DEVELOPMENT OF A MASSIVELY PARALLEL NANOELECTRONIC MODELING TOOL AND ITS APPLICATION TO QUANTUM COMPUTING DEVICES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Sun Hee Lee

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2011

Purdue University

West Lafayette, Indiana

To my parents, my wife Jun Young . . . and our son Jayin

ACKNOWLEDGMENTS

As I finish another chapter in my life at Purdue, I would like to thank all the people who have encouraged and inspired me to finish the long journey of a doctoral study.

First and foremost I offer my sincerest gratitude to my advisor, Prof. Gerhard Klimeck, who has supported me throughout my Ph.D. both academically and financially. He waited for me with patience and guided me in gradually evolving from a software engineer to a computational scientist. I appreciate his time and critical ideas in tool development, research and the writing of this thesis. I am also thankful for what I learned from his passion for work and interactions with peers, as well as from his dedication to the success of nanoHUB. I learned how to make my career successful as well as to succeed in life. I feel blessed to have him as my advisor, my mentor and my friend.

Prof. Mark Lundstrom, Prof. Leonid Rokhinson, Prof. Alejandro Strachan and Prof. Michelle Simmons deserve a special thanks as my thesis committee members. It indeed was the most challenging yet invaluable experience in my life to present and defend my work in front of some of the world's most renowned researchers. I would also like to pay my respect to Prof. Mark Lundstrom, Prof. Muhammad Alam, Prof. Supriyo Datta and Prof. Alejandro Strachan for offering insightful courses on solid-state and device physics.

I cannot help mentioning my only partner in the `NEMO3D-peta` development team, Dr. Hoon Ryu. He sticked with me through the times in Purdue, good or bad. I could never be such a passionate and brilliant person like him. I wish him all the best in his current and future endeavors. I also want to thank Zhengping Jiang for undertaking the initial code benchmarking work without any complaints. He is the most positive guy I have ever seen despite his overwhelming work load! I am sure he will become a great engineer as well as Ph.D. I owe Yui-Hong Matthias Tan for helping me in revising this thesis and my first journal paper. A hardworking student like him will eventually succeed for sure. I appreciate Junzhe Geng for helping me with the quantum dot simulations. I would like to thank Abhijeet for showing me how to *survive* in research by example. His endless efforts resulted in three valuable publications on interface charges in FinFET devices.

I would like to acknowledge Prof. Michelle Simmons for giving me a chance to collaborate with one of the best experimental quantum computing group in the world and for their hospitality during my stay in Sydney. I enjoyed working with such a passionate group of people - Dr. Suddhasatta Mahapatra, Bent Weber, Dr. Jill Miwa, Martin Fuechsle and many others. I hope to see many groundbreaking papers in near future! In addition, I enjoyed interacting with Prof. Andrew Dzurak and Dr. Andrea Morello as well as Dr. Wee Han Lim and Fahd Mohiyaddin for the past two semesters. Although it was a little bit of a pain for me to run a set of simulations that took a couple of weeks to complete, to generate single plot, they always made me feel that I meant something to them. I look forward to continuing the work with these great people at UNSW.

I have been blessed for being with a fantastic group of fellow students in our lab. Guys like Parijat Sengupta, Mehdi Salmani, Yahoua Tan, Sunggeun Kim, Honghyun Park, Saumitra Mehrotra and Ganesh Hegde always listened to me with open hearts and shared many thoughts not limited on nanoelectronics. Thank you and best wishes to you all. Since I came here, my high school (SSHS) and college (SNU) alumni have helped me a lot mentally to overcome many difficulties in Ph.D. life. Especially, I appreciate the hospitality offered from Hoosang Ko, Changwook Jeong and Yongwook Choi's family. Junior alumni, Hwun Park, Jayoung Park and Seokmin Hong were always beside me throughout my Ph.D. years. I wish you the very best in your future.

I would like to express my deepest gratitude to my family, especially to my parents. My father, Sang Eun, who is the most skilled Urology surgeon in Korea has always been my role model. His diligence and hard work to be the best in his field has always inspired me to be a good engineer. As I move on, I will follow his spirit to be the best engineer in the field. My mother, Eun Jin, showed the greatest love of all and stood by my side all the time. A life without her unconditional support is impossible to imagine. I also wish all the best for my younger sister Jin Hee, her husband Soo Heon and my lovely nephews, Yuri and Yumin.

Last but not least, I cannot imagine my life without my wife Jun Young and my son Jayin who came into my life when I was studying at Purdue. They made my Ph.D. years the busiest, yet the happiest moments of my life.

Thank you all.
Sun Hee Lee
On a cold and windy day of November at Purdue

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

Appendix Table Page

LIST OF FIGURES

Figure Page

Appendix Figure Page

Appendix Figure Page

Appendix Figure Page

ABSTRACT

Lee, Sun Hee Ph.D., Purdue University, December 2011. Development of a Massively Parallel Nanoelectronic Modeling Tool and Its Application to Quantum Computing Devices . Major Professor: Gerhard Klimeck.

The rapid progress in nanofabrication technologies has led to the possibility of realizing scalable solid-state quantum computers (QC) which have the potential to outperform conventional microprocessors. Using STM patterning followed by low temperature MBE, phosphorus doping in silicon can be controlled with atomic precision to fabricate Si:P delta doping layers for semi-metallic contacts and leads as well as for few-donor quantum dots (QDs). To model Si:P quantum dot systems in realistic domains, we have developed a nanoelectronic modeling tool (NEMO3D-peta) to expand existing NEMO3D capabilities with advanced parallelization schemes and to provide more flexibility to the code through an object-oriented design approach. Benchmark studies are performed on various aspects and NEMO3D-peta is proven to scale up to 32,000 processors. Furthermore, a variety of applications including a charge-potential self-consistent module are implemented. In the first step of QD system modeling with NEMO3D-peta, the electronic structure of delta doped contacts under equilibrium condition is computed self-consistently based on an atomistic $sp^3d^5s^*$ tight-binding Hamiltonian. Bandstructure results are compared against previous *ab-initio* studies and shown to be in good agreement in terms of valley minima and Fermi level positions. In addition, the effect of random dopant disorder in the delta layers is investigated in extended domains. Although fluctuations in the density of states (DOS) in a disordered supercell are expected, a high DOS will exhibit little effects on the conduction properties of Si:P delta doped layers. Finally, work is in progress on predicting valley splitting (VS) and excited states in electrostatically

defined silicon QDs in the single electron regime. From the simulation studies with NEMO3D-peta, VS in the dot can be controlled by tuning the barrier, top gates and top gate geometry. Correct prediction of valley states may lead to a noise-tolerant QC platform that exploits the valley degree of freedom in silicon QDs. With the potential of NEMO3D-peta, we expect to provide useful insight and expertise to the quantum electronics community.

1. INTRODUCTION

Building quantum computers (QC) in silicon has intrigued many researchers around the world since the proposal of Kane to use the nuclear spin of phosphorus donor as the basic quantum information unit - called *qubit* [1]. Silicon is the most promising semiconductor material for making spin-based quantum computers since over 90% of the isotopes in nature have zero net spin (^{28}Si, $I = 0$), and with additional purification processes, the composition of ^{28}Si can be upgraded to up to 99.9% [2]. Having a net spin zero environment significantly reduces decoherence for spin based qubits and retains spin information for timescales exceeding milliseconds. Moreover, silicon technology continuously developed for last 40 years and expertise in silicon fabrication meets the demands of most state-of-the-art technologies. The advantages and maturity of silicon technologies increase the possibility of realizing scalable quantum computers that may integrate and operate on $10^3 \sim 10^6$ qubits simultaneously.

There are several notable proposals of realizing qubit in silicon. Kane first proposed the use of the nucleus spin of phosphorus donors in silicon for a qubit - with the electrons used for mediating nucleus spin interaction and for spin information readout [1]. Hollenberg et al. proposed a way to utilize the charging state of two buried dopants as qubits. This proposal might resolve the difficulties in nuclear spin readout in Kane's architecture and enable fabricating scaled qubit architectures [3]. Another widely referenced scheme proposes to use the electron spin for a qubit confined in a quantum dot [4–7]. Using the electron spin as a qubit, in particular, has a couple of advantages over nuclear spin QC; the qubit operation is much faster, and spin state readout circuits and CNOT gate realization becomes simpler because an additional mediating process using electrons is required to read the qubit information coded in the nuclear spins [5, 7]. Irrespective of the different qubit types, it is evident that in silicon QC structures the basic idea includes implementation of quantum dots - either

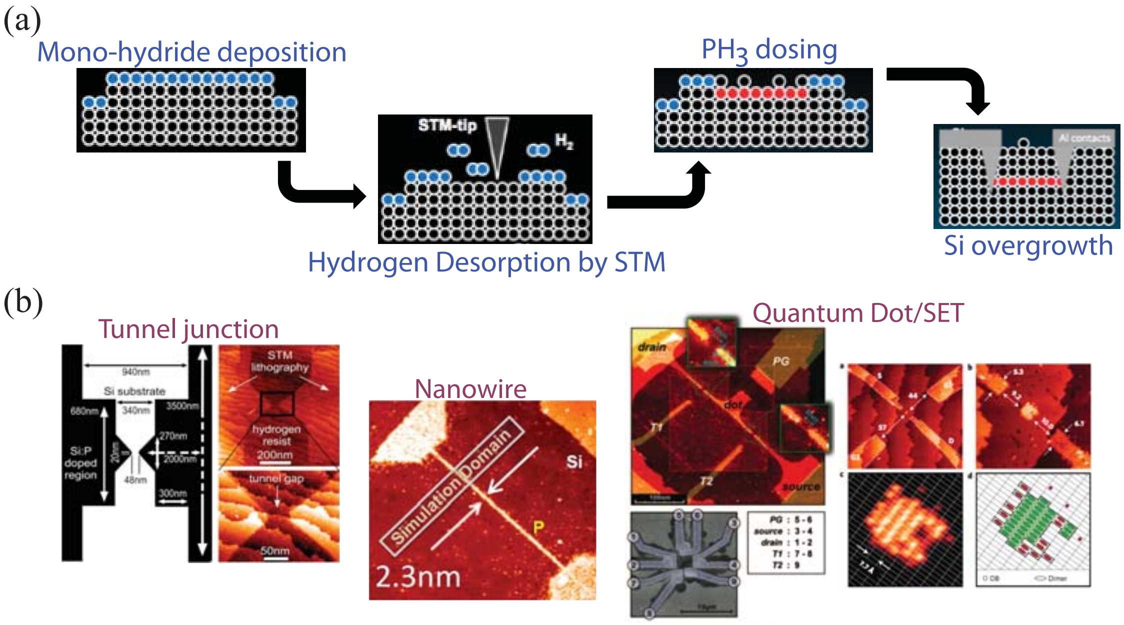


Fig. 1.1. (a) Simplified STM fabrication steps. Starting with a hydrogen passivated silicon surface, hydrogen atoms are removed from silicon using STM lithography. Then the region is dosed with phosphine gas (PH_3) for subsequent phosphorus adsorption. Finally, the silicon-phosphorus (Si:P) layer is capped with silicon to complete a planar structure. (b) Evolving device geometry (2D→1D→0D) towards quantum dot design using STM+MBE technology [8–11]. (Ref. [8] reprinted with author's permission - waiting for author's response. Ref. [9] reprinted with author's permission. Ref. [10] reprinted with permission. Copyright 2009, American Chemical Society. Ref. [11] reprinted by permission from Macmillan Publishers Ltd: Nature Nanotechnology, copyright 2010.)

by ionized donors, electrostatic potential or self-assembly processes - to manipulate single electron or nucleus states.

One of the most challenging aspects of fabricating donor based qubits is the precise donor placement in both vertical and lateral directions as well as uniform gate patterning. Fortunately, recent advances in nanofabrication technology are bringing donor based nucleus and electron qubits closer to reality. One of the techniques utilizes scanning tunneling microscopy (STM) to carefully pattern and place dopant atoms with few atomic monolayer precision and low temperature molecular beam epitaxy (MBE) for silicon encapsulation [12, 13]. The overall fabrication steps are explained in Fig. 1.1(a). A highly optimized STM lithography technique is used to selectively tip-off hydrogen atoms passivated on the silicon surface. The exposed parts of the silicon surface are then adsorbed with phosphine gas (PH_3). In the next step, annealing to temperatures of $350\sim500$°C removes all hydrogen atoms from the surface. Finally, low temperature MBE is applied to encapsulate the phosphorus dosed plane (Si:P δ-layer) with silicon and to prevent donors from diffusing out of the plane. This technology facilitates production of a variety of structures with ultra-high doping densities of up to $2.3\times10^{14}(\mathrm{cm}^{-2})$ and with atomic precision as shown in Fig. 1.1(b). Furthermore, this technology has now reached the stage, which enables single donor quantum dot devices with in-plane bias control.

From a theoretical point of view, single phosphorus donor in bulk silicon is carefully analyzed and successfully characterized with tight-binding parameters to match the shallow donor levels with excited states by Rahman et al. This parameterization has led to the successful explanation of experimentally measured energy spectra in finFET devices in which single donor is embedded near the $\mathrm{Si{-}SiO_2}$ interface [14]. Building on the knowledge of single donor physics, the patterned δ-doped Si:P systems used in contacts and leads for quantum dot systems is modeled self-consistently under the mean-field approximation [15]. The main advantage of adopting a tight-binding approach over previously studied methods is that the computational burden becomes manageable enough to handle both 2D (contacts) [16] and 1D (leads) [17] periodic structures within reasonable times. The applied tight-binding approach is based on correct silicon bulk bandstructure information and the self-consistently computed bandstructure results are in good agreement with previous calculations. In addition, it is also possible to explore the effect of true disorder in extended simulation domains and to investigate whether disorder in such semi-metallic layers will affect the electronic properties of the contacts and leads. As a result, setting up reasonable methodologies and understanding the electronic structure of phosphorus

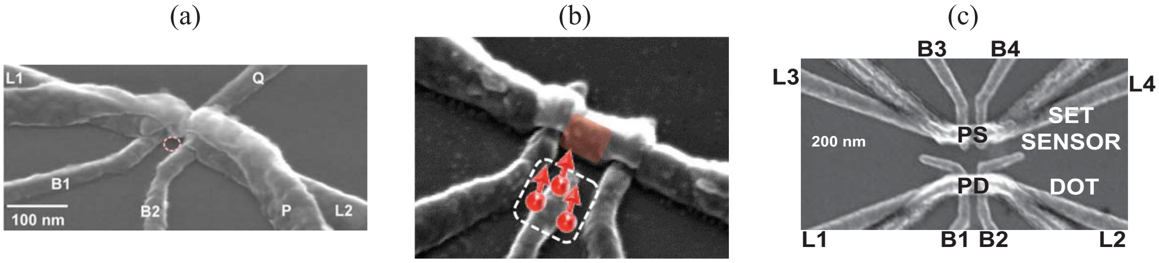


Fig. 1.2. STM images of a silicon quantum dot (Si QD) fabricated using metal-oxide-semiconductor technology. (a) Single QD [18] (reprinted with author's permission), (b) donor implanted QD with spin read-out design [19] (reprinted by permission from Macmillan Publishers Ltd: Nature, copyright 2010) and (c) double Si QD with SET sensor [20].

δ-doped layers in silicon plays a fundamental role in modeling donor-based quantum dot systems.

In contrast to recent bottom-up fabrication technologies applied for building donor based qubits, there are efforts being made to build gate defined silicon quantum dots (Si QDs) using state-of-the-art metal-oxide-semiconductor (MOS) technology [18, 21, 22] as shown in Fig. 1.2. Despite the advantage of long spin coherence times of electrons in ^{28}Si, the operability and quality of silicon quantum dots is hampered by interface states at the oxide interface and six-fold valley degeneracy [23]. Recent efforts in making clean oxide interfaces and designing top metal gates [18, 24–27] enabled fabrication of high-quality Si QD and control of the electron occupancy in the dot down to single electron by using proper gating schemes [18]. Moreover, detailed excitation spectra, which are essential for controlling spin-filling in the dot, can be observed in the fabricated QD via magnetospectroscopy measurements [26]. To support experimental observations of the spectrum in Si QD with simulations, the following critical elements have to be considered. First, the simulation should be able to capture the correct valley physics of silicon. Next, a self-consistent scheme to determine the potential shape of the quantum dot is required to work at low-temperature conditions. Finally, the simulator should be capable of including the whole quantum dot

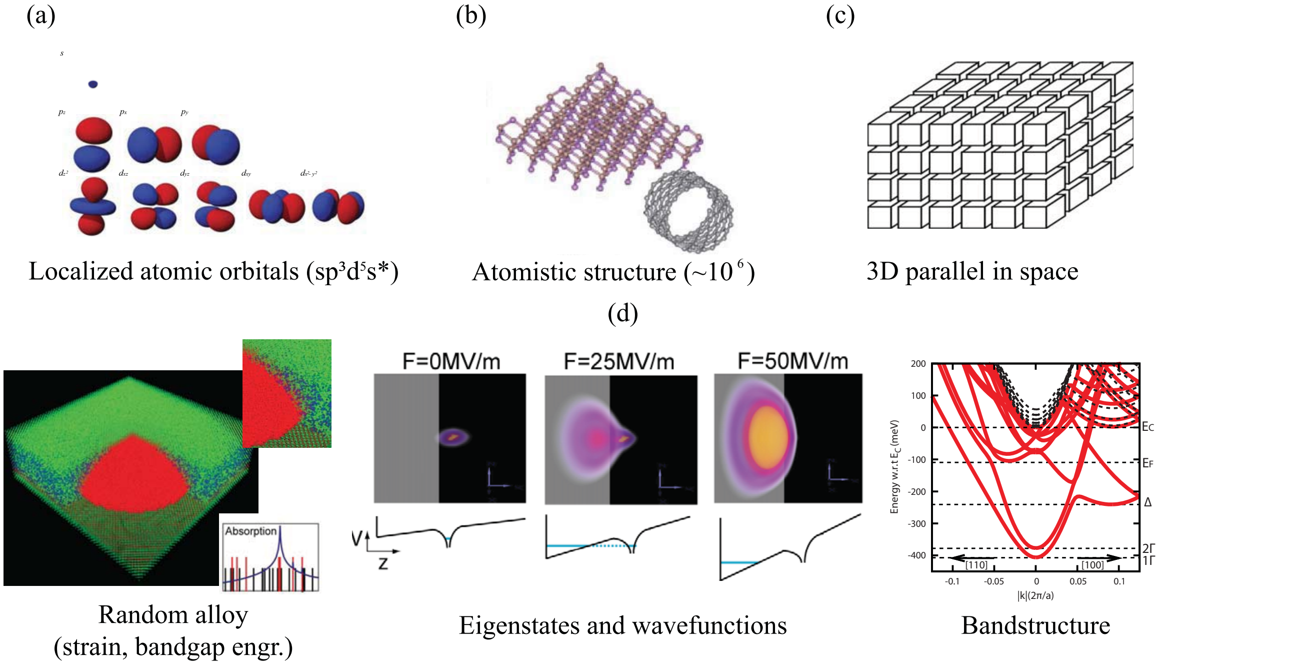


Fig. 1.3. (a) The wavefunctions for each atom are represented as localized orbitals (s, p, d like orbitals). (b) The device with realistic sizes is constructed atomistically, which in many cases exceeds $\sim 10^6$ atoms. (c) The simulation domain is spatially distributed over multiple processors for parallel operation. (d) As a result, various quantum effects in nanoscale devices such as random alloy effects and confined quantum states can be resolved.

region into the simulation domain. In summary, the simulator must be massively parallelized to handle matrices with multi-million degrees of freedom.

To solve challenging problems as discussed previously, a significant amount of time and effort was spent to implement the nano-electronic modeling tool `NEMO3D-peta`, which includes all features of `NEMO3D` [28–30] as well as significant improvements in code implementation, performance and physics. Similar to `NEMO3D`, `NEMO3D-peta` uses an atomistic tight-binding model for electronic structure calculations (Fig. 1.3 (a)∼(b)) with parallelization engine to handle realistic devices such as quantum dots (Fig. 1.3(c)). Such atomistic description enables examination of various nanoscale features, such as, strain in random alloys or confined quantum states. (Fig. 1.3(d)).

Several key features of NEMO3D-peta include 1) a significantly improved 3D parallelization engine for better scalability in a growing supercomputing environment, 2) greater flexibility by adopting an object-oriented coding scheme, 3) an improved and stable eigenvalue solver for million atom structures and 4) a general self-consistent loop, which can also be applied for identifying the electronic structure of Si:P δ-layers or for silicon quantum dot simulations. The great potential of NEMO3D-peta not only contributed to solving the specific challenges as mentioned above but also helped establish collaborations with peers and experimentalists [31–33] and foster the broader nanoelectronic community through dissemination of a user-friendly 1D heterostructure simulation tool, 1d-hetero via nanoHUB [34, 35].

Furthermore, important quantum computing related physics concepts are blended into NEMO3D-peta. The bandstructure of bulk silicon is carefully parameterized for an atomistic tight-binding formalism to match experimentally observed bandgap and effective masses using genetic algorithm by Klimeck et al. [36–38]. Rahman et al. parameterized phosphorus in silicon with ionized screened potentials and successfully predicted ground and excited state spectra under external electric fields [14, 39, 40]. Boykin et al. studied valley splitting (VS) in silicon or strained silicon quantum well for a variety of cases in atomistic tight-binding [41–44] and successfully explained experimentally observed VS in 2° miscut Si-SiGe quantum wells [45, 46].

In summary, NEMO3D-peta is a powerful electronic structure calculation tool that can compute eigenstates and bandstructure of realistic structures in an atomistic manner. A self-consistent scheme is included by default based on the mean-field theory. NEMO3D-peta is also a massively parallelized tool and compatible with numerous available clusters [47]. The tool incorporates essential physics from NEMO3D relevant for many quantum computing applications, and has proven to be successful in modeling and predicting numerous experimental studies over the past years. Therefore, NEMO3D-peta will be introduced, evaluated and applied on a number of physics problems of interest in the latter part of this document. Readers can find relevant publications involving NEMO3D optimization in Refs. [29, 48–50]. Also, simulation studies and experimental collaborations using `NEMO3D-peta` are documented in Refs. [9, 16, 17, 31–33, 35, 47, 51].

This thesis is organized as follows. Chapter 2 discusses design and performance aspects of `NEMO3D-peta`. Chapter 3 introduces the basics and optimization work of the Lanczos eigenvalue solver algorithm - the most critical component in solving the Schrödinger equation in multi-million atom systems. Chapter 4 presents the self-consistent methodology used to identify the electronic structure of phosphorus δ-doped layers in silicon. Chapter 5 shows recent progress on setting up the self-consistent methodology for measuring valley splitting in gate defined silicon quantum dots.

In addition, several self-study topics that may help readers broaden basic physical and numerical insight are discussed in the Appendix. Appendix A introduces the tight-binding formalism and its implementation into the code. Appendix B discusses Newton's method for solving the Poisson equation on a 1D example. A numerical DOS calculation is presented in Appendix C with an example Matlab code. Finally, Coulomb and exchange integrals used in multi-electron Hamiltonians and exciton energy calculations used in tight-binding are summarized in Appendix D.

2. DEVELOPMENT OF THE NANO-ELECTRONIC MODELING TOOL, NEMO3D-PETA

2.1 Introduction

Need for Atomistic Modeling: As semiconductor structures are scaled down to deca-nano sizes the underlying material can no longer be considered continuous. The number of atoms in the active device region becomes countable in the range of 50,000 to around 1 million and their local arrangement becomes critical in interfaces, alloys, and strained systems. An atomistic modeling approach needs to be used to capture such discreteness as well as quantum mechanical effects. Most experimentally relevant structures are not infinitely periodic, but finite in size and contain contacts. A local orbital basis is favored in these geometries over a plane wave basis, which implies infinite periodicity. Furthermore, we are primarily interested in stable semiconductor structures with well-established bonds, which lessens or even eliminates the need to be able to compute the formation of bonds with a full *ab-initio* methodology.

Multi-Million Atom Simulations: `NEMO3D` [28,29] uses empirical sp^3s^* and $sp^3d^5s^*$ tight binding models that have been carefully calibrated against bulk materials in III-V [52] and Si/Ge [38,53] material systems under various bulk strain and composition configurations. This bulk parameterization is transferred to the nanoscale under the assumption of weak charge redistributions. Weak piezo-electric effects in InGaAs systems can be captured through strain derived charge and electrostatic potential corrections [30,50]. Transferability of the bulk parameters to nanometer devices was demonstrated by experimentally verified multi-million atom calculations for valley splitting in Si on SiGe [46], single impurities in Si FinFETs, and InAs quantum dots in an InGaAs buffer matrix [54]. In these simulations none of the bulk parameters were modified and nominal device dimensions were used to obtain quantitative agreement with experiment. The simulations also showed that it is essential to include millions of atoms in the simulation domain and that simplified effective mass models lead to wrong conclusions.

Computational Cost: Multi-million atom calculations in NEMO3D come, however, at a typical computational price of 4-10 hours runtime on 20-64 cores on a standard cluster for a single evaluation of the eigenvalue spectrum. Inclusion of a one pass electronic structure calculation into a self-consistent Poisson solution is possible, but increases the computation time up by another factor of 6-20. This pushes the total computational simulation time into the realm of days, rather than hours. In spite of huge investments made into Peta-Scale computing and availability of over 100,000 cores on a single supercomputer, efficient parallelism is still vital for computational speed-up. We have been able to demonstrate NEMO3D scaling to up to 8,196 processors [49]. However, such high levels of scaling can only be achieved for unrealistically long and essentially 1-D structures, due to the 1-D spatial parallel decomposition of NEMO3D. NEMO3D therefore needs significant improvements in terms of parallelization schemes, data handling, post-processing, and code maintainability.

NEMO3D-peta: The major purpose of restructuring the engine of NEMO3D-peta emerged out from the need for expandability in growing processor-rich environments. NEMO3D-peta is equipped with a more powerful parallelization engine as well as a 3-D domain decomposition scheme and support of general multi-level parallelism. In addition, self-consistent charge calculations, which need additional computational power, are built in to facilitate various kinds of scientific simulations from impurity physics to device applications.

This chapter is organized as follows. Section 2.2 discusses the overall design schematics of NEMO3D-peta. In Section 2.3, the parallelization schemes in NEMO3D-peta are presented. Benchmark results are discussed in Section 2.4. We will go through an example of a self-consistent calculation in Section 2.5. Finally, we will summarize and conclude this chapter in Section 2.6.

2.2 Implementation hierarchy in NEMO3D-peta

The design block of `NEMO3D-peta` is shown in Fig. 2.1. `NEMO3D-peta` is built on an object-oriented concept using C++, which makes the code flexible enough to implement various kinds of quantum mechanics application in a plug-and-play manner. A *class* is a minimal set of variables and functions required to perform certain task and stores information. For example, the task of the geometry class is to construct specific device structures as defined in the input file in either atomistic or (second) nearest neighbor cubic grid and also holds the position information. The class may be interdependent, i.e. it may be needed by other classes to generate meaningful information required by other classes. For example, if the Hamiltonian class is to create and store Hamiltonian, the geometry and material information is definitely required. A *module* is defined as an independent application that connects relevant classes to operate in an end-to-end operation and to generate final outputs (converged potentials and charge profiles, eigenstates, relaxed atom positions and etc.). For instance, the eigenstates of a silicon quantum well from an atomistic tight-binding Hamiltonian can be computed in the bandstructure module. The bandstructure module first constructs a geometry class for an atomistic grid (i.e. zincblende crystal) with periodicity information. Then the module constructs a Hamiltonian class and links it with the previously created geometry class to build a tight-binding Hamiltonian. The Hamiltonian class is then fed into the eigenvalue solver to perform the eigenvalue calculation. Finally, the bandstructure module generates the output file that can be handled by the user. Adjustable parameters, such as, the structure geometry and different materials, input and output options, numerical parameters and etc., are specified in the input file and each option is selectively used by relevant classes.

The numerical aspects of `NEMO3D-peta` are described in Fig. 2.2. Since the code has to handle matrices of exceptionally large size ($10^7 \sim 10^8$), any computational issues, such as, parallelization, efficiency (speed) and storage (memory) as well as interfacing with external numerical packages should be readily taken into account.

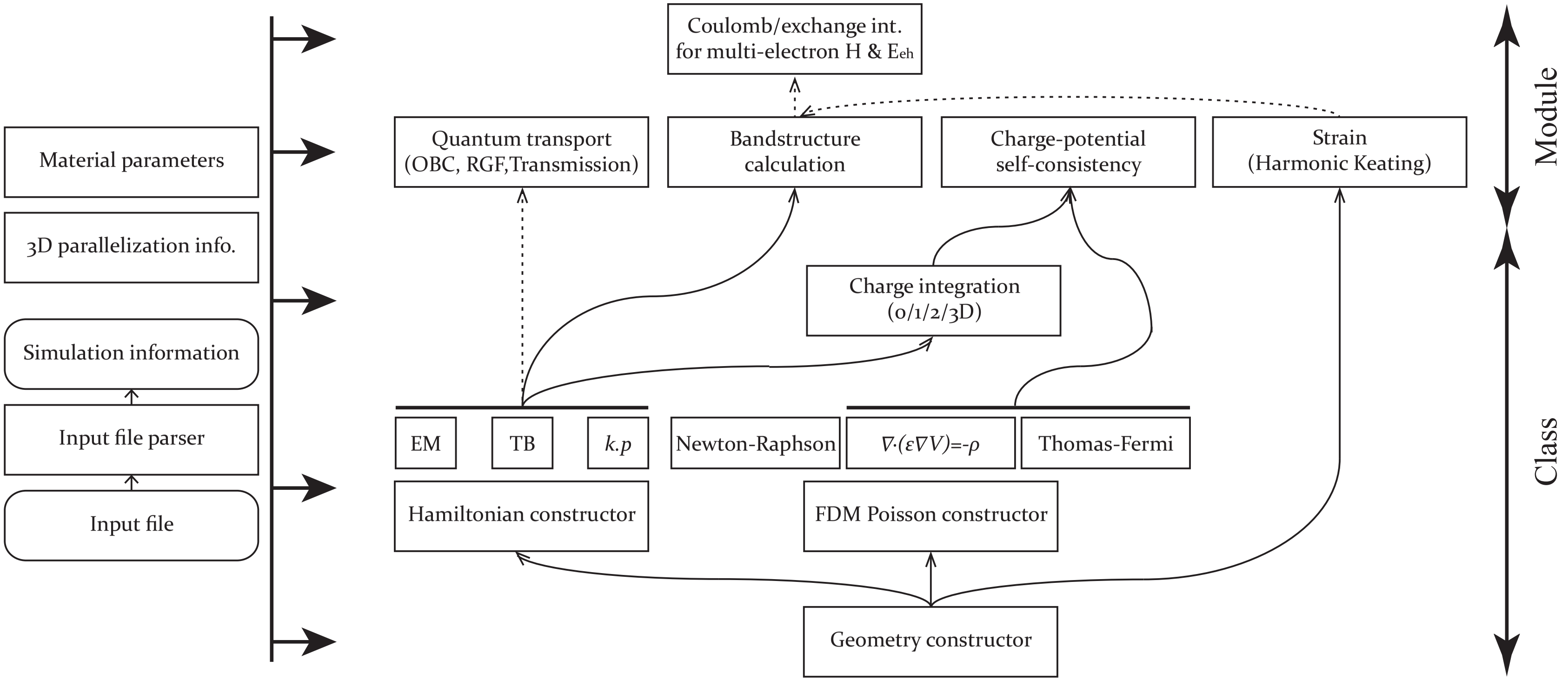


Fig. 2.1. A connectivity diagram of NEMO3D-peta prototype. It is implemented in an object-oriented manner to provide a simpler and more flexible implementation framework. Each *module* is comprised of multiple *classes* to perform an independent end-to-end task in a parallel computing environment. The dashed lines indicate either limited or soon-to-be implemented features.

As shown in Fig. 2.2, three different kinds of computational problems are categorized in `NEMO3D-peta` – eigenvalue problems in electronic structure calculations, the differential equation solver for Poisson and drift-diffusion problems and the minimization routine using conjugate gradient method used in solving strain energy minimization.

First, the electronic structure calculation in `NEMO3D-peta` is essentially solving a time-independent Schrödinger equation $H\Psi = E\Psi$, which is directly converted to an eigenvalue problem of a sparse Hermitian matrix. There are two different types of eigenvalue solvers embedded; the Lanczos algorithm [55] and the PARPACK package utilizing an implicitly restarted Arnoldi method [56]. Both solvers have similar background algorithms but follow different numerical paths depending on the size of the Hamiltonian as indicated in Fig. 2.2. The PARPACK solver is usually applied to relatively small devices to quickly obtain eigensolutions. It is fast but requires more memory space for matrix inversions and its application is therefore limited by the matrix dimension. On the other hand, the Lanczos algorithm is mostly used to solve for the eigenvalues of matrices with multi-million degrees of freedom. It is ten times slower than PARPACK but highly optimized with native implementations to occupy minimal memory space. The details of the Lanczos algorithm customized for million-atom quantum dot applications is further discussed in Chapter 3.

The next class of implementation is a general non-linear differential equation solver, which is used in solving Poisson equations as well as continuity equations for drift-diffusion. It is designed to solve $Ax - f(x) = 0$ where A is a linear differential operator and $f(x)$ is a non-linear function of x. The solution can be easily achieved using iterative schemes such as Newton's method [57]. The differential equation solver is not as computationally intensive as the eigenvalue solver, but it is memory limited due to fine grid spacing used in the finite difference scheme. The Poisson equation and Newton's method are further discussed in Appendix B.

Finally, the conjugate gradient (CG) minimization algorithm is used to find optimal atom configurations with a minimum of overall strain energy. Atoms positions are considered to be stable and are generally distorted within 10% under strain. Similar to the Newton's method, the CG algorithm is an iterative scheme that gradually converges to the minimum point in energy.

In summary, NEMO3D-peta is designed to be modular and object oriented to provide more flexibility NEMO3D failed to provide. NEMO3D-peta is capable of solving three major classes of numerical problems – eigenvalue, differential equation and minimization solvers. Every component in NEMO3D-peta is massively parallelized, compatible with many different clusters and allows handling of exceptionally large quantum dot systems containing multi-million atoms.

2.3 Parallelization Scheme in NEMO3D-peta

The major feature in NEMO3D-peta is its enhanced parallelization algorithm. Possible solutions include adding an additional degree of freedom in the spatial domain space and distributing more processors in k-space used for bandstructure and charge calculation, which will be discussed in the following subsections.

2.3.1 3-D Spatial Domain Decomposition

The domain decomposition method is a widely used concept in solving boundary value problems in partial differential equations. It subdivides the problem into fragments of smaller problems and coordinates the solution between them to obtain a true solution. In the spatial domain decomposition method, the problem is spatially decomposed into smaller problems. This method utilizes parallel computation to reduce execution times. NEMO3D-peta which is in essence a partial differential equation solver (Schrödinger equation, Poisson equation and etc.), can immediately benefit from this method since stable atom positions or grid points are assumed and mostly sparse matrices are encountered. For simple atomic structures it is possible to conceive a domain decomposition scheme as shown in Fig. 2.3.

The existing NEMO3D-peta utilized a 1-D spatial decomposition scheme for parallelism. NEMO3D-peta has been tested on many supercomputers and has proven to

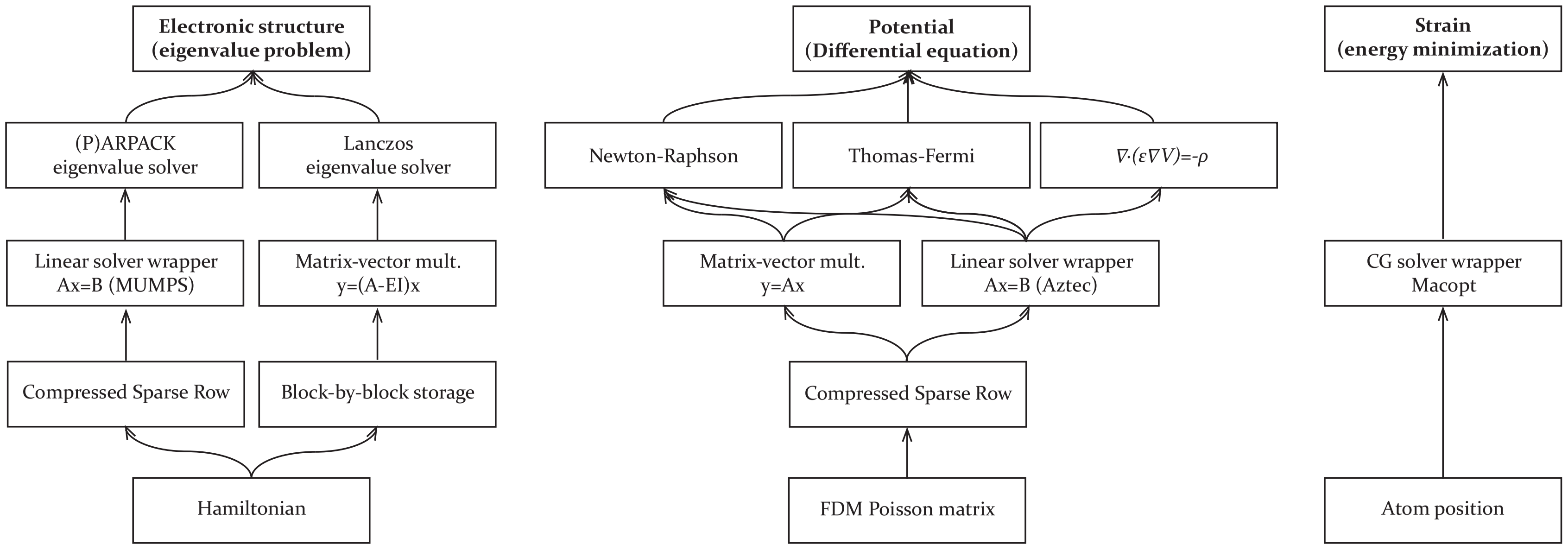


Fig. 2.2. The three different classes of numerical solvers implemented in `NEMO3D-peta`. (Left) Eigenvalue solver using either a native implementation of the Lanczos algorithm for large systems or the PARPACK solver package for fast computation of relatively small matrices. (Center) Non-linear differential equation solver using Newton-Raphson's method applied to Poisson equation. (Right) Energy minimization of the strained system using conjugate gradient method. Displaced atom positions can be computed to achieve minimum energy of the system.

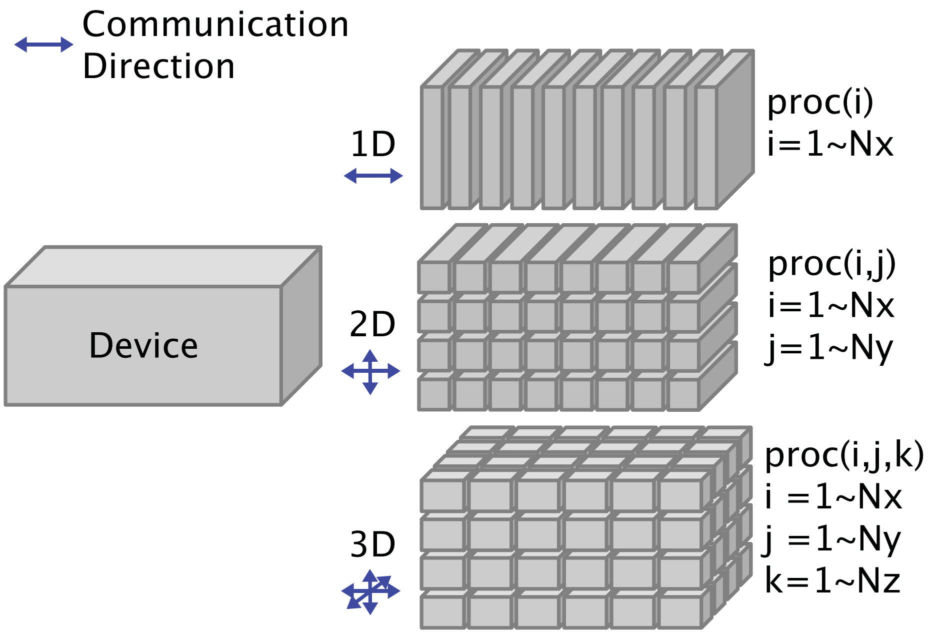


Fig. 2.3. Graphical representation of 1-D/2-D/3-D spatial domain decomposition scheme. Only 1-D domain decomposition was implemented in the existing `NEMO3D-peta`. Communication complexity increases by 2×(dimension).

show close to perfect scalability. The maximum utilizable processors, however, was strongly limited by the geometry in the 1-D decomposition. Regardless of the size of the structure, the maximum number of processors could not exceed the number of unit cells in x direction.

To reduce compute times when using more than 10,000 cores, a new domain decomposition scheme is introduced in `NEMO3D-peta`. In `NEMO3D-peta`, devices of any shapes can be spatially decomposed in three dimensions and each subdomain is assigned to a corresponding processor. An example of a 2-D parallel geometry construction decomposed by 16 processors is shown in Fig. 2.4.

a. Set up a geometry as defined in the input deck. Total number of unit cells and atoms are estimated based on the definition of the domain.

b. Since the processors already have the information of the global device structure, boundaries are independently determined by each processor. The region each processor covers in the simulation domain is defined as *sub-domain*.

c. For instance, let us investigate how each sub-domain is constructed. In this example, CPU (3,2) is considered.

d. Fill up atoms based on the unit cell definition (e.g. Zincblende) and store atom positions, atom types, ion types and domain identifiers. Note that the global list of atom information is not needed; each processor only considers atoms belonging to its own sub-domain.

e. Each atom is connected with other atoms which lie in the sub-domain based on relative positions of its neighbors. Depending on how the unit cell is defined, n-th nearest neighbor configuration is also possible.

f. Since each processor does not have any knowledge about atoms outside its own sub-domain, exchange of information is needed between processors. Fortunately, most atom configurations are limited by nearest neighbors; thus only the outermost atoms are considered and communication between adjacent processors - in this example, only CPUs (2,3), (3,3), (4,3), (2,2), (4,2), (2,1), (3,1), (4,1) are involved in the communication process. The Message Passing Interface (MPI) communication scheme is used to exchange atom information.

g. After the communication, remaining connectivity between outermost atoms with atoms adjacent to this sub-domain is completed, which finalizes the device construction.

The advantage of the proposed scheme is four-fold; first, each processor only has the list of information of the atoms in its subdomain and neighbor atoms from adjacent sub-domains. Since no processor holds only local information and atom (grid) information obtained from adjacent sub-domains, it is not bound by memory per core (typically 2GB per core), which enables to simulate multi-million atom systems, such as, quantum dots. Secondly, we are not limited by the geometry of the device since this scheme is designed to decompose the domain arbitrarily in all three dimensions. Third, this scheme is not limited to only atomistic geometry as shown in the above example; it can be generally used to generate uniform spatial grids for finite difference method (FDM) solvers, such as, Poisson solvers, drift-diffusion (DD) solvers, effective mass and $k \cdot p$ Hamiltonians, all of which are already implemented and active in `NEMO3D-peta`. Finally, it is applicable to any growing supercomputing environment [58] and has already been tested on various platforms, such as Jaguar [59], Kraken [60], Ranger [61], Trestles [62] and various Purdue clusters [63].

The drawback for 3-D parallelization is the increase of complexity in communication by $O(n^{N_{Dim}})$. The increased coupling among processors may cause significant performance degradation. In fact, there is a trade-off between reducing the computational burden and increasing the communication overhead. However, benchmark results indicated the average time consumed in the MPI communication to be typically 5% of the total simulation time. Moreover, from `NEMO3D-peta` cases, it was shown that the total simulation time was not bound by communication as long as the ratio of the number of surface atoms to the total number of atoms in each sub-domain is kept sufficiently small.

2.3.2 Multi-Level Parallelism

`NEMO3D-peta` also has a programmable interface ready for multi-level parallelism as depicted in Fig. 2.5 to achieve extra performance enhancement. In contrast to the spatial domain decomposition, where processors are coupled to each other by MPI communication. This hierarchical parallelism solves the task *independently*, with different parameters assigned for each group. Since communication is minimized, it is usually referred to as the *embarrassingly parallel* scheme, which shows perfect scalability (Fig.2.5). k-space grouping, for example, can be useful when bandstructure

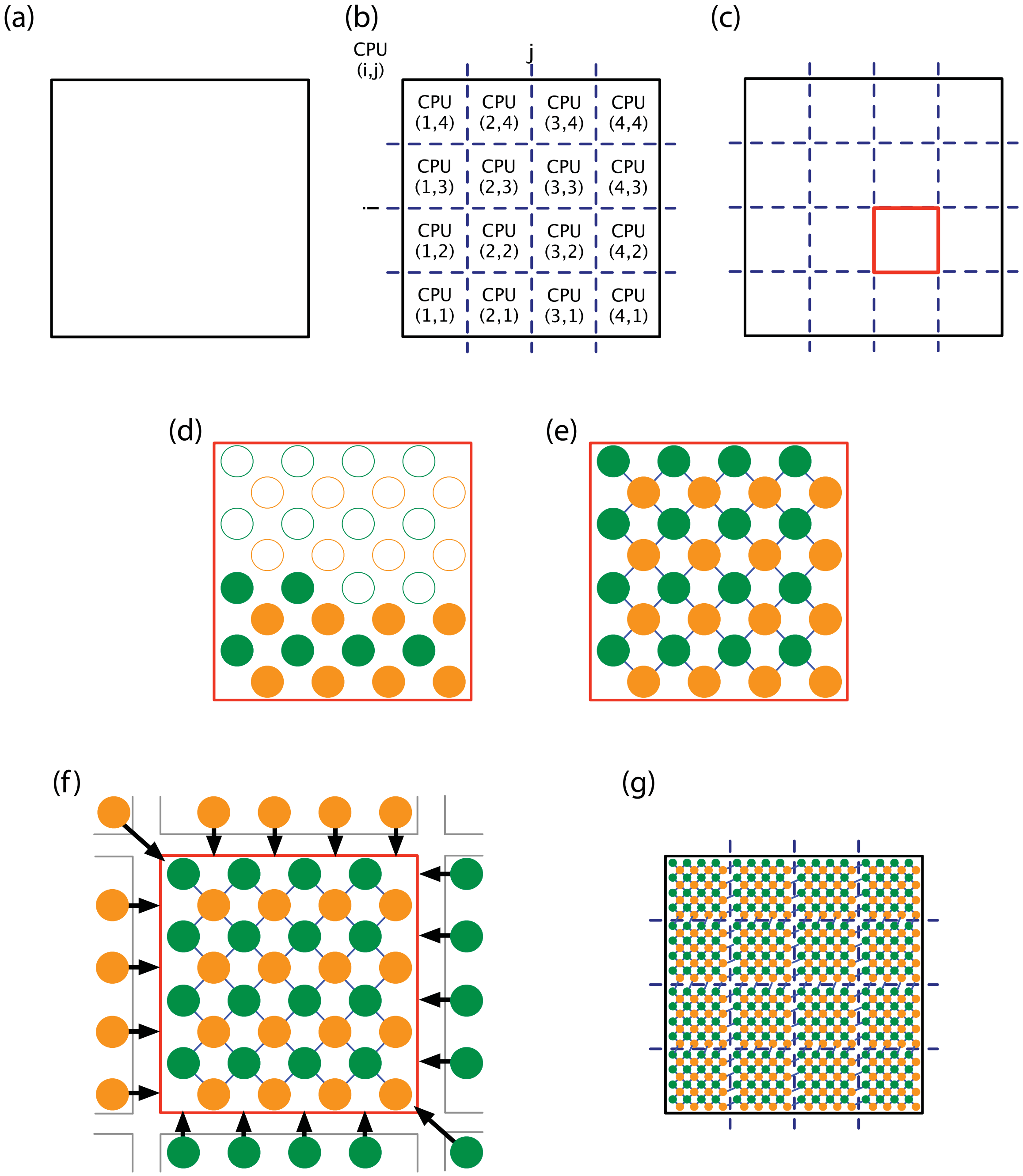


Fig. 2.4. An example of a parallel 2-D geometry construction. (a) Initial domain defined in input deck. (b) Decomposition of the domain into subdomains by number of processors and (c) taking one of the sub-domain as an example. (d) Filling up cells (atoms) that belong to certain subdomain. (e) Internal connection of bonds. (f) External connection of bonds outside its own sub-domain using MPI communication. (g) Final result of geometry construction.

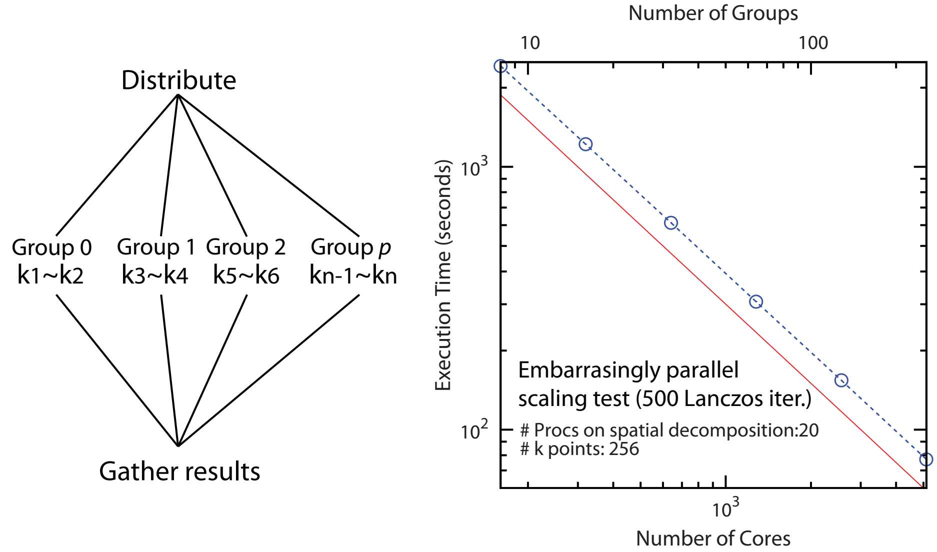


Fig. 2.5. Graphical representation of multi-level parallelism. The problem is decomposed and independently solved, involving minimal communication between groups. Perfect scalability is expected as the number of groups increases.

or charge calculations are needed. Additional groups can be added for simulations that may involve external electrical or magnetic fields. Depending on the application, `NEMO3D-peta` can provide multiple levels of additional parallelism to utilize more computational resources.

2.4 Benchmark Results

To examine the advantage of parallelism implemented in `NEMO3D-peta`, various benchmark tests are introduced in this section. Except for end-to-end tests, 500 iterations of the Lanczos eigenvalue solver algorithm are used to measure execution times with varying numbers of processors.

Fig. 2.6 shows the strong scaling[1] plot of 1-D parallelism for elongated systems of different number of atoms. The strong scaling plot indicates that with minimal

[1] *Strong scaling* means that we increase the number of processors and measure execution time or CPU utilization rate on a fixed size problem. In contrast, *weak scaling*is measured with maintaining

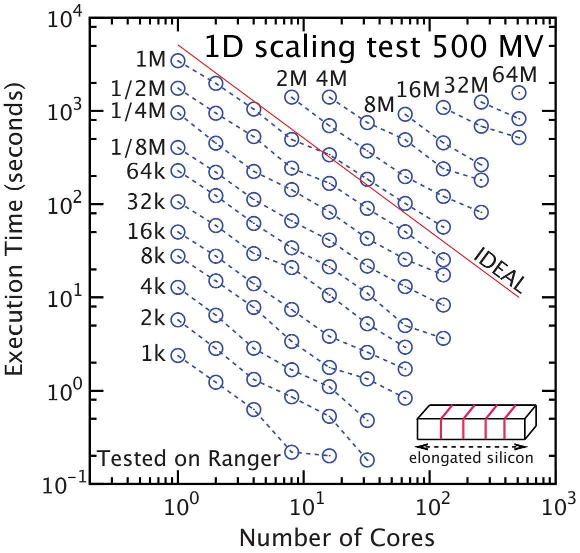


Fig. 2.6. Strong scaling benchmark results of a 1-D Decomposition scheme in `NEMO3D-peta`. 500 Lanczos iteration performance is measured on elongated silicon structures (inset). The number of atoms ranges from 1,000 to 64 million.

load of communication, `NEMO3D-peta` shows reasonable scalability up to structures containing 32 million atoms with 512 processors on Ranger. However, with smaller number of atoms per subdomain, fluctuations in performance are observed as we increase the number of processors. The instability arises when the communication load becomes comparable to the load of computational operations.

Fig. 2.7 shows the performance of `NEMO3D-peta` is benchmarked against `NEMO3D` on a realistic InAs quantum dot structure containing 6 million atoms with 20 basis functions per atom (matrix dimension: $(6{\times}10^6) \times 20 = 12 \times 10^6$) [54]. Both tool shows ideal scaling behavior to up to reasonable extent and then starts to deviate from ideal slope. `NEMO3D` scaling hits the limit when the number of processors meet the number of unit cells (88 in this simulation) along x direction since `NEMO3D` uses 1D spatial decomposition. The communication load also prevents from following ideal

a constant number of atoms per processor. Ideally, strong scaling plots show a constant $1/n$ slope on a log-log scale, while weak scaling plots yield flat lines.

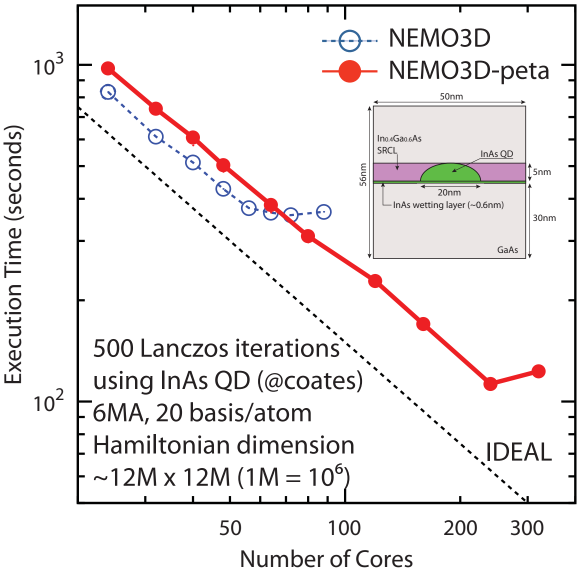


Fig. 2.7. Strong scaling result of NEMO3D-peta benchmarked against NEMO3D using InAS quantum dot [54]. NEMO3D-peta is 15% slower than NEMO3D on average. The scalability of NEMO3D is limited by 1D spatial parallelization scheme and only able to use 88 processors at maximum, equal to the number of unit cells along x direction. In contrast, NEMO3D-peta can utilize more processors using 3D spatial parallelization to compensate for the speed and extend the scalability. The scaling starts to deviate from the ideal slope and to fluctuate when the communication load becomes comparable to the computation time and the load balancing becomes critical.

scaling slope and the scaling starts to saturate around 56 processors. NEMO3D-peta also shows ideal scaling behavior up to 240 processors using 3D spatial parallelization scheme. However, NEMO3D-peta exhibits lower performance than NEMO3D as shown in Fig. 2.7. To quantify the performance degradation of NEMO3D-peta with respect to NEMO3D, the *degrading factor* is defined as following:

$$f(p) = 1 - \frac{t_{\text{NEMO3D}}(24) \times \frac{24}{p}}{t_{\text{NEMO3D-peta}}(p)} \tag{2.1}$$

where t^p_{tool} is the time taken for 500 matrix vector multiplies with p processors with “tool”. The degrading factor indicates how much the scaling is off from the ideal slope. In this benchmark study, the degrading factor gradually increases from 0.15 to 0.25 since the communication load also increases as increasing number of processors.

There are two reasons for NEMO3D-peta to show lower performance. First, the communication scheme is more complicated in NEMO3D-peta since 6-way communication with adjacent processors is needed for every matrix-vector multiplication while only 2-way communication is needed in NEMO3D. Such increased communication complexity increases overall simulation time. The communication instability also increases as more processors are utilized, as shown in Fig. 2.7 around 320 processors. [2] Second, NEMO3D-peta uses general linear algebra package (LAPACK/BLAS) for matrix-vector multiplies while NEMO3D uses manually optimized computation routines using SSE3 instruction sets. [3] Using SSE3 instruction sets is generally more effective than using standard packages. However, SSE3 is architecture dependent instruction set and may not be compatible on clusters that have different hardware architectures. NEMO3D-peta is more focused on the compatibility and transferability of the code on different platforms, rather than maximizing the performance on particular platforms. In conclusion, NEMO3D-peta shows extended scalability but may need additional optimization on block matrix-vector multiplication routines for better performance.

The benefit of 3-D domain decomposition is clearly indicated in Fig. 2.8. The structure under test is a $44 \times 44 \times 44(\mathrm{nm}^3)$ silicon cube, which has 4 million atoms (80 unit cells in each direction). In the existing NEMO3D-peta, the maximum number of processors cannot exceed 80 since the parallelism is limited by the number of unit cells in one direction. On the other hand, 2-D and 3-D parallelization enables us to assign more processors for the calculation. In principle, 2-D and 3-D parallelization utilizes 80×80=6,400 and $80\times80\times80$=512,000 processors, respectively. Additional degrees of freedom in parallelism dramatically improve simulation time as shown in Fig. 2.8, where the performance is improved by 50 times. In other words, we can obtain the

[2] For this reason, initial benchmark is recommended to find the optimal number of processors before running any simulations.

[3] SSE stands for Streaming SIMD (Single Instruction Multiple Data) Extension, which allows to compute multiple arithmetic operations within single instruction cycle. It has upgrade to 4th version (SSE4) and supports all intel compatible processors.

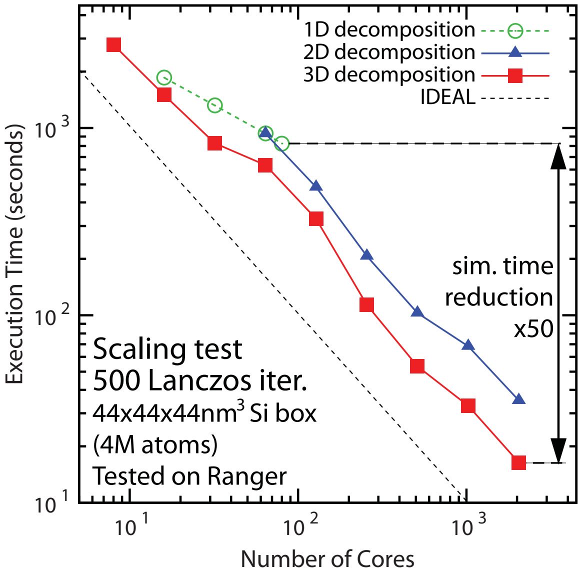


Fig. 2.8. Strong scaling comparison between 1/2/3-D spatial decomposition. Performance of 500 Lanczos iterations is measured on a $44 \times 44 \times 44(\text{nm}^3)$ silicon cube (4 million atoms). For the 2-D case, the processors are assigned as $(\text{c}_\text{x}, \text{c}_\text{y}, \text{c}_\text{z})$=(16,$2^i$,1), $i = 0, \cdots, 4$. And for 3-D case, $(\text{c}_\text{x}, \text{c}_\text{y}, \text{c}_\text{z})$=$(2^i, 2^j, 2^k)$, $i, j, k = 1, 2, 3$.

result of a simulation run in a couple of minutes that will normally take an hour to finish by using this 3-D decomposition scheme. In addition, it is proven that by using the 3-D domain decomposition method, we can utilize as many as 32,768 processors on systems comprising 1 billion atoms without losing any scalability as shown in Fig. 2.9. [4] In Fig. 2.10, an end-to-end scalability was examined using strain calculation in GaAs-InAs quantum dot (QD). The sample structure is a cylindrical InAs QD of size 20nm(D) × 5nm(H) embedded in $68 \times 68 \times 68(\text{nm}^3)$ GaAs buffer, which is comprised of 13 million atoms. This calculation requires minimization of the total strain energy using an atomistic valence force field (VFF) [28] and a conjugate gradient (CG) minimization algorithm. CG does not need intensive computation compared to the matrix-vector multiplier used in the eigenvalue problem. However, it requires frequent

[4] This scaling example is not for real use since in practice we do not need an atomistic tight-binding approach for such large domains. It is a demonstration to show that the 3-D domain decomposition algorithm is not limited by either the method itself or the memory limit.

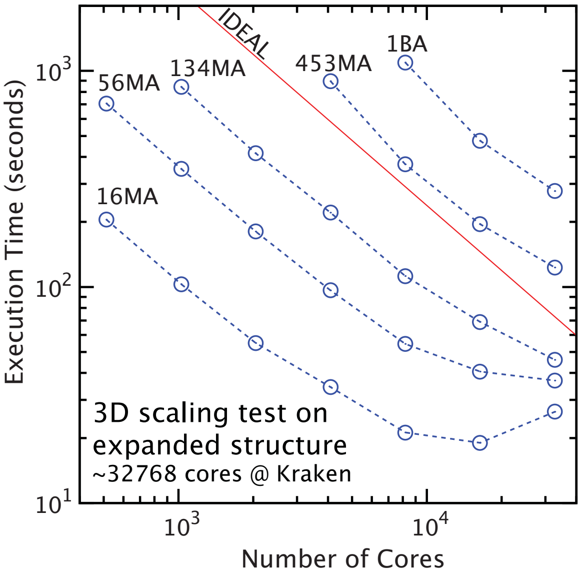


Fig. 2.9. Strong scaling benchmark result on huge structures using `NEMO3D-peta`. 500 Lanczos iteration performance is measured on the large cubic box that extends up to $100 \times 100 \times 100(\text{nm}^3)$. The number of atoms in the structure is as large as 1 billion and the maximum number of cores used in this test is 32,768.

vector dot product operations, which need larger numbers of MPI calls. Therefore, the CG method generally shows poor scalability. It is also shown in Fig. 2.10 that the scale plot saturates faster than in the Lanczos benchmark tests. Although communication overhead is larger, we can benefit from the 3-D decomposition method to achieve extended scalability up to a factor of 3.5 by allocating 4 times more processors.

The effect of the embarrassingly parallel method is shown in Fig. 2.11. The structure is a 1D periodic slab of 16×16 silicon nanowire, which is comprised of about 10,000 atoms. A k-space integration will be needed for charge calculation; k points can be distributed evenly among processors and each processor solves an eigenvalue problem independently. In this example, perfect scaling can be achieved using k-space distribution to up to 64 processors. Beyond this limit, we can also adopt 3-D decomposition scheme simultaneously to achieve additional scalability to up to 256 processors. In summary, two different parallelization schemes can contribute simul-

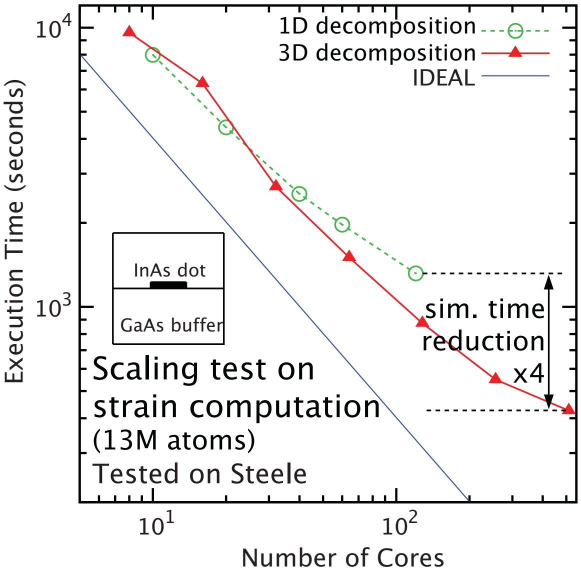


Fig. 2.10. Comparison of strain performance between 1-D and 3-D decomposition. A cylindrical InAs QD of size 20nm(D)×5nm(H) is encapsulated in 68×68×68 (nm^3) GaAs buffer. This structure has 13 million atoms.

taneously to improving the performance of `NEMO3D-peta`. As mentioned above and seen from the scaling plots, the major cause of performance degradation in the 3-D decomposition method is communication overhead. In Fig. 2.12, a simple example is used to quantify the relationship between communication overhead and the number of processors. An end-to-end self-consistent simulation was performed on a slab of 8.7×8.7(nm^2) silicon wire with single phosphorus impurity placed in the middle of the channel. For the scaling result, we measured the end-to-end simulation time. To quantify the communication load, we computed the ratio of the number of surface atoms in each sub-domain (outermost atoms in each sub-domain) to the total number of atoms. Since the communication overhead is mostly caused by domain decomposition, the embarrassingly parallel scheme is ignored. The scaling curve starts to deviate from the ideal slope bound by communication load. Increasing the number of processors will reduce the size of each sub-domain, leading the relative ratio of the number of surface atoms to total number of atoms to increase and it can be di-

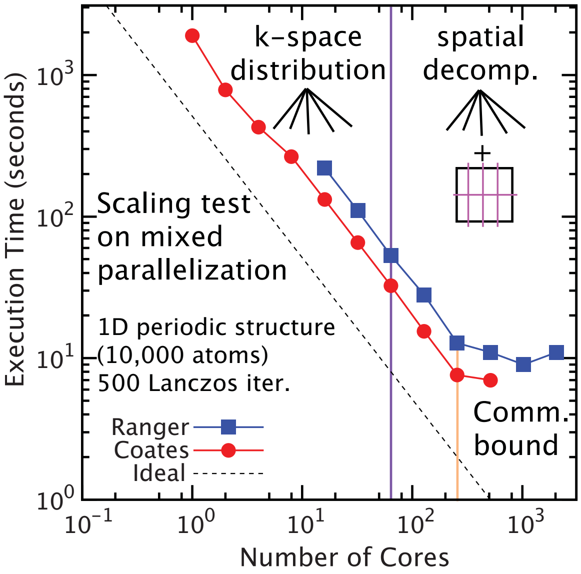


Fig. 2.11. Scaling test on multiple level parallelism on different platforms. 500 Lanczos iteration performance is measured on the bandstructure calculation. First part of the simulation only considers distributing processors in k space and further scaling is achieved by adding spatial decomposition algorithm until communication overhead starts to dominate.

rectly mapped to the communication load. From Fig. 2.12, it turns out that the scale plot starts to saturate when the ratio is greater than 20%. Further scaling can be achieved, however, CPU utilization efficiency will clearly drop significantly and may result in a waste of computational resources. Even though the exact ratio may vary depending on the device geometry, it is recommended to keep the ratio under 20% in any circumstances based on this example.

2.5 Application: the Schrödinger-Poisson solver

One of the first applications of `NEMO3D-peta` is the self-consistent charge and potential computation module, known as the Schrödinger-Poisson solver, which was *not* present in `NEMO3D-peta`. It fully utilizes the parallel algorithms introduced in previous sections. Detailed steps for the self-consistent loop are discussed below in

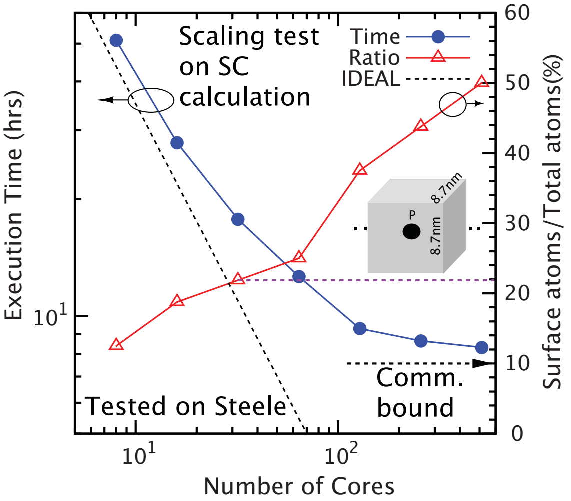


Fig. 2.12. Scaling test on an end-to-end self-consistent calculation on a Si:P wire depicted in the inset. As a measure of communication load, the ratio of atoms adjacent to neighboring sub-domains to the total number of atoms is considered. If the ratio exceeds 20%, the scaling curve starts to be bounded by the communication overload.

terms of parallelization. The physical interpretation of the Schrödinger-Poisson self-consistent loop is discussed in the next chapter.

1. *Charge integration*: For the quantum charge calculation, we first need to decompose the problem in k-space using multi-level parallelism as discussed in Section 2.2.2. For every k-point, the eigenstates are evaluated by solving $(H(\mathbf{k})+U)\Psi_n(\mathbf{k}) = E_n\Psi_n(\mathbf{k})$ and utilizing spatial domain decomposition. Finally, the density of states (DOS) and local density of states (LDOS) results from all k groups are aggregated using MPI_Allreduce. The charge profile is gathered from all processor groups (k-groups) and identically distributed back, resulting in same charge profiles in all groups. A graphical representation of the charge integration is shown in Fig. 2.13(a).

2. *Atomistic grid - Continuum grid mapping (Fig. 2.13(b))*: Due to discrepancies between the atomistic grid (Schrödinger equation) and continuum grid (Poisson

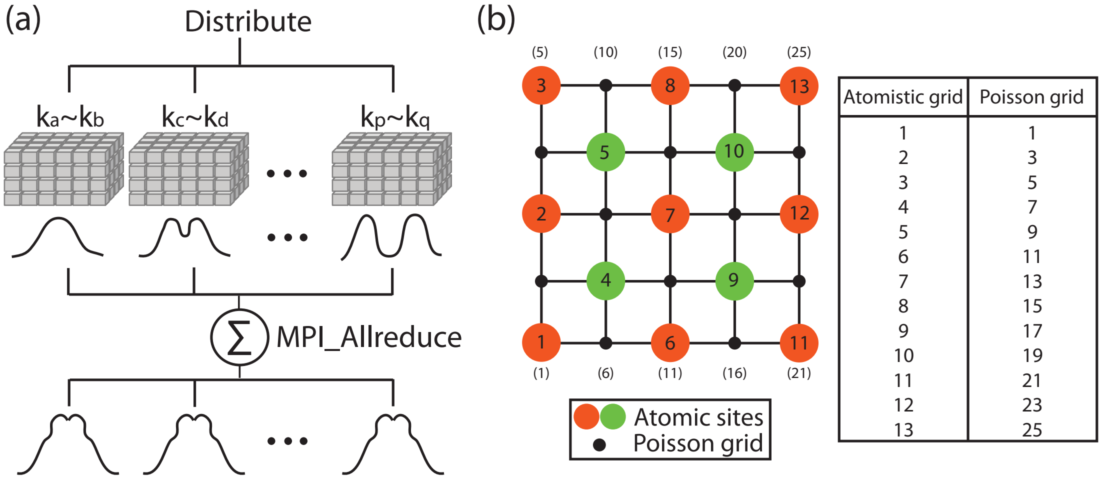
equation), a mapper between the two different grids is required. The domain decomposition scheme is also applied in such a way that each electronic sub-domain is built with respect to the Poisson sub-domain, which has finer grid [5].

3. *Poisson Solver*: Once the charge is mapped and correctly loaded into the Poisson grid, we either solve a linear equation ($\nabla \cdot (\epsilon \nabla U) = \rho$) or apply Newton-Raphson's method (or non-linear Poisson solver) [57] to obtain the electrostatic potential. The Poisson solver is implemented using an FDM grid and the Aztec parallel iterative linear solver package in conjunction with our spatial domain decomposition scheme, which also exhibits good scalability [64]. The converged potential is applied back into the Schrödinger equation ($U = \alpha U_{\text{new}} + (1-\alpha) U_{\text{old}}$) to complete the self-consistent loop. This loop is repeated until the self-consistent potential has converged.

2.6 Conclusion

The new nanoelectronic simulator `NEMO3D-peta` has been developed to overcome the limitations of `NEMO3D-peta` in a processor-rich environment. The 3-D parallel geometry constructor introduced in `NEMO3D-peta` is carefully benchmarked in various aspects. As a result, our parallelization scheme not only exhibits better scalability than the old `NEMO3D-peta` up to 32,000 processors but also effectively handles communication overhead although the complexity of MPI communication is increased. In particular, this engine is highly suitable for quantum dot simulations that contain multi-million atoms. In addition, a self-consistent loop is embedded into the new engine and applied to compute self-consistent bandstructures of Si:P systems. This work will allow us to perform `NEMO3D-peta` like calculations in minutes rather than days.

[5] In general, the electronic domain is a subset of Poisson domain; regardless of the grid used (FDM or FEM), the Poisson grid needs to fill the gap between atomic sites to ensure the solution (potential) to satisfy smoothness.



Atomistic grid	Poisson grid
1	1
2	3
3	5
4	7
5	9
6	11
7	13
8	15
9	17
10	19
11	21
12	23
13	25

Fig. 2.13. (a) Computational flow of the charge integration. First, distribute the problem into groups of different k and by using spatial domain decomposition compute the charge with respect to the Fermi level. Finally, gather the sum of local density of states (LDOS) using MPI_Allreduce. As a result, we obtain an identical charge profile throughout the distributed groups. (b) An atomistic grid - Poisson grid mapper example.

3. NUMERICAL RECIPE: LANCZOS EIGENSOLVER FOR LARGE-SCALE SIMULATION

3.1 Introduction

The heart of identifying an atomistic description of the electronic structure is to solve the time-independent Schrödinger equation, which from a numerical standpoint leads to the Hermitian eigenvalue problem,

$$H\Psi_k = E_k\Psi_k \tag{3.1}$$

where $H \in \mathbb{C}^{n\times n}$ is large, sparse and Hermitian, $\Psi_k \in \mathbb{C}^{n\times 1}$ and the scalar $E_k \in \mathbb{R}$ with $k = 1 \ldots n$.

The eigenvalue spectrum of the Hamiltonian matrix H in Eqn. (3.1) describes the physical system and has a gap in the interior of the spectrum in the range of $[-1\text{eV}, 2\text{eV}]$. Usually, a small set of eigenpairs is sought, immediately above and below the gap. The eigenvalues correspond to energy levels in the conduction and valence bands, whereas the eigenvectors correspond to electron and hole wavefunctions. These wavefunctions are often spatially confined to a small region of the overall device. Spin or valley degeneracies may introduce multiplicities (degeneracies) in the energy levels. Magnetic fields, lack of crystal symmetry, atomic disorder or piezoelectric effects may split the degeneracies [48]. Different physical conditions and simulation goals determine the required accuracy of the eigenvalue calculations. It may sometimes be sufficient just to know the energy levels, irrespective of their multiplicities and without the states. In other cases, however, knowledge of the degeneracies and wavefunctions may be required.

Additionally, the dimension of the Hamiltonian matrix can exceed 10^7 in quantum dots, requiring a massively parallelizable eigensolver. There are variety of fast and efficient eigensolver algorithms, such as, the implicitly restarted Arnoldi method with shift-and-invert algorithm and the Jacobi-Davidson method [65]. None of the algorithms, however, are suitable for solving eigenvalues of huge systems due to the parallelization difficulty and the memory limitation. The Lanczos algorithm, on the other hand, is one of the simplest eigenvalue solvers based on the Arnoldi method to solve extremal eigenstates with a least amount of storage. The method generates a sequence of tridiagonal matrices $T_k \in \mathbb{R}^{k\times k}$, which have the property that their eigenvalues approach the eigenvalues of the original matrix. The Lanczos algorithm tends to converge faster to extremum eigenvalues but is also shown in practice to be successful in computing interior eigenvalues within reasonable number of iterations when solving the Schrödinger equation with matrix sizes exceeding billion degrees of freedom. Computationally, it can be easily implemented in parallel using a parallelized matrix-vector multiplier. The Lanczos algorithm and its implementation is described in sections 3.2 and 3.3, respectively.

Numerically, however, the bare Lanczos algorithm is very slow compared to the implicitly restarted Arnoldi method with shift-and-invert algorithm by a factor of 10 when solving eigenstates of small matrices, and it inherently shows numerical instability since the basis vectors gradually lose orthogonality. Moreover, there are a couple of practical issues in the Lanczos algorithm when applied to electronic structure calculations. The Lanczos algorithm is unable to resolve degeneracies and its computed eigenvalues are not guaranteed to be sequential, which can be critical when computing optical properties of quantum dots. To circumvent these problems, two simple algorithms are introduced in sections 3.4 $\sim$ 3.5 to guarantee sequential eigenvalues with multiplicity. The effectiveness of the implemented algorithms is benchmarked in section 3.6. Section 3.7 summarizes and concludes the chapter. All items discussed in this chapter is implemented in `lanczos.hpp` in `NEMO3D-peta`.

3.2 Basic Algorithm

[1]Let $A \in \mathbb{R}^{n\times n}$ be sparse and Hermitian and $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \cdots \geqslant \lambda_n$ be its eigenvalues. From Schur's theorem , the eigenvalues of A are *real* and there exists a unitary matrix Q such that $Q^H AQ = \Lambda$ where $\Lambda = \mathrm{diag}(\lambda_1 \cdots \lambda_n)$. In addition, min−max theorem shows for Hermitian matrix the maximum and minimum values of Rayleigh quotient $r(x) = x^H Ax/x^H x, x \neq 0$ are λ_1 and λ_2, respectively. Let's define $Q_k = [q_1 q_2 \cdots q_k], q_i \subseteq \mathbb{R}^{n\times 1}$ and two scalars m_k and M_k given by

$$\begin{aligned} M_k = \lambda_1(Q^H AQ) = \max_{y\neq 0} \frac{y^H(Q_k^H AQ_k)y}{y^H y} = \max_{\|y\|_2=1} r(Q_k y) \leq \lambda_1(A) \\ m_k = \lambda_k(Q^H AQ) = \min_{y\neq 0} \frac{y^H(Q_k^H AQ_k)y}{y^H y} = \min_{\|y\|_2=1} r(Q_k y) \geq \lambda_n(A) \end{aligned} \tag{3.2}$$

From Eqn. 3.2, the Lanczos algorithm can be restated as finding an orthonormal vector set Q_k to acquire better estimates of $\lambda_1(A)$ and $\lambda_n(A)$.

Let $u_k(v_k) \in \mathrm{span}\{q_1, q_2, \cdots, q_{k+1}\}$ such that $M_k = r(u_k)$ $(m_k = r(v_k))$. Since $r(x)$ will increase (decrease) along the direction of the gradient [2] $\nabla r(x)$ $(-\nabla r(x))$, if $\exists\ q_{k+1}$ such that $\nabla r(u_k),\ \nabla r(v_k) \in \mathrm{span}\{q_1, \cdots, q_{k+1}\}$ then it is possible to satisfy both $M_k \leq M_{k+1} (\leq \lambda_1(A))$ and $m_k \geq m_{k+1} (\geq \lambda_n(A))$. Since $\nabla r(x) \in \mathrm{span}\{x, Ax\}$ from the gradient expression in footnote 2, it is clear that if we choose q_{k+1} such that

$$\nabla r(u_k),\ \nabla r(v_k) \in \mathrm{span}\{q_1, \cdots, q_{k+1}\} = \mathrm{span}\{q_1, Aq_1, \cdots, A^k q_1\}$$

[1] This section is adopted and modified from Ref. [55].

[2]

$$\begin{aligned} \nabla r(x) &= \left[\frac{\partial r(x)}{\partial x_1}, \cdots, \frac{\partial r(x)}{\partial x_n}\right] \\ \frac{\partial r(x)}{\partial x_j} &= \frac{\partial}{\partial x_j}\left(\frac{x^H Ax}{x^H x}\right) = \frac{(\frac{\partial}{\partial x_j} x^H Ax)x^H x - x^H Ax(\frac{\partial}{\partial x_j} x^H x)}{(x^H x)^2} \\ \frac{\partial}{\partial x_j} x^H Ax &= 2\mathrm{Re}(Ax)_j \\ \frac{\partial}{\partial x_j} x^H x &= 2\mathrm{Re}(x)_j \\ \frac{\partial r(x)}{\partial x_j} &= \frac{2}{x^H x}(\mathrm{Re}\{(Ax)_j\} - r(x)\mathrm{Re}\{(x_j)\}) \\ \nabla r(x) &= \frac{2}{x^H x}(\mathrm{Re}(Ax) - r(x)\mathrm{Re}(x)) \end{aligned}$$

then it is possible to make a significantly reduced (but growing) matrix $Q_k^H AQ_k$ with eigenvalues approaching the actual eigenvalues of A. Now the problem translates to computing orthonormal bases for the *Krylov subspace.*

$$\mathcal{K}(A, q_1, k) = \text{span}\{q_1, Aq_1, \cdots, A^{n-1}q_1\} \tag{3.3}$$

To find this basis efficiently, it is necessary to find connection between the tridiagonalization of A and the QR factorization of $K(A, q_1, n)$. If $Q^H AQ = T$ is tridiagonal with $Qe_1 = q_1$, then $K(A, q_1, n) = Q[e_1, Te_1, T^2 e_1, \cdots, T^{n-1}e_1]$ becomes the QR factorization of $K(A, q_1, n)$. Therefore, the q_k can be generated by tridiagonalizing A with an orthogonal matrix whose first column is q_1. Setting $Q = [q_1, q_2, \cdots, q_n]$ and

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \cdots & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \cdots & \vdots \\ 0 & \cdots & \cdots & \beta_{n-1} & \alpha_n \end{bmatrix}$$

and convert columns in $AQ = QT$, following equation can be easily derived for $k = 1 \sim n - 1$.

$$Aq_k = \beta_{k-1}q_{k-1} + \alpha_k q_k + \beta_k q_{k+1} \quad \beta_0 = 1, q_0 = \vec{0} \tag{3.4}$$

Since q_k are orthonormal to one another, if we multiply q_k, or *Lanczos vector at iteration* k to both sides of Eqn. (3.4)

$$q_k^H Aq_k = \alpha_k \tag{3.5}$$

After k-th iteration, the remaining vector can be defined as $r_k = Aq_k - \alpha_k q_k - \beta_{k-1}q_{k-1}$ and the next Lanczos vector q_{k+1} can be obtained as $q_{k+1} = r_k/\beta_k,\ \beta_k = \|r_k\|_2$. As long as sufficient information for invariant subspace information is acquired, r_k will never be zero. As a result, the lanczos iteration for obtaining the matrix elements of T_k is shown in Eqn. (3.6).

```
$r_0 = q_1$ (a random vector with $\|q_1\|_2 = 1$)
$\beta_0 = 1, \; q_0 = 0, \; k = 0$
**while** $\beta_k \neq 0$
    $q_{k+1} = r_k / \beta_k$
    $k = k + 1$                                        (3.6)
    $\alpha_k = q_k^H A q_k$
    $r_k = (A - \alpha_k I) q_k - \beta_{k-1} q_{k-1}$
    $\beta_k = \|r_k\|_2$
**end**
```

At any step k, the eigensolution of T_k can be calculated $T_k u_j^{(k)} = u_j^{(k)} \theta_j^{(k)}$, where $\theta_j^{(k)}$ is the Ritz value and the Ritz vector can be calculated as $x_j^{(k)} = Q_k u_j^{(k)}$. The Ritz pair $(\theta_j^{(k)}, x_j^{(k)})$ is a good approximation to the eigenstate of the larger matrix A if the norm of the residual meets convergence criteria, typically $\|z_j^{(k)}\|_2 \leqslant 10^{-8} \sim 10^{-6}$.

The residual - the difference between true value and Ritz value - for the Ritz pair $(\theta_j^{(k)}, x_j^{(k)})$ can be calculated as

$$z_j^{(k)} = A x_j^{(k)} - x_j^{(k)} \theta_j^{(k)} = A Q_k u_j^{(k)} - Q_k u_j^{(k)} \theta_j^{(k)} = (A Q_k - Q_k T_k) u_j^{(k)} = \beta_k q_{k+1} u_{j,k}^{(k)}$$

The norm of residual is $\|z_j^{(k)}\|_2 = |\beta_k u_{j,k}^{(k)}|$ and $u_{j,k}^{(k)}$ is the last element (k-th element) of the eigenvector $u_j^{(k)}$.

3.3 Computational Aspects

Unlike Arnoldi iterations with restarting the T matrix grows with number of iterations. Typically, when dealing with million-atom quantum dots, it takes 15,000 $\sim$ 25,000 iterations to obtain converged eigenvalues and it is computationally impossible to store all Lanczos vectors in the memory. To save memory it is required to run the same Lanczos iterations *twice* for computing eigenvectors. In the eigenvalue calculation stage, the procedures described in the previous section are carried out. Assume all t desired eigenvalues are found at iteration k. Since the same Lanczos iterations are needed for the eigenvector calculation, the initial random vector (q_1), the tridiagonal matrix (T_k), the converged eigenstates $(\theta_i^{(m_i)}, u_i^{(m_i)}),\ i = 1 \cdots t,\ m_i \leqslant k$ and (i, m_i) should be stored in advance. m_i is the iteration number of the i-th eigenvalue since each eigenvalue converges at a different iteration count. Since the desired eigenvector can be expressed as $x_i^{m_i} = Q_{m_i} u_i^{(m_i)} = [q_1\ q_2\ \cdots\ q_{m_i}] u_i^{(m_i)}$ the eigenvector can be reconstructed using the following formula

$$x_i^{m_i} = \sum_{n=1}^{m_i} q_n u_{i,n}^{(m_i)} \tag{3.7}$$

where $u_{i,n}^{(m_i)}$ is n-th element of i-th converged eigenvector $u_i^{(m_i)}$ which is converged at m_i-th iteration. Each q_n is obtained from Lanczos iteration as Eqn. (3.6). The overall computation procedure is described in Fig. 3.1~3.2.

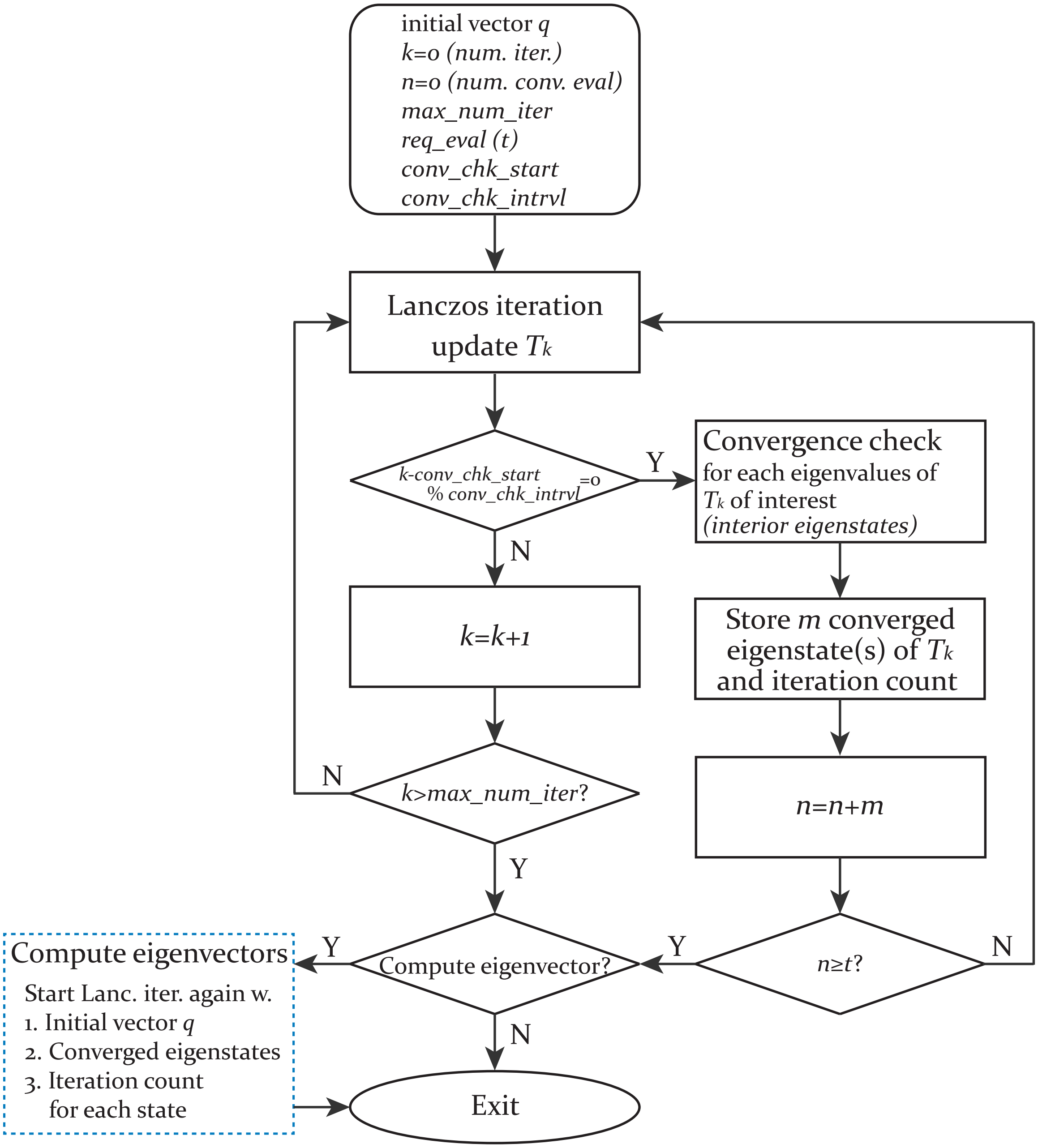
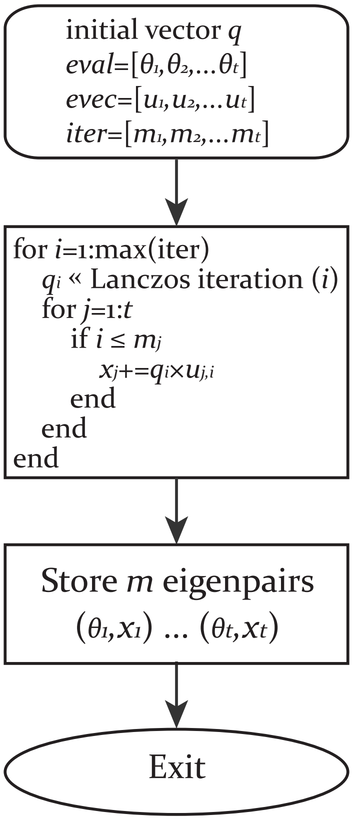

initial vector q
k=0 (num. iter.)
n=0 (num. conv. eval)
max_num_iter
req_eval (t)
conv_chk_start
conv_chk_intrvl
Lanczos iteration update Tk
k-conv_chk_start % conv_chk_intrvl =0
Y
N
Convergence check for each eigenvalues of Tk of interest (interior eigenstates)
k=k+1
Store m converged eigenstate(s) of Tk and iteration count
N
k>max_num_iter?
n=n+m
Y
Compute eigenvectors
Start Lanc. iter. again w.
1. Initial vector q
2. Converged eigenstates
3. Iteration count for each state
Y
Compute eigenvector?
Y
n≥t?
N
N
Exit

Fig. 3.1. Flow chart of finding eigenvalues using Lanczos algorithm. A Lanczos iteration is computed according to Eqn. (3.6) and the norm of the residual determines the convergence of eigenstates (θ, u) of T_k at every convergence check stage. An eigenvector calculation flow marked with a dashed box is shown in Fig. 3.2.


initial vector q
$eval$=[θ_1,θ_2,...θ_t]
$evec$=[u_1,u_2,...u_t]
$iter$=[m_1,m_2,...m_t]
Store m eigenpairs
(θ_1,x_1) ... (θ_t,x_t)
Exit

```
for $i$=1:max(iter)
   $q_i$ « Lanczos iteration ($i$)
   for $j$=1:$t$
      if $i \le m_j$
         $x_j$+=$q_i \times u_{j,i}$
      end
   end
end
```

Fig. 3.2. Flow of finding eigenvectors based on the eigenvalue calculation described in Fig. 3.1. Initial vector q, converged eigenstates information of T_k, (θ_i, u_i, m_i), $i = 1 \cdots t$ are used.

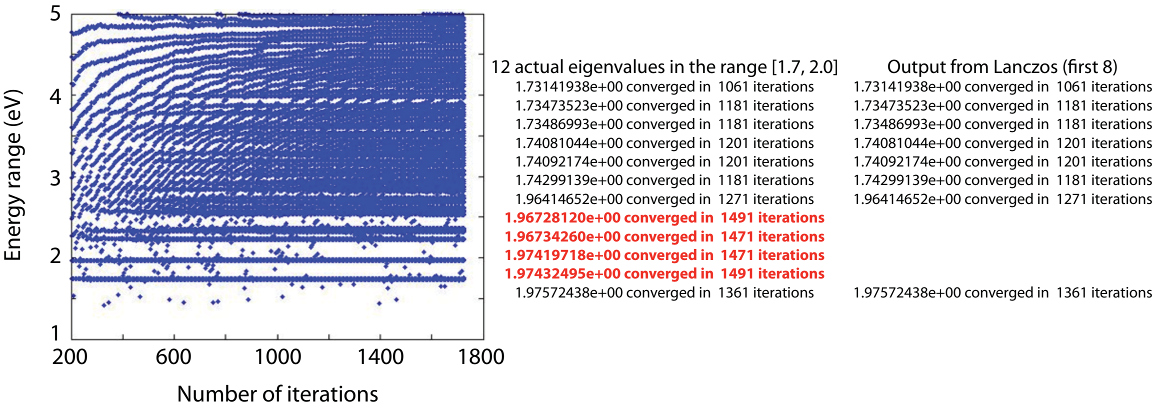
3.4 Finding eigenvalues in sequential order

The optical and electronic properties of semiconductors are primarily determined by the band-gap (the difference between the conduction band and valence band edge) and states right above and below the band edges. Therefore, the eigenvalue solvers should be capable of resolving all interior eigenstates in the range of interest. Furthermore, the eigensolver must prioritize the sequence of eigenvalues; e.g. when finding the eigenvalues near the conduction (valence) band edge, smaller (larger) eigenvalues should have higher priority. Unfortunately, however, the Lanczos algorithm does not guarantee a sequential convergence when finding the interior eigenvalues (Fig. 3.3). Lanczos converges faster towards the extremum eigenstates but the convergence for the interior eigenvalues is random in nature. To resolve this issue, a simple algorithm to detect whether there are any missing eigenvalues in the range of interest is devised. The key feature of this algorithm, or *binning method* is summarized as follows.

1. Keep track of all eigenstates in the range of interest at every convergence check period and determine whether those eigenvalues are destined to converge and predict possible number of converged eigenvalues in the range of interest.

2. When the requested number of eigenvalues has been obtained or when the maximum iteration has been reached, it is determined whether to continue iterating to find missing eigenstates.

Examining the convergence pattern of eigenstates like in Fig. 3.3 indicates that eigenstates around the bottom of the conduction band generally converge faster than compared to other eigenstates that are less significant, which makes it easier to keep track of those values. The energy range of interest ($[E_{min}\ E_{max}]$) is split into n small *bins*, $B = \{b_1,\ b_2, \cdots,\ b_n\}$ where b_i is the number of eigenvalues of matrix T that belongs in the bin with interval $[E_{i-1}\ E_i)$ $(i = 1, \cdots, n,\ E_0 = E_{min},\ E_n = E_{max})$. At every convergence check period k, compute all the intermediate eigenvalues in the energy range and create $B_k = \{b_1^k,\ b_2^k, \cdots,\ b_n^k\}$. After a sufficient number of



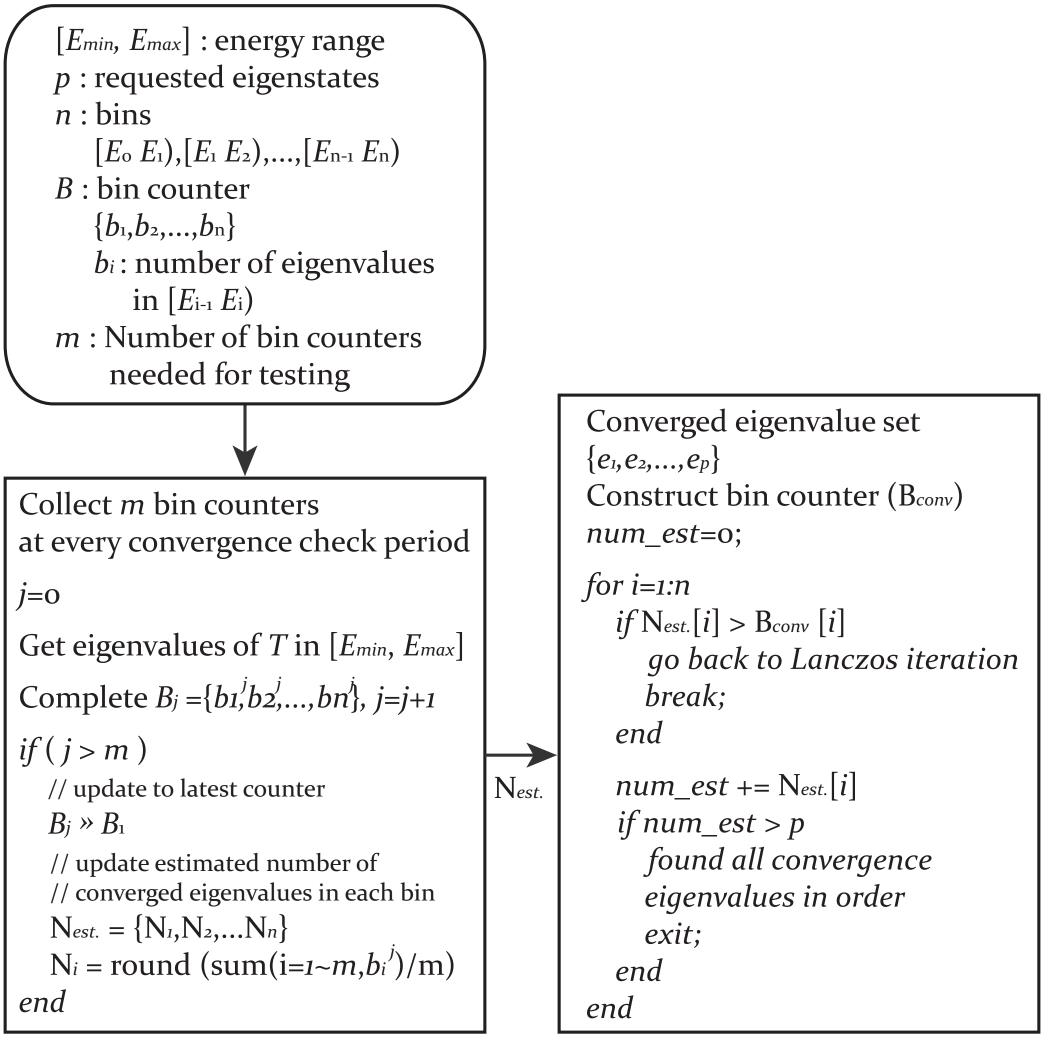
```
12 actual eigenvalues in the range [1.7, 2.0]           Output from Lanczos (first 8)
   1.73141938e+00 converged in  1061 iterations     1.73141938e+00 converged in  1061 iterations
   1.73473523e+00 converged in  1181 iterations     1.73473523e+00 converged in  1181 iterations
   1.73486993e+00 converged in  1181 iterations     1.73486993e+00 converged in  1181 iterations
   1.74081044e+00 converged in  1201 iterations     1.74081044e+00 converged in  1201 iterations
   1.74092174e+00 converged in  1201 iterations     1.74092174e+00 converged in  1201 iterations
   1.74299139e+00 converged in  1181 iterations     1.74299139e+00 converged in  1181 iterations
   1.96414652e+00 converged in  1271 iterations     1.96414652e+00 converged in  1271 iterations
  1.96728120e+00 converged in  1491 iterations
  1.96734260e+00 converged in  1471 iterations
  1.97419718e+00 converged in  1471 iterations
  1.97432495e+00 converged in  1491 iterations
   1.97572438e+00 converged in  1361 iterations     1.97572438e+00 converged in  1361 iterations
```

Fig. 3.3. (Left) Convergence pattern of eigenvalue spectrum tested on $2.2 \times 2.2 \times 2.2$ (nm^3) silicon box and first 12 values from conduction band minimum. Due to the random convergence nature of Lanczos algorithm, intermediate states (highlighted in red) are sometimes missing.

convergence check procedures (m), the number of eigenvalues in each bin interval can be predicted using a simple formula.

$$\mathcal{N}_i \equiv \text{Number of eigenvalues in } [E_{i-1}\ E_i) = \text{round}\left(\frac{\sum_{j=1}^{m} b_i^j}{m}\right),\ i = 1, \cdots, n$$

As mentioned above, $\mathcal{N}_0$ has the highest priority for conduction band eigenvalues and vice versa for the valence band case. After completing the requested number of eigenvalues, an additional check process whether to proceed with more iterations for missing eigenvalues should be performed by comparing the expected number of eigenvalues in each bin ($\mathcal{N}_i$) with the actual converged count ($\mathcal{N}_i^*$) starting from $\mathcal{N}_0$. If $\mathcal{N}_i > \mathcal{N}_i^*$ additional Lanczos iterations are performed, otherwise compare the next bin. If the first few bin counts match with predicted values, complete the Lanczos iteration. The overall procedure is shown in Fig. 3.4. The binning algorithm provides a simple and efficient way of finding eigenvalues near the band edge in a sequential manner and is particularly important in finding correct optical gaps in quantum dots. However, it tends to over-extend Lanczos iterations by a factor of 1.5 on average and therefore is hard to apply to self-consistent simulations, which require multiple iterations to obtain a converged solution. Additional optimization



```
[$E_{min}$, $E_{max}$] : energy range
$p$ : requested eigenstates
$n$ : bins
    [$E_0$ $E_1$),[$E_1$ $E_2$),...,[$E_{n-1}$ $E_n$)
$B$ : bin counter
    {$b_1$,$b_2$,...,$b_n$}
    $b_i$ : number of eigenvalues
        in [$E_{i-1}$ $E_i$)
$m$ : Number of bin counters
      needed for testing
```

```
Collect $m$ bin counters
at every convergence check period
$j$=0
Get eigenvalues of $T$ in [$E_{min}$, $E_{max}$]
Complete $B_j$ ={$b1^j$,$b2^j$,...,$bn^j$}, $j$=$j$+1
if ( $j > m$ )
    // update to latest counter
    $B_j$ » $B_1$
    // update estimated number of
    // converged eigenvalues in each bin
    $N_{est.}$ = {$N_1$,$N_2$,...$N_n$}
    $N_i$ = round (sum(i=1~$m$,$b_i^j$)/m)
end
```

```
Converged eigenvalue set
{$e_1$,$e_2$,...,$e_p$}
Construct bin counter ($B_{conv}$)
num_est=0;
for i=1:n
    if $N_{est.}[i] > B_{conv}[i]$
        go back to Lanczos iteration
        break;
    end
    num_est += $N_{est.}[i]$
    if num_est > p
        found all convergence
        eigenvalues in order
        exit;
    end
end
```

Fig. 3.4. Pseudocode description of the binning algorithm. The energy range of interest is divided into n small bins and gathers the information of the bin counter (B) at every convergence check period for m times. The number of expected eigenvalues ($N_{est.}$) is defined as the average value of the bin counter. $N_{est.}$ is compared with the actual converged eigenvalue count to determine whether the converged eigenvalues are in sequential order. If the bin count is less than the expected count in each bin, more Lanczos iterations are run to find more eigenvalues.

can be done, for instance, by additionally storing m sets of eigenvalue information and perform one-to-one comparisons. Additional optimization schemes are not discussed further in this paper.

3.5 Finding degeneracy

Many quantum mechanics problems require resolving degenerate eigenstates since each state is by default spin degenerate in the absence of external fields. In the case of harmonic potentials modeled for many quantum dots multiplicity increases as a function of energy level. In physics, the easiest way to obtain degenerate eigenstates with a priori knowledge of a given Hamiltonian having (spin) degenerate eigenvalues is to apply time-reversal symmetry (Kramer's degeneracy) to create the degenerate eigenstate pairs [66,67]. In general cases, however, it is difficult to expect the number of degenerate eigenvalues other than spin multiplicity. Therefore, it is critical to explore all-purpose degenerate eigenvalue solvers.

As previously mentioned the Lanczos algorithm fails to resolve degenerate eigenvalues since the upper Hessenberg matrix T_k is *unreduced* [55]. Fortunately, similar Lanczos procedure can be adopted to resolve this difficulty. The block Lanczos algorithm which is a block (matrix) version of the Lanczos iteration [68] provides additional bandwidth for T_k making a *block* tridiagonal matrix $\bar{T}_k$ with $p \times p$ block components ($\alpha_k \to A_k,\ \beta_k \to B_k$). The block dimension p should be set to at least the largest expected degeneracy. At the same time the computational effort increases by p^2 to construct and solve for $\bar{T}_k$. In practice, however, the block Lanczos algorithm tends to lose orthogonality between Lanczos vectors faster than Lanczos, which makes difficult to apply to matrices requiring large numbers of Lanczos iterations ($\sim 10^4$) for interior eigenvalue problems. Therefore, an additional (re)orthogonalization technique should be applied for every Lanczos iteration.

It is also possible to take advantage of the Lanczos algorithm to find degeneracy without using block Lanczos, but by repeating Lanczos run multiple times with different initial random vectors as mentioned in Ref. [69]. The overall process is described in Fig. 3.5. For every Lanczos run, a new random initial vector is created to explore a different Krylov subspace. After each Lanczos operation to obtain eigenstates, an orthogonality test and refinement procedure (Gram-Schmidt orthogonalization) is

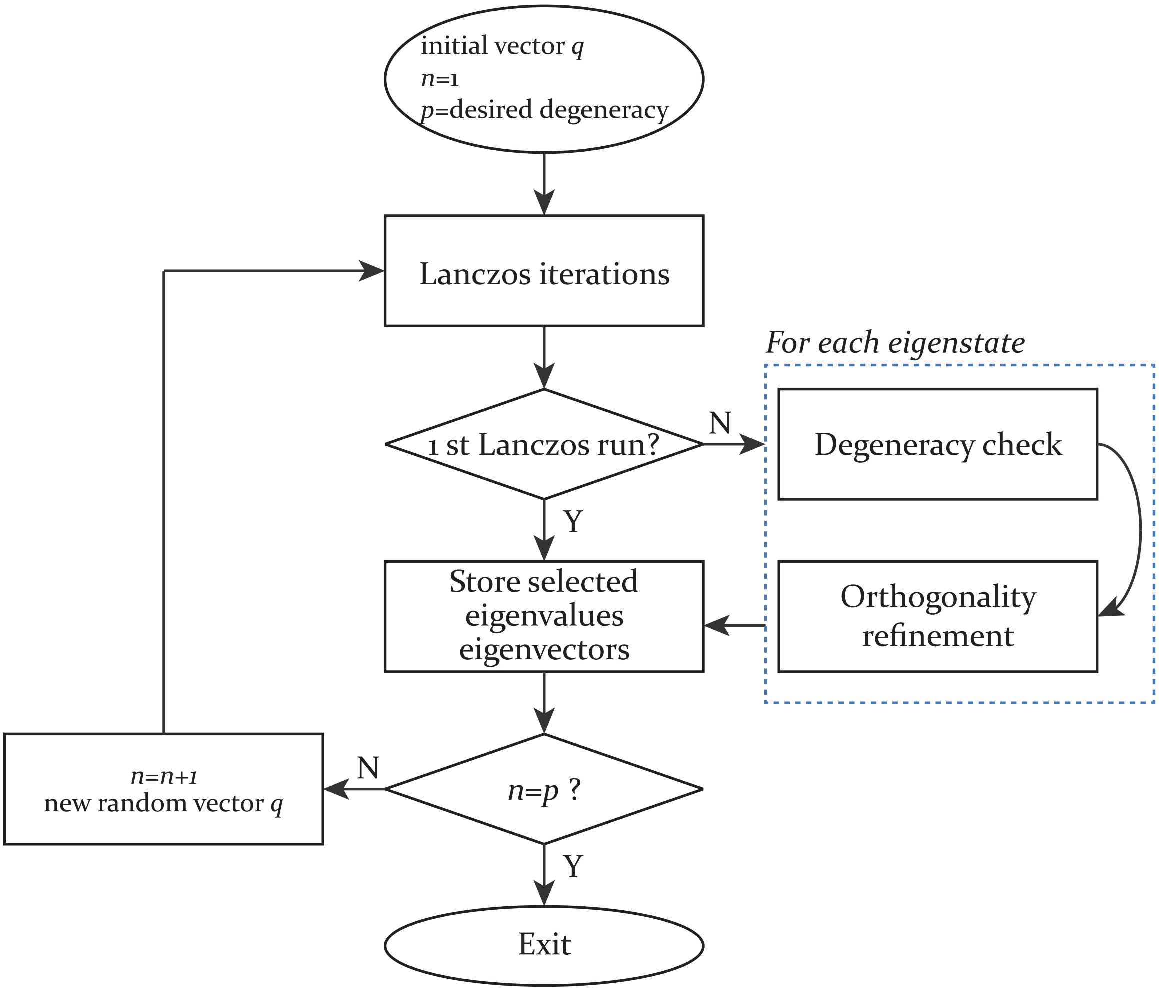
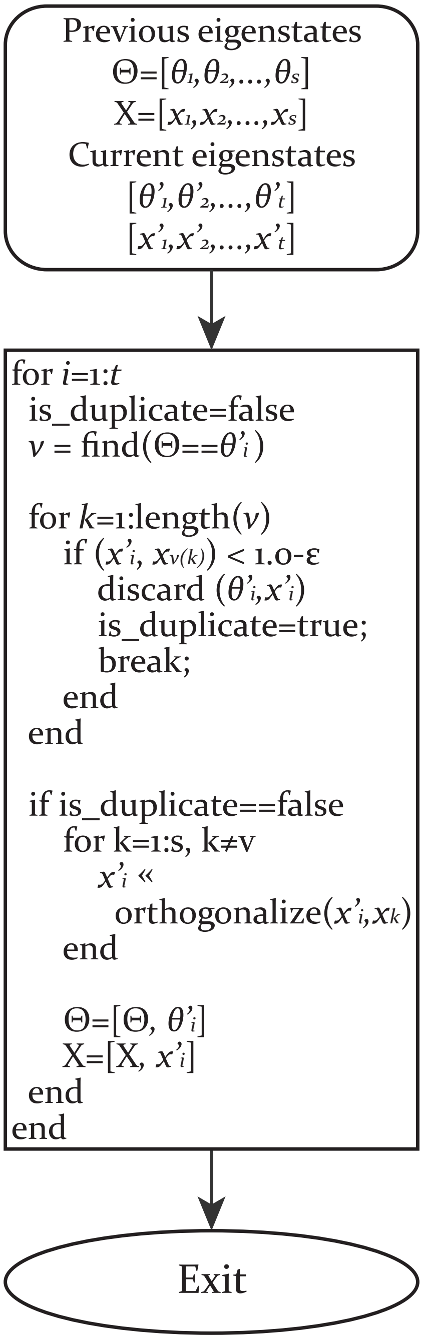


Fig. 3.5. Flow of finding degenerate eigenvalues using multiple Lanczos operations. It is simply running the Lanczos algorithm multiple times, however, between each Lanczos run, an orthogonality check is performed to sort out duplicate eigenvalues. The orthogonality check procedure is described in Fig. 3.6.



```
for $i$=1:$t$
  is_duplicate=false
  $v$ = find($\Theta$==$\theta'_i$)

  for $k$=1:length($v$)
     if ($x'_i$, $x_{v(k)}$) < 1.0-$\varepsilon$
        discard ($\theta'_i$,$x'_i$)
        is_duplicate=true;
        break;
     end
  end

  if is_duplicate==false
     for k=1:s, k$\neq$v
        $x'_i$ «
          orthogonalize($x'_i$,$x_k$)
     end

     $\Theta$=[$\Theta$, $\theta'_i$]
     X=[X, $x'_i$]
  end
end
```

Fig. 3.6. Procedure for detecting multiplicity of each eigenvalue obtained after single Lanczos run. For every new eigenvalue find a match among previously acquired values and compute inner product with all eigenvectors. If the inner product result is exactly the same (inner product = 1) discard the new eigenpair, otherwise it is considered as a degenerate eigenvalue. The degenerate eigenvector then goes through the orthogonalization process (Gram-Schmidt) again and all existing eigenvectors except for degenerate ones are stored in memory.

performed against previously computed eigenvectors (Fig. 3.6). In the orthogonality test phase, first check for each eigenvalue whether same values exist from a previous Lanczos run. Next, for eigenvectors with same eigenvalues compute the inner-product values to determine whether it is a duplicate pair. The new computed eigenpair is discarded if the inner-product is close to 1. Otherwise the eigenvalue is considered to be degenerate and we proceed with the refinement process using Gram-Schmidt orthogonalization against all other eigenvectors. In this process, degenerate eigenpairs do not participate since degenerate subspace vectors are only required to be orthogonal with vectors complement to the degenerate subspace. Finally, compute the Ritz value $x^H Ax/x^H x$ and compare it with the originally computed eigenvalue for a sanity check. Although this simple Lanczos procedure takes 2 (eigenvalue + eigenvector) $\times$ (Lanczos iteration) $\times p$ (multiplicity factor), it provides robust eigenstate solutions without significant numerical errors.

3.6 Results

The effect of each algorithm on Lanczos is benchmarked against PARPACK results on a 3.3×3.3×3.3 (nm^3) silicon quantum dot using a $sp^3d^5s^*$ tight-binding Hamiltonian with spin orbit coupling which creates degeneracy. The maximum iteration count is 3,000 and the convergence check is performed every 20 iterations after the 200-th iteration. The effect of the binning algorithm is first shown in Table 3.6. It clearly shows the random convergence nature of the Lanczos algorithm, which misses eigenvalues. The binning algorithm, on the other hand, actually detects whether there are eigenvalues missing in the first few bins and runs more iterations to find the missing eigenvalues.

Another benchmark result on finding degeneracy with multiple Lanczos runs is shown in Table 3.6. The result shows that multiple Lanczos runs with different initial random vectors actually explore different Krylov subspaces and compute eigenvalues that are degenerate with previous values but result in orthogonal eigenvectors (L2

Table 3.1
Eigenvalue spectrum results for different eigenvalue solvers. The Lanczos solver is benchmarked against PARPACK results. The original Lanczos algorithm (L) is missing eigenvalues within certain iterations while the binning algorithm (LB) detects and successfully finds the missing values by running more iterations.

PARPACK	Lanczos(L)	Lanczos (LB)
1.434398	1.434398	1.434398
1.434398		
1.434788	1.434788	1.434788
1.434788		
1.434796		1.434796
1.434796		
1.437094	1.437094	1.437094
1.437094		
1.437098		1.437098
1.437098		
1.437226		1.437226
1.437226		
1.541293	1.541293	1.541293
1.541293		
1.541930	1.541930	1.541930
1.541930		
1.541945		1.541934
1.541934		

Table 3.2

Benchmark of the eigenvalues computed from multiple Lanczos runs. L2 indicates Lanczos run is performed twice and LB2 stands for two runs of the Lanczos algorithm with binning. The degeneracy of the eigenvalues is resolved by running multiple Lanczos runs. Orthogonality tests effectively filter out duplicate eigenvalues at the third and fourth Lanczos runs. On the contrary, the original Lanczos fails to find the complete spectrum in the range of interest even when running the Lanczos algorithm for four times (L4).

PARPACK	L	LB	L2	LB2	L3	LB3	L4	LB4
1.434398	1.434398	1.434398	1.434398	1.434398	1.434398	1.434398	1.434398	1.434398
1.434398			1.434398	1.434398	1.434398	1.434398	1.434398	1.434398
1.434788	1.434788	1.434788	1.434788	1.434788	1.434788	1.434788	1.434788	1.434788
1.434788			1.434788	1.434788	1.434788	1.434788	1.434788	1.434788
1.434796		1.434796		1.434796	1.434796	1.434796	1.434796	1.434796
1.434796				1.434796		1.434796	1.434796	1.434796
1.437094	1.437094	1.437094	1.437094	1.437094	1.437094	1.437094	1.437094	1.437094
1.437094			1.437094	1.437094	1.437094	1.437094	1.437094	1.437094
1.437098		1.437098		1.437098	1.437098	1.437098	1.437098	1.437098
1.437098				1.437098		1.437098	1.437098	1.437098
1.437226		1.437226		1.437226		1.437226		1.437226
1.437226				1.437226		1.437226		1.437226
1.541293	1.541293	1.541293	1.541293	1.541293	1.541293	1.541293	1.541293	1.541293
1.541293			1.541293	1.541293	1.541293	1.541293	1.541293	1.541293
1.541930	1.541930	1.541930	1.541930	1.541930	1.541930	1.541930	1.541930	1.541930
1.541930			1.541930	1.541930	1.541930	1.541930	1.541930	1.541930
1.541934		1.541934		1.541934	1.541934	1.541934	1.541934	1.541934
1.541934				1.541934		1.541934	1.541934	1.541934

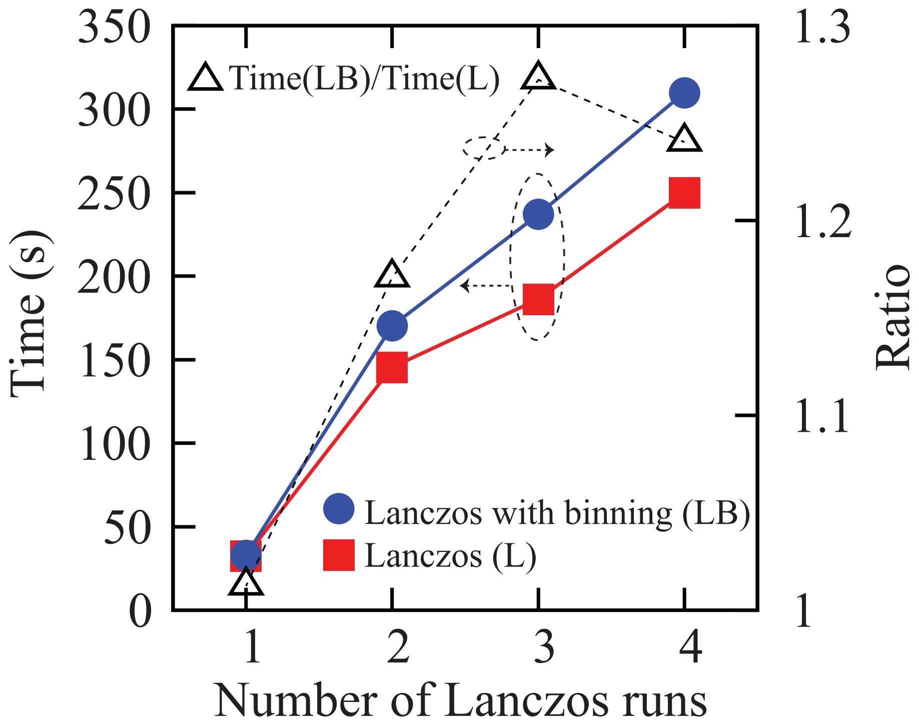


Fig. 3.7. Time comparison between different Lanczos test runs. Blue with circle plot indicates the compute time for the binning algorithm and red with square shows the original Lanczos compute time. The time increases with the number of Lanczos runs and the binning algorithm generally slows down overall performance by 30%.

and LB2 in the Table 3.6). The orthogonality test successfully filters out duplicate eigenpairs in 3 and 4 Lanczos runs (L3, LB3, L4 and LB4) since all degeneracy is resolved after second Lanczos run. In these test cases, the binning algorithm should be used simultaneously since the original Lanczos algorithm still fails to find all eigenvalues even with multiple Lanczos runs (L2, L3 and L4).

Fig. 3.7 shows a computation time comparison between different cases of Lanczos algorithms. Binning generally results in 30% longer compute times as compared to the original Lanczos since the algorithm searches for missing eigenvalues and runs more Lanczos iterations to find missing eigenvalues. On the other hand, it is clear that multiple Lanczos runs take more time to compute degenerate eigenvalues and the compute time is almost proportional to the number of runs. The reason for the steeper increase (factor of 4) between single Lanczos run and double Lanczos run is that eigenvectors are not computed in the eigenvalue calculation while eigenvector is required to resolve degeneracy in the multiple runs. A perfect proportionality relationship is not guaranteed since the convergence pattern is random and differs from each Lanczos run. In practice, a factor of 4∼6 of compute time is needed to guarantee the sequence of the eigenvalues with degeneracy.

3.7 Conclusion

The Lanczos eigenvalue solver algorithm is simple and very useful for computing eigenstates or band structures of large quantum dots or periodic structures on massively parallel computing environments due to its efficient memory use. However, it is slower than typical shift-and-invert Arnoldi solvers like PARPACK, has possibility of missing important eigenvalues near band edges and fails to find multiplicity. A simple binning algorithm is discussed to predict the number of spectrum in a small energy range and to determine whether Lanczos solvers fail to find within limited iterations. Multiplicity can easily be found by running the Lanczos algorithm multiple times with simple degeneracy checks and orthogonalization procedures between each Lanczos solver run. Although two additional algorithms make the Lanczos algorithm slower by at least 30% (binning algorithm only) and 4∼6 times (number of repeated eigensolver runs), they are required for certain applications, such as, finding correct optical gaps or bandstructure calculations of periodic structures with large supercell sizes, to compute correct spectrum of interest. From an implementation point of view, the Lanczos solver embedded in `NEMO3D-peta` does not involve any external packages other than open source packages such as MPI [70] and LAPACK [71], which can be easily migrated to any computing system.

4. ELECTRONIC PROPERTIES OF DISORDERED SI:P δ-LAYERS

4.1 Introduction

There has been rapid progress using scanning tunneling microscopy (STM) to pattern phosphorus donors in silicon using phosphine gas and then encapsulating them with low temperature molecular beam epitaxy (MBE) [12, 13]. The combination of these two technologies has created the possibility for controlling dopant placement in silicon with atomic-scale precision in all three dimensions. Using this technology, experimentalists have built a variety of planar, highly doped phosphorus δ-doped devices embedded in silicon (Si:P) such as tunnel junctions, quantum dots and nanowires [8, 11, 72]. More importantly, the precise incorporation of donors enables the potential realization of donor-based quantum computers [1, 3, 5, 7, 73].

Central to these planar Si:P device architectures is a highly conductive 2D δ-doped sheet [74–77]. By patterning the 2D δ-doped layers into specific geometries, they can act as both Ohmic contacts as well as gates for the control of electron and spin transport through singly placed donor impurities in quantum computing architectures. Understanding the impact of impurity placement and position both within the plane and vertically in 2D δ-doped sheets is important for understanding the electron transport in highly confined, STM-patterned architectures and essential for continuing efforts in experimental device design.

Over the past few years, a detailed understanding of the incorporation mechanism of P atoms into silicon using phosphine gas has been developed [78–80]. From this understanding it has been possible to use an STM to lithographically position single P atoms into the top atomic layer of silicon by opening a hole in a hydrogen resist and annealing to temperatures of 350°C [80]. This anneal causes the phosphine gas to loose its hydrogen atoms on the surface before a single P atom can incorporate into the top layer of silicon, displacing a silicon atom. The incorporation anneal can also be performed on a phosphine saturation dosed sample. At saturation dosing after room-temperature exposure, depending on the dosing conditions the surface coverage will be a disordered alloy of PH_2+H and PH+2H. However, after the incorporation anneal the final P dopant density invariably takes the nominal value of 0.20∼0.25 ML with the P atoms located in random positions within the top layer of the Si surface [79]. The high doping density means that the P atoms are typically 1nm apart, which is much smaller than the Bohr radius. As a result, one can expect extensive wavefunction overlap and metallic-like behavior within atomically controlled 2D nanostructured domains.

There have been efforts made to experimentally identify the electronic structure of MBE fabricated δ-doped layer in silicon by Eisele *et al.* using resonant tunneling spectroscopy [81–83]. The fabricated δ-doped layer was about 2 nm thick with the doping density exceeding 10^{13} cm^{-2} and it showed quantized energy levels originating from 2D sub-bands confined in the layer. To compute the electronic structure of δ-doped layers, several theoretical studies have been published. Initially, the potential profile was computed using a Thomas-Fermi approximation and then superimposed to the diagonal elements of the Schrödinger equation to compute confined energy levels in different types of δ-layers in silicon [84–87]. More rigorous self-consistent calculations were carried out by Qian *et al.* by computing such systems using planar Wannier orbitals based on an empirical pseudopotential method (PP) and parabolic dispersions of silicon sub-bands [88]. Cartoixà *et al.* focused on determining Fermi level fluctuations with temperature and doping density in the Si:P δ-layer structure using atomistic sp^3s^* tight-binding (TB) calculations with an anti-bonding orbital model [89]. However the sp^3s^* model is well known to misrepresent X points in the conduction band [37], and it is clear that the basis size is too small to describe L and X valleys [36–38]. Carter *et al.* has calculated bandstructures of Si:P δ-doped layers using different atom configurations within density functional theory (DFT) [90].

The computational burden of DFT calculations, however, prevents the method from extending to 2D systems and structures with realistic degrees of disorder, which require large supercell domains. To overcome the computational limitation, a mixed-atom pseudopotential (MP) was recently examined, which reduces the level of *ab-initio* input. This method has been shown to compare well with earlier theoretical studies [91]. In the MP approach, however, atomistic effects cannot be handled since the model assumes an averaged nuclear charge between silicon and phosphorus in the δ-layer.

Thus, to date all theoretical works have focused on either ordered configurations or limited disorder using small domain sizes; none of them have been able to investigate the effect of realistic disorder in Si:P δ-doped layers using an *extended* domain, which is critical for a reasonable approximation of random dopant placement. Examining how disorder plays a role in the electronic properties of Si:P δ-doped layers will ultimately provide a critical theoretical background for experimentalists. Therefore, the focus of this paper is to develop the methodology to handle sufficiently large domains to validate this approach against others, and then to investigate the effect of horizontal and vertical disorder on the electronic structure in highly doped monolayer systems.

Atomistic representation in realistically extended spatial domains is essential to represent dopant disorder effects. The Nanoelectronic Modeling Tool (NEMO3D) [28–30, 50] can simulate atomistic structures of realistic size and include non-parabolicity of bulk materials automatically by using an empirical $sp^3d^5s^*$ TB model. NEMO3D has been successful in modeling a spectrum of systems in which atomistic details and interface effects are important to understand device behavior, such as phosphorus impurities in silicon devices [14, 39, 40, 92, 93], valley splitting in miscut Si/SiGe quantum wells [46] and InGaAs embedded InAs quantum dots [54]. Having demonstrated our ability to model phosphorus in silicon accurately using an atomistic approach, we expanded NEMO3D's capabilities to run efficiently on peta-scaled computer systems and included a charge-potential self-consistent loop [28–30, 50]. We now apply

NEMO3D-peta to highly doped Si:P system to explore atomistic effects on the electronic structure of δ-doped layers.

This chapter is organized as follows. Section 2 summarizes the simulation methodology and structure modeling. Section 3 discusses the electronic structures of Si:P δ-doped layers and shows the effect of disorders in bandstructures. Section 4 summarizes and concludes the chapter.

4.2 Methodology

Simulation structure: The physical structure used in this work is depicted in Fig. 4.1(a). To represent the infinite sheet of the Si:P δ-layer buried in silicon, periodic boundary conditions are applied to both in-plane directions ([100]/[010]). The silicon substrate and encapsulation layer along [001] is assumed to be intrinsic. We have calculated that at 4K, a minimum confinement thickness of 120 monolayers (ML), or approximately 16nm is needed to avoid hard wall boundary effects. At such large encapsulation thicknesses we find variations of eigenenergies smaller than 3 meV upon further increase in buffer size. Fig. 4.1(b) shows the top view of δ-layer with 1/4ML ($1.7{\times}10^{14}(\mathrm{cm}^2)$) doping density in a *perfectly ordered* $p(8 \times 8)$ unit cell. In such an ordered configuration, the minimum supercell that we can use is $p(2 \times 2)$, which corresponds to 960 atoms with a 120ML buffer. The $p(8 \times 8)$ cell corresponds to a total atom count of 15,360 atoms. Fig. 4.1(c) shows an example of a disordered dopant configuration. Scrambling the dopant placement can be readily achieved due to the nature of atomistic simulations. However, disorder simulations require a larger supercell of at least $p(8 \times 8)$ to account for random effects in bandstructures. A similar approach can be applied for electronic structure simulations of III-V or SiGe alloys [94]. The vertical straggle of dopants can also be simulated by using Gaussian distributions to model diffusive penetration of dopants from the δ-doped plane (Fig. 4.1(d)). We use a measurement temperature of 4K for all numerical experiments. At 4K it is valid to only consider the conduction band occupation since electrons

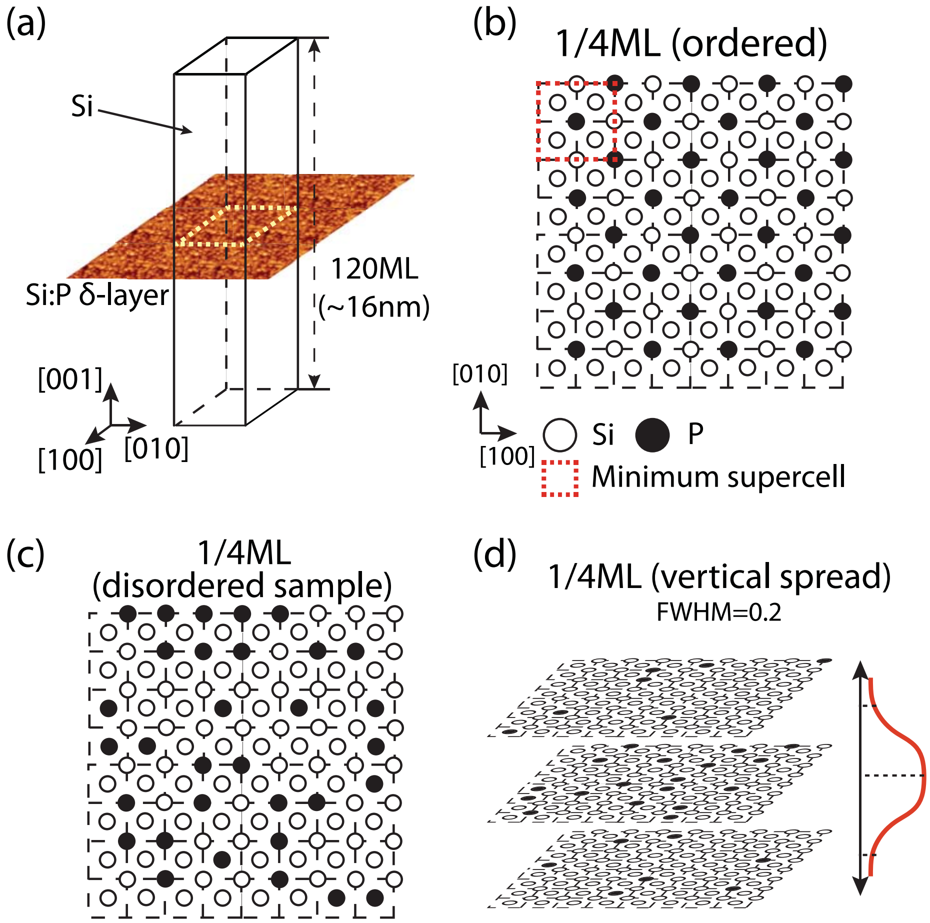


Fig. 4.1. (Color online) (a) The simulation structure used to represent a 2D Si:P δ-doped layer encapsulated by silicon of thickness 120ML. 2D periodic boundary conditions are imposed along the doping plane. (b) A perfectly ordered 1/4ML (1.7×10^{14} (cm^{-2})) Si:P supercell used for the atomistic simulations. In this case, $p(2\times2)$ represents the smallest supercell marked in red. (c) An example of disordered supercell, where $p(8\times8)$ is used for disorder simulations. (d) An example representing vertical segregation of the dopants with a Gaussian distribution (FWHM=0.2 (nm)). Note: the lattice constant of silicon is $a = 0.54$ (nm).

are only coming from the donor levels and not from thermal excitation from valence bands. Bandgap and valence electrons are therefore ignored in the charge calculations presented here, even though our 20 band $sp^3d^5s^*$ model provides accurate results.

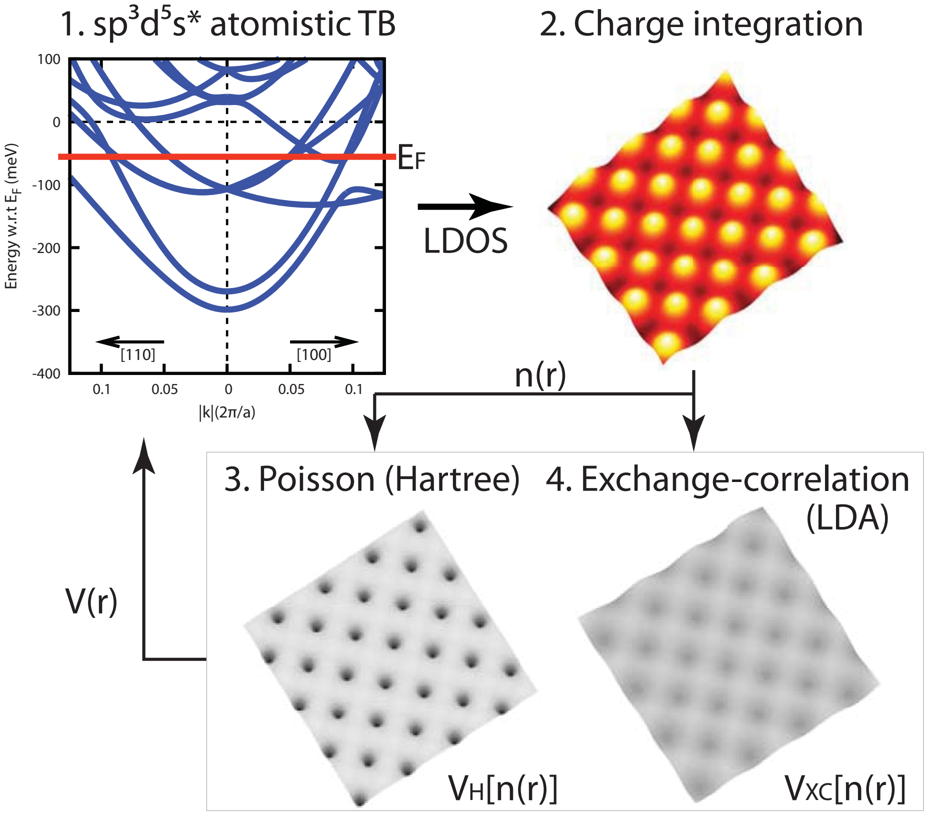


Fig. 4.2. The graphical representation of self-consistent bandstructure calculation. Eigenstates are computed out of the single electron Schrödinger equation for each k point and sub-band. The equation is comprised of an atomistic TB Hamiltonian and external potential profiles. The Fermi level (E_F) is determined at the charge neutrality condition which equates the number of electron charge to the number of total impurity charge. The charge is then integrated to determine the electron density $(n(\mathbf{r}))$. The potential profile is subsequently computed at a given electron density. The electrostatic potential $(V_H[n(\mathbf{r})])$ is computed by solving Poisson's equation based on a finite difference grid. A correction is included for electron-electron interaction effects $(V_{XC}[n(\mathbf{r})])$ based on LDA [95]. The potential profile is the sum of electrostatic and exchange-correlation potentials and it is finally fed back into the Schrödinger solver for the next iteration of the charge calculation.

Self-consistent procedure: The simulation approach used in this work is to obtain the potential profile and bandstructure with charge density based self-consistent calculation using `NEMO3D-peta` [47,51]. The self-consistent loop has been applied previously to study the temperature dependence of electronic properties in Si:P δ-doped layers [51]. The graphical representation of the self-consistent methodology is shown in Fig. 4.2. First, eigenstates of the δ-doped layer from an atomistic $sp^3d^5s^*$ TB single electron Hamiltonian [36] are computed over the first Brillouin zone (BZ) in the discretized 2D k-space. Only conduction band states are of interest in these simulations, since the intrinsic carriers are frozen out at low temperature and all carriers in the δ-doped layer are provided by the donors. We can therefore reduce the basis size from 20 to 10 in the $sp^3d^5s^*$ model by only considering a single spin explicitly. This reduces the computational burden by at least a factor of 2. The Fermi level (E_F) of the system is determined iteratively by the charge neutrality condition, which assumes the total number of electrons to be equal to the number of donors. The local density of states (LDOS) is obtained by binning k states and by integrating the LDOS over the occupied 2D k-space, resulting in the electron charge profile ($n(\mathbf{r})$) of Si:P δ-doped layer.

The electron charge density profile is used to determine two different terms that enter the Schrödinger equation: the Hartree potential and the exchange-correlation potential. The Hartree term ($V_H[n(\mathbf{r})]$) can be obtained by solving Poisson's equation with a given electron and donor ion profile. All charge contributions are treated as local point charges on a zincblende lattice. Electron-electron interactions must also be taken into account in such many-electron systems. To first order, we include an analytical form of exchange and correlation functionals ($V_{XC}[n(\mathbf{r})]$) [95] based on the local density approximation (LDA) [96], which lowers the total energy of the system and modifies the wavefunction [97]. To obtain the local charge density in space from the point charge in the atomistic grid, the charge is assumed to be uniform within a finite volume around each atom. The volume around each site is computed as the volume of the unit cell divided by number of atoms per unit cell. Therefore, the local electron charge density used for the exchange-correlation potential can be calculated as the amount of charge at each site divided by its surrounding volume. Note that V_{XC} is treated as a linear function of electron density and therefore is computed only for electrons from donors, assuming the exchange-correlation effect of the frozen-out valence electrons ($V_{XC}[n_V(\mathbf{r})]$) of silicon is inherently included in the TB Hamiltonian ($V_{XC} \approx V_{XC}[n(\mathbf{r})] + V_{XC}[n_V(\mathbf{r})]$). More rigorous calculation considering the nonlinear behavior of V_{XC} in atomistic TB formalism is beyond the scope of this paper but can be found in other density-functional based methodologies that also utilize a TB scheme [98].

In general, convergence is difficult to achieve because of the sharp potential variations around each impurity location and the low temperature condition. Therefore, the under-relaxed potential ($V(\mathbf{r})$) is updated for the next iteration of the charge calculation and the charge-potential loop is continued until the mean square value of the potential is converged to within 0.1meV. Convergence is typically achieved in about 25∼35 iteration steps. The computations are carried out on state-of-the-art cluster machines [58] with 2GB of memory per core and a typical total compute time of 48 hours using 256 cores. Furthermore, with the computational advantage over other *ab-initio* approaches, this methodology can also be expanded to compute the electronic and low-bias conduction properties of 1D periodic Si:P nanowires [17].

4.3 Results

4.3.1 Equilibrium properties of the ordered Si:P δ-layer

Bandstructure: Fig. 4.3(a) compares the equilibrium bandstructure of the 1/4ML Si:P δ-doped layer plotted with respect to the *silicon bulk band minimum* with the bandstructure of the pure silicon structure (without the δ-doped layer) of the same minimal unit cell $p(2\times2)$ as depicted. The δ-doped layer creates a strong confinement (Fig. 4.3(b)) and pulls down the bands significantly, causing large splitting between confined sub-bands. The positions of the 1Γ, 2Γ, Δ valleys and Fermi level (E_F) all reside under the silicon bulk band edge and are within the confinement potential created by the δ-doped layer. The first few sub-bands in this bandstructure can be easily interpreted by the band projection of the bulk silicon valleys as shown in Fig.

4.4. The two out-of-plane valleys marked in dark color are projected to the $k_x - k_y$ plane to form 1Γ and 2Γ bands, while the remaining four in-plane valleys are projected along $k_{x,y} = \pm 0.18 \times 2\pi/a$ in the reduced zone scheme. The lower quantization energy ($\propto m^{*-1}$) of Γ-projected valleys is always observed since the confinement mass is larger in Γ-projected valleys ($m_l = 0.91 m_0$) than in the remaining in-plane valleys ($m_t = 0.19 m_0$). The sharp electrostatic confinement in the z-direction due to the screened donor potential creates a very narrow quantum well, which results in a large valley splitting (VS, $E_{2\Gamma} - E_{1\Gamma}$) of ∼25 (meV). This VS in a V-shaped QW [41] is an order of magnitude larger than the VS of typical Si-SiGe quantum wells [42].

Table 4.1 compares the Fermi level and valley minimum values with results from other methods for the ordered 1/4ML δ-doped layers. The overall comparison results show that the atomistic TB approach predicts reasonable values for the band minima and Fermi level with respect to other methods. Computed Fermi levels stay very close to each other except for the MP method. Since the MP method uses an averaged dopant representation, it may result in a weakly confined potential, which causes a

Table 4.1
Energies obtained from different models of the 1/4ML Si:P 2D δ-layers showing the Fermi energy, 1Γ, 2Γ and 1Δ bands in meV (Reference: Silicon E_C=0.0 (meV))

Approach	E_F	1Γ	2Γ	1Δ
This work	**-110**	**-394**	**-369**	**-242**
Wannier/PP [88]	-111	-410	-400	-270
TB (sp^3s^*) [89]	-110	N/A	N/A	N/A
DFT [90]	-110	-540	-420	-210
DFT/MP [91]	-62	-445	-425	-236

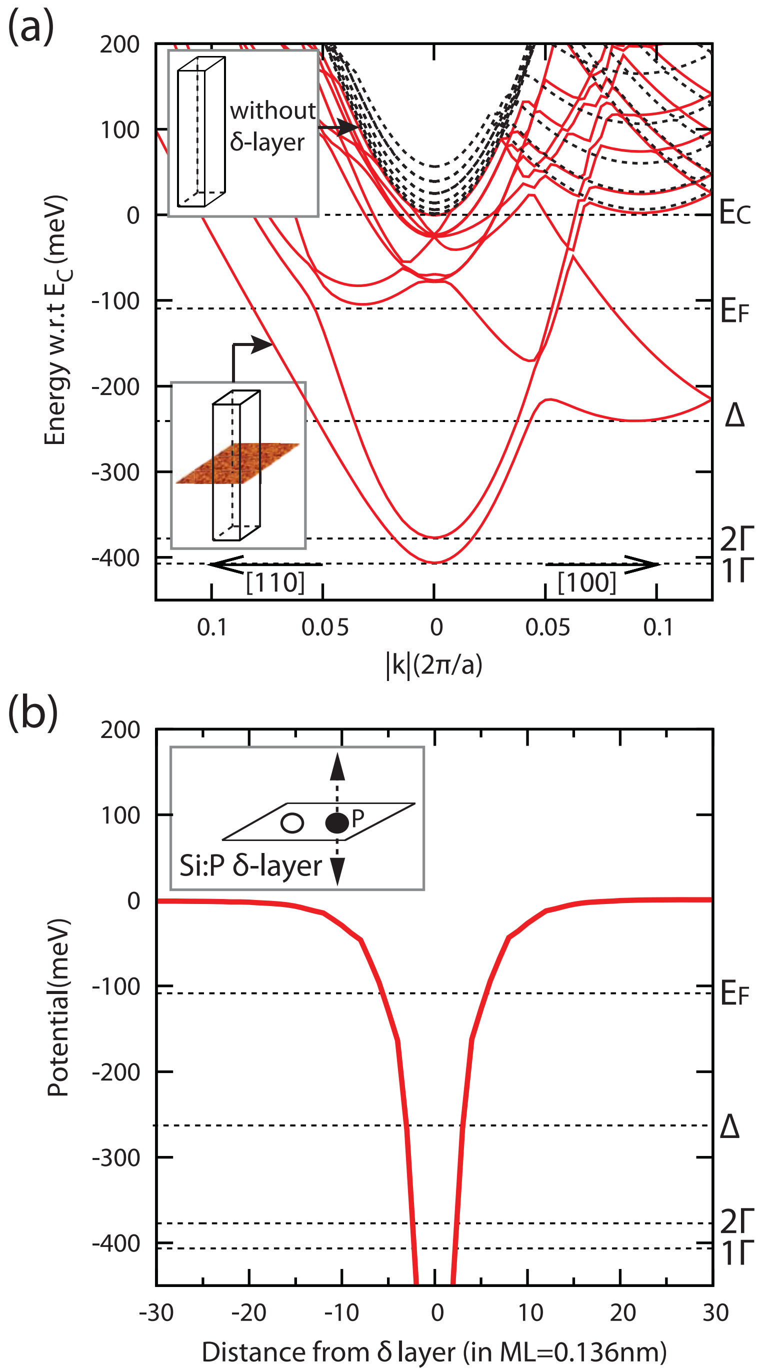


Fig. 4.3. (Color online) (a) Equilibrium bandstructure of an ideal ordered 1/4ML Si:P δ-doped layer at 4K. The bandstructure of a pure silicon structure with the same dimension without the δ-doped layer is plotted (dashed line) for comparison. (b) Potential profile plotted along confinement direction ([001]) passing through impurity site. The relative position of the valley minimum point and Fermi level with respect to the silicon conduction band minimum is indicated.

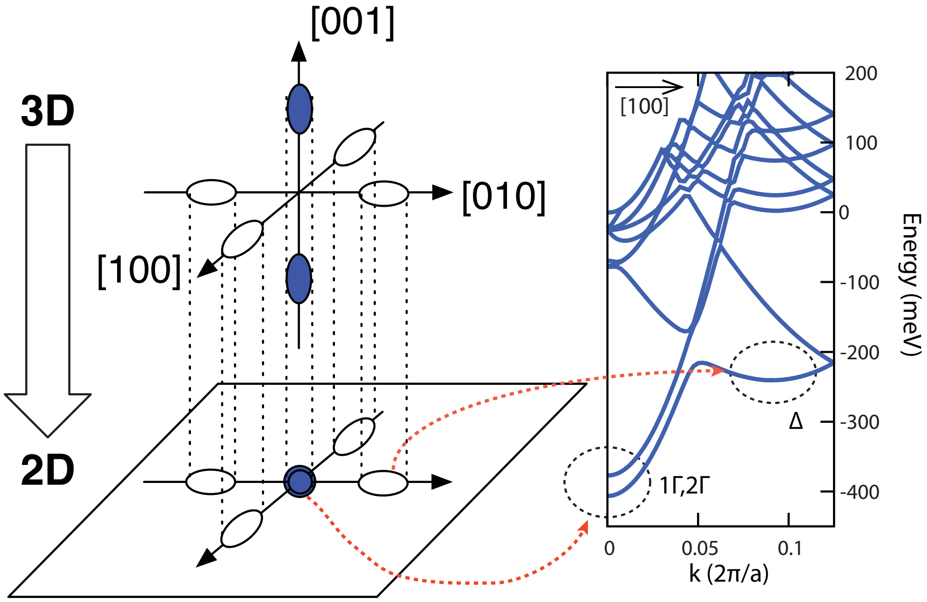


Fig. 4.4. Band projection diagram for the highly confined 2D δ-layers. 2 valleys (dark) along the z-confinement direction are projected to the Γ point and 4 other valleys are projected to their own positions.

Fermi level shift and reduced valley splitting [91]. DFT predicts a larger VS and lower Γ valleys but a higher Δ valley. PP shows similar values to our results but smaller VS. A major difference of the PP method can be seen by comparing the details of the bandstructure, which for PP is based on a parabolic band assumption [88]. Details, such as, non-parabolicity and band anti-crossing, which may cause non-linear modulation of low-bias conductance, are not taken into account in the PP method.

Potential and charge: The charge profile along the confinement or z-direction is shown in Fig. 4.5. 96% of the charge is confined vertically within 21 monolayers (<3 (nm)) from the donor positions, leaving the donor charges in the δ-doped layer perfectly screened. Therefore, the potential profile along the confinement direction decays faster than the ideal Coulombic potential ($\propto r^{-1}$) and the local field vanishes within a range of ±20 monolayers (±3 (nm)) from the δ-doped layer. Our charge distribution (FWHM = 0.81 (nm) = 7ML) agrees well with both DFT and MP calculations, which predict 0.67 (nm) (DFT), 0.84 (nm) (MP), respectively [90, 91].

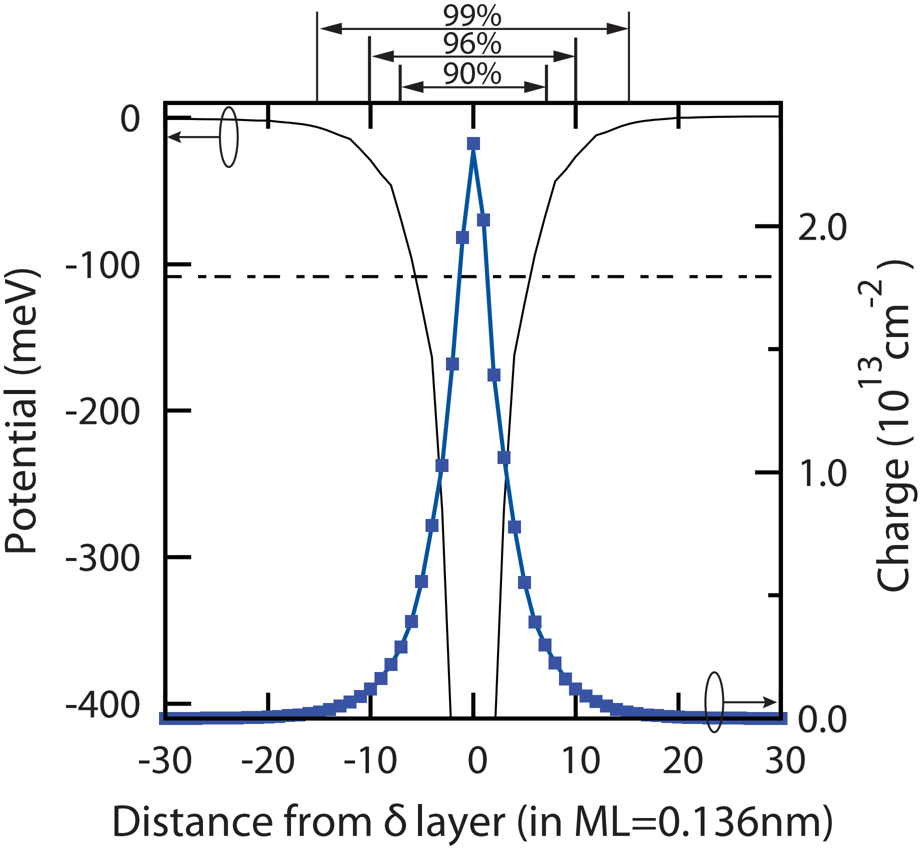


Fig. 4.5. (Color online) Charge profile of a perfectly ordered 1/4ML Si:P δ-doped layer at 4K. 90%, 96% and 99% of the charge is confined within 15 (2.03 (nm)), 21 (2.85 (nm)) and 31 (4.21 (nm)) (ML), respectively. For reference, FWHM=7ML (0.81 (nm)) and it contains 66% of the charge.

The temperature dependence of the charge screening in such Si:P δ-doped layers is described further in Ref. [51].

4.3.2 The effect of doping density in the Si:P δ-layer

To examine the doping density dependence on the Si:P δ-doped layer bandstructure, different numbers of impurities are placed in a $4a \times 4a$ simulation domain. The doping density conversion table which relates the discrete atomistic density with the more common per-square-cm density is provided in Table 4.2. As the doping density increases, more sub-bands are occupied under the Fermi level to maintain charge neutrality. Stronger electrostatic coupling contributes to a further down-shift of the sub-bands, promoting an increased number of occupied sub-bands (Fig. 4.6(a)). Our model predicts a linear dependence of the critical 1Γ, 2Γ, and Δ energies as a function of doping density as shown in Fig. 4.6(b). A slight increase in VS as the doping density increase indicates that the potential confinement of the δ-doped layer is also becoming stronger. A more gradual trend of the Δ valley compared to Γ valleys is predicted since the density of states (DOS) effective mass of the Δ valley is larger. In other words, even a small inclusion of sub-bands originating from Δ valleys causes a larger increase in DOS and occupied states compared to the lighter Γ valley sub-bands.

4.3.3 The effect of disorder in the Si:P δ-layer

In reality, it is impossible to make a perfectly arranged infinite δ-doped layer. Random dopant incorporation causes a disordered donor configuration within the Si:P δ-doped layer. The Si:P system can be viewed as a *random alloy* system with a different set of bonding parameters (Si-P, P-Si, Si-Si) and an additional electrostatic potential caused by the donor charges screened by their electrons. This is analogous to typical III-V and Si-Ge alloys in heterostructures that have different TB and strain

Table 4.2
Conversion table between doping constant and number of impurities in the $4a \times 4a$ supercell. Total number of atoms in the δ-layer is 32.

P coverage	1/6.4	1/5.3	1/4.6	**1/4.0**	1/3.6	1/3.2	1/2.9
Doping (10^{14} cm^{-2})	1.06	1.27	1.48	**1.70**	1.91	2.12	2.33
No. impurities	5	6	7	**8**	9	10	11

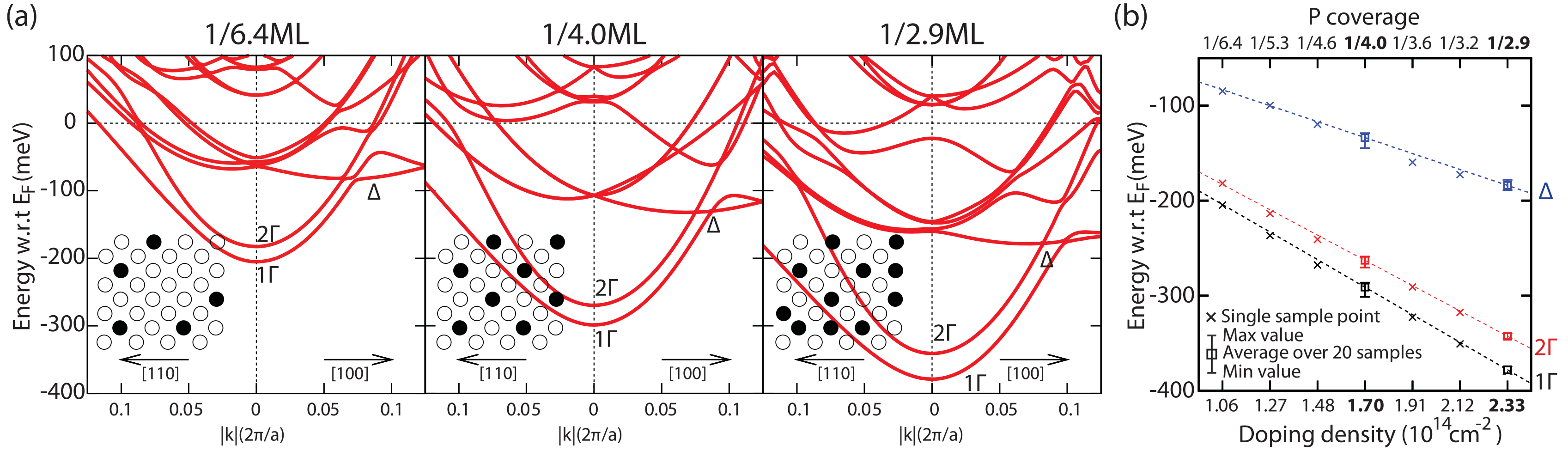


Fig. 4.6. (Color online) (a) Bandstructure results of 1/6.4ML, 1/4.0ML and 1/2.9ML δ-doped layers. Inset in each bandstructure plot provides dopant placement used for the simulation. Due to odd number of dopants in 1/6.4ML and 1/2.9ML cases, unavoidable disorders are present. (b) Valley minimum values of different doping densities plotted with respect to the Fermi level. Statistical samples of 1/4.0ML and 1/2.9ML cases are plotted against other doping cases.

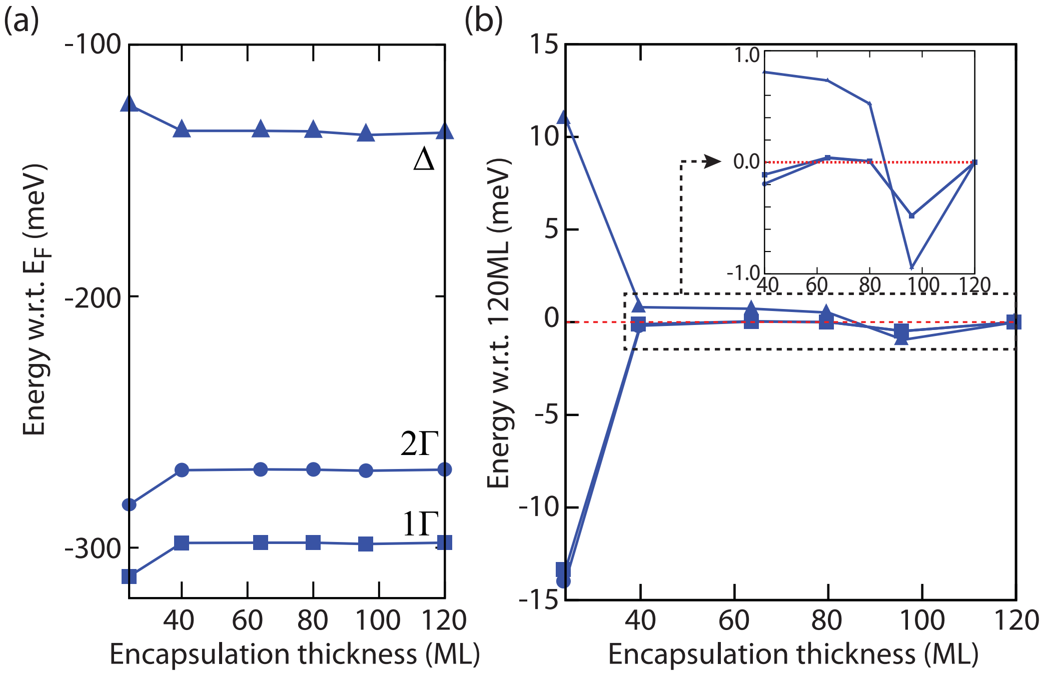


Fig. 4.7. (a) Eigenvalues of valley minima with respect to Fermi levels vs. encapsulation thickness. The relative position stays constant down to 64ML regardless of the thickness. (b) Energy change of 1Γ, 2Γ and Δ measured against 120ML structure. Inset provides a zoom-in view of energy difference down to 40ML and difference is less than 1meV.

parameters and bond lengths. For δ-doped layers, however, it is computationally more intensive to obtain the dispersion since the potential has to be computed self-consistently. The simulation of such random alloy systems is generally performed with repeated supercells that represent the randomness. While true bandstructures only exist for large supercells, the existence of bandgaps and effective masses in alloys validates the concept of an approximate bandstructure [99].

Supercell geometry: To represent the electronic properties of a realistic alloy system, a large enough supercell is needed to mimic the random nature of the target system and to consider a sufficient number of statistical samples. Since the potential profile in a Si:P δ-doped layer requires heavy computation, it is difficult to collect enough samples without a reduced supercell. Therefore, to set up a reasonable supercell geometry, the size is first increased to $8a \times 8a$ in the periodic plane to appropriately represent randomness (Fig. 4.1). The cladding thickness is reduced to save the overall computational burden and to enable the collection of a sufficient number of samples. To determine the minimum cladding thickness to satisfy the above conditions without losing any physical meaning due to artificial domain boundaries, the valley minimum with respect to corresponding Fermi level is compared with varying encapsulation thicknesses (120ML, 96ML, 80ML, 64ML, 40ML, and 24ML). As shown in Fig. 4.7, valley energies vary little within 1meV indicating the minimal effect on the bandstructure down to 40ML. Therefore, it is reasonable to reduce the encapsulation thickness from 120ML to 64ML in modeling disordered samples, which reduces the overall computational burden by 50%.

In-plane disorder: Initially, we considered a δ-doped layer with a disordered dopant configuration within the atomic dopant plane. Fig. 4.8 compares the effect of disorder with a 1/4ML ordered supercell. Multiple band crossings of the Γ and Δ originating bands occur due to the repeating supercell structure as shown in Fig. 4.8(a). In contrast, disordered configurations couple some of these bands such that they anti-cross. The disordered dopant configuration used for the disordered sample (Fig. 4.8(a)) also breaks the translational symmetry along [100]/[010], introducing distortions to every sub-band. AC1$\sim$AC2 and AC3$\sim$AC4 labeled in the bandstructures in Fig. 4.8(a) are major anti-crossings in impurity sub-bands for ordered and disordered cases, respectively. Fig. 4.8(b) and (c) compare the effects of disorder on the potential and charge density in the impurity plane. The disordered charge shows a significant charge accumulation in impurity clusters, while the ordered array shows a much smoother background charge distribution. However, the randomness causes little change in the positions of the 1Γ, 2Γ and Δ valleys, as seen in Fig. 4.6(b) where the statistical results of valley energies for 1/2.9ML and 1/4.0ML are indicated.

Since the bandstructure of the large supercell is complicated without much insight beyond the lowest 2 band edges, we also study the density of states (DOS). Fig. 4.9 compares the DOS between the ordered and disordered Si:P δ-doped layers. The

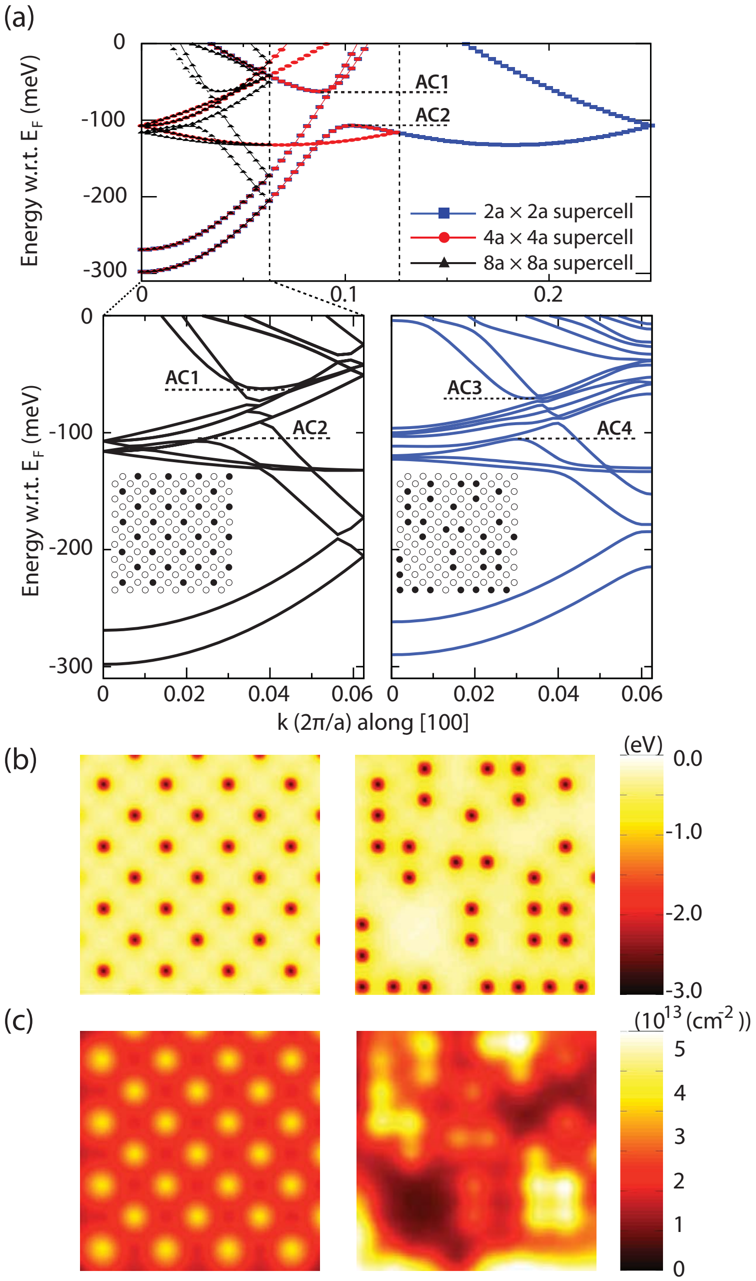


Fig. 4.8. (Color online) (a) Bandstructure comparison between ordered and an example of disordered supercell. Zone unfolding relationship in the perfectly ordered case is displayed (top) and the bandstructure of ordered 8a×8a supercell (bottom left) is shown next to the disordered supercell (bottom right) for direct comparison. AC1∼AC2 and AC3∼AC4 indicate gaps due to the band anti-crossing in the ordered and disordered supercell, respectively. The anti-crossings affect the density of states distribution, which will be discussed in Fig. 4.9. Comparison of the (b) Potential and (c) charge profile between ordered and in-plane disordered supercell shown in (a).

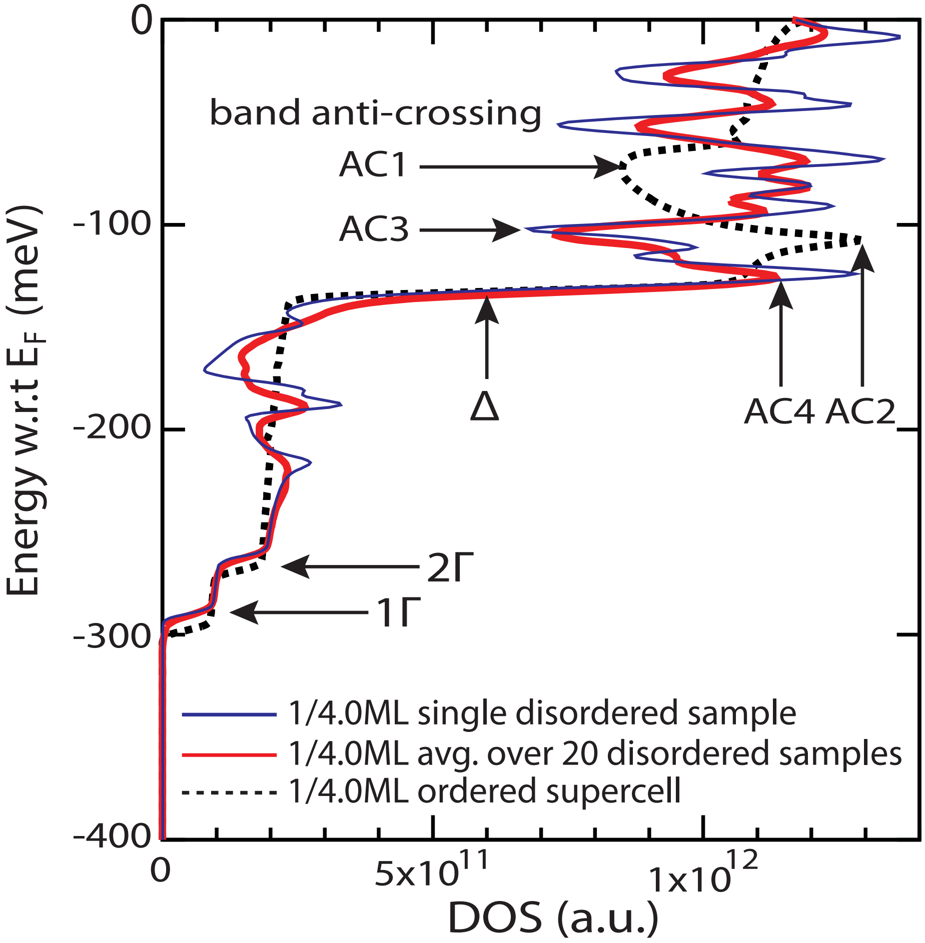


Fig. 4.9. (Color online) DOS comparison between ordered (dashed line) and disordered Si:P δ-doped layers (solid lines). Fluctuations in DOS for ordered supercell is shown between AC1 and AC2 which is caused by band anti-crossing labeled in Fig.4.8(a). Similar fluctuation is also captured in disordered supercell between AC3 and AC4. For disordered cases, the DOS for single sample (blue) and the DOS is averaged over 20 samples (red) are indicated.

ordered layer (Fig. 4.9 - dashed lines), shows a non-parabolicity of the first two sub-bands as seen from the gradual increase in the DOS. A perfectly parabolic dispersion would show a flat DOS. The steep increase in the DOS at -130 (meV) indicates the turn-on of the Δ bands, which have a larger DOS mass ($m_l = 0.9 > m_t = 0.19$). The first sub-band (1Γ) turns off at around -100meV due to band anti-crossing, resulting in a decrease of the DOS at this energy as observed. At the AC1∼AC2 gap in Fig. 4.9, the DOS is lowered due to anti-crossing of the impurity bands indicated in Fig. 4.8(a).

The DOS for the disordered δ-doped layer (red line) has a couple of interesting features compared to the ordered case. The disordered DOS in Fig. 4.9 is averaged over 20 statistical samples. Again, the Δ valley contribution can be identified easily regardless of complicated sub-band splitting. This sub-band splitting in addition to the gaps in the BZ boundaries create additional fluctuations in the DOS as indicated in the energy range AC3∼AC4 of Fig. 4.8(a) and 4.9. Despite these additional DOS fluctuations, the overall DOS appears very similar to the ideally ordered DOS and the DOS is large enough to constantly provide electrons to the attached narrow leads, which clearly have a smaller DOS. We note that it is the DOS fluctuation at the Fermi level that will modulate the conductance in the low bias regime. However, the variation of the conductance is expected to be minimal among samples fabricated under same doping density.

Out-of-plane (vertical) disorder: Finally to simulate what would happen if there were dopant diffusion leading to disorder out of the δ-doped layer, a Gaussian distribution of dopants with a varying FWHM is considered assuming a vertical segregation of no more than 7 layers, or 0.81 (nm). To date the maximum limit of vertical segregation of δ-doped layers encapsulated at low temperatures of 250 (°C) has been measured experimentally to be 0.58(nm) [74]. To mimic this finding in simulation, 1/4ML is taken into account and an ensemble of 20 samples for every case - FWHM=0.0, 0.15, 0.2, 0.3 and 0.4 (nm) - is considered.

Fig. 4.10(a) compares the valley minimum values of these samples. Spreading out the doping out of the central layer can be associated with a weak doping reduction in that particular layer. Such doping reduction allows the impurity bands to rise slightly in energy (compare with Fig. 4.6). A more significant effect is seen by the reduction of the strong confinement, which will be most evident in the VS between the Γ valleys (Fig 4.10(b)). A perfect Si:P δ-doped layer exhibits large VS (∼27 (meV)) with small

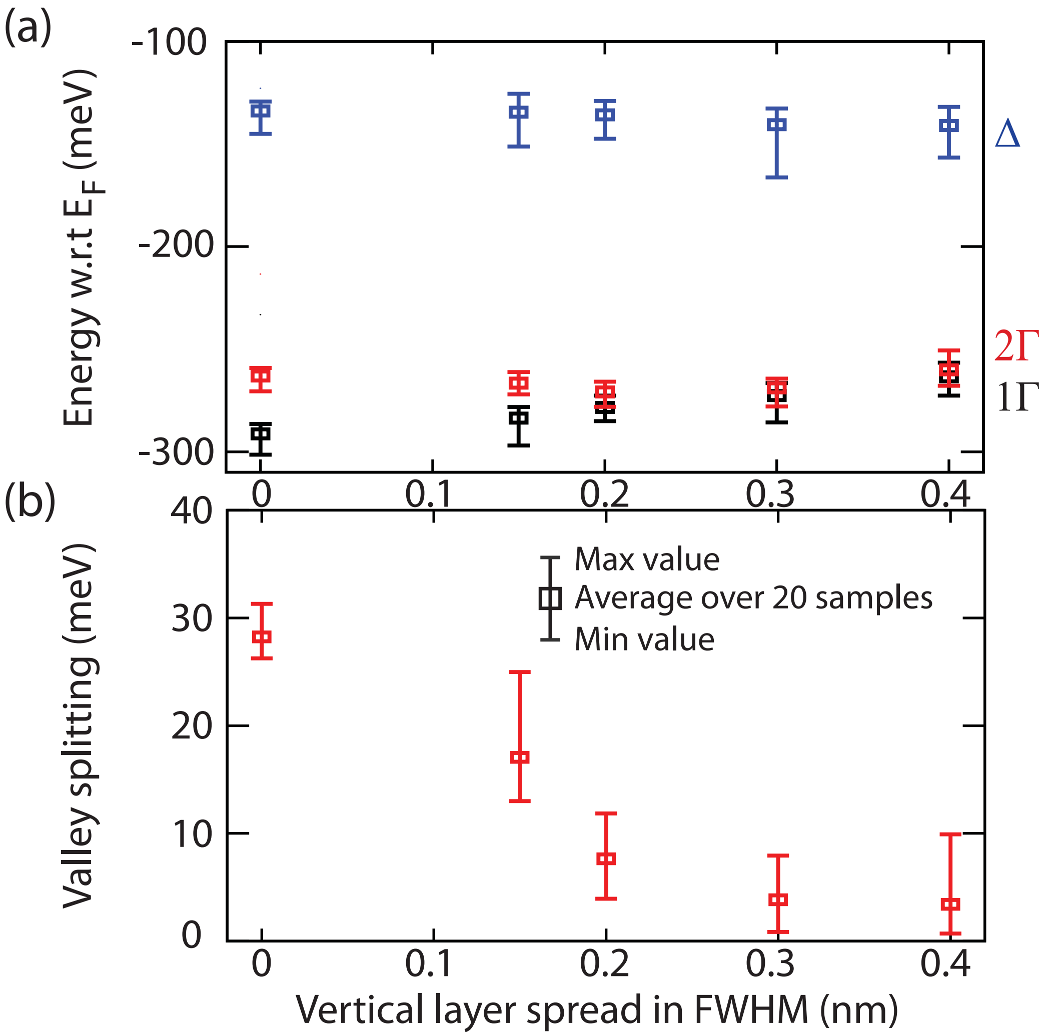


Fig. 4.10. (Color online) (a) 1Γ, 2Γ and Δ valley minimum values are plotted with respected to the Fermi level versus vertical impurity segregation. Statistical samples of vertical segregation of dopant atoms with a nominal 1/4ML doping is considered. (b) Valley splitting as a function of vertical segregation.

variations. On the contrary, as the vertical impurity segregation increases, the VS decreases significantly. As a result measuring the VS experimentally, such as by using Schottky-barrier tunneling spectroscopy, [81, 83, 100, 101] can be used to determine the degree of vertical dopant diffusion.

4.4 Conclusion

We use an empirical tight binding, self-consistent potential approach to model realistically extended Si:P δ-doped layer structures. The methodology is validated against other approaches, such as, DFT and pseudo-potential methods. The scalability of the NEMO methodology enables us to study supercells that resemble realistically disordered systems. We compare statistical samples for dopant disorder in the doping plane and out of the doping plane, and study the sensitivity to doping density. The δ-doped layer creates a Coulombic quantum well that confines electrons in a dense quasi-metallic impurity band under the standard silicon conduction band. An increased doping density is found to increase the confinement and to lower the impurity band energies. The 1Γ, 2Γ, and Δ bands all depend linearly on the doping, but react to doping changes at a different rate, mainly due to the DOS effective mass difference. In-plane disorder is predicted to only weakly affect the VS and DOS of the quasi-metallic sheet. Doping disorder leads to an increased DOS modulation close to the Fermi energy, thus in turn leading to stronger conductance variation with device gating. Out-of-plane disorder shows a significant effect on the band edges and VS. VS is predicted to be reduced with increased out-of-plane disorder. The strong VS modulation may serve as a metrology tool to gauge vertical doping straggle in a well-controlled sequence of experiments. With extensive simulation results, we provide new information about the properties of these highly confined sheets that will guide experimentalists in understanding and validating electronic properties of Si:P δ-doped layers.

5. IN PROGRESS: VALLEY SPLITTING IN THE HIGHLY TUNABLE SILICON QUANTUM DOT IN THE SINGLE-ELECTRON REGIME

5.1 Introduction

Solid-state quantum dots (QDs) are considered as artificial atoms due to their ability to confine countable number of electrons. Nowadays these dots are of great interest in building quantum computing (QC) devices for storing, manipulating and reading spin information encoded in electrons or nuclei. Silicon, in particular, is the most promising solid-state material for fabricating QDs for QC devices due to its long spin coherence times [1, 3, 102]. From a fabrication point of view, the silicon metal-oxide-semiconductor (Si MOS) structure has been the dominant transistor design used by industry for over 40 years and is still the most preferred structure in ultra-scaled devices. Utilizing the most advanced technology will thus increase the chance of integrating uniform, scalable silicon qubit architectures. Despite the advantages of silicon, the perfect control of electron filling in Si QD is hampered by disorders at silicon-oxide interfaces and inherent six-fold valley degeneracy in the bulk silicon bandstructure. The valley degeneracy in silicon, in particular, is a major hurdle for QC devices that use electron spin as a qubit - spin qubits require the lowest two-spin states to be far away from other states to avoid spin decoherence [23, 103].

In an electrostatically defined Si QD, the six-fold degeneracy can be split to 2 (lower) and 4 (higher) states by creating confinement using top gates. The explanation of such a large separation between orbital states is illustrated in Fig. 5.1. The confinement strength of the dot is dominated by the electric field along the [001] direction. From the valley orientation with respect to [001], the six valleys behave differently due to differences in confinement masses. The two valleys along [001] have

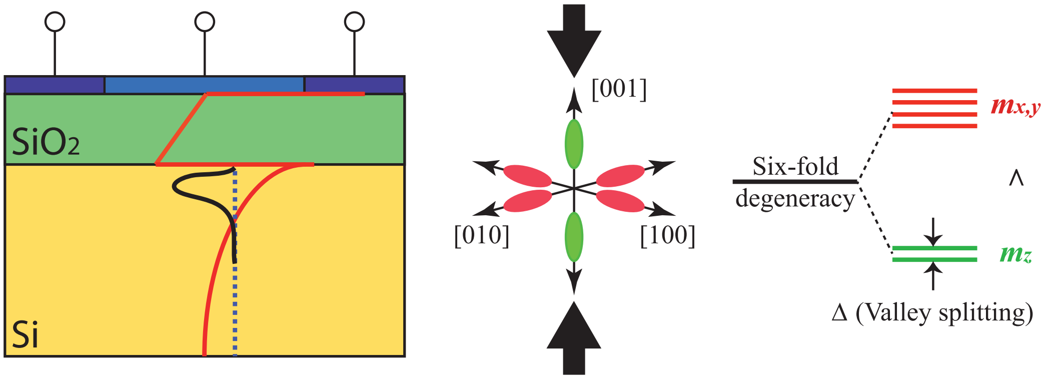


Fig. 5.1. Valley projection diagram in the presence of 3D confinement. The ground state of each valley is mainly influenced by the confinement along [001]. Since the confinement mass of the two valleys along [001] is larger than the other four in-plane valleys, the six-fold degeneracy is split into 2+4 and additional asymmetries in the confinement direction further split each degeneracy.

a larger confinement mass than the other four in-plane valleys. Since the quantization energy is inversely proportional to the effective mass, the two k_z valleys will always be lower in energy compared to the other four valleys. Therefore, the six fold degeneracy in silicon is split into 2-fold and 4-fold degenerate states and further asymmetries of the potential in the dot split each degeneracy up to a few meV's. The lowest degenerate states are again split mostly due to confinement effects in the dot. The difference between the two lowest valley states in the QD, or valley splitting (VS), is a critical quantity to determine whether the QD can be safely used for holding electron spin information. Due to its significance in QC applications, VS in silicon has long been studied [43, 104] and the theory of VS is compared to actual QDs and quantum wells (QWs) [46, 105, 106]. Particularly, an $sp^3d^5s^*$ atomistic tight-binding (TB) approach using `NEMO3D` is of interest since NEMO3D starts from the correct silicon bulk bandstructure, and automatically includes confinement, strain and atomistic disorder effects in the Hamiltonian [28–30]. Boykin et al. extensively investigated VS on various cases using TB, [41–45] According to the simulation results, the VS decays in an oscillatory manner with an atomic layer resolution as the width of QW increases.

A strong lateral confinement potential also turns out to increase the magnitude of VS. Moreover, recent simulation studies revealed that disorder effects influence VS as well. Kharche et al. studied VS of (100) and (111) 2° tilted strained silicon QWs and successfully explained that alloy and step disorder plays a significant role in such structures [45, 46] Jiang et al. examined the effect of interface disorder in SiGe-Si-SiGe QWs and showed that VS is a strong function of the width of the QW, the external field, the alloy composition and atomistic details near the interfaces of SiGe layers [107]. Despite of many simulation efforts aimed at explaining VS phenomena in confined nanostructures in silicon, the actual observation of VS varies from a few tens of μeVs to a few meVs. These variations attribute to the fact that VS is sensitive to the barrier height, QD geometry and bias conditions which vary from device to device. Therefore, exploring VS under a range of bias points through simulation may provide useful insight to QD design and help to determine proper gating scenarios for multiple qubit couplings in silicon QC devices.

Recently, a Si MOS QD with low interface disorder has been reported. The study showed that the QD can be electrostatically controlled to operate down from the few-electron state to the single-electron regime [26]. VS in the dot is extracted from magnetospectroscopy measurements and found to be around 100 (μeV) in the single electron regime, which is roughly one order of magnitude larger than the thermal energy at 100 mK ($k_BT = 8.7$ μeV). The achievement of fabricating a clean Si MOS QD operating in the single-electron regime clearly puts electrostatically defined QDs as one of the strong candidates for building scalable quantum computers.

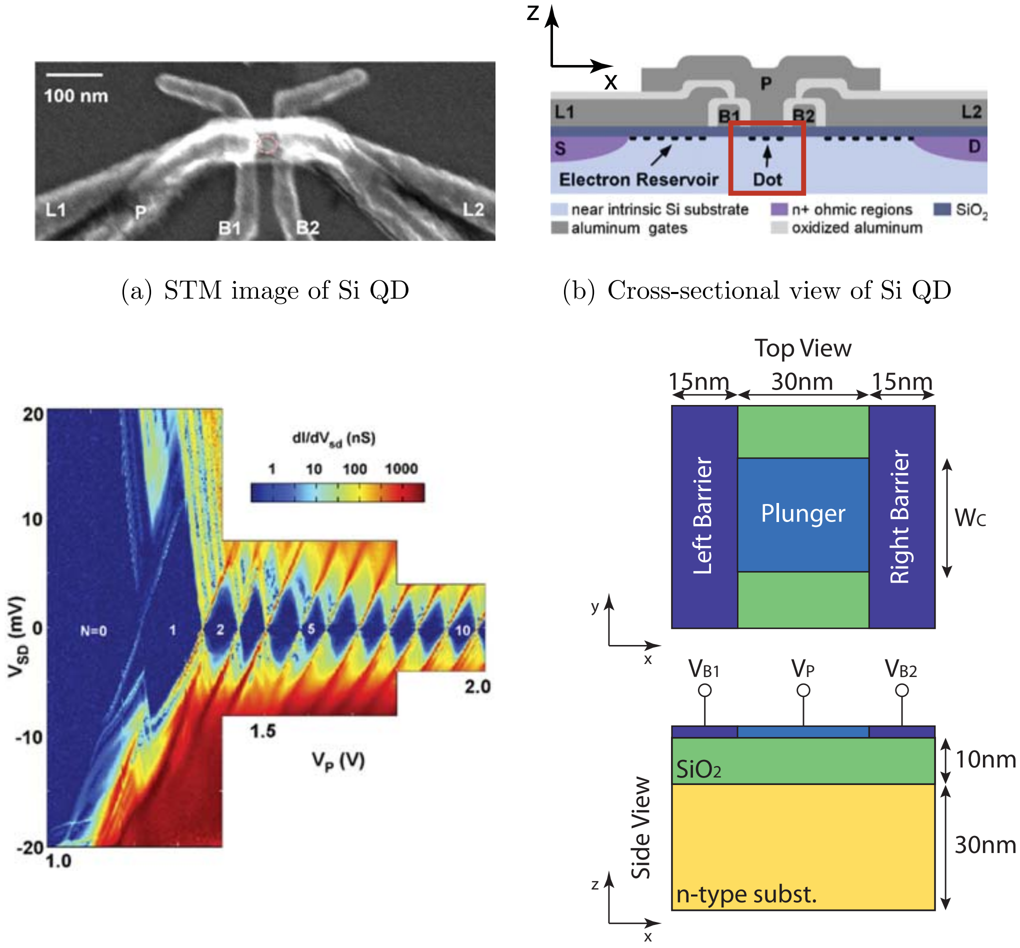
For a complete understanding of the dot under bias tuning, an elaborate simulation study is required to provide insight to the following questions that cannot be fully covered by experiments: What is the possible range of VS in the QD with proper biasing conditions? What are the possible factors influencing VS? Is it possible to map VS as a function of the external field and barrier height? Do the simulation results show reasonable agreement with experimental results? Can we suggest desirable device geometries and bias ranges?

This chapter is organized as follows to provide reasonable answers to the questions addressed above. Section 5.2 discusses the self-consistent procedure to extract VS in the single-electron regime. Section 5.3 shows and discusses the simulation results of the QD for different cases of biasing conditions. Section 5.4 summarizes and concludes the chapter.

5.2 Methodology

The Si MOS QD of interest is shown in Fig. 5.2(a) and 5.2(b). This Si MOS QD is controlled by five independent gates. The lead gates labeled L1 (L2) are used to create source (drain) reservoirs to supply (eject) electrons to (from) the QD. The reservoirs are then separated by the depletion region formed by the barrier gates (B1/B2) which serve as tunnel barriers. Finally, the quantum dot is created by applying a positive bias to the plunger gate (P). More importantly, the first diamond is observed to be open even at high V$_{DS}$ and before entering the next transition point, which indicates that the device is capable of operating in the single-electron regime. The simulation structure used for `NEMO3D-peta` is shown in Fig. 5.2(d). The domain is restricted to the central region of the device which includes the dot and part of the left and right barrier regions. The SiO$_2$ thickness is roughly 10 (nm). The oxidized aluminum surrounding each metal gate is ignored in the domain. As seen in the STM image (Fig. 5.2(a)), the plunger gate region is roughly 30×60 (nm^2), but to investigate geometry effects on the VS, the width of the plunger gate is adjusted to four different values (W_C = 30, 40, 50 and 60 nm). The overall electronic domain is fixed to 60×90×40 (nm^3) and is filled by 8 million atoms.

The simulation flow in the single-electron regime is described in Fig. 5.3. For the electronic structure calculation, an atomistic $sp^3d^5s^*$ tight-binding Hamiltonian without spin-orbit coupling is used to automatically account for the correct VS in Si QDs [46]. The eigenstates obtained from the Schrödinger equation (ϵ_i, $\Psi_i(\vec{r})$) are used to compute the total number of charge (n_Q) and charge profile ($n(\vec{r})$) in the



(a) STM image of Si QD

(b) Cross-sectional view of Si QD

(c) Coulomb diamond for Si QD

(d) `NEMO3D-peta` modeling

Fig. 5.2. (a), (b) Physical structure of Si MOS QD. Five top gates marked L1/L2/B1/B2/P are on top of the oxide interface to create electron reservoir (L1/L2), barrier between QD and the reservoir (B1/B2) and QD (P). (c) Charge stability diagram of the device in the few electron regime. (d) Simulation domain used in `NEMO3D-peta` for self-consistent simulation. Central region of the device (solid red box in (b)) is taken and it includes plunger gate (P) and part of barrier gates (B1/B2). The total simulation domain size is 60×90×40 (nm^3) (8 million atoms). Images adopted from Ref. [26] with author's permission.

dot. The Fermi-Dirac distribution function becomes a rectifying function around the measurement temperature of 100 (mK). Therefore, the total number of integrated charge is very close to an integer value. Any other single-electron charging effects are not taken into account because only one electron is assumed to be in the system. The charge profile is then fed into the FDM Poisson solver to obtain the potential profile of the system. Since the Poisson equation is a nonlinear equation as a function of the potential, the Newton-Raphson's method is used to calculate the potential and relaxed charge profile by starting from the quantum charge as an initial guess [57].[1] As a result of solving the Poisson equation, the potential profile ($V(\vec{r})$) as well as the total number of *corrected* charge (n_C) are obtained. In such low temperature simulations even small fluctuations in the potential profile may create relatively large deviations in the total charge from iteration to iteration, which in turn show oscillatory behavior in the convergence pattern. We therefore apply weaker convergence criteria for this self-consistent simulation due to the fact that it is a low-temperature simulation. The convergence criteria is set by comparing the difference between the number of quantum charge (n_Q) and relaxed charge (n_C). If the difference is within 30%, the self-consistent procedure is halted and proceeds to computing the final eigenstates and corresponding valley splitting. Although the convergence process is not as strict as in normal room-temperature simulations, VS turns out to be less sensitive to small potential fluctuations near true convergence state.

5.3 Results

In this section, the range of VS is measured for two biasing schemes, which are used for different measurement techniques. First, the barriers are kept symmetric ($V_{B1} = V_{B2}$), similar to the barrier setting performed in Ref. [26] and predict VS using magnetospectroscopy. Next, a large barrier is created by decreasing V_{B2} while fixing V_{B1} for electron shuttle experiments, which are used for investigating excited state spectra as well as VS while minimizing the interference by leads [108]. In both

[1] The details of the Newton-Raphson's method implemented in `NEMO3D-peta` are documented in Appendix B.

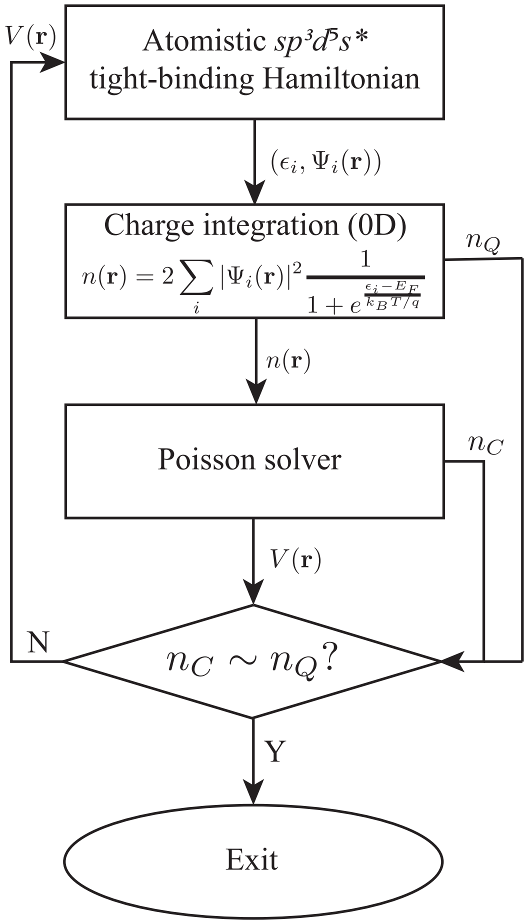

$V(\mathbf{r})$
Atomistic sp^3d^5s*
tight-binding Hamiltonian
$(\epsilon_i, \Psi_i(\mathbf{r}))$
Charge integration (0D)
$n(\mathbf{r}) = 2\sum_i |\Psi_i(\mathbf{r})|^2 \frac{1}{1+e^{\frac{\epsilon_i - E_F}{k_B T/q}}}$
n_Q
$n(\mathbf{r})$
Poisson solver
n_C
$V(\mathbf{r})$
N
$n_C \sim n_Q?$
Y
Exit

Fig. 5.3. Self-consistent simulation scheme at very low temperatures (<1K). Quantum mechanical charge profile is computed from an atomistic $sp^3d^5s^*$ tight-binding Hamiltonian. The charge profile ($n(\vec{r})$) and the total number of charge (n_Q) serve as inputs to the Poisson solver, which utilizes Newton-Raphson's iterative scheme. Subsequently, the potential profile ($V(\vec{r})$) with the corresponding number of charge (n_C) is extracted. Convergence is met when n_Q and n_C is differ only within 20 percent. Otherwise, update the external potential to the Schrödinger solver for the next iteration.

cases, the effects of the QD area and potential barrier height on VS are explored to provide insight for experimentalists.

5.3.1 Quantum dot with symmetric barrier height

Bias conditions in the single-electron regime for different QD sizes and resulting VS is summarized in Table 5.1. The VS values vary from 95 to 470 μeV, which is comparable to experimentally measured value (100 μeV). It is clearly shown that, as the barrier height increases (lower V_B), more V_P is required for the ground state to cross the Fermi level, which creates stronger confinement and subsequently larger VS. A smaller plunger gate area creates a smaller QD, whose geometry properties are expected to create stronger confinement. Analogous to the particle-in-a-box problem, the eigenstates within the dot are more quantized, requiring larger V_P to pull the states below the Fermi level. As a result, larger VS is expected as the plunger gate region is reduced, which indicates that smaller dots are desirable for maximizing VS. Using the simulation results in Table 5.1, the VS map as a function of V_P and V_B in the single-electron regime can be translated as shown in Fig. 5.4. This figure can be used for designing a quantum dot structure with desirable VS and biasing sensitivity or for predicting the effective area of the plunger gate pad by measuring the VS spectrum at varying barrier gate biases. High barriers are desirable for forcing larger VS but will also reduce the tunneling rate significantly and hamper electron injection. A trade-off between barrier height and plunger gate geometry should be carefully examined to determine the operation region of the quantum dot. The effect of VS can be viewed differently by investigating the relationship between the barrier height or electric field at the oxide interface and VS. Fig. 5.5(a) shows the effect of barrier height on VS for different QDs sizes. A lower barrier height is directly related to weaker confinement and generally reduces VS. At fixed barrier heights, smaller QDs exhibit larger VS due to the stronger confinement. The relationship between the electric field at the oxide interface and VS is shown in Fig. 5.5(b). Consistent with previous discussions, smaller QDs require stronger electric fields for more band bending near the oxide interface, resulting in larger VS. Regardless of the QD size, however, the graph shows a strong dependence between the electric field and VS.

Table 5.1
List of bias conditions to fill one electron in the dot with different sizes and corresponding valley splitting (VS).

Dot size (L×W)(nm^2)	No.	$V_{B1/B2}$ (V)	V_P (V)	VS (μeV)
30×30	1	0.79	1.29	471.78
	2	0.80	1.25	416.62
	3	0.83	1.21	324.85
	4	0.86	1.15	233.70
	5	0.89	1.09	169.84
30×40	1	0.79	1.23	390.05
	2	0.80	1.20	347.74
	3	0.83	1.14	279.58
	4	0.86	1.08	217.29
	5	0.89	1.04	141.83
30×50	1	0.79	1.18	315.37
	2	0.80	1.15	280.08
	3	0.83	1.10	228.44
	4	0.86	1.06	158.08
	5	0.89	1.02	95.91
30×60	1	0.79	1.14	285.74
	2	0.80	1.11	226.29
	3	0.83	1.08	186.92
	4	0.86	1.05	162.89
	5	0.89	1.02	96.85

This relationship indicates that VS is determined by the electric field (strength of the confinement) along the substrate direction, which is again determined by the plunger gate area and potential barrier height.

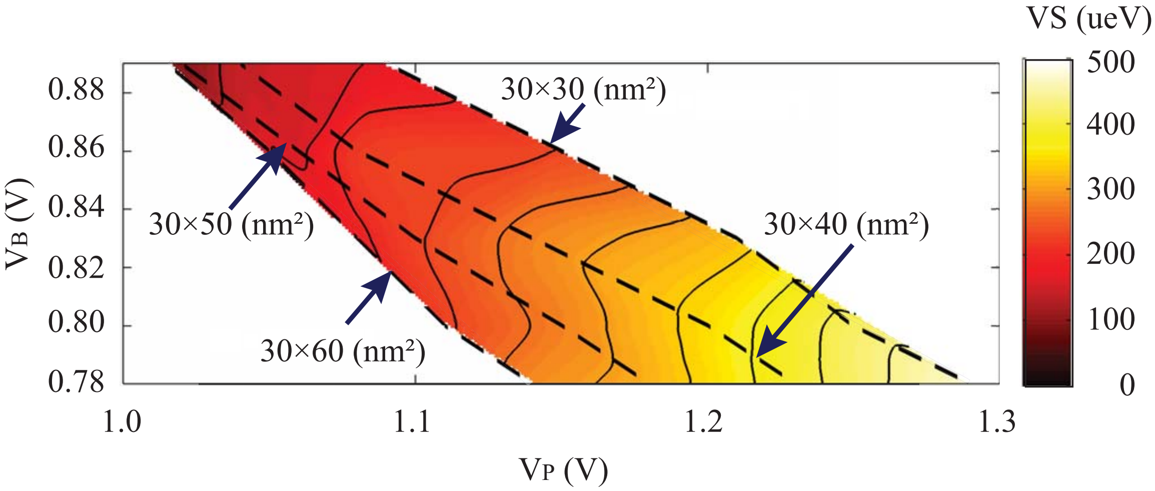
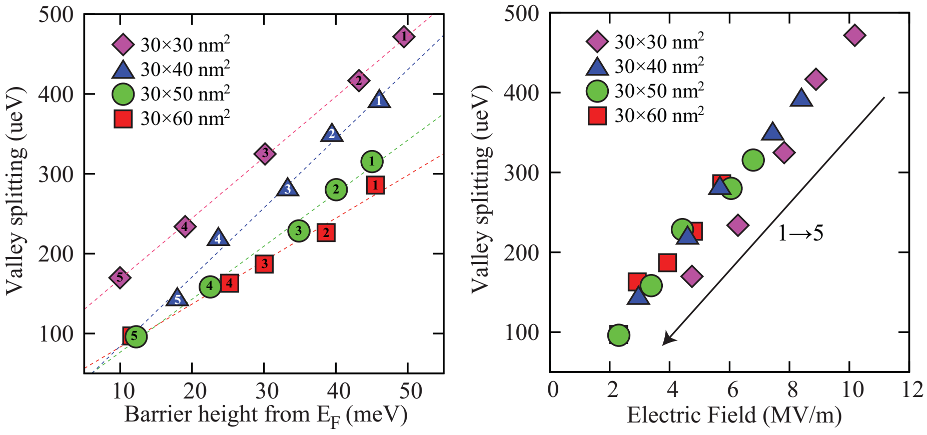


Fig. 5.4. Interpolated color map plot of VS distribution as a function of V_P and V_B in the single-electron state based on Table 5.1. Dashed line indicates (V_P, V_B) for different QD sizes as indicated in the plot. V_P in the figure is the bias required to fill in a single electron with given V_B and size of the dot.

5.3.2 Asymmetric barrier height manipulation for electron shuttling experiment

Another way of resolving discretized states in a quantum dot experimentally is to use charge-detection measurements with pulsed voltages [108]. In this experiment, electrons are not tunneling through the source and drain, but are rather tunneling in and out of the quantum dot through single barriers connected to the reservoir. These tunneling phenomena in turn modify the quantum point contact (QPC) current running from source to drain. From the conductance measurement the excited spectrum in the dot can then be extracted.

Similarly, the simulation can be carried out by using identical simulation domains as shown in Fig. 5.2(d) but in this case, only the right barrier (V_{B2}) is raised with a fixed V_{B1} = 0.89(V). The color map of the VS as a function of V_{B2}, V_P and the plunger gate area is shown in Fig. 5.6. The right potential barrier height is larger in this simulation since electrons are shuttled in and out only through the left barrier.



(a)

(b)

Fig. 5.5. (a) The effect of the barrier height on VS for different QD sizes. VS decreases as the barrier height is lowered due to weaker confinement in the dot. On the contrary, smaller QDs will generally have larger VS because of stronger confinement. (b) The electric field dependence of VS for different QD sizes. The electric field is related to the confinement along the substrate direction and smaller VS is expected for weaker electric fields.

VS varies from 200 to 1,600 (ueV) since the right potential barrier creates stronger confinement to the dot as compared to previous simulation; in turn a higher plunger gate bias is required to fill an electron into the dot. The size dependence of VS is also consistent with previous simulation results; smaller QDs require more plunger gate bias to confine electrons, which results in larger VS. VS as a function of the barrier height in 30 × 40 (nm^2) QD is shown in Fig. 5.3.2(a). As in the symmetric QD case, higher barriers induce a stronger plunger gate, which fills the dot with an electron, resulting in larger VS. However, VS is less sensitive to the barrier height when compared to the symmetric barrier QD since only V_{B2} is controlling the confinement in the dot. Consequently, a moderate increase in the VS with increased barrier height is expected as plotted in Fig. 5.3.2(a). On the other hand, VS shows a consistent trend as a function of the electric field at the oxide interface under plunger gate.

Table 5.2
List of bias conditions with different sizes and corresponding VS at single-electron regime on asymmetric barrier QD. V_{B1} is fixed to 0.79V.

Dot size (L × W) (nm^2)	No.	V_{B2} (V)	V_P (V)	VS (ueV)
30×30	1	0.79	1.18	296.71
	2	0.69	1.24	433.38
	3	0.59	1.34	519.30
	4	0.39	1.49	836.10
	5	0.19	1.64	1086.22
	6	-0.19	1.92	1610.89
30×40	1	0.79	1.13	221.72
	2	0.69	1.20	333.55
	3	0.59	1.26	434.96
	4	0.39	1.39	654.47
	5	0.19	1.52	877.68
	6	-0.19	1.71	1293.44
30×50	1	0.79	1.09	174.15
	2	0.69	1.15	265.70
	3	0.59	1.20	349.23
	4	0.39	1.31	532.65
	5	0.19	1.40	701.87
	6	-0.19	1.59	1042.76
30×60	1	0.79	1.07	143.85
	2	0.69	1.12	211.63
	3	0.59	1.17	298.30
	4	0.39	1.25	437.02
	5	0.19	1.33	574.85
	6	-0.19	1.50	858.18

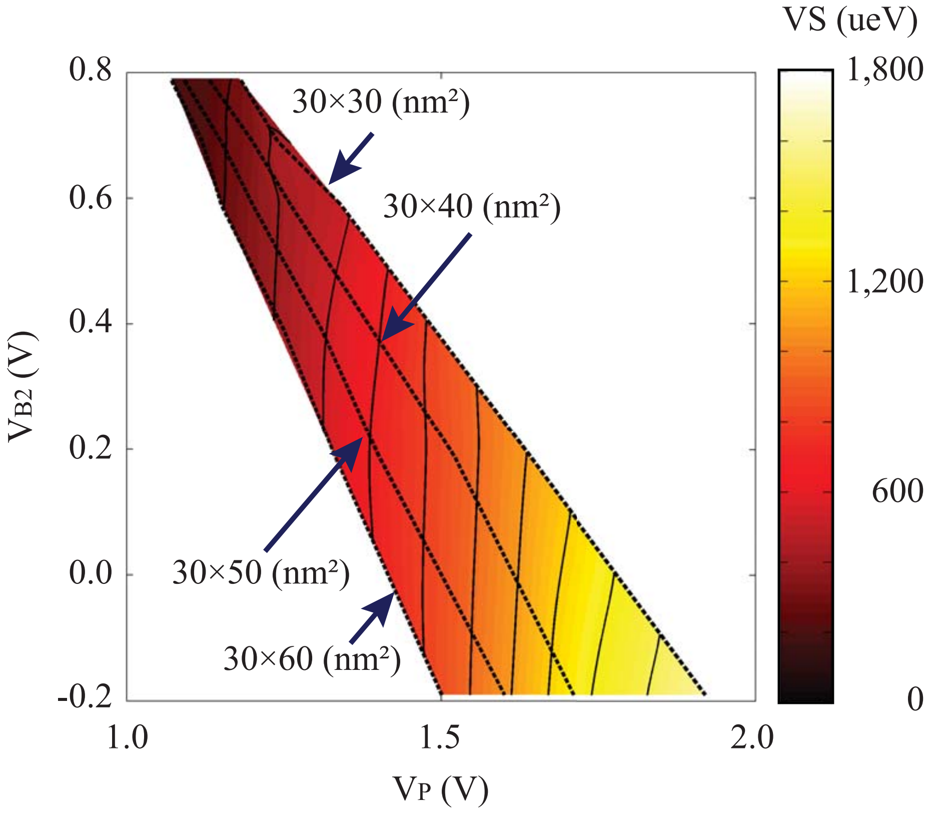


Fig. 5.6. Interpolated color map plot of VS distribution as a function of V_{B2} and V_P in the single-electron state for asymmetric gating cases. Similar to Fig. 5.4, the dashed line indicates (V_{B2},V_P) for different QD sizes. In all simulation cases, V_{B1} is fixed to 0.79 V.

To examine the first few excited confined states, the spectrum of the first four eigenvalues in the single electron state is shown in Fig. 5.8. The splitting between eigenstates is also consistent with the confinement strength of the dot; larger barrier height and smaller plunger gate area also split the first few excited states. As seen in the figure, the second orbital excited states (3rd and 4th states) are clearly separated from the first orbital states (1st and 2nd states, or valley states) up to 8 (meV), which is desirable for minimizing valley decoherence in qubit operations [23].

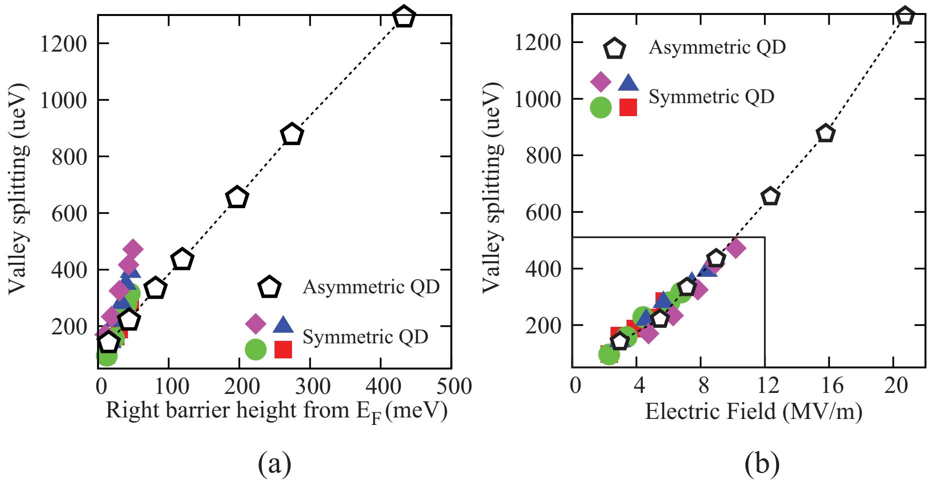


Fig. 5.7. (a) The effect of the barrier height on VS in the 30×40 (nm^2) QD. VS decreases as the barrier height is lowered due to weaker confinement in the dot, which is consistent with the symmetric barrier cases (colored dots from Fig. 5.5(a)). The slope is less than for the symmetric barrier QD since confinement is only controlled by the single gate. (b) Electric field dependence of VS. VS exhibits a stronger electric field dependence regardless of QD geometry - consistent with Fig. 5.5(b).

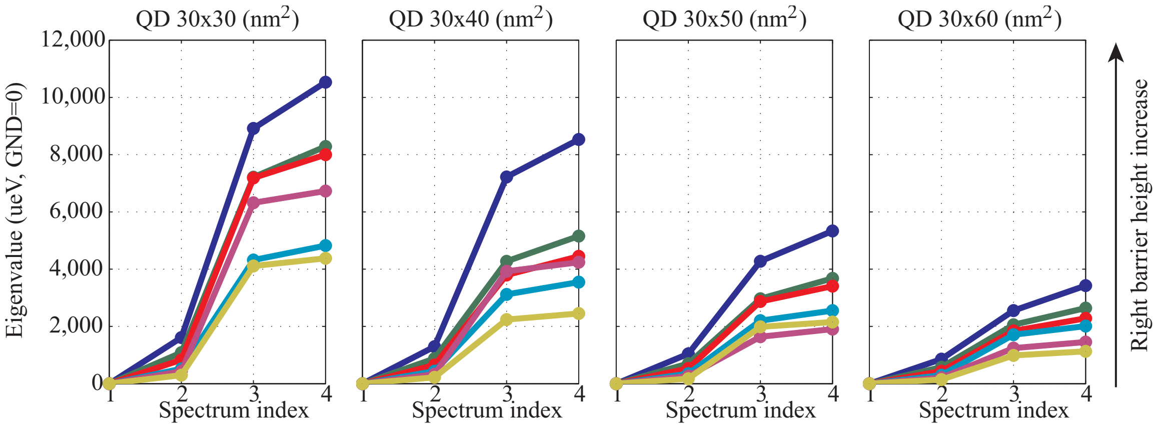


Fig. 5.8. First four eigenvalues plotted with respect to the ground state eigenvalue for the cases listed in Table 5.2. Typically, larger splitting in the eigenvalue spectrum is seen when decreasing the QD area, since smaller QD areas require stronger electric field under the plunger gate.

5.4 Conclusion

In conclusion, the eigenspectrum of a MOS based silicon QD in the single electron regime is investigated using `NEMO3D-peta` with charge-potential self-consistent capabilities. From the simulation studies, it is possible to manipulate the VS by controlling the barrier height and quantum dot size. VS is a strong function of the electric field along the plunger gate, but the plunger gate is determined by the barrier height and gate geometry. VS is maximized under strong confinement, which is created by higher barrier and smaller quantum dots. The trade-off between barrier height and VS should be taken into account in designing charge-based qubits. The splitting between the first two valley states and the next orbital states can be explained through the confinement mass along [001], which results in 2+4 separation. From the simulations using `NEMO3D-peta`, we expect to provide insight to experimentalists and expertise in Si MOS QD device development.

LIST OF REFERENCES

LIST OF REFERENCES

[1] B. Kane, “A silicon-based nuclear spin quantum computer,” *Nature*, vol. 393, p. 133, May 1998.

[2] A. Gusev and A. Bulanov, “High-purity silicon isotopes ^{28}Si, ^{29}Si and ^{30}Si,” *Inorganic Materials*, vol. 44, pp. 1395–1408, 2008. 10.1134/S0020168508130013.

[3] L. C. L. Hollenberg, A. S. Dzurak, C. Wellard, A. R. Hamilton, D. J. Reilly, G. J. Milburn, and R. G. Clark, “Charge-based quantum computing using single donors in semiconductors,” *Physical Review B*, vol. 69, p. 113301, Mar. 2004.

[4] D. Loss and D. P. DiVincenzo, “Quantum computation with quantum dots,” *Physical Review A*, vol. 57, pp. 120–126, Jan. 1998.

[5] R. Vrijen, E. Yablonovitch, K. Wang, H. W. Jiang, A. Balandin, V. Roychowdhury, T. Mor, and D. DiVincenzo, “Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures,” *Physical Review A*, vol. 62, p. 012306, Jun. 2000.

[6] R. de Sousa, J. D. Delgado, and S. Das Sarma, “Silicon quantum computation based on magnetic dipolar coupling,” *Physical Review A*, vol. 70, p. 052304, Nov. 2004.

[7] C. D. Hill, L. C. L. Hollenberg, A. G. Fowler, C. J. Wellard, A. D. Greentree, and H. S. Goan, “Global control and fast solid-state donor electron spin quantum computing,” *Physical Review B*, vol. 72, p. 045350, Jul. 2005.

[8] F. J. Rueß, W. Pok, K. E. J. Goh, A. R. Hamilton, and M. Y. Simmons, “Electronic properties of atomically abrupt tunnel junctions in silicon,” *Physical Review B*, vol. 75, p. 121303, Mar. 2007.

[9] B. Weber, S. Mahapatra, W. R. Clarke, R. H., L. S., G. Klimeck, L. C. L. Hollenberg, and M. Y. Simmons, “Quantum transport in atomic-scale silicon nanowires,” in *Silicon Nanoelectronics Workshop (SNW), 2010*, pp. 1 –2, Jun. 2010.

[10] A. Fuhrer, M. Fuechsle, T. C. G. Reusch, B. Weber, and M. Y. Simmons, “Atomic-scale, all epitaxial in-plane gated donor quantum dot in silicon,” *Nano Letters*, vol. 9, no. 2, pp. 707–710, 2009.

[11] M. Fuechsle, S. Mahapatra, F. A. Zwanenburg, M. Friesen, M. A. Eriksson, and M. Y. Simmons, “Spectroscopy of few-electron single-crystal silicon quantum dots,” *Nature Nanotechnology*, vol. 5, p. 502, May 2010.

[12] M. Y. Simmons, S. R. Schofield, J. L. O'Brien, N. J. Curson, L. Oberbeck, T. Hallam, and R. G. Clark, "Towards the atomic-scale fabrication of a silicon-based solid state quantum computer," *Surface Science*, vol. 532-535, pp. 1209–1218, 2003.

[13] F. J. Rueß, L. Oberbeck, M. Y. Simmons, K. E. J. Goh, A. R. Hamilton, T. Hallam, S. R. Schofield, N. J. Curson, and R. G. Clark, "Toward atomic-scale device fabrication in silicon using scanning probe microscopy," *Nano Letters*, vol. 4, no. 10, pp. 1969–1973, 2004.

[14] G. P. Lansbergen, R. Rahman, C. J. Wellard, I. Woo, J. Caro, N. Collaert, S. Biesemans, G. Klimeck, L. C. L. Hollenberