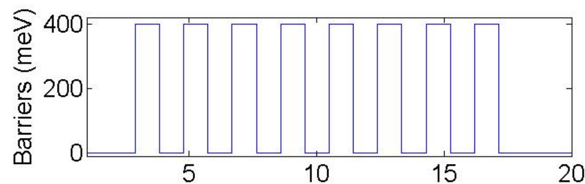

Figure 3.2: Barrier structure for a model super-lattice device.

A simulation with a five-point stencil discretization on a grid with spacing dx = dy = 0.1nm is made for the Fermi energy E_f = 140meV and 500 energy points uniformly distributed between 0 to 500meV. The density of states, electron density, and current are calculated by the RGF method and the extended-HSC method. The output electron density and its linear distribution in y direction is plotted in Figure 3.3. The figures indicate that the charge distribution in the barrier-well multi-layer junctions are symmetric, as expected.

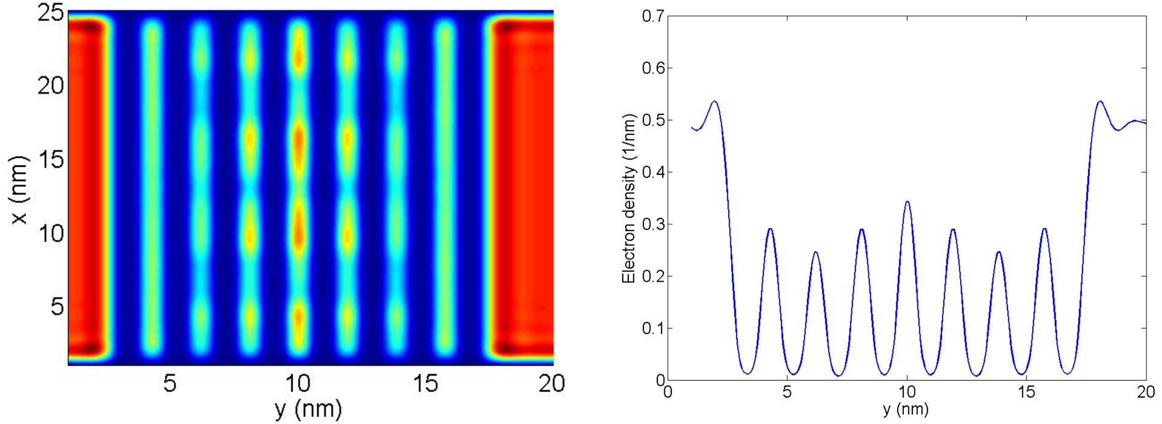

Figure 3.3: Electron density profile and electron density in y direction for a model superlattice device.

Next devices of lengths $l_x = N_x \times 0.1$nm and $l_y = N_y \times 0.1$nm are used to compare the two algorithms. The number of barriers is kept at 8. The lengths for the two-sides flat region are adjusted according to the lengths of device l_x and l_y. The other parameters remain unchanged.

In Figure 3.4(a), diagonal self-energy matrices are used for the left and right contacts. Calculation times are compared for square systems— *i.e.* $N_x = N_y = N$ — and plotted in Figure 3.4(a). As expected, the HSC-extension exhibits smaller CPU times and a complexity of $\mathcal{O}(N^3)$ while RGF's complexity is $\mathcal{O}(N^4)$.

In Figure 3.4(b), CPU times for square systems with dense self-energy matrices for both contacts are plotted. Here again the HSC-extension exhibits smaller CPU times. A com-

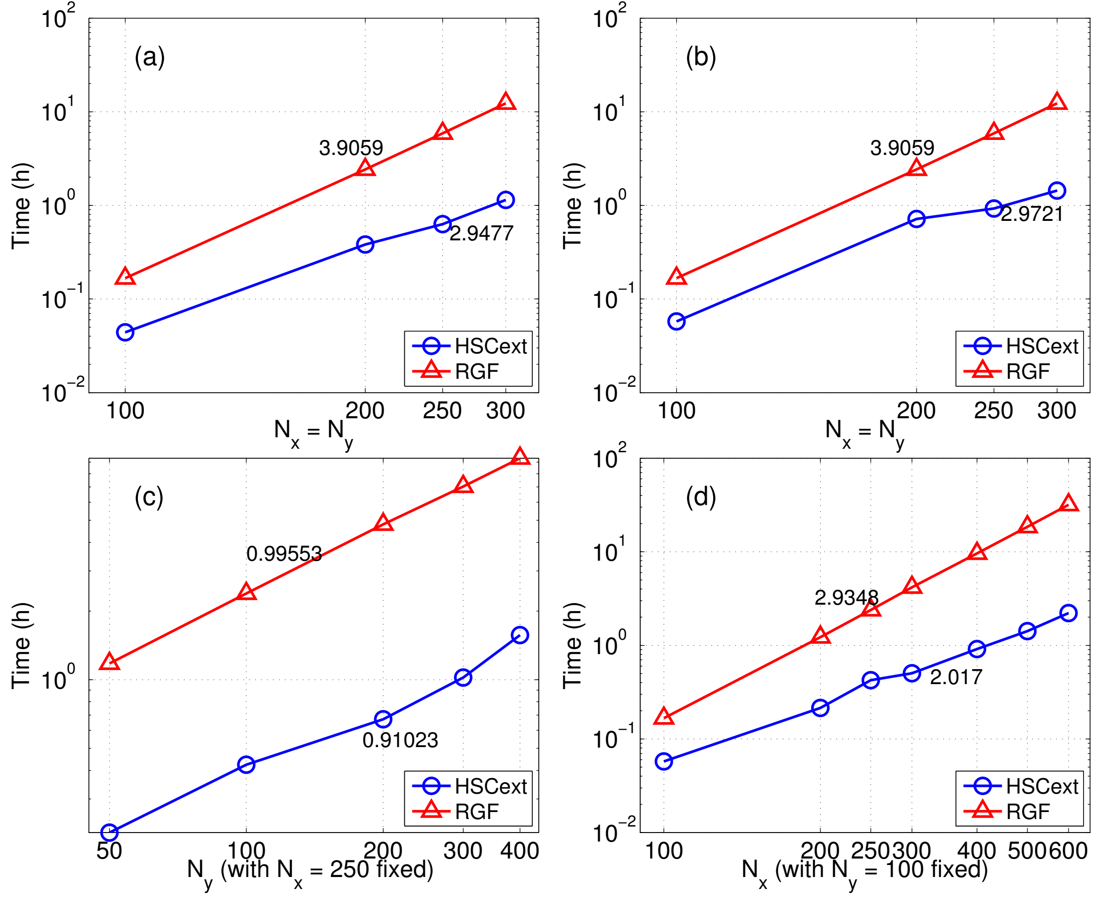

Figure 3.4: Superlattice device NEGF simulation computation time comparison for RGF and our methods, all systems grid spacing is 0.1nm. (a) Square system of with diagonal self-energy matrix; (b) Square system of with dense self-energy matrix; (c) For systems in this plot, the length in the x-direction is fixed at 25 nm while the length in the y-direction is increased. (d) For systems in this plot, the length in the y-direction is fixed at 10 nm while the length in the x-direction is increased. Dense self-energy matrices are used in (c) and (d) devices.

plexity $\mathcal{O}(N^3)$ for HSC-extension compared with $\mathcal{O}(N^4)$ for RGF can be seen.

Figure 3.4(c) plots CPU times for rectangular devices where $l_x = 25$nm (or $N_x = 250$) and the length in the y-direction is varied. Dense self-energy blocks for the left and right contacts are employed. The implementation of RGF is biased towards the x-direction so that its complexity is $\mathcal{O}(N_y)$. The linear trend is clearly present in the plot. For the HSC-extension, similarly to the cost of a sparse LDL^T factorization, the computational cost is $\mathcal{O}(N_y)$. Clearly, the constant for RGF is larger for this device.

To illustrate the dependence of this constant with respect to N_x, Figure 3.4(d) plots the CPU times when N_y is fixed and N_x is varied. The recorded CPU times illustrate that the RGF method has an asymptotic complexity $\mathcal{O}(N_x^3)$, while the HSC extension exhibits a complexity $\mathcal{O}(N_x^2)$. So, for rectangular devices, the RGF method has a complexity $\mathcal{O}(N_x^3 N_y)$ and the HSC-extension a complexity $\mathcal{O}(N_x^2 N_y)$.

3.4 Graphene

Graphene is one of the most promising next-generation materials. Its remarkable electric properties, such as high carrier mobility and zero band gaps, generate a rapidly increasing interest in the electronic device community. Since 2007, many advances in graphene-based transistor development have been reported. [61]

The NEGF simulation of graphene transport is based on tight binding method, which yields a four-point-stencil Hamiltonian due to system decomposition of carbon atoms coupling (see Figure 3.5). In the numerical experiments, armchair planar graphene nanoribbon structures are simulated. The on-site energy for each carbon atom is 0 eV and the hopping parameter between two nearest carbon atoms is -3.1 eV. The Fermi energy is set to 0 eV. The simulation is run for only one energy point $E = 0.5$ eV.

Simulation timings are plotted in Figure 3.6 for graphene structures of different sizes. To minimize the dimension of blocks to invert in RGF, one layer of hexagonal structure is divided into four layers (see the dashed lines in Figure 3.5). Conclusions on the asymptotic complexity remain unchanged. Namely, the HSC extension is more efficient than the RGF

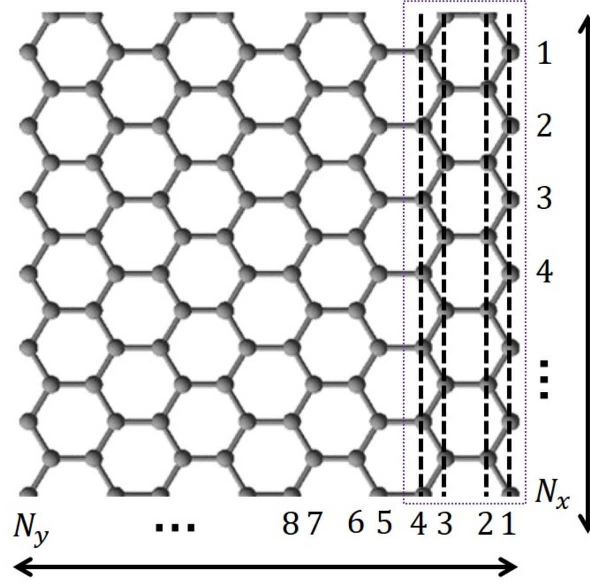

Figure 3.5: Graphene hexagonal structure decomposed by tight biding method. Dashed rectangular illustrates one repeating hexagon layer. Dashed lines represent inner four atom layers, showing the atoms ordering in tight binding Hamiltonian construction.

method for square and rectangular structures. The complexity of RGF for four-point stencil behaves as $\mathcal{O}(N_x^3 N_y)$ with the same constant as five-point stencil, which is expected due to the layered system partition. In Figure 3.6(a) and (b), for a four-point stencil system, our HSC-extension exhibits a complexity of $\mathcal{O}(N^3)$ for square system. Figure 3.6(c) and (d) also demonstrate a complexity growing linearly with N_y and, respectively quadratically with N_x, when N_x, respectively N_y, is fixed. The final result is a complexity $\mathcal{O}(N_x^2 N_y)$. The constant in front of $N_x^2 N_y$ for the extended HSC method is smaller for the graphene structures than for the superlattice devices. This reduction is explained by a more efficient partitioning of four-point stencil systems by METIS .

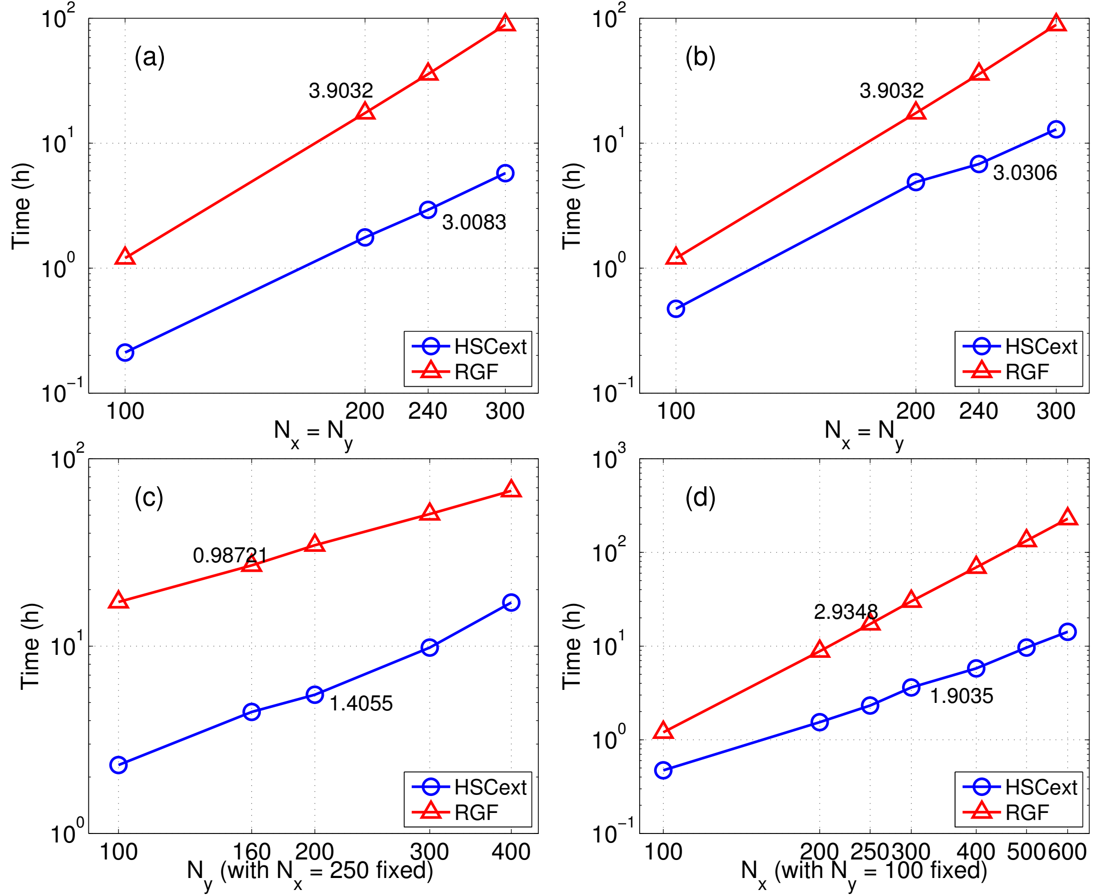

Figure 3.6: Graphene device NEGF simulation computation time comparison for RGF and the HSC extension, based on tight binding theory. (a) Square system of with diagonal self-energy matrix. (b) Square system of with dense self-energy matrix. (c) For systems in this plot, the number of atoms in the x-direction is set to $N_x = 250$, while the number of atoms in the y-direction is increased. (d) For systems in this plot, the number of atoms in the y-direction is set to $N_y = 100$, while the number of atoms in the x-direction is increased. Dense self-energy matrices are used in (c) and (d) devices.

Chapter 4

NUMERICAL EXPERIMENTS FOR 3D STRUCTURES

We have shown that the HSC-extension approach has significantly accelerated the evaluation of the retarded Green's function, particularly the lesser Green's function, for two-dimensional nanoscale devices. In this chapter, the HSC-extension is applied to determine the solution of NEGF equations on three-dimensional nanoscale devices [93]. Operation counts and runtimes are also studied for three-dimensional nanoscale devices of practical interest: a graphene-boron nitride-graphene multilayer system, a silicon nanowire, and a DNA molecule.

4.1 Motivation

Though achieving great success in predicting electronic transport performance in 2D devices, NEGF method is associated with the large computational cost which currently prevents a broader use to large scale 3D simulations. Researchers and analysts have used several approximations to alleviate this cost. Examples include the mode-space approximation [97], which couples 1D NEGF transport simulation to transverse states from solving 2D Schrödinger equation on the cross-section, and the restriction to levels of lower accuracy (such as tight-binding or effective mass levels). Although these approximations of NEGF have successfully predicted the transport characteristics of nanoscale devices [98, 99], these alternatives remain unable to capture atomic-scale inhomogeneities, such as surface roughness, unintentional doping, and trapped charges (see, for example, [100] for an illustration of the resulting inaccuracy). The modeling of such inhomogeneities requires a full 3D real-space NEGF simulation. Recent studies [50, 51, 101, 102, 103] have performed 3D NEGF simulations to handle these inhomogeneities but with "coarse discretization" (*i.e.* small number of atomic

orbitals or small number of grid points) to control the computational cost. To enhance the predictability of these 3D NEGF simulations, finer discretizations have to be considered. Therefore the computational cost of NEGF needs to be addressed. The goal of this chapter is to extend the HSC-extension approach to 3D scenarios and demonstrate its ability of significantly reducing the cost of 3D NEGF simulations.

4.2 Operation Count Analysis for 3D Brick-like Devices

In the last chapter, We have performed a cost analysis for a 2D rectangular system. Given the working knowledge of RGF and HSC-extension algorithms, we can then extend the operational cost discussion to a 3D brick-like devices. Consider a cuboid device, covered by a three-dimensional orthogonal mesh with N_x, N_y, and N_z grid points per direction. The discretization of the Hamiltonian is obtained via a 7-point stencil. The self-energy matrix $\mathbf{\Sigma}^r$ is assumed to be represented via a similar 7-point stencil (for example, with a crude diagonal approximation or with a PML-like approximation [104]). The resulting matrix $\mathbf{A}$ is of dimension $N_x \times N_z \times N_y$, where the y-direction is the transport direction.

The RGF approach groups the grid points into N_y disjoint layers, each layer holding $N_x \times N_z$ grid points. Fig. 4.1 illustrates these layers for $N_x = 3$, $N_z = 3$, and $N_y = 5$. By ordering the grid points one layer at a time, the matrix $\mathbf{A}$ exhibits the block-tridiagonal structure, required by the RGF approach. The operation count for the RGF method on this cuboid device is $\mathcal{O}\left(N_x^3 N_z^3 N_y\right)$.

The HSC-extension employs a multilevel nested dissection to gather the grid points into a hierarchy of clusters. Fig. 4.2 illustrates the separators obtained at each level and the eight subdomains. The resulting binary tree is also depicted in Fig. 4.2 with matching colors for the separators. The operation count for the HSC-extension is derived in the appendix. For the sake of conciseness, the final value is summarized in Table 4.1.

Remark: *When $N_z = 1$, the operation counts for RGF and HSC reduce to their expression for 2D devices (with mesh $N \times N$). Namely, for RGF, the operation*

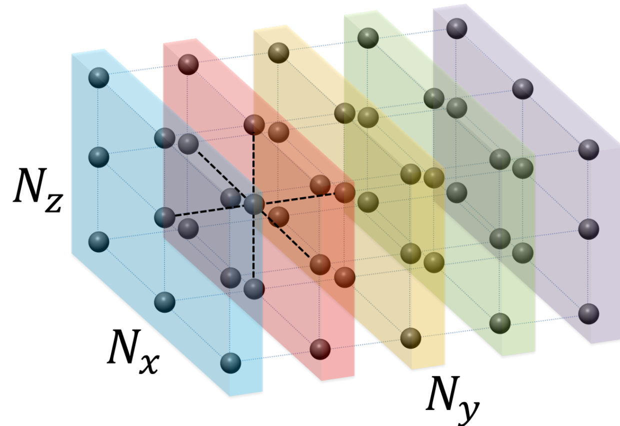

Figure 4.1: A Cartesian 3D mesh with 7-point-stencil discretization of dimension $N_x \times N_z \times N_y$ (the y-direction is the transport direction.). The colored layers along the y-direction show the layered-structure organization of grid points for the RGF approach.

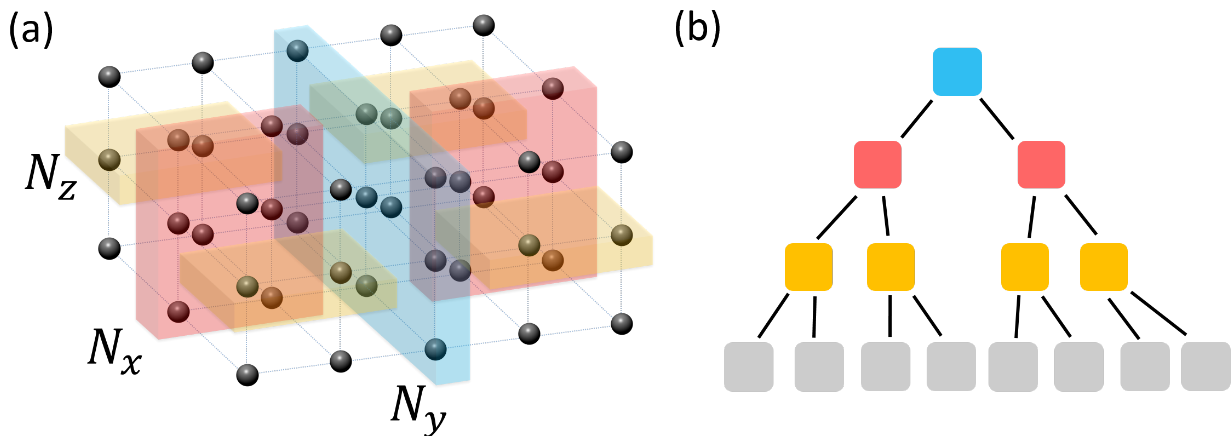

Figure 4.2: (a) The partition of grid points obtained from the multilevel nested dissection. The colored clusters show the separators defined at each level. (b) A binary tree representing the clustering. Each colored block matches a colored separator.

Configuration	HSC-extension	RGF
Cubic mesh ($N_x = N_z = N_y = N$)	$\mathcal{O}(N^6)$	$\mathcal{O}(N^7)$
Elongated mesh ($N_x = N_z = N \ll N_y$)	$\mathcal{O}(N^5 N_y)$	$\mathcal{O}(N^6 N_y)$
Flattened mesh ($N_z \ll N_x = N_y = N$)	$\mathcal{O}(N_z^3 N^3)$	$\mathcal{O}(N_z^3 N^4)$

Table 4.1: Operation counts of HSC-extension and RGF for various configurations of cuboid mesh.

count becomes $\mathcal{O}(N^4)$ *and, for HSC,* $\mathcal{O}(N^3)$.

In practice the self-energy matrix $\mathbf{\Sigma}^r$ contains dense blocks for the grid points on open boundary conditions. The analysis in the appendix and the counts (Table 4.1) do not cover such cases. The next section will study numerically the operation counts for practical nanoscale devices with open boundary conditions.

4.3 Numerical Experiments

Next we demonstrate and analyze numerically the performance of HSC-extension approach when evaluating entries of the matrices $\mathbf{G}^r$ and $\mathbf{G}^<$. First a cuboid device is used to illustrate the cost analysis presented in Section 4.2. The impact of dense blocks in $\mathbf{\Sigma}^r$ when modeling open boundary conditions is also discussed. Then three nanoscale devices of practical importance are considered: a graphene-boron nitride-graphene multilayer system, a silicon nanowire (SiNW), and a DNA molecule. The discretizations for these three devices yield matrices with different sparsity pattern and provide different challenges for the HSC-extension.

For reference, timings for the LU-factorization (`lu` routine in MATLAB 2011b [105] calling UMFPACK v5.0.4 [107]) of $\mathbf{A}$ are also included. The operation count for the LU-factorization represents the optimal cost complexity because every solution of a linear system with $\mathbf{A}$ requires, at least, the cost of one LU-factorization.

4.3.1 Cuboid Nanoscale Device

As described in section 4.2, a cuboid nanoscale device is considered where the Hamiltonian is constructed by effective-mass approximation and with 7-point stencil finite difference. A three-dimensional orthogonal mesh is used with N_x, N_y, and N_z grid points per direction. Two distinct treatments for the self-energy matrices are studied: a diagonal approximation, referred to as *SPARSE*, and a *DENSE* approximation for modeling open boundary conditions.

Results with SPARSE Self-energy Matrices

Here diagonal self-energy matrices are considered. The matrix **A** has the same sparsity as the Hamiltonian matrix **H**. Fig. 4.3 illustrates the pattern of non-zero entries in the matrix **A**, when the grid points are ordered one layer at a time (as in Fig. 4.1).

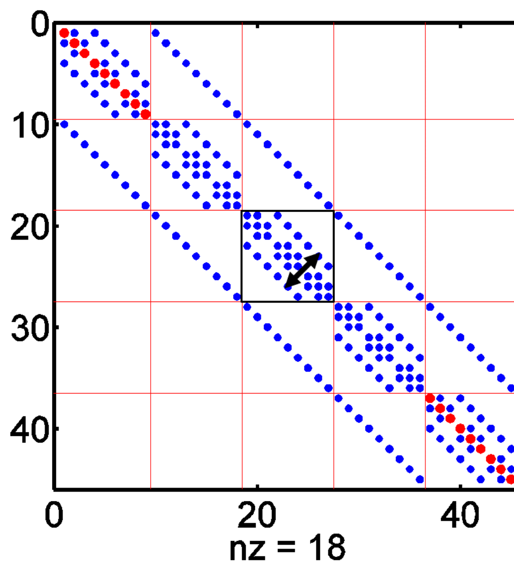

Figure 4.3: Non-zero pattern of **A** for a 3D cuboid system with $N_x = 3$, $N_z = 3$ and $N_y = 5$. Entries for the diagonal self-energy approximation are marked in red. The matrix exhibits a block-tridiagonal structure, where each block is of dimension $N_xN_z \times N_xN_z$. The arrow highlights the diagonal width in each block controlled by the ratio N_x/N_z. In all the matrix pattern graphs, `nz` specifies the number of non-zero entries.

First, when $N_x = N_y = N_z = N$, the CPU times for evaluating $\mathbf{G}^r$ and $\mathbf{G}^<$ at one energy point are plotted as a function of N in Fig. 4.4.

The slopes are consistent with the analysis of section 4.2, namely $\mathcal{O}(N^7)$ for RGF and $\mathcal{O}(N^6)$ for HSC-extension. The LU-factorization of $\mathbf{A}$ exhibits also a complexity $\mathcal{O}(N^6)$. When $N = 32$, the HSC-extension exhibits a speed-up of 10 times. Note that, on this 12GB machine, RGF can solve problems only up to $N = 32$ (the resulting matrix $\mathbf{A}$ is of dimension 32,768), while the HSC-extension can solve these problems up to $N = 40$ (dimension of matrix $\mathbf{A}$ is 64,000).

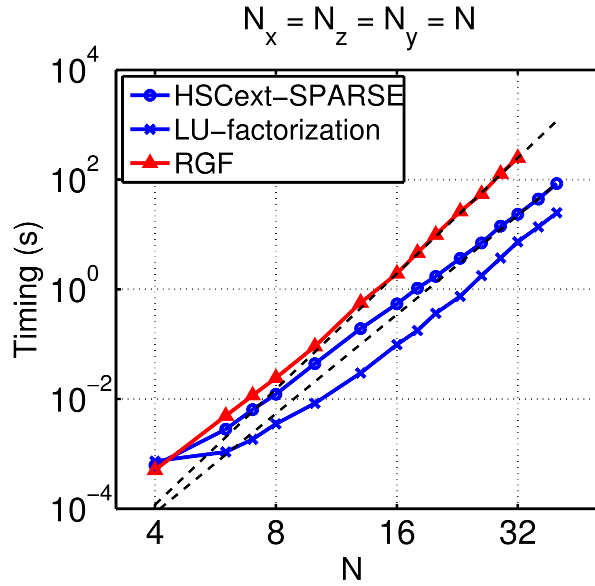

Figure 4.4: CPU timing for cubic system versus the dimension N_x. For all plots in result section, the timing includes the $\mathbf{G}^r$ and $\mathbf{G}^<$ calculation at one energy point. The runtime of RGF, HSC-extension and LU-factorization for $\mathbf{A}$ with *SPARSE* self-energy are presented. For comparison, we also plot black dashed curves, reflecting the theoretically asymptotic slopes for HSC-extension: $\mathcal{O}(N^6)$, and for RGF: $\mathcal{O}(N^7)$.

Next the case of an elongated device is considered, *i.e.* $N_x = N_z = N \ll N_y$. Fig. 4.5 illustrates timings for different elongated devices with square cross-section ($N_x = N_z = N$). The asymptotic slopes (black dashed curves) match the analysis, namely $\mathcal{O}(N^5 N_y)$ for the

HSC-extension and $\mathcal{O}(N^6 N_y)$ for RGF. Here again the HSC-extension and the LU-factorization have numerically the same complexity. When N_y is fixed at 200, the CPU time of HSC-extension is initially higher than the one for RGF at small cross-sections and becomes smaller than the one for RGF when $N_x = N_z \geq 12$, eventually reaching a speed-up of 2 for the largest structure studied.

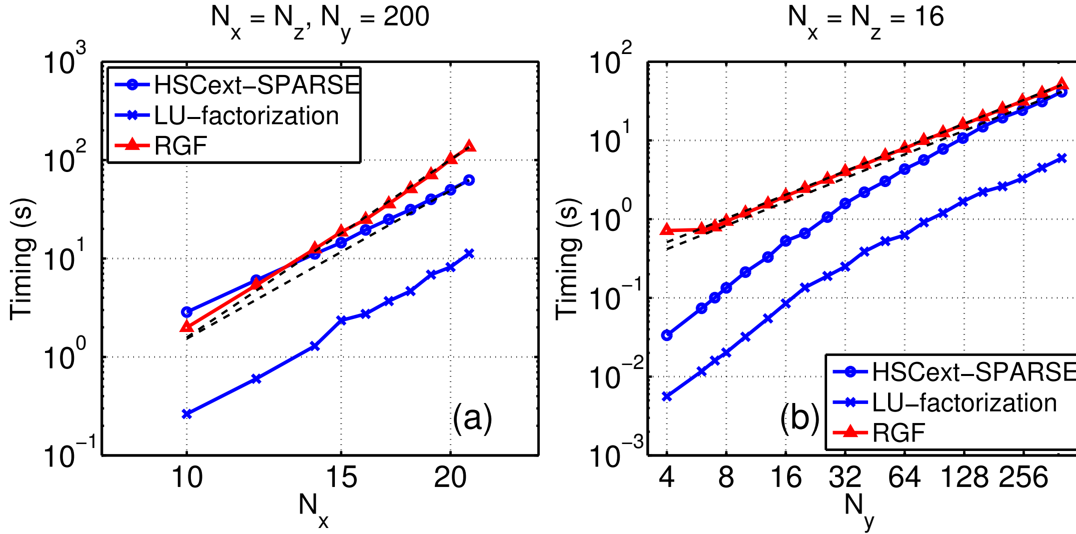

Figure 4.5: (a) CPU timing for elongated mesh versus N_x with fixed $N_x = N_z = N \ll N_y$, $N_y = 200$. (b) CPU timing for elongated mesh versus N_y with fixed $N_x = N_z = 16$. The theoretically asymptotic slopes (black dashed curves) for HSC-extension correspond to Table 4.1, $\mathcal{O}(N^5 N_y)$, and for RGF to $\mathcal{O}(N^6 N_y)$.

