LOW-RANK APPROXIMATIONS IN QUANTUM TRANSPORT SIMULATIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Daniel A. Lemus

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Prof. Tillmann Kubis, Co-chair

School of Electrical and Computer Engineering

Prof. Gerhard Klimeck, Co-chair

School of Electrical and Computer Engineering

Prof. David Gleich

Department of Computer Science

Prof. Milind Kulkarni

School of Electrical and Computer Engineering

Approved by:

Prof. Dimitrios Peroulis

Head of the School of Electrical and Computer Engineering

ACKNOWLEDGMENTS

I would like to thank my family, friends at Purdue and at home, and my advisors for keeping me motivated to reach a successful end to what I had once considered an insurmountable goal.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

ABBREVIATIONS

GPU	graphics processing unit
I-V	current-voltage
NEGF	Nonequilibrium Green's function
LRA	low-rank approximation
MIC	Many Integrated Core
MOSFET	metal-oxide-semiconductor field-effect transistor
MPI	Message Passing Interface
RGF	recursive Green's function
TB	tight binding
TFET	tunneling field-effect transistor
UTB	ultra-thin body

ABSTRACT

Lemus, Daniel A., Ph.D., Purdue University, May 2020. Low-rank Approximations in Quantum Transport Simulations. Major Professors: Tillmann Kubis, Gerhard Klimeck.

Quantum-mechanical effects play a major role in the performance of modern electronic devices. In order to predict the behavior of novel devices, quantum effects are often included using Non-Equilibrium Green's Function (NEGF) methods in atomistic device representations. These quantum effects may include realistic inelastic scattering caused by device impurities and phonons. With the inclusion of realistic physical phenomena, the computational load of predictive simulations increases greatly, and a manageable basis through low-rank approximations is desired.

In this work, low-rank approximations are used to reduce the computational load of atomistic simulations. The benefits of basis reductions on simulation time and peak memory are assessed. The low-rank approximation method is then extended to include more realistic physical effects than those modeled today, including exact calculations of scattering phenomena. The inclusion of these exact calculations are then contrasted to current methods and approximations.

1. INTRODUCTION

Numerical simulations are commonly used in the scientific and engineering world to model complex physical phenomena in nature. Using well-known theory, the behavior of state-of-the-art technologies can be predicted through simulations. The design of electronic devices, such as semiconductor transistors, has taken this route, since the cost of experimentation and fabrication of devices often greatly outweighs that of simulations for device behavior prediction.

Today's state-of-the-art computer chips are fabricated to contain billions of densely packed transistors, each with dimensions in the nanometer scale. It is well known that at the nano-scale, where atoms are countable, quantum effects may drastically change device performance [1–4]. Atomistic models, the simulations of quantum effects at the subatomic scale, have thus become a requirement for the effective simulation of novel nano-devices.

The NanoElectronic MOdeling (NEMO5) software suite [5, 6] is an in-house software designed by the iNEMO group at Purdue University, and is one of the world's most flexible quantum transport software packages. It is currently being used worldwide by various semiconductor device design companies and academic institutions to predict the performance of- and design state-of-the-art electronic devices before they begin to be fabricated. The various methods described in this thesis document have been implemented into the NEMO5 software and can be used by users of NEMO5.

1.1 How this thesis document is organized

One purpose of this thesis is to describe methods for solving computationally intensive equations involved in atomistic device simulations, the computational burdens of which will be described in chapter 2 of this thesis document. The first method is

described in chapter 3, which involves the use of coprocessors to solve large linear systems of equations for these atomistic models through a high degree of parallelization through multithreading. In chapter 4, the use of highly parallel and powerful supercomputers through multiprocessing and high performance computing is explored. High performance computing environments and heterogeneous systems with coprocessors may be combined to solve quantum transport systems.

In chapter 5 of this document, the solution for solving the complex and computationally intensive atomistic quantum transport equations is flipped, and the problem is made smaller instead of using a large amount of computational resources. A low-rank approximation (LRA) method called the mode space method [7] is introduced, which allows for the solution of the quantum transport equations with incoherent scattering with limited resources. In chapter 6, the mode space method is assessed in terms of its physical correctness and performance improvements which include faster time to solution and lower peak memory.

One major significance of basis reductions shown in this thesis is the ability to simulate large devices as well as more complex and more exact physics. In chapter 7, the mode space method is extended by including the solution of realistic physics through exact retarded scattering self-energies via the Kramers-Kronig relations. This is normally a highly computationally intensive process, but with basis reductions can be done in a reasonable amount of time. This novel method has been included and tested in NEMO5. Chapter 8 reports on another extension to LRA capabilities, accompanied with an extension [8] to the recursive Green's function (RGF) [9] algorithm that provides the capability of accurately solving nonlocality in quantum transport.

Chapter 9 outlines the impact of this PhD work and provides insight to the future work that may result from this PhD work.

2. DEVICE MODELING AND COMPUTATIONAL BURDENS

Atomistic computational models come with the advantage of providing a realistic model for the nano-scale devices in today's computer chips, but also with the disadvantage of having linear systems with many degrees of freedom and interdependent differential equations. In this chapter, the many layers of complexity in solving quantum transport in an atomistic system are laid out, as well as their computational burdens and solutions to these burdens which this thesis document will aim to provide.

2.1 The Non-equilibrium Green's function (NEGF) formalism

The characteristic length scale of state-of-the-art logic devices has reached dimensions with a countable number of atoms [8,10]. At this scale, quantum effects such as tunneling, interference and confinement drastically change device performance [1–4]. Understanding and optimizing these effects almost always requires predictive models. The non-equilibrium Green's function (NEGF) formalism is the well-accepted method of modeling of coherent and incoherent electron transport [11–14] in molecules [15], carbon nanotubes [16], MOSFETs [17] and many other nano-scale devices. NEGF has been solved for nanodevices represented in realistic basis sets [18–21]. Important device parameters such as electron density and current can be calculated using an interplay of quantum transport and electrostatics through the NEGF and Poisson equations [11]. Characteristic nanoelectronic device dimensions contain a countable number of atoms, but a typical transistor contains hundreds to thousands of atoms in the volume of only a few cubic nanometers. Accurate basis representations such as the empirical tight binding method [22,23] usually contain tens of matrix elements per

atom representing atomic orbitals [24]. Solving the NEGF equations in a tight binding basis can be computationally cumbersome due to the required matrices consisting of thousands of rows and columns [25, 26].

2.2 Computational burdens of atomistic simulations

In the following subsections, the various contributors to computational complexity in quantum transport are introduced and explained.

2.2.1 Atomistic basis sets

Devices at the nano-scale may have variations that may span the space between atoms, introducing quantum-mechanical effects that must be captured at subatomic resolutions [19, 23]. These variations may be introduced by quantum confinement [27, 28], material impurities [29, 30], surface and interface roughness [31, 32] and atomic variations of material alloys [33–35]. All effects must be captured in nanoscale device simulations to maintain physical accuracy. Some semiclassical methods may be used for modeling nanoscale devices such as the WKB approximation [36], but these methods may fail to capture some quantum effects [8, 37, 38].

In the simplest atomistic cases [39], matrices representative of the device used in NEGF may contain only hundreds of rows. However, accurate basis representations such as the empirical tight binding (ETB) model [22, 24] may introduce tens of matrix elements per atom representing atomic orbitals. When introducing spin-orbit coupling, the degrees of freedom double [24], therefore doubling the rank of matrices. Although nanoelectronic devices contain a countable number of atoms, a typical transistor may contain hundreds to thousands of atoms in the volume of only a few cubic nanometers. For all of the materials mentioned in this thesis, which have either a zincblende or diamond lattice crystal configuration, the number of atoms in a device can be determined by

$$N_{atoms} = \frac{w_1 \times w_2 \times l}{a^3} \times 8, \tag{2.1}$$

where w_1 and w_2 are widths of the device cross-section, l is the length and a is the lattice constant of the material. This value is multiplied by 8 since zincblende and diamond lattice configurations have 8 atoms per cubic unit cell. This number must of course be an integer. The total degrees of freedom of the system are thus the sum of all orbitals of every atom in the system $N_{atoms} \times m$ where m is the number of orbitals per atom. Solving the NEGF equations in a realistic tight binding (TB) basis can be computationally cumbersome due to matrices with ranks ranging in the tens-of-thousands to hundreds-of-thousands [25].

The parameters for atomistic basis sets are provided from prior scientific knowledge by fitting to physical observables and to other parametrization methods [40–42]. Ab initio "first principles" methods also exist, which calculate electronic behavior starting from fundamental physical parameters [43, 44]. These methods, such as density functional theory (DFT) [45, 46], can be very computationally expensive for realistically-sized devices [19,20]. Other atomistic bases include the effective mass approximation [4, 42] for which each atom is given a single degree of freedom, but only provides a simplistic parabolic band structure (more details given in section 5.1) and maximally localized Wannier functions [47,48], which are calculated from first principles and parameters must therefore be pre-determined before transport calculations can begin [49–51].

2.2.2 Solving NEGF with the recursive Green's function (RGF) algorithm

The solution of the NEGF equations, in the simplest case, would involve the solution of linear systems of equations with dense matrices representative of every atom and (depending on the basis) electronic orbitals in the device. For $\tilde{N} = N_{atoms} \times m$ where m are the degrees of freedom per atom corresponding to atomic orbitals in

a tight binding representation, a direct solution to the linear system of equations involves dense matrix operations that scale on the order $O(\tilde{N}^3)$ for time to solution and $O(\tilde{N}^2)$ for system memory. Because $\tilde{N}$ can be as large as tens-of-thousands to hundreds-of-thousands, only very small devices with simple bases like the effective mass approximation may be solved by "direct" NEGF [14].

To ease numerical load, the recursive Green's function method (RGF) [9, 52] provides a block-wise recursive solution for NEGF equations that can be discretized with block-tridiagonal sparse matrices [53–55]. With RGF, realistically-sized systems may be solved using a realistic atomistic basis such as empirical tight binding (ETB) [22,24]. The size of each matrix block for block-row I and block-column J of a Green's function matrix depends on the cross-section and size of the tight binding basis, and results in square blocks of rank N. In the RGF algorithm the goal is to solve each block $G^R_{I,J}$ and $G^<_{I,J}$ of a block tri-diagonal Green's function matrix starting from a block tri-diagonal Hamiltonian H as shown in figure 2.1. The device Hamiltonian H is a matrix that represents the energy states of the entire device. The algorithm is divided into two portions, the forward RGF portion where the functions $g^R_{I,J}$ and $g^<_{I,J}$ are solved recursively from the upper-left to the bottom-right of the matrix, and the backward RGF portion where the functions $G^R_{I,J}$ and $G^<_{I,J}$ are solved. Details of this process can be found in various publications, such as [9, 56, 57]. The method for solving the RGF algorithm will also be detailed in section 5.4.

With the RGF algorithm, matrices are divided into blocks of rank N. The rank N of these blocks is determined in zincblende and diamond lattice configurations by

$$N = \frac{w_1 \times w_2}{a^2} \times 2n_l \times m, \tag{2.2}$$

where n_l layers along the device length are used for a block such that $N \ll \tilde{N}$. The number 2 is used since a single atomic layer contains 2 atoms in the zincblende and diamond lattice configurations. Time to solution of matrix operations on these blocks scales on the order of $O(N^3)$ and memory usage scales on the order of $O(N^2)$. Any number of layers n_l can be chosen for a block layer in RGF, so RGF would need to

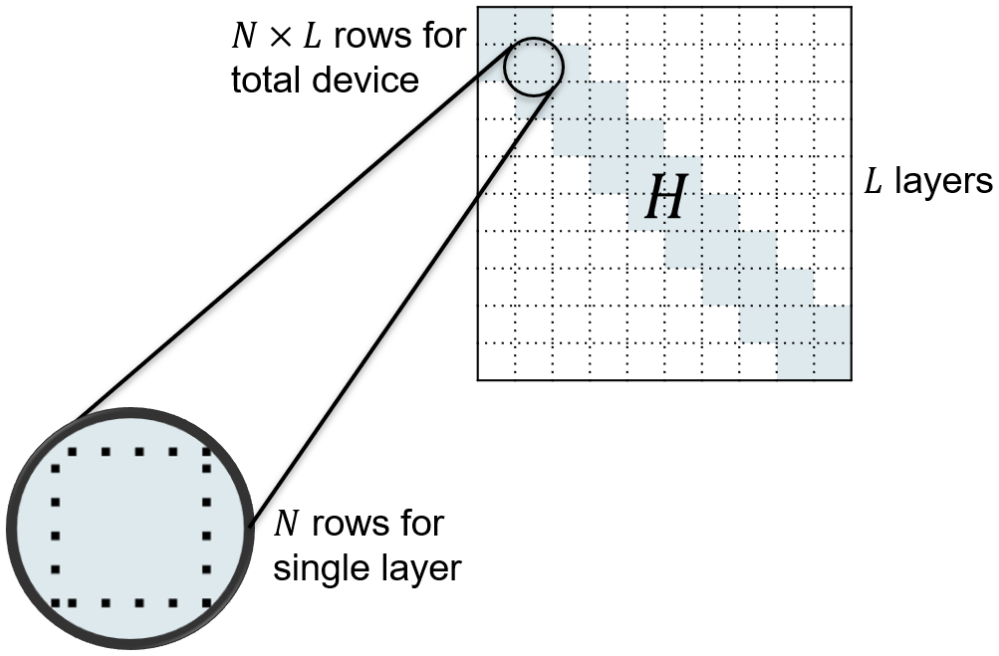


Figure 2.1. A block tri-diagonal Hamiltonian of total $N \times L$ rows and columns for L device layers. Each block in this matrix has N rows and columns

solve fewer recursive iterations when more layers are used, such as a unit cell thickness $n_l = 4$. However, the RGF method iterates along the length of the device, so length contributes linearly to the complexity of the RGF solution of the entire device. The most computationally efficient option would therefore be to keep $n_l = 1$. From this point forward in this thesis, the solution of Green's functions with RGF described will indicate the solution of $G^R_{I,J}$ and $G^<_{I,J}$ of rank N, rather than the Green's functions of the entire device which would be of rank $N \times L$ for L layers.

2.2.3 Calculation of multiple energies and momentums

The solution of the above must be performed independently on many energies of interest which, depending on the device material, geometry, temperature, and applied

voltage, may number in the thousands. In NEMO5 these energies are determined inhomogeneously using an adaptive energy grid that detects the energies at which transport occurs based on band structures [8]. This way, energies at which electronic transmission is unlikely to occur are omitted, while resonant energy ranges have a dense energy presence. Figure 2.2 shows an example of the energy distribution of 377 energies for an InAs tunneling field-effect transistor (TFET) in its ON-state, where darker horizontal lines correspond to a higher density of energies. Some devices, such as ultra-thin body (UTB) devices, must be solved independently for tens to hundreds of k-points (directly proportional to momentum) in reciprocal space [3, 4]. The number of times the NEGF equations must be solved in parallel depends directly on the number of (E, k) points being solved. More on this type of device will be discussed in section 4.3.

2.2.4 Self-energies and incoherent scattering

Simulated nanoelectronic devices such as MOSFETs most often consist of a central semiconductor device represented as an open system. Contacts on both ends of the semiconductor device are represented by semi-infinite leads. These leads represent infinite reservoirs of electrons controlled by an applied potential [4]. Since modeling infinite contacts is unfeasible, boundary conditions are applied to the semiconductor device through self-energies, which represent interactions between the device and external sources. These interactions can represent electrons entering and exiting the device through the contacts as well as perturbations from finite temperatures through phonons and device impurities. Incoherent scattering on phonons can thus be introduced through self-energies, which are included in the quantum transport model of the NEGF formalism [12, 58].

Fabrication of nanoelectronic devices is not perfect, and structure uncertainties exist in final products. These imperfections may include roughness, alloy disorder, and geometrical errors. These non-uniformities in the material's crystalline structure

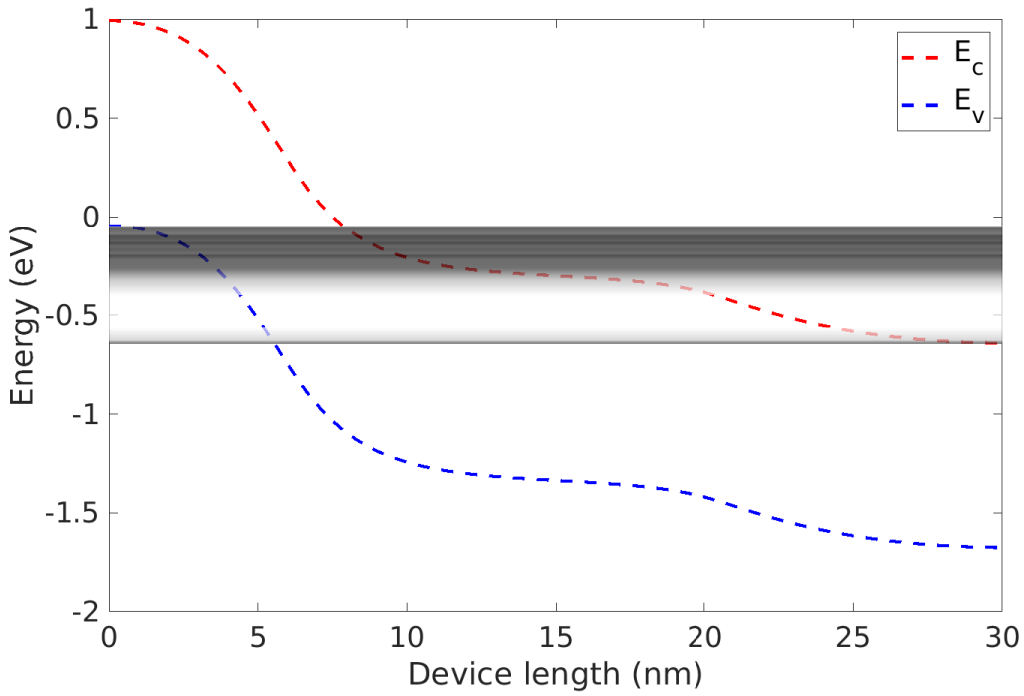


Figure 2.2. Conduction and valence band edges and inhomogeneous energy distribution for a TFET in its ON-state. Darker horizontal lines correspond to a higher density of energies placed by the NEMO5 adaptive energy grid

result in scattering centers which may alter an electron's phase, energy and momentum [8, 27, 29]. An advantage of the NEGF and RGF methods is the ability to introduce incoherent scattering through self-energies, which represent device structure uncertainties such as roughness, alloy disorder and geometric errors, and temperature fluctuations through phonons [11, 12, 27, 28, 30, 34, 59–63]. Phonons, quasi-particles that represent vibrations in the crystal lattice of the device material, contribute to temperature fluctuations in the device and exchange energy with electrons of the device. Incoherent scattering due to phonons alters device performance and may not be ignored in realistic device modeling at finite temperatures [11, 27, 28, 30, 34, 60–65].

Two main computational challenges exist when solving the self-energy equations. The first is the complexity of the integrals involved, which will be explored in sec-

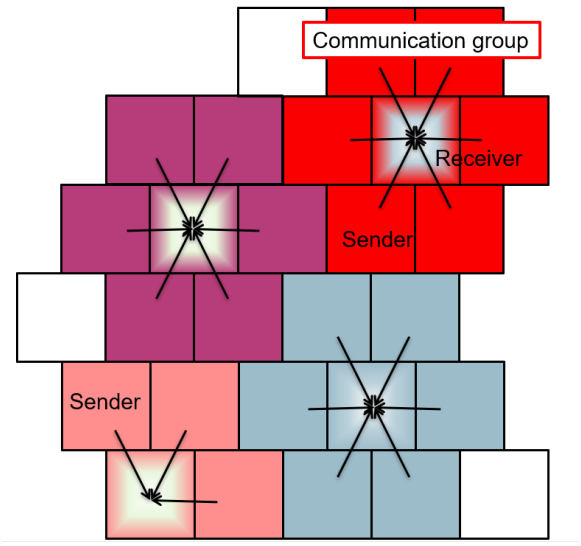
tion 7.2. The second computational challenge is the communication required between energies and between momentums. To take advantage of the mostly independent energies E and momentums k, NEMO5 and other quantum transport simulators use multiprocessing through MPI to solve (E, k) tuples in parallel. This independence breaks down by the introduction of scattering, since electrons may be transferred from one (E, k) to another, and may therefore require communication from one process to another. Communication may be unpredictable and highly complex due to overlaps in communication patterns required by each (E, k) tuple, as shown by figure 2.3.a. Care must be taken when planning communication so that load imbalances may be minimized and deadlocks may be avoided [8, 25]. This is done in NEMO5 by sorting (E, k) tuples into independent communication groups as depicted in figure 2.3.b. These tuples are sorted in such a way that as many groups simultaneously solve the self-energy equations using the required (E, k) tuples from neighboring processes as possible. This reduces idling by processes, thus improving scaling capabilities [8, 25].

On top of the burden of communication and computation of self-energies, the introduction of incoherent scattering into the RGF solution introduces yet another degree of complexity through the self-consistent solution of retarded and lesser Green's functions $G^{R,<}$ and the corresponding scattering self-energies $\Sigma^{R,<}$.

2.2.5 Self-consistent equations

The solution of the NEGF equations alone does not immediately result in the solution of measurable quantities of a device. The NEGF equations must be self-consistently solved with the Poisson equation that represents the electrostatic effects caused by the quantum mechanical evolution of the system [11,66,67]. This introduces a degree of complexity to the solution of NEGF, since solving the equations is required multiple times.

(a) Overlap of (E, k) tuple communication groups



(b) Communication of tuple groups in NEMO5

Figure 2.3. Depiction of (a) overlap in the complex communication of scattering calculations required by the (E, k) tuple at the center of each communication group, and depiction of (b) communication of (E, k) tuples divided into groups which perform communication independently in NEMO5. *Image courtesy of Tillmann Kubis*

The introduction of inelastic scattering into the NEGF solution introduces yet another degree of complexity through the self-consistent solution of Green's functions

$$G^R = (E - H - \Sigma^R)^{-1} \tag{2.3}$$

$$G^< = G^R \Sigma^< G^{R\dagger} \tag{2.4}$$

which represent electron density of states and occupied electron density respectively, and the self-energies

$$\Sigma^< = G^< D^< \tag{2.5}$$

$$\Sigma^R = G^R D^R + G^R D^< + G^< D^R \tag{2.6}$$

which represent the perturbations of the electrons. H is the Hamiltonian and D is the sum of environmental Green's functions with phonon, impurity and roughness information [12, 68]. The self-consistent Born method [69] provides the self-consistency needed to solve these important equations. Within the self-consistent Born approximation the scattering self-energies $\Sigma^{R,<}$ and Green's functions $G^{R,<}$ are solved iteratively to achieve particle number conservation [12, 58, 70]. When combined, two loops of self-consistency exist, as depicted in figure 2.4. This interdependence can result in the solution of dozens of iterations for a single NEGF calculation.

It is worth mentioning that some alternatives to the self-consistent Born approximation of scattering exist, such as low-order approximations [71–73], the Büttiker probe scattering model [11, 65, 74] and the multi-scale approach of reference [75]. Other layers of self-consistency may also exist, such as in calculations of self-heating, which may self-consistently couple thermal conduction, quantum dot gain equations and carrier transport equations [76–78]. Calculations of resistive RAM (RRAM) devices may include self-consistency between carrier transport and heat conduction [79, 80] as well as ion and electron transport [81, 82].

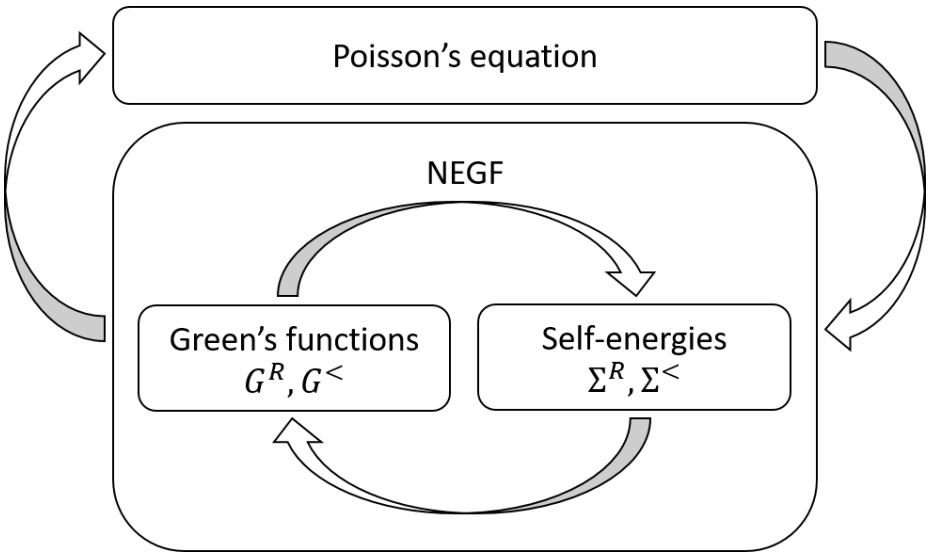


Figure 2.4. Diagram showing the two levels of self-consistent calculations required for the solution of incoherent scattering in NEGF. The first layer consists of the interaction between Green's functions and self-energies for the solution of NEGF. The second layer consists of the interactions between the NEGF solution for quantum mechanical evolution of the system and Poisson's equation for electrostatic effects

2.2.6 Calculation of multiple bias points

An important measure of electronic device performance is the response of current to applied voltage. The shape of the current-voltage (I-V) response curve is often a central figure of merit in transistor design. The subthreshold slope (SS), which is inversely proportional to the slope of this curve, demonstrates the speed at which a transistor switches when a bias is applied to a terminal [3, 83, 84]. Assessment of current response of the device to a dozen or more applied voltage biases is often needed to understand device switching performance. Figure 2.5 shows an example of such a curve, with 14 points calculated independently (This specific I-V curve is shown later in figure 7.3 for "zero Σ^R real part"). Only by calculating this amount of points can the shape of the I-V curve be determined, therefore all of the aforementioned collection of expensive computations must be performed >10 times for device engineering.

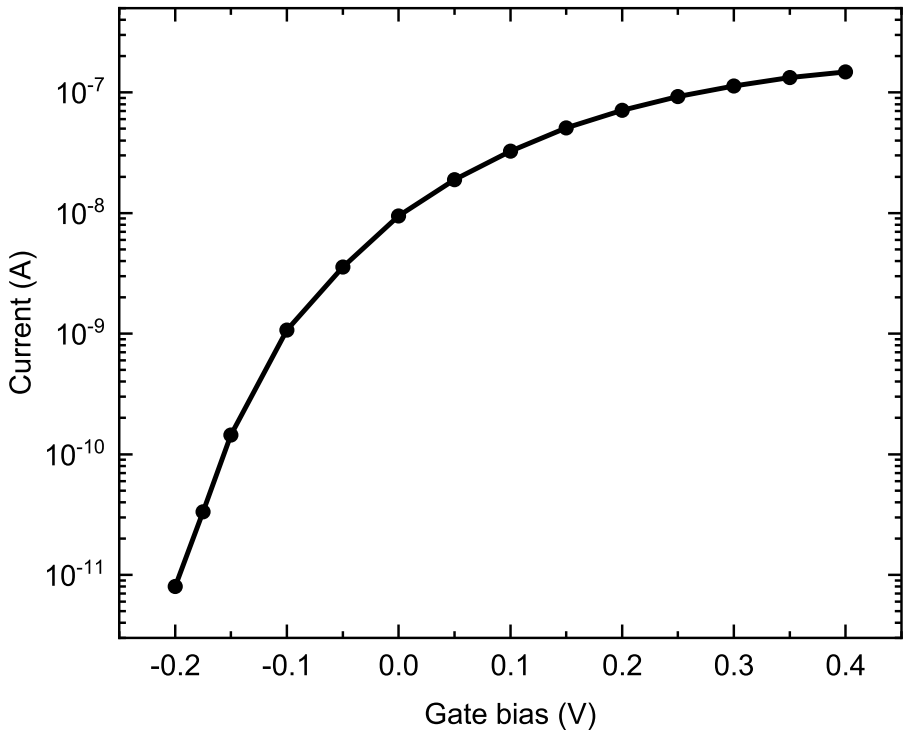


Figure 2.5. Current-voltage (I-V) characteristic curve of a 3.64 nm × 3.64 nm × 30.29 nm InAs TFET device showing the need to solve 14 independent points to determine curve shape

2.2.7 Calculation of various device materials and geometries

Often different materials and device geometries are required to understand the behavior of a new electronic device design. Modeling of the above simulations may be needed for multiple devices to properly assess performance behavior. Different semiconductor materials have a vast range of performance capabilities, such as III-V semiconductors like InAs and GaSb that are suitable for tunneling devices [26, 30, 85] due to inherent material properties that influence a low subthreshold slope but result in low ON/OFF current ratios [83]. This is as opposed to Si MOSFETS and TFETs which can achieve higher ON/OFF current ratios at the expense of lower switching speeds (higher SS) [86, 87]. Geometry variations such as device width also have an effect on device performance due to confinement having an effect on electron

mobility [27, 51, 88]. Material orientations may also result in varying device performance [24, 89–91]. Since all of these variations in material design have significant performance effects, having overly-expensive simulations would hinder the ability to quickly modify device configurations during device design.

2.2.8 Calculation of imperfections

Modeling of imperfections may be required to statistically predict the performance of a real-world nanoelectronic device that has been manufactured with imperfections. These imperfections may include surface roughness [27, 61], interface roughness [31, 32, 92, 93] and alloy disorder [33, 34, 94]. Statistical analysis to determine the average performance of a device with these imperfections may require the solution of hundreds of data points [25,95], therefore all aspects of the complex simulations described above may need to be performed hundreds of times. The introduction of these imperfections may not always be explicitly modeled and statistically analyzed, as there also exist implicit models using incoherent scattering [8, 68] and approximations such as the virtual crystal approximation (VCA) [96] which implicitly models alloy disorder.

2.3 Solutions to computational burdens

Although including the aforementioned effects into NEGF device simulations can result in heavy computational burdens, there are multiple solutions to this issue.

2.3.1 Ballistic simulations

Ballistic simulations, which represent the transport of electrons without the inclusion of scattering on phonons, are often used to avoid the computational burden of scattering calculations as described above. Very small devices are ballistic to a significant degree, since the mean-free-path of traveling electrons may exceed the channel length of the semiconductor device [4]. However, realistically, imperfect nano-devices

in a finite-temperature setting are affected by perturbations which make ballistic simulations inaccurate [8, 60, 97].

2.3.2 Approximations to scattering simulations

The solution through the NEGF formalism often includes several approximations to allow for faster system solutions. Details on these approximations can be found in the assessment by Kubis et. al [12]. One of these approximations, which involves the removal of a principal value integral from the calculation of retarded scattering self-energies [8, 68, 97, 98], is discussed in detail in section 7.2. Another significant approximation, which will be approached in chapter 8 of this thesis report, is the diagonal self-energy approximation. This approximation allows for a significant decrease in complexity, but may result in significant deviations from experiment [12].

2.3.3 Low-rank approximations

Many discretized degrees of freedom are common in atomistic representations, as well as multiple layers of self-consistency, that result in heavy computational burdens. To ease this burden, incoherent scattering effects are often neglected in NEGF transport calculations [4, 99–102]. In the case of atomistic representations, even ballistic NEGF calculations often yield large computational loads. Such situations have motivated the introduction of a low-rank approximation (LRA) [103] into NEGF [54, 59, 100, 104–106] that is often called the mode space approach [26, 53, 99, 107]. Since scattering phenomena are important to retain in quantum transport simulations, the goal of this work is to introduce a LRA that accurately retains scattering phenomena and remains based on an atomistic device representation. Atomistic bases may be reduced with matrix transformations through LRA. Some reduction methods may reduce matrix rank to under 10% of the size of the original system [7, 104, 108]. Some previous methods of LRA have been performed in atomistic simulations: In the works of references [109–113], the contact block reduction (CBR) method divides the

device Hamiltonian into inner-device and boundary partitions and a subset of propagating modes are chosen. Similarly, the quantum transmitting boundary method (QTBM) shown in references [34] and [114] solves transport for propagating modes while excluding vanishing modes. The work of reference [115] constructs a rectangular transformation matrix that reduces the device Hamiltonian. This transformation matrix is created by adding columns with spatial and energy information until a residual is smaller than a chosen tolerance.

Although many methods of LRA such as the aforementioned exist, the principal method of LRA in this thesis is the mode space method of reference [7], which Mil'nikov et. al call the "equivalent model." In reference [104], an effective mass approximation is used and reduced using eigenvectors corresponding to eigenvalues in a desired energy range. This range is often from the conduction band edge energy E_0 and several k_BT above E_0, where k_B is the Boltzmann constant and T is the temperature of the device. Reference [116] shows this mode space method in an effective mass approximation with scattering. The work of Mil'nikov et. al [7] expands on the mode space method by introducing it to a tight binding basis, and includes scattering. This will be further elaborated on in chapter 5. The works of references [26, 53, 85, 107, 108, 117] and [118] employ this method.

2.3.4 Highly parallel computing

Due to the high degree of parallelism in the solution of quantum transport in nanoelectronic devices, parallel solutions of the equations through multiprocessing and multithreading are often necessary. The availability of powerful supercomputers with hundreds of thousands of CPUs allows for highly complex problems such as NEGF to be solved well within our lifetimes, and scaling capabilities of simulations on these machines determines the capability of simulation models and software to effectively use this technology. The creators of many scientific software products make great efforts to make their software as scalable as possible [119–123] to the

extent that annual competitions such as the Gordon Bell Prize competition [124] exist to showcase the most highly scalable scientific software. The highly scalable NEMO5 similarly provides the capability of modeling nano-scale devices using atomistic bases on hundreds of thousands of CPU cores [25]. In the next two chapters, this capability is explored.

3. HETEROGENEOUS COMPUTING

As mentioned in section 2.3, the solution of the NEGF equations afford the need for a high degree of parallelism. However, the parallelism explored in this chapter will not be of energies or momentums, rather within individual mathematical matrix operations. The solution of the quantum transmitting boundary method (QTBM) in NEMO5 for ballistic transport prediction involves the solution of a linear system of equations in the form of $Ax = b$. The solution of this linear system of equations was a candidate for optimization by improving time to solution using the Intel Xeon Phi coprocessors, which could provide highly parallel solutions. The algorithm used to solve this linear system of equations was the Compression Algorithm, implemented in NEMO5 and based on the optimized renormalization method of Boykin et. al [125]. This algorithm involved the solution of a linear system of equations.

3.1 The Intel Xeon Phi coprocessor

The Intel Xeon Phi Knight's Corner (KNC) was a coprocessor introduced in 2012 that, similarly to a general-purpose GPU (GPGPU), provided computing clusters with a 61-core alternative to the typical 16-24 core Sandy Bridge CPUs available. Although Sandy Bridge CPUs were faster and more suited for sequential work, the Many Integrated Core (MIC) architecture of Xeon Phi KNC coprocessors allowed for highly parallel computation which was needed by the solution of the NEGF equations. For each hardware core, the coprocessor had 4 hardware threads, allowing for 244 total threads for solving parallel tasks such as matrix operations and vectorizable for-loops.

3.2 Description of linear system

The first software tool that was analyzed was the MKL BLAS GEMM (general matrix-matrix multiply) functions using MKL Automatic Offload. The solution of QTBM for ballistic transport simulations requires the solution of a linear system $Ax = b$, where A is a block-tridiagonal matrix. x and b are rectangular matrices, or block "vectors." The square matrix size of A could range between 50,000 rows to 500,000 rows, depending on the atomistic basis, cross-section and length of the device.

An important thing to note is that GEMM functions lie at the core of many linear solve algorithms [126], and those in NEMO5 are no exception. Due to the abundance of these BLAS functions in the algorithms, one could use MKL Automatic Offload or Compiler Assisted Offload to take advantage of the availability of a Xeon Phi coprocessor card in computing systems such as Stampede2 [127].

3.3 Compression algorithm

The Compression algorithm takes advantage of the linear system in such a way that the solution is performed in a computationally efficient manner. The reason that attempts were made to optimize these algorithms was that much of the computation in a typical NEMO5 simulation was spent within the linear solve. The device being simulated is divided in slabs, with the block rows of matrix A corresponding to each slab. The matrix A is mostly hermitian and in some simulation cases, mostly real. The exceptions to this are the corner matrices (block rows **a** and **i** as shown in figure 3.1), which are complex and non-hermitian. These blocks correspond to the contact self-energies and prevent us from using the hermitian and potentially real properties of the matrix interior. By decoupling the interior of the matrix from the corner blocks, however, we can still use any properties that may be exploited for computational speedup. Decoupling the inner blocks involves applying a renormalization algorithm [125] that decouples inner layers from its neighbors. One performs renor-

malization by applying the following operations to each block of matrix A, where i (not to be confused with the depicted layer **i** shown in figures 3.1 and 3.2) is the corresponding atomic layer being operated on:

$$\hat{A}_{i-1,i+1} = -A_{i-1,i}(A_{i,i})^{-1}A_{i,i+1}, \tag{3.1}$$

$$\hat{A}_{i+i,i-1} = (\hat{A}_{i-1,i+1})^{\dagger}, \tag{3.2}$$

$$\hat{A}_{i-1,i-1} = -A_{i-1,i-1} - A_{i-1,i}(A_{i,i})^{-1}A_{i,i-1}, \tag{3.3}$$

$$\hat{A}_{i+1,i+1} = -A_{i+1,i+1} - A_{i+1,i}(A_{i,i})^{-1}A_{ii+1} \tag{3.4}$$

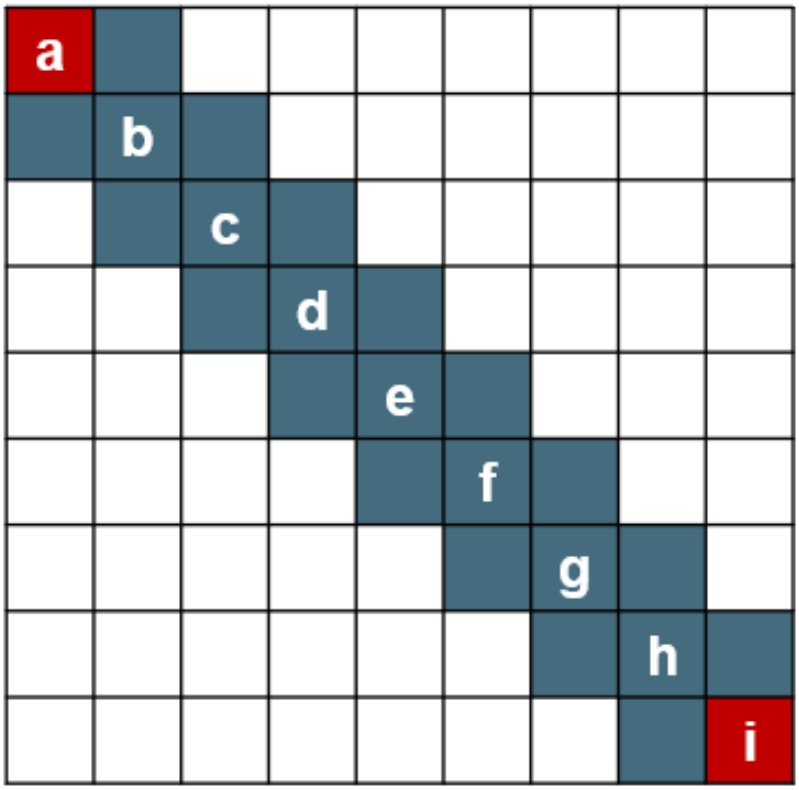
The order of these operations is not from the top-left corner of the block matrix to the bottom-right corner. Figure 3.2 shows an example of the order in which the layer-wise operations must be performed, and shows how each layer is modified, e.g. **a** into **a'**. The resulting matrix $\hat{A}$ is block-diagonal, as opposed to the block tri-diagonal matrix A. $\hat{A}$ also has properties which allow for obtaining the solution to x in a much more computationally efficient manner. Solving for x in the aforementioned $Ax = b$ system involves matrix-vector multiplication with the inverse of $\hat{A}$, as well as a matrix transformation. However, the bulk of the computation occurs when converting matrix A to $\hat{A}$.

When decoupling layers of the system, it is important to note that each layer is dependent of its nearest neighbors, but not any further layer. Since each alternating layer is independent of the others, renormalization can occur in parallel. NEMO5 is capable of performing parallelization of the renormalization process using MPI. With MPI, the workload of the entire matrix is divided among the MPI ranks by block rows. When parallelized, each process has a nearly equal number of block rows to perform renormalizations on.

3.4 Automatic offload to Xeon Phi

Like most algorithms that involve the solution of a linear system, BLAS GEMM functions are called very often in the Compression Algorithm, especially DGEMM,

(a) Depiction of partitioning in a nanowire device



(b) Block tri-diagonal matrix corresponding to device (a)

Figure 3.1. Nanowire device depiction with (a) atomic layers labeled from **a** to **i**, and corresponding block-tridiagonal matrix (b). This form of block tri-diagonal matrix is solved using the NEMO5 Compression Algorithm for the quantum transmitting boundary method (QTBM) model

the real double-precision matrix-matrix multiplication routine. In fact, most of the algorithm's execution time is spent in the DGEMM function. This yields the opportunity to use MKL Automatic Offload to improve performance in the presence of a Xeon Phi coprocessor card. A simple test was performed to measure the performance boost available through MKL Automatic Offload to a single coprocessor card, in which the Compression Algorithm was executed with various simulation sizes on a single pro-

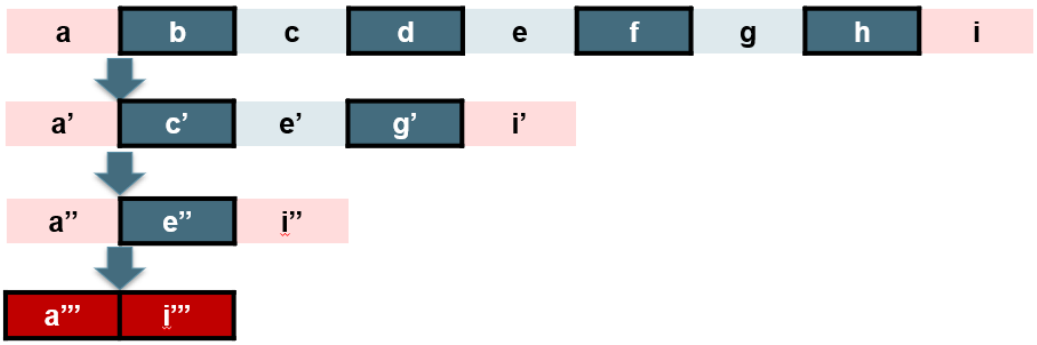


Figure 3.2. A simplified depiction of the order in which the compression algorithm modifies the blocks of the block tri-diagonal matrix of figure 3.1 using the operations shown in equations 3.1-3.4

cess, single Sandy Bridge core, single OpenMP thread and 240 OpenMP threads on the coprocessor. The ranks of the matrices ranged from 320 to 2880. MKL workload between the CPU and MIC was not explicitly set. Resulting execution times for the offload and no-offload case for each simulation size, shown in figure 3.3, show that MKL Automatic Offload indeed improves performance in the simplest case, when only a single process from the host Sandy Bridge, without OpenMP multithreading, offloads to a single coprocessor with 240 OpenMP threads. As is evident from figure 3.3, offloads occur only when the offloaded functions use matrices that exceed a certain size threshold of 2000 rows [128]. This test was performed on the Intel Endeavour supercomputer.

Since the algorithm contains MPI parallelization capabilities, it may be tempting to use both MPI and Automatic Offload to further improve performance. One must be careful when using MPI with Automatic Offload, since this could yield a sharp decrease in performance when a large number of processes offload large workloads. The reason for this is oversubscription of resources. When every process attempts to offload, there are not enough resources on the coprocessor for all offloads to be executed. This could be because all MIC OpenMP threads are in use by a process when another process attempts to offload. Another reason could be that memory

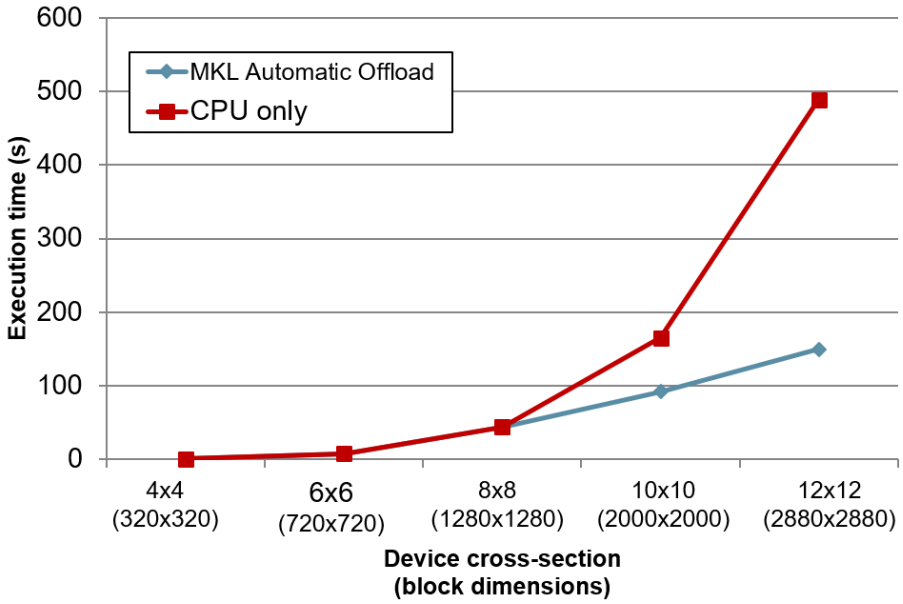


Figure 3.3. Comparison of DGEMM of various matrix sizes in Si unit cells between automatic offload to a single Intel Xeon Phi KNC coprocessor and a single process without multithreading. The coprocessor performed DGEMM operations on 240 OpenMP threads

resources are filled when a certain number of offloads are being performed on the coprocessor. The number of MIC OpenMP threads may vary depending on the number of coprocessor cards available on the system and the optimal number of threads to use with BLAS Automatic Offload is $(n-1) \times 4$ where n is the number of cores on the coprocessor.

Since figure 3.3 shows only a decrease in compute time when using a single process, it would remain to be seen whether a coprocessor could compete with a full Sandy Bridge CPU with 16 cores. To test this, a "realistic" comparison test was performed with a 16-core Sandy Bridge CPU host with 16 OpenMP threads and 2 coprocessors with 240 OpenMP threads each. This test was performed on the Purdue RCAC Conte computing cluster. The largest device from the previous test, a 12×12 unit cell (2880×2880 matrices) device was used with the compression algorithm. First, 1 to 16 MPI processes ran in parallel, and only the first two MPI ranks (if only one was available,

only one offloaded) offloaded automatically to the coprocessors after using the MKL routine `mkl_mic_enable()`. This divided the workload between the CPU host and the coprocessors: while the coprocessors performed parallelized matrix operations, the remaining host CPUs also performed matrix operations. Since multithreading was available through OpenMP and all 16 CPU cores were available, the CPU processes were able to use these threads to parallelize some operations. Secondly, the same test was run with 1 to 16 MPI processes, but with the offloading capability removed. The 16 OpenMP threads remained. Figure 3.4 shows the results of this comparison. The results show that for 1 and 2 MPI processes, time to solution is lower when CPUs simply use host threads to perform parallel work. This may be because of overhead in sending data to and from the coprocessor. For more than 2 processes, time to solution closely matches. Due to the lack of improved performance, and because Automatic Offloads lacked the fine-tuning capability to improve performance further, another method of manual offloading was turned to: MKL Compiler Assisted Offloading.

3.5 Compiler Assisted Offload

Intel MKL Compiler Assisted Offloading allows the software developer to have fine-tuned control of how work is sent to the Intel Xeon Phi KNC coprocessor. Fine-tuning capabilities include limiting the times that memory allocations occur on the coprocessor, control over which subroutines occur on the coprocessor, and load balancing between the coprocessor and the host CPUs. Load balancing is an important factor in improving time to solution performance, since the coprocessors are more capable than the host of performing large parallel operations and less capable of performing fast sequential work. By balancing the workload such that the coprocessor is able to complete its task at the same time as the host, idling time at an MPI barrier can be minimized and resource use can be optimized. Figure 3.5 shows an example of a load distribution among 6 MPI processes, with the first two offloading to a many-integrated-core (MIC) coprocessor and the rest performing work on host

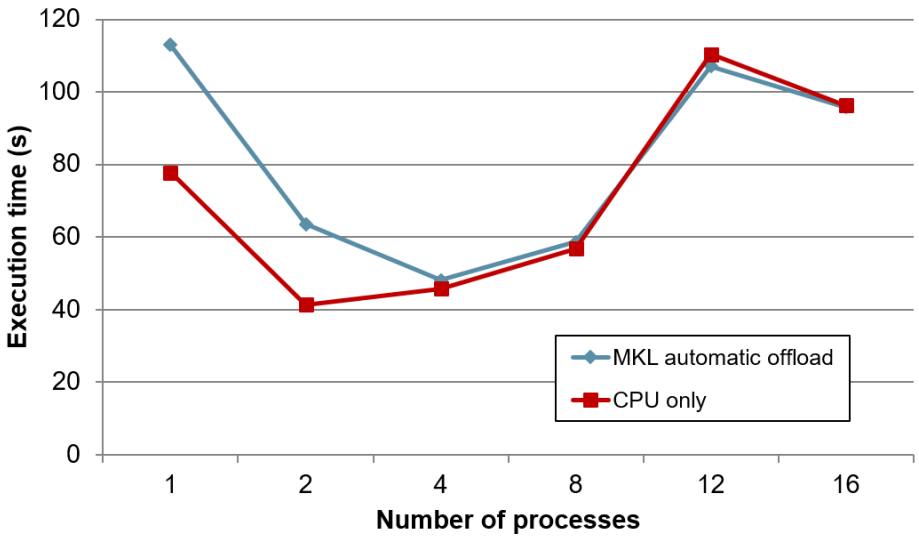


Figure 3.4. Performance test of 16 MPI processes working in parallel, with the first two processes offloading to a Xeon Phi KNC coprocessor using MKL Automatic offload. 16 OpenMP threads were available on the host CPU, so host processes also parallelized matrix operations when cores and threads were available. This was compared to a CPU-only test with 16 OpenMP threads

cores. The area of each square represents the amount of work needed to be done for a task. For example, a task could be a single matrix-matrix multiplication through a BLAS DGEMM routine call, and the area could represent the combined rank of the matrices. Host-only processes should ideally perform smaller, less parallelizable operations, but more of them sequentially, while the offloading processes should offload highly parallelizable operations while performing fewer of them. With the ideal workload distribution, no difference in time to solution occurs and therefore no idling time occurs on any process. Note that when solving the NEGF equations with the RGF algorithm, all matrices for a device which does not change in dimensions from source to drain will have equally-sized block matrices, therefore another method of efficient load balancing may be achieved by distributing a larger number of energies to processes which are able to perform mathematical operations more quickly, such as processes with a coprocessor.

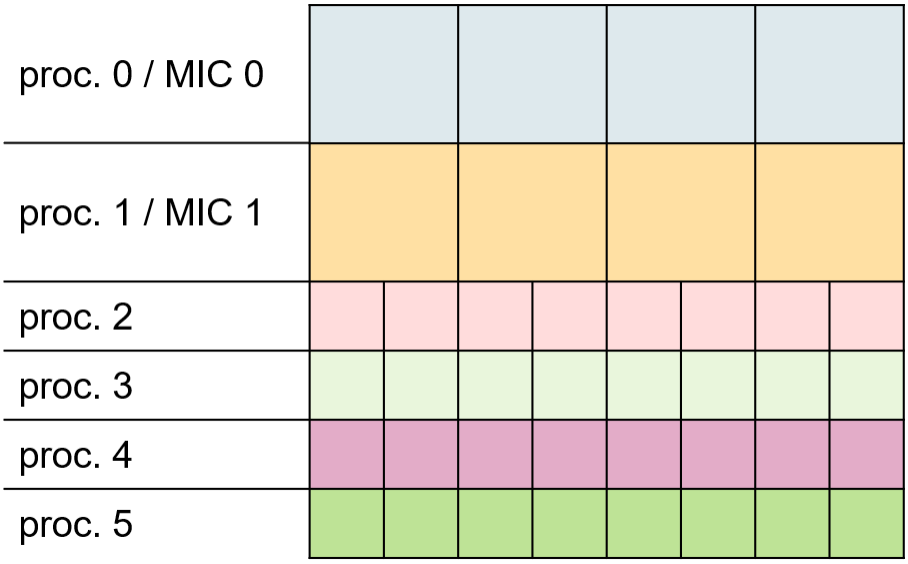


Figure 3.5. An example of an ideal workload distribution for 6 processes, two of which are capable of offloading work to a many integrated core (MIC) coprocessor. Coprocessors are capable of performing highly parallel computations, so operations such as large matrix multiplications should be performed there. Since mathematical operations may have a shorter time to solution on coprocessors, another method of load balancing may be to distribute a larger amount of tasks, e.g. energies, to offloading processes. The ideal load distribution would have each process complete its task in the same amount of time for minimal idling

3.6 Optimized dense matrix multiplication

For all further Intel Xeon Phi KNC tests with Compiler Assisted Offload, operations were done on the RGF algorithm, which include the dense matrix operations ZGEMM (complex matrix-matrix multiplication) and ZGESV (complex linear system of equations). The reason for moving on from tests with the Compression Algorithm was that the algorithm was used for the quantum transmitting boundary method (QTBM), which is only valid for ballistic simulations. RGF is a useful tool for its ability to include scattering effects.

With the availability of fine-tuned control of offload events, one can use computational techniques to improve performance by using available resources more efficiently. One such example is tiling, which uses available resources to perform computation and memory access simultaneously. In this case, a tiling algorithm was written such that access to a coprocessor was performed as simultaneously as possible to matrix multiplication on the coprocessor. This was done using OpenMP multithreading. Figure 3.6 shows a depiction of a matrix multiplication $A \times B = C$. The first step is to send n rows of matrix A and n columns of matrix B as rectangular matrices via an offload operation to the coprocessor. The next step in the process is two-fold: while the dense matrix multiplication is performed on the coprocessor (The result of this is an $n \times n$ block of matrix C stored on the coprocessor), the host offloads another n rows of matrix A and n columns of matrix B to the coprocessor simultaneously using a second OpenMP thread. This process continues until every block of C has been computed. Alongside this communication/computation overlapping model, offloading CPUs performed some computation on the host side while waiting for computation to complete on the coprocessor. This host-side workload was limited to 720 rows and columns for all matrix sizes. This optimized dense matrix mulitplication method was implemented into NEMO5 in 2015 with the collaboration of the Intel Parallel Computing Group (PCL), and replaced BLAS ZGEMM routine calls when using the RGF algorithm and when Intel Xeon Phi KNC coprocessors were present.

To test the effectiveness of this tiling method, the NEGF equations were solved on a 20.6 nm Si nanowire of various cross-sections using the RGF algorithm. This test was performed on the Purdue RCAC Conte computing cluster. The operations included depend on the number of blocks in the Hamiltonian, which correspond to the number of atomic layers. For the 20.6 nm Si device, which corresponds to 38 unit cells of diamond lattice Si with a lattice constant of $a = 0.543$ nm (the length at which the crystal lattice repeats), each atomic layer would have a length of ¾. Therefore the block tri-diagonal Hamiltonian and subsequent Green's functions have $l = 152$ square diagonal blocks corresponding to atomic layers. The operations performed

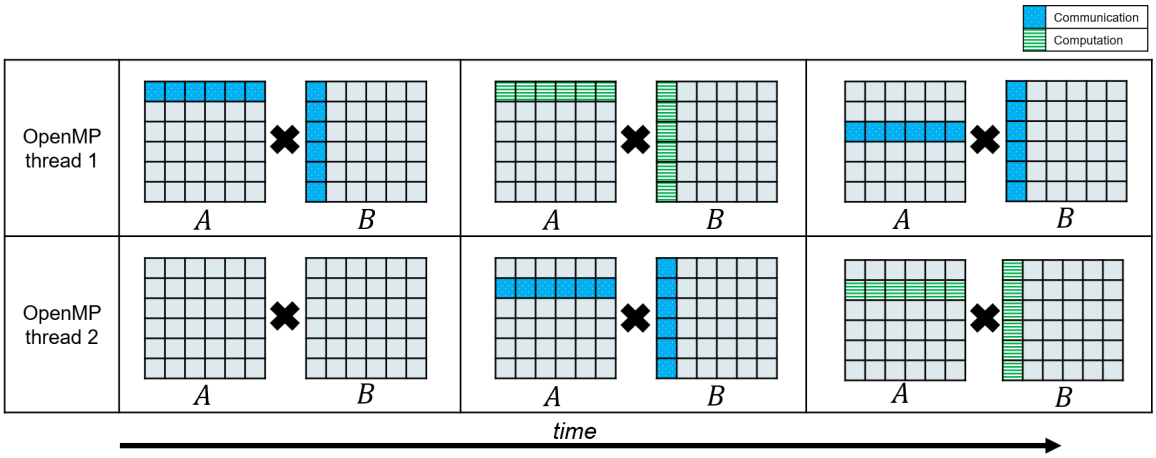


Figure 3.6. Depiction of the custom tiling dense matrix multiplication method in NEMO5 which overlaps communication (sending matrix data to a coprocessor) and computation (computing matrix products on the coprocessor)

in a single RGF calculation are $4l - 1$ dense matrix multiplications, $2l - 2$ sparse-dense-sparse matrix multiplications, l matrix inversions (linear system solve), $l - 1$ sparse-dense matrix multiplications, l diagonal matrix multiplications, l sparse-dense matrix additions, 3 dense matrix additions, 1 matrix trace, and $2l - 1$ dense matrix shifts. Figure 3.7 visually depicts the distribution of operations in RGF with blocks of rank 2880.

Figure 3.8 shows the time to solution of the RGF portion of the NEGF calculation (excluding all other portions of NEGF such as the Poisson equation calculation) with all dense matrix multiplications performed by the optimized, tiled and offloaded ZGEMM routine. Operations were performed with 16 MPI processes on either a 16-core CPU with 16 available threads, or as a hybrid model with both 16 cores (16 threads) and 2 coprocessors (480 threads) used to solve the equations. As expected, the larger block size 2880 $\times$ 2880, corresponding to a cross-section of 12 $\times$ 12 unit cells or 6.52 nm $\times$ 6.52 nm, obtains the greatest speedup from offloading to the coprocessor, since these are ideal for a highly parallelizable system. A speedup of

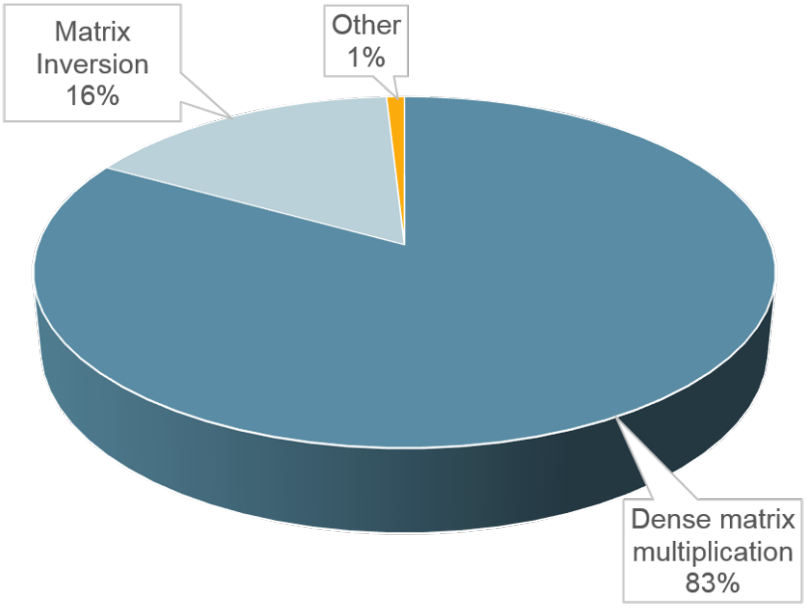


Figure 3.7. Distribution of mathematical operations in RGF with blocks of rank 2880

about 2.8 times was obtained for this cross-section. Also expected is the result of the smallest block size 320 × 320, corresponding to a cross-section of 4 × 4 unit cells or 2.17 nm × 2.17 nm. This case is too small to benefit from a highly parallel coprocessor.

3.7 Outcomes of heterogeneous computing work

From the results shown in this chapter, the iNEMO group concluded that the resulting performance improvements obtained for very specific devices may not be worth the amount of optimization required to achieve such improvements. At the time of these tests, the Intel Xeon Phi Knight's Corner (KNC) coprocessor was relatively new, having only been released a year prior. Many supercomputers migrated to systems that contained KNC coprocessors, but since 2016 have migrated to less heterogeneous Knight's Landing (KNL) systems, which also have a Many Integrated

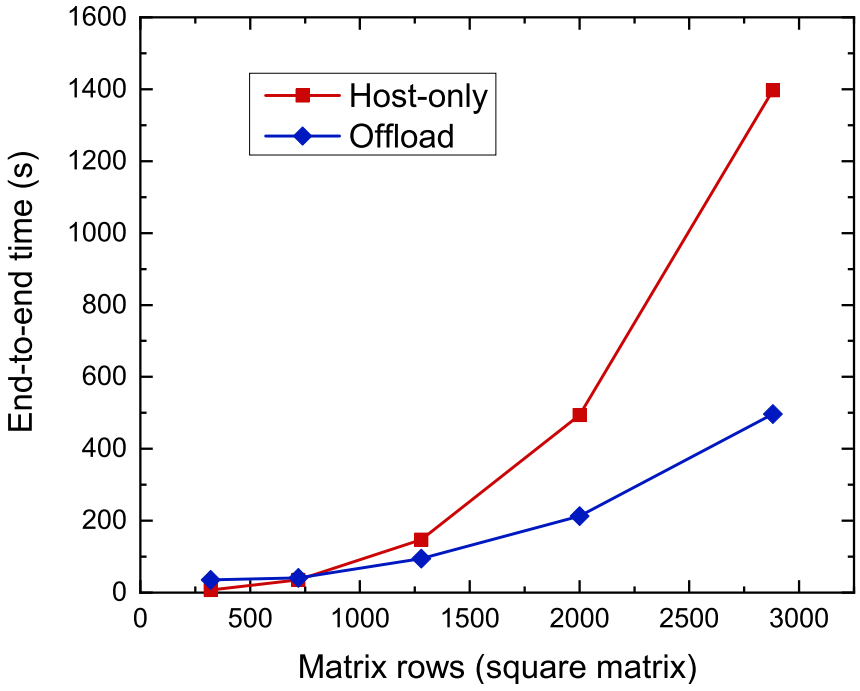


Figure 3.8. Intel Xeon Phi Compiler Assisted Offload performance improvement for various matrix sizes. The largest speedup obtained was of about 2.8 times. This speedup was obtained with a highly optimized dense matrix multiplication routine in collaboration with the Intel Parallel Computing Lab (PCL)

Core (MIC) architecture, but are often treated as separate nodes from a typical CPU node. One example is the Stampede2 supercomputer, which hosts 4,200 KNL nodes and 1,736 Skylake CPU nodes separately [127]. This migration from heterogeneous Intel Xeon Phi coprocessors may have been prompted by less-than-optimal speedups from a large amount of work by scientific groups around the world, including the iNEMO group.

A positive outcome of this research was aiding the Intel Numerical Device Modeling group, with whom the iNEMO closely collaborated with, to decide on requesting a purely homogeneous compute cluster with 30,000 cores rather than a heterogeneous system with Intel Xeon Phi KNC coprocessors [129].

4. HIGH PERFORMANCE COMPUTING

The degree of parallelism provided by the NEGF equations allows them to be solved in a highly parallel computing environment. Supercomputers such as Blue Waters [130] and Stampede2 [127] provide the computational power to solve these equations using multiprocessing, multithreading and sometimes heterogeneous computing to solve these equations on up to hundreds of thousands of CPU cores.

In this chapter, the physics of a realistic UTB device are tested by use of some of the world's most powerful supercomputers: Stampede, Blue Waters, Stampede2, and Tianhe-2. NEMO5's boundaries of scalability are pushed by simulating this device on up to 356,352 CPU cores.

4.1 Parallelism in the NEGF equations

The NEGF equations must be solved for many energies and in some cases, many k-points. The number of energies and k-points depends on device geometry and electrostatic configuration, but often reaches the range of up to a thousand energies and hundreds of k-points. With the nanoelectronic modeling software suite NEMO5 [5, 6, 131], each MPI process can solve the NEGF equations for a minimum of a single energy-k-point ((E, k) point) tuple. For simulations with a single confinement direction and reciprocal space (k-space) such as those used for modeling UTB devices, this may mean parallelism of up to hundreds of thousands of processes is available to a NEMO5 user. Most simulations shown in this thesis were performed on a device with two confinement directions and no k-space. This type of device is often called a nanowire. Although a smaller degree of parallelism of under a thousand processes is available for this type of device, the OpenMP [132] multithreading environment is available for use in NEMO5. This multithreading environment is most often used

for mathematical operations on matrices, often through the linear algebra packages BLAS and LAPACK [133].

Due to the high degree of computational resources needed to solve the NEGF equations, particularly with incoherent scattering, atomistic simulations in a tight binding basis (without reductions or approximations) must be solved on a supercomputer with a large number of nodes and memory, such as Blue Waters [130] and Stampede2 [127]. Efforts have been made in optimizing performance of NEMO5 for high performance computing (HPC). This includes scaling on up to hundreds of thousands of CPU cores and heterogeneous computing by offloading to coprocessors.

4.2 The Gordon Bell Prize Competition

The Gordon Bell Prize is an annual award presented by the Association for Computing Machinery to the most innovative software application that uses state-of-the-art parallel computing technology [124]. Its purpose is to keep track of progress in parallel computing each year. In this competition, various groups from many science and computing fields compete for a $10,000 prize and, more importantly, recognition. The Gordon Bell Prize is the most prestigious high performance computing award, and NEMO5 was entered because winning, being a finalist, or even just competing in the competition would give NEMO5 visibility in the computing world. Although NEMO5 is currently known to be one of the few go-to NEGF-based atomistic device modeling tools, the Gordon Bell Prize competition was an opportunity to show the HPC world how scalable NEMO5 is when performing complex calculations and producing scientifically relevant physical predictions. Another reason for submitting NEMO5 to the Gordon Bell Prize competition was proving scalability. When requesting access to large supercomputers, proof of scalability is requested, and competitively scalable software is the ideal use of highly parallel computing resources. NEMO5, as shown in chapter 3, is capable of efficient heterogeneous computing, a resource available in many of the largest supercomputers through GPUs and Intel Xeon

Phi Knight's Corner (KNC) and Knight's Landing (KNL) coprocessors [127, 130]. In 2014, the year prior to the submission of NEMO5, 2 of 5 Gordon Bell finalists used a heterogeneous supercomputer: Titan with GPUs and Tianhe-2 with Intel Xeon Phi coprocessors [124].

In 2015, NEMO5 was entered as a submission to the Gordon Bell Prize competition by simulating incoherent scattering with a large degree of communication [25]. The largest simulation was performed on the Tianhe-2 supercomputer at the National Supercomputer Center in Guangzhou, China on 356,352 cores. The specific publication submitted can be found as reference [25]. In this chapter, motivation for the need for highly parallel supercomputing resources is presented, as well as the high degree of scalability of NEMO5 in solving the NEGF equations.

4.3 Computational burdens from alloy disorder and k-space

As mentioned in chapter 2, today's devices feature a countable number of atoms, therefore simulations must include detailed calculations at a subatomic resolution to include realistic device physics. At this scale, generalized material properties are insufficient since device imperfections such as alloy disorder [33–35], varying dopant distributions and roughness are present and affect device performance [31, 32]. Phonons are present at any finite temperature in addition to these these imperfections, and must also be modeled [11, 12, 27, 28, 30, 34, 59–63]. All of these effects are modeled in NEGF with scattering by the inclusion of scattering self-energies, which broaden predictions of energy and momentum of electrons in transport. Reliable predictions of device performance must include a consistent consideration of these effects.

The device considered for this chapter is a double-gate ultra-thin-body (UTB) transistor. Figure 4.1 shows an example of this device with a randomized alloy disorder. The UTB device used in this work was 28 nm (precisely 28.23 nm) in length and 3 nm (precisely 3.26 nm) in width. Unlike a nanowire device, UTB devices are periodic along one of the directions perpendicular to the transport direction. Because

of this, UTB simulations often have a thickness of a single unit cell which is repeated periodically along that direction. However, due to the need to keep alloy disorders as random as possible, device thickness in this work was extended beyond the usual single unit cell, as thicker devices along the periodic direction more accurately represent realistic randomness [25]. In the device simulated in this chapter, that thickness ranged from 6 to 10 Si unit cells, or 3.26 nm to 5.43 nm. This results in a device similar to a nanowire with a rectangular cross-section with dimensions of 6 $\times$ 6 unit cells to 6 $\times$ 10 unit cells, or 3.26 nm $\times$ 3.26 nm to 3.26 nm $\times$ 5.43 nm. The device material is made up of 90% Si and 10% Ge, and atom properties are modeled using the virtual crystal approximation (VCA) [96], which creates fictitious atoms with 90% Si and 10% Ge properties. The basis of the device was an $sp3d5s^*$ empirical tight binding basis, where each atom hosts 10 orbitals. Inelastic optical phonon scattering and elastic acoustic phonon scattering were included in these simulations, as well as a self-consistent Poisson equation solution.

As mentioned in section 2.2.5, the inclusion of scattering and the solution of the Poisson equation means that two layers of self-consistent loops must be solved until both converge. In addition to the computational burden introduced by self-consistency, alloy disorder simulations include randomness when the disorder is included explicitly (without VCA), so a statistical assessment would need to be performed with the data output from at least 100 cases [25]. In this assessment, only a single case was run with the VCA approximation, but an ideal design model would include explicit Ge atoms and over 100 statistical cases. In addition to the alloy disorder of the device, the general lead method of reference [34] allows for the inclusion of alloy disorder in the leads, as well as any other roughness, impurities and contact shapes. This general lead method was the method of solving the contact self-energies for the simulations of this chapter. The method involves a recursive RGF-like solution of many layers of the contact material and results in a single surface Green's function for each contact [34].

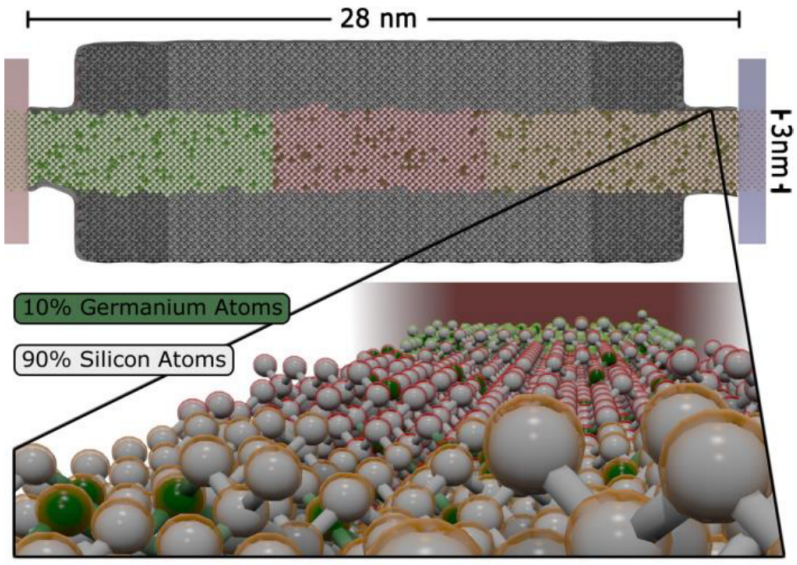


Figure 4.1. 3D rendering of a UTB device for which the NEGF equations are solved in this chapter. The image shows alloy disorder of Ge atoms in a Si material as well as surface roughness in the inset. *Material from: R. Andrawis, J. D. Bermeo, J. Charles, J. Fang, J. Fonseca, Y. He, G. Klimeck, Z. Jiang, T. Kubis, D. Mejia, D. Lemus, M. Povolotskyi, S. A. P. Rubiano, P. Sarangapani, and L. Zeng, 'NEMO5 : Achieving High-end Internode Communication for Performance Projection Beyond Moore's Law,' 2015 Gordon Bell Prize Submission, 2015.*

The solution of a nanowire, with its confinement in all directions perpendicular to the transport direction, would require several hundred energy points to be solved with NEGF. When modeling a UTB device the inclusion of periodicity, and therefore a momentum- or k-space in a direction perpendicular to transport, requires the solution of all combinations of energies and k-points. This creates a grid of (E, k) tuples that require solving. In the case of the device of this chapter, up to 16,000 (E, k) tuples required solving which allowed for 16,000 parallel MPI processes to run in parallel, thus increasing our need for massively parallel computing systems. Although the solution of these many (E, k) tuples for UTB devices usually involves the solution of many small matrices due to their single-unit-cell thickness, alloy disorder calls for the solution of a larger thickness along the periodic direction, resulting in block

matrices of a higher rank being solved in the RGF algorithm. This combination of the requirement to perform math on large matrices and many (E, k) tuples necessitates a large degree of both local parallelism (multithreading) and MPI parallelism.

4.4 Communication in NEGF equations

Although highly parallel in nature, the solution of the NEGF equations with scattering could not be labeled "embarrassingly parallel" because of the need for communication with every inelastic scattering self-energy solution. Due to the presence of inelastic scattering, electrons may be shifted unpredictably from one energy to another, requiring communication to occur between processes which must modify the Green's functions that describe electron occupancy for specific energies [8,25,68]. NEMO5 determines scattering communication patterns during the solution of the scattering self-energies. These depend on the unpredictable broadening effects of electrons scattering on phonons [8, 25]. The resulting communication is therefore greatly influenced by the type of scattering model and scattering strength. As mentioned in section 2.2.4, the complex and overlapping communication patterns of self-energy calculations must be properly managed in order to avoid excessive idling that diminishes scaling capabilities. The sorting algorithm described section 2.2.4 and blocking communication were thus used for these simulations to improve scaling capabilities. In addition to scattering calculations, the calculation of charge density requires communication among all processes to add the total charge for all energies. The communication pattern for this communication event is much more predictable than scattering communication events, as it involves all processes and occurs only once per Poisson iteration.

4.5 Scaling results on supercomputers

In HPC, there are two ways of measuring the scalability of software: strong and weak scaling:

Strong scaling is defined as the capability of a problem to scale to an increasingly large degree of resources while maintaining a fixed problem size. For example, for an embarrassingly parallel problem, the number of processes p would decrease the time to solution for the number of required tasks n proportionally. In NEGF, these n tasks would each correspond to a (E, k) tuple, and the time to solution of n fixed (E, k) tuples would ideally be decreased by a factor of p. Realistically, however, this is not possible due to existing sequential sections of the algorithm and the need to communicate at each scattering iteration. Figure 4.2 shows the results of a strong scaling test that was performed on the Stampede supercomputer on up to 32,768 cores. The single Intel Xeon Phi KNC coprocessor on each of the nodes of this machine was also used. The simulation run on Stampede was the 3.26 nm $\times$ 5.43 nm UTB device, and the operation performed was a single scattering iteration.

Weak scaling is defined as the capability of a problem to scale to an increasingly large degree of resources while also increasing the problem size proportionally. For an embarrassingly parallel problem, the time to solution would remain identical for all cases. In NEGF, n tasks would correspond to (E, k) tuples and the ideal time to solution t for n tasks on p processes would be $t(n) = \alpha p$ where α is constant. Realistically, increasing the parallelism of a problem introduces finite communication time due to less-than-ideal communication patterns and imperfect load balancing. Figure 4.3 shows the weak scaling trend on the Tianhe-2 supercomputer. The simulation run on Tianhe-2 was the 3.26 nm $\times$ 3.26 nm UTB device due to walltime limitations, and the operation performed was a single scattering iteration.

4.6 Outcomes of high performance computing work

The most tangible outcome of the HPC work was the submission of reference [25] to the Gordon Bell Prize competition of 2015. And although NEMO5 was not nominated to win the Gordon Bell Prize, an arguably more important outcome was achieved, which was the gradual optimization and improvement of the parallel capabilites of

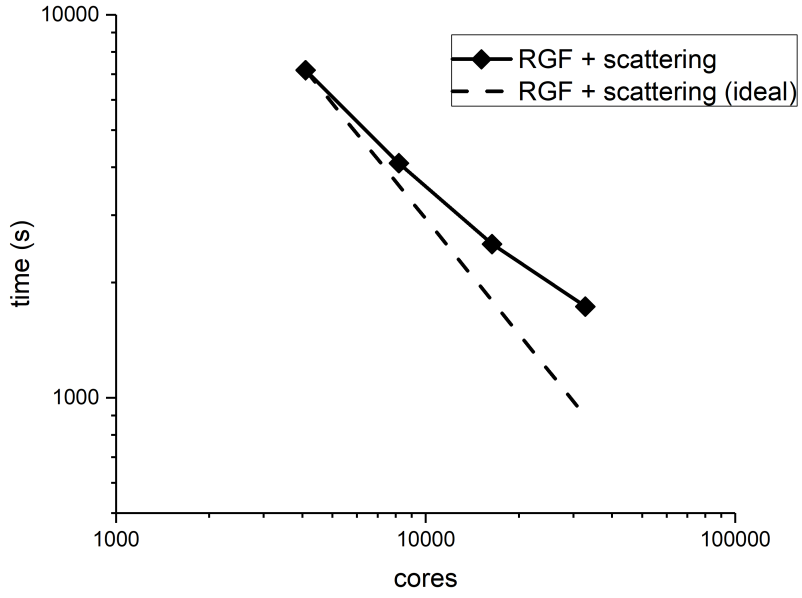


Figure 4.2. Strong scaling of a scattering simulation in NEMO5 on up to 32,768 cores on the Stampede supercomputer

NEMO5. Great improvement was seen in the months of work leading up to the Gordon Bell Prize submission in 2015, and NEMO5 was proven to be highly scalable given how complex the communication patterns of inelastic scattering are.

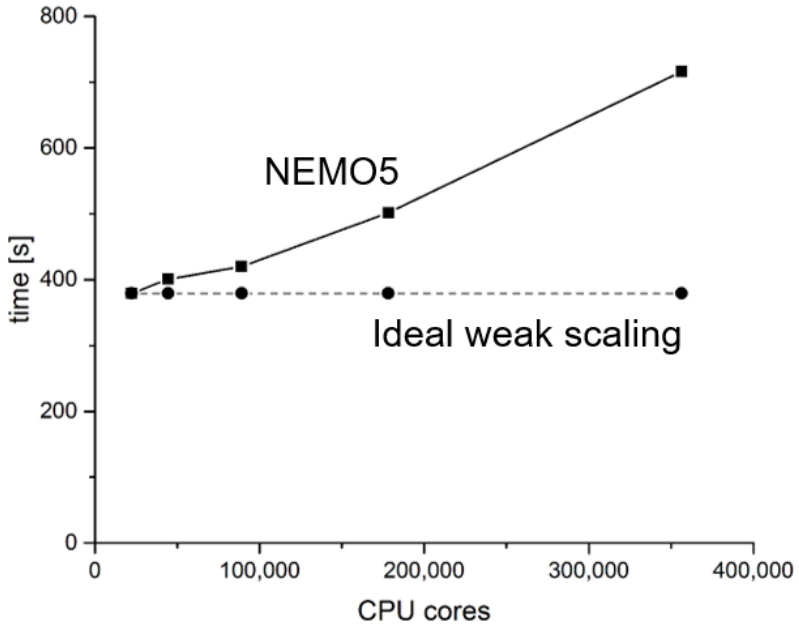


Figure 4.3. Weak scaling of a scattering simulation in NEMO5 on up to 356,353 cores on the Tianhe-2 supercomputer

5. LOW-RANK APPROXIMATIONS IN NEGF

Opposite to the method of using a large number of computational resources to solve the NEGF equations lies the method of reducing the computational burden of these equations using approximations. Approximations to physics, when properly used, are capable of reducing matrix sizes in the NEGF equations, therefore reducing the need for computational resources greatly while maintaining accurate physics.

In this chapter the mode space method is shown from its simplest form to its most complex form that is capable of incoherent scattering. The recursive Green's function method with mode space approximations is detailed, along with modifications to the scattering self-energies which allow for their calculation in mode space.

5.1 Mode space approach for basis reduction

The dispersion relation of a simulated system represents the available electronic states for various energy-momentum, (E, k), configurations of the system. The dispersion relation, often shown via a band diagram or band structure, represents the various ways that an electron can propagate along the crystal of devices such as silicon nanowires. Often a very complex relation in realistic bases such as tight binding, in the effective mass approximation, the dispersion relation for semiconductor devices can be approximated as

$$E(k) = E_0 + \frac{\hbar^2 k^2}{2m*}, \tag{5.1}$$

which is parabolic in nature [134]. Here, $\hbar$ is Planck's constant and k is the wave vector such that the momentum $p = \hbar k$. E_0 is the band-edge, which corresponds to the vertex of the parabola of the effective mass approximation. m^* is the effective

mass, which is the mass of an electron in a given material. This effective mass is a material parameter and corresponds to the shape of the band structure.

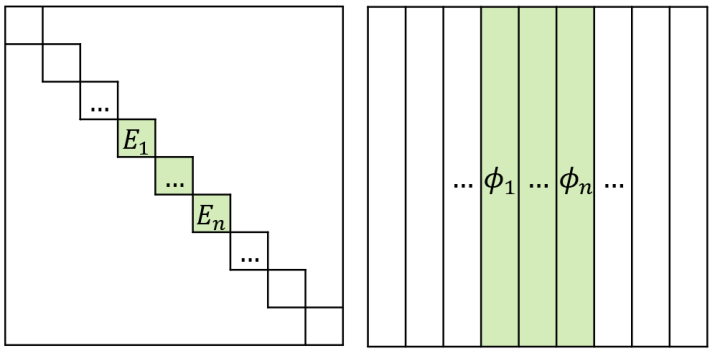
Due to the simplicity of an effective mass system, methods such as the mode space method [104, 116] can reduce the available band states of the system, thus reducing the rank of the system Hamiltonian dramatically by removing states that are unlikely for an electron to occupy. For example, if a material is known to be more likely to host electrons within a range of energies, choosing the eigenvectors of the Hamiltonian by their corresponding energies near the band-edge will provide a transformation matrix that can be used to reduce the rank of the device Hamiltonian such that

$$h = \Phi_{eig}^{T} H \Phi_{eig} \tag{5.2}$$

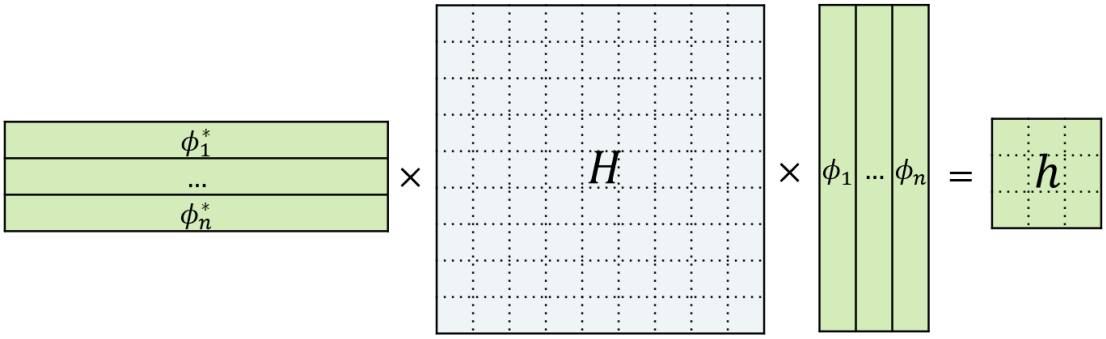
where H is the initial device Hamiltonian, Φ_{eig} is a rectangular matrix constructed from the desired eigenvectors, and h is the reduced Hamiltonian. Figure 5.1 offers a depiction of the chosen eigenvalues E_i, their corresponding eigenvectors ϕ_i, and a transformation from H to h with the resulting transformation matrix.

5.2 Low-rank approximations in atomistic tight binding basis

In a realistic tight-binding basis, using the mode space method directly to reduce the rank of the system Hamiltonian is not possible due to the appearance of spurious states that create non-existent electronic band states and cause deviations from experimental data [7]. An example of spurious states is shown in figure 5.2, where figure 5.2.a shows the states obtained when using equation 5.2 directly. The method developed by Mil'nikov et. al. [7] provides a solution for removing these spurious states. The first step of the method is to obtain the eigenvectors ϕ_i from the desired energy interval $[E_1, E_2]$, similarly to the traditional mode space approach. The full basis Hamiltonian H is transformed to a reduced (mode space) basis h using a transformation matrix Φ_{eig} constructed from the eigenvectors ϕ_i as shown in equation 5.2. At this stage, the reduced Hamiltonian h yields several unphysical states. A modified



(a) Chosen eigenvalues E_i and eigenvectors ϕ_i



(b) Transformation of Hamiltonian with chosen eigenvectors

Figure 5.1. Depiction of (a) the eigenvalues E_i and corresponding eigenvectors ϕ_i chosen to represent the reduced basis and (b) a transformation of a full basis Hamiltonian H to a reduced basis Hamiltonian h

reduced Hamiltonian $\widetilde{h}$ is created by adding new orthogonal basis states $\widetilde{\Phi}$ such that $\Phi^T\widetilde{\Phi} = 0$ and

$$\widetilde{h} = \begin{vmatrix} h & X \\ X^\dagger & H_{\widetilde{\Phi}\widetilde{\Phi}} \end{vmatrix} \tag{5.3}$$

where

$$X = \Phi_{eig}^T H \widetilde{\Phi}. \tag{5.4}$$

The added states $\widetilde{\Phi}$ do not deteriorate the basis and have no effect on non-spurious band states due to the already complete basis of h. The purpose of the new state $\widetilde{\Phi}$ is to remove the spurious states, so $\widetilde{\Phi}$ must be chosen carefully such that it reduces the

number of states in the band structure. Since adding states to the modified reduced Hamiltonian h cannot modify correct physical states [7], the correct solution is such that $\widetilde{h}$ generates the fewest band states.

To find this, we first introduce a function of energy

$$N(E) = \left\langle \frac{z - E_c}{z - E} \right\rangle = \frac{1}{2n_z} \sum_{j=1}^{2n_z} . \frac{z_j - E_c}{z_j - E} \tag{5.5}$$

that gives $\sim$1 for energies within the window of E_1 and E_2 and $\ll$1 otherwise. Therefore a sum of $N(E)$ for various energies E would give the number of states in a system. Here, $E_c = (E_1 + E_2)/2$ and $z_j = E_c + \rho e^{\frac{i\pi}{n_z}\left(j - \frac{1}{2}\right)}$. The sum occurs on a complex contour in the complex z plane with $2n_z$ points, center E_c, and radius $\rho = (E_2 - E_1)/2$. This function is used in the functional

$$\begin{aligned} F[\widetilde{\Phi}] &= \sum_{i=1}^{n_k} \sum_{\nu} N\left(E_\nu\left(k_i, [\widetilde{\Phi}]\right)\right) \\ &= \left\langle \sum_i \mathrm{Tr}\left[\frac{1}{z - \tilde{h}\left(k_i\right)}\right](z - E_c) \right\rangle \end{aligned} \tag{5.6}$$

where $\widetilde{h}$ is given by equation 5.3. This functional provides the total number of states at n_k wave numbers k_i, which each correspond to a chosen set of values in k-space. $F[\tilde{\Phi}]$ is equal to the the original number of states F_0 (without $\tilde{\Phi}$) plus the change to the number of states, or cost function $\Delta F[\tilde{\Phi}]$:

$$F[\widetilde{\Phi}] = F_0 + \Delta F[\widetilde{\Phi}]. \tag{5.7}$$

The next step is to find $\tilde{\Phi}$ such that

$$\widetilde{\Phi} = \frac{1}{\sqrt{C^T C}} \Xi C \tag{5.8}$$

and the following cost function ΔF is minimized:

$$\Delta F(C) = \frac{1}{2n_z} \sum_{i=1}^{n_k} \sum_{j=1}^{2n_z} \frac{C^T A\left(k_i, z_j\right) C}{C^T B\left(k_i, z_j\right) C} \left(z_j - E_c\right) + \left(C^T C - 1\right)^2 \tag{5.9}$$

where Ξ is obtained by orthogonalizing the columns of the matrix [7, 107]

$$\left[\left(1 - \Phi\Phi^T\right) H(k=0)\Phi, \left(1 - \Phi\Phi^T\right) H(k=\pi)\Phi\right] \tag{5.10}$$

and obtaining M' columns. C is a vector of dimension M' which contains the expansion coefficients of Ξ. Matrices A and B are

$$A(k,z) = I_{M' \times M'} + \Xi^T H(k) \Phi [z - h(k)]^{-2} \Phi^T H(k) \Xi, \tag{5.11}$$

$$B(k,z) = z I_{M' \times M'} - \Xi^T H(k) \Xi - \Xi^T H(k) \Phi [z - h(k)]^{-1} \Phi^T H(k) \Xi. \tag{5.12}$$

This method is repeated until no new (E, k) states have been added when adding a state $\tilde{\Phi}$ to the basis Φ, signaling that the cost function ΔF has been minimized. From this point forward in the thesis, the final mode space transformation matrix that results from the removal of spurious states will be denoted simply as Φ. Figure 5.2 shows the evolution of the mode space band structure, along with the "correct" full basis tight binding band structure for a 2.17 nm $\times$ 2.17 nm $sp3d5s^*$ device. Note that $E_1 = E_c + 0.5$ eV and $E_2 = E_v - 0.5$ eV. After 90 iterations, band states match well within this energy window $[E_1, E_2]$.

5.3 Generation of basis states in NEMO5

Generation of mode space basis states is provided by the ModeSpace solver in the NEMO5 software suite [5, 6, 131]. This library-like module of NEMO5 provides a basis that can be used for matrix transformations which may reduce matrix ranks

down to 10% of their original size or lower. This has been used in the past to obtain speedups for ballistic simulations of up to 10,000 times [107, 108]. The ModeSpace solver uses the method described in section 5.2. The solver interface includes controls for the lower and upper bounds of the energy window of interest $[E_1, E_2]$. It then attempts to reduce the number of spurious states in the band structure. This process can typically take a few minutes on devices with small cross-sections, especially when only a small energy range is required. For cross-sections greater than 5 nm $\times$ 5 nm, however, the process of obtaining a reduced basis with a matching band structure may take hours, and multiple attempts may be needed to obtain a suitable basis. Reference [107] details some improvements to the algorithm of reference [7], including MPI parallelization and a method of detecting spurious states.

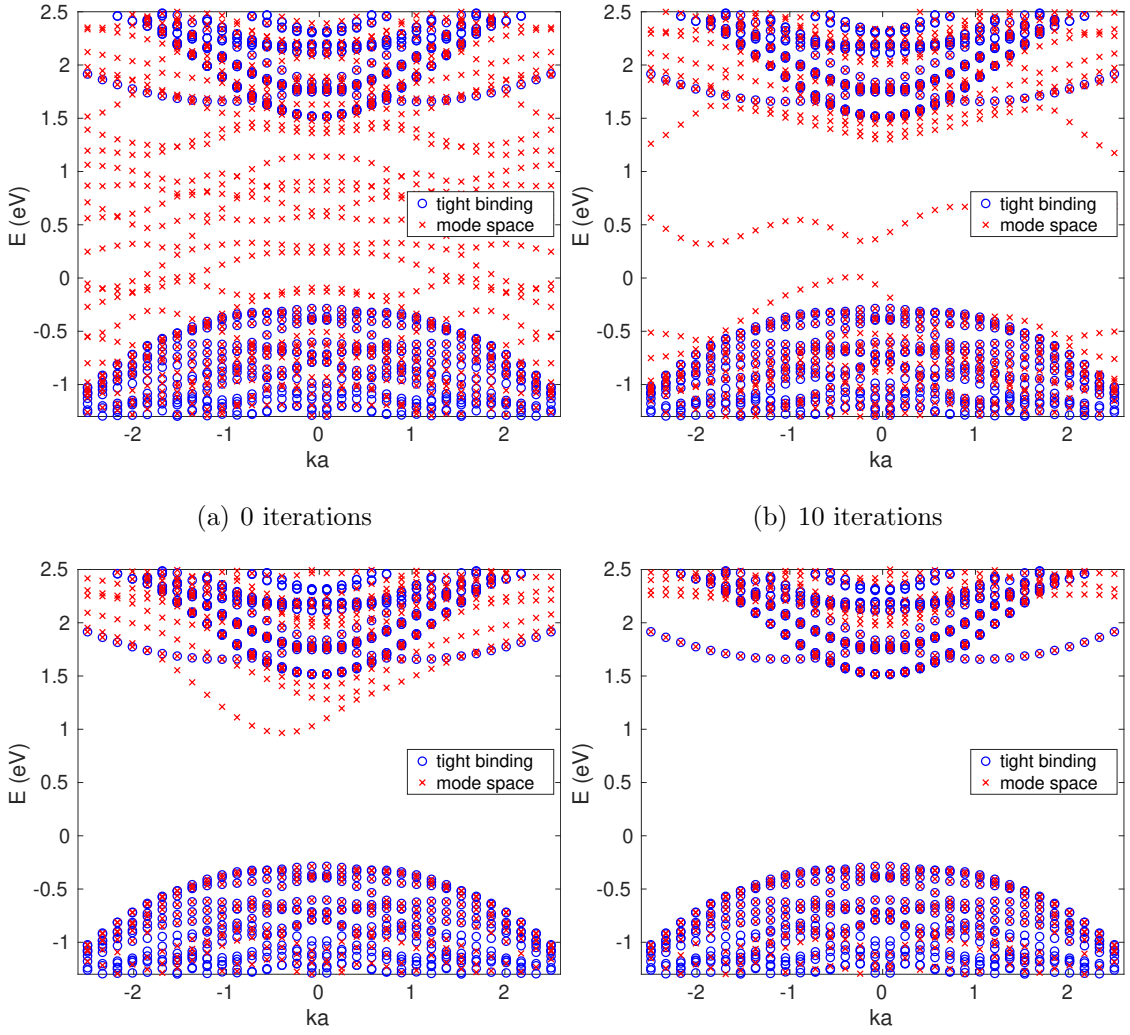
5.4 RGF method and LRA application

After the transformation matrix Φ is created, the blocks of the block tri-diagonal device Hamiltonian H are transformed into a reduced device Hamiltonian h using the equation

$$h_{I,J} = \Phi^T H_{I,J} \Phi \tag{5.13}$$

for block indices I and J. Then the RGF algorithm may be performed on reduced matrices.

The calculation of the Green's functions G^R using the RGF algorithm [9, 135] can be divided into two main steps: the "forward" calculation in which the blocks of G^R are solved recursively from the top-left block to the bottom-right which results in a single-block-sized matrix $g^R_{L,L}$ for L device layers, and the "backward" calculation which, from the bottom-right to the top-left recursively creates the resulting matrix G^R.



(a) 0 iterations

(b) 10 iterations

(c) 45 iterations

(d) 90 iterations

Figure 5.2. Evolution of mode space band structure compared to full basis $sp3d5s^*$ tight binding of a 2.17 nm $\times$ 2.17 nm cross-section Si device. Energy window was set to a range 0.5 eV above conduction band edge to 0.5 eV below valence band edge

In the standard block tri-diagonal and local RGF algorithm [9, 135], the forward Green's function block $g^R_{I,I}$ for device layer and block index I is calculated using the equation

$$g^R_{I,I} = \left(A_{I,I} + h_{I,I-1} g^R_{I-1,I} h_{I-1,I}\right)^{-1} \tag{5.14}$$

where

$$A_{I,I} = \left(E - h_{I,I} - \Sigma^R_{I,I}\right)^{-1}. \tag{5.15}$$

Note that the Hamiltonian h is the reduced version, and has a reduced rank n according to the number of modes found by the method shown in section 5.2. The forward Green's function for the first layer and block index $I = 1$ is

$$g^R_{1,1} = \left(A_{1,1}\right)^{-1}. \tag{5.16}$$

Σ^R is an addition of scattering self-energies and contact self-energies when they exist at index I. The calculation of the contact self-energies in a reduced basis is straightforward after a reduction of the Hamiltonian [7], however, calculating the contact self-energy portion in a reduced basis is not trivial and will be discussed in section 5.5.

Continuing into the "backward" portion of RGF for L total layers,

$$G^R_{L,L} = g^R_{L,L} \tag{5.17}$$

and

$$G^R_{I,I} = g^R_{I,I} + g^R_{I,I}\left(h_{I,I+1} G^R_{I+1,I+1} h_{I+1,I}\right) g^R_{I,I}. \tag{5.18}$$

The lower ($G^R_{I+1,I}$) and upper ($G^R_{I,I+1}$) offdiagonal blocks must also be calculated, since they are needed for the calculation of $G^<$:

$$\begin{aligned} G^R_{I+1,I} &= -G^R_{I+1,I+1} h_{I+1,I} g^R_{I,I}, \\ G^R_{I,I+1} &= -g^R_{I,I} h_{I,I+1} G^R_{I+1,I+1}. \end{aligned} \tag{5.19}$$

The lesser Green's function $G^<$ can similarly be calculated recursively, with $g^<_{I,I}$ for layer and block index I being

$$\begin{aligned} g^<_{I,I} = g^R_{I,I} \Big[&\Sigma^<_{I,I} + A_{I,I-1} g^<_{I-1,I-1} A^\dagger_{I-1,I} \\ &+ \Sigma^<_{I,I-1} g^A_{I-1,I-1} A^\dagger_{I-1,I} + A_{I,I-1} g^R_{I-1,I-1} \Sigma^<_{I-1,I} \Big] g^A_{I,I} \end{aligned} \tag{5.20}$$

and

$$g^<_{1,1} = g^R_{1,1} \Sigma^<_{1,1} g^A_{1,1} \tag{5.21}$$

for layer and block index $I = 1$. "Backward" RGF for $G^<$ is then performed with

$$G^<_{L,L} = g^<_{L,L} \tag{5.22}$$

for the last layer L and

$$\begin{aligned} G^<_{I,I} = &g^<_{I,I} + g^<_{I,I} A^\dagger_{I,I+1} G^A_{I+1,I} \\ &+ g^R_{I,I} \Sigma^<_{I,I+1} g^A_{I+1,I+1} A^\dagger_{I+1,I} G^A_{I,I} + g^R_{I,I} A_{I,I+1} G^<_{I+1,I} \end{aligned} \tag{5.23}$$

for the diagonal blocks. The mode space LRA approach by Mil'nikov et. al. [7] was applied to tight binding bases, and is included in the NEMO5 software suite [107, 108]. The RGF method [9] with mode-space-reduced tight binding bases shown in section 5.4 with incoherent scattering is performed in this work, and physics and performance results will be shown in chapter 6.

5.5 Expanding low-rank approximations to incoherent scattering simulations

Calculation of scattering self-energies requires real-space information. Real space is represented by the vectors **r** and **r'** in the lesser scattering self-energy

$$\Sigma^{<}(\mathbf{r},\mathbf{r'},E) = \frac{1}{(2\pi)^3}\int d\mathbf{q}|U_q|^2 e^{i\mathbf{q}(\mathbf{r}-\mathbf{r'})}[n_q G^{<}(\mathbf{r},\mathbf{r'},E-\hbar\omega_q) \\ +(1+n_q)G^{<}(\mathbf{r},\mathbf{r'},E+\hbar\omega_q)] \tag{5.24}$$

where $\mathbf{q}$ are phonon momentums, E is the electron energy, n_q is the system's Bose distribution, $\hbar$ is Planck's constant, U_q is a constant scattering potential and ω_q is the phonon frequency. Transformation of Green's functions to mode space means that matrices no longer contain position information. One solution to this problem involves the up-conversion of Green's functions before the solution of scattering self-energies like equation 5.24.

5.5.1 The Green's function upconversion method

After calculating Green's functions in a reduced mode space basis, the transformation matrix Φ can be used to upconvert each block at block indices I and J of a Green's function

$$G^{R,<}_{I,J,full} = \Phi G^{R,<}_{I,J,MS}\Phi^T \tag{5.25}$$

after which the self-energies $\Sigma^{R,<}$ are solved in the full (real space) basis by replacing the Green's functions of equation 5.24 with $G^{R,<}_{full}$, then converted back into mode space in a block-wise fashion:

$$\Sigma^{R,<}_{I,J,MS} = \Phi^T \Sigma^{R,<}_{I,J,full}\Phi. \tag{5.26}$$

This, however, involves costly transformations that happen for every iteration of

the self-consistent Born method, the computational cost of which will be shown in chapter 6, table 6.1. This also means that any improvements in memory footprint can be completely eliminated since matrices are restored to their original size in the middle of the calculation.

5.5.2 The form factor transformation method

The way to avoid upconversion of matrices in the middle of the calculation is through the introduction of a form factor transformation as described in reference [116]. The form factor is a four-dimensional tensor that contains an overlap of all available modes integrated in real space transverse to the transport direction of the device.

$$F_{i,j,k,l} = \sum_{\nu} \phi_i(\nu)\phi_j(\nu)\phi_k(\nu)\phi_l(\nu) \tag{5.27}$$

where each index i, j, k, l exists for n modes, or columns of the mode space transformation matrix Φ. The index ν is iterated through to the N rows of Φ. This sum is equivalent to a real space integral for every possible combination of modes. Each element of

$$\Sigma^{R,<}_{\text{acoustic}}(\mathbf{r}, \mathbf{r'}, E) = \frac{D^2 k_B T}{\rho v_s^2} \delta_{\mathbf{r},\mathbf{r'}} G^{R,<}(\mathbf{r}, \mathbf{r'}, E), \tag{5.28}$$

$$\begin{aligned} \Sigma^{<}_{\text{optical}}(\mathbf{r}, \mathbf{r'}, E) = & \frac{\hbar D_{op}^2 k_B T}{2\rho\omega_q} \delta_{\mathbf{r},\mathbf{r'}} \\ & \times \left[n_q G^{<}(\mathbf{r}, \mathbf{r'}, E - \hbar\omega_q) \right. \\ & \left. + (1 + n_q) G^{<}(\mathbf{r}, \mathbf{r'}, E + \hbar\omega_q) \right] \end{aligned} \tag{5.29}$$

and

$$\begin{aligned}\Sigma^{R}_{\text{optical}}(\mathbf{r},\mathbf{r'},E) =& \frac{\hbar D_{op}^2 k_B T}{2\rho\omega_q}\delta_{\mathbf{r},\mathbf{r'}} \\ & \times\Big[(1+n_q)\,G^R(\mathbf{r},\mathbf{r'},E-\hbar\omega_q) \\ & +n_q G^R(\mathbf{r},\mathbf{r'},E+\hbar\omega_q) \\ & +\frac{1}{2}G^<(\mathbf{r},\mathbf{r'},E-\hbar\omega_q) \\ & -\frac{1}{2}G^<(\mathbf{r},\mathbf{r'},E+\hbar\omega_q)\Big]\end{aligned} \tag{5.30}$$

can be defined as $\Sigma_{i,j}$ and each element of a Green's function matrix $G^{R,<}(\mathbf{r},\mathbf{r'},E)$ as $G_{k,l}$. $\mathbf{r}$ and $\mathbf{r'}$ denote perturbations from position $\mathbf{r}$ to $\mathbf{r'}$, which both correspond to a specific atomic orbital. D is the deformation potential constant, D_{op} is the optical deformation potential constant, k_B is Boltzmann's constant, T is temperature, ρ is the density of the material, v_s is the sound velocity in the material, $\hbar$ is Planck's constant, ω_q is the phonon frequency and n_q is the system's Bose distribution. Note that equation 5.30 contains an approximation which will be discussed in section 7.2. For simplicity we define C as the product of all scalar factors involved in each of equations 5.28-5.30. The form factor elements $F_{i,j,k,l}$ are applied to the Green's function elements $G_{k,l}$ as follows:

$$\Sigma_{i,j} = \sum_l \sum_k C F_{i,j,k,l} G_{k,l}. \tag{5.31}$$

The transformation described above avoids upconversion of Green's functions; all matrices remain in their reduced rank for the duration of the simulation, keeping the memory footprint low.

5.5.3 Approximation of form factor

Because the tensor $F_{i,j,k,l}$ is four-dimensional and depends on the dimension n of the reduced basis, its memory footprint scales rapidly on the order of $O(n^4)$, which can become unwieldy for bases of only over 100 modes. The time for construction of $F_{i,j,k,l}$ scales on the order of $O(n^4 N)$, and time for application scales on the order of $O(n^4)$.

A larger number of modes can easily result in the form factor application taking a significant amount of time and memory footprint. Similarly to reference [116], it will be shown in section 6.3 that eliminating offdiagonal elements of the form factor F, such that $F_{i,j,k,l} = 0$ for $i \neq j$ and $k \neq l$, provides reasonable physical results. This approximation corresponds to the lack of interaction between modes. Therefore, no intra-mode scattering takes place when the form factor is diagonal. This provides a much more memory-thin form factor that only contains the "diagonal" (in four-dimensions) elements. Construction complexity of the form factor also is reduced to approximately $O(n^2N)$ and application is reduced to $O(n^2)$. In chapter 6, physical results provided by the approximate form factor are compared to the results of the full form factor as well as the full basis calculation. Also in chapter 6 the time-to-solution with the Green's function upconversion method, full form factor calculation and approximate form factor will be compared.

6. ASSESSMENT OF LOW-RANK APPROXIMATIONS

To ensure that the application of low-rank approximations provides a valid basis for modeling the physics of a nanowire device with incoherent scattering, validation tests were performed with NEMO5 by comparing current-voltage (I-V) characteristic curves for sweeping gate bias voltages using both reduced (mode space) and full basis (tight binding) simulations.

After validation of physics was done, computational performance after basis reductions was assessed and compared to that of the tight binding basis representation. Baseline computational measurements were performed: time to solution and peak memory of a single scattering iteration. Since a production simulation includes more aspects to the solution of the NEGF equations, however, a test with aspects such as the Poisson equation and density calculation is shown, and from this production scale simulations can be projected for both tight binding and mode space representations.

6.1 Simulation setup

The device used for both validation and performance tests was a $w \times w \times 20.65$ nm silicon nanowire as shown in figure 6.1, where w is the variable width in nm of the square cross-section of the device. The device had a 1 nm gate oxide layer surrounding the central region. The original basis was a 10-orbital $sp^3d^5s^*$ tight binding model using the parameter set of reference [136]. A source-drain bias of 0.2 V was applied to the device. Note that the applied source-drain bias does not affect the validity of the presented method, and mode space calculations with higher source-drain voltages can be found in references [107] and [108]. The device was *NIN* doped, with the $s = 5.97$ nm source and $d = 6.66$ nm drain regions having a 10^{20} cm^{-3} doping density and the central $c = 8.02$ nm intrinsic region having a 10^{15} cm^{-3} doping density. The

lengths s, d and c are labeled in figure 6.1. Simulations of Si devices included both inelastic optical phonon and elastic acoustic phonon deformation potential scattering, applied to the NEGF equations through self-energies in the self-consistent Born approximation [8, 60]. The energy grid was generated using an adaptive grid generator in NEMO5.

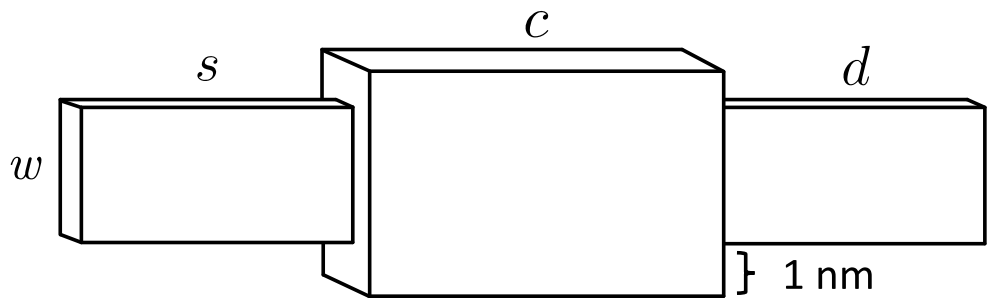


Figure 6.1. Schematic of the nanowire devices considered in this work with a $w \times w$ cross-section and a 1 nm gate oxide layer surrounding the center of the device. s labels the source length, c the channel length and d the drain length of the device. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

6.2 Validation of mode space simulation results

For validation, a silicon nanowire of width $w = 3.26$ nm was used (see figures 6.1 and 6.2). The mode space simulation had a reduction ratio n/N of 2.8%, transforming matrix blocks from 2880 $\times$ 2880 matrices to 81 $\times$ 81 matrices. NEGF was solved using the scattering-compatible RGF algorithm [60]. Figure 6.3 shows the current-voltage (I-V) characteristic curves of both the original tight binding basis and mode space basis for sweeping gate biases ranging from -0.1 V to 0.5 V. The mode space scattering results of figure 6.3 were obtained using the full form factor as described in section 5.5.2. The virtually identical results of mode space and tight binding show that the mode space low-rank approximation provides a valid and highly efficient model

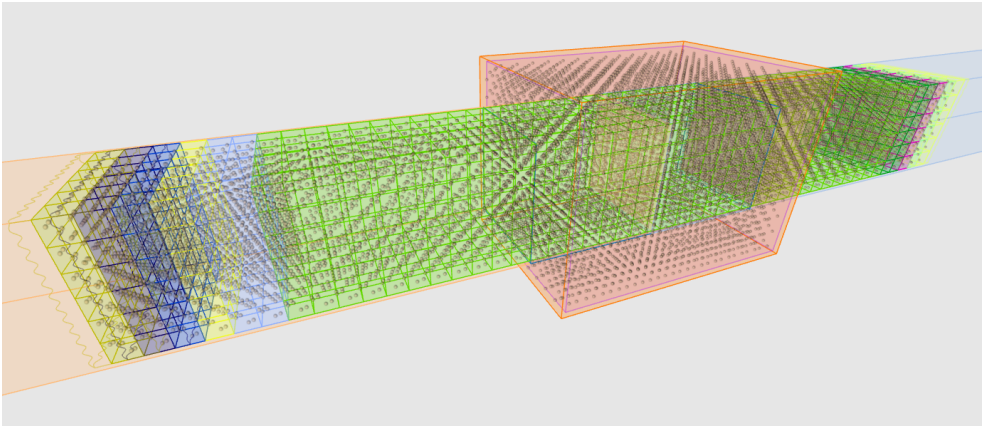


Figure 6.2. 3.26 nm × 3.26 nm × 20.65 nm silicon nanowire device with a gate oxide layer surrounding the center of the device, used to compare physical results for reduced basis simulation with full-basis results. For performance tests, devices similar in geometry to this, but with varying cross-sections, were used. Device structure visualization was generated using the NEMO5 graphical interface NemoViz

for quantum transport simulations with incoherent scattering. Figure 6.4 shows that the mode space approach with approximate form factors, as discussed in section 5.5.3, also yields results very close to those of the original basis calculations. To further justify the use of the approximate form factor, table 6.1 compares its time to solution and peak memory in a single self-consistent scattering iteration to the full form factor and Green's function upconversion method discussed in section 5.5.1. In this table the form factor rows include the form factor generation and application time as discussed in section 5.5.2. This test was performed using the $w = 3.26$ device on 2 MPI processes, 24 OpenMP threads per process, and a total of 4 energies solved. This small number of energies was chosen due to the large memory footprint of the Green's function upconversion method. The OpenMP threads were used to parallelize the generation and application of the form factor elements. From this comparison the benefits of both the full and approximate form factors are immediately evident, as iteration time and peak memory are an order of magnitude larger when using the Green's function upconversion method.

Figure 6.5 shows a contour plot of the potential profile of the center cross-section of the device for a tight binding simulation at the applied gate bias of 0.5 V. Contour lines show the relative absolute error of the mode space potential profile results relative to the original tight binding data. Note that the mode-space method agrees with NEGF calculations in the original tight binding representation for many wire cross-sections as similarly well as those shown in figures 6.3 and 6.4. Similar benchmark data can be found in references [7], [107] and [108].

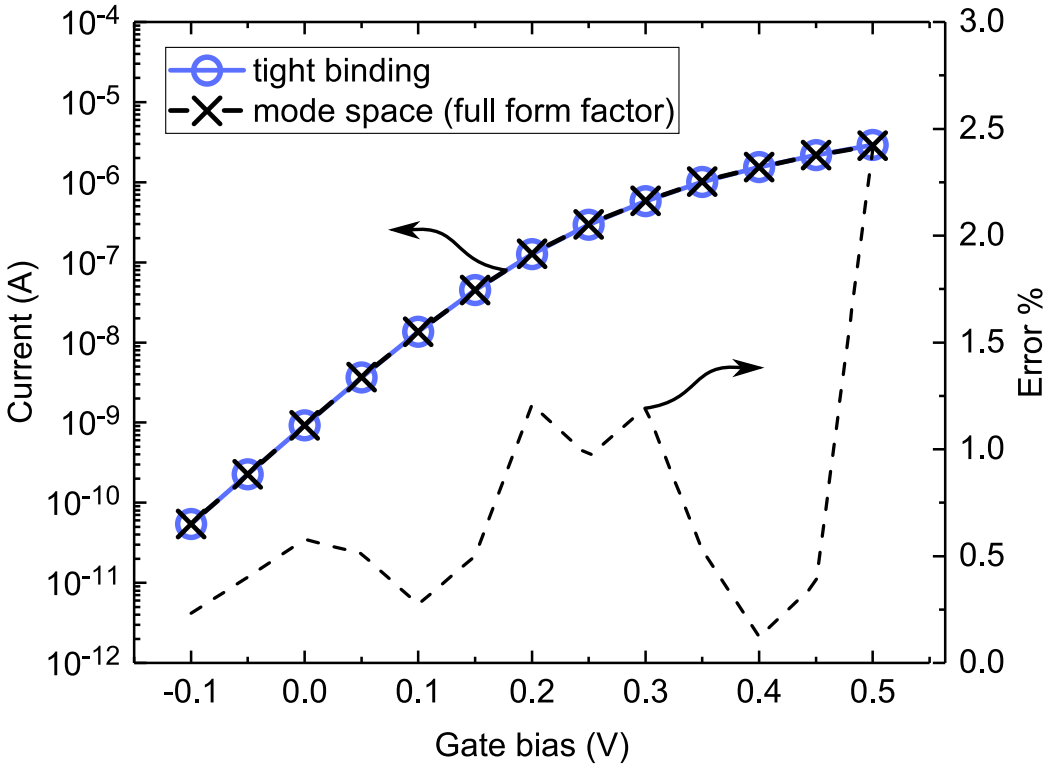


Figure 6.3. Current-gate-voltage (I-V) characteristic curve of a 3.26 nm × 3.26 nm × 20.65 nm silicon nanowire. The agreeing results prove the mode space approach provides a valid physical model. All simulations include inelastic scattering on phonons. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

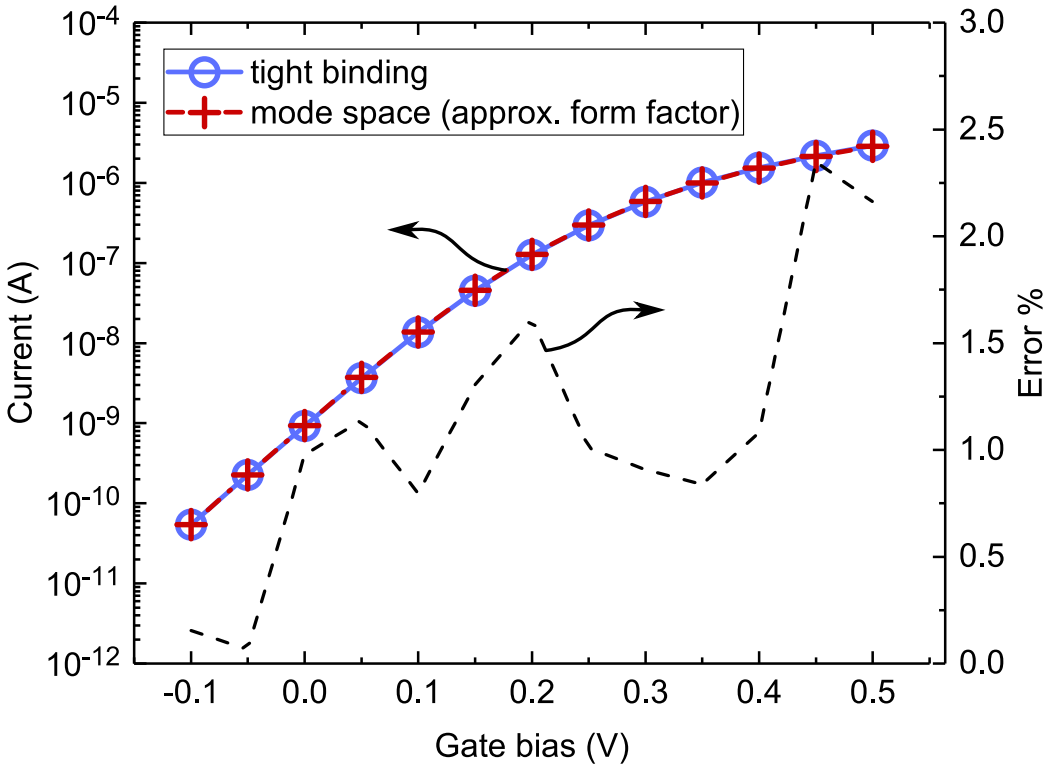


Figure 6.4. I-V curve of the 3.26 nm × 3.26 nm × 20.65 nm silicon nanowire of Fig. 6.3 with an approximate form factor. The agreeing results justify the form factor approximation. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

6.3 Assessment of computational performance

The device in figure 6.1 was used with varying widths w to measure performance improvements in NEMO5 time to solution and peak memory. Each width also had a corresponding mode space transformation matrix with its respective number of modes. Correspondingly, the reduction ratios n/N in figures 6.6 and 6.10 vary. The exact width w values simulated were 4, 6, 8, 10 and 12 silicon unit cells and the respective reduction ratios n/N were 5.6%, 2.8%, 2.9%, 2.8% and 3.0% with a silicon lattice parameter of 0.543 nm. In other words the widths simulated, in nm, were 2.17

Table 6.1.
Time to solution and peak memory of three methods for including real-space information into mode space calculations of scattering self-energies. For the form factor method, times of generation of the form factor F and application to Green's functions are shown. Form factor generation time is not included in iteration time, since it only occurs once at the beginning of the simulation. This data corresponds to a basis reduction from a rank of 2880 to 81

method	iteration time (s)	FF application time (s)	FF generation time (s)	peak memory (GB)
$G^{R,<}$ upconversion	64.19	N/A	N/A	25.23
full form factor	5.69	0.76	12.67	1.42
approx. form factor	4.95	0.05	0.01	1.07

nm, 3.26 nm, 4.34 nm, 5.43 nm and 6.516 nm. All performance simulations were performed with the same inputs used for validation in section 6.2, with the exception that a fixed number of 256 energies were simulated. Since results for the approximate form factor have been shown in figure 6.4 to closely match those of the full form factor, mode space data for performance comparisons in this section were generated using the approximate form factor.

6.3.1 Time to solution assessment for a single scattering iteration

The Green's functions were solved for 256 energies with 1 energy per MPI process. Each MPI process was designated a 32-core node on the Blue Waters petascale supercomputer at the University of Illinois at Urbana-Champaign. Each MPI process was designated to a 32-core node on the Blue Waters petascale supercomputer at the University of Illinois at Urbana-Champaign [130]. Each MPI process was assigned 32 OpenMP threads for multithreaded matrix operations, as well as form factor construction and application. Figure 6.6 shows the average time (of 6 iterations) to

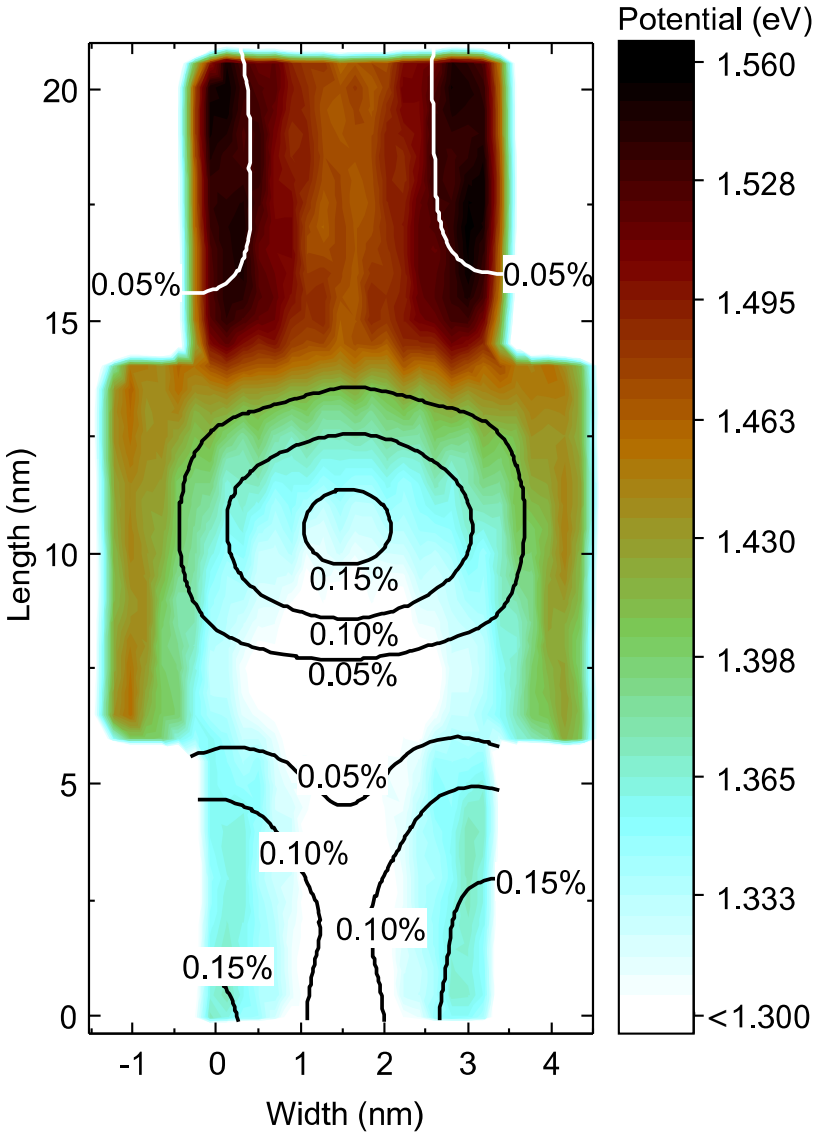


Figure 6.5. Potential profile (contour plot) of the center cross-section of the simulated 3.26 nm × 3.26 nm × 20.65 nm silicon nanowire device in original tight binding basis. Contour lines represent the relative absolute error of the potential in mode space compared to tight binding representation. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

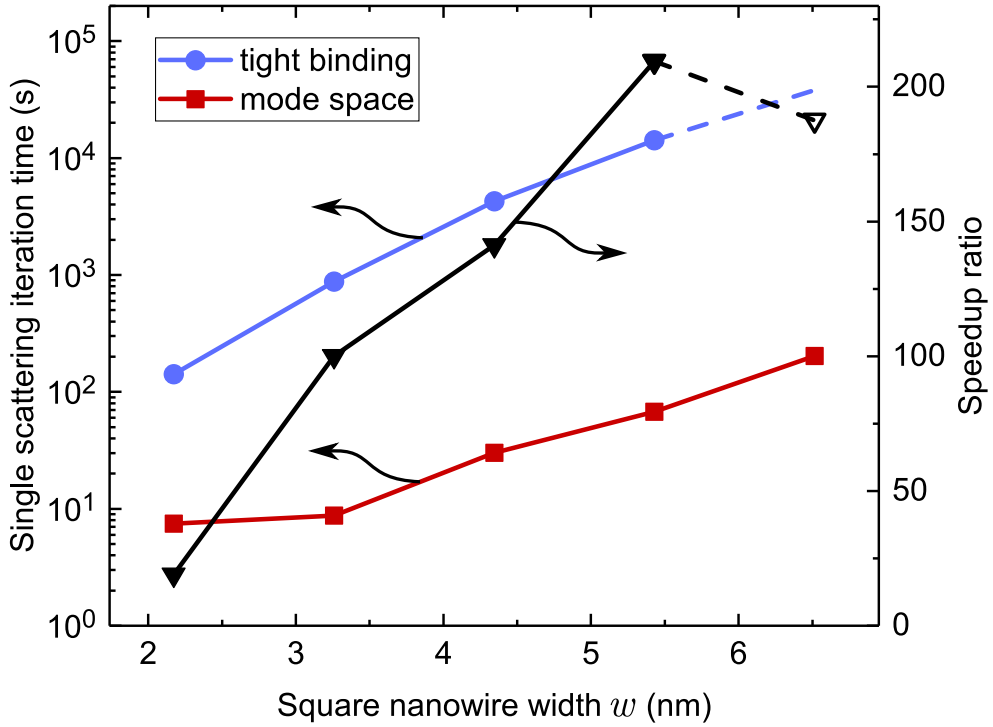


Figure 6.6. Time to solution for a single self-consistent Born iteration (left) and speedup ratio (right) with low-rank approximations for the 20.65 nm silicon nanowire of Fig. 6.1 for various widths w. The tight binding timing data was extrapolated beyond $w = 5.43$ nm using a power fitting function shown as a dashed line. All simulations include inelastic scattering. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

compute a single self-consistent Born iteration. Each self-consistent Born iteration includes the time to compute the RGF algorithm as well as the time to compute lesser scattering self-energies $\Sigma^<$ and retarded scattering self-energies Σ^R for optical and acoustic deformation potential inelastic scattering. The calculation of scattering self-energies involves a large degree of communication between MPI processes as discussed in reference [8].

The maximum speedup obtained with low-rank approximations for an iteration in this work was of 209.5 times. Due to computational limitations, the tight binding simulation for the point $w = 6.52$ nm was not assessed, since a single iteration would have taken about 38,000 seconds according to a power fitting function of the existing data. By extrapolating the data, the speedup for $w = 6.52$ nm is predicted to be of 187.5 times, as is shown in figure 6.6. It can be noted that this is lower than the speedup of $w = 5.43$ nm. This is likely due to the fact that the reduction ratio for $w = 6.52$ nm is slightly higher at 3.0% than for $w = 5.43$ nm at 2.8%.

The timing shown in figure 6.6 does not include the calculation of other aspects of quantum transport such as charge density and potential with Poisson's equation and the generation of the adaptive energy grid. The exclusion of these calculations can be justified since the time to solution of Poisson calculations is negligible when compared to the solution of NEGF in production scale simulations that include hundreds of energies, dozens of Poisson iterations, and hundreds of scattering iterations. In production runs, these calculations are performed only a small fraction of times when compared to the multiple self-consistent Born iterations per Poisson iteration.

6.3.2 Time to solution assessment for NEGF simulation walltime

As a preview of the timing breakdown of a production simulation with many iterations, figure 6.7 shows the walltime of the NEMO5 simulations of figure 6.6 from beginning to end. These simulations, however, do not reflect a full-scale production simulation, as only 2 Poisson iterations with 3 scattering iterations each were performed. A total of 6 scattering iterations were therefore performed. Along with these calculations, a ballistic RGF iteration and contact self-energy (using the Sancho Rubio method [137]) calculations were also performed in the reduced mode space basis. Note that the largest width to complete in a reasonable time for the full tight binding basis was $w = 4.34$ nm. The speedup for this largest possible comparison was of 80.52×, and much larger speedups can be expected for devices of larger cross-sections.

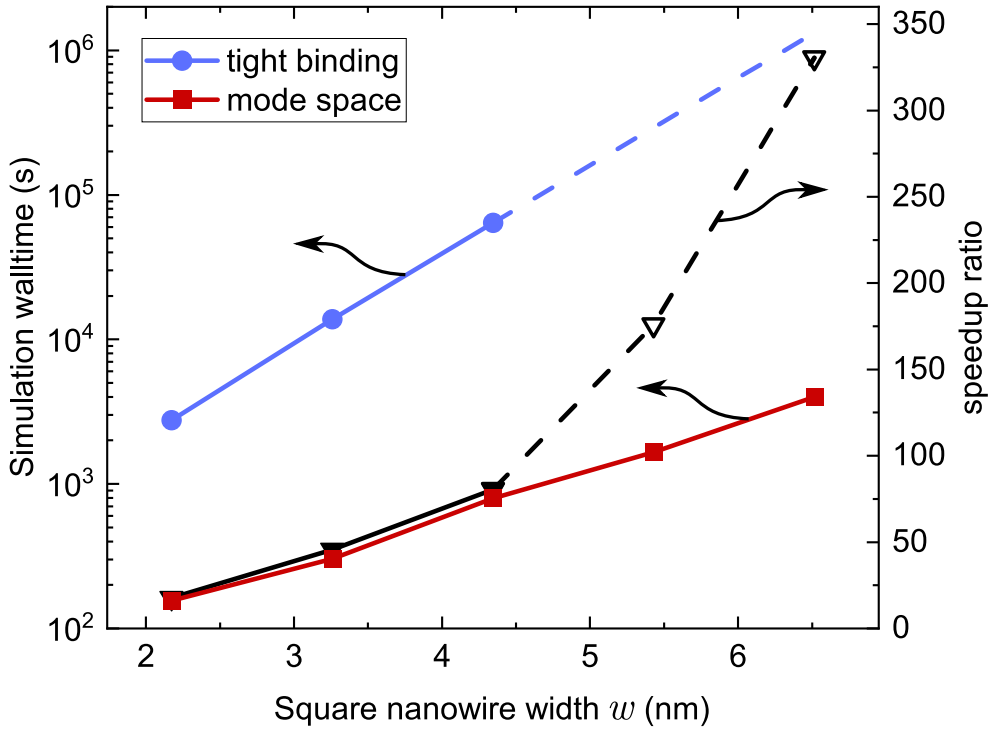


Figure 6.7. Full simulation walltime for the same simulations of figure 6.6 including 6 scattering iterations and all other portions of the NEGF calculation. Note that all 6 iterations could not be completed for the two largest widths due to the required computational resources. The dashed lines represent predictions for these two largest widths

Similarly to figure 6.6, a power fitting function was used to predict the simulation walltimes for the full tight binding basis due to time restrictions.

6.3.3 Timing breakdown of simulations

To analyze the timing breakdown of a full simulation such as those shown in figure 6.7, the simulation of width $w = 3.26$ nm was chosen. A NEMO5-internal profiling tool was used to measure and analyze the timing spent on various portions of the NEGF simulations. This breakdown is shown for both the original tight binding basis and reduced mode space basis as pie charts in figure 6.8. Note that for the tight

binding simulation shown in figure 6.8.a, the majority of the time is spent in the *leads + 7 RGF* portion. This portion of the pie chart includes 6 RGF iterations with scattering, 1 ballistic RGF iteration and contact self-energy calculations. Note that while RGF with scattering may occur dozens of times per Poisson iteration, the ballistic RGF and contact self-energy calculations occur only once per Poisson iteration. The *leads + 7 RGF* portion of the simulation is greatly reduced in mode space, as shown by figure 6.8.b. In addition to the *leads + 7 RGF* portion, the other portions of the NEGF calculation shown in figure 6.8.b that have been reduced in mode space include *density (MPI)* and *scattering (MPI)*. These portions, the calculation of charge density and scattering self-energies, involve frequent MPI communication [8,25]. Mode space basis reductions reduce the size of data blocks being communicated via MPI, making communication more efficient and faster. The *output current* section of figure 6.8 represents the output of energy-resolved and slab-resolved current that is written to file for every scattering iteration. Although this is not necessary to achieve results from NEGF, it was included in this test because it is a commonly analyzed metric that exhibits the convergence behavior of the self-consistent Born approximation method. Note also that mode space reduces the time spent on the slab-resolved current output. This is because there are fewer device "slabs" in the mode space basis, as well as fewer elements to sum to obtain current. In the tight binding representation of Si, each unit cell of the material contains 4 atomic layers, which would be reduced to a single slab in mode space. Other portions of the NEGF simulation remain virtually unchanged, including *semiclassical* which provides an initial potential guess at the beginning of the simulation, and *source/drain bands* which provides a bandstructure solution for the adaptive energy grid. These two tasks are only performed once per simulation, regardless of the number of energies. The transformation of the Hamiltonian H is only performed in the mode space version of the simulation, and it is only performed once per Poisson iteration for each energy.

A breakdown of only 6 iterations, however, does not reflect on the timing breakdown or total time to solution of a production run, since this includes sequential

code and calculations that are performed only once. Since a production run in a full tight binding basis would require many compute resources, figure 6.9 shows an ideal timing projection of the $w = 3.25$ nm device for a production run on a per-scattering-iteration basis. This was obtained by using the timing breakdown of figure 6.8 and multiplying the portions which are repeated in a production run by a typical number of iterations. For this example, 10 Poisson iterations, with 10 self-consistent Born scattering iterations per Poisson iteration (total 100 scattering iterations) would be performed. This timing breakdown assumes that every MPI process solves a single energy point. The portions *semiclassical*, *source/drain bands*, *transform H* and *other* would only be performed once. The portion *density (MPI)* would only be performed 10 times, once per Poisson iteration. The portion *leads + RGF* would change according to how often the contact self-energy and ballistic calculation would be performed: 10 times, while the scattered RGF portion would be performed 100 times. *scattering (MPI)* and *output current* would be performed 100 times, once per scattering iteration. The total corresponding walltimes for the full basis and mode space basis simulations would be 131,414 seconds (36.5 hours) and 2305 seconds (38.4 minutes) respectively, a speedup of 57×. A similar projection for a 2.17 nm × 2.17 device results in a predicted production walltime for full basis and mode space of 23,516 seconds (6.5 hours) and 863 seconds (14.4 minutes) respectively, a speedup of 27×.

6.3.4 Memory assessment

The simulations performed for peak memory assessment were the same simulations of figure 6.6. Figure 6.10 shows that the maximum peak memory reduction was of 7.14×. Similarly to figure 6.6, a power fitting function was used to predict that for a device of $w = 6.52$ nm, the speedup would be of 5.67×. Peak memory was assessed using NEMO5-internal code.

6.4 Simulating beyond existing capabilities

With the time to solution and memory footprint significantly reduced, the opportunity to simulate larger devices with complex physical phenomena such as incoherent scattering of multiple types (phonons, roughness, impurities) is now accessible. Reference [8] describes the simulation of a circular nanowire, with acoustic and optical deformation potential scattering and a 10-band tight binding basis. The diameter of the cross-section of this device was 3 nm, and the device length was 27 nm. Solution of an I-V characteristic curve took approximately 275 hours on 330 cores on the Blue Waters petascale supercomputer [130]. The peak memory of a process was 60 GB per node, close to the maximum node memory of 64 GB. This device therefore approaches the limit of what can be simulated in a full basis representation such as tight binding. To demonstrate the capability of solving larger devices in a reduced basis, a full I-V curve was generated for a square nanowire of figure 6.1 with $w = 5.43$. Due to the different cross-sectional geometry this nanowire has over 4 times more atoms in the cross-section than the circular nanowire of reference [8]. The reduction ratio n/N for the square nanowire was of 2.8%. Figure 6.11 shows an I-V characteristic curve for optical and acoustic phonon deformation potential scattering, compared to that of a ballistic simulation. As expected, the on-current density is reduced by the inelastic scattering on phonons [8, 68, 97]. The scattered transport simulation of the $w = 5.43$ nm device took approximately 160 total hours on 16,384 cores (2.62 million core hours) on the Blue Waters supercomputer. Based on previous performance comparisons it can be estimated that the same I-V calculation would take about 550 million core hours and 168 GB of memory in the original tight binding basis representation.

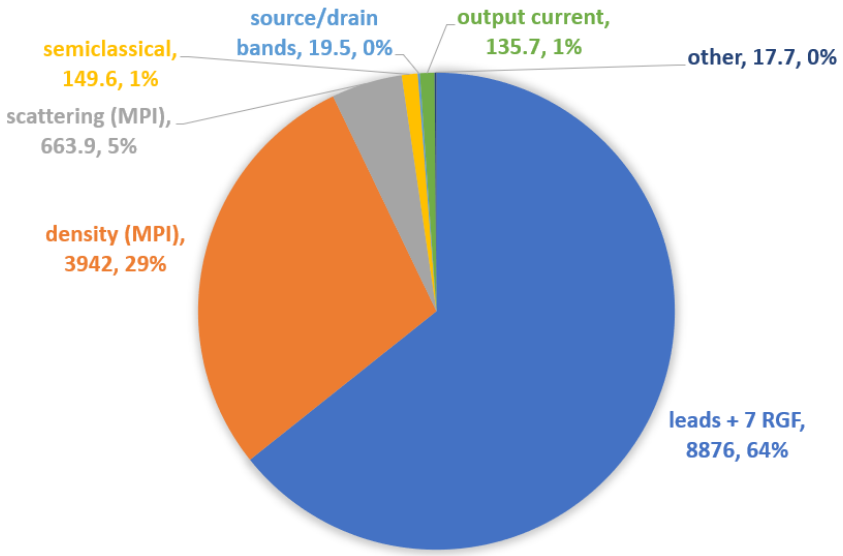
6.5 Outcomes of low-rank approximation work

The iNEMO group collaborated with the Taiwan Semiconductor Manufacturing Company to develop the mode space method with scattering that is available in

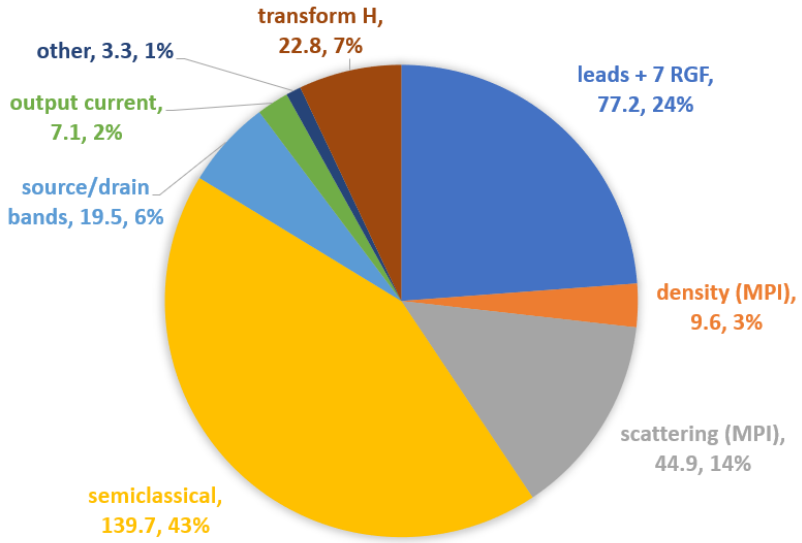
NEMO5 from 2014 to 2017. Publications by TSMC affiliates that used the NEMO5 mode space framework include references [26, 53, 85, 108, 117, 138].

The inclusion of mode space with scattering in NEMO5 has an impact on industry. The NEMO5 software, as of 2020, is licensed by Silvaco Inc. for use by semiconductor manufacturing companies as a commercial tool [139]. The performance improvements obtained by low-rank approximations were a motivator for Silvaco to choose to include NEMO5 in its lineup of TCAD products for simulating quantum transport, and these basis reductions continue to be expanded to include more realistic physics.

Another important outcome of the LRA implementation in NEMO5 is the newly available capability of performing complex calculations in reduced time and with reduced memory footprint. This not only reduces the resources required for existing quantum transport models, but allows for use of models that were previously realistically unattainable. In the next two chapters, two computationally expensive models will be shown, which have been implemented into NEMO5 through an extension of LRA capabilities.

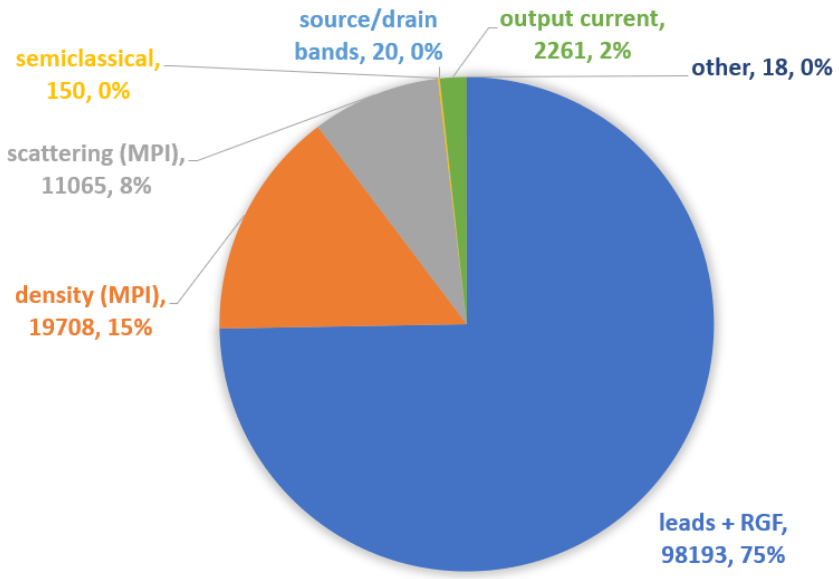


(a) Timing breakdown of NEGF in full tight binding basis

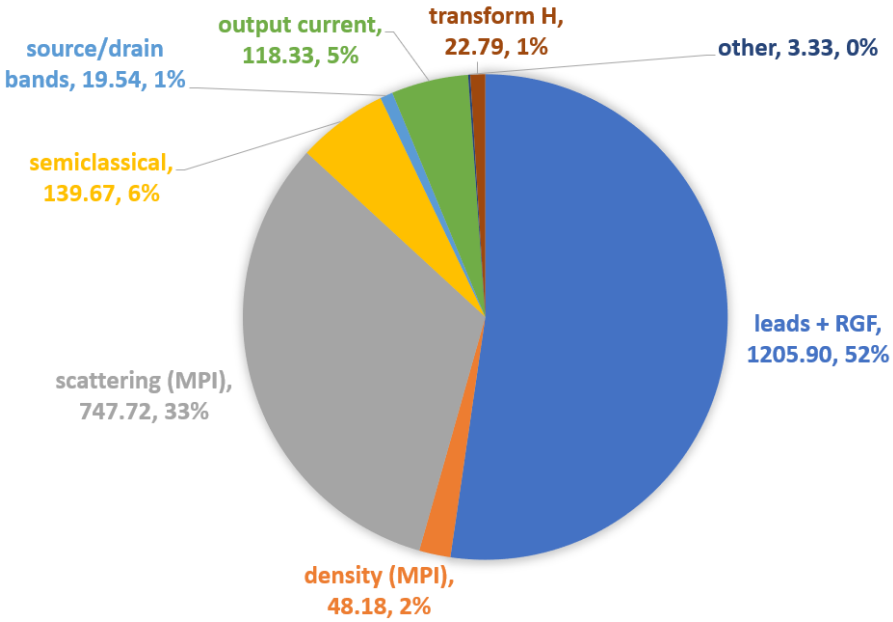


(b) Timing breakdown of NEGF in mode space basis

Figure 6.8. Breakdown of the timing in seconds spent on various portions of the NEGF calculation for a 3.25 nm $\times$ 3.25 nm $\times$ 20.65 nm device. These simulations performed 6 scattering iterations in (a) full basis tight binding, and (b) mode space



(a) Projected timing breakdown of NEGF in full tight binding basis



(b) Projected timing breakdown of NEGF in mode space basis

Figure 6.9. Projected timing breakdown for a full-scale production run of a 3.25 nm $\times$ 3.25 nm $\times$ 20.65 nm Si device in full tight binding basis and mode space for 10 Poisson iterations and 100 total scattering iterations

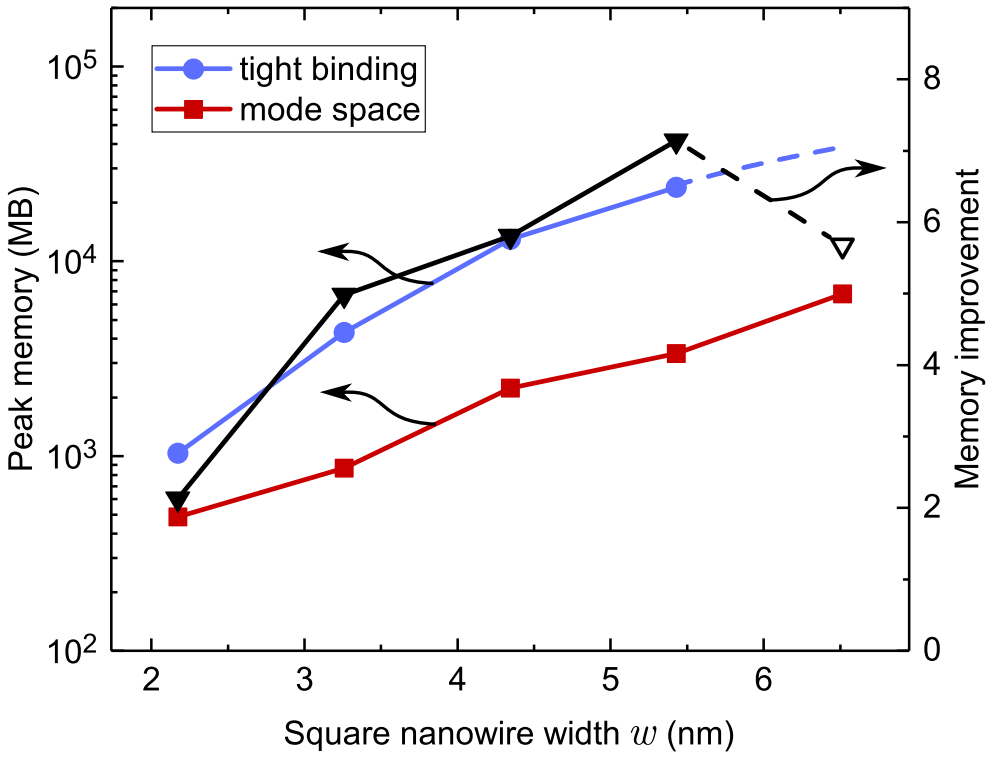


Figure 6.10. Peak memory (left) and memory improvement ratio (right) with low-rank approximations for 20.65 nm silicon nanowires of figure 6.1 for various widths w. All simulations include inelastic scattering. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

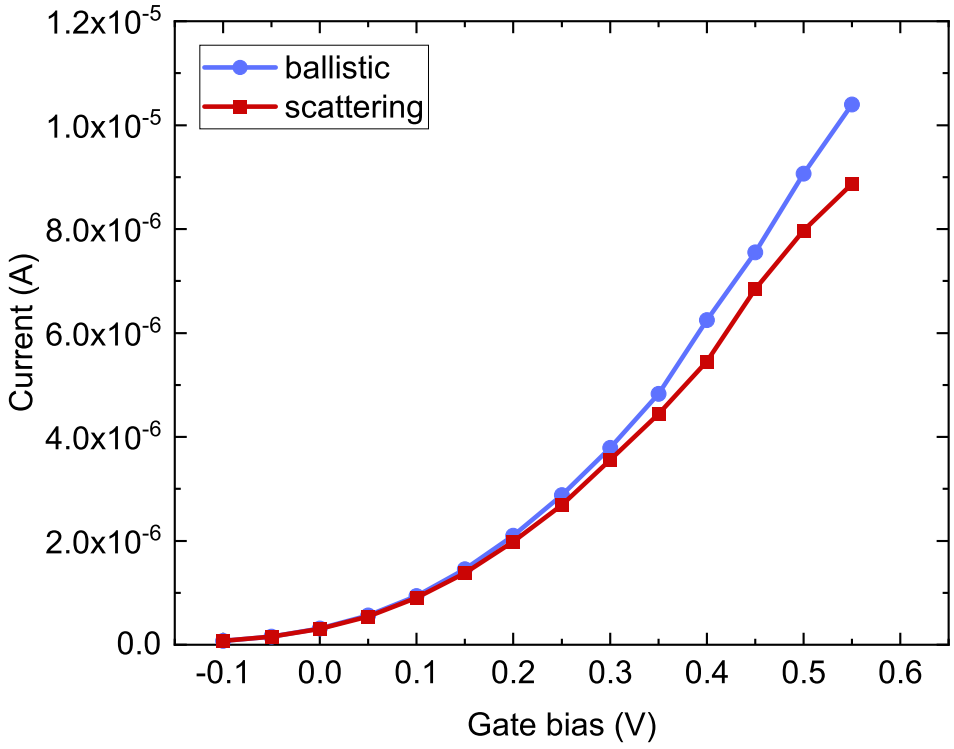


Figure 6.11. Comparison of I-V characteristics for a 5.43 nm × 5.43 nm × 20.65 nm n-type FET device for simulations with and without scattering. The reduction ratio n/N for this simulation was 2.8%. This device size significantly exceeds the largest nanowires possible to resolve in a scattered NEGF calculation in the original atomic representation. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

7. NOVEL AND EXACT IMPLEMENTATION OF RETARDED SCATTERING SELF-ENERGIES USING THE KRAMERS-KRONIG RELATIONS IN MODE SPACE

The general form of the retarded scattering self-energy Σ^R includes a principal value integral $\mathcal{P}$ of large computational burden [7, 8, 97, 98, 135, 140]. Typically, the real part of the retarded self-energy is entirely excluded, and although the approximation often yields reasonable physical results [7, 135], exclusion of the real part causes deviations. In particular, OFF-state current densities are underestimated in this approximation [8, 97, 98]. The real part of retarded self-energies shifts resonance energies and thus influences band edges and threshold voltages [70]. In this chapter, the exact real parts of the retarded scattering self-energies are obtained using the Kramers-Kronig relations [141].

7.1 Method for obtaining the real part of retarded scattering self-energies

$\Sigma^R(\mathbf{r}, \mathbf{r'}, E)$, the retarded self-energy for a perturbation from position $\mathbf{r}$ to a position $\mathbf{r'}$ and energy E, can be obtained by its separate real and imaginary parts [97, 98, 135]

$$\begin{aligned} Re[\Sigma^R(\mathbf{r}, \mathbf{r'}, E)] &= i\mathcal{P}\int \frac{dE'}{2\pi}\frac{\Sigma^>(\mathbf{r}, \mathbf{r'}, E') - \Sigma^<(\mathbf{r}, \mathbf{r'}, E')}{E - E'} \\ &= i\mathcal{P}\int \frac{dE'}{\pi}\frac{Im[\Sigma^R(\mathbf{r}, \mathbf{r'}, E')]}{E - E'}, \end{aligned} \tag{7.1}$$

$$Im[\Sigma^R(\mathbf{r}, \mathbf{r'}, E)] = \frac{1}{2}(\Sigma^>(\mathbf{r}, \mathbf{r'}, E) - \Sigma^<(\mathbf{r}, \mathbf{r'}, E)). \tag{7.2}$$

For each matrix element $\Sigma^R_{i,j}(\mathbf{r}, \mathbf{r'}, E)$ of a retarded self-energy at row i and column j, its real part $\Sigma(\mathbf{r}, \mathbf{r'}, E)^R_{i,j,real}$ is obtained by applying the Kramers-Kronig relation

on its imaginary part $\Sigma(\mathbf{r}, \mathbf{r'}, E)^R_{i,j,imag}$. Using a Hilbert transform $\mathcal{H}$, the real part becomes:

$$\Sigma(\mathbf{r}, \mathbf{r'}, E)^R_{i,j,real} = \mathcal{H}(\Sigma(\mathbf{r}, \mathbf{r'}, E)^R_{i,j,imag}) \tag{7.3}$$

which can be obtained with the following operations:

$$\mathcal{H}(\Sigma(\mathbf{r}, \mathbf{r'}, E)^R_{i,j,imag}) = \mathcal{F}^{-1}\left(-S(m) \cdot \mathcal{F}(\Sigma(\mathbf{r}, \mathbf{r'}, E)^R_{i,j,imag})\right) \tag{7.4}$$

where $\mathcal{F}$ is a Fourier transform, $\mathcal{F}^{-1}$ is an inverse Fourier transform, m is the energy index, and

$$S(m) = \begin{cases} 1 & \text{for } m = 0, \dfrac{N_E}{2} \\ 2 & \text{for } m = 1, 2, 3, ..., \dfrac{N_E}{2} - 1 \\ 0 & \text{for } m > \dfrac{N_E}{2} \end{cases} \tag{7.5}$$

for N_E total energies [142]. Put simply, this Hilbert transform is performed using a fast Fourier transform (FFT), a multiplication in the Fourier space, and an inverse FFT afterwards [143].

7.2 Approximations of retarded scattering self-energies

Note that in chapter 6, scattered NEGF calculations in mode-space did not include the real part of Σ^R, and the same is true for other works with scattering in mode space [7, 26]. Many publications [8, 9, 12, 25, 30, 97, 98] use an approximation that removes a principal value integral from the calculation of Σ^R, though it is not the principal value integral shown in equation 7.1. The first step in obtaining this approximation is by expanding the equation for Σ^R so that it only depends on G^R and $G^<$ and not $G^>$ [8, 12] using the relation

$$G^>(E) = G^R(E) - G^A(E) + G^<(E) \tag{7.6}$$

where $G^A(E)$ is the advanced Green's function

$$G^A(E) = \left(G^R(E)\right)^\dagger . \tag{7.7}$$

From this expansion, Σ^R becomes:

$$\begin{aligned}
&\Sigma^R(\mathbf{r},\mathbf{r'},E) = \int \frac{d\mathbf{q}}{(2\pi)^3} e^{i\mathbf{q}(\mathbf{r}-\mathbf{r'})}|M_q|^2 \cdot \\
&\Bigg[(n_q+1)G^R(\mathbf{r},\mathbf{r'},E-\hbar\omega_q) + n_q G^R(\mathbf{r},\mathbf{r'},E+\hbar\omega_q) \\
&+ \frac{1}{2}\big(G^<(\mathbf{r},\mathbf{r'},E-\hbar\omega_q) - G^<(\mathbf{r},\mathbf{r'},E+\hbar\omega_q)\big) \\
&+ i\mathcal{P}\int \frac{dE'}{2\pi}\frac{G^<(\mathbf{r},\mathbf{r'},E-E')}{E'-\hbar\omega_q} - \frac{G^<(\mathbf{r},\mathbf{r'},E-E')}{E'+\hbar\omega_q}\Bigg]
\end{aligned} \tag{7.8}$$

where perturbations occur from atom and orbital positions $\mathbf{r}$ to $\mathbf{r'}$ and energy E. $\mathbf{q}$ are phonon momentums, n_q is the Bose distribution, ω_q is the phonon frequency and M_q are the constants corresponding to the type of scattering. The approximation is attained by removing the principal value integral in the last line of equation 7.8 and using the resulting Σ^R. The removal of the principal value integral makes equation 7.8 equivalent to equation 5.30 for optical phonon scattering. Transport with this approximated Σ^R which includes a non-zero real part is compared to a Σ^R with the real part completely removed, as well as transport with an exactly calculated real part of Σ^R in references [8, 97, 98]. These show differences in the current output of a nanowire, including underestimations of OFF-current.

To show that the real part of Σ^R cannot be used in mode space, a test was conducted using a 20 nm n-type Si nanowire with a 2.2 nm $\times$ 2.2 nm cross-section. The initial tight binding basis was a 5-band $sp3s^*$. From figure 6.1, the device had an n-type $s = d = 6$ nm source and drain region, doped at 10^{20} cm^{-3}, and a $c = 8$ nm central intrinsic region. A source-drain bias V_{DS} of 0.2 V was applied to the drain and a gate bias V_G of 0.7 V was applied. For the mode space reduction, matrices were reduced from a rank of 640 to 149. This test compared transport in three different cases in the original TB basis and mode space basis for a total of six cases:

1. Zero real Σ^R, in full TB basis
2. Zero real Σ^R, in mode space
3. Non-zero real and approximate Σ^R, in full TB basis
4. Non-zero real and approximate Σ^R, in mode space
5. Non-zero real and exact Σ^R from Kramers-Kronig relations, in full TB basis
6. Non-zero real and exact Σ^R from Kramers-Kronig relations, in mode space

Figure 7.1 shows the result of the six above cases and several observations can be made. First, it is evident that the mode space approximations closely match the results of the full TB basis when no zero real part of Σ^R is included. Second, when the approximate non-zero real part of Σ^R is included, the full basis density results differ slightly from the zero real part results. However, it is evident that when using this approximation, the mode space results do not match the full basis results, especially in the drain section of the device. Lastly, the results with an exact real Σ^R match closely for mode space and full basis, proving that this method of obtaining an exact Σ^R works in a reduced mode space basis. Although it has been proven that the non-zero Σ^R approximation does not work with mode space basis reductions, mode space reductions allow us to perform computationally expensive operations such as Hilbert transforms on greatly reduced matrices, negating the need for approximations.

7.3 Assessment of the real part of retarded scattering self-energies on a TFET device

To assess the scattering effects of the real part of the retarded self-energies Σ^R on a real device, full I-V characteristic curves were obtained for TFET devices. The material of the transistor in figure 6.1 was chosen to be InAs, with two tested device widths $w = 2.42$ nm and $w = 3.63$ nm. Both devices had an $s = 5.97$ nm p-type source doped at 5×10^{19} cm^{-3}, an n-type $d = 9.66$ nm drain doped at 2×10^{19} cm^{-3}

and a $c = 14.66$ nm central undoped region. A source-drain bias of 0.3 V was applied. Since TFETs require the occupation of both electrons and holes, the method of reference [7] was applied to obtain modes for a wide energy window that included bands near the conduction and valence band edges. The inclusion of holes also necessitated a proper definition of electrons and holes as states tunnel from valence band to conduction band in the TFET. An interpolation method was applied as defined by reference [8] to avoid sharp transitions from holes to electrons or vice versa. Simulations included optical phonon, acoustic phonon and polar optical phonon (POP) scattering [144] to represent the polar nature of InAs. Due to the non-local nature of polar optical phonon scattering, such a calculation would be very expensive even in a reduced basis [30]. To avoid this, a local scattering calculation was performed using a cross-section-dependent compensation factor defined in reference [30]. With this compensation, scattering operations can be treated as local. Compensating scaling factors of 30.0 and 26.56 were used in the calculation of polar optical phonon scattering for the $w = 2.42$ nm and $w = 3.63$ nm devices respectively. Note, the diagonal form factor approximation as described in section 5.5.3 was not performed in this case.

The 2-norms of the real and imaginary parts of the retarded self-energy Σ^R can show the amplitude of their relative contributions. Comparing the 2-norms of fully charge-self-consistent calculations is misleading, however, since scattering impacts the density of states: The Poisson potential would compensate some of the density of state differences to accommodate the device's doping profile. Therefore, for this comparison only, scattering self-energies and Green's functions were solved self-consistently with a fixed Poisson potential. That potential was deduced from a converged ballistic transport solution of the same device. The calculations were performed for the ON-state bias of 0.4 V. Table 7.1 shows the 2-norm values of the real and imaginary parts of the Σ^R when the Kramers-Kronig relation is observed and when the real part is set to 0. In both of the simulated cross-sections, the norm of the real part is comparable to the norm of the imaginary part.

Table 7.1.
2-norms of the retarded scattering self-energies Σ^R solved in NEGF simulations of two InAs TFETs with a width w and an applied gate bias of 0.4 V. The norm of the real part, calculated using the Kramers-Kronig relations, is comparable to the norm of the imaginary part, and must have a similar significance to simulation results

width w (nm)	zero real Σ^R		Kramers-Kronig	
	real	imag.	real	imag.
2.42	0	0.1184	0.0965	0.1130
3.64	0	0.1080	0.0920	0.1104

Figures 7.2 and 7.3 show the I-V characteristics of the $w = 2.42$ nm and $w = 3.64$ nm devices respectively. Both figures show the differences of the two scattering models (with and without the real part of Σ^R) when compared to ballistic transport. Incoherent scattering increases the OFF-current density due to scattering-supported gate leakage and decreases the ON-current density due to stronger back-scattering. This is in agreement with findings in literature [8, 57, 58, 97, 135].

The impact of the real part of Σ^R becomes more apparent in situations with larger scattering strengths, e.g. when higher temperatures, impurity scattering, or surface roughness scattering are present. Figure 7.4 shows the I-V characteristics of the device in figure 7.3 solved with NEGF when all electron-phonon scattering self-energies are multiplied by 2. More significant gate leakage and back-scattering effects can be observed than that shown in figure 7.3. More importantly, however, figure 7.4 shows that the exact Σ^R with a non-zero real part provides even higher scattering strengths than the approximate, zero real part case.

7.4 Performance of a TFET simulation with Hilbert transforms

Similarly to section 6.3.3, a NEGF simulation was run to measure the time to solution of the various portions of the NEGF calculation on a TFET with Hilbert

transforms. The device used was the same as that of section 7.2, and like in section 6.3.3, 6 iterations were run and included various other portions of the NEGF calculation. This simulation, however, was run on 48 MPI processes, each with 2 OpenMP threads, on the RCAC Brown cluster [145] at Purdue University. 240 energies were simulated on these 48 processes.

These TFET simulations include three major performance differences to the performance tests of chapter 6: The first is that they do not include the diagonal form factor approximation of section 5.5.2. Because of this, the form factor generation and application to Green's functions shown in equation 5.31 can take a significant amount of time. Figure 7.5 shows that the pie chart portions *generate FF* and *apply FF*, which correspond to form factor generation and application to Green's functions respectively, can take a much more significant amount of time than when the approximation of 5.5.2 is applied. Fortunately, the form factor generation must only be performed a single time per mode space basis, per simulation. The form factor application, however, must be performed for every self-consistent scattering iteration and for every energy, and is thus needed for the entirety of the calculation of the NEGF equations.

The second major performance difference to the tests of chapter 6 is of course the inclusion of Hilbert transforms, which must be performed for the retarded scattering self-energy calculation of every scattering iteration. This is shown in figure 7.5 in the pie chart portion labeled *Hilbert transform*.

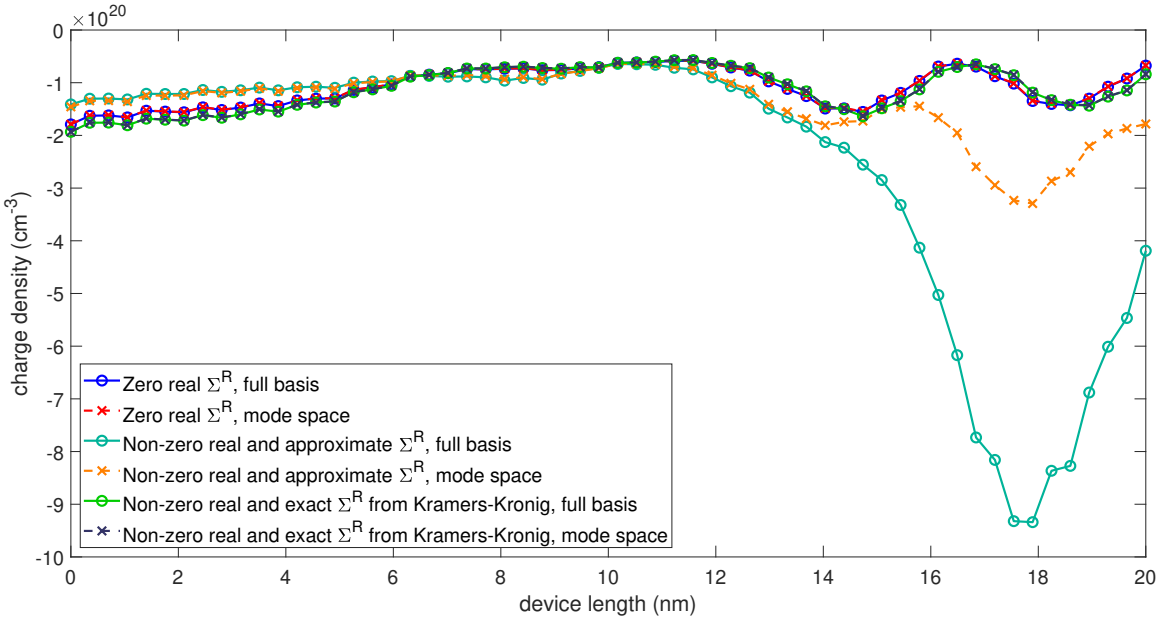
The third performance difference is the increased time spent in the scattering calculation as shown by the *scattering (MPI)* section when compared to that of figure 6.3.3. The reason for this is the inclusion of polar optical phonon scattering. Although the scalar compensation factor approximation of reference [30] reduces communication greatly, self-energies in mode space are block-dense and thus require large blocks to be communicated for each scattering self-energy calculated.

The rest of the pie chart portions in figure 7.5 show a similar time to solution distribution as figure 6.8.b, with the exception that these portions of the NEGF

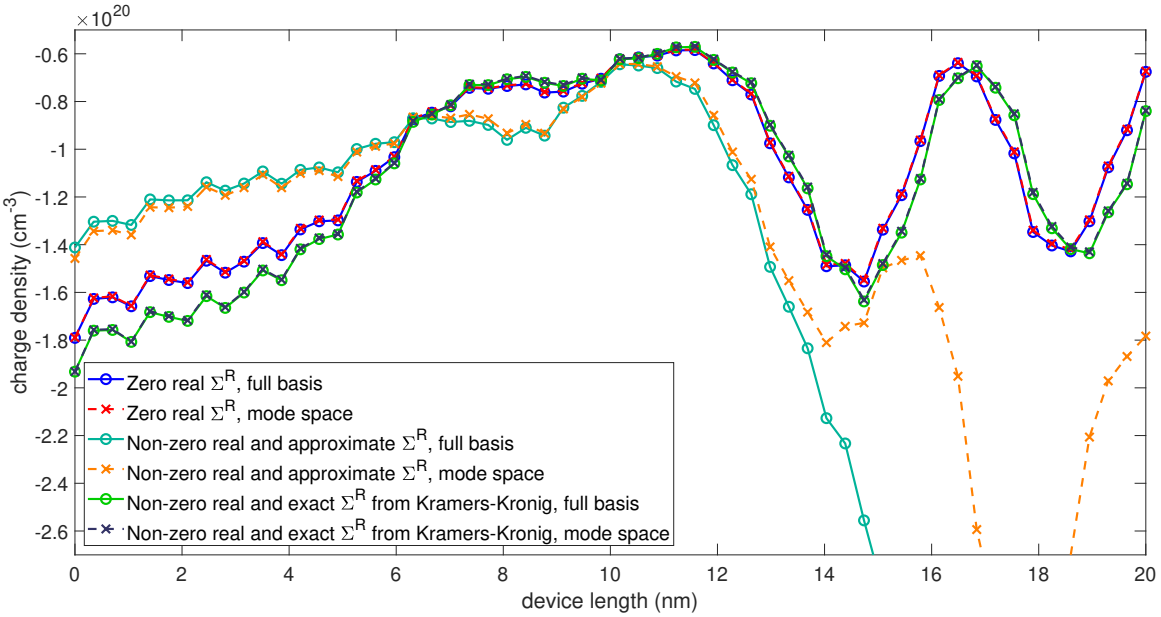
solution were performed faster, despite the mode space basis for this TFET having 101 modes as opposed to the 81 mode basis of the $w = 3.26$ nm device of chapter 6. The reason for this could be a combination of newer hardware being used and a better hybrid distribution of MPI processes and OpenMP threads. Figure 7.6 shows a smaller $w = 2.42$ nm device, for which a full basis comparison can be seen.

7.5 Outcomes of exact retarded scattering self-energies in mode space

The most important outcome of the work shown in this chapter is the inclusion of an exactly calculated real part of retarded scattering self-energies Σ^R. Although this would be most conveniently used in a reduced basis due to large computational burdens in a full basis, it is now available to all users of NEMO5. The approximation shown in references [8, 12, 97, 98] may be valid for many device simulations, but the availability of exact calculations of the real part of Σ^R in NEMO5 should negate the need for this approximation for any future scattering simulations.



(a) All cases compared in full basis and mode space basis



(b) Plot of (a) zoomed in on correctly matching mode space results

Figure 7.1. One-dimensional charge density along the center of a 20 nm × 2.2 nm × 2.2 nm Si nanowire. Three cases are tested in the full TB basis and mode space: With a zero real part of Σ^R, a non-zero real part of Σ^R calculated via an approximation, and the real part of Σ^R calculated with the Kramers-Kronig relations

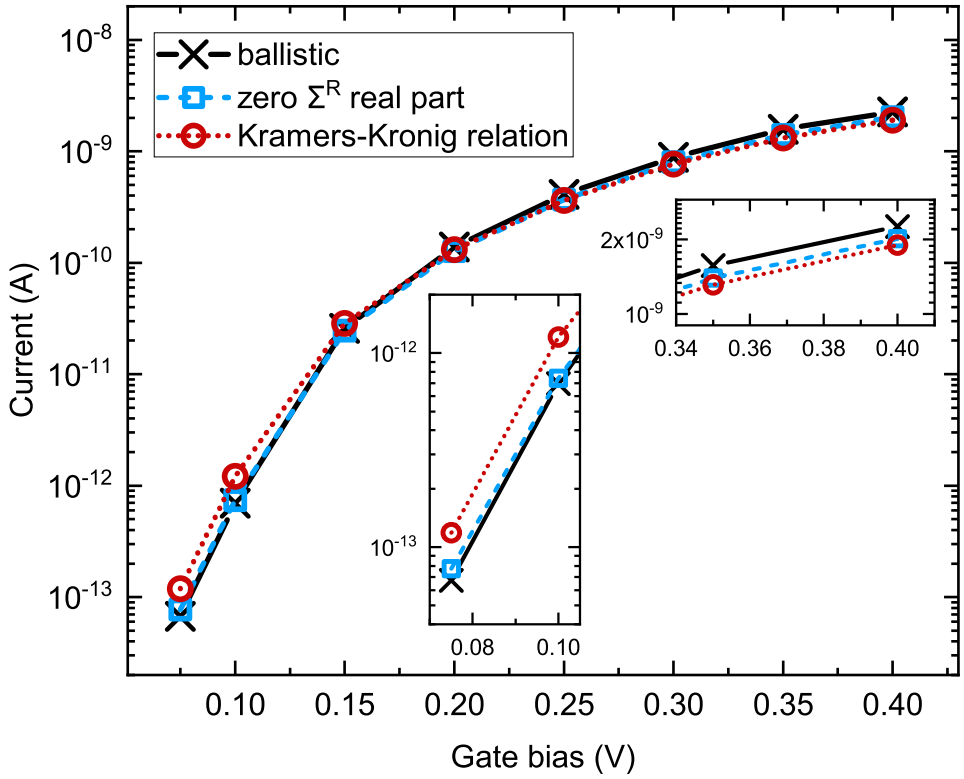


Figure 7.2. I-V characteristics for a 2.42 nm × 2.42 nm × 30.29 nm InAs TFET device solved in NEGF including incoherent scattering on polar optical phonons, acoustic phonons and optical deformation potential phonons. Scattering, even without a real part of Σ^R, increases the OFF-current densities and lowers ON-current densities. When the real part of the retarded self-energy Σ^R is included, the Kramers-Kronig relations are obeyed and scattering shows an even larger impact. The insets zoom into the first two and the last two points of the curves. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

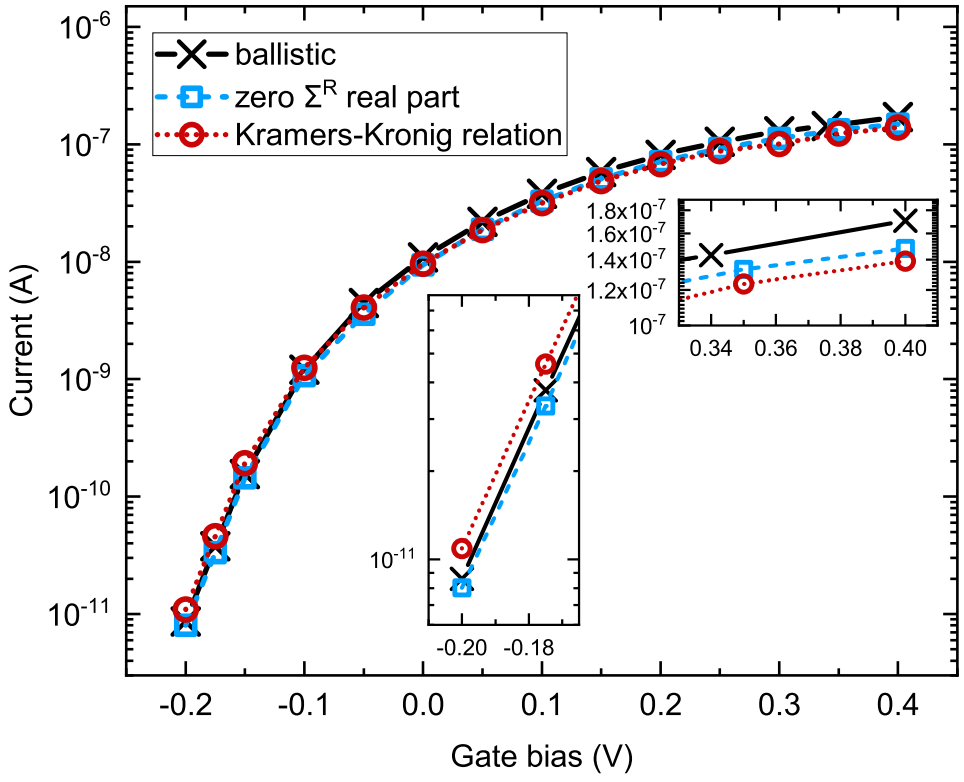


Figure 7.3. Similar to figure 7.2, I-V characteristics of a 3.64 nm $\times$ 3.64 nm $\times$ 30.29 nm InAs TFET device. The effects of scattering with and without a real part of Σ^R are larger than in the smaller wire of figure 7.2. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

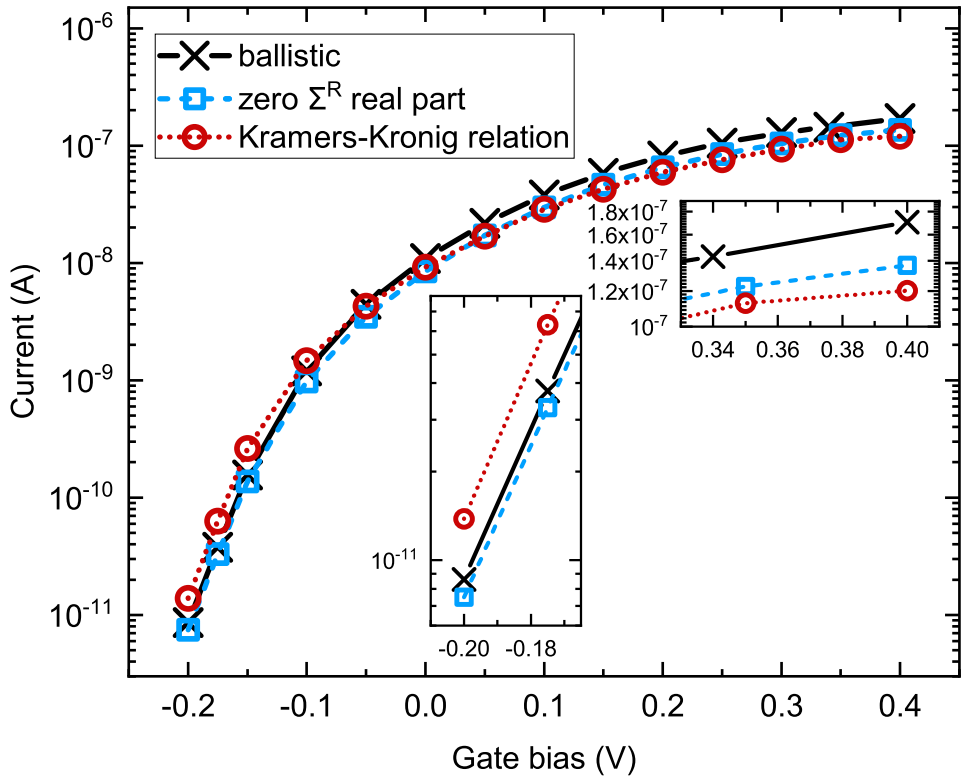


Figure 7.4. Similar to figure 7.3, I-V characteristics of a 3.64 nm × 3.64 nm × 30.29 nm InAs TFET device, but with scattering self-energies multiplied by 2. *Material from: 'D. A. Lemus, J. Charles, T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," Journal of Computational Electronics, submitted 2020, Springer'*

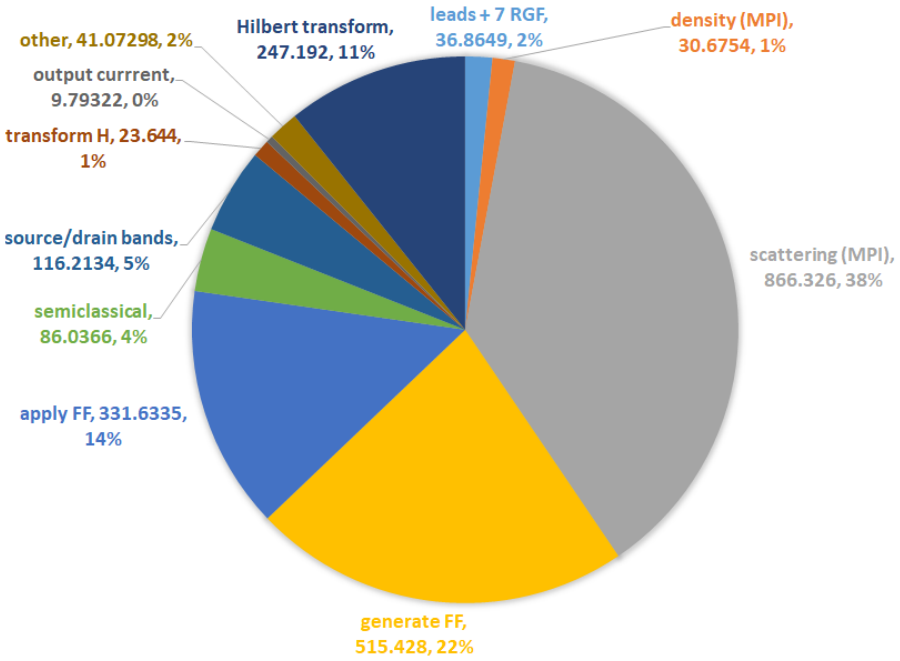
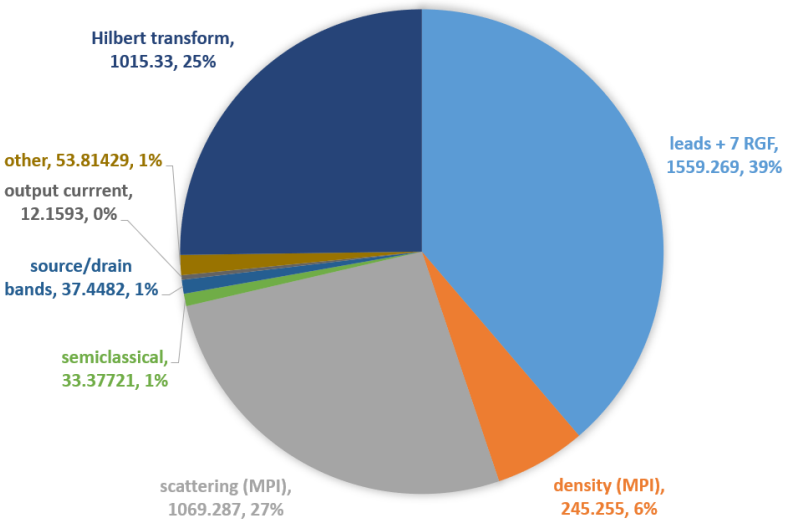
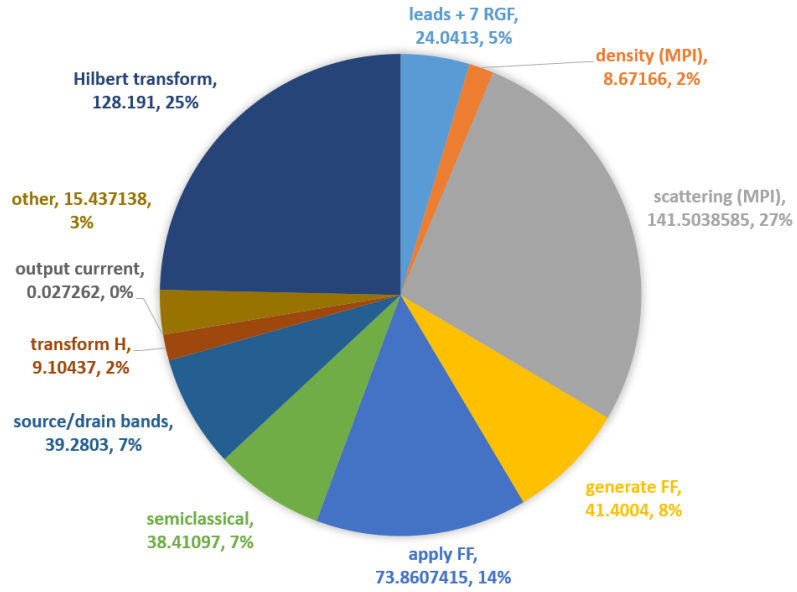


Figure 7.5. Breakdown of the timing spent on various portions of the NEGF calculation for a 3.64 nm × 3.64 nm × 30.29 nm TFET in mode space basis with the real part of scattering self-energies calculated using Kramers-Kronig relations



(a) Timing breakdown of NEGF in full tight binding basis



(b) Timing breakdown of NEGF in mode space basis

Figure 7.6. Breakdown of the timing in seconds spent on various portions of the NEGF calculation with the inclusion of Hilbert transforms and full form factor calculations for a 2.42 nm × 2.42 nm × 30.29 nm device. These simulations performed 6 scattering iterations in (a) full basis tight binding, and (b) mode space

8. EXTENDING LOW-RANK APPROXIMATIONS TO NONLOCAL SCATTERING

Various types of electron-on-phonon scattering such as elastic acoustic deformation potential scattering and optical phonon scattering shown in chapter 6 can be approximated and treated as local [12] to reduce computational complexity and communication. However, some types of scattering such as polar optical phonon scattering must often be treated as nonlocal in nature [12, 30, 57]. An exception to the rule is when an approximation such as that of section 7.2 is introduced, which allows for a local scattering environment and applies a scalar compensation factor to the polar optical phonon scattering self-energies to compensate. The scalar compensation factor, detailed in reference [30], is the result of the ratio of the nonlocal scattering rate and local scattering rate calculated using Fermi's Golden Rule. Other types of nonlocal scattering include roughness scattering due to device imperfections [31, 32]. Nonlocal scattering in NEGF presents a computational challenge due to its high requirement of computational resources and high time to solution.

8.1 Computational burden of nonlocal scattering calculations

Existing solutions of incoherent scattering in NEGF use some approximations, one of them being the use of diagonal self-energies [12]. Although this allows for lower computational complexity, predictions with local scattering may deviate from experimental results [12, 146]. Figure 8.1 shows that a numerical calculation of nonlocal scattering provides a scattering rate prediction closer to theoretical predictions via Fermi's Golden Rule when polar optical phonon scattering is included [146].

Implementation of nonlocal scattering through the newly developed nonlocal RGF algorithm has been completed in NEMO5 [57]. It extends the recursive Green's

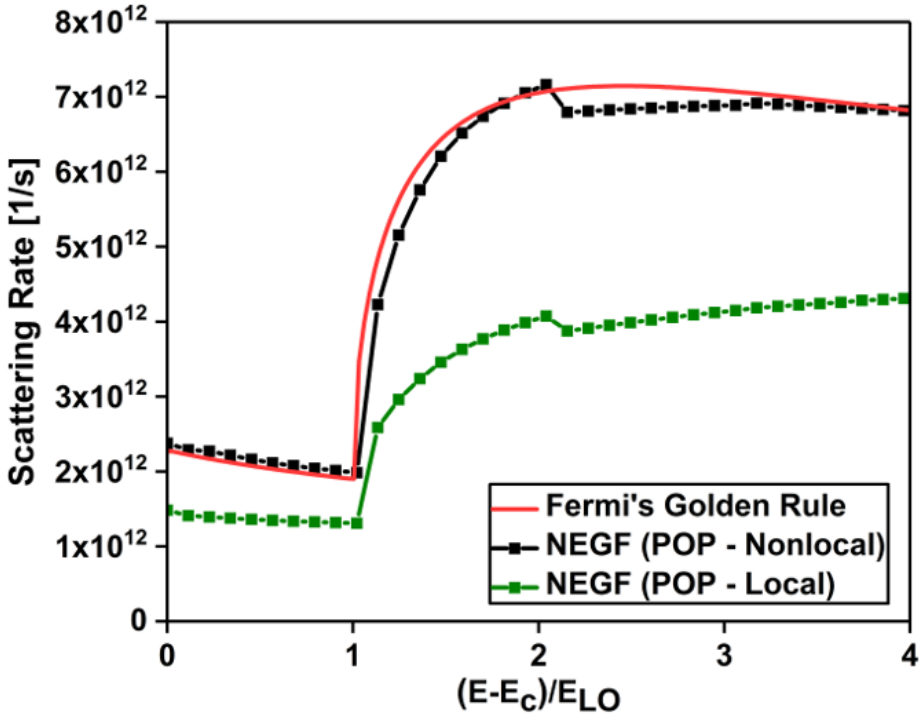


Figure 8.1. Scattering rate for local and nonlocal scattering simulations compared to analytical solution via Fermi's Golden Rule. *Image courtesy of Prasad Sarangapani* [146]

function algorithm [9, 52] and allows for the calculation of offdiagonal blocks of the Green's functions for the entire device. Therefore, instead of being block tridiagonal, Green's functions, and by extension self-energies, may have any number of offdiagonal blocks. The algorithm can be found in detail in reference [57]. Figure 8.2 shows a timing comparison for local and nonlocal RGF calculations for a variable nonlocality range in nm. The range of 1.9 nm, for example, would correspond to 14 nonzero offdiagonal blocks in Green's functions and self-energies. The significant ratio of up to 150 times for the longest nonlocality tested shows that these calculations have an unreasonably long time to solution for realistically sized devices. Figure 8.3 shows a memory comparison. The largest ratio is of almost 8 times, which places an easily reached limitation on what can be modeled with nonlocal RGF. These tests were

performed on a 2.17 nm × 2.17 nm × 20.63 nm silicon device in a 10-orbital $sp3d5s^*$ tight binding basis.

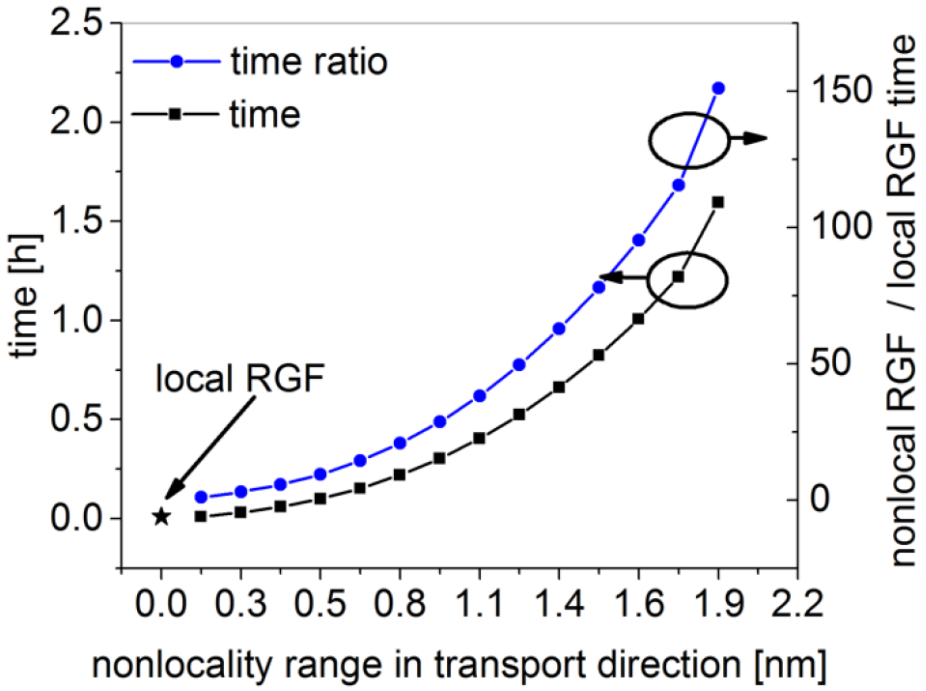


Figure 8.2. Time to solution for a nonlocal RGF calculation with variable nonlocality range (in black) and timing ratio (in blue) when compared to the local calculation (shown as a star). *Image courtesy of James Charles* [57]

By expanding the methods of basis reduction detailed in section 5.5, modeling of nonlocal scattering in nanoelectronic devices becomes more feasible than existing full-basis solutions. The computational requirements of nonlocal RGF with scattering make it unusable in a full atomistic basis for the simulation of anything but the smallest devices, like the 2.17 nm × 2.17 nm × 20.63 nm device described. Low-rank approximations are required for reduced matrix sizes so that computation is feasible.

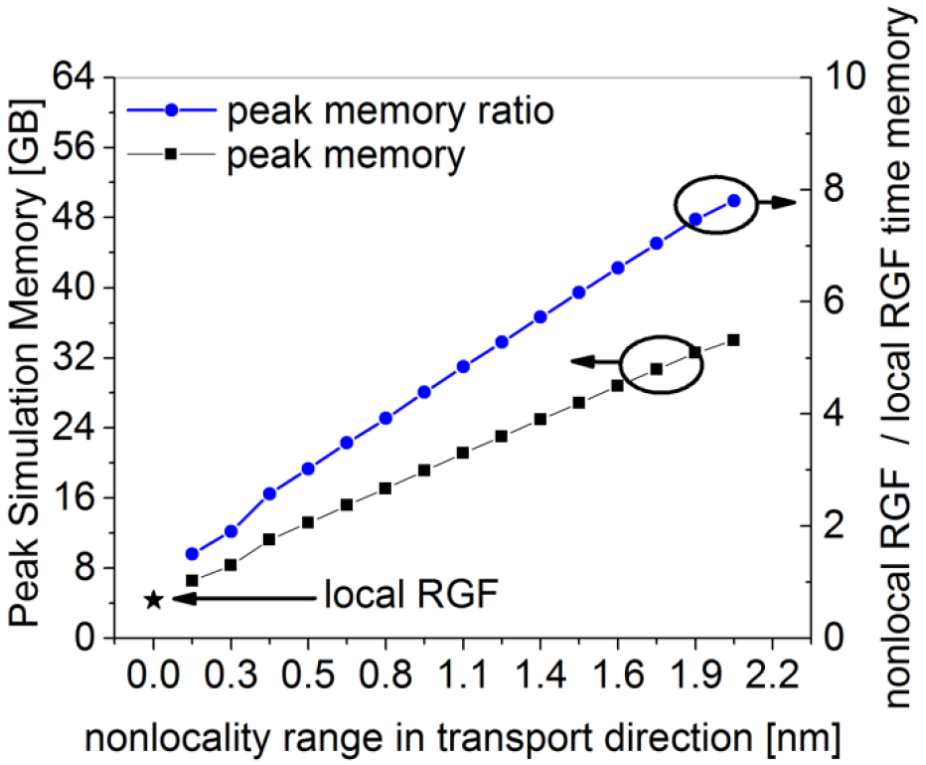


Figure 8.3. Peak memory for a nonlocal RGF calculation with variable nonlocality range (in black) and timing ratio (in blue) when compared to the local calculation (shown as a star). Image by James Charles [57]

8.2 Nonlocal RGF method and LRA application

The traditional block tridiagonal RGF algorithm with mode space basis reductions was shown in section 5.4. Similarly to the block tridiagonal Green's function, the device Hamiltonian H must be reduced to the mode space Hamiltonian h using equation 5.13 so that

$$\left(G^R\right)^{-1} = E - h - \Sigma^R. \tag{8.1}$$

Derived in reference [57], an extension to the RGF algorithm capable of solving any

number of off-diagonal blocks is shown now for block indices I and J. To begin with the nonlocal RGF algorithm we first define the equations

$$D_{I,I} = \left(G^R_{I,I}\right)^{-1} - \sum_{K=0}^{I-1} L_{I,K} D_{K,K} \left(L_{I,K}\right)^T, \tag{8.2}$$

$$L_{I,J} = \left[\left(G^R_{I,J}\right)^{-1} - \sum_{K=0}^{J-1} L_{I,K} D_{K,K} \left(L_{J,K}\right)^T\right] D^{-1}_{J,J} \tag{8.3}$$

which correspond to the LDL decomposition $(G^R)^{-1} = LDL^T$. T is the transpose operator. We then define

$$g^R_{I,I} = \left[A_{I,I} - \sum_{K=I-1-N_l}^{I-1} \tilde{L}_{I,K} g^R_{K,K} \left(\tilde{L}_{I,K}\right)^T\right]^{-1} \tag{8.4}$$

where N_l is the total number of layers in the device and

$$\tilde{L}_{I,K} = A_{I,K} - \sum_{K'=I-1-N_l}^{K-1} \tilde{L}_{I,K'} g^R_{K',K'} \left(\tilde{L}_{K,K'}\right)^T \tag{8.5}$$

with

$$\tilde{L}_{I,J} = L_{I,J} g^R_{I,J}. \tag{8.6}$$

$g^<_{I,J}$ can be calculated for the **diagonal** blocks when $I = J$:

$$\begin{aligned} g^<_{I,I} = {} & g^R_{I,I} \sum_{K=I-N_l}^{I-1} \tilde{L}_{I,K} \left(-g^<_{I,K}\right)^\dagger \\ & + g^R_{I,I} \sum_{K=I-N_l}^{I-1} \Sigma^<_{I,K} \sum_{M=K}^{I-1} g^A_{K,M} \tilde{L}_{M,I} g^A_{I,I} + g^R_{I,I} \Sigma^<_{I,I} g^A_{I,I} \end{aligned} \tag{8.7}$$

and for the **offdiagonal** blocks when $I < J$:

$$g^<_{I,J} = g^R_{I,I} \sum_{K=I-N_l}^{I-1} \tilde{L}_{I,K} g^<_{K,J} + g^R_{I,I} \sum_{K=I-N_l}^{J-1} \Sigma^<_{I,K} g^A_{K,J} + g^R_{I,I} \Sigma^<_{I,J} g^A_{J,J} \tag{8.8}$$

where

$$g^A_{I,J} = \left(g^R_{I,J}\right)^\dagger . \tag{8.9}$$

G^R is calculated in the "backward" RGF portion as follows for the **diagonal** blocks:

$$G^R_{I,I} = g^R_{I,I} - g^R_{I,I} \sum_{K=I+1}^{I+1+N_l} \tilde{L}_{I,K} G^R_{K,I} \tag{8.10}$$

and the **offdiagonal** blocks when $I < J$

$$G^R_{I,J} = -g^R_{I,I} \sum_{K=I+1}^{I+1+N_l} \tilde{L}_{I,K} G^R_{K,J}. \tag{8.11}$$

Lastly, the blocks of $G^<$ are calculated in "backward" RGF for the **diagonal** blocks:

$$\begin{aligned} G^<_{I,I} &= g^<_{I,I} + g^R_{I,I} \sum_{K=I+1}^{I+1+N_l} L^T_{I,K} G^<_{K,I} \\ &+ \sum_{K=I+1}^{I+1+N_l} g^<_{I,K} \sum_{l=I+1}^{I+1+N_l} L^\dagger_{K,l} G^A_{l,I} + \sum_{K=I+1}^{I+1+N_l} g^R_{I,K} \sum_{l=K+1}^{K+1+N_l} \Sigma^<_{K,l} G^A_{l,I} \end{aligned} \tag{8.12}$$

and the **off-diagonal** blocks when $I < J$:

$$\begin{aligned} G^<_{I,J} &= g^<_{I,J} + g^R_{I,I} \sum_{K=I+1}^{I+1+N_l} L^T_{I,K} G^<_{K,J} \\ &+ \sum_{K=I-N_l}^{J} g^<_{I,K} \sum_{l=J+1}^{J+1+N_l} L^\dagger_{K,l} G^A_{l,I} + \sum_{K=I-N_l}^{I} g^R_{I,K} \sum_{M=J+1}^{J+1+N_l} \Sigma^<_{K,M} G^A_{M,I} \end{aligned} \tag{8.13}$$

where the advanced Green's function blocks

$$G^A_{I,J} = \left(G^R_{I,J}\right)^\dagger . \tag{8.14}$$

The above equations can be used to calculate any upper-diagonal block of G^R and $G^<$, and can be chosen by the user to calculate up to a given nonlocality range.

After the Hamiltonian is reduced, Green's function calculations are performed in a reduced basis using equations 8.2- 8.14. As was shown in section 5.5, complications enter when using self-energies $\Sigma^{R,<}$ in mode space. To calculate and apply a form factor as shown in section 5.5.2, equations 5.27 and 5.31 no longer apply with nonlocal scattering. These equations correspond to local scattering, and a generalized form factor must therefore be constructed which corresponds to the overlap of modes in a nonlocality range $|\mathbf{r'} - \mathbf{r}|$:

$$F_{i,j,k,l}(\mathbf{r}, \mathbf{r'}) = \sum_v \phi_i(\nu, \mathbf{r})\phi_j(\nu, \mathbf{r'})\phi_k(\nu, \mathbf{r})\phi_l(\nu, \mathbf{r'}). \tag{8.15}$$

The corresponding form factor application is performed on the Green's function elements as follows:

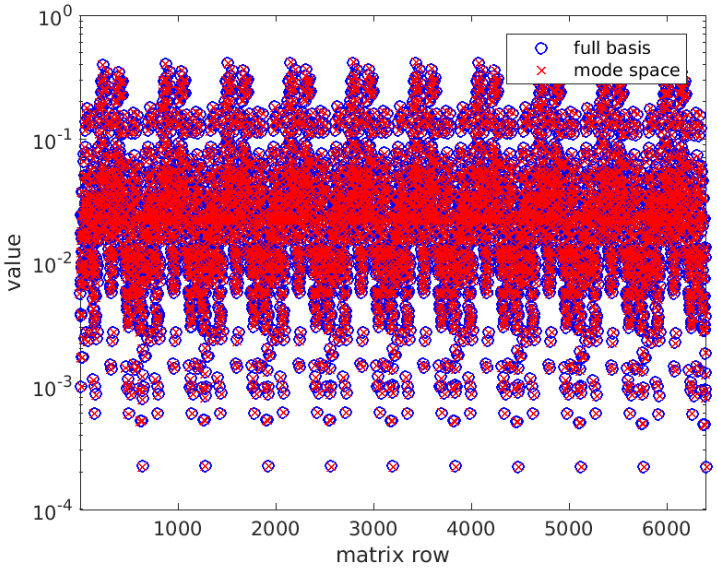
$$\Sigma_{i,j}(\mathbf{r}, \mathbf{r'}) = \sum_l \sum_k CF_{i,j,k,l}(\mathbf{r}, \mathbf{r'})G_{k,l}(\mathbf{r}, \mathbf{r'}). \tag{8.16}$$

For a basis reduction from rank N to rank n the generation and application times would greatly increase, becoming $O(n^6N)$ and $O(n^6)$ respectively. Memory scaling would become $O(n^6)$. It is possible that using the approximation of section 5.5.3, time to solution and memory would decrease significantly, but that approximation relies on the lack of intra-mode overlap. This is not the case for nonlocal scattering, where modes must interact between layers and self-energies and Green's functions are not diagonal. Because of these difficulties, all scattering self-energies in this chapter were solved using the Green's function upconversion method of section 5.5.1 with equations 5.25 and 5.26.

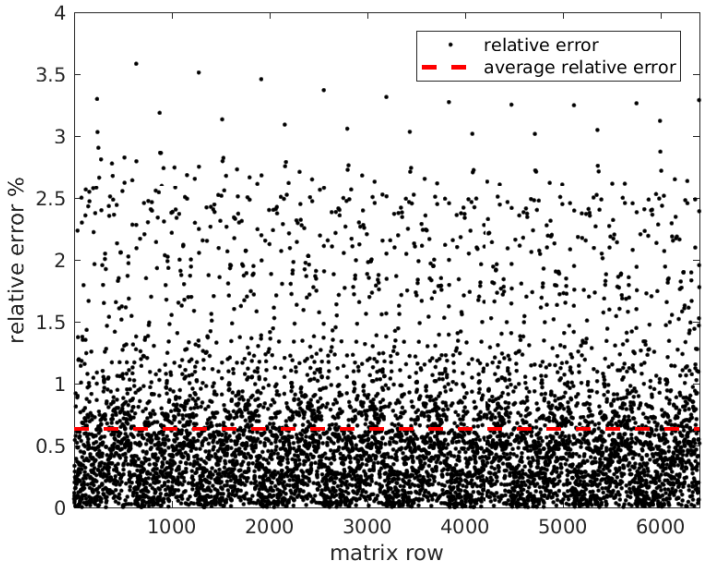
8.3 Assessment of results

The simplest test to assess physical result correctness of nonlocal RGF with scattering in a reduced mode space basis is the use of a homogeneous nanowire with the use of a scalar applied to Green's functions to obtain scattering self-energies. A GaSb device of dimensions 2.42 nm × 2.42 nm × 6.05 nm in a mode-space-reduced 5-band $sp3s^*$ basis and 16 homogeneous energy points was used for this test. The device was completely homogeneous, having a uniform 5×10^{19} cm^{-3} n-type doping, and no biases were applied to the terminals. A constant scalar factor λ was applied to Green's functions to generate the scattering self-energies such that $\Sigma^{R,<} = \lambda \cdot G^{R,<}$. Two scattering iterations were performed with a nonlocality of up to two offdiagonal blocks. This corresponds to two unit cells of nonlocality, or a 1.22 nm nonlocality range. After calculations had been performed in mode space, the reduced Green's functions were upconverted using the same mode space basis transformation to compare to the full basis Green's functions. Figure 8.4 shows the diagonal elements of the resulting $G^<$ matrix after 2 iterations, and figure 8.5 shows the diagonal elements of the second offdiagonal blocks, or the diagonal shifted upward by 1280 rows. The figures show that a very close agreement was obtained from nonlocal implementation of mode space transformations in NEMO5, as they both show an average relative error of less than 1%.

The next step after this simple test was to test nonlocal RGF in mode space on an inhomogeneous nanowire with an applied bias and realistic scattering. The device chosen for this was an InAs nanowire of dimensions 2.42 nm × 2.42 nm × 9.69 nm with, according to figure 6.1, $s = c = d = 3.23$ nm. The source and drain were n-type doped at 5×10^{19} and the central region was intrinsic InAs. A $V_{DS} = 0.3$ V was applied to the drain, and $V_G = 0$ V was applied to the gate terminal. The original basis was a 5-orbital $sp3s^*$. The types of scattering included were elastic acoustic phonon scattering, inelastic optical phonon scattering, and most importantly, nonlocal polar optical phonon (POP) scattering. Unlike the polar optical phonon



(a) Diagonal values of $G^<$



(b) Relative error of diagonal values of $G^<$

Figure 8.4. (a) Diagonal values of $G^<$ compared in full TB basis and mode space after two $\lambda \cdot G^<$ scattering iterations and after upconversion of mode space $G^<$. (b) The relative error of these values

calculations of chapter 7, this case did not include the scalar compensation factor of reference [30], and rather calculated the nonlocal scattering self-energies directly. Like the previous homogeneous device, two offdiagonal blocks were included in the nonlocal RGF calculation, corresponding to a 1.21 nm nonlocality range. Figure 8.6 shows the sparsity pattern of a $G^<$ matrix with 2 offdiagonal blocks. For this test, the viability of nonlocal RGF in mode space was tested, with the resulting current being tested for up to 3 scattering iterations. Figure 8.7 shows the resulting currents for mode space and full basis tight binding. An error of under 10% shows the viability of mode space for nonlocal RGF and nonlocal scattering calculations.

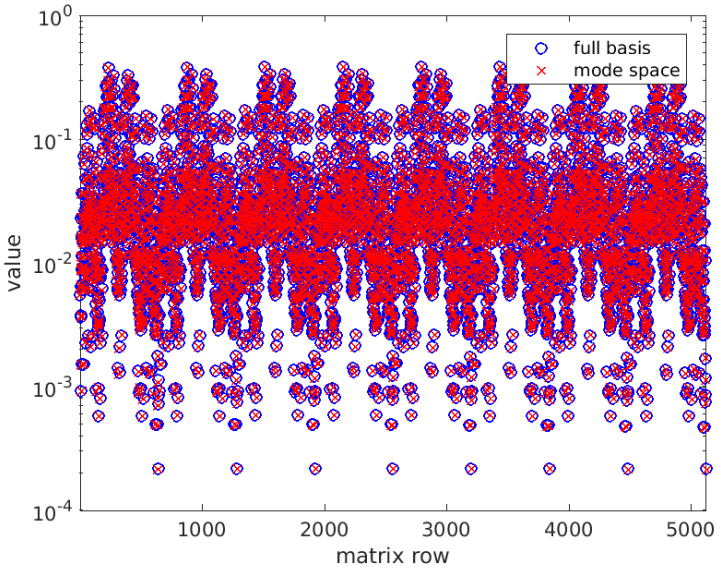
Another important factor to compare is the speedup factor obtained from basis reductions down to 110 modes from 640 degrees of freedom in $sp3s*$. Table 8.1 shows the timing results from this basis reduction. Basis reductions tend to be more significant (under 90% reductions) when the initial basis is a 10-orbital basis, so larger speedups can be expected then. The decrease in speedups between the ballistic and scattering iterations can easily be explained by the inclusion of scattering, which now includes the more complex nonlocal polar optical phonon.

Table 8.1.
Single-iteration time to solution results in mode space and full basis of nonlocal RGF with 2 offdiagonal blocks, acoustic phonon scattering, optical phonon scattering, and nonlocal polar optical phonon scattering

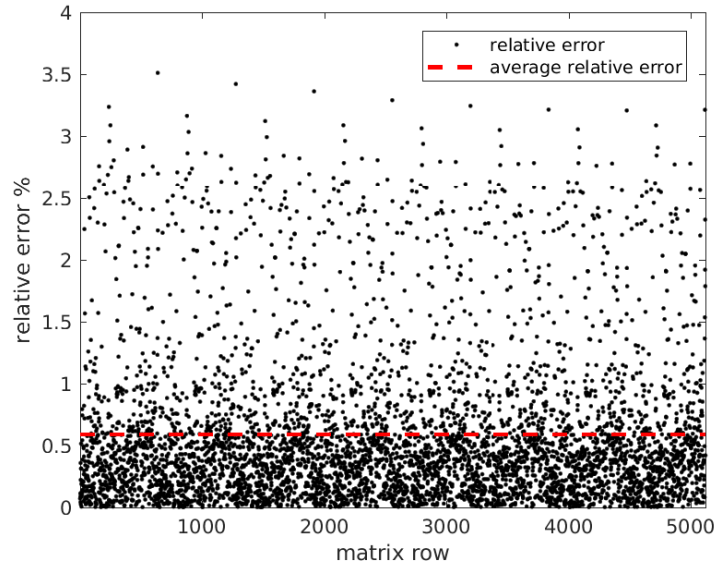
	MS time (s)	full basis time (s)	speedup
ballistic	58.83	1553.88	26.41
scattering	378.43	1785	4.72

8.4 Outcomes of low-rank approximations in nonlocal scattering

The outcome of the inclusion of basis reductions to nonlocal RGF is the availability for future work that includes nonlocal scattering which would otherwise be infeasible in a full atomistic basis. The inclusion of basis reductions opens the possibility to simulate scattering effects such as explicit roughness and device impurities and paves the way for device engineering that includes these realistic effects.



(a) Diagonal values of second offdiagonal block of $G^<$



(b) Relative error of diagonal values of second offdiagonal block of $G^<$

Figure 8.5. (a) Diagonal values of the second offdiagonal block (upwards shift of 1280 rows) of $G^<$ compared in full TB basis and mode space after two $\lambda \cdot G^<$ scattering iterations and upconversion of mode space $G^<$. (b) The relative error of these values

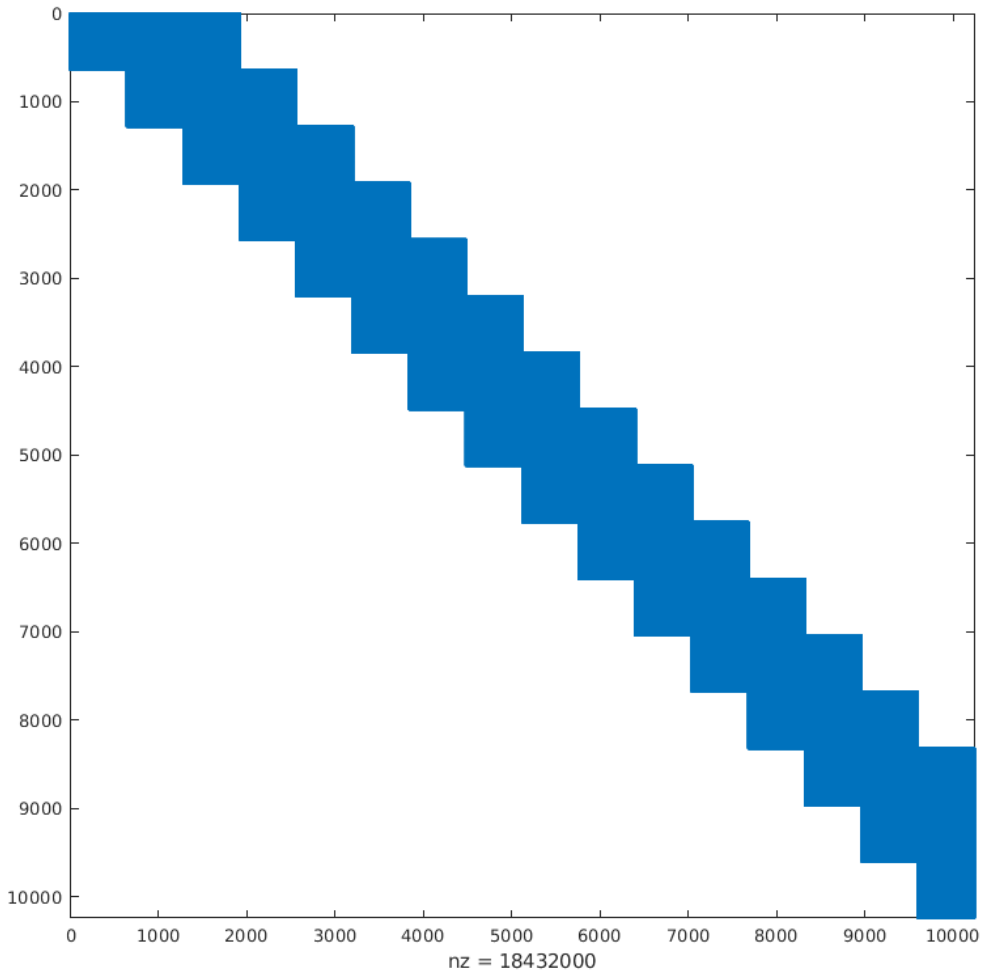


Figure 8.6. Sparsity pattern of $G^<$ matrices with 2 offdiagonal blocks in a 2.42 nm $\times$ 2.42 nm $\times$ 9.69 nm InAs nanowire device after upconversion from mode space

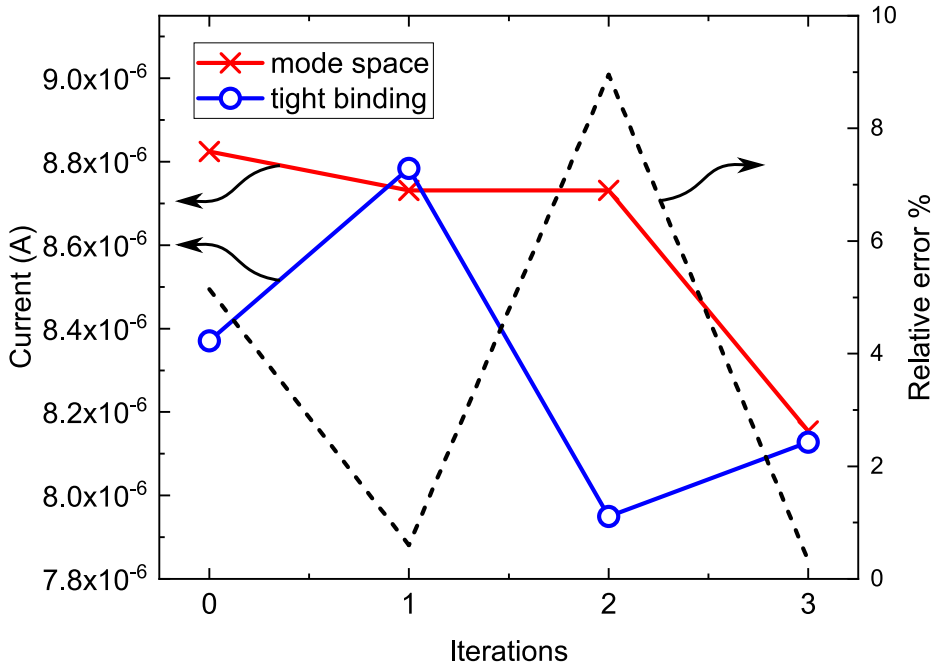


Figure 8.7. Differences of current for an nanowire device with nonlocal RGF and nonlocal scattering in mode space and full basis tight binding for various scattering iterations, with 0 being the ballistic iteration. Error under 10% shows that mode space reductions are viable even for nonlocal calculations

9. CONCLUSION AND IMPACT OF THIS WORK

The overarching message of this thesis is that complex and computationally intensive simulations may be solved in one of two ways: brute force as shown in chapters 3 and 4 through highly parallel heterogeneous computing and supercomputer simulations that scale up to hundreds of thousands of CPU cores, and by reducing problems to a more manageable size through low-rank approximations that maintain device physics as shown in chapters 5 and 6 by the mode space basis reduction method.

These two ways do not have to be mutually exclusive, however, as basis reductions allow for a more effective use of highly parallel systems for solving realistic physical phenomena. In this thesis, this includes the exact solution of retarded scattering self-energies and nonlocal scattering through the nonlocal RGF algorithm in a reduced basis, which may be used for future device engineering with reasonable computational expense.

9.1 Summary of PhD impact

Along with these features which are newly available to device engineers that use NEMO5 for TCAD simulations, the accomplishments and impact of the PhD work outlined in this thesis document can be summarized in the following points:

- Contributed to the development and upkeep of the atomistic electronic device modeling code NEMO5, including compilation issues, memory leaks and performance improvements
- Led the porting of NEMO5 to many environment configurations, including the supercomputers Blue Waters and Stampede2, and portable Ubuntu builds

- Introduced heterogeneous computing capabilities to NEMO5, which include Intel Xeon Phi coprocessors and general-purpose GPUs
- Tested the limits of Intel Xeon Phi capabilities and contributed to the decision by the Intel Numerical Device Modeling group to request a homogeneous CPU-only compute cluster in 2014 [129]
- Submitted NEMO5 scaling capabilities with scattering to the Gordon Bell Prize competition of 2015
- Contributed to the mode space scattering implementation in NEMO5 that is used in references [26, 53, 85, 108, 117, 138]
- Supported Silvaco Inc. in porting NEMO5 to Victory Atomistic as a commercial tool [139]
- Introduced to NEMO5 a novel implementation of the exact solution of retarded scattering self-energies using the Kramers-Kronig relations in a reduced basis
- Introduced to NEMO5 a novel expansion of low-rank approximation capabilities to include the nonlocal RGF method of reference [57]
- Filed patent "Method of modeling many particle systems" with T. Kubis and J. Charles, publication number 2020-0104442, 2020
- Filed patent "System architecture and methods of determining device behavior," with T. Kubis and J. Charles, application number 16/588,046, 2020

9.2 Future work

The addition of the Kramers-Kronig relations to mode space calculations allows for the solution of realistically-sized devices without approximations to the solution of scattering self-energies. This will allow for modeling of scattering phenomena in a manner more consistent with experimental results, and will be useful for testing

devices and physics that would otherwise be too computationally intensive to model. The extension of mode space capabilities to nonlocal scattering will also grant the ability to simulate very computationally intensive physical phenomena that involve long range scattering on phonons and device roughness. The new ability to perform these calculations should be used for device simulations with realistic physics. The low-rank approximation framework in NEMO5 is also potentially generalizable to a vast number of methods and models, including other basis sets and transport of particles other than electrons, and should be used in future work to reduce the computational burden of any otherwise impractical simulation model.

REFERENCES

REFERENCES

[1] S. Ciraci, A. Buldum, and I. P. Batra, "Quantum effects in electrical and thermal transport through nanowires," *Journal of Physics: Condensed Matter*, vol. 13, no. 29, pp. R537–R568, jul 2001.

[2] G. Jing, A. Javey, H. Dai, and M. Lundstorm, "Performance analysis and design optimization of near ballistic carbon nanotube field-effect transistors," *IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004.*, pp. 703–706, 2004.

[3] R. F. Pierret, *Semiconductor device fundamentals*. Pearson Education India, 1996.

[4] S. Datta, *Quantum transport: atom to transistor*. Cambridge University Press, 2005.

[5] "Nemo5," https://engineering.purdue.edu/gekcogrp/software-projects/nemo5/, 2017, [Online].

[6] J. Sellier, J. Fonseca, T. C. Kubis, M. Povolotskyi, Y. He, H. Ilatikhameneh, Z. Jiang, S. Kim, D. Mejia, P. Sengupta *et al.*, "Nemo5, a parallel, multiscale, multiphysics nanoelectronics modeling tool," in *Proc. SISPAD*, 2012, pp. 1–4.

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

[8] J. Charles, P. Sarangapani, R. Golizadeh-Mojarad, R. Andrawis, D. Lemus, X. Guo, D. Mejia, J. E. Fonseca, M. Povolotskyi, T. Kubis, and G. Klimeck, "Incoherent transport in NEMO5: realistic and efficient scattering on phonons," *Journal of Computational Electronics*, 2016.

[9] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices," *Journal of Applied Physics*, vol. 81, no. 12, p. 7845, 1997.

[10] "International technology roadmap for semiconductors." https://www.itrs2.net/, 2017, [Online].

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

[12] T. Kubis and P. Vogl, "Assessment of approximations in nonequilibrium Green's function theory," *Physical Review B - Condensed Matter and Materials Physics*, vol. 83, no. 19, pp. 1–12, 2011.

[13] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations," *Physical Review B*, vol. 74, no. 20, p. 205323, nov 2006.

[14] M. Luisier, "Quantum transport beyond the effective mass approximation," Ph.D. dissertation, ETH Zurich, 2007.

[15] A. R. Rocha, V. M. García-Suárez, S. W. Bailey, C. J. Lambert, J. Ferrer, and S. Sanvito, "Towards molecular spintronics." *Nature materials*, vol. 4, no. 4, pp. 335–9, apr 2005.

[16] J. Guo, S. Datta, and M. Lundstrom, "Toward Multiscale Modeling of Carbon Nanotube Transistors," *International Journal for Multiscale Computational Engineering*, vol. 2, no. 2, pp. 257–276, 2004.

[17] S. Jin, Y. J. Park, and H. S. Min, "A three-dimensional simulation of quantum transport in silicon nanowire transistor in the presence of electron-phonon interactions," *Journal of Applied Physics*, vol. 99, no. 12, pp. 1–10, 2006.

[18] D. Valencia, E. Wilson, P. Sarangapani, G. A. Valencia-Zapata, G. Klimeck, M. Povolotskyi, and Z. Jiang, "Grain boundary resistance in nanoscale copper interconnections," in *Simulation of Semiconductor Processes and Devices (SISPAD), 2016 International Conference on.* IEEE, 2016, pp. 105–108.

[19] K. C. Wang, T. K. Stanev, D. Valencia, J. Charles, A. Henning, V. K. Sangwan, A. Lahiri, D. Mejia, P. Sarangapani, M. Povolotskyi, A. Afzalian, J. Maassen, G. Klimeck, M. C. Hersam, L. J. Lauhon, N. P. Stern, and T. Kubis, "Control of interlayer physics in 2H transition metal dichalcogenides," *Journal of Applied Physics*, vol. 122, no. 22, 2017.

[20] H. Ilatikhameneh, Y. Tan, B. Novakovic, G. Klimeck, R. Rahman, and J. Appenzeller, "Tunnel field-effect transistors in 2-d transition metal dichalcogenide materials," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 12–18, 2015.

[21] F. W. Chen, M. Manfra, G. Klimeck, and T. Kubis, "Nemo5: Why must we treat topological insulator nanowires atomically?" in *Proc. IWCE*, 2015.

[22] T. B. Boykin, J. P. Van der Wagt, and J. S. Harris Jr, "Tight-binding model for gaas/alas resonant-tunneling diodes," *Physical Review B*, vol. 43, no. 6, p. 4777, 1991.

[23] Y. P. Tan, M. Povolotskyi, T. Kubis, T. B. Boykin, and G. Klimeck, "Tight-binding analysis of Si and GaAs ultrathin bodies with subatomic wave-function resolution," *Physical Review B - Condensed Matter and Materials Physics*, vol. 92, no. 8, pp. 1–11, 2015.

[24] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the s p 3 d 5 s* tight-binding formalism: From boundary conditions to strain calculations," *Physical Review B*, vol. 74, no. 20, p. 205323, 2006.

[25] R. Andrawis, J. D. Bermeo, J. Charles, J. Fang, J. Fonseca, Y. He, G. Klimeck, Z. Jiang, T. Kubis, D. Mejia, D. Lemus, M. Povolotskyi, S. A. P. Rubiano, P. Sarangapani, and L. Zeng, "NEMO5: Achieving High-end Internode Communication for Performance Projection Beyond Moore's Law," oct 2015. [Online]. Available: http://arxiv.org/abs/1510.04686

[26] A. Afzalian, J. Huang, H. Ilatikhameneh, J. Charles, D. Lemus, J. B. Lopez, S. P. Rubiano, T. Kubis, M. Povolotskyi, G. Klimeck *et al.*, "Mode space tight binding model for ultra-fast simulations of iii-v nanowire mosfets and heterojunction tfets," in *Computational Electronics (IWCE), 2015 International Workshop on*. IEEE, 2015, pp. 1–3.

[27] E. B. Ramayya, D. Vasileska, S. M. Goodnick, and I. Knezevic, "Electron transport in silicon nanowires: The role of acoustic phonon confinement and surface roughness scattering," *Journal of Applied Physics*, vol. 104, no. 6, 2008.

[28] W. Zhang, C. Delerue, Y. M. Niquet, G. Allan, and E. Wang, "Atomistic modeling of electron-phonon coupling and transport properties in n -type [110] silicon nanowires," *Physical Review B - Condensed Matter and Materials Physics*, vol. 82, no. 11, pp. 2–8, 2010.

[29] E. Conwell and V. F. Weisskopf, "Theory of impurity scattering in semiconductors," *Physical Review*, vol. 77, no. 3, pp. 388–390, 1950.

[30] P. Sarangapani, Y. Chu, J. Charles, G. Klimeck, and T. Kubis, "Band-tail Formation and Band-gap Narrowing Driven by Polar Optical Phonons and Charged Impurities in Atomically Resolved III-V Semiconductors and Nanodevices," *Physical Review Applied*, vol. 12, no. 4, p. 1, 2019.

[31] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy, and O. L. Krivanek, "Surface roughness at the Si(100)-SiO2 interface," *Physical Review B*, vol. 32, no. 12, pp. 8171–8186, 1985.

[32] R. Lake, G. Klimeck, R. C. Bowen, C. Fernando, T. Moise, Y. C. Kao, and M. Leng, "Interface roughness, polar optical phonons, and the valley current of a resonant tunneling diode," *Superlattices and Microstructures*, vol. 20, no. 3, pp. 279–285, 1996.

[33] J. W. Harrison and J. R. Hauser, "Alloy scattering in ternary III-V compounds," *Physical Review B*, vol. 13, no. 12, pp. 5347–5350, 1976.

[34] Y. He, T. Kubis, M. Povolotskyi, J. Fonseca, and G. Klimeck, "Quantum transport in NEMO5: Algorithm improvements and high performance implementation," *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, no. 13, pp. 361–364, 2014.

[35] T. A. Ameen, H. Ilatikhameneh, P. Fay, A. Seabaugh, R. Rahman, and G. Klimeck, "Alloy engineered nitride tunneling field-effect transistor: A solution for the challenge of heterojunction tfets," *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 736–742, 2019.

[36] D. J. Griffiths and D. F. Schroeter, *Introduction to Quantum Mechanics*, 3rd ed. Cambridge University Press, 2018.

[37] M. Luisier and G. Klimeck, "Simulation of nanowire tunneling transistors: From the Wentzel-Kramers- Brillouin approximation to full-band phonon-assisted tunneling," *Journal of Applied Physics*, vol. 107, no. 8, 2010.

[38] H. Ilatikhameneh, G. Klimeck, and R. Rahman, "Can Homojunction Tunnel FETs Scale below 10 nm?" *IEEE Electron Device Letters*, vol. 37, no. 1, pp. 115–118, 2016.

[39] M. A. Lampert, "Mobile and immobile effective-mass-particle complexes in nonmetallic solids," *Physical Review Letters*, vol. 1, no. 12, pp. 450–453, 1958.

[40] G. Klimeck, R. C. Bowen, T. B. Boykin, C. Salazar-Lazaro, T. A. Cwik, and A. Stoica, "Si tight-binding parameters from genetic algorithm fitting," *Superlattices and Microstructures*, vol. 27, no. 2, pp. 77–88, 2000.

[41] D. M. York and W. Yang, "A chemical potential equalization method for molecular simulations," *The Journal of Chemical Physics*, vol. 104, no. 1, p. 159, 1996.

[42] T. B. Boykin, G. Klimeck, R. C. Bowen, and R. Lake, "Effective-mass reproducibility of the nearest-neighbor sp3s* models: Analytic results," *Physical Review B - Condensed Matter and Materials Physics*, vol. 56, no. 7, pp. 4102–4107, 1997.

[43] W. J. Hehre, "Ab Initio Molecular Orbital Theory," *Accounts of Chemical Research*, vol. 9, no. 11, pp. 399–406, 1976.

[44] S. Kuzmin and W. W. Duley, "Ab initio Calculations of Some Electronic and Vibrational Properties of Molecules Based on Multi-Layered Stacks of Cyclic C 6," *Fullerenes, Nanotubes and Carbon Nanostructures*, vol. 20, no. 8, pp. 730–736, nov 2012.

[45] M. Shin, W. J. Jeong, J. Lee, and J. Seo, "First principles based NEGF simulations of Si nanowire FETs," *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, pp. 217–219, 2016.

[46] P. Itskowitz and M. L. Berkowitz, "Chemical potential equalization principle: Direct approach from density functional theory," *Journal of Physical Chemistry A*, vol. 101, no. 31, pp. 5687–5691, 1997.

[47] N. Marzari and D. Vanderbilt, "Maximally localized generalized Wannier functions for composite energy bands," *Physical Review B*, vol. 56, no. 20, pp. 12 847–12 865, nov 1997.

[48] I. Souza, N. Marzari, and D. Vanderbilt, "Maximally localized Wannier functions for entangled energy bands," *Physical Review B*, vol. 65, no. 3, p. 035109, 2001.

[49] G. Pizzi, V. Vitale, R. Arita, S. Blügel, F. Freimuth, G. Géranton, M. Gibertini, D. Gresch, C. Johnson, T. Koretsune, J. Ibañez-Azpiroz, H. Lee, J.-M. Lihm, D. Marchand, A. Marrazzo, Y. Mokrousov, J. I. Mustafa, Y. Nohara, Y. Nomura, L. Paulatto, S. Poncé, T. Ponweiser, J. Qiao, F. Thöle, S. S. Tsirkin, M. Wierzbowska, N. Marzari, D. Vanderbilt, I. Souza, A. A. Mostofi, and J. R. Yates, "Wannier90 as a community code: new features and applications," *Journal of Physics: Condensed Matter*, vol. 32, no. 16, p. 165902, jan 2020.

[50] K. C. Wang, D. Valencia, J. Charles, Y. He, M. Povolotskyi, G. Klimeck, J. Maassen, M. Lundstrom, and T. Kubis, "NEMO5: Predicting MoS2 heterojunctions," *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, pp. 221–224, 2016.

[51] P. Sarangapani, Y. Chu, K. C. Wang, D. Valencia, J. Charles, and T. Kubis, "Nonequilibrium Green's function method: Transport and band tail predictions in transition metal dichalcogenides," *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2018-September, pp. 38–39, 2018.

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

[53] A. Afzalian, T. Vasen, P. Ramvall, and M. Passlack, "An efficient tight-binding mode-space NEGF model enabling up to million atoms III-V nanowire MOSFETs and TFETs simulations," 2017. [Online]. Available: http://arxiv.org/abs/1705.00909

[54] Q. Chen, J. Li, C. Yam, Y. Zhang, N. Wong, and G. Chen, "An approximate framework for quantum transport calculation with model order reduction," *Journal of Computational Physics*, vol. 286, pp. 49–61, 2015.

[55] A. Kuzmin, M. Luisier, and O. Schenk, "Fast methods for computing selected elements of the green's function in massively parallel nanoelectronic device simulations," in *European Conference on Parallel Processing*. Springer, 2013, pp. 533–544.

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

[57] J. Charles, "Modeling Nonlocality in Quantum Systems," Ph.D. dissertation, Purdue University, 2018.

[58] M. Luisier and G. Klimeck, "Simulation of nanowire tunneling transistors: From the Wentzel-Kramers- Brillouin approximation to full-band phonon-assisted tunneling," *Journal of Applied Physics*, vol. 107, no. 8, 2010.

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

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

[61] V. H. Nguyen, F. Triozon, F. D. Bonnet, and Y. M. Niquet, "Performances of strained nanowire devices: Ballistic versus scattering-limited currents," *IEEE Transactions on Electron Devices*, vol. 60, no. 5, pp. 1506–1513, 2013.

[62] M. Aldegunde, A. Martinez, and J. R. Barker, "Study of individual phonon scattering mechanisms and the validity of Matthiessen's rule in a gate-all-around silicon nanowire transistor," *Journal of Applied Physics*, vol. 113, no. 1, 2013.

[63] R. Valin, M. Aldegunde, A. Martinez, and J. R. Barker, "Quantum transport of a nanowire field-effect transistor with complex phonon self-energy," *Journal of Applied Physics*, vol. 116, no. 8, 2014.

[64] H. Ehrenreich, "Screening effects in polar semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 8, pp. 130–135, 1959.

[65] K. Miao, S. Sadasivam, J. Charles, G. Klimeck, T. S. Fisher, and T. Kubis, "Büttiker probes for dissipative phonon quantum transport in semiconductor nanostructures," *Applied Physics Letters*, vol. 108, no. 11, 2016.

[66] S. Datta, "The non-equilibrium green's function (negf) formalism: An elementary introduction," in *Electron Devices Meeting, 2002. IEDM'02. International*. IEEE, 2002, pp. 703–706.

[67] A. Trellakis, A. Galick, A. Pacelli, and U. Ravaioli, "Iteration scheme for the solution of the two-dimensional schrödinger-poisson equations in quantum structures," *Journal of Applied Physics*, vol. 81, no. 12, pp. 7880–7884, 1997.

[68] T. Kubis, C. Yeh, P. Vogl, A. Benz, G. Fasching, and C. Deutsch, "Theory of nonequilibrium quantum transport and energy dissipation in terahertz quantum cascade lasers," *Physical Review B*, vol. 79, no. 19, p. 195323, 2009.

[69] M. Born, "Quantenmechanik der Stoßvorgänge," *Zeitschrift für Physik*, vol. 38, p. 803, 1926.

[70] T. Kubis, "Quantum transport in semiconductor nanostructures," Ph.D. dissertation, Technical University of Munich, 2009.

[71] Y. Lee, M. Bescond, N. Cavassilas, D. Logoteta, L. Raymond, M. Lannoo, and M. Luisier, "Quantum treatment of phonon scattering for modeling of three-dimensional atomistic transport," *Physical Review B*, vol. 95, no. 20, pp. 1–6, 2017.

[72] Y. Lee, M. Lannoo, N. Cavassilas, M. Luisier, and M. Bescond, "Efficient quantum modeling of inelastic interactions in nanodevices," *Physical Review B*, vol. 93, no. 20, pp. 1–14, 2016.

[73] H. Mera, M. Lannoo, C. Li, N. Cavassilas, and M. Bescond, "Inelastic scattering in nanoscale devices: One-shot current-conserving lowest-order approximation," *Physical Review B - Condensed Matter and Materials Physics*, vol. 86, no. 16, pp. 2–5, 2012.

[74] Y. Chu, J. Shi, K. Miao, Y. Zhong, P. Sarangapani, T. S. Fisher, G. Klimeck, X. Ruan, and T. Kubis, "Thermal boundary resistance predictions with non-equilibrium Green's function and molecular dynamics simulations," *Applied Physics Letters*, vol. 115, no. 23, 2019.

[75] J. Geng, P. Sarangapani, K. C. Wang, E. Nelson, B. Browne, C. Wordelman, J. Charles, Y. Chu, T. Kubis, and G. Klimeck, "Quantitative Multi-Scale, Multi-Physics Quantum Transport Modeling of GaN-Based Light Emitting Diodes," *Physica Status Solidi (A) Applications and Materials Science*, vol. 215, no. 9, pp. 1–7, 2018.

[76] J. Piprek, P. Abraham, and J. E. Bowers, "Self-consistent analysis of high-temperature effects on strained-layer multiquantum-well ingaasp-inp lasers," *IEEE Journal of Quantum Electronics*, vol. 36, no. 3, pp. 366–374, 2000.

[77] M. Clemens, E. Gjonaj, P. Pinder, and T. Weiland, "Self-consistent simulations of transient heating effects in electrical devices using the finite integration technique," *IEEE Transactions on Magnetics*, vol. 37, no. 5 I, pp. 3375–3379, 2001.

[78] D. W. Xu, S. F. Yoon, and C. Z. Tong, "Self-Consistent Analysis of Carrier Confinement and Output Power in 1.3-μm InAs–GaAs Quantum-Dot VCSELs," *IEEE Journal of Quantum Electronics*, vol. 44, no. 9, pp. 879–885, 2008.

[79] U. Russo, D. Ielmini, C. Cagli, and A. L. Lacaita, "Self-accelerated thermal dissolution model for reset programming in unipolar resistive-switching memory (RRAM) devices," *IEEE Transactions on Electron Devices*, vol. 56, no. 2, pp. 193–200, 2009.

[80] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive switching by voltage-driven ion migration in bipolar RRAMPart II: Modeling," *IEEE Transactions on Electron Devices*, vol. 59, no. 9, pp. 2468–2475, 2012.

[81] T. Sadi, L. Wang, L. Gerrer, V. Georgiev, and A. Asenov, "Self-consistent physical modeling of SiOx-based RRAM structures," *18th International Workshop on Computational Electronics, IWCE 2015*, no. 1, pp. 1–4, 2015.

[82] A. Padovani, L. Larcher, O. Pirrotta, L. Vandelli, and G. Bersuker, "Microscopic modeling of HfOx RRAM operations: From forming to switching," *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 1998–2006, 2015.

[83] W. Y. Choi, B. Park, J. D. Lee, and T. K. Liu, "Tunneling field-effect transistors (tfets) with subthreshold swing (ss) less than 60 mv/dec," *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 743–745, 2007.

[84] A. Seabaugh, C. Alessandri, M. A. Heidarlou, H. M. Li, L. Liu, H. Lu, S. Fathipour, P. Paletti, P. Pandey, and T. Ytterdal, "Steep slope transistors: Tunnel FETs and beyond," *European Solid-State Device Research Conference*, vol. 2016-Octob, pp. 349–351, 2016.

[85] A. Afzalian, G. Doornbos, T. M. Shen, M. Passlack, and J. Wu, "A High-Performance InAs/GaSb Core-Shell Nanowire Line-Tunneling TFET: An Atomistic Mode-Space NEGF Study," *IEEE Journal of the Electron Devices Society*, vol. 7, no. August 2018, pp. 111–117, 2019.

[86] Y. Li and H. M. Chou, "A comparative study of electrical characteristic on sub-10-nm double-gate MOSFETs," *IEEE Transactions on Nanotechnology*, vol. 4, no. 5, pp. 645–647, 2005.

[87] L. Knoll, M. Schmidt, Q. T. Zhao, S. Trellenkamp, A. Schäfer, K. K. Bourdelle, and S. Mantl, "Si tunneling transistors with high on-currents and slopes of 50 mV/dec using segregation doped NiSi2 tunnel junctions," *Solid-State Electronics*, vol. 84, pp. 211–215, 2013.

[88] Y. Lu, A. Seabaugh, P. Fay, S. J. Koester, S. E. Laux, T. W. Haensch, and S. O. Koswatta, "Geometry dependent tunnel FET performance - Dilemma of electrostatics vs. quantum confinement," *Device Research Conference - Conference Digest, DRC*, pp. 17–18, 2010.

[89] S. Takagi, a. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's:\nPart II-effects of surface orientation," *IEEE Transactions on Electron Devices*, vol. 41, no. 12, 1994.

[90] J. Z. Huang, P. Long, M. Povolotskyi, G. Klimeck, and M. J. Rodwell, "P-Type Tunnel FETs with Triple Heterojunctions," *IEEE Journal of the Electron Devices Society*, vol. 4, no. 6, pp. 410–415, 2016.

[91] K. Alam, "Orientation Engineering for Improved Performance of a Ge-Si Heterojunction Nanowire TFET," *IEEE Transactions on Electron Devices*, vol. 64, no. 12, pp. 4850–4855, 2017.

[92] R. Gottinger, A. Gold, G. Abstreiter, G. Weimann, and W. Schlapp, "Interface roughness scattering and electron mobilities in thin gaas quantum wells," *Epl*, vol. 6, no. 2, pp. 183–188, 1988.

[93] S. G. Kim, M. Luisier, T. B. Boykin, and G. Klimeck, "Effects of interface roughness scattering on radio frequency performance of silicon nanowire transistors," *Applied Physics Letters*, vol. 99, no. 23, 2011.

[94] M. Littlejohn, J. Hauser, T. Glisson, D. K. Ferry, and J. W. Harrison, "Alloy scattering and high field transport in ternary and quaternary III–V semiconductors," *Solid-State Electronics*, vol. 21, pp. 107–114, 1978.

[95] Y. Ye, F. Liu, M. Chen, S. Nassif, and Y. Cao, "Statistical modeling and simulation of threshold variation under random dopant fluctuations and line-edge roughness," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 6, pp. 987–996, 2011.

[96] L. Bellaiche and D. Vanderbilt, "Virtual crystal approximation revisited: Application to dielectric and piezoelectric properties of perovskites," *Physical Review B - Condensed Matter and Materials Physics*, vol. 61, no. 12, pp. 7877–7882, 2000.

[97] M. Frey, "Scattering in Nanoscale Devices," Ph.D. dissertation, ETH Zurich, 2010.

[98] A. Esposito, M. Frey, and A. Schenk, "Quantum transport including non-parabolicity and phonon scattering: Application to silicon nanowires," *Journal of Computational Electronics*, vol. 8, no. 3-4, pp. 336–348, 2009.

[99] R. Venugopal, M. Paulsson, S. Goasguen, S. Datta, and M. Lundstrom, "A simple quantum mechanical treatment of scattering in nanoscale transistors," *Journal of Applied Physics*, vol. 93, no. 9, pp. 5613–5625, 2003.

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

[101] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom, "Theory of ballistic nanotransistors," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1853–1864, 2003.

[102] X. Shao and Z. Yu, "Nanoscale finfet simulation: A quasi-3d quantum mechanical model using negf," *Solid-State Electronics*, vol. 49, no. 8, pp. 1435–1445, 2005.

[103] I. Markovsky, *Low rank approximation: algorithms, implementation, applications*. Springer Science & Business Media, 2011.

[104] L. Zeng, Y. He, M. Povolotskyi, X. Liu, G. Klimeck, and T. Kubis, "Low rank approximation method for efficient Green's function calculation of dissipative quantum transport," *Journal of Applied Physics*, vol. 113, no. 21, p. 213707, 2013.

[105] U. Hetmaniuk, D. Ji, Y. Zhao, and M. P. Anantram, "A reduced-order method for coherent transport using green's functions," *IEEE Transactions on Electron Devices*, vol. 62, no. 3, pp. 736–742, 2015.

[106] S. Birner, C. Schindler, P. Greck, M. Sabathil, and P. Vogl, "Ballistic quantum transport using the contact block reduction (cbr) method," *Journal of computational electronics*, vol. 8, no. 3, pp. 267–286, 2009.

[107] J. Z. Huang, H. Ilatikhameneh, M. Povolotskyi, and G. Klimeck, "Robust Mode Space Approach for Atomistic Modeling of Realistically Large Nanowire Transistors," *Journal of Applied Physics*, vol. 044303, pp. 1–9, 2017.

[108] A. Afzalian, T. Vasen, P. Ramvall, T. M. Shen, J. Wu, and M. Passlack, "Physics and performances of III-V nanowire broken-gap heterojunction TFETs using an efficient tight-binding mode-space NEGF model enabling million-atom nanowire simulations," *Journal of Physics Condensed Matter*, vol. 30, no. 25, 2018.

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

[110] D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl, "Contact block reduction method for ballistic transport and carrier densities of open nanostructures," *Physical Review B - Condensed Matter and Materials Physics*, vol. 71, no. 24, pp. 1–14, 2005.

[111] T. Zibold, P. Vogl, and A. Bertoni, "Theory of semiconductor quantum-wire-based single- and two-qubit gates," *Physical Review B - Condensed Matter and Materials Physics*, vol. 76, no. 19, pp. 1–14, 2007.

[112] S. Birner, C. Schindler, P. Greck, M. Sabathil, and P. Vogl, "Ballistic quantum transport using the contact block reduction (CBR) method," *Journal of Computational Electronics*, vol. 8, no. 3-4, pp. 267–286, oct 2009.

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

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

[115] U. Hetmaniuk, D. Ji, Y. Zhao, and M. P. Anantram, "A reduced-order method for coherent transport using green's functions," *IEEE Transactions on Electron Devices*, vol. 62, no. 3, pp. 736–742, 2015.

[116] A. Afzalian, "Computationally efficient self-consistent born approximation treatments of phonon scattering for coupled-mode space non-equilibrium Green's function," *Journal of Applied Physics*, vol. 110, no. 9, p. 094517, 2011.

[117] A. Afzalian, M. Passlack, and Y. C. Yeo, "Atomistic simulation of gate-All-Around GaSb/InAs nanowire TFETs using a fast full-band mode-space NEGF model," *2016 International Symposium on VLSI Technology, Systems and Application, VLSI-TSA 2016*, pp. 1–2, 2016.

[118] D. A. Lemus, J. Charles, and T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations," *Journal of Computational Electronics*, 2020, submitted. [Online]. Available: http://arxiv.org/abs/2003.09536

[119] J. Rudi, A. C. I. Malossi, T. Isaac, G. Stadler, M. Gurnis, P. W. Staar, Y. Ineichen, C. Bekas, A. Curioni, and O. Ghattas, "An extreme-scale implicit solver for complex PDEs: Highly heterogeneous flow in earth's mantle," *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, vol. 15-20-November-2015, 2015.

[120] C. Yang, W. Xue, H. Fu, H. You, X. Wang, Y. Ao, F. Liu, L. Gan, P. Xu, L. Wang, G. Yang, and W. Zheng, "10M-Core Scalable Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics," *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, vol. 0, no. November, pp. 57–68, 2016.

[121] H. Fu, C. He, B. Chen, Z. Yin, Z. Zhang, W. Zhang, T. Zhang, W. Xue, W. Liu, W. Yin, G. Yang, and X. Chen, "18.9-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of 18-Hz and 8-Meter Scenarios," *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2017*, 2017.

[122] W. Joubert, D. Weighill, D. Kainer, S. Climer, A. Justice, K. Fagnan, and D. Jacobson, "Attacking the opioid epidemic: Determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction," *Proceedings - International Conference for High Performance Computing, Networking, Storage, and Analysis, SC 2018*, pp. 717–730, 2019.

[123] A. N. Ziogas, T. Ben-Nun, G. I. Fernández, T. Schneider, M. Luisier, and T. Hoefler, "A data-centric approach to extreme-scale ab initio dissipative quantum transport simulations," *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2019.

[124] "Acm gordon bell prize," https://awards.acm.org/bell, 2020, [Online].

[125] T. Boykin, M. Luisier, and G. Klimeck, "Multiband transmission calculations for nanowires using an optimized renormalization method," *Physical Review B*, vol. 77, no. 16, p. 165318, apr 2008.

[126] "Lapack - linear algebra package," http://www.netlib.org/lapack/, 2020, [Online].

[127] "Stampede2 - texas advanced computing center," https://portal.tacc.utexas.edu/user-guides/stampede2, 2020, [Online].

[128] J. Jeffers and J. Reinders, *Intel Xeon Phi Coprocessor High Performance Programming*. Morgan Kaufmann, 2013.

[129] G. Klimeck, private communication, 2020.

[130] "About blue waters," http://www.ncsa.illinois.edu/enabling/bluewaters, 2020, [Online].

[131] S. Steiger and M. Povolotskyi, "Nemo5: a parallel multiscale nanoelectronics modeling tool," *Nanotechnology, ...*, vol. 10, no. 6, pp. 1464–1474, 2011.

[132] "About us - openmp," https://www.openmp.org/about/about-us/, 2020, [Online].

[133] "The netlib," http://www.netlib.org/, 2020, [Online].

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

[135] A. Svizhenko and M. P. Anantram, "Role of scattering in nanotransistors," *IEEE Transactions on Electron Devices*, vol. 50, no. 6, pp. 1459–1466, 2003.

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

[137] M. P. Lopez Sancho, J. M. Lopez Sancho, and J. Rubio, "Quick iterative scheme for the calculation of transfer matrices: Application to Mo (100)," *Journal of Physics F: Metal Physics*, vol. 14, no. 5, pp. 1205–1215, 1984.

[138] A. Afzalian, M. Passlack, and Y. C. Yeo, "Scaling perspective for III-V broken gap nanowire TFETs: An atomistic study using a fast tight-binding mode-space NEGF model," *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 30.1.1–30.1.4, 2016.

[139] "Silvaco, purdue team up to bring scalable atomistic tcad solutions for next generation semiconductor devices and materials," https://www.silvaco.com/news/pressreleases/2018_08_24_01.html, 2018, [Online].

[140] A. Wacker, "Semiconductor superlattices: A model system for nonlinear transport," *Physics Reports*, vol. 357, no. 1, pp. 1–111, 2002.

[141] R. Kronig, "Optical Society of America Review of Scientific Instruments," *Journal of the Optical Society of America*, vol. 12, no. 6, pp. 459–463, 1925.

[142] G. Todoran, R. Holonec, and C. Iakab, "Discrete Hilbert Transform . Numeric Algorithms," *Acta Electrotehnica*, vol. 49, no. 4, pp. 485–490, 2008.

[143] V. Čížek, "Discrete Hilbert Transform," *IEEE Transactions on Audio and Electroacoustics*, vol. 18, no. 4, pp. 340–343, 1970.

[144] B. L. Gelmont, M. Shur, and M. Stroscio, "Polar optical-phonon scattering in three- and two-dimensional electron gases," *Journal of Applied Physics*, vol. 77, no. 1, 1995.

[145] "Itap reseach computing - brown," https://www.rcac.purdue.edu/compute/brown/, 2020, [Online].

[146] S. Sarangapani, Y. Chu, J. Charles, G. Klimeck, and T. Kubis, "Non-equilibrium greens function method: Band tail formation in non-local polar optical phonon scattering." *International Workshop on Computational Nanotechnology (IWCN)*, 2017.

PUBLICATIONS

- D. A. Lemus, J. Charles, and T. Kubis, "Mode-space-compatible inelastic scattering in atomistic nonequilibrium Green's function implementations,", *Submitted to Journal of Computational Electronics*, 2020.
- T. Kubis, J. Charles, D. Lemus, "Method of modeling many particle systems", patent, publication number 2020-0104442, 2020
- T. Kubis, J. Charles, D. Lemus, "System architecture and methods of determining device behavior," patent, application number 16/588,046, 2020
- X. Guo, K. Wang, J. Charles, J. Geng, D. Mejia, D. Valencia, D. Lemus, J. E. Fonseca, G. Klimeck, T. Kubis, "NEMO5, Xeon Phi and hStreams: Physics of Ultrascaled 2D Nanotransistors," SC19, 2019.
- A. Afzalian, T. Vasen, P. Ramvall, D. Lemus, T. Kubis, M. Passlack, T. Shen, and J. Wu, "Performance evaluation of III-V nanowire broken-gap TFETs including elec-tron-phonon scattering using an atomistic mode space NEGF technique enabling million atoms NW simulations," Extended Abstracts of the 2017 International Conference on Solid State Devices and Materials, Sendai, 2017
- J. Charles, P. Sarangapani, R. Golizadeh-Mojarad, R. Andrawis, D. Lemus, X. Guo, D. Mejia, J. E. Fonseca, M. Povolotskyi, T. Kubis, and G. Klimeck, "Incoherent transport in NEMO5: realistic and efficient scattering on phonons," J. Comput. Electron., vol. 15, no. 4, pp. 1123–1129, 2016.
- A. Afzalian, J. Huang, H. Ilatikhameneh, J. Charles, D. Lemus, J. B. Lopez, S. P. Rubiano, T. Kubis, M. Povolotskyi, G. Klimeck, M. Passlack, and Y. C. Yeo, "Mode space tight binding model for ultra-fast simulations of III-V nanowire

MOSFETs and heterojunction TFETs," 18th Int. Work. Comput. Electron. IWCE 2015, pp. 4–6, 2015.

- J. D. Bermeo, J. Charles, J. Fang, J. Fonseca, Y. He, G. Klimeck, T. Kubis, D. Mejia, D. Lemus, M. Povolotskyi, S. A. Pérez, P. Sarangapani, and L. Zeng, "NEMO5 : Achieving High-end Internode Communication for Performance Projection Beyond Moore's Law," 2015 Gordon Bell Prize Submiss., 2015.
- Z. Jiang, M. Povolotskyi, N. Onofrio, D. Guzman, D. Lemus, "Multi-Scale Quantum Simulations of Conductive Bridging RAM." Oral presentation, 18th International Workshop on Computational Electronics, West Lafayette, Indiana, September 2-4, 2015.

VITA

VITA

Daniel A. Lemus was born in El Paso, Texas on February 6th, 1990 and began attending the University of Texas at El Paso in 2008. Majoring in Electrical and Computer Engineering with a concentration in computer engineering, Daniel gained a liking for programming and began working on TCP/IP socket programming and Linux embedded systems. After a summer spent with the Purdue Summer Undergraduate Research Fellowships (SURF) program, he learned about the NEMO5 software that the iNEMO group developed. In 2012, before graduating that year, Daniel applied to the iNEMO group and would be accepted soon after.

In 2012, Daniel began his PhD working with the iNEMO group helping develop the NEMO5 software. Although the iNEMO group worked with topics far outside his expertise, Daniel wanted his work to be a mix of what iNEMO was good at: quantum transport method development, and computer-engineering-oriented topics like high performance computing and low-rank approximations.

Upon graduation in 2020, Daniel hopes to continue interdisciplinary work by stepping into unfamiliar territory and applying his knowledge to new fields and topics.