Finally, for the case of flattened devices, the CPU results are shown in Fig. 4.6. When $N_z = 4$ and $N_x = N_y = N \gg N_z$, the costs of $\mathcal{O}(N^3)$ for HSC-extension and $\mathcal{O}(N^4)$ for RGF are observed. These asymptotic behaviors are consistent with the analysis in section 4.2 and with the conclusions for 2D devices with $N_x \times N_y$ grid points [32].

Our numerical experiments in Fig. 4.3-4.6 illustrate the asymptotic operation count of HSC-extension as a function of system dimensions for various cuboidal shapes. In all three cases, the HSC-extension and the LU-factorization have identical asymptotic operation counts. These numerical experiments strongly suggest that the HSC-extension reaches

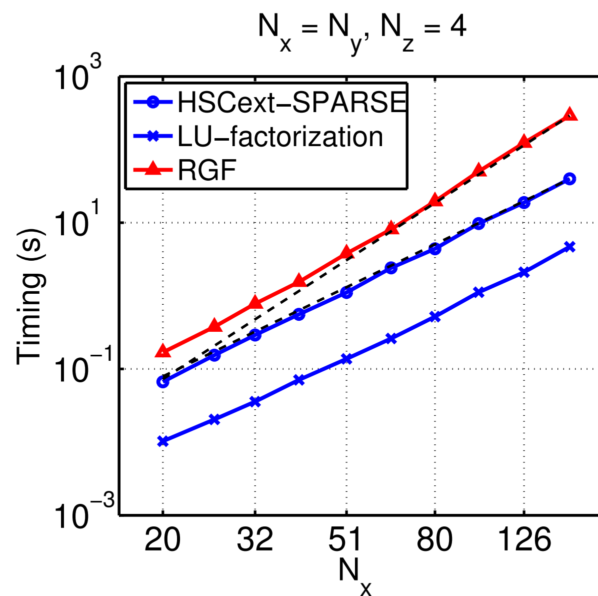

Figure 4.6: CPU timing for flattened mesh versus N_x with $N_x = N_y = N \gg N_z$, $N_z = 4$. The theoretically asymptotic slope (black dashed curves) for flattened mesh is $\mathcal{O}(N^3)$ for HSC-extension, and $\mathcal{O}(N^4)$ for RGF.

the ideal complexity for 3D nanoscale devices.

Effect of DENSE Self-energy Matrices

Next dense self-energy matrices are considered to model open boundary conditions. Fig. 4.7 illustrates the pattern of non-zero entries in the matrix $\mathbf{A}$, when the grid points are ordered one layer at a time (as in Fig. 4.1).

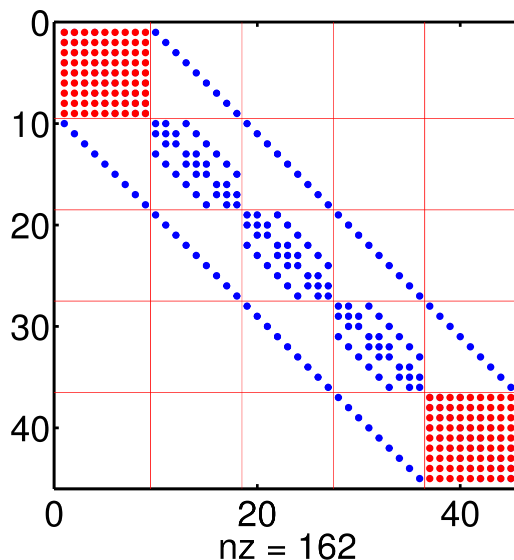

Figure 4.7: Non-zero pattern of $\mathbf{A}$ for a 3D cuboid system with $N_x = 3$, $N_z = 3$ and $N_y = 5$. Entries for the diagonal self-energy approximation are marked in red. The matrix exhibits a block-tridiagonal structure, where each block is of dimension $N_xN_z \times N_xN_z$.

RGF does not exploit the sparsity present in most diagonal blocks of matrix $\mathbf{A}$. So using a dense self-energy matrix does not impact the performance of RGF. On the other hand, the HSC-extension aims to exploit as much as possible the sparsity of $\mathbf{A}$. So it is important to study the impact of a dense self-energy matrix on the HSC-extension. The analysis in section 4.2 does not handle this situation.

First consider the case where $N_x = N_y = N_z = N$. Fig. 4.8 plots the CPU timings as a function of N. Here the RGF calculation stops at $N = 32$ due to memory limitation, while

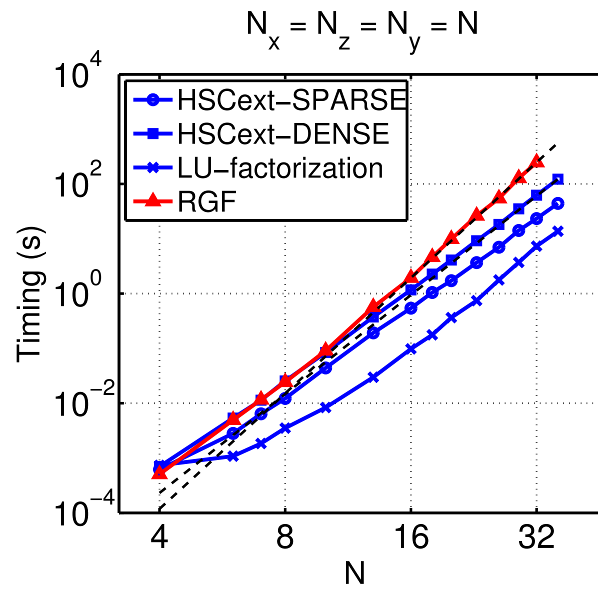

Figure 4.8: CPU timing for cubic system $N_x = N_y = N_z = N$ with both *DENSE* and *SPARSE* self-energies. The black dashed curve shows the asymptotic rates, namely $\mathcal{O}(N^6)$ for HSC-extension and $\mathcal{O}(N^7)$ for RGF.

the HSC-extension can solve problems up to $N = 36$ with dense self-energy matrices. The presence of a dense self-energy matrix yields larger CPU times for the HSC-extension but the asymptotic operation count is not modified.

Fig. 4.9 plots the CPU timings when N_y is modified and N_x and N_z remain constant. Timings for RGF, for HSC-extension with sparse self-energy matrix, for HSC-extension with dense self-energy matrix, and for the LU-factorization of $\mathbf{A}$ are reported. Note that in this

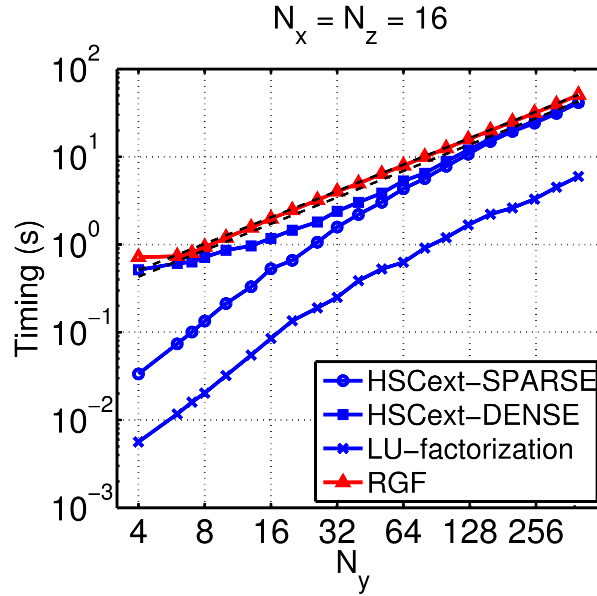

Figure 4.9: CPU timing for different N_y and fixed $N_x = N_z = 16$. The black dashed curve shows the asymptotic rate $\mathcal{O}(N_y)$.

numerical experiment, the layered-structure decomposition employed in RGF is kept along y-direction (even when $N_y < N_x$). When N_y is comparable to N_x and N_z, the speed-up for HSC-extension over RGF is reduced when a dense self-energy matrix is considered. As N_y gets larger, the timings of HSC-extension with the two forms of self-energy are closer and, asymptotically, approaching the complexity $\mathcal{O}(N_y)$.

4.3.2 Graphene - Boron Nitride - Graphene Multilayer System

Graphene, stacked with boron nitride insulating material, is a promising material to build next generation transistors because of its extraordinary thermal and electronic properties [11]. Here a multilayer heterostructure is considered, as shown in Fig. 4.10.

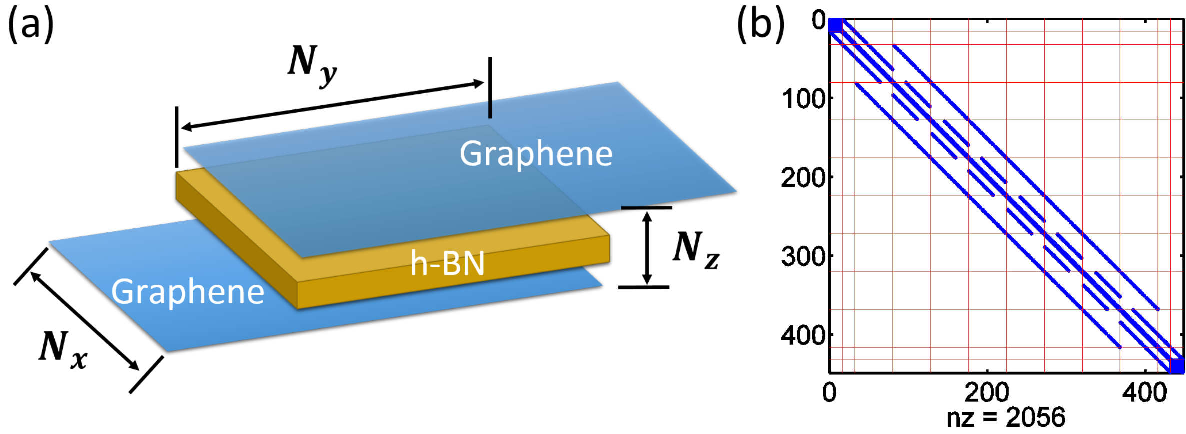

Figure 4.10: (a) Schematic view of graphene-hBN-graphene multilayer heterostructure. Two graphene layers are semi-infinitely long used as contacts. (b) The non-zero pattern of the **A** matrix with $N_x = 16$, $N_y = 8$ and $N_z = 3$.

The device consists of two semi-infinitely long monolayer armchair-edged graphene nanoribbon (AGNR) electrodes sandwiching an ultra-thin hexagonal boron nitride (hBN) multilayer film, yielding a vertical tunneling heterostructure with hBN acting as a potential barrier [8]. The hBN film is a few atomic layers thick and the central graphene-hBN-graphene (G-BN-G) overlapping heterostructure/multilayer region is stacked in AB-order (Bernal stacking). For this problem, the number of 2D vertical layers is denoted by N_z in units of atomic layers. The system width is N_x and the length of the multilayer stacking region is N_y, also in units of atomic layers. The semi-infinitely long AGNR monolayer electrodes at the top and bottom layers are treated as open boundary conditions, their effect is folded into dense self-energy blocks (extreme blocks) of dimension $N_x \times N_x$.

The system Hamiltonian is constructed using the nearest neighbor tight binding approx-

imation with parameters from [91]. Only the low energy p_z orbitals are considered here; thus the Hamiltonian has the same dimension as the total number of atoms simulated. The geometric lattice complexity of the multilayer system yields an average 5-point stencil Hamiltonian sparsity (multiple hexagonal-meshed layers stacked in AB order). The complexity of RGF remains $\mathcal{O}\left(N_x^3 N_z^3 N_y\right)$ because the sparsity inside each block is not exploited. The complexity of HSC-extension is studied numerically.

Fig. 4.11 plots the CPU timings when $N_x = N_y$ and $N_z = 5$. The HSC-extension and the LU-factorization of $\mathbf{A}$ have the same asymptote, indicating an operation count $\mathcal{O}(N_x^3)$.

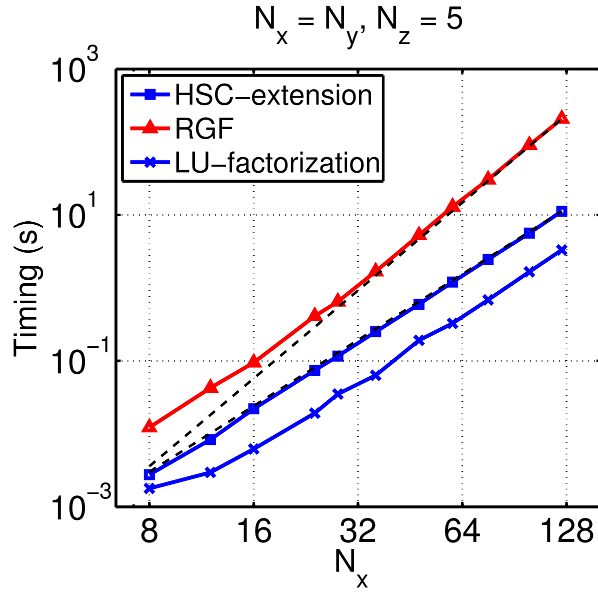

Figure 4.11: CPU timing for G-BN-G system as a function of $N_x = N_y$ and fixed $N_z = 5$. Dashed curves illustrate asymptotic rates, namely $\mathcal{O}(N_x^3)$ for HSC-extension and $\mathcal{O}(N_x^4)$ for RGF.

Fig. 4.12 plots the CPU timings for different configurations of N_x, N_y, and N_z. The experiments illustrate that HSC-extension still exhibits a complexity similar to the LU-factorization of $\mathbf{A}$. For the largest devices simulated in Fig. 4.12(a), $N_x = 256$, the HSC-extension method offers a speed-up of 3 orders of magnitude over RGF.

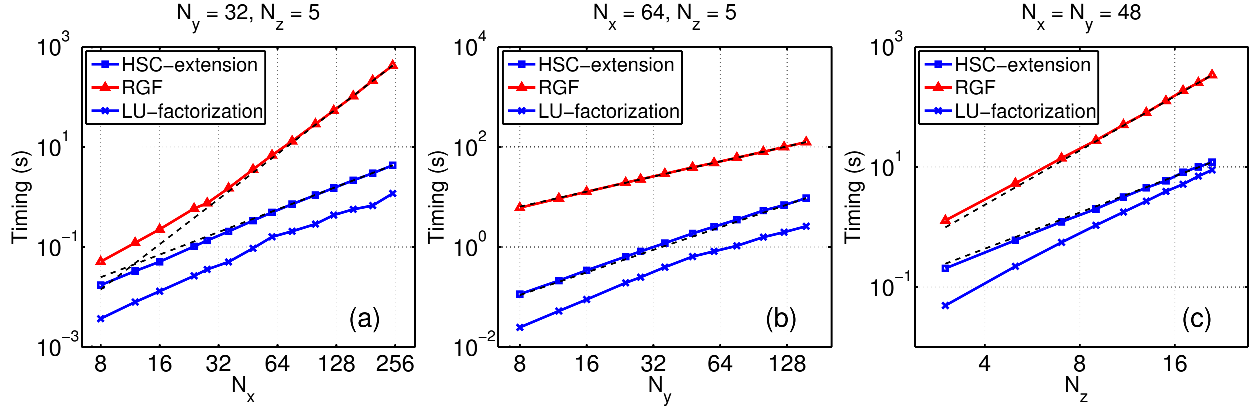

Figure 4.12: (a) CPU timings for G-BN-G system as a function of N_x and fixed $N_y = 32$, $N_z = 5$. (b) CPU timings for different N_y and fixed $N_x = 64$, $N_z = 5$. (c) CPU timings as a function of N_z and fixed $N_x = N_y = 48$. The dashed curves indicates the asymptotic operation counts. For HSC-extension, they are $\mathcal{O}(N_x^{1.5})$ for (a), $\mathcal{O}(N_y^{1.5})$ for (b), and $\mathcal{O}(N_z^2)$ for (c). The operation counts for RGF are as follows: $\mathcal{O}(N_x^3)$ for (a), $\mathcal{O}(N_y)$ for (b), and $\mathcal{O}(N_z^3)$ for (c).

The numerical experiments indicate that the asymptotic cost of HSC-extension is $\mathcal{O}\left(N_x^{1.5}N_z^2N_y^{1.5}\right)$. This cost of HSC-extension can be compared to the cost (Table 4.1) for the flattened device. The term $\mathcal{O}(N_x^{1.5}N_y^{1.5})$ matches well with $\mathcal{O}(N^3)$ by assuming $N_x = N_y = N$. The term $\mathcal{O}(N_z^2)$ is due to the Bernal stacking order for the multilayer structure in the z-direction.

Finally, Fig. 4.13 illustrates CPU timings for an elongated device, *i.e.* $N_x \ll N_y$. The timings for the HSC-extension behave like $\mathcal{O}(N_x^2N_y)$, demonstrating a lower order of complexity over the RGF method.

As a summary, the runtime cost of HSC-extension for the G-BN-G multilayer structure is

$$T = \begin{cases} \mathcal{O}(N_x^{1.5}N_y^{1.5}N_z^2) & \text{when } N_z \ll N_x \simeq N_y \\ \mathcal{O}(N_x^2N_yN_z^2) & \text{when } N_z \ll N_x \ll N_y \end{cases} \tag{4.1}$$

while the runtime cost of RGF behaves like $\mathcal{O}(N_x^3N_z^3N_y)$.

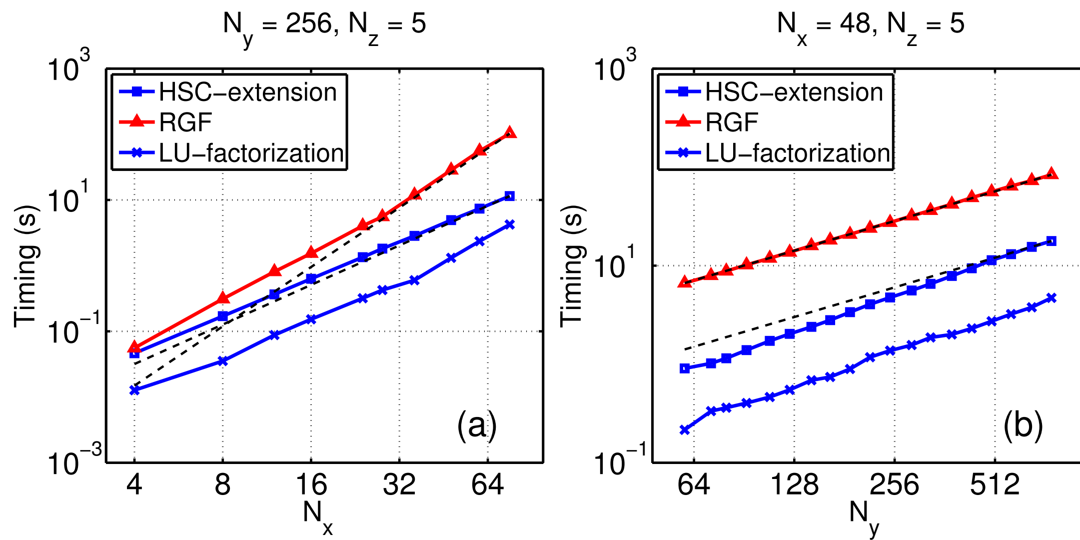

Figure 4.13: (a) CPU timing for G-BN-G system with different N_x and fixed $N_y = 256$, $N_z = 5$. (b) CPU timing for different N_y and fixed $N_x = 48$, $N_z = 5$. The dashed curves indicates the asymptotic operation counts. For HSC-extension, they are $\mathcal{O}(N_x^2)$ for (a) and $\mathcal{O}(N_y)$ for (b). The operation counts for RGF are as follows: $\mathcal{O}(N_x^3)$ for (a) and $\mathcal{O}(N_y)$ for (b).

4.3.3 Silicon Nanowire Structure

Silicon nanowire devices have shown promises to become key components in the next generation computer chips [108]. Solving efficiently the NEGF equations for such devices is therefore important.

In order to investigate the scaling of computational runtime as a function of SiNW lateral dimensions, specifically the number of atoms in each layer and the number of unit cells, we consider a SiNW device depicted in Fig. 4.14(a). The number of atomic layers in y-direction is denoted as N_y and the number of silicon atoms within each atomic layer (cross-section) is denoted as N_{cs}. So a $N_{cs} \times N_y$ SiNW structure contains an array of $N_y/2$ unit cells with $2N_{cs}$ atoms per unit cell. Next the $sp^3d^5s^*$ tight-binding formalism [109] is used to

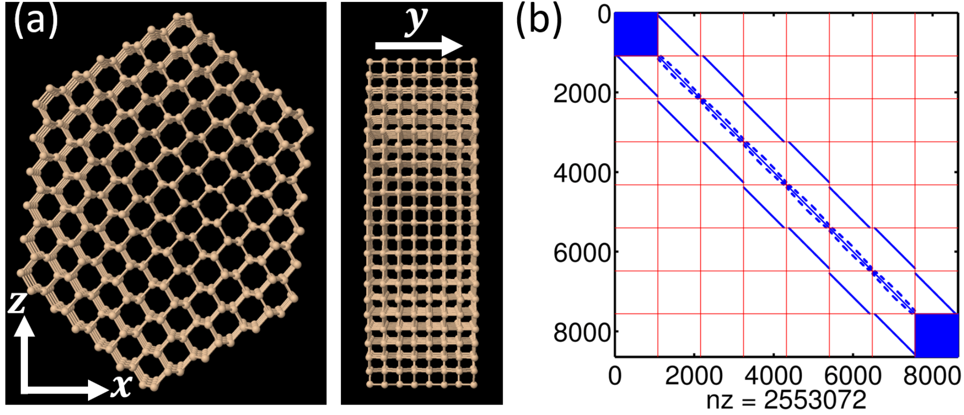

Figure 4.14: (a) Atomic view of a silicon nanowire example with 4 unit cells. Each unit cell has two atomic layers and hexagonal cross-section shape, with each atomic layer containing 108 Si atoms. Cross-section is along $x - z$ plane and transport direction is along y direction. This example corresponds to $N_{cs} = 108$ and $N_y = 8$. (b) The non-zero pattern of the $\mathbf{A}$ matrix with $N_{cs} = 108$ and $N_y = 8$.

discretize the system. Each silicon atom is represented by a 10×10 diagonal block, thereby interconnecting with up to 40 orbitals of the nearest-neighbor silicon atoms. The resulting Hamiltonian matrix exhibits a stencil involving more than 40 points, which results in a particular computational challenge. Dense self-energy matrices of dimension $10N_{cs} \times 10N_{cs}$

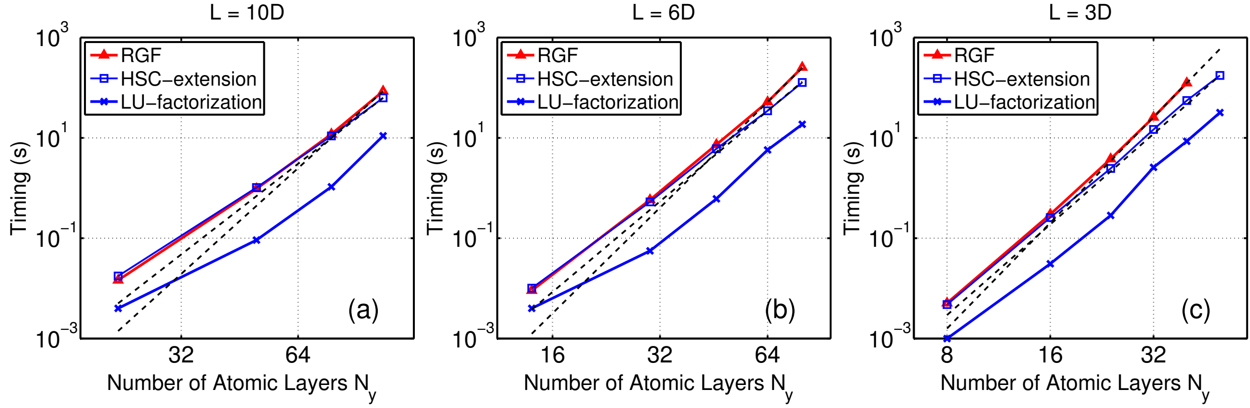

Figure 4.15: CPU timing for SiNW system with (a) $L = 10D$ with largest $L = 20$nm, (b) $L = 6D$ with largest $L = 15$nm and (c) $L = 3D$ with largest $L = 9.3$nm. Note that $N_y \propto L$ and $N_{cs} \propto D^2$. The dashed curves represent asymptotes: $\mathcal{O}(N_y^6)$ for HSC-extension and $\mathcal{O}(N_y^7)$ for RGF.

are employed. Fig. 4.14(b) illustrates the sparsity of $\mathbf{A}$ when $N_{cs} = 108$ and $N_y = 8$.

In this section, we consider nanowire whose length L is proportional to the diameter D of the cross-section, namely $L = \alpha D$. When ordering the atoms one layer at a time, the Hamiltonian matrix, as well as matrix $\mathbf{A}$, has a block-tridiagonal structure, where each block is of dimension $40N_{cs} \times 40N_{cs}$. The operation count for RGF becomes $\mathcal{O}\left(N_{cs}^3 N_y\right)$. Since $N_y = \mathcal{O}(L)$, $N_{cs} = \mathcal{O}(D^2)$, and $L = \alpha D$, the operation count for RGF is $\mathcal{O}(N_y^7)$. The operation count for HSC-extension will be studied numerically.

Fig. 4.15 plots CPU timings as a function of N_y for structures shaped by $L = 10D$, $L = 6D$ and $L = 3D$. Our numerical experiments exhibit an asymptotic cost of $\mathcal{O}(N_y^7)$ for RGF and $\mathcal{O}(N_y^6)$ for HSC-extension. We would like to emphasize that the complexity $\mathcal{O}(N_y^6)$ is valid for HSC-extension as long as $L = \alpha D$, independent of the value α. The HSC-extension has the same asymptotic behavior as the LU-factorization of $\mathbf{A}$.

In practice, analysts may consider nanowires of 20nm length. Table 4.2 lists CPU timings when the length is 20nm. Because these simulations did not fit in the 12GB RAM of our machine, the CPU timings were extrapolated from the asymptotes in Fig. 4.15. This extrapolation indicates that the speedup of HSC-extension over RGF improves as α decreases.

Shapes	HSC-extension (s)	RGF (s)	speed-up
$L = 10D$	62.9	85.3	1.4
$L = 6D$	1,064	3,030	2.8
$L = 3D$	19,920	147,706	7.4

Table 4.2: Extrapolated CPU timings of transmission calculation for SiNW devices at one energy point with $L = 20$nm for various shapes.

The reduction of α enlarges the cross-section N_{cs} while L is fixed, yielding a higher speedup of HSC-extension, which is consistent with the other experiments.

4.3.4 DNA Molecule

Finally we test the algorithm for DNA-based structure, which represents a complex organic system. It has been shown that DNA is one of the promising candidates in the molecule devices [110]. The study of electronic structures can be used to develop new sequencing techniques [111], acting as DNA fingerprints. Another application is disease detection [112]. Many diseases are linked with the mutation in DNA bases, resulting in different electronic properties which can be used to distinguish the mutated DNA. In our numerical experiments, the DNA molecule is described by density functional theory (DFT) method. Although the number of atoms contained in single DNA molecule is not huge, the decomposed Hamiltonian matrices are relatively dense, thus impeding an effective decomposition using the multilevel nested dissection.

The DNA molecule in our simulation is a double-helix structure containing 7 – 15 base pairs in each strand sketched in Fig. 4.16(a). The Hamiltonian matrices are generated by DFT package GAUSSIAN 09 [106] at HG/6-31G (d, p) level [113]. The number of orbitals (matrix dimension) for each base is about 250. For example, for a 9-mer DNA molecule, the Hamiltonian is of dimension 4500×4500 as shown in Fig. 4.16(b). The Hamiltonian can be decomposed by treating each base pair as one layer, yielding a block tri-diagonal shape

with 9 layers. Different from the structures studied above, the diagonal and nearest neighbor off-diagonal blocks in the Hamiltonian are fully dense.

The CPU timing results for various DNA molecules are summarized in Table 4.3. For

Number of Base Pairs	7	9	11	13	15
HSC-extension (s)	5.0	8.7	11.3	12.5	14.6
RGF (s)	5.5	7.8	9.7	11.6	13.7

Table 4.3: CPU timings of transmission calculation for DNA molecules at one energy point.

these configurations, the HSC-extension seems to be less efficient than RGF.

To better understand these CPU times, it is important to look at the clusters defined by the multilevel nested dissection. Fig. 4.17 illustrates the clusters used for the HSC-extension when the multilevel nested dissection is applied *blindly* to the graph of $\mathbf{A}$, and the layers used for RGF. RGF employs layers with one base pair and the resulting blocks are of dimension 500×500. The multilevel nested dissection works at the level of base pairs because one base pair corresponds to one fully-dense diagonal block in the matrix $\mathbf{A}$. The resulting partition introduces clusters with one base pair except for two bottom level (level 3) clusters that can not be partitioned with nested dissection. These two clusters result in two block matrices of dimension 1000×1000. The time discrepancy arises from these two distinct choices of row gathering.

To further illustrate the impact of the row numbering or partitioning, Table 4.4 lists timings for two additional approaches. The four different partitionings used for these approaches are depicted on Fig. 4.17. The *customized* RGF method with two 2-pairs layers gathers, twice, 2-pairs into one layer. This particular choice of layers indicates the impact of two blocks of dimension 1000×1000 on the overall CPU time, *i.e.* an increase in CPU time. The *customized* partition for the HSC-extension allows a degenerate sub-tree[1] to avoid

[1]Any cluster with 2-pairs is partitioned according to a degenerate tree, where each parent node has only one child.

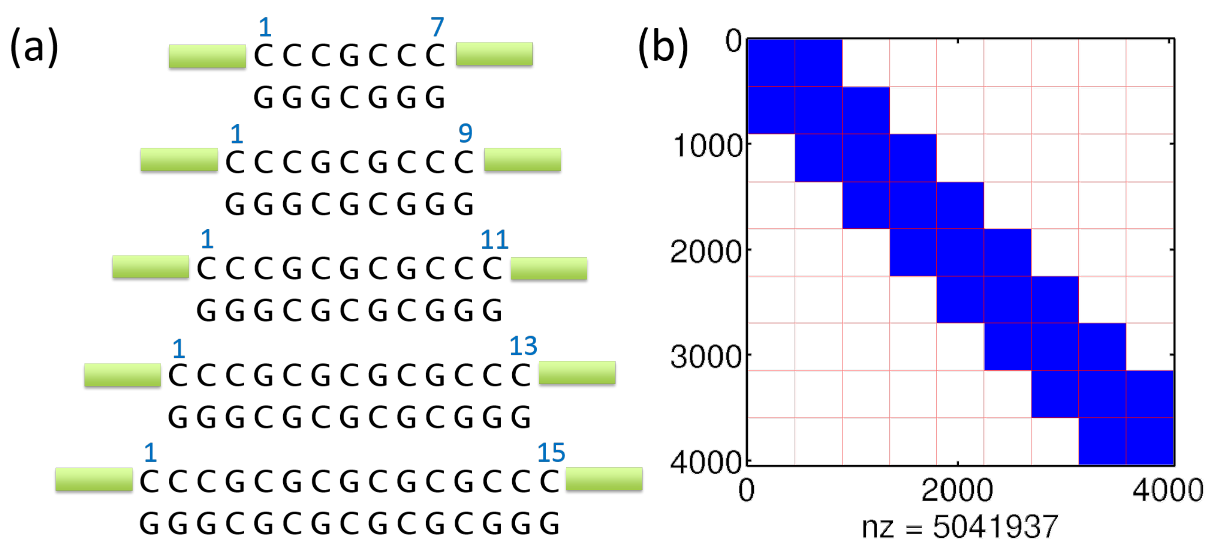

Figure 4.16: (a) Sketch of the simulated DNA sequence with 7, 9, 11, 13 and 15 base pairs respectively. Cytosine (C) and guanine (G) are two types of bases in DNA. The left/right contacts are connected to the bases on one strand. (b) The corresponding non-zero pattern of the **A** matrix for the 9 base pairs DNA. All tri-diagonal blocks are fully dense.

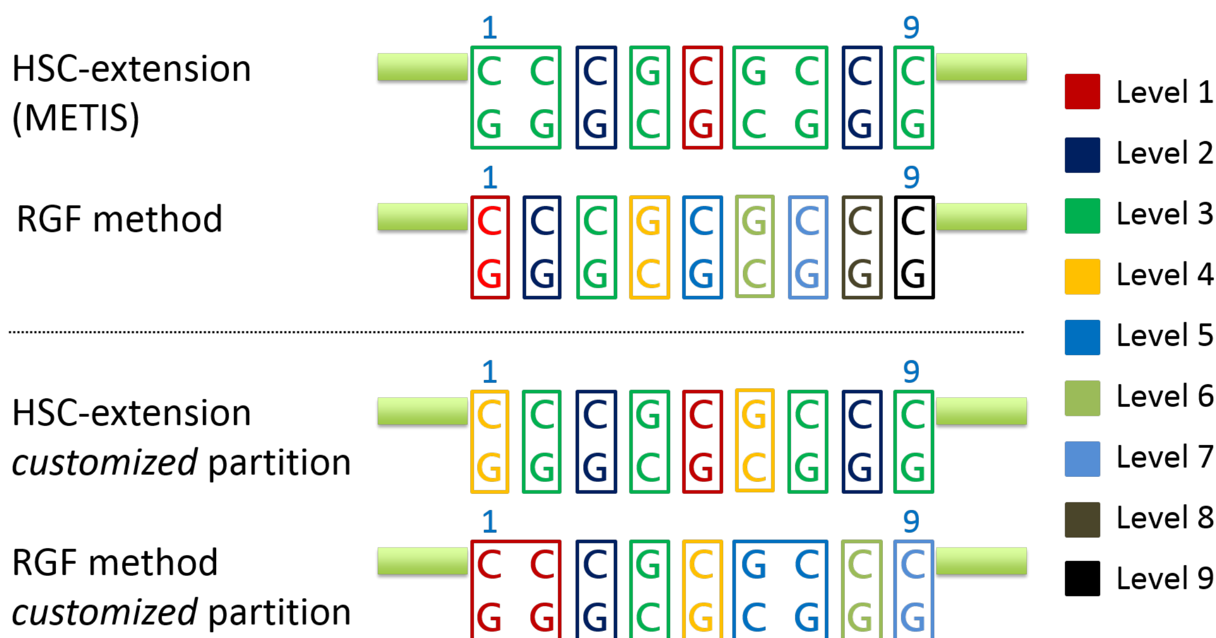

Figure 4.17: Cluster definitions for HSC-extension, for RGF, and for two customizations.

Number of Base Pairs	9
HSC-extension (s)	8.7
RGF (s)	7.8
HSC-extension with *customized* partition (s)	7.8
Customized RGF with two 2-pairs layers (s)	11.8

Table 4.4: CPU timings of transmission calculation for DNA molecules at one energy point.

any 2-pairs cluster. This *customized* partition makes the HSC-extension operate on blocks of dimension, at most, 500×500. This choice results in a lower CPU time, on par with the original RGF approach. A similar behavior was observed for other DNA molecules, where METIS gathers 2-pairs into one cluster.

These numerical experiments suggest that the HSC-extension, combined with a multilevel nested dissection, is an efficient approach even for smaller but denser matrices. When the graph partitioning is allowed to insert degenerate sub-trees, the performance is comparable with that of the performance of RGF.

4.4 Summary

In this chapter, we demonstrate the HSC-extension based NEGF solver as a working methodology for various 3D systems. The cost analysis for HSC-extension is performed on a cuboid structure. HSC-extension exhibits operation count of $\mathcal{O}(N^6)$ when simulating cubic device with dimension $N \times N \times N$, whereas a $\mathcal{O}(N^7)$ count is observed for RGF. We also illustrate various asymptotic costs of HSC-extension when the device has an elongated shape (N_x, $N_z \ll N_y$), when the device is flattened ($N_z \ll N_x$, N_y), and when a dense self-energy is used to model the open boundary conditions.

The runtime performance of HSC-extension is further investigated for nano-electronic devices of practical interest: graphene-hBN-graphene multilayer heterostructure, silicon nanowire and DNA molecule. These devices exhibit distinct atomistic sparsity, indicating differ-

ent computational efficiency for HSC-extension. The numerical experiments suggest that the HSC-extension exhibits asymptotic runtimes and operation counts proportional to the runtime of the LU-factorization. For all the nano-electronic devices considered, the HSC-extension becomes faster than the RGF method as the device gets larger. A 1,000 speed-up is observed for a graphene-hBN-graphene multilayer device with 40,000 atoms. Since the HSC-extension requires less operations than RGF, these speed-ups will increase as the device gets larger.

Chapter 5

DEVELOPMENT OF NEGF – POISSON SOLVER

While previous chapters are focused on an efficient numerical scheme for solving NEGF equations, in this chapter we will provide a comprehensive description of key components in a complete implementation of NEGF – Poisson simulator, including system decomposition, pre and post-processing of NEGF solver, and the calculation of non-linear Poisson's equation.

Section 5.1 gives a schematic overview of the NEGF – Poisson simulatior. Section 5.2, section 5.3 and section 5.4 then present the detailed description of the simulation process as well as the simulation results for a 2D dual-gate MOSFET, a 3D gate all-around MOSFET, and a graphene FET, which are widely-used representative structures corresponding to 2D effective-mass, 3D effective-mass and 3D tight-binding models respectively.

5.1 NEGF – Poisson Simulator

The flowchart of a complete NEGF – Poisson simulation is shown in Figure 5.1. The simulation starts with a prediction of electrostatic potential profile, which is often derived from the initial conditions or analytical equations. For each external bias point (e.g. gate bias, source-drain bias), the computational procedure consists of two iterative loops:

1. The inner iteration within NEGF calculation is to obtain the charge distribution with self-consistently including electron-phonon scattering self-energies.

2. The outer iteration self-consistently couples NEGF and Poisson solvers via the electrostatic potential profile

The NEGF formalism and modeling process is similar to the description in section 1.2.2. The electrostatics is solved by Poisson's equation and included in system Hamiltonian. NEGF

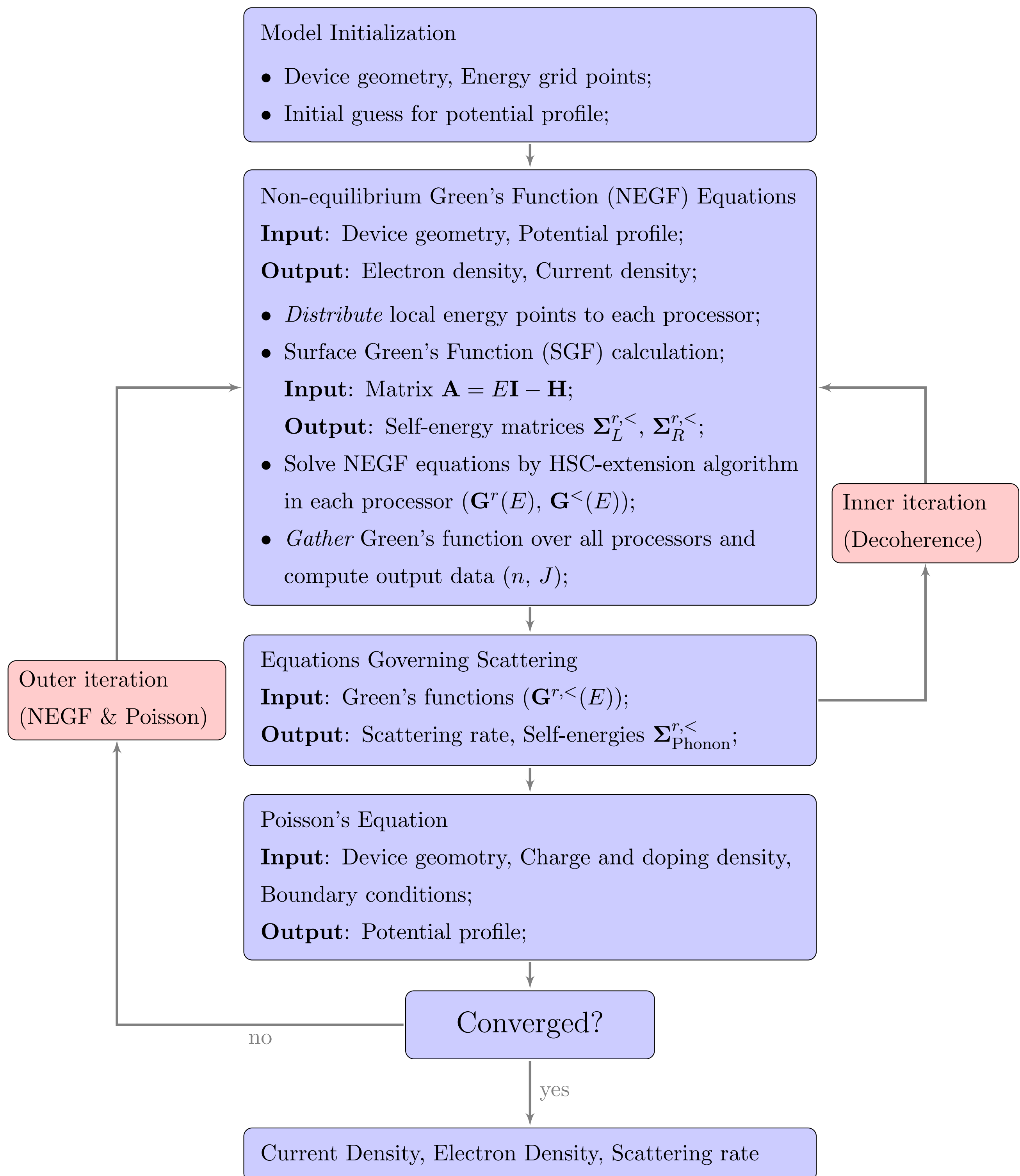

Figure 5.1: Flowchart for NEGF-Poisson self-consistent simulation, including self-consistent solution of decoherence self-energy.

then calculates physical quantities such as electrons / holes concentration, which is involved as input of Poisson's equation, yielding a self-consistent calculation between NEGF formalism and Poisson's equations.

5.1.1 Poisson's Equation

In semiconductor transport simulation, instead of solving the conventional Poisson's equation $-\nabla\cdot(\varepsilon\nabla V) = \rho$, people commonly tend to solve the non-linear version of this equation which reads:

$$-\nabla \cdot (\varepsilon \nabla V) = \rho(V) \tag{5.1}$$

where the charge distribution in the right-hand-side of the equation explicitly depends on the potential profile. The employment of non-linear Poisson's equation can effectively enhance the stability for the convergence between the transport solver (e.g. NEGF equations) and the Poisson solver.

> *Although this chapter is focused to describe the numerical implementation for the Poisson solver, the computation process of the NEGF equations (not fully identical to the previous chapters) will also be presented in order to provide readers a comprehensive understanding of the coupling between NEGF and Poisson's equations.*

5.2 Development of 2D NEGF – Poisson Solver

This section begins with a glossary of useful physical quantities in the follow sections. Then we present the development of non-linear Poisson solver and apply it on a conventional 2D silicon-on-insulator (SOI) dual-gate MOSFET whose Hamiltonian is discretized using effective-mass scheme. In the subsequent sections, we will extend our development to both 3D effective-mass (3D MOSFET) and tight-binding (graphene) scenarios. For the sake of conciseness, only coherent electron transport is considered in our description and numerical experiments. Although ignored for simplicity, hole transport can be easily included as well.

5.2.1 *Physical Quantities and Useful Equations*

Simple but useful in terms of practical applications, a Poisson solver for a 2D SOI-based MOSFET is developed first. We start by declaring a set of physical quantities and several useful relationships to derive classical expression for key parameters such as charge density and Fermi levels.

N_c, N_v: effective density of states for conduction (valence) band, only determined by material and temperature.

m_e, m_h: effective mass of electron (hole), only determined by material and temperature.

N_d, N_a: doping concentration for n-type (p-type) region, user defined quantities.

a: mesh spacing of finite difference, a user defined quantity (for simplicity, uniform meshing is assumed).

ε_r: relative permittivity of semiconductor and oxide.

E_f: Fermi energy of the system at equilibrium, a user defined quantity.

E_{fs}, E_{fd}: Fermi energy at source and at drain ends.

E_{fi}: quasi-Fermi energy profile, usually as a function of coordinates.

n, p: electron (hole) density.

ρ: total charge density, $\rho = q(N_d - N_a - n + p)$.

V: electrostatic potential profile, usually as a function of coordinates.

E_c, E_v: energy of bottom of conduction band (top of valence band).

V_g, V_b: gate bias, source-drain bias.

J: current density profile, usually as a function of coordinates.

For a bulk semiconductor with given E_f and E_C, the electron density is expressed as:

$$n = N_c F_{1/2}\left(\frac{E_f - E_c}{k_B T}\right) \tag{5.2}$$

Conversely,

$$E_c = E_f - F_{1/2}^{-1}\left(\frac{n}{N_c}\right) k_B T \tag{5.3}$$

$F_{\pm 1/2}$ is half-order Fermi-Dirac integral. $F_{1/2}^{-1}$ is its inverse function. The definition of complete Fermi-Dirac integral and its numerical approximation can be found at [114]. The electrostatic potential is given by

$$V = -\frac{1}{q}\left(E_c + E_{\text{ref}}\right) \tag{5.4}$$

where E_{ref} is an arbitrarily chosen constant (reference potential). This equation defines the relation between E_c and V, which is used throughout the following description.

5.2.2 Poisson Solver for a 2D MOSFET

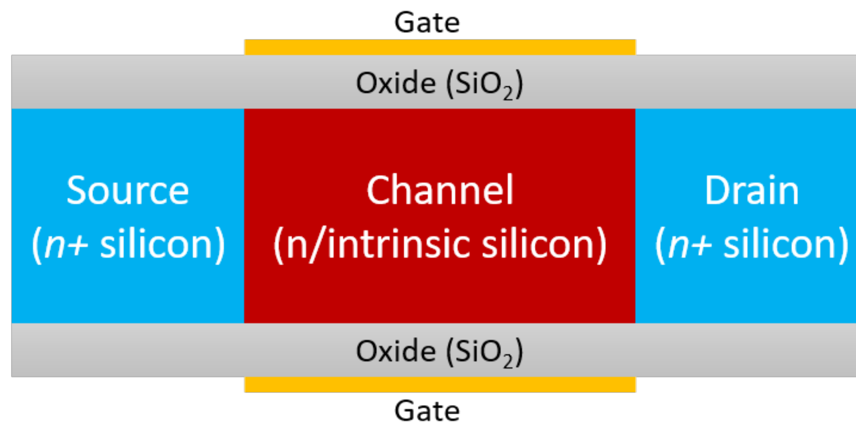

Figure 5.2: A 2D dual-gate SOI MOSFET toy model employed as an example in the development of Poisson solver.

A prototypical device for a 2D dual-gate SOI-based MOSFET is shown in Figure 5.2. The toy system is a thin body device with intrinsic (or low level n-type doped) channel and

heavily n-type doped source and drain. A complete and comprehensive description of the full 2D NEGF – Poisson simulation process is presented first:

Define system

- Define the system geometry: including oxide thickness, lengths of source, drain and channel, meshing profile (a uniform mesh is assumed here).

- Define the doping strategy: doping concentration $N_d(x, y)$ of n^+ regions and channel region (intrinsic or low level n-doping is assumed here).

- Initial guess for electrostatic potential.
  One common choice for the initial guess is obtained by assuming $n = N_d$, and compute initial potential $V_{guess}(x, y)$ using equations 5.3 and 5.4.

For each V_g and V_b (or other external bias), the solver self-consistently process step 2 and step 3 until converged.

NEGF calculation

- Hamiltonian $H = -\frac{q^2}{2m_e}\nabla^2 + E_c$ is defined only in semiconductor domain (source, drain and channel), which is discretized by finite difference approximation, and a homogeneous boundary condition is applied for all edges.

- Energy window is determined by: $E_{min} = \min(E_c) - 3k_BT$, and $E_{max} = \max(E_c) + 10k_BT$. Energy spacing d$E$ is typically less than 0.5 meV.

- Fermi level: $E_{fs} = E_f$ and $E_{fd} = E_f - V_b$.

- For E from E_{min} to E_{max}:

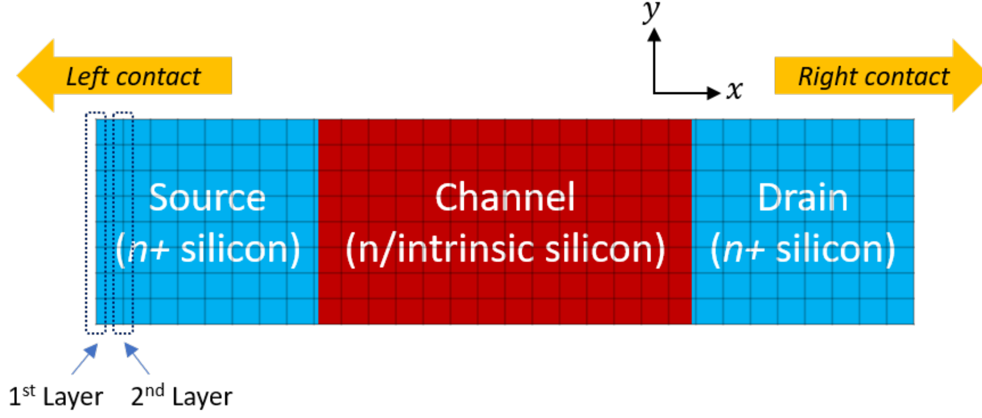

Figure 5.3: The computational domain for the NEGF calculation part. Note that we only discretize the semiconductor region, since the oxide regions do not hold charge and thus have no contribution to the transport.

- Contact self-energy matrices $\Sigma^r_{L,R}(E)$ are obtained (from $E\mathbf{I} - \mathbf{H}$) at left and right ends assuming open-boundary conditions at each energy point. Approach has been described in section 1.2.3.
- $\mathbf{A}(E) = E\mathbf{I} - \mathbf{H} - \mathbf{\Sigma}^r_L(E) - \mathbf{\Sigma}^r_R(E)$
- $\Sigma^< = -2i\,\mathrm{Im}[\Sigma^r_L]F_{-1/2}\left(\frac{E_{fs}-E}{k_BT}\right) - 2i\,\mathrm{Im}[\Sigma^r_R]F_{-1/2}\left(\frac{E_{fd}-E}{k_BT}\right)$
- $\mathbf{G}^r(E) = \mathrm{inv}(\mathbf{A})$
- $\mathbf{G}^<(E) = \mathbf{G}^r(E)\Sigma^<\left(\mathbf{G}^r(E)\right)^\dagger$
- $n(E,j) = -i\mathbf{G}^<(E)_{j,j}$ for each grid point j
- $J(E) = (2q/\hbar)\,\mathrm{Tr}\left[\mathbf{H}_{\{1,2\}}\mathbf{G}^<_{\{2,1\}} - \mathbf{G}^<_{\{1,2\}}\mathbf{H}_{\{2,1\}}\right]$, where 1, 2 denotes the first and second layer of grid points along transport direction (x). $\{1,2\}$ represents the off-diagonal block.

- Electron density at each grid point (m^{-3}):

$$n(j)_{\mathrm{NEGF}} = \frac{1}{a^2}\sqrt{\frac{m_e k_B T}{2\pi^3\hbar^2}}\int n(E,j)\mathrm{d}E$$

- Current density (A/m):

$$J = \sqrt{\frac{m_e k_B T}{2\pi^3 \hbar^2}} \int J(E) \mathrm{d}E$$

Poisson calculation

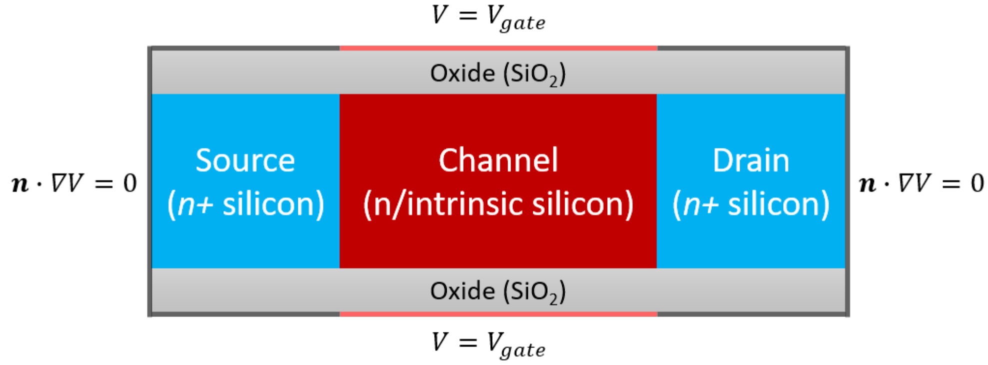

Figure 5.4: The computational domain and boundary conditions for the Poisson calculation.

- First we compute quasi-Fermi energy profile (only has definition in semiconductor regime)

$$E_{fi} = F_{1/2}^{-1}\left(\frac{n_{\text{NEGF}}}{N_c}\right) k_B T + E_c$$

where E_c is the one used in NEGF part (Hamiltonian).

- Non-linear Poisson equation is defined in the whole domain (semiconductor and oxide):

$$-\nabla \cdot (\varepsilon_0 \varepsilon_r \nabla V) = \rho(V) = q\left[N_d - n(V)\right]$$

Note that $n(V)$ is not directly obtained from NEGF calculation (n_{NEGF}). Instead, we use equation 5.2 as a classical predictor of the electron concentration:

$$n(V) = N_c F_{1/2}\left(\frac{E_{fi} + qV - E_{\text{ref}}}{k_B T}\right)$$

$\rho = 0$ for grid points in oxide.

- Boundary conditions are shown in Figure 5.4:
  - Potential at "gate" boundaries satisfies $V = V_{\text{gate}} = -(E_{\text{flat-band}} + E_{\text{ref}})/q + V_g$, where $E_{\text{flat-band}} = E_f - F_{1/2}^{-1}(n_{\text{channel}}/N_c)\, k_B T$.
  - Potential at boundaries elsewhere satisfies $\hat{\mathbf{n}} \cdot \nabla V = 0$ when ε_r is scalar (isotropic permittivity); and satisfies $\hat{\mathbf{n}} \cdot (\varepsilon_r \nabla V) = 0$ when ε_r is a matrix (anisotropy permittivity). $\hat{\mathbf{n}}$ is the unit outer vector normal to the boundary.
- This non-linear equation can be solved with different methods, such as iterative solver:

$$-\nabla \cdot \varepsilon_0 \varepsilon_r \nabla \left(V_{\text{old}} + \delta V\right) = \rho \left(V_{\text{old}} + \delta V\right)$$

$$\left(-\nabla \cdot \varepsilon_0 \varepsilon_r \nabla - \frac{\partial \rho}{\partial V}\right) \delta V = \nabla \cdot \varepsilon_0 \varepsilon_r \nabla V_{\text{old}} + \rho \left(V_{\text{old}}\right)$$

  where

$$\frac{\partial \rho}{\partial V} = -\frac{q^2 N_c}{k_B T} F_{-1/2} \left(\frac{E_{fi} + qV - E_{\text{ref}}}{k_B T}\right)$$

  Boundary condition for δV: $\delta V = V_{\text{gate}} - V_{\text{old}}$ at "gate" boundaries; $\hat{\mathbf{n}} \cdot (\varepsilon_r \nabla \delta V) = -\hat{\mathbf{n}} \cdot (\varepsilon_r \nabla V_{\text{old}})$ at boundaries elsewhere.

5.2.3 Results

The NEGF – Poisson solver is utilized to simulate the I-V characteristics of three types of 2D SOI-based MOSFET structure: thick body, thin body, and short channel length. Simulation results are presented below.

Thick Body

The first numerical experiment is performed on a thick body device which mimics the bulk MOSFET case. The body thickness is 18nm; the channel length and source (drain) length is 70nm and 40nm respectively; the oxide thickness is 3nm. Channel is lightly n-type doped with a concentration of 10^{13}cm^{-3} and source / drain are heavily doped with a concentration of 10^{18}cm^{-3}. The output characteristics curves ($I_{ds} - V_{ds}$) and band diagram at various bias

points are shown in Figure 5.5 for $V_g = 0$ and 0.1V. Triode and saturation regions are clearly shown in the plot with slight short channel effect in saturation region, indicating that the gates still do not have the full control of conduction in the channel due to the relative large body thickness.

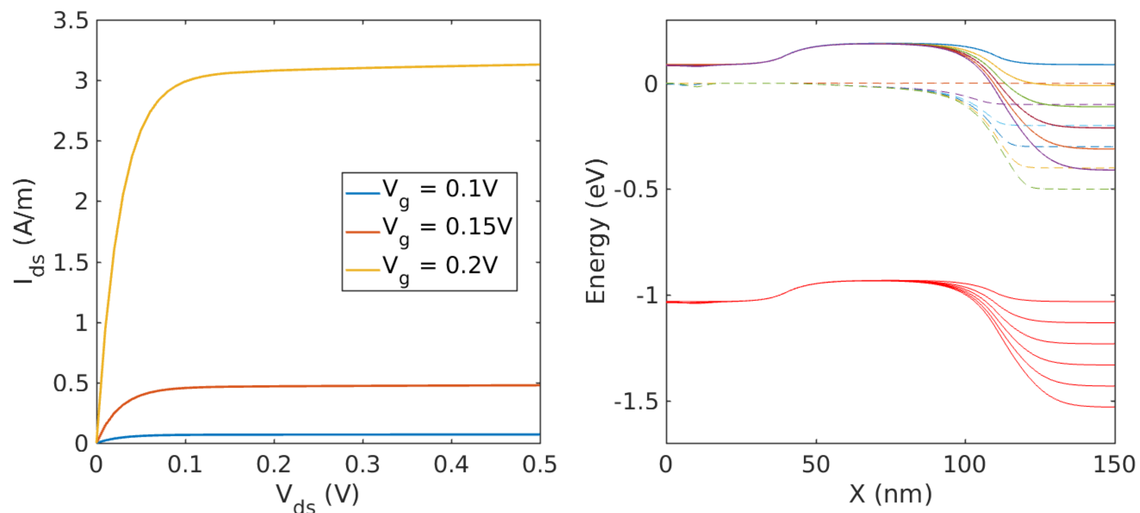

Figure 5.5: Simulation results for a thick body 2D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.2V$.

Thin Body

In the second experiment, the body thickness is decreased to 3nm, yielding a ultra-thin body device which is similar to the more advanced SOI-based MOSFET. It is shown (see Figure 5.6) in the band diagrams that the conduction band (and valence band) keeps unaffected by gate-drain bias inside the channel region, indicating that the channel region has been fully depleted. The current-voltage curves also imply that gate control has been improved for the thin body structure, since the saturation current remains unchanged after pinch-off when compared to the thick-body simulation results.

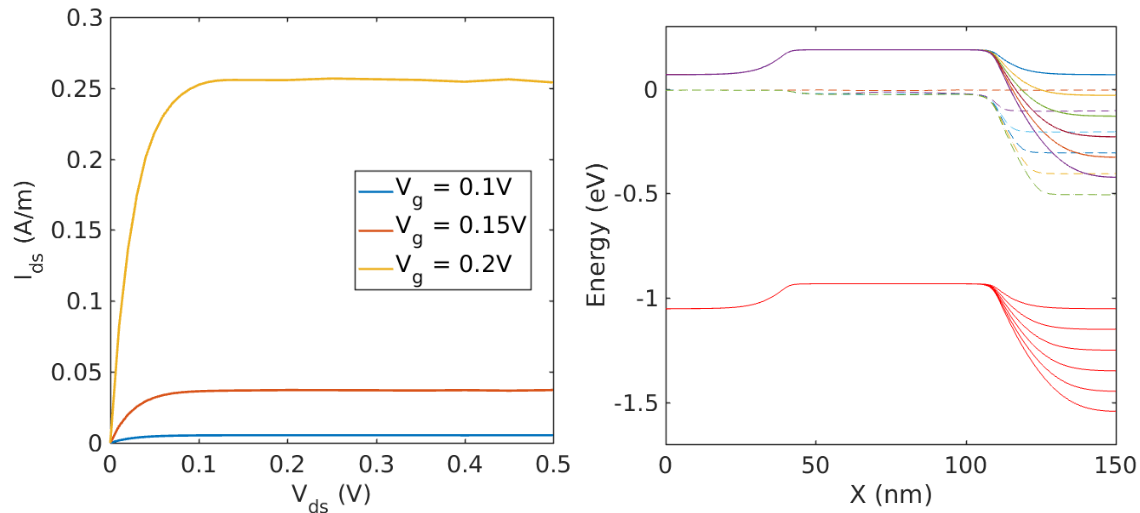

Figure 5.6: Simulation results for a thin body 2D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.2V$.

Short Channel Effect

Short channel effects becomes severe as the channel length further scales down. A quantum scale simulation is performed on short-channel MOSFET device to demonstrate the capability of our NEGF – Poisson solver in simulating ultra-small devices. The body thickness is 3nm; the channel lengths and source (drain) lengths are all 10nm; the oxide thickness is 1nm. The n-doping concentration at source (drain) regions is 1E20cm^{-3} and the channel region is intrinsic without doping. In quantum scale simulation, the convergence between NEGF and Poisson's equation is relatively more difficult to achieve. A fine mesh (spacing less than 0.1nm) and Anderson mixture among electrostatic potentials from different iterations are usually required.

The simulation results are shown in Figure 5.7. When the channel length shrinks to quantum regime, the drain induced barrier lowering (DIBL) becomes significant as shown in the band diagram. As the result, the drain-source current starts to apparently increase with drain bias in saturation region, suggesting a notable decrease in the output resistance

of MOSFET and further loss of control of the gates.

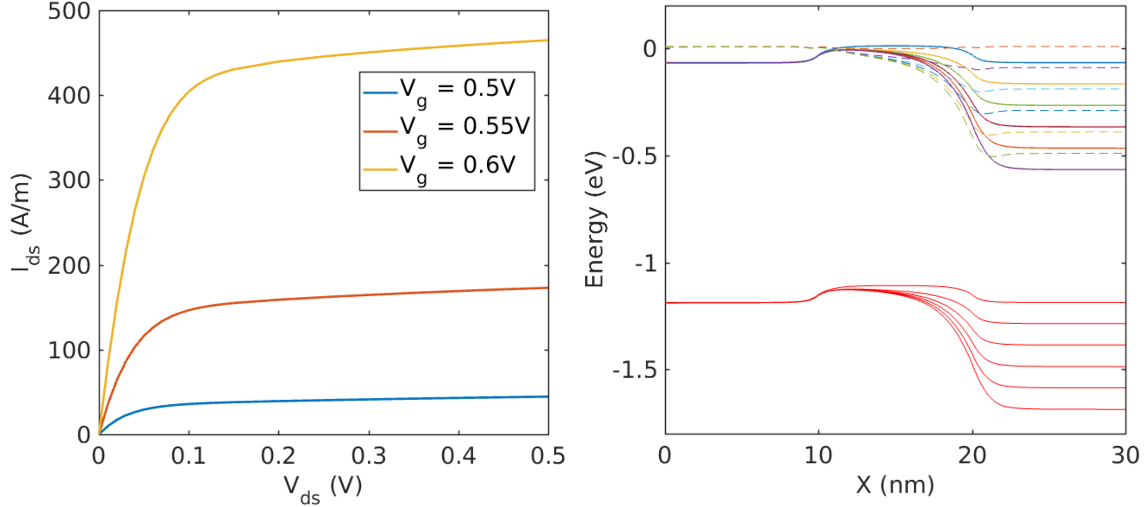

Figure 5.7: Simulation results for a short channel 2D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.6V$.

5.3 *Poisson Solver for 3D Gate All-around MOSFET*

Next, we extend our development of Poisson solver for 2D toy model to a more practical 3D structure. A prototypical device for a 3D gate all-around SOI MOSFET is shown in Figure 5.8. The cross section of the 3D MOSFET is identical to the 2D system discussed previously. Note that the channel region is wrapped all-around by a thin layer of oxide and gate electrode. The same doping strategy $n^+ - n - n^+$ is applied. Here, we only discuss the major difference in terms of numerical implementation in each step, since the simulation process for the 3D system are similar to the 2D simulation.

In NEGF calculation, the Fermi-Dirac integral is reduced to Fermi distribution function in 3D systems. Therefore, all the relevant equations need to be modified accordingly.

- The self-energy matrices

$$\Sigma^< = -2i\,\mathrm{Im}[\Sigma_L^r]f(E - E_{fs}) - 2i\,\mathrm{Im}[\Sigma_R^r]f(E - E_{fd})$$

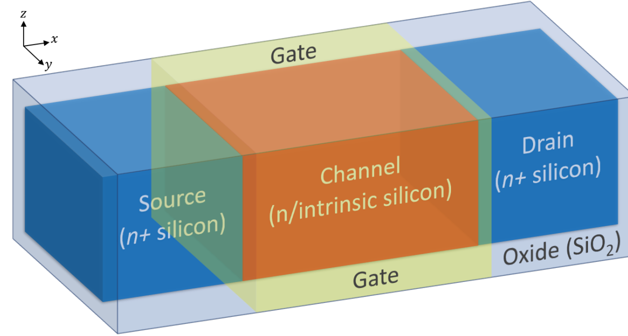

Figure 5.8: A 3D gate all-around SOI MOSFET device, which is a straightforward extension from the 2D dual-gate MOSFET.

where $f(E) = [1 + \exp(E/k_B T)]^{-1}$ is the Fermi distribution function.

- Electron density at each grid point (m^{-3}):

$$n(j)_{\mathrm{NEGF}} = \frac{1}{2\pi a^3} \int n(E, j) \mathrm{d}E$$

- Current density ($\mathrm{A/m}^2$):

$$J = \int J(E) \mathrm{d}E$$

The steps of solving Poisson's equation remains unchanged mathematically. Note that in 3D system, the "gate" boundaries actually represent 4 surfaces surrounding the channel region.

5.3.1 Results

A quantum scale simulation is performed on a ultra-small 3D MOSFET using our 3D NEGF – Poisson solver. The semiconductor cross section is a 2nm × 2nm square; the surrounding gate oxide layer is 1nm thick; the source (drain) and channel lengths are all 10nm. The n-doping concentration at source (drain) regions is 1E20cm^{-3} and intrinsic doping is assumed in the channel region.

The result is shown in Figure 5.9. When compared to the simulation results of 2D MOSFET, the DIBL effect is much suppressed. The control ability of gate is greatly enhanced due to the significant increase in contact area between gate electrode and channel region. It is demonstrated that the introduction of surrounding gate technology can effectively reduce the impact of DIBL and suppress the short channel effect.

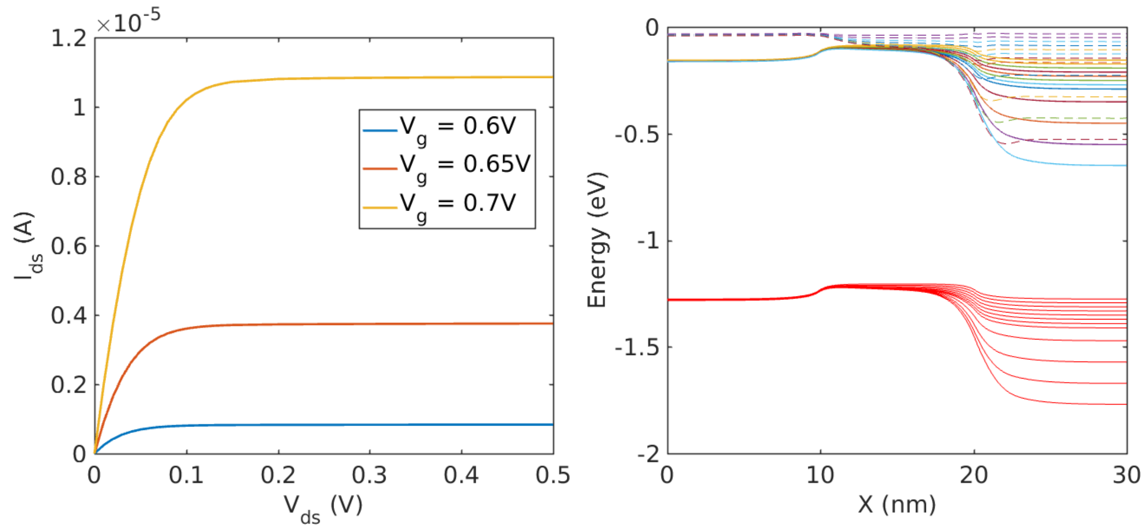

Figure 5.9: Simulation results for a gate all-around 3D MOSFET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.6V$.

5.4 *Poisson Solver for Graphene FET*

The presented Poisson solver can be extended to be applied in tight-binding simulation. A three-terminal graphene-based FET device (shown in Figure 5.10) consisting of graphene monolayer sandwiched by top and bottom insulating layers (usually SiO_2 or hBN) is simulated to illustrate the process. Key elements for modeling such devices using the developed NEGF – Poisson framework involve:

- In NEGF calculation, we only consider the graphene layer (carbon atoms), which is

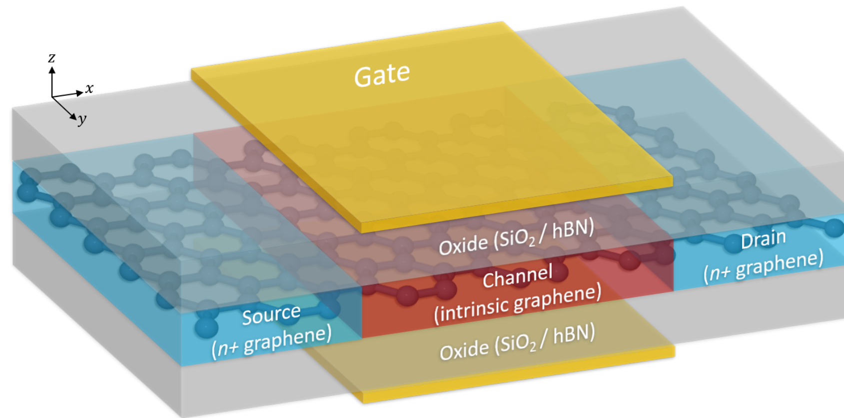

Figure 5.10: A 3D graphene FET device formed by a single graphene layer sandwiched by two thin oxide layers. Left graph shows the graphene monolayer and the right graph shows the lateral view of the device.

straightforwardly discretized using tight-binding approximation, so that the charge density is obtained at each atomic site.

- In Poisson calculation, due to the non-regular mesh grids constructed in the semiconductor domain, the system decomposition is usually performed by finite element approximation.

- Armchair-edged graphene nanoribbon is used here so that one can determine whether the graphene layer is a semiconductor or a semi-metal by modulating its lateral width. The graphene-FET structure requires that the graphene to be semiconductor.

- Similar to the MOSFET structures, the doping strategy is n^+-i-n^+, and the external gate electrodes only covers the intrinsic area.

Given the tight-binding decomposition of the device Hamiltonian, the NEGF calculation process and the governing equations are identical to what we described in the last section. In the rest of the section, we will focus on the numerical details in solving Poisson's equation.

5.4.1 Discretization

Since the tight-binding grids are usually non-regular, a finite-difference implementation of Poisson solver requires certain approximation such as the interpolation of the charge density from a non-regular mesh to a regular mesh. Typically, a generic Poisson solver is implemented using the Finite Element Method (FEM), which is capable of handling mesh grids in arbitrary pattern. The discretization for the graphene-based FET device consists of two parts:

1. The insulating layers are discretized using regular mesh (cuboidal 7-point-stencil grids).

2. The tight-binding grid (atomic sites) of the graphene layer used in NEGF calculation can be directly adopted in Poisson solver. Sometimes, one might also include interbond grid points (usually a few grid points between each C–C bond) in order to improve the stability of the self-consistent iteration when solving non-linear Poisson's equation. In this work, we include 2 additional grid points between each C–C bond as interbond grid points. The impact of the interbond grid points will be discussed in later section. A sample grid points constructed for the graphene layer is show in Figure 5.11(a).

The regular mesh discretizing insulating layers and the non-regular mesh of the graphene layer are then coerced together to create the grids for the graphene-FET device. The 3D tetrahedron FEM mesh can then be generated using the well-known Delaunay triangulation algorithm. A sample mesh is illustrated in Figure 5.11(b) with graphene layer located at $z = 0$.

5.4.2 FEM Formalism

Here we consider the FEM formalism for the non-linear Poisson's equation $-\nabla \cdot (\varepsilon \nabla V) = \rho(V)$. Using the iterative scheme, we can write the equation as:

$$-\nabla \cdot \varepsilon_0 \varepsilon_r \nabla \left(V_{\text{old}} + \delta V\right) = \rho \left(V_{\text{old}} + \delta V\right) \tag{5.5}$$

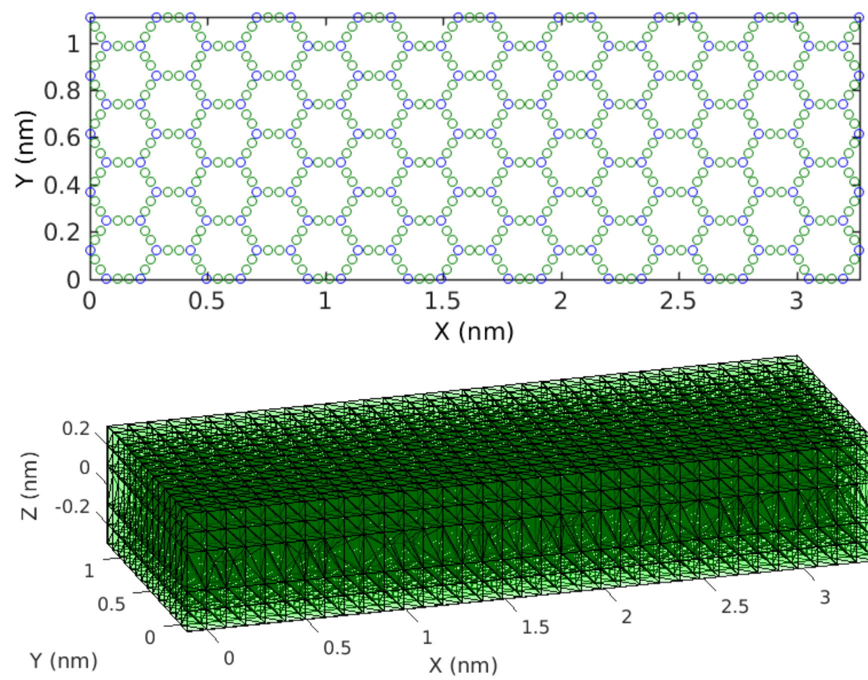

Figure 5.11: (a) Illustration of grid points for the graphene layer. Blue circles denote the atomic sites and red circles denote the interbond grid points. (b) A final tetrahedron FEM mesh generated for the graphene-FET device consisting of a graphene monolayer vertically sandwiched by two insulating layers. In the sample device, the graphene layer contains 160 carbon atoms and locates at $z = 0$.

In each iteration, potential profile can be updated by solving a linear system regarding δV:

$$-\left(\nabla \cdot \varepsilon_0 \varepsilon_r \nabla + \frac{\partial \rho}{\partial V}\right) \delta V = \rho\left(V_{\text{old}}\right) + \nabla \cdot \varepsilon_0 \varepsilon_r \nabla V_{\text{old}} \tag{5.6}$$

Next, it is straightforward to apply the finite element approximation. Given a triangulation discretization of the computation domain Ω, we use $\{\varphi_i\}_{i=1}^N$ to denote the FEM basis functions, typically a set of continuous piecewise linear functions. N is the number of nodes with the triangulation. By assuming the Neumann boundary condition $\nabla V = 0$ on boundary $\partial\Omega$, the weak form of equation 5.6 is equivalent to

$$\int_\Omega \varepsilon_0 \varepsilon_r \nabla \delta V \cdot \nabla \varphi_i dx - \int_\Omega \frac{\partial \rho}{\partial V} \delta V \varphi_i dx = \int_\Omega \rho\left(V_{\text{old}}\right) \varphi_i dx - \int_\Omega \varepsilon_0 \varepsilon_r \nabla V_{\text{old}} \cdot \nabla \varphi_i dx \tag{5.7}$$

holes for each basis function φ_i $(i = 1, 2, \ldots, N)$.

FEM assumes that the solution δV can be written as a linear combination of the basis functions

$$\delta V = \sum_{j=1}^{N} c_j \varphi_j \tag{5.8}$$

for some cooeffcients c_j to be determined. Equation 5.7 ends up with a $N \times N$ system of linear equations

$$\sum_{j=1}^{N} c_j \left(\int_\Omega \varepsilon_0 \varepsilon_r \nabla \varphi_j \cdot \nabla \varphi_i dx - \int_\Omega \frac{\partial \rho}{\partial V} \varphi_j \varphi_i dx\right) = \int_\Omega \rho\left(V_{\text{old}}\right) \varphi_i dx - \int_\Omega \varepsilon_0 \varepsilon_r \nabla V_{\text{old}} \cdot \nabla \varphi_i dx \tag{5.9}$$

for $i = 1, 2, \ldots, N$, which can be expressed in a matrix form.

Therefore, by numerically solving equation 5.9, we can update potential profile iteratively. Dirichlet boundary condition for δV can also be easily addressed by $c_j = V_{\text{gate}} - V_{\text{old}}$ for FEM nodes j located at "gate" boundaries.

5.4.3 Results

The graphene based FET device we modeled is made up of a 1.2nm × 34nm graphene single layer and two 10nm thick insulating layers. The width of the graphene layer has been chosen

so that the graphene nanoribbon behaves as a semiconductor of bandgap around 0.89eV. The doping density of source (drain) area is 1E16cm^{-3}, and the channel region is intrinsic whose length is 10nm.

The simulated transport properties and the corresponding band diagrams are shown in Figure 5.12. Similar to the thin body MOSFET device, the channel region is fully depleted due to the 2D nature of the graphene monolayer and its low density of states near charge neutrality point. As reflected in the current-voltage characteristics, a strong gate control is observed even in the short-channel system, making graphene-based device a promising candidate for the next-generation integrated circuits.

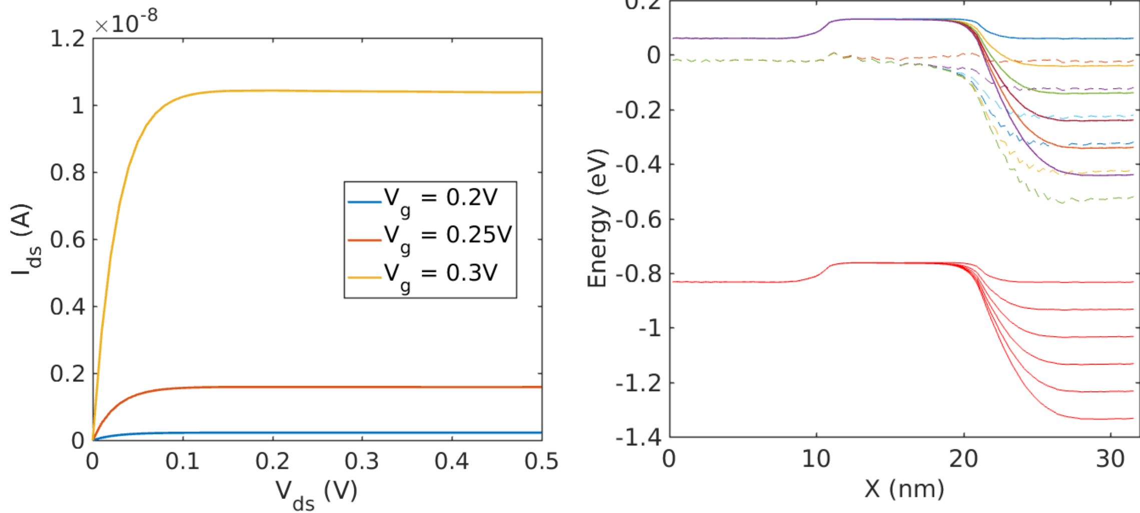

Figure 5.12: Simulation results for a 3D graphene-FET structure. Left figure shows the I-V characteristics for various gate voltages. Right figure shows the conduction band, valence band, and quasi-Fermi level for source-drain bias from 0 to 0.5V at $V_g = 0.3V$.

5.4.4 Discussion

Generating FEM mesh is a critical step in the coupled simulation between NEGF formalism and Poisson equation. In our Poisson model, point-charge approximation is adopted, suggesting that the net charge associated with each atom is approximated by a point charge

located at the position of the atom, instead of distributed as atomic orbitals. Therefore, the Poisson calculation does not require extra grid points besides the atomic sites theoretically. However, in practical simulations, adding extra interbond grid points (between C–C bond in our device) can smooth the potential profile, thus effectively increasing numerical accuracy in Poisson calculation.

We show the test result in Figure 5.13, which repeats the above graphene-FET simulation with various interbond grid points in FEM mesh. It is shown that by adding extra points between atomic bonds, the saturation current tends to converge better and the most significant convergence occurs when the number of interbond grid points changes from 0 to 2. Therefore, in the above section, we choose to add 2 extra grid points between C–C bonds to increase the numerical precision while not sacrificing too much computational cost.

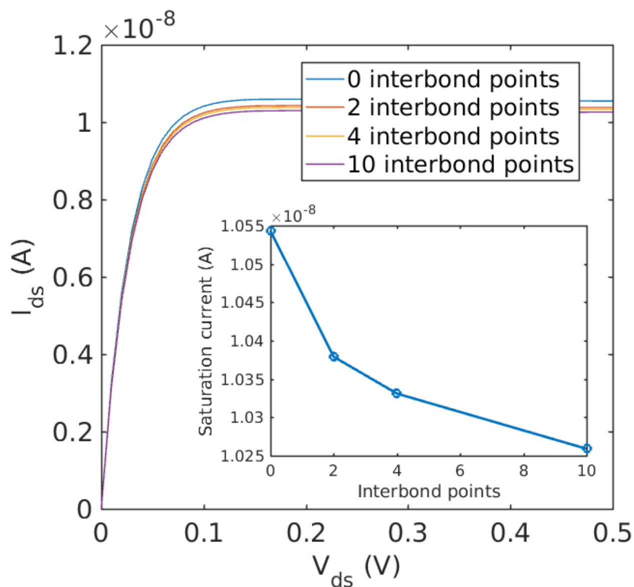

Figure 5.13: Simulation results for a 3D graphene-FET structure for $V_g = 0.3$V with different numbers of interbond grid points.

Chapter 6

APPLICATION: TRANSPORT SIMULATION IN EMERGING 2D ELECTRONIC DEVICES

Hexagonal boron nitride (hBN) is drawing increasing attention as an insulator and substrate material to develop next generation graphene-based electronic devices. As the first application of HSC-extension approach in realistic device modeling, in this chapter, we investigate the quantum transport in heterostructures consisting of a few atomic layers thick hBN film sandwiched between graphene nanoribbon electrodes. The application work involves two main sections:

1. In the first part, we present a gate-controllable graphene-hBN-graphene vertical transistor exhibiting strong negative differential resistance (NDR) effect with multiple resonant peaks, which stay pronounced for various device dimensions. We find two distinct mechanisms that are responsible for NDR, depending on the gate and applied biases, in the same device. The origin of first mechanism is a Fabry-Pérot like interference and that of the second mechanism is an in-plane wave vector matching when the Dirac points of the electrodes align. The hBN layers can induce an asymmetry in the current-voltage characteristics which can be further modulated by an applied bias. We also show that the NDR features are tunable by varying device dimensions.

2. In the second part, the distinct NDR mechanism arising from interlayer angular rotation in the three-terminal graphene-hBN-graphene heterostructures, as a function of both the twisting angle and gate bias, is simulated and analyzed. Analytical expressions for the positions of the NDR peaks in the I-V characteristics are developed. To capture the degradation of peak-to-valley ratios observed in experiment at room temperature,

electron-phonon scattering has been added to the simulation and a good agreement with experiment is achieved, indicating a robust preservation of NDR feature when temperature increases.

6.1 Negative Differential Resistance in Boron Nitride Graphene Heterostructures: Physical Mechanisms and Size Scaling Analysis

6.1.1 Motivation

Graphene, a two-dimensional material with unique mechanical, thermal and electronic transport properties [11] is a promising candidate for nanodevices as it is deeply scaled in one dimension and the lithography offers scaling in the other two dimensions. Building devices based on graphene is, however, partially impeded by the lack of compatible insulating substrate. Hexagonal boron nitride (hBN) has an atomically smooth two dimensional (2D) layered structure with a lattice constant very similar to that of graphene (1.8% mismatch), sufficiently large electrical band gap (4.7eV), and excellent thermal and chemical stability [16], allowing it to be stacked with graphene to build device structures with desired functionalities. Also, hBN reduces the surface roughness of graphene without degrading its giant mobility [17, 77]. The nano-scale devices based on graphene employing atomically thin hBN with novel electrical and optical properties have recently been reported [3, 9, 10, 15, 26, 38, 39, 68, 72, 76, 82]. An appearance of negative differential resistance (NDR) in such devices further interests the researchers as it could potentially impact the number of applications such as high-speed IC circuits, signal generators, data storage, and so on [52].

NDR in double barrier resonant tunneling diodes (DB-RTD), appears when the quasi-bound levels can no longer enhance the tunneling resonantly [20]. Recent theoretical investigations report the appearance of NDR features in pure graphene based devices, involving nanoribbon superlattice [22], doped junctions [18, 23, 33, 57, 74], tunnel-FET [2, 54], and MOSFET structures [75]. These structures typically employ graphene with fine-tuned bandgap, such that graphene behaves more like a semiconductor. NDR effect is also being re-

ported in a single as well as multilayer heterostructure of graphene-hBN-graphene [21, 53, 62]. Reference [28] models NDR peak in a near metallic bi-layer graphene device. Apart from this, such devices could also find applications in multi-valued memory [24, 45]. The multilayer graphene-hBN-graphene heterostructure based electronic devices particularly attract the attention of engineers due to their relatively simpler fabrication [8, 55, 70].

The NDR features in multilayer based devices are to be investigated by exploring current-voltage characteristics as a function of (i) number of hBN layers, (ii) lateral dimensions in determining both the voltage location of NDR peaks and the peak-to-valley ratio, which are essential in the device design, (iii) the role of the asymmetric band offset between hBN and graphene, and (iv) defects and scattering. This, however, necessitates further research to rationalize the underlying physics of the NDR effect and gain insight on how to control its critical properties mentioned above.

In this section, we focus on a prototypical multilayer device is shown in Figure 6.1(a), which consists of layers of graphene and hBN that are vertically stacked. The graphene layers serve as conducting electrodes with a unique band structure while the hBN layers are tunnel barriers. We model the electron transport in these devices by atomistic non-equilibrium Green's function (NEGF) method. Additionally, we demonstrate how the magnitude of current, locations of resonant peaks, and peak-to-valley ratio (PVR) values can be tuned by the device parameters. The modeled devices range from a small system with 6,000 atoms to experimentally feasible sizes up to 70,000 atoms (lateral dimensions 24.6nm $\times$ 27nm).

Next, section 6.1.2 defines our method by discussing the underlying Hamiltonians and the methodology for the computation of the quantum transport. Section 6.1.3 demonstrates the results and discusses the NDR effects with two underlying mechanisms. Section 6.1.7 presents the size scaling analysis.

6.1.2 Method

A prototypical heterostructure consists of two semi-infinitely long monolayer armchair-edged graphene nanoribbon (AGNR) electrodes sandwiching an ultra-thin hBN film, with a ver-

tically applied external gate electric field as shown in Figure 6.1(a). AGNR is employed because it can be engineered as an intrinsic conductor. This forms a vertical tunneling heterostructure with hBN acting as a potential barrier. The hBN film is sandwiched between a bottom and top AGNR, forming a central overlapping heterostructure/multilayer region stacked in AB order (Bernal stacking). The lattice constant mismatch between hBN and graphene is negligibly small, 1.8%, therefore, we build the device structure with the uniform lattice constant of graphene (2.46 Å) only. The system Hamiltonian is constructed using the nearest neighbor tight binding approximation, with the parameters [58, 68]:

$$E^{\mathrm{C}}_{\mathrm{on-site}} = 0, E^{\mathrm{B}}_{\mathrm{on-site}} = 3.34\mathrm{eV}, E^{\mathrm{N}}_{\mathrm{on-site}} = -1.4\mathrm{eV},$$

and

$$t^{\mathrm{C-C}}_{\mathrm{intralayer}} = 2.64\mathrm{eV}, t^{\mathrm{B-N}}_{\mathrm{intralayer}} = 2.79\mathrm{eV}, t^{\mathrm{B-N}}_{\mathrm{interlayer}} = 0.60\mathrm{eV}, t^{\mathrm{C-B/N}}_{\mathrm{interlayer}} = 0.43\mathrm{eV}.$$

Only the low energy p_z orbitals are considered here; so that the Hamiltonian has the same dimension as the total number of atoms simulated. The effect of number of tunneling hBN layers (N_z), the system width (N_x) and the length of the multilayer stacking region (N_y), where units of N_x and N_y are number of atoms, on the device performance is investigated. The nanostructure thickness is $(N_z + 2)$ in units of atomic layers, which includes the two monolayer graphene sheets at the ends.

The bias across the heterostructure is applied by rigidly shifting the electrostatic potential of the bottom graphene electrode by the amount equal to the applied bias as the metallic graphene layers have much higher conductivity than hBN. The electrostatic potential energy of the bottom layer is $U = -eV_b$, where V_b is the applied bias, and the electrostatic potential of the top graphene layer remains zero. The electrostatic potential at each of the sandwiched hBN layers are determined by linearly increasing/decreasing the potential from top to the bottom graphene layer, because the c-axis (out of plane) conductivity of hBN is orders of magnitude smaller than the in-plane conductivity of graphene. The chemical potential of contacts are controlled by bias voltage, namely $\mu_B = -eV_b$ and $\mu_T = 0$ for bottom and top graphene leads respectively. The gate voltage is modeled by shifting the electrostatic

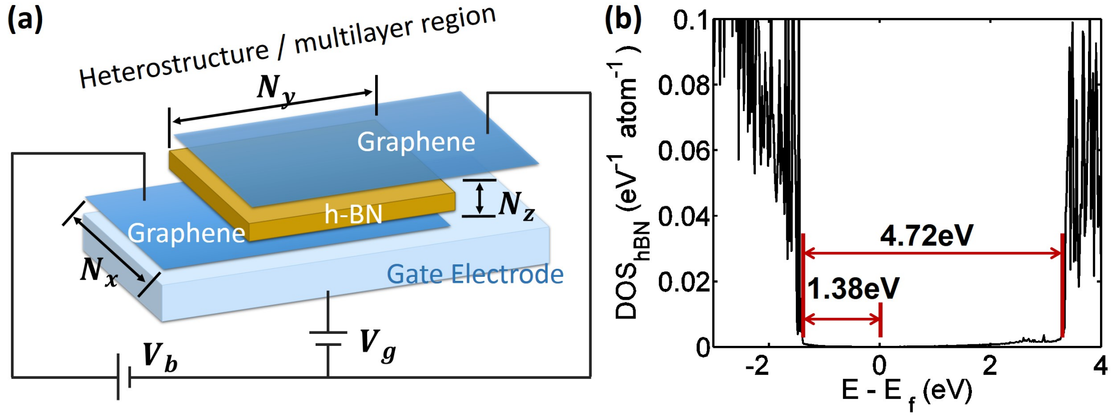

Figure 6.1: A schematic view of the heterostructure device. N_x and N_y represent the width and stacking length of the device respectively. N_z is the number of hBN layers sandwiched between the two AGNR ribbons. All the dimensions are in unit of atomic layers. (b) The average DOS versus Energy of hBN for device with $N_x = 200$, $N_y = 32$ and $N_z = 3$. This shows a 4.72eV bandgap of atomically thin hBN material and a 1.38eV valence band-offset between graphene and hBN stacking structure.

potential at the bottom electrode by $\Delta U = -0.01\text{eV}_\text{g}$, where V_g is the gate voltage. We choose N_x equal to $3n + 2$, where n is an arbitrary positive integer [41, 64], such that the AGNR have zero bandgap. All calculations were performed at 300K.

We simulate the transport properties of the device by using the NEGF formalism described in the above chapters. Here the incorporation of HSC-extension realizes the simulation of requisite large scale systems. In Figure 6.1(b), we plot the DOS for a structure with a width of $N_x = 200$, $N_y = 32$, and with three hBN layers $N_z = 3$, at zero bias. The bandgap of hBN is found to be around 4.72eV, and the valence band offset between hBN and graphene is around 1.38eV, which is consistent with the prior work [37].

6.1.3 Mechanisms

6.1.4 Origin of Multiple NDR Peaks (Mechanism 1)

Figure 6.2(a) presents the computed current-voltage characteristics of the heterostructure with a transverse width of $N_x = 62$ (7.6nm), stacking length $N_y = 32$ (6.8nm) and three hBN layers (1.4nm) serving as the tunneling barrier. First we consider the highlighted curve with $V_g = 0$, in which case two NDR peaks emerge at $V_b = 0.3$V and 0.66V, respectively. We attribute the formation of these multiple NDR peaks to the Fabry-Pérot like interference in the multilayer region (mechanism 1), as rationalized below.

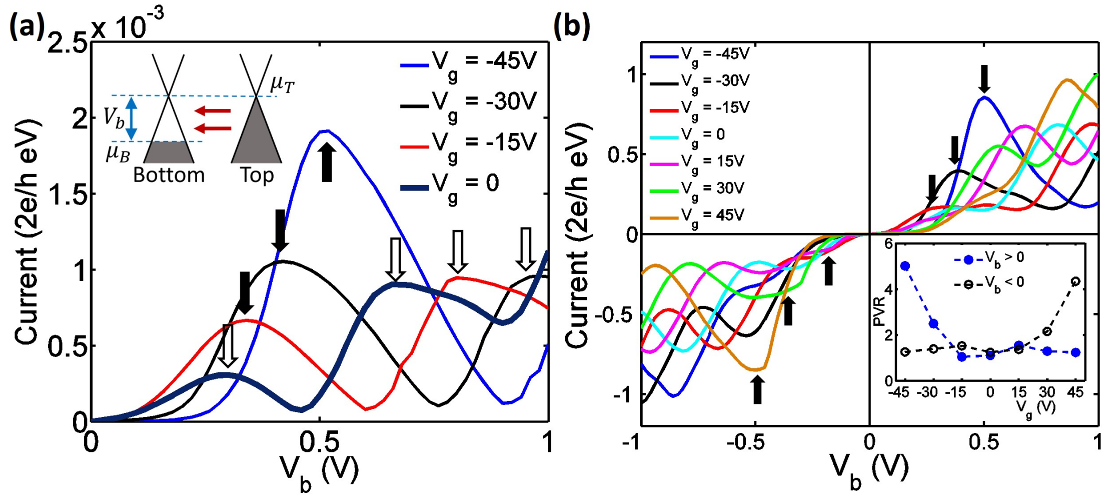

Figure 6.2: (a) Current versus drain voltage for a device with $N_x = 62$, $N_y = 32$ and $N_z = 3$. V_g varies from -45V to 0. The black solid arrows in the four plots mark the current resonant peaks due to mechanism 2, and the empty arrows marks the NDR peak due to mechanism 1. The inset explains the resonant tunneling induced from mechanism 2. The difference between Fermi energy and Dirac point in bottom graphene is induced by gate potential. When $V_b = -0.01V_g$, the electronic spectra of top and bottom electrodes are tuned into alignment, allowing the resonant tunneling. (b) Current versus drain voltage for large device with $N_x = 200$, $N_y = 32$ and $N_z = 1$. Here V_g varies from -45V to +45V. Inset shows an asymmetric PVR relationship with the applied vertical gate potential.

To understand the multiple NDR peaks, we calculate the transmission and the average density of states (DOS_g and $\mathrm{DOS}_{\mathrm{hBN}}$) at V_b =0.3V, 0.46V and 0.66V, corresponding to

the first peak, the first valley and the second peak in the current-voltage characteristics. In the DOS_g curves (Figure 6.3), the average DOS of bottom and top graphene layers are plotted separately. In particular, the huge peaks in DOS_g are marked as peak S, which captures the edge states due to the zigzag-shaped cut-ends of graphene ribbons. The blue DOS_{hBN} curves show the average DOS of the three hBN barrier layers. For the sake of comparison, the transmission coefficient at equilibrium is also plotted with the black dashed curves. The chemical potentials of the bottom and top AGNR are marked as vertical black lines (μ_B and μ_T). The transmission and DOS_g show a strong Fabry-Pérot like resonant feature in the low energy window. The semi-infinite top and bottom AGNRs couple with hBN at the central heterostructure (multilayer) region. The potential discontinuity caused by the interaction between the hBN cut-ends and the graphene layers create a resonant cavity in the overlapping region at both the top and bottom graphene layer. When electrons transport across the boundaries between graphene monolayer and hBN multilayer regions, partial reflections occur at the interfaces. As a result, the transmission is oscillatory with peaks and valleys corresponding to constructive and destructive interferences.

The current is determined from the area enveloped by the transmission curve in the energy window bounded by the Fermi levels of two electrodes, μ_B and μ_T (black dash-dot lines in Figure 6.3). The transmission peak (P) at $E = -0.3\text{eV}$ in Figure 6.3(a) mainly contributes to the tunneling current. At low bias regime ($V_b < 0.18\text{V}$), we find that this transmission peak P is enhanced, resulting in the increase of current with applied bias. The resonant tunneling occurs when the constructive quantum interference assists the tunneling of electrons from the top to bottom electrodes at specific energies. When the bias is further increased (until 0.46V), the transmission peak P is reduced due to destructive interference despite the fact that the energy window for carrying current enlarges. This transmission reduction begins to dominate after $V_b = 0.3\text{V}$, which induces a drop in current. At $V_b = 0.46\text{V}$, there is a large suppression of transmission within the bias window, which creates a large tunneling gap, leading to the current valley as reflected in the highlighted curve of Figure 6.2(a). Note that the density of states is large in both graphene electrodes even when the transmission

is small as see in Figure 6.3(b). Then, the transmission starts to increase again at around $V_b = 0.66$V due to the constructive interference.

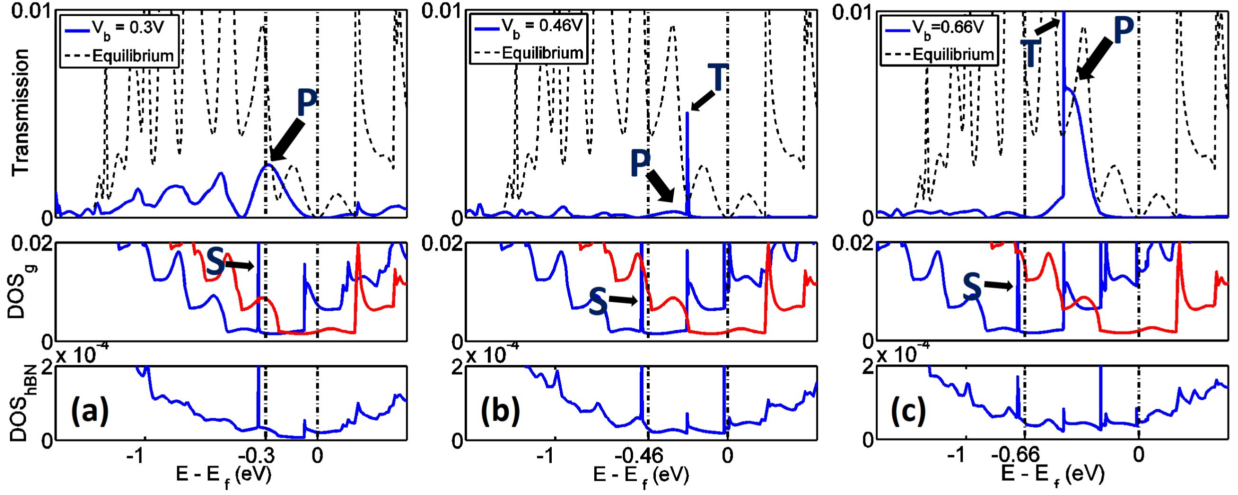

Figure 6.3: Transmission and DOS plot at various biases for I-V curve ($V_g = 0$) in Figure 6.2(a). (a)-(c) specifies the bias potential $V_b = 0.3$V, 0.46V and 0.66V respectively. In the transmission plots, black dashed curves are transmission coefficient when $V_b = 0$. In DOS$_g$ plots, blue and red curves represent DOS of bottom and top graphene sheets respectively. Vertical dash-dot lines give the chemical potentials at both graphene ends μ_B and μ_T, which determines the bias window. S mark the DOS peaks resulting from the zigzag shaped edges of graphene cut ends. P mark the transmission peaks that mainly contribute to the current. T represent the tunneling peaks due to the energy alignment of subbands in top and bottom graphene contacts; they do not contribute significantly to current. Units of DOS are number of states per atom per eV.

An interesting feature of Figure 6.3(a) is that only one transmission peak is observed at around $E = -0.3$eV (μ_T), while a symmetric peak at $E = 0$eV (μ_B) is clearly absent. This is due to the fact that the presence of hBN layers break the symmetry. We could understand this from DOS$_{\text{hBN}}$ plot at $V_b = 0.3$V (Figure 6.3(a) DOS$_{\text{hBN}}$ curve), which shows a peak near $E = -0.3$eV but a valley at $E = 0$eV. This means that electrons at $E = 0$eV see a stronger barrier when tunneling between AGNR layers, and suggests the breaking of the $\pi - \pi^*$ symmetry. This argument is tested by considering a symmetric tunnel barrier, where such an asymmetry in transmission does not exist. In Figure 6.3(b) and (c), sharp

transmission peaks (marked as peak T) are observed. This significant tunneling enhancement results from the energy level alignment between the subbands of top and bottom graphene electrodes, as reflected in the corresponding DOS_g features.

6.1.5 Gate Induced NDR Peak (Mechanism 2)

We next discuss the second mechanism that leads to single intense NDR peak by investigating the operational behavior of the heterostructure in the presence of an external gate voltage (V_g). Figure 6.2 shows the current-voltage characteristics for a family of V_g ranging from -45V to +45V for two devices with different sizes. Take the current-voltage curve at $V_g = -45$V as an example; At $V_b = 0$, the negative gate voltage shifts the energy of Dirac point in the bottom AGNR electrode to $U = 0.45$eV at equilibrium, while preserving the chemical potentials from two contacts at $\mu_B = \mu_T = 0$. At $V_b = 0.45$V (see Figure 6.2(a) inset), the Dirac points of bottom and top AGNR electrodes are aligned. As a result, electrons can tunnel from the valence band of the top graphene layer to the conduction band of the bottom graphene layer owing to the in-plane wave vector conservation [8]. This particular mechanism (mechanism 2) induces the resonant transmission and results in the large current peaks marked by solid arrows in Figure 6.2. It is noticeable that current peak positions are shifted from the theoretical prediction ($V_b = -0.01V_g$) based on mechanism 2 only. This occurs when the strength of mechanism 2 is comparable to that of mechanism 1, when the voltage at which the peak current occurs is influenced by the Fabry-Pérot like interference. The superposition of mechanisms 1 and 2 leads to the current peak displacement, which is larger at low gate voltage (for example, $V_g = -15V$).

6.1.6 Gate response

Gate voltage has different impacts on NDR induced by two distinct mechanisms discussed above. In Figure 6.2(a), the peak current (solid arrows) increases with V_g as the peak current is proportional to the number of carriers between μ_B and μ_T when Dirac points of the top and bottom graphene align (see inset of Figure 6.2(a)). In contrast to this, we find that in

Figure 6.2(a), the peak current of the NDR peaks induced by mechanism 1 (empty arrows) are relatively insensitive to V_g. In addition, the PVR values of these NDR peaks are also V_g-insensitive because the vertical gate potential tunes the resonant energies for constructive interference, but do not affect the number of tunneling carriers. Consequently, for two types of NDR effects in a single device, the amplitudes of current peaks for mechanism 2 is weaker than that for mechanism 1 at low V_g, but can become significantly stronger at large gate voltage, as shown in Figure 6.2(a) when $V_g = -45$V. When the device structure is enlarged to $N_x = 200$ and $N_y = 32$, the current-voltage curves for various gate voltages (Figure 6.2(b)) show that the multiple NDR peaks stay clearly defined and their locations are strongly gate-controlled. We point out that the current-voltage curves are asymmetric for positive and negative biases even at $V_g = 0$ as the hBN layers breaks the $\pi - \pi^*$ symmetry in the multilayer system. We also note that after the NDR peak, our calculations clearly show a trend of increase in current with increase in drain voltage, in a manner qualitatively similar to the experiments. [8]

6.1.7 Size Scaling Analysis

System dimensions are the key ingredients in engineering the device performance. In this particular multilayer heterostructure, for instance, the device width determines the number of subbands in ANGR electrodes and the heterostructure length determines the length of interference region. Based on the two distinct mechanisms responsible for the multiple NDR peaks, it is intuitive that the device dimensions have significant and non-trivial influence on the NDR features rather than simply tuning the current magnitude by following quantum mechanical rules or Ohm's law. In order to comprehend such influences, the scaling analysis of the device dimensions, namely the lateral (x, y) dimensions which defines the overlap area between hBN and graphene layers and the z-direction (number of hBN layers), is performed in this section. However, we do not consider the electron-phonon scattering effects during this analysis as they do not significantly alter the outcomes of the analysis.

6.1.8 Tunneling Barrier Thickness (N_z)

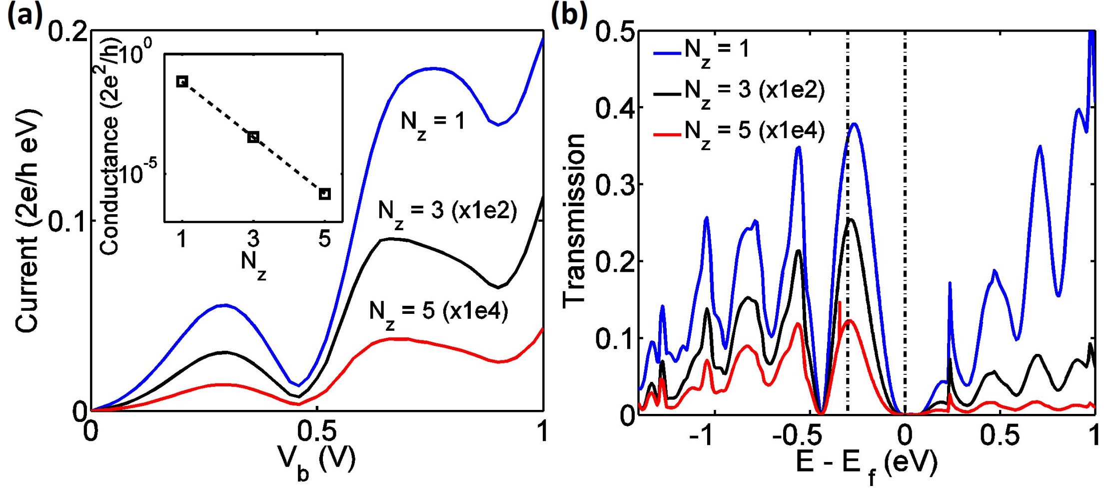

Figure 6.4: (a) Current-voltage curves for devices with different N_z, with fixed $N_x = 62$ and $N_y = 32$. Here the current value for cases when $N_z = 3$ and $N_z = 5$ are scaled by 1E2 and 1E4 respectively. The inset plots the low bias conductance of the three current-voltage curves. (b) Transmission relationship for devices with different N_z and fixed $N_x = 62$ and $N_y = 32$ at $V_b = 0.3V$, corresponding to the first current peaks shown in (a). Again, the transmission coefficient value for cases when $N_z = 3$ and $N_z = 5$ are scaled by 1E2 and 1E4 respectively.

Representing the thickness of tunneling barrier, N_z homogeneously modifies the current magnitude at different applied voltages, whereas has little effects on the peak positions and corresponding PVR. In Figure 6.4(a), the hBN thickness (N_z) is varied from 1 layer (0.6nm) to 5 layers (2nm), while both N_x and N_y are fixed. The magnitudes of current are scaled by a multiplicative factor to present results on the same plot for different values of N_z. The transmission versus energy (Figure 6.4(b)) shows that while the magnitude of transmission depends strongly on N_z, the locations of peaks depend weakly on N_z. Note that the dependence of current magnitude on N_z lose its validity in the case when all the incident modes can tunnel through a thin barrier. This is because a thinner barrier only increases the tunneling probability of electrons without affecting the number of incident modes.

6.1.9 Width of AGNR (N_x)

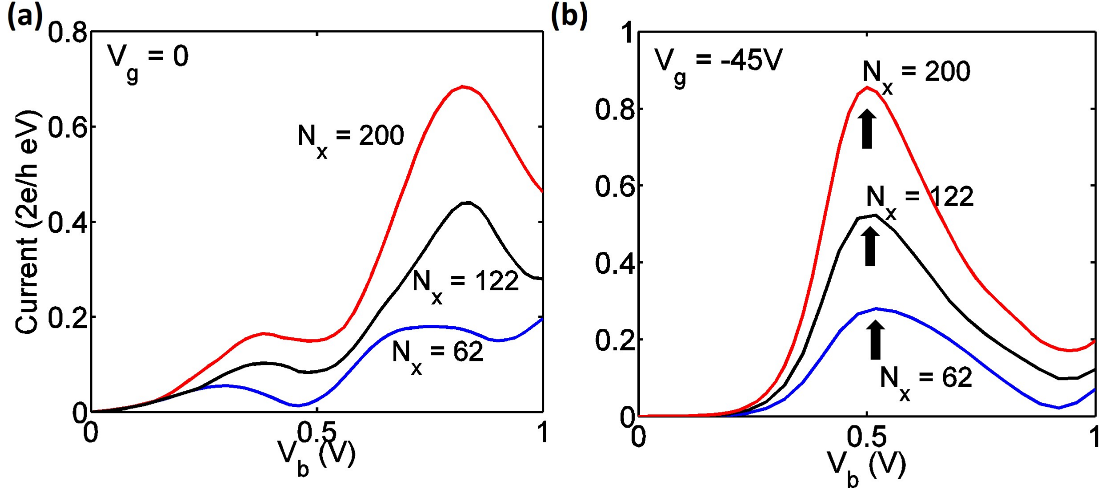

Figure 6.5: (a) Current versus drain voltage at $V_g = 0$ for devices with various N_x, with fixed $N_y = 32$ and $N_z = 1$. (b) Current versus drain voltage at $V_g = -45V$ for devices with various N_x, with $N_y = 32$ and $N_z = 1$. Black arrows mark the NDR peaks due to mechanism 2.

For the mechanism 1 (at $V_g = 0$), the density of subbands for the monolayer AGNR electrodes and the heterostructure region depends on the graphene nanoribbon width, i.e. the energy intervals between subbands for the structure with $N_x = 200$ are about three times smaller than that for the structure with $N_x = 62$. Therefore, a larger number of subbands contribute to current under lower biases, resulting in initial increase in current with N_x, as seen in Figure 6.5(a). When the gate voltage is -45V, the NDR peaks induced by the mechanism 2 are observed near $V_b = 0.45$V for different N_x in Figure 6.5(b). The heights of these peaks increase with device widths because the number of subbands carrying current between μ_B and μ_T grows with the width of the AGNR electrode (inset of Figure 6.2(a)). We summarize the peak currents and PVR values for both mechanisms in Table 6.1. Although the peak current is larger for the wider device, a rapidly decreasing PVR value can be observed. This is because of the stronger band-to-band tunneling between two AGNR

contacts with a larger width, arising from the smaller subband spacing.

Table 6.1 exhibits a PVR value up to 13 for $N_x = 62$, which can be further increased to over 60 when N_x shrinks to 14, showing the potential for the heterostructure to be utilized in both digital logic and memory. However, in reality to achieve the large PVR values will require a downscaling of N_x and minimization of decoherence.

N_x ($V_g = 0$)	62	122	200
Peak Current ($2e^2/h$)	0.05	0.10	0.16
PVR	4.2	1.19	1.06
N_x ($V_g = -45V$)	62	122	200
Peak Current ($2e^2/h$)	0.28	0.53	0.85
PVR	13	5.9	5.0

Table 6.1: Peak current and PVR values as a function of for both mechanisms. (I-V curves from Figure 6.5)

6.1.10 Length of the Heterostructure (N_y)

The length of the central multilayer region determines the number of incident carriers, and also characterizes the size of the Fabry-Pérot like interference cavity. For mechanism 1, when the heterostructure length N_y changes from 16 (3.4nm) to 64 (13.6nm), the number of transmission peaks increase as shown in Figure 6.6(a). The NDR peaks appear at $V_b =$ 0.38V, 0.8V for $N_y = 32$, which shifts to lower V_b i.e. at 0.2V, 0.4V respectively, when $N_y = 64$. This is because the resonant transmission appears at various energies, which vary inversely proportionally with the length of the interference cavity N_y. Experiments where the overlap between two graphene nanoribbons are altered should be able to reveal the differences in oscillations of I-V characteristics as a function of N_y. We note that experiments with changing overlap have been performed in carbon nanotubes before [12, 35] and future experiments in BN-graphene heterostructures should be useful in studying these features.

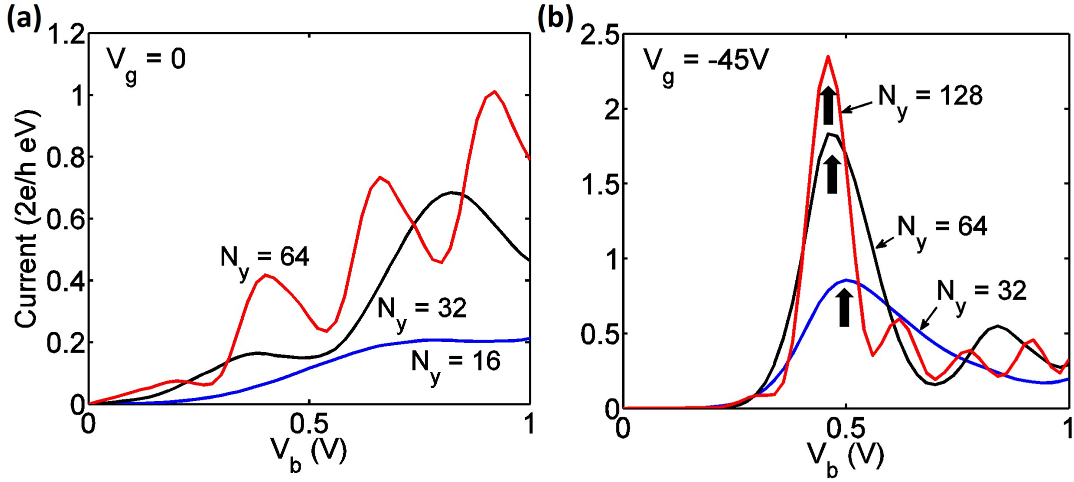

Figure 6.6: (a) Current versus drain voltage at $V_g = 0$ for devices with various N_y , with fixed $N_x = 200$ and $N_z = 1$. (b) Current versus drain voltage at $V_g = -45V$ for devices with various N_y, with $N_x = 200$ and $N_z = 1$. Black arrows mark the NDR peaks due to mechanism 2.

With an ultrathin tunneling barrier ($N_z = 1$), electrons have high tunneling probabilities and thus the current is mainly limited by the number of modes incident within the energy window. Graphene layer has low DOS near Dirac point, yielding the saturation of peak current at largeN_y. It is also observed that for larger N_z ($N_z \geq 3$, results not shown), the peak current increases with N_y rapidly without saturation. This is consistent with our previous discussion since a thicker hBN tunneling barrier greatly suppresses the overall tunneling transport probability and therefore the peak current magnitude is a strong function of the number of carriers in graphene.

6.1.11 Summary

In the first part of the application of the HSC-extension simulator, we have systematically investigated the charge transport properties of a three-terminal graphene-hBN-graphene multilayer heterostructure device as a function of device dimensions, so as to further understand the underlying mechanisms for negative differential resistance. The prototypical graphene-

hBN-graphene multilayer heterostructure has two distinct mechanisms that can introduce NDR behavior.

- The first mechanism involves a Fabry-Pérot like resonant feature due to interference in the multilayer heterostructure region, which can produce multiple current peaks.

- In the presence of an external gate, resonant tunneling can also occur when the electronic spectrum (Dirac points) of the top and bottom graphene electrodes align, which leads to a second mechanism for resonant tunneling.

Both mechanisms respond to gate voltage distinctly. Gate voltage only controls the locations of NDR peaks from mechanism 1 while can tune both PVR and locations of NDR peaks due to mechanism 2.

Size scaling analysis provides insight into the device physics that determines the number of NDR peaks, the variation of peak current and PVR value with change in device dimensions.

- The hBN thickness exponentially controls the magnitude of current without significantly affecting the NDR features.

- For devices with larger widths (N_x), the multiple current peaks preserve but with decreasing PVR values for both mechanisms.

- For mechanism 1, the bias voltages at which multiple current peaks, the number of peaks increase with length. In contrast to this, location of single peak originated from mechanism 2 is independent of the length.

We believe that the negative differential resistance's sensitivity to the system dimensions will provide additional insights for future theoretical and experimental investigations.

6.2 Negative Differential Resistance in Graphene Boron Nitride Heterostructure Controlled by Twist and Phonon-Scattering

6.2.1 Motivation

In the previous section, we have theoretically rationalized two distinct physical mechanisms that are responsible for the NDR effects in graphene-hBN-graphene heterostructure, namely the Fabry-Pérot like quantum interference and the bias controlled Dirac cone alignment [94]. Our work assumes a perfect "AB" lattice structure between the hBN and graphene sheets. However, the lattice misorientation between stacked 2D atomic crystals is unavoidable during fabrication [90]. In this section, by introducing a tunable angular misorientation between graphene and hBN layers, we investigate the transport properties for a twisted graphene-hBN-graphene three-terminal device [92].

6.2.2 Methods

The device [see Figure 6.7(a)] consists of two semi-infinitely long monolayer armchair-edged graphene nanoribbon (AGNR) electrodes sandwiching a single layer hBN film as a tunneling barrier [8, 90]. An external gate electric field is applied vertically to the heterostructure. The system construction here is different from the one used in the last section in two ways:

1. The top graphene layer is rotated by a small tunable angle θ with respect to the central hBN. The sizes of the bottom AGNR and hBN sheets are 22.6nm (L_x along transverse direction) $\times$ 13.4nm (L_y along transport direction), and the size of the top AGNR sheet is 14.9nm $\times$ 13.4nm.

   The system Hamiltonian is constructed by considering a single p_z orbital for C, B and N atoms [96]. Different from the tight-bindin Hamiltonian in the last section, we adopt a Slater-Koster model [95] to capture the modulation of the interlayer hopping amplitude due to the lattice misorientation.

2. The second difference is the electrostatic model. Given the values of the bias voltage

(V_b) and the gate voltage (V_g), the chemical potentials of top and bottom AGNR electrodes are determined by solving the following equations [86]:

$$\Delta\varphi_b + \mu_T - \mu_B = eV_b \tag{6.1}$$

$$\Delta\mu_B - \varphi_g = eV_g \tag{6.2}$$

In equation 6.1, the first term $\Delta\varphi_b = e^2 d_{BN} n_T/\epsilon_{BN}$ is the electrostatic energy difference between graphene electrodes (or equivalently the energy difference between two Dirac points). d_{BN} and ϵ_{BN} are the thickness and dielectric constant of hBN barrier. $n_{T(B)}$ is the electron concentration in top (bottom) graphene sheet. The second and third terms μ_T and μ_B are the chemical potentials of graphene electrodes defined by $\mu_{T(B)} = \pm\hbar v_F\sqrt{\pi\left|n_{T(B)}\right|}$ with v_F being the Fermi velocity of graphene. Note that the chemical potential in this paper is defined as the energy difference from the Fermi-level to the Dirac points. In the equation 6.2, $\Delta\varphi_g = e^2 d_{OX} n_{ext}/\epsilon_{OX}$ is the electrostatic energy difference between bottom graphene and gate electrode. d_{OX} is the thickness of gate oxide. n_{ext} denotes the gate-induced charge density on gate electrode (typically n-Si), satisfying $n_B + n_T + n_{ext} = 0$.

The quantum transport simulation is carried out within the NEGF framework by adopting the efficient HSC-extension solver.

6.2.3 Features of NDR Peaks

We start by analyzing the twisted device with zero external gate voltage. The simulated I-V curves, shown in Figure 6.8, exhibit strong NDR peaks, whose location and peak current depends on the misalignment angles.

In the untwisted system ($\theta = 0$), the I-V curves in Figure 6.8 show multiple current peaks which are fully induced by a Fabry-Pérot interference mechanism. When θ deviates from perfect alignment and increases, the oscillations gradually disappear. We explain this quenching of current peaks at non-zero twisting angles by looking at the resonant condition

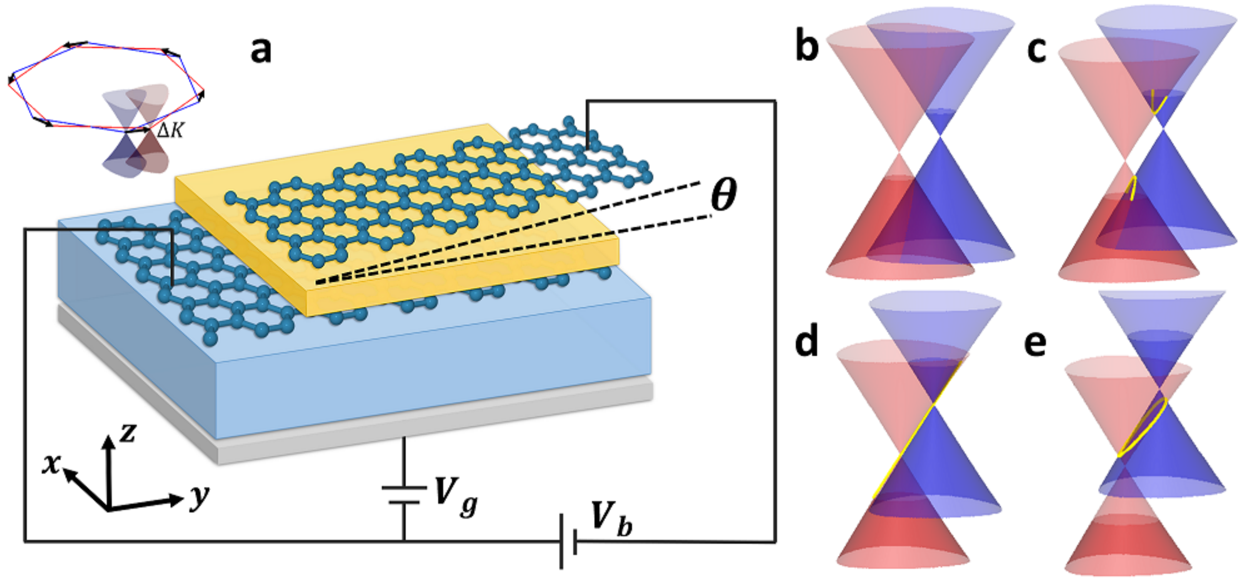

Figure 6.7: (a) A schematic view of the twisted heterostructure device. An external gate electrode is applied on bottom graphene sheet. The top graphene layer is rotated with hBN insulator by an exaggerated angle θ. Inset: The Brillouin zones for bottom and top graphene layers in momentum space. The neutrality points of different graphene layers are displaced by ΔK. (b)-(e): The horizontal distance between neutrality points is determined by the rotation angle θ and the vertical distance between them are determined by the applied gate voltage. (b) depicts the situation of $V_b = V_b^R$. (c)-(e) correspond to situations of $V_b < V_b^P$, $V_b = V_b^P$ and $V_b > V_b^P$. The red and blue cones represent the energy dispersions of bottom and top graphene layers respectively. Occupied and unoccupied states are distinguished by different transparency. The transmissive states that can carry tunnel current is highlighted by yellow curves.

of Fabry-Pérot like interference. In the case of perfect lattice alignment, the transmission states lie on a circular curve with wavevectors at the same energy. When the energy of these states satisfies the resonant condition for Fabry-Pérot interference, all states along the circular curve are capable of carrying current. However, the angular misorientation between graphene layers creates a displacement between two Dirac cones in momentum space. As a result of the conic intersection, the transmission states lie on a hyperbolic [Figure 6.7(c)] or elliptic [Figure 6.7(e)] curve without sharing the same energy. Therefore, the number of transmissive states that can tunnel resonantly with the assistance of Fabry-Pérot like interference are greatly suppressed, leading to the damping of current oscillations.

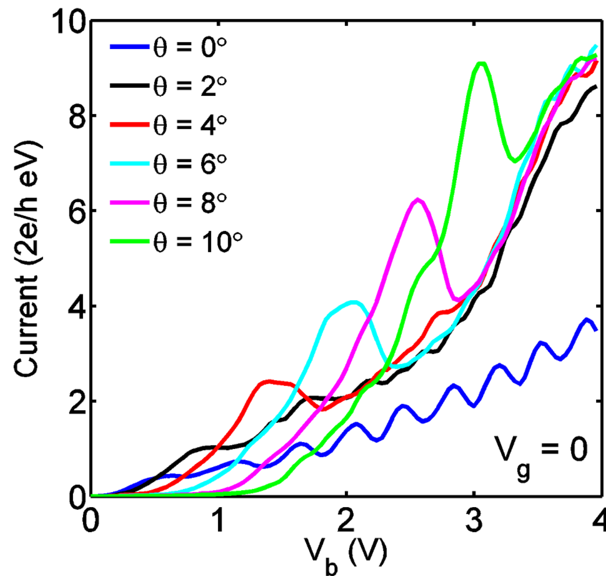

Figure 6.8: Calculated I-V curves for a family of twisting angle without external gate electrode.

At non-zero θ, the current is close to zero at small biases in Figure 6.8, and the current rapidly grows after a particular bias voltage V_b^R. We explain this feature by depicting the conic dispersions of the two graphene layers at $V_b = V_b^R$ in Figure 6.7(b). When $V_b < V_b^R$, although the Dirac cones intersect along a hyperbolic curve, all transmissive states are occupied in both top and bottom graphene layers, yielding a zero tunneling current. At

$V_b = V_b^R$, the occupied / unoccupied states of the bottom (red) / top (blue) Dirac cones intersect only at two points shown in Figure 6.7(b) (also shown in Figure 6.9 bottom inset). When $V_b > V_b^R$, a fraction of the states in the hyperbolic intersection is unoccupied at top layer [Figure 6.7(c)], resulting in a rapid increase of current.

To explain how V_b^R changes as a function of θ, we provide a formula for V_b^R by solving equations 6.1-6.2 under the situation displayed in Figure 6.9, that is $\Delta\varphi_b+\mu_T+\mu_T \sim \hbar v_F \Delta K$, where $\Delta K = \frac{4\pi}{3a}\theta$ we obtain that at small θ ($V_g = 0$):

$$V_b^R = \frac{\hbar v_F}{e}\Delta K = \frac{4\pi}{3a}\frac{\hbar v_F}{e}\theta \tag{6.3}$$

Next, as the source-drain bias becomes larger, strong NDR peaks occur at V_b^P, when the intersection between two Dirac cones becomes a straight line [86**?**]. V_b^P as a function of twisting angle can also be evaluated by solving equations 6.1-6.2 corresponding to Figure 6.7(d) (see illustration in Figure 6.9 inset), and using $\Delta\varphi_b = \hbar v_F \Delta K$, the value of V_b^P can be expressed as a function of θ:

$$V_b^P = \frac{\hbar v_F}{e}\Delta K + \frac{(\hbar v_F)^{3/2}}{e^2}\sqrt{\frac{\pi\epsilon_0\epsilon_{BN}}{d_{BN}}} \left(\sqrt{\Delta K} + \sqrt{\Delta K + \frac{d_{BN}}{d_{OX}}\left(\frac{\hbar v_F \pi\epsilon_0\epsilon_{BN}}{4e^2 d_{OX}} - \frac{eV_g}{\hbar v_F}\right)}\right) - \frac{(\hbar v_F)^2\pi\epsilon_0\epsilon_{BN}}{2e^3 d_{OX}} \tag{6.4}$$

The estimations of V_b^R and V_b^P are compared against the simulated values in Figure 6.9, indicating a quantitative match between the analytical formula and numerical results.

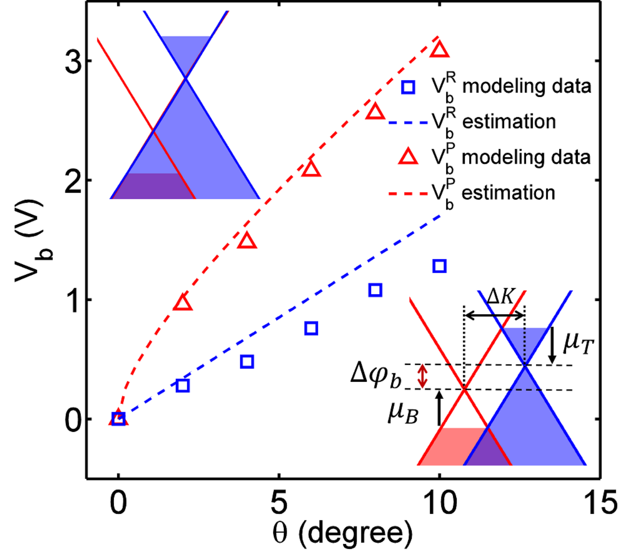

Figure 6.9: V_b^R and V_b^P as a function of θ for both simulated results (markers) and analytical (dashed) estimations. Top inset: illustration of the situation at V_b^P, a 2D version of Figure 6.7(d). Bottom inset: illustration of the situation at V_b^R where tunneling current begins to increase rapidly from zero, a 2D version of Figure 6.7(b). The electrostatic model is defined in equations 6.1-6.2.

6.2.4 Gate Controllability

For a twisted heterostructure with fixed angle $\theta = 4°$, we model the current-voltage characteristics with various values of V_g in Figure 6.10. Pronounced resonant peaks whose locations and amplitudes vary as a function of gate voltage are seen as the gate electrode modulates the electrostatic potentials by changing the carrier concentration in graphene layers. As a result, the gate electrode alters the energy difference between the Dirac points on the two sheets, thereby shifting the value of V_b^P. Our calculated dependence on V_g is qualitatively consistent with experimental results in reference [90].

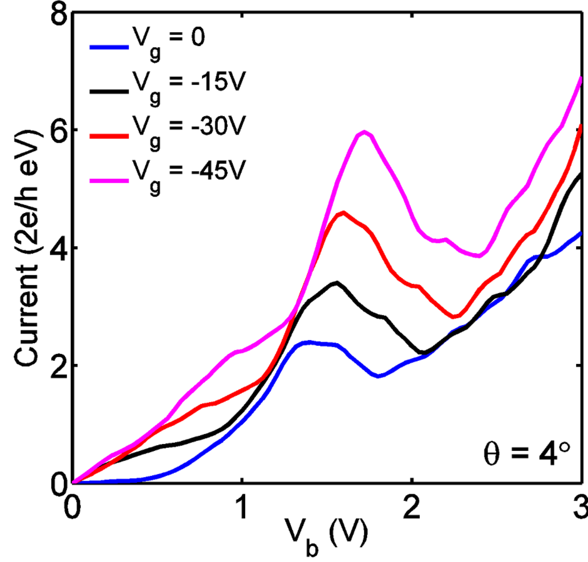

Figure 6.10: Calculated I-V curves for a fixed non-zero twisting angle $\theta = 4^{\circ}$ with various gate voltages.

6.2.5 Impact of Phonon Scattering

According to reference [90] when the environmental temperature increases from 2K to room temperature, the measured peak-to-valley ratio (PVR) values are reduced by 10% - 15%. We have verified that when decoherence is absent, the difference between I-V curves at low and high temperatures is negligible, indicating that thermal smearing is not responsible for PVR reduction observed in experiments.

To better interpret the experimental measurements, we include the electron-phonon scattering in top and bottom AGNR within the NEGF framework. Depending on different mechanisms, scattering can be elastic (acoustic phonon) and inelastic (optical phonon): $\mathbf{\Sigma}_{ph}^{r,<,>} = \mathbf{\Sigma}_{el}^{r,<,>} + \mathbf{\Sigma}_{inel}^{r,<,>}$. Following the Born approximation [67], the scattering self-energies can be solved self-consistently with Green's function:

$$\mathbf{\Sigma}_{el}^{r,<,>} = D_{el}\mathbf{G}^{r,<,>} \tag{6.5}$$

$$\begin{aligned}\mathbf{\Sigma}_{inel}^{<} = D_{inel}\{&\left[n_B\left(\hbar\omega\right)+1\right]\mathbf{G}^{<}\left(\varepsilon+\hbar\omega\right)\\ &+n_B\left(\hbar\omega\right)\mathbf{G}^{<}\left(\varepsilon-\hbar\omega\right)\}\end{aligned} \tag{6.6}$$

$$\begin{aligned}\mathbf{\Sigma}_{inel}^{>} = D_{inel}\{&\left[n_B\left(\hbar\omega\right)+1\right]\mathbf{G}^{>}\left(\varepsilon-\hbar\omega\right)\\ &+n_B\left(\hbar\omega\right)\mathbf{G}^{>}\left(\varepsilon+\hbar\omega\right)\}\end{aligned} \tag{6.7}$$

$$\mathrm{Im}\left[\mathbf{\Sigma}_{ph}^{r}\right] = \left[\mathbf{\Sigma}_{ph}^{>} - \mathbf{\Sigma}_{ph}^{<}\right]/2i \tag{6.8}$$

Here, n_B is the Boltzmann distribution; $\hbar\omega$ is the phonon energy; and D_{el} and D_{inel} are the electron-phonon deformation potentials. This model characterizes the scattering of electrons by three parameters, which are determined to satisfy the experimentally measurable electron mean free path in graphene: $D_{el} = 0.01\text{eV}^2$, $D_{inel} = 0.07\text{eV}^2$ and $\hbar\omega = 180\text{meV}$ [1]. Phenomenologically, larger deformational potentials reflect stronger electron-phonon scattering, thus shorter electron mean free path. The mean free path obtained from our calculations is about $1.42\mu m$, consistent with the measured mean free path of graphene deposited on hBN substrate (around $1.5\mu m$).

Oriented Case

We first examine the case without misorientation to show how the decoherence impacts the two resonant mechanisms we studied in the first part of this chapter. The current-voltage curves are plotted in Figure 6.11. Apparently, electron-phonon scattering suppresses both NDR mechanisms and therefore, PVR values of these NDR peaks are reduced. However, the suppression of gate-induced NDR peak is not as substantial as the interference induced ones, since the quantum interference is more vulnerable to decoherence introduced by electron-phonon scattering. This might explain the absence of NDR peaks due to quantum interference in the experiments of reference [8]. However, in an experiment with sufficiently smaller devices, both mechanisms leading to the multiple NDR peaks can occur.

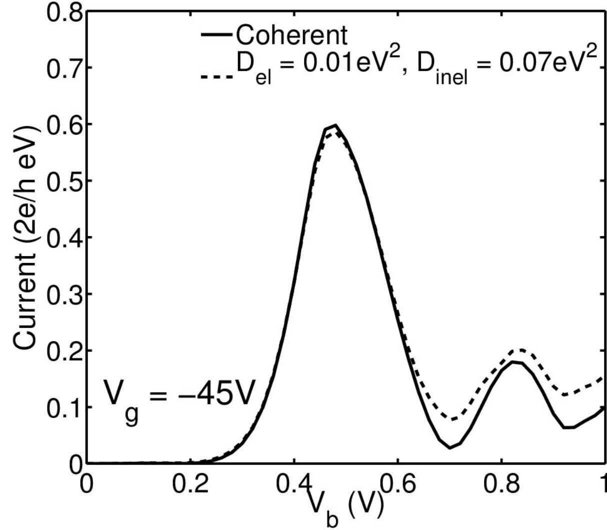

Figure 6.11: Calculated I-V curves at $V_g = -45V$ for devices ($\theta = 0$) with $N_z = 1$, $N_x = 62$ and $N_y = 64$ with consideration of electron-phonon scattering.

Misoriented Case

The simulation results for misoriented devices with electron-phonon scattering are plotted in Figure 6.12 (dashed lines). For the twisted heterostructures, the phonon-mediated current as a function of drain voltage preserves the NDR features. The PVR decreases compared to the case of coherent tunneling, whereas the magnitude of both peak and valley current is larger. When electron-phonon scattering exists, the conservation of wavevectors required for the tunneling of electrons between the two layers is weakened, resulting in the rise of tunneling current.

The reduction of PVR values observed in experiment is clearly captured in our simulation. In Figure 6.12 inset, we plot the PVR values of the current peaks as a function of rotation angle in the coherent case and with phonon-scattering, where a 15% - 25% reduction of PVR values is observed. Therefore, the modeled results are in a reasonable agreement with the observations in experiments, demonstrating that the suppression of NDR features introduced by higher temperature is mainly due to a stronger decoherence mechanism including electron-

phonon scattering.

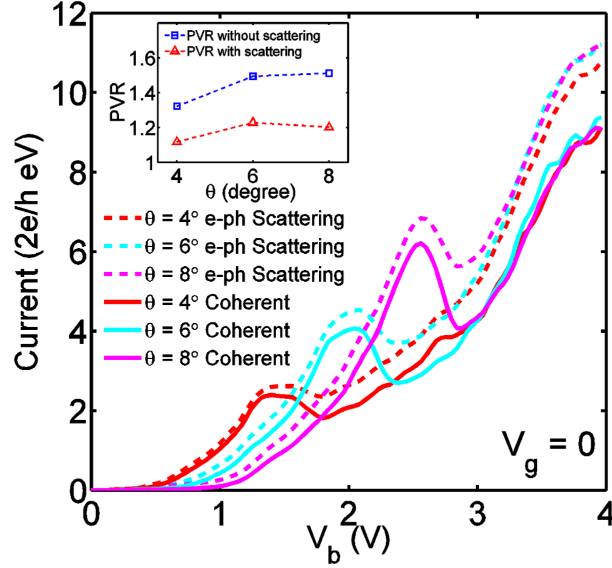

Figure 6.12: Calculated I-V curves for a family of θ at $V_g = 0$ with electron-phonon scattering (dashed curves). For comparison, the corresponding results of coherent transport (from Figure 6.8) are plotted as solid curves.

6.2.6 Summary

In the second part of the realistic application, we model the electron transport properties of a three-terminal tunnel-FET device built with twisted graphene layers sandwiching hBN barrier. Robust NDR features in I-V characteristics are captured by the numerical simulation and distinct mechanisms are responsible for the resonant tunneling. The Fabry-Pérot like quantum interference vanishes at larger twisting angles. NDR peaks arising in the case of twisted graphene layers are controllable by both gate voltage and twisting angle. Analytical equations for V_b^R and V_b^P are derived. Moreover, the role of phonon induced decoherence is also numerically simulated to capture the effects of temperature increase in experiments. In the case of twisted graphene sheet, the NDR survives electron-phonon scattering but the peak-to-valley ratios are slightly reduced, consistent with experimental works.

Chapter 7

CONCLUSION

In this Ph.D. work, a numerical approach to calculate physical quantities (such as density of states and charge density) in nanoscale devices, within the context of the non-equilibrium Green's function approach is presented, namely HSC-extension. This work exploits recent advances to use an established graph partitioning method (nested dissection). This contribution does not require any processing of the partition and it can handle open boundary conditions, represented by full self-energy matrices. The key ingredients are an efficient sparse block LDL^T-factorization and an appropriate order of operations to preserve the sparsity as much as possible. The resulting algorithm was illustrated on a quantum well superlattice and a graphene nanoribbon, which are represented by a continuum and tight binding Hamiltonian respectively, and demonstrated a significant speed up over the recursive method RGF.

In order to extend the HSC-extension approach to 3D scenarios, a variety of numerical experiments are carried out. While greatly depending on the system complexity, HSC-extension preserves significant speed up over RGF for structures like graphene based multilayer heterostructures. Whereas for silicon nanowire and DNA molecure electronic devices, the acceleration is suppressed.

Next, we apply our numerical modeling approach on a large scale realistic simulation of a graphene-hBN-graphene multilayer heterostructure device, which exhibits multiple negative differential resistance peaks. We rationalize three distinct underlying mechanisms, which are sensitively controllable by gate bias and angular misorientation between graphene and hBN layers. In addition, the electron-phonon scattering decoherence calculation is incorporated into the NEGF solver (with HSC-extension) in a self-consistent manner. The consideration of electron-phonon scattering contributes in explanation of the NDR peaks degradation which

is observed in the experiments performed in room temperature.

Appendix A

DERIVATION OF HSC-EXTENTION (SIMPLE CASE)

With the partition shown in Figure 2.1, the matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{LL} & \mathbf{0} & \mathbf{A}_{LS} \\ \mathbf{0} & \mathbf{A}_{RR} & \mathbf{A}_{RS} \\ \mathbf{A}_{LS}^T & \mathbf{A}_{RS}^T & \mathbf{A}_{SS} \end{bmatrix}$$

Note that matrix $\mathbf{A}$ is typically complex symmetric. The block LDL^T-factorization of $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{LS}^T\mathbf{A}_{LL}^{-1} & \mathbf{A}_{RS}^T\mathbf{A}_{RR}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{LL} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{RR} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\mathbf{A}}_{SS} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{A}_{LL}^{-1}\mathbf{A}_{LS} \\ \mathbf{0} & \mathbf{I} & \mathbf{A}_{RR}^{-1}\mathbf{A}_{RS} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}$$

where $\widehat{\mathbf{A}}_{SS}$ is the Schur complement,

$$\widehat{\mathbf{A}}_{SS} = \mathbf{A}_{SS} - \mathbf{A}_{LS}^T\mathbf{A}_{LL}^{-1}\mathbf{A}_{LS} - \mathbf{A}_{RS}^T\mathbf{A}_{RR}^{-1}\mathbf{A}_{RS}.$$

The matrix $\mathbf{G}^r$ satisfies the relation

$$\mathbf{G}^r = \left(\mathbf{I} - \mathbf{L}^T\right)\mathbf{G}^r + \mathbf{D}^{-1}\mathbf{L}^{-1} \quad \text{with} \quad \mathbf{A} = \mathbf{LDL}^T \tag{A.1}$$

(described in Takahashi et al. [69] and Erisman and Tinney [19]). The block notation yields

$$\mathbf{G}^r = -\begin{bmatrix} \mathbf{A}_{LL}^{-1}\mathbf{A}_{LS}\mathbf{G}_{SL}^r & \mathbf{A}_{LL}^{-1}\mathbf{A}_{LS}\mathbf{G}_{SR}^r & \mathbf{A}_{LL}^{-1}\mathbf{A}_{LS}\mathbf{G}_{SS}^r \\ \mathbf{A}_{RR}^{-1}\mathbf{A}_{RS}\mathbf{G}_{SL}^r & \mathbf{A}_{RR}^{-1}\mathbf{A}_{RS}\mathbf{G}_{SR}^r & \mathbf{A}_{RR}^{-1}\mathbf{A}_{RS}\mathbf{G}_{SS}^r \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$
$$+ \begin{bmatrix} \mathbf{A}_{LL}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{RR}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\mathbf{A}}_{SS}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{LS}^T\mathbf{A}_{LL}^{-1} & -\mathbf{A}_{RS}^T\mathbf{A}_{RR}^{-1} & \mathbf{I} \end{bmatrix}.$$

This equation indicates

$$\mathbf{G}^r_{SS} = \left(\widehat{\mathbf{A}}_{SS}\right)^{-1},$$
$$\mathbf{G}^r_{LS} = -\mathbf{A}^{-1}_{LL}\mathbf{A}_{LS}\mathbf{G}^r_{SS},$$
$$\mathbf{G}^r_{RS} = -\mathbf{A}^{-1}_{RR}\mathbf{A}_{RS}\mathbf{G}^r_{SS}.$$

The diagonal blocks $\mathbf{G}^r_{LL}$ and $\mathbf{G}^r_{RR}$, for the regions L and R, respectively, are computed independently of each other,

$$\mathbf{G}^r_{LL} = \mathbf{A}^{-1}_{LL} - \mathbf{A}^{-1}_{LL}\mathbf{A}_{LS}\left(\mathbf{G}^r_{LS}\right)^T = \mathbf{A}^{-1}_{LL} + \mathbf{A}^{-1}_{LL}\mathbf{A}_{LS}\mathbf{G}^r_{SS}\mathbf{A}^T_{LS}\mathbf{A}^{-1}_{LL}$$
$$\mathbf{G}^r_{RR} = \mathbf{A}^{-1}_{RR} - \mathbf{A}^{-1}_{RR}\mathbf{A}_{RS}\left(\mathbf{G}^r_{RS}\right)^T = \mathbf{A}^{-1}_{RR} + \mathbf{A}^{-1}_{RR}\mathbf{A}_{RS}\mathbf{G}^r_{SS}\mathbf{A}^T_{RS}\mathbf{A}^{-1}_{RR}$$

(where the symmetry of $\mathbf{G}^r$ has been exploited). For this simple case, the resulting algorithm matches exactly the HSC method [46].

For calculating entries in the correlation matrix $\mathbf{G}^<$, The block LDL^T-factorization of $\mathbf{A}$ yields

$$\mathbf{G}^< = -\begin{bmatrix} \mathbf{A}^{-1}_{LL}\mathbf{A}_{LS}\mathbf{G}^<_{SL} & \mathbf{A}^{-1}_{LL}\mathbf{A}_{LS}\mathbf{G}^<_{SR} & \mathbf{A}^{-1}_{LL}\mathbf{A}_{LS}\mathbf{G}^<_{SS} \\ \mathbf{A}^{-1}_{RR}\mathbf{A}_{RS}\mathbf{G}^<_{SL} & \mathbf{A}^{-1}_{RR}\mathbf{A}_{RS}\mathbf{G}^<_{SR} & \mathbf{A}^{-1}_{RR}\mathbf{A}_{RS}\mathbf{G}^<_{SS} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$
$$+ \begin{bmatrix} \mathbf{A}^{-1}_{LL} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}_{RR} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \left(\widehat{\mathbf{A}}_{SS}\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{A}^T_{LS}\mathbf{A}^{-1}_{LL} & -\mathbf{A}^T_{RS}\mathbf{A}^{-1}_{RR} & \mathbf{I} \end{bmatrix} \boldsymbol{\Sigma}^< \left(\mathbf{G}^r\right)^\dagger.$$

Parts of $\mathbf{G}^r$ are computed with the previous algorithm, namely, $\mathbf{G}^r_{LL}$, $\mathbf{G}^r_{RR}$, $\mathbf{G}^r_{SS}$, $\mathbf{G}^r_{RS}$, and $\mathbf{G}^r_{LS}$. By assumption, $\boldsymbol{\Sigma}^<$ is a block-diagonal skew-Hermitian matrix with purely imaginary

entries. The partial matrix multiplication gives

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{LS}^T\mathbf{A}_{LL}^{-1} & -\mathbf{A}_{RS}^T\mathbf{A}_{RR}^{-1} & \mathbf{I} \end{bmatrix} \mathbf{\Sigma}^< (\mathbf{G}^r)^\dagger =$$

$$\begin{bmatrix} \mathbf{\Sigma}_{LL}^< (\mathbf{G}_{LL}^r)^\dagger & * & \mathbf{\Sigma}_{LL}^< (\mathbf{G}_{SL}^r)^\dagger \\ * & \mathbf{\Sigma}_{RR}^< (\mathbf{G}_{RR}^r)^\dagger & \mathbf{\Sigma}_{RR}^< (\mathbf{G}_{RS}^r)^\dagger \\ * & * & \mathbf{\Sigma}_{SS}^< (\mathbf{G}_{SS}^r)^\dagger - \mathbf{A}_{LS}^T\mathbf{A}_{LL}^{-1}\mathbf{\Sigma}_{LL}^< (\mathbf{G}_{SL}^r)^\dagger - \mathbf{A}_{RS}^T\mathbf{A}_{RR}^{-1}\mathbf{\Sigma}_{RR}^< (\mathbf{G}_{SR}^r)^\dagger \end{bmatrix}$$

(where the starred blocks are not computed). This relation indicates

$$\begin{aligned} \mathbf{G}_{SS}^< &= \mathbf{G}_{SS}^r \left(\mathbf{\Sigma}_{SS}^< (\mathbf{G}_{SS}^r)^\dagger - \mathbf{A}_{LS}^T\mathbf{A}_{LL}^{-1}\mathbf{\Sigma}_{LL}^< (\mathbf{G}_{SL}^r)^\dagger - \mathbf{A}_{RS}^T\mathbf{A}_{RR}^{-1}\mathbf{\Sigma}_{RR}^< (\mathbf{G}_{SR}^r)^\dagger \right), \\ \mathbf{G}_{LS}^< &= -\mathbf{A}_{LL}^{-1}\mathbf{A}_{LS}\mathbf{G}_{SS}^< + \mathbf{A}_{LL}^{-1}\mathbf{\Sigma}_{LL}^< (\mathbf{G}_{SL}^r)^\dagger, \\ \mathbf{G}_{RS}^< &= -\mathbf{A}_{RR}^{-1}\mathbf{A}_{RS}\mathbf{G}_{SS}^< + \mathbf{A}_{RR}^{-1}\mathbf{\Sigma}_{RR}^< (\mathbf{G}_{SR}^r)^\dagger. \end{aligned}$$

Finally, the diagonal blocks $\mathbf{G}_{LL}^<$ and $\mathbf{G}_{RR}^<$, for the regions L and R, respectively, are computed independently of each other,

$$\begin{aligned} \mathbf{G}_{LL}^< &= \mathbf{A}_{LL}^{-1}\mathbf{\Sigma}_{LL}^< (\mathbf{G}_{LL}^r)^\dagger - \mathbf{A}_{LL}^{-1}\mathbf{A}_{LS} (\mathbf{G}_{LS}^<)^\dagger \\ \mathbf{G}_{RR}^< &= \mathbf{A}_{RR}^{-1}\mathbf{\Sigma}_{RR}^< (\mathbf{G}_{RR}^r)^\dagger - \mathbf{A}_{RR}^{-1}\mathbf{A}_{RS} (\mathbf{G}_{RS}^<)^\dagger \end{aligned}$$

(where the skew-Hermitian property of $\mathbf{G}^<$ has been exploited).

Appendix B

DESCRIPTION OF THE ALGORITHM FOR A THREE-LEVEL TREE

In order to make the extension more comprehensive, a description of the HSC extension is given for a three-level system (see Figure B.1).

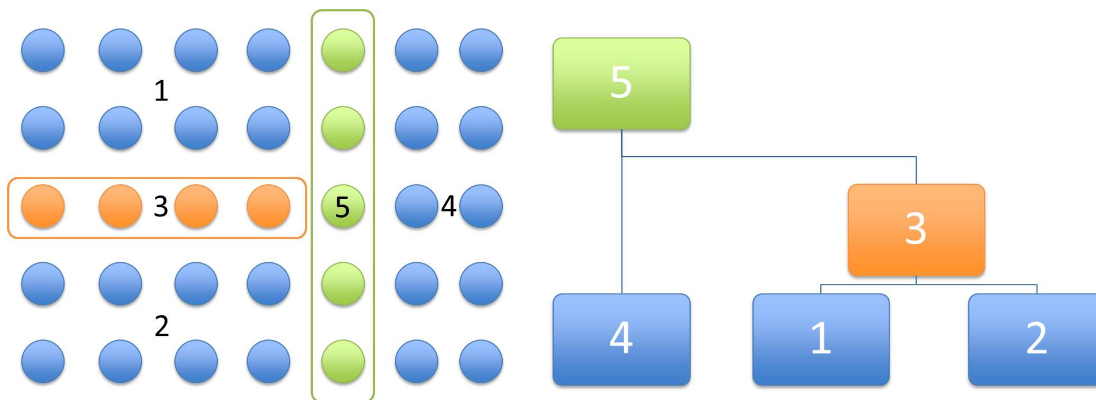

Figure B.1: Partition for a three-level system.

The size of the partition is chosen to make the description as relevant as possible without becoming overcomplicated. The first level contains regions 1, 2, and 4. The second-level refers to region 3 and the top level is the root or region 5. To illustrate the algorithm, steps from Section 2.2 are described for this particular device.

When a five-point stencil is used for discretization, the structure of the matrix $\mathbf{A}$, after

re-ordering, is

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{A}_{13} & \mathbf{0} & \mathbf{A}_{15} \\ \mathbf{0} & \mathbf{A}_{22} & \mathbf{A}_{23} & \mathbf{0} & \mathbf{A}_{25} \\ \mathbf{A}_{13}^T & \mathbf{A}_{23}^T & \mathbf{A}_{33} & \mathbf{0} & \mathbf{A}_{35} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{44} & \mathbf{A}_{45} \\ \mathbf{A}_{15}^T & \mathbf{A}_{25}^T & \mathbf{A}_{35}^T & \mathbf{A}_{45}^T & \mathbf{A}_{55} \end{bmatrix}$$

The blocks $\mathbf{A}_{11}$ and $\mathbf{A}_{44}$ are dense for representing the contacts with the semi-infinite leads.

Recall that $\mathbf{A}^{(0)}$ is equal to $\mathbf{A}$. Then inner points in regions 1, 2, and 4 are eliminated by block Gaussian elimination — the effects of the inner points in regions 1, 2, and 4 are folded over their boundary. This first step yields the matrix $\mathbf{A}^{(1)}$

$$\mathbf{A}^{(1)} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{33}^{(1)} & \mathbf{0} & \mathbf{A}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{44} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \left(\mathbf{A}_{35}^{(1)}\right)^T & \mathbf{0} & \mathbf{A}_{55}^{(1)} \end{bmatrix}$$

where the updated matrices are

$$\begin{aligned} \mathbf{A}_{33}^{(1)} &= \mathbf{A}_{33} - \mathbf{A}_{13}^T\mathbf{A}_{11}^{-1}\mathbf{A}_{13} - \mathbf{A}_{23}^T\mathbf{A}_{22}^{-1}\mathbf{A}_{23} \\ \mathbf{A}_{35}^{(1)} &= \mathbf{A}_{35} - \mathbf{A}_{13}^T\mathbf{A}_{11}^{-1}\mathbf{A}_{15} - \mathbf{A}_{23}^T\mathbf{A}_{22}^{-1}\mathbf{A}_{25} \\ \mathbf{A}_{55}^{(1)} &= \mathbf{A}_{55} - \mathbf{A}_{15}^T\mathbf{A}_{11}^{-1}\mathbf{A}_{15} - \mathbf{A}_{25}^T\mathbf{A}_{22}^{-1}\mathbf{A}_{25} - \mathbf{A}_{45}^T\mathbf{A}_{44}^{-1}\mathbf{A}_{45} \end{aligned}$$

After folding the effects of regions 1 and 2, block $\mathbf{A}_{35}^{(1)}$ is now a dense block. Figure B.2 illustrates the change of sparsity between $\mathbf{A}$ and $\mathbf{A}^{(1)}$.

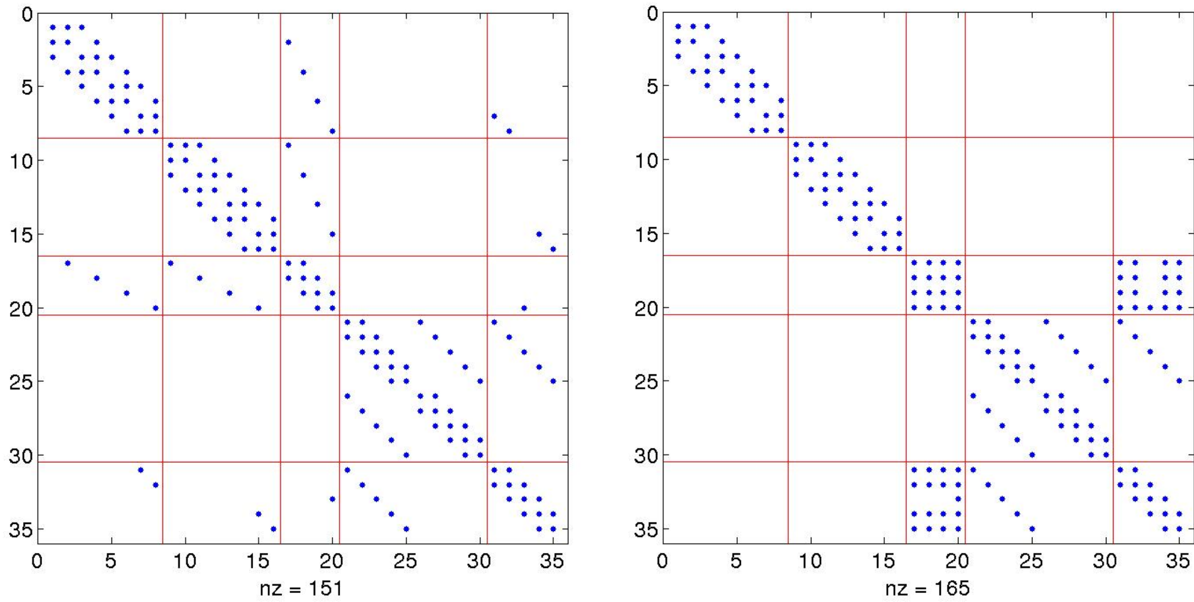

Figure B.2: Sparsity of matrix $\mathbf{A}$ and matrix $\mathbf{A}^{(1)}$.

In the next step, the remaining off-diagonal blocks are eliminated to obtain the matrix $\mathbf{A}^{(2)}$,

$$\mathbf{A}^{(2)} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{33}^{(1)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{44} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{55}^{(2)} \end{bmatrix}$$

where the block $\mathbf{A}_{55}^{(2)}$ is

$$\mathbf{A}_{55}^{(2)} = \mathbf{A}_{55}^{(1)} - \left(\mathbf{A}_{35}^{(1)}\right)^T \left(\mathbf{A}_{33}^{(1)}\right)^{-1} \mathbf{A}_{35}^{(1)}.$$

The next step is written as the inversion of $\mathbf{A}^{(2)}$, which is symmetric and block diagonal,

$$\mathbf{G}^{(2)} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \left(\mathbf{A}_{33}^{(1)}\right)^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{44}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \left(\mathbf{A}_{55}^{(2)}\right)^{-1} \end{bmatrix}.$$

This operation requires only the inversion of the block $\mathbf{A}_{55}^{(2)}$. All the other blocks have been inverted during the folding steps.

Next diagonal blocks of $\mathbf{G}^r$ are extracted one level at a time. Starting from the main root (or separator), blocks at level 2 are updated to obtain

$$\mathbf{G}^{(1)} = \begin{bmatrix} \mathbf{G}_{11}^{(2)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{22}^{(2)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{33}^{(1)} & \mathbf{0} & \mathbf{G}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{G}_{44}^{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \left(\mathbf{G}_{35}^{(1)}\right)^{T} & \mathbf{0} & \mathbf{G}_{55}^{(2)} \end{bmatrix}$$

with

$$\mathbf{G}_{35}^{(1)} = -\left(\mathbf{A}_{33}^{(1)}\right)^{-1} \mathbf{A}_{35}^{(1)} \mathbf{G}_{55}^{(2)} = \mathbf{\Psi}_{35} \mathbf{G}_{55}^{(2)}$$

$$\mathbf{G}_{33}^{(1)} = \mathbf{G}_{33}^{(2)} - \left(\mathbf{A}_{33}^{(1)}\right)^{-1} \mathbf{A}_{35}^{(1)} \left(\mathbf{G}_{35}^{(1)}\right)^{T} = \mathbf{G}_{33}^{(2)} + \mathbf{\Psi}_{35} \left(\mathbf{G}_{35}^{(1)}\right)^{T}$$

Finally, blocks for regions 1, 2, and 4 are updated, yielding the matrix $\mathbf{G}^{(0)}$,

$$\mathbf{G}^{(0)} = \begin{bmatrix} \mathbf{G}_{11}^{(0)} & \mathbf{0} & \mathbf{G}_{13}^{(0)} & \mathbf{0} & \mathbf{G}_{15}^{(0)} \\ \mathbf{0} & \mathbf{G}_{22}^{(0)} & \mathbf{G}_{23}^{(0)} & \mathbf{0} & \mathbf{G}_{25}^{(0)} \\ \left(\mathbf{G}_{13}^{(0)}\right)^{T} & \left(\mathbf{G}_{23}^{(0)}\right)^{T} & \mathbf{G}_{33}^{(1)} & \mathbf{0} & \mathbf{G}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{G}_{44}^{(0)} & \mathbf{G}_{45}^{(0)} \\ \left(\mathbf{G}_{15}^{(0)}\right)^{T} & \left(\mathbf{G}_{25}^{(0)}\right)^{T} & \left(\mathbf{G}_{35}^{(1)}\right)^{T} & \left(\mathbf{G}_{45}^{(0)}\right)^{T} & \mathbf{G}_{55}^{(2)} \end{bmatrix}.$$

Blocks for region 4 are satisfying

$$\begin{aligned}
\mathbf{G}_{45}^{(0)} &= -\left(\mathbf{A}_{44}^{(0)}\right)^{-1}\mathbf{A}_{45}^{(0)}\mathbf{G}_{55}^{(2)} = \mathbf{\Psi}_{45}\mathbf{G}_{55}^{(2)} \\
\mathbf{G}_{44}^{(0)} &= \mathbf{G}_{44}^{(2)} - \left(\mathbf{A}_{44}^{(0)}\right)^{-1}\mathbf{A}_{45}^{(0)}\left(\mathbf{G}_{45}^{(0)}\right)^{T}
\end{aligned}$$

Blocks for region 1 are defined by

$$\begin{aligned}
\mathbf{G}_{15}^{(0)} &= -\left(\mathbf{A}_{11}^{(0)}\right)^{-1}\mathbf{A}_{15}^{(0)}\mathbf{G}_{55}^{(2)} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1}\mathbf{A}_{13}^{(0)}\mathbf{G}_{35}^{(1)} \\
\mathbf{G}_{13}^{(0)} &= -\left(\mathbf{A}_{11}^{(0)}\right)^{-1}\mathbf{A}_{13}^{(0)}\mathbf{G}_{33}^{(1)} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1}\mathbf{A}_{15}^{(0)}\left(\mathbf{G}_{53}^{(1)}\right)^{T} \\
\mathbf{G}_{11}^{(0)} &= \left(\mathbf{A}_{11}^{(0)}\right)^{-1} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1}\mathbf{A}_{13}^{(0)}\left(\mathbf{G}_{13}^{(0)}\right)^{T} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1}\mathbf{A}_{15}^{(0)}\left(\mathbf{G}_{15}^{(0)}\right)^{T}
\end{aligned}$$

and blocks for region 2

$$\begin{aligned}
\mathbf{G}_{25}^{(0)} &= -\left(\mathbf{A}_{22}^{(0)}\right)^{-1}\mathbf{A}_{25}^{(0)}\mathbf{G}_{55}^{(2)} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1}\mathbf{A}_{23}^{(0)}\mathbf{G}_{35}^{(1)} \\
\mathbf{G}_{23}^{(0)} &= -\left(\mathbf{A}_{22}^{(0)}\right)^{-1}\mathbf{A}_{23}^{(0)}\mathbf{G}_{33}^{(1)} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1}\mathbf{A}_{25}^{(0)}\left(\mathbf{G}_{53}^{(1)}\right)^{T} \\
\mathbf{G}_{22}^{(0)} &= \left(\mathbf{A}_{22}^{(0)}\right)^{-1} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1}\mathbf{A}_{23}^{(0)}\left(\mathbf{G}_{23}^{(0)}\right)^{T} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1}\mathbf{A}_{25}^{(0)}\left(\mathbf{G}_{25}^{(0)}\right)^{T}
\end{aligned}$$

Figure B.3 displays the sparsity of the resulting matrix $\mathbf{G}^{(0)}$.

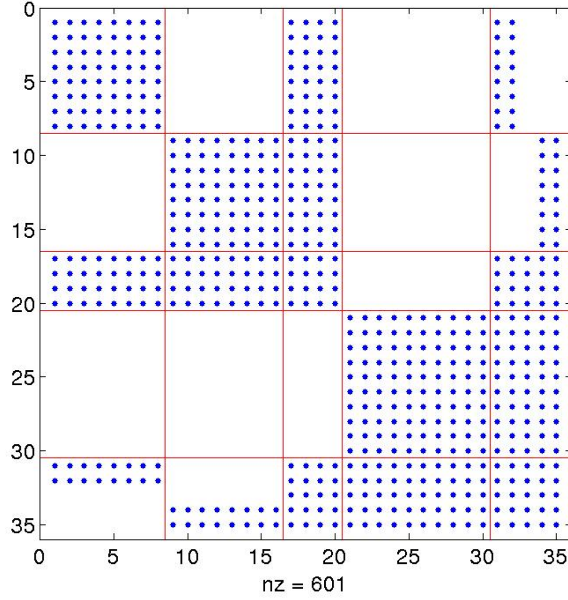

Figure B.3: Sparsity of matrix $\mathbf{G}^{(0)}$.

All the entries in $\mathbf{G}^{(0)}$ are equal to their corresponding entries in $\mathbf{G}^r$.

The computation of diagonal blocks in $\mathbf{G}^<$ are described for the same device. First the matrix $\mathbf{N}$ is computed,

$$\mathbf{N} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^< \overline{\mathbf{G}_{11}^{(0)}} & \mathbf{0} & \boldsymbol{\Sigma}_{11}^< \overline{\mathbf{G}_{13}^{(0)}} & \mathbf{0} & \boldsymbol{\Sigma}_{11}^< \overline{\mathbf{G}_{15}^{(0)}} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^< \overline{\mathbf{G}_{22}^{(0)}} & \boldsymbol{\Sigma}_{22}^< \overline{\mathbf{G}_{23}^{(0)}} & \mathbf{0} & \boldsymbol{\Sigma}_{22}^< \overline{\mathbf{G}_{25}^{(0)}} \\ \boldsymbol{\Sigma}_{33}^< \left(\mathbf{G}_{13}^{(0)}\right)^\dagger & \boldsymbol{\Sigma}_{33}^< \left(\mathbf{G}_{23}^{(0)}\right)^\dagger & \boldsymbol{\Sigma}_{33}^< \overline{\mathbf{G}_{33}^{(0)}} & \mathbf{0} & \boldsymbol{\Sigma}_{33}^< \overline{\mathbf{G}_{35}^{(0)}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{44}^< \overline{\mathbf{G}_{44}^{(0)}} & \boldsymbol{\Sigma}_{44}^< \overline{\mathbf{G}_{45}^{(0)}} \\ \boldsymbol{\Sigma}_{55}^< \left(\mathbf{G}_{15}^{(0)}\right)^\dagger & \boldsymbol{\Sigma}_{55}^< \left(\mathbf{G}_{25}^{(0)}\right)^\dagger & \boldsymbol{\Sigma}_{55}^< \left(\mathbf{G}_{35}^{(0)}\right)^\dagger & \boldsymbol{\Sigma}_{55}^< \left(\mathbf{G}_{45}^{(0)}\right)^\dagger & \boldsymbol{\Sigma}_{55}^< \overline{\mathbf{G}_{55}^{(0)}} \end{bmatrix}.$$

Set $\mathbf{N}^{(0)} = \mathbf{N}$. Next the lower level clusters are folded into the higher ones to obtain the

matrix $\mathbf{N}^{(1)}$

$$\mathbf{N}^{(1)} = \begin{bmatrix} \mathbf{N}_{11}^{(0)} & \mathbf{0} & \mathbf{N}_{13}^{(0)} & \mathbf{0} & \mathbf{N}_{15}^{(0)} \\ \mathbf{0} & \mathbf{N}_{22}^{(0)} & \mathbf{N}_{23}^{(0)} & \mathbf{0} & \mathbf{N}_{25}^{(0)} \\ \mathbf{N}_{31}^{(0)} & \mathbf{N}_{32}^{(0)} & \mathbf{N}_{33}^{(1)} & \mathbf{0} & \mathbf{N}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N}_{44}^{(0)} & \mathbf{N}_{45}^{(0)} \\ \mathbf{N}_{51}^{(0)} & \mathbf{N}_{52}^{(0)} & \mathbf{N}_{53}^{(1)} & \mathbf{N}_{54}^{(0)} & \mathbf{N}_{55}^{(1)} \end{bmatrix}$$

where the updated blocks are

$$\begin{aligned}
\mathbf{N}_{33}^{(1)} &= \mathbf{N}_{33}^{(0)} - (\mathbf{A}_{13})^T \mathbf{A}_{11}^{-1}\mathbf{N}_{13}^{(0)} - (\mathbf{A}_{23})^T \mathbf{A}_{22}^{-1}\mathbf{N}_{23}^{(0)} \\
\mathbf{N}_{55}^{(1)} &= \mathbf{N}_{55}^{(0)} - (\mathbf{A}_{15})^T \mathbf{A}_{11}^{-1}\mathbf{N}_{15}^{(0)} - (\mathbf{A}_{25})^T \mathbf{A}_{22}^{-1}\mathbf{N}_{25}^{(0)} - (\mathbf{A}_{45})^T \mathbf{A}_{44}^{-1}\mathbf{N}_{45}^{(0)} \\
\mathbf{N}_{35}^{(1)} &= \mathbf{N}_{35}^{(0)} - (\mathbf{A}_{13})^T \mathbf{A}_{11}^{-1}\mathbf{N}_{15}^{(0)} - (\mathbf{A}_{23})^T \mathbf{A}_{22}^{-1}\mathbf{N}_{25}^{(0)} \\
\mathbf{N}_{53}^{(1)} &= \mathbf{N}_{53}^{(0)} - (\mathbf{A}_{15})^T \mathbf{A}_{11}^{-1}\mathbf{N}_{13}^{(0)} - (\mathbf{A}_{25})^T \mathbf{A}_{22}^{-1}\mathbf{N}_{23}^{(0)}
\end{aligned}$$

For the top level, the block for region 5 is updated

$$\mathbf{N}^{(2)} = \begin{bmatrix} \mathbf{N}_{11}^{(0)} & \mathbf{0} & \mathbf{N}_{13}^{(0)} & \mathbf{0} & \mathbf{N}_{15}^{(0)} \\ \mathbf{0} & \mathbf{N}_{22}^{(0)} & \mathbf{N}_{23}^{(0)} & \mathbf{0} & \mathbf{N}_{25}^{(0)} \\ \mathbf{N}_{31}^{(0)} & \mathbf{N}_{32}^{(0)} & \mathbf{N}_{33}^{(1)} & \mathbf{0} & \mathbf{N}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{N}_{44}^{(0)} & \mathbf{N}_{45}^{(0)} \\ \mathbf{N}_{51}^{(0)} & \mathbf{N}_{52}^{(0)} & \mathbf{N}_{53}^{(1)} & \mathbf{N}_{54}^{(0)} & \mathbf{N}_{55}^{(2)} \end{bmatrix}$$

with

$$\mathbf{N}_{55}^{(2)} = \mathbf{N}_{55}^{(1)} - \left(\mathbf{A}_{35}^{(1)}\right)^T \left(\mathbf{A}_{33}^{(1)}\right)^{-1} \mathbf{N}_{35}^{(1)}.$$

The next step is a block-diagonal multiplication

$$\mathbf{P}^{(2)} = \begin{bmatrix} \mathbf{A}_{11}^{-1}\mathbf{N}_{11}^{(0)} & \mathbf{0} & \mathbf{A}_{11}^{-1}\mathbf{N}_{13}^{(0)} & \mathbf{0} & \mathbf{A}_{11}^{-1}\mathbf{N}_{15}^{(0)} \\ \mathbf{0} & \mathbf{A}_{22}^{-1}\mathbf{N}_{22}^{(0)} & \mathbf{A}_{22}^{-1}\mathbf{N}_{23}^{(0)} & \mathbf{0} & \mathbf{A}_{22}^{-1}\mathbf{N}_{25}^{(0)} \\ \left(\mathbf{A}_{33}^{(1)}\right)^{-1}\mathbf{N}_{31}^{(0)} & \left(\mathbf{A}_{33}^{(1)}\right)^{-1}\mathbf{N}_{32}^{(0)} & \left(\mathbf{A}_{33}^{(1)}\right)^{-1}\mathbf{N}_{33}^{(1)} & \mathbf{0} & \left(\mathbf{A}_{33}^{(1)}\right)^{-1}\mathbf{N}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{44}^{-1}\mathbf{N}_{44}^{(0)} & \mathbf{A}_{44}^{-1}\mathbf{N}_{45}^{(0)} \\ \left(\mathbf{A}_{55}^{(2)}\right)^{-1}\mathbf{N}_{51}^{(0)} & \left(\mathbf{A}_{55}^{(2)}\right)^{-1}\mathbf{N}_{52}^{(0)} & \left(\mathbf{A}_{55}^{(2)}\right)^{-1}\mathbf{N}_{53}^{(1)} & \left(\mathbf{A}_{55}^{(2)}\right)^{-1}\mathbf{N}_{54}^{(0)} & \left(\mathbf{A}_{55}^{(2)}\right)^{-1}\mathbf{N}_{55}^{(2)} \end{bmatrix}.$$

Finally, Step 4 extracts blocks one level at a time, defining first

$$\mathbf{P}^{(1)} = \begin{bmatrix} \mathbf{P}_{11}^{(2)} & \mathbf{0} & \mathbf{P}_{13}^{(2)} & \mathbf{0} & \mathbf{P}_{15}^{(2)} \\ \mathbf{0} & \mathbf{P}_{22}^{(2)} & \mathbf{P}_{23}^{(2)} & \mathbf{0} & \mathbf{P}_{25}^{(2)} \\ \mathbf{P}_{31}^{(2)} & \mathbf{P}_{32}^{(2)} & \mathbf{P}_{33}^{(1)} & \mathbf{0} & \mathbf{P}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{P}_{44}^{(2)} & \mathbf{P}_{45}^{(2)} \\ \mathbf{P}_{51}^{(2)} & \mathbf{P}_{52}^{(2)} & \mathbf{P}_{53}^{(1)} & \mathbf{P}_{54}^{(2)} & \mathbf{P}_{55}^{(2)} \end{bmatrix}$$

where the updated blocks are

$$\begin{aligned} \mathbf{P}_{35}^{(1)} &= \mathbf{P}_{35}^{(2)} - \left(\mathbf{A}_{33}^{(1)}\right)^{-1} \mathbf{A}_{35}^{(1)} \mathbf{P}_{55}^{(2)} \\ \mathbf{P}_{53}^{(1)} &= -\left(\mathbf{P}_{35}^{(1)}\right)^{\dagger} \\ \mathbf{P}_{33}^{(1)} &= \mathbf{P}_{33}^{(2)} - \left(\mathbf{A}_{33}^{(1)}\right)^{-1} \mathbf{A}_{35}^{(1)} \mathbf{P}_{53}^{(1)} \end{aligned}$$

Finally, blocks for regions 1, 2, and 4 are updated, yielding the matrix $\mathbf{P}^{(0)}$,

$$\mathbf{P}^{(0)} = \begin{bmatrix} \mathbf{P}_{11}^{(0)} & \mathbf{0} & \mathbf{P}_{13}^{(0)} & \mathbf{0} & \mathbf{P}_{15}^{(0)} \\ \mathbf{0} & \mathbf{P}_{22}^{(0)} & \mathbf{P}_{23}^{(0)} & \mathbf{0} & \mathbf{P}_{25}^{(0)} \\ -\left(\mathbf{P}_{13}^{(0)}\right)^{\dagger} & -\left(\mathbf{P}_{23}^{(0)}\right)^{\dagger} & \mathbf{P}_{33}^{(1)} & \mathbf{0} & \mathbf{P}_{35}^{(1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{P}_{44}^{(0)} & \mathbf{P}_{45}^{(0)} \\ -\left(\mathbf{P}_{15}^{(0)}\right)^{\dagger} & -\left(\mathbf{P}_{25}^{(0)}\right)^{\dagger} & -\left(\mathbf{P}_{35}^{(1)}\right)^{\dagger} & -\left(\mathbf{P}_{45}^{(0)}\right)^{\dagger} & \mathbf{P}_{55}^{(2)} \end{bmatrix}.$$

Blocks for region 4 are satisfying

$$\begin{aligned} \mathbf{P}_{45}^{(0)} &= \mathbf{P}_{45}^{(1)} - \left(\mathbf{A}_{44}^{(0)}\right)^{-1} \mathbf{A}_{45}^{(0)} \mathbf{P}_{55}^{(2)} \\ \mathbf{P}_{44}^{(0)} &= \mathbf{P}_{44}^{(1)} + \left(\mathbf{A}_{44}^{(0)}\right)^{-1} \mathbf{A}_{45}^{(0)} \left(\mathbf{P}_{45}^{(0)}\right)^{\dagger} \end{aligned}$$

Blocks for region 1 are defined by

$$\begin{aligned} \mathbf{P}_{15}^{(0)} &= \mathbf{P}_{15}^{(1)} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1} \mathbf{A}_{15}^{(0)} \mathbf{P}_{55}^{(2)} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1} \mathbf{A}_{13}^{(0)} \mathbf{P}_{35}^{(1)} \\ \mathbf{P}_{13}^{(0)} &= \mathbf{P}_{13}^{(0)} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1} \mathbf{A}_{13}^{(0)} \mathbf{P}_{33}^{(1)} - \left(\mathbf{A}_{11}^{(0)}\right)^{-1} \mathbf{A}_{15}^{(0)} \left(\mathbf{P}_{53}^{(1)}\right)^{T} \\ \mathbf{P}_{11}^{(0)} &= \mathbf{P}_{11}^{(0)} + \left(\mathbf{A}_{11}^{(0)}\right)^{-1} \mathbf{A}_{13}^{(0)} \left(\mathbf{P}_{13}^{(0)}\right)^{\dagger} + \left(\mathbf{A}_{11}^{(0)}\right)^{-1} \mathbf{A}_{15}^{(0)} \left(\mathbf{P}_{15}^{(0)}\right)^{\dagger} \end{aligned}$$

and blocks for region 2

$$\mathbf{P}_{25}^{(0)} = \mathbf{P}_{25}^{(1)} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1} \mathbf{A}_{25}^{(0)}\mathbf{P}_{55}^{(2)} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1} \mathbf{A}_{23}^{(0)}\mathbf{P}_{35}^{(1)}$$

$$\mathbf{P}_{23}^{(0)} = \mathbf{P}_{23}^{(0)} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1} \mathbf{A}_{23}^{(0)}\mathbf{P}_{33}^{(1)} - \left(\mathbf{A}_{22}^{(0)}\right)^{-1} \mathbf{A}_{25}^{(0)} \left(\mathbf{P}_{53}^{(1)}\right)^{T}$$

$$\mathbf{P}_{22}^{(0)} = \mathbf{P}_{22}^{(0)} + \left(\mathbf{A}_{22}^{(0)}\right)^{-1} \mathbf{A}_{23}^{(0)} \left(\mathbf{P}_{23}^{(0)}\right)^{\dagger} + \left(\mathbf{A}_{22}^{(0)}\right)^{-1} \mathbf{A}_{25}^{(0)} \left(\mathbf{P}_{25}^{(0)}\right)^{\dagger}$$

All the entries in $\mathbf{P}^{(0)}$ are equal to their corresponding entries in $\mathbf{G}^{<}$.

Appendix C

COMPLEXITY DERIVATION OF HSC-EXTENSION FOR 3D CUBOIDAL STRUCTURES

To analyze the runtime complexity of HSC-extension, we consider a cuboid device with $N_x \times N_y \times N_z$ grid points per direction (Figure C.1 illustrates a partition for a toy-model cuboid device of size $(2a+1) \times (2a+1) \times (2a+1)$ and the corresponding binary tree[1].

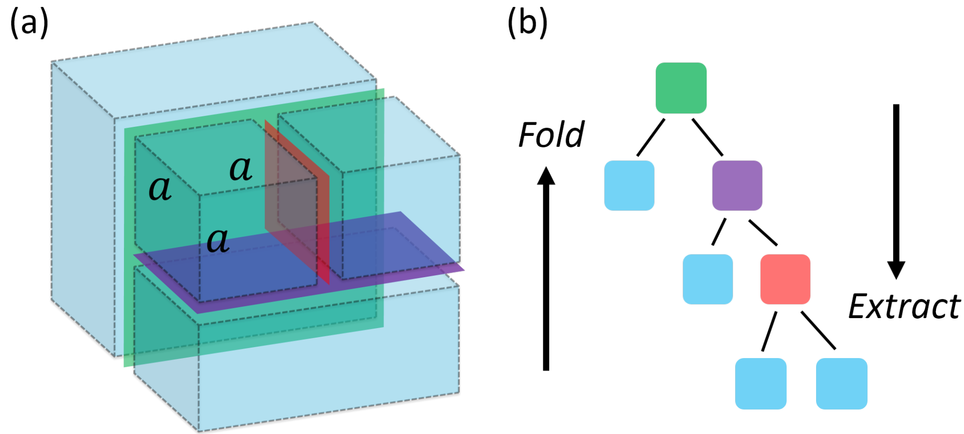

Figure C.1: (a) The domain decomposition from cube of dimension $2a + 1$ to cubes of dimension a. Three levels of separators are colored by red, purple and green respectively. (b) The multilevel binary tree corresponding to the cuboid decomposition. The three levels of separators are depicted with matching colors. The blue blocks denote the corresponding blue clusters.

First we discuss the case of a cubic mesh, *i.e.* $N_x = N_y = N_z = N$. The operation count for evaluating diagonal entries in $\mathbf{G}^r$ was discussed by Lin *et al.* [46, section 2.5.3]. According to their analysis, the operation count grows as $\mathcal{O}(N^6)$. For the HSC-extension,

[1] In practice, the binary tree is likely to be balanced.

the complexity for evaluating diagonal entries of $\mathbf{G}^{<}$ is identical to the complexity for $\mathbf{G}^{r}$ (as discussed in Hetmaniuk *et al.* [32] for two-dimensional devices). The overall operation count will grow as $\mathcal{O}(N^6)$.

Next we consider the case of an elongated device, where the numbers of grid points per direction satisfy $N_x = N_z = N \ll N_y$. The multilevel nested-dissection will identify N_y/N_z subdomains, each discretized with $N \times N \times N$ grid points. When evaluating diagonal entries in $\mathbf{G}^r$, the operation count for each cubic subdomains will grow as $\mathcal{O}(N^6)$. The remaining operations will involve dense matrices for the separators of dimension $N_x \times N_z$. The algebraic operations for one separator will include dense matrix-matrix multiplications and dense matrix inversions, yielding an asymptotic cost $\mathcal{O}(N_x^3 N_z^3) = \mathcal{O}(N^6)$. The number of separators is also $\mathcal{O}(N_y/N)$. The overall operation count will grow as $\mathcal{O}(N^6)\mathcal{O}(N_y/N) = \mathcal{O}(N^5 N_y)$.

Finally we consider the case of a flattened device, where the number of grid points per direction satisfy $N_z \ll N_x = N_y = N$. As discussed by Lin *et al.* [46, section 2.5.3], this configuration is similar to a two-dimensional problem with $N \times N$ grid points. The prefactor will depend on N_z. The HSC algorithm and our extension can proceed as if the device is two-dimensional by replacing scalar algebraic operations with block algebraic operations (each block being of dimension $N_z \times N_z$). These block operations will cost $\mathcal{O}(N_z^3)$. So the overall operation count will grow as $\mathcal{O}(N_z^3)\mathcal{O}(N^3) = \mathcal{O}(N_z^3 N^3)$.

BIBLIOGRAPHY

[1] N. D. Akhavan, G. Jolley, G. A. Umana Membreno, J. Antoszewski, and L. Faraone. Phonon limited transport in graphene nanoribbon field effect transistors using full three dimensional quantum mechanical simulation. *J. Appl. Phys.*, 112(9):094505, 2012.

[2] A. Alarcón, V-H Nguyen, S. Berrada, D. Querlioz, J. Saint-Martin, A. Bournel, and P. Dollfus. Pseudosaturation and negative differential conductance in graphene field-effect transistors. *IEEE T. Electron Dev.*, 60:985–991, 2013.

[3] F. Amet, J. R. Williams, A. G. F. Garcia, M. Yankowitz, K. Watanabe, T. Taniguchi, and D. Goldhaber-Gordon. Tunneling spectroscopy of graphene-boron-nitride heterostructures. *Phys. Rev. B*, 85(7):073405, 2012.

[4] M. Anantram, M. Lundstrom, and D. Nikonov. Modeling of nanoscale devices. *Proc. IEEE*, 96(9):1511–1550, 2008.

[5] M. P. Anantram and S. Datta. Effect of phase breaking on the AC-response of mesoscopic systems. *Phys. Rev. B*, 51(12):7632, 1995.

[6] M. P. Anantram and A. Svizhenko. Multidimensional modeling of nanotransistors. *IEEE E Electron. Dev.*, 54(9):2100–2115, 2007.

[7] M. Bescond, N. Cavassilas, K. Kalna, K. Nehari, L. Raymond, J. L. Autran, M. Lannoo, and A. Asenov. Ballistic transport in Si, Ge, and GaAs nanowire MOSFETs. *IEDM Tech. Dig*, 533, 2005.

[8] L. Britnell, R. V. Gorbachev, A. K. Geim, L. A. Ponomarenko, A. Mishchenko, M. T. Greenaway, T. M. Fromhold, K. S. Novoselov, and L. Eaves. Resonant tunnelling and negative differential conductance in graphene transistors. *Nat. comm.*, 4:1794, 2013.

[9] L. Britnell, R. V. Gorbachev, R. Jalil, B. D. Belle, F. Schedin, M. I. Katsnelson, L. Eaves, S. V. Morozov, A. S. Mayorov, N. M. R. Peres, et al. Electron tunneling through ultrathin boron nitride crystalline barriers. *Nano Lett.*, 12(3):1707–1710, 2012.

[10] S. Bruzzone, G. Fiori, and G. Iannaccone. Tunneling properties of vertical heterostructures of multilayer hexagonal boron nitride and graphene. *arXiv preprint arXiv:1212.4629*, 2012.

[11] A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim. The electronic properties of graphene. *Rev. Mod. Phys.*, 81(1):109–162, 2009.

[12] J. Cumings and A. Zettl. Resistance of telescoping nanotubes. *AIP Conf. Proc.*, 633(1):227–230, 2002.

[13] S. Datta. The non-equilibrium Green's function (NEGF) formalism: An elementary introduction. In *Electron Devices Meeting, 2002. IEDM'02. International*, pages 703–706. IEEE, 2002.

[14] S. Datta. Nanoscale device modeling: the Green's function method. *Superlattice. Microst.*, 28:253–278, 2000.

[15] C. Dean, A. F. Young, L. Wang, I. Meric, G. H. Lee, K. Watanabe, T. Taniguchi, K. Shepard, P. Kim, and J. Hone. Graphene based heterostructures. *Solid State Commun.*, 152(15):1275–1282, 2012.

[16] C. R. Dean, A. F. Young, I. Meric, C. Lee, L. Wang, S. Sorgenfrei, K. Watanabe, T. Taniguchi, P. Kim, and K. L. Shepard. Boron nitride substrates for high-quality graphene electronics. *Nat. Nanotechnol.*, 5(10):722–726, 2010.

[17] R. Decker, Y. Wang, V. W. Brar, W. Regan, H.-Z. Tsai, Q. Wu, W. Gannett, A. Zettl, and M. F. Crommie. Local electronic properties of graphene on a BN substrate via scanning tunneling microscopy. *Nano Lett.*, 11(6):2291–2295, 2011.

[18] V. N. Do and P. Dollfus. Negative differential resistance in zigzag-edge graphene nanoribbon junctions. *J. Appl. Phys.*, 107(6):063705, 2010.

[19] A. M. Erisman and W. F. Tinney. On computing certain elements of the inverse of a sparse matrix. *Commun. ACM*, 18(3):177–179, March 1975.

[20] L. Esaki. New phenomenon in narrow germanium p-n junctions. *Phys. Rev.*, 109(2):603–604, 1958.

[21] R. M. Feenstra, D. Jena, and G. Gu. Single-particle tunneling in doped graphene-insulator-graphene junctions. *J. Appl. Phys.*, 111(4):043711, 2012.

[22] G. J. Ferreira, M. N. Leuenberger, D. Loss, and J. C. Egues. Low-bias negative differential resistance in graphene nanoribbon superlattices. *Phys. Rev. B*, 84(12):125453, 2011.

[23] G. Fiori. Negative differential resistance in mono and bilayer graphene pn junctions. *IEEE Electr. Device L.*, 32(10):1334–1336, 2011.

[24] K.-J. Gan, C.-S. Tsai, and D.-S. Liang. Novel multiple-selected and multiple-valued memory design using negative differential resistance circuits suitable for standard sige-based bicmos process. *Analog Integr. Circ. S.*, 59(2):161–167, 2009.

[25] A. George. Nested dissection of a regular finite element mesh. *SIAM J. Numer. Anal.*, 10(2):345–363, 1973.

[26] G. Giovannetti, P. A. Khomyakov, G. Brocks, P. J. Kelly, and J. van den Brink. Substrate-induced band gap in graphene on hexagonal boron nitride: Ab initio density functional calculations. *Phys. Rev. B*, 76(7):073103, 2007.

[27] D. C. Guhr, D. Rettinger, J. Boneberg, A. Erbe, P. Leiderer, and E. Scheer. Influence of laser light on electronic transport through atomic-size contacts. *Phys. Rev. Lett.*, 99:086801, Aug 2007.

[28] K. M. M. Habib, F. Zahid, and R. K. Lake. Negative differential resistance in bilayer graphene nanoribbons. *Appl. Phys. Lett.*, 98(19):192112, 2011.

[29] R. Haydock. *Solid state physics*, volume 35. Academic Press, 1980.

[30] R. Haydock, V. Heine, and M. Kelly. Electronic structure based on the local atomic environment for tight-binding bands. *J. Phys. C: Solid State Phys.*, 5(20):2845, 1972.

[31] R. Haydock and C. Nex. A general terminator for the recursion method. *J. Phys. C: Solid State Phys.*, 18(11):2235, 1985.

[32] U. Hetmaniuk, Y. Zhao, and M. P. Anantram. A nested dissection approach to modeling transport in nanodevices: Algorithms and applications. *Int. J. Numer. Meth. Eng.*, 95(7):587–607, 2013.

[33] V. Hung Nguyen, F. Mazzamuto, J. Saint-Martin, A. Bournel, and P. Dollfus. Giant effect of negative differential conductance in graphene nanoribbon p-n hetero-junctions. *Appl. Phys. Lett.*, 99(4):042105, 2011.

[34] A.-P. Jauho, N. S. WinGreen, and Y. Meir. Time-dependent transport in interacting and noninteracting resonant-tunneling systems. *Phys. Rev. B*, 50(8):5528, 1994.

[35] K. Jeong Won and J. Qing. Electrostatically telescoping nanotube nonvolatile memory device. *Nanotechnology*, 18(9):095705, 2007.

[36] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, 1998.

[37] N. Kharche and S. K. Nayak. Quasiparticle band gap engineering of graphene and graphone on hexagonal boron nitride substrate. *Nano Lett.*, 11(12):5274–5278, 2011.

[38] S. A. Khorasani. Tunable spontaneous emission from layered graphene/dielectric tunnel junctions. *IEEE J. Quantum Elect.*, 50(5):307–313, 2014.

[39] S. B. Kumar, G. Seol, and J. Guo. Modeling of a vertical tunneling graphene heterojunction field-effect transistor. *Appl. Phys. Lett.*, 101(3):033503, 2012.

[40] R. Lake, G. Klimeck, R. Bowen, and D. Jovanovic. Single and multi-band modeling of quantum electron transport through layered semiconductor devices. *J. Appl. Phys.*, 81:7845–7869, 1997.

[41] K.-T. Lam and G. Liang. An ab initio study on energy gap of bilayer graphene nanoribbons with armchair edges. *Appl. Phys. Lett.*, 92(22):223106, 2008.

[42] S. Li, S. Ahmed, G. Klimeck, and E. Darve. Computing entries of the inverse of a sparse matrix using the FIND algorithm. *J. Comp. Phys.*, 227:9408–9427, 2008.

[43] S. Li and E. Darve. Extension and optimization of the FIND algorithm: Computing Green's and less-than Green's functions. *J. Comp. Phys.*, 231(4):1121 – 1139, 2012.

[44] S Li, W Wu, and E Darve. A fast algorithm for sparse matrix computations related to inversion. *J. Comp. Phys.*, 242:915–945, 2013.

[45] D.-S. Liang, K.-J. Gan, L.-X. Su, C. Chen, C. Hsiao, C. Tsai, Y. Chen, S. Wang, S. Kuo, and F. Chiang. Four-valued memory circuit designed by multiple-peak MOS-NDR devices and circuits. In *Fifth International Workshop on System-on-Chip for Real-Time Applications*, pages 78–81. IEEE.

[46] L. Lin, J. Lu, L. Ying, R. Car, and W. E. Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure analysis of metallic systems. *Commun. Math. Sci.*, 7(3):755–777, 2009.

[47] L. Lin, C. Yang, J. Meza, J. Lu, L. Ying, and W. E. SelInv – An algorithm for selected inversion of a sparse symmetric matrix. *ACM Trans. Math. Softw.*, 37(40), 2011.

[48] D. Mamaluy, M. Sabathil, and P. Vogl. Efficient method for the calculation of ballistic quantum transport. *J. Appl. Phys.*, 93(8):4628–4633, 2003.

[49] D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl. Contact block reduction method for ballistic transport and carrier densities of open nanostructures. *Phys. Rev. B*, 71(24):245321, 2005.

[50] A. Martinez, M. Bescond, J. R. Barker, A. Svizhenko, M. P. Anantram, C. Millar, and A. Asenov. A self-consistent full 3-D real-space NEGF simulator for studying nonperturbative effects in nano-MOSFETs. *IEEE E Electron. Dev.*, 54(9):2213–2222, 2007.

[51] A. Martinez, A. R. Brown, A. Asenov, and N. Seoane. A comparison between a fully-3D real-space versus coupled mode-space NEGF in the study of variability in gate-all-around Si nanowire MOSFET. In *Simulation of Semiconductor Processes and Devices, 2009. SISPAD'09. International Conference on*, pages 1–4. IEEE, 2009.

[52] H. Mizuta and T. Tanoue. *The physics and applications of resonant tunnelling diodes*, volume 2. Cambridge University Press, 2006.

[53] V. H. Nguyen, F. Mazzamuto, A. Bournel, and P. Dollfus. Resonant tunnelling diodes based on graphene/h-bn heterostructure. *J. Phys. D Appl. Phys.*, 45(32):325104, 2012.

[54] V. H. Nguyen, Y. M. Niquet, and P. Dollfus. Gate-controllable negative differential conductance in graphene tunneling transistors. *Semicond. Sci. Tech.*, 27(10):105018, 2012.

[55] M. Nobuya, Ed. Takuya, K. Yoshinari, and E. Laurence. Nonequilibrium Green's function simulations of graphene-nanoribbon resonant-tunneling transistors. *JPN J. Appl. Phys.*, 53(4S):04EN04, 2014.

[56] D. Petersen, S. Li, K. Stokbro, H. Sorensen, P. Hansen, S. Skelboe, and E. Darve. A hybrid method for the parallel computation of Green's functions. *J. Comp. Phys.*, 228:5020–5039, 2009.

[57] H. Ren, Q. Li, Y. Luo, and J. Yang. Graphene nanoribbon as a negative differential resistance device. *Appl. Phys. Lett.*, 94(17):173110, 2009.

[58] R. M. Ribeiro and N. M. R. Peres. Stability of boron nitride bilayers: Ground-state energies, interlayer distances, and tight-binding description. *Phys. Rev. B*, 83(23):235312, 2011.

[59] M. P. Lopez Sancho, J. M. L. Sancho, and J. Rubio. Quick iterative scheme for the calculation of transfer matrices: application to Mo (100). *J. Phys. F: Met. Phys.*, 14(5):1205, 1984.

[60] M. P. L. Sancho, J. M. Lopez Sancho, J. M. L. Sancho, and J. Rubio. Highly convergent schemes for the calculation of bulk and surface Green's functions. *J. Phys. F: Met. Phys.*, 15(4):851, 1985.

[61] F. Schwierz. Graphene transistors. *Nat. Nanotechnol.*, 5:487–496, 2010.

[62] B. Sensale-Rodriguez. Graphene-insulator-graphene active plasmonic terahertz devices. *Appl. Phys. Lett.*, 103(12):123109, 2013.

[63] F. Sols, M. Macucci, U. Ravaioli, and K. Hess. Theory for a quantum modulated transistor. *J. Appl. Phys.*, 66(8):3892–3906, 1989.

[64] Y.-W. Son, M. L. Cohen, and S. G. Louie. Energy gaps in graphene nanoribbons. *Phys. Rev. Lett.*, 97(21):216803, 2006.

[65] S. K. Sundaram and E. Mazur. Inducing and probing non-thermal transitions in semiconductors using femtosecond laser pulses. *Nat. Mat.*, 1(4):217–224, 2002.

[66] A. Svizhenko, M. Anantram, T. Govindam, B. Biegel, and R. Venugopal. Two-dimensional quantum mechanical modeling of nanotransistors. *J. Appl. Phys.*, 91:2343, 2002.

[67] A. Svizhenko and M. P. Anantram. Effect of scattering and contacts on current and electrostatics in carbon nanotubes. *Phys. Rev. B*, 72(8):085430, 2005.

[68] J. Sławińska, I. Zasada, and Z. Klusek. Energy gap tuning in graphene on hexagonal boron nitride bilayer system. *Phys. Rev. B*, 81(15):155433, 2010.

[69] K. Takahashi, J. Fagan, and M. S. Chin. Formation of a sparse bus impedance matrix and its application to short circuit study. In *Eighth PICA Conference*, 1973.

[70] F. T. Vasko. Resonant and nondissipative tunneling in independently contacted graphene structures. *Phys. Rev. B*, 87:075424, 2013.

[71] B. Wang, J. Wang, and H. Guo. Current partition: A nonequilibrium Green's function approach. *Phys. Rev. Lett.*, 82(2):398, 1999.

[72] H. Wang, T. Taychatanapat, A. Hsu, K. Watanabe, T. Taniguchi, P. Jarillo-Herrero, and T. Palacios. BN/graphene/BN transistors for RF applications. *arXiv preprint arXiv:1108.2021*, 2011.

[73] Y. Wang, C.-Y. Yam, T. Frauenheim, G. H. Chen, and T. A. Niehaus. An efficient method for quantum transport simulations in the time domain. *Chem. Phys.*, 391(1):69–77, 2011.

[74] Z. F. Wang, Qunxiang Li, Q. W. Shi, Xiaoping Wang, Jinlong Yang, J. G. Hou, and Jie Chen. Chiral selective tunneling induced negative differential resistance in zigzag graphene nanoribbon: A theoretical study. *Appl. Phys. Lett.*, 92(13):133114, 2008.

[75] Y. Wu, D. B. Farmer, W. Zhu, S. Han, C. D. Dimitrakopoulos, A. A. Bol, P. Avouris, and Y. Lin. Three-terminal graphene negative differential resistance devices. *ACS Nano*, 6(3):2610–2616, 2012.

[76] Y. Xu, Z. Guo, H. Chen, Y. Yuan, J. Lou, X. Lin, H. Gao, H. Chen, and B. Yu.

In-plane and tunneling pressure sensors based on graphene/hexagonal boron nitride heterostructures. *Appl. Phys. Lett.*, 99(13):133109, 2011.

[77] J. Xue, J. Sanchez-Yamagishi, D. Bulmash, P. Jacquod, A. Deshpande, K. Watanabe, T. Taniguchi, P. Jarillo-Herrero, and B. J. LeRoy. Scanning tunnelling microscopy and spectroscopy of ultra-flat graphene on hexagonal boron nitride. *Nat. Mater.*, 10(4):282–285, 2011.

[78] C. Yam, X. Zheng, G. Chen, Y. Wang, T. Frauenheim, and T. A. Niehaus. Time-dependent versus static quantum transport simulations beyond linear response. *Phys. Rev. B*, 83(24):245448, 2011.

[79] T. Yamamoto, K. Sasaoka, and S. Watanabe. Universal transition between inductive and capacitive admittance of metallic single-walled carbon nanotubes. *Phys. Rev. B*, 82(20):205404, 2010.

[80] Y. Yu, B. Wang, and Y. Wei. ac response of a carbon chain under a finite frequency bias. *J. Chem. Phys.*, 127(10):104701, 2007.

[81] X. Zheng, F. Wang, C. Y. Yam, Y. Mo, and G. Chen. Time-dependent density-functional theory for open systems. *Phys. Rev. B*, 75(19):195127, 2007.

[82] X. Zhong, R. G. Amorim, R. H. Scheicher, R. Pandey, and S. P. Karna. Electronic structure and quantum transport properties of trilayers formed from graphene and boron nitride. *Nanoscale*, 4(17):5490–5498, 2012.

[83] L. Brey. Coherent tunneling and negative differential conductivity in a graphene/h-BN/graphene heterostructure. *Phys. Rev. A*, 2(1):014003, 2014.

[84] L. Britnell, R. M. Ribeiro, A. Eckmann, R. Jalil, B. D. Belle, A. Mishchenko, Y.-J. Kim, R. V. Gorbachev, T. Georgiou, S. V. Morozov, A. N. Grigorenko, A. K. Geim, C. Casiraghi, A. H. Castro Neto, and K. S. Novoselov. Strong light-matter interactions in heterostructures of atomically thin films. *Science*, 340(6138):1311–1314, 2013.

[85] S. C. de la Barrera, Q. Gao, and R. M. Feenstra. Theory of grapheneinsulatorgraphene tunnel junctions. *J. Vac. Sci. Technol. B.*, 32(4):04E101, 2014.

[86] J. Gaskell, L. Eaves, K. S. Novoselov, A. Mishchenko, A. K. Geim, T. M. Fromhold, and M. T. Greenaway. Graphene-hexagonal boron nitride resonant tunneling diodes as high-frequency oscillators. *Appl. Phys. Lett.*, 107(10):103105, 2015.

[87] T. Georgiou, R. Jalil, B. D. Belle, L. Britnell, R. V. Gorbachev, S. V. Morozov, Y. J. Kim, A. Gholinia, S. J. Haigh, O. Makarovsky, L. Eaves, L. A. Ponomarenko, A. K. Geim, K. S. Novoselov, and A. Mishchenko. Vertical field-effect transistor based on graphene-WS_2 heterostructures for flexible and transparent electronics. *Nat Nano*, 8(2):100–103, 2013.

[88] M. T. Greenaway, E. E. Vdovin, A. Mishchenko, O. Makarovsky, A. Patane, J. R. Wallbank, Y. Cao, A. V. Kretinin, M. J. Zhu, S. V. Morozov, V. I. Falko, K. S. Novoselov, A. K. Geim, T. M. Fromhold, and L. Eaves. Resonant tunnelling between the chiral landau states of twisted graphene lattices. *Nat. Phys.*, 11(12):1057–1062, 2015.

[89] A. S. Mayorov, R. V. Gorbachev, S. V. Morozov, L. Britnell, R. Jalil, L. A. Ponomarenko, P. Blake, K. S. Novoselov, K. Watanabe, and T. Taniguchi. Micrometer-scale ballistic transport in encapsulated graphene at room temperature. *Nano lett.*, 11(6):2396–2399, 2011.

[90] A. Mishchenko, J. S. Tu, Y. Cao, R. V. Gorbachev, J. R. Wallbank, M. T. Greenaway, V. E. Morozov, S. V. Morozov, M. J. Zhu, and S. L. Wong. Twist-controlled resonant tunnelling in graphene/boron nitride/graphene heterostructures. *Nat. nanotechnol.*, 9(10):808–813, 2014.

[91] Y. Zhao, Z. Wan, X. Xu, S. R. Patil, U. Hetmaniuk, and M. P. Anantram. Negative

differential resistance in boron nitride graphene heterostructures: Physical mechanisms and size scaling analysis. *Scientific Reports*, 5:10712, 2015.

[92] Y. Zhao, Z. Wan, U. Hetmanuik, and M. P. Anantram. Negative differential resistance in graphene boron nitride heterostructure controlled by twist and phonon-scattering. *IEEE Electron Dev. Lett.*, 37(9):1242–1245, 2016.

[93] Y. Zhao, U. Hetmaniuk, S. R. Patil, J. Qi, and M. P. Anantram. Nested dissection solver for transport in 3D nano-electronic devices. *J. Comput. Electronics*, 15(2):708–720, 2016.

[94] Y. Zhao, Z. Wan, X. Xu, S. R. Patil, U. Hetmaniuk, and M. P. Anantram. A modeling study of mechanisms for NDR in graphene-bn-graphene heterostructures. In *Nanotechnology Materials and Devices Conference (NMDC), 2015 IEEE*, pages 1–2. IEEE, 2015.

[95] G. Trambly de Laissardiere, D. Mayou, and L. Magaud. Localization of Dirac electrons in rotated graphene bilayers. *Nano Lett.*, 10(3):804–808, 2010.

[96] P. Moon and M. Koshino. Electronic properties of graphene/hexagonal-boron-nitride Moiré superlattice. *Phys. Rev. B*, 90(15):155406, 2014.

[97] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom. nanoMOS 2.5: a two-dimensional simulator for quantum transport in double-gate MOSFETs. *IEEE T. Electron Dev.*, 50(9):1914–1925, 2003.

[98] J. R. Barker, J. Pepin, M. Finch, and M. Laughton. Theory of non-linear transport in quantum waveguides. *Solid-state electron.*, 32(12):1155–1159, 1989.

[99] 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. *J. Appl. Phys.*, 100(4):043713, 2006.

[100] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva. Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs. *IEEE T. Electron Dev.*, 50(9):1837–1852, 2003.

[101] A. Martinez, K. Kalna, J. R. Barker, and A. Asenov. A study of the interface roughness effect in Si nanowires using a full 3D NEGF approach. *Physica E*, 37(1):168–172, 2007.

[102] A. Martinez, J. R. Barker, A. Asenov, M. Bescond, A. Svizhenko, and A. Anantram. Development of a full 3D NEGF nano-CMOS simulator. In *Simulation of Semiconductor Processes and Devices, 2006 International Conference on*, pages 353–356. IEEE, 2006.

[103] A. Martinez, N. Seoane, A. R. Brown, and A. Asenov. A detailed 3D-NEGF simulation study of tunnelling in n-Si nanowire MOSFETs. In *Silicon Nanoelectronics Workshop (SNW), 2010*, pages 1–2. IEEE, 2010.

[104] A. Nissen and G. Kreiss. An optimized perfectly matched layer for the Schrödinger equation. *Commun. Comput. Phys.*, 9:147–179, 2011.

[105] The MathWorks Inc. *MATLAB Release 2011b*. Natick, Massachusetts, United States, 2011.

[106] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J.

Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian-09 Revision A.1. Gaussian Inc. Wallingford CT 2009.

[107] T. Davis. *Direct Methods for Sparse Linear Systems*. SIAM, 2006.

[108] Y. Cui and C. M. Lieber. Functional nanoscale electronic devices assembled using silicon nanowire building blocks. *Science*, 291(5505):851–853, 2001.

[109] 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. *Phys. Rev. B*, 74(20):205323, 2006.

[110] B. Göhler, V. Hamelbeck, T. Z. Markus, M. Kettner, G. F. Hanne, Z. Vager, R. Naaman, and H. Zacharias. Spin selectivity in electron transmission through self-assembled monolayers of double-stranded DNA. *Science*, 331(6019):894–897, 2011.

[111] H. W. C. Postma. Rapid sequencing of individual DNA molecules in graphene nanogaps. *Nano Lett.*, 10(2):420–425, 2010.

[112] M. Tsutsui, K. Matsubara, T. Ohshiro, M. Furuhashi, M. Taniguchi, and T. Kawai. Electrical detection of single methylcytosines in a DNA oligomer. *J. Am. Chem. Soc.*, 133(23):9124–9128, 2011.

[113] J. Qi, N. Edirisinghe, M. G. Rabbani, and M. P. Anantram. Unified model for conductance through DNA with the Landauer-Büttiker formalism. *Phys. Rev. B*, 87(8):085404, 2013.

[114] D. Bednarczyk and J. Bednarczyk. The approximation of the Fermi-Dirac integral $f_{12}(\eta)$. *Phys. Lett. A*, 64(4):409–410, 1978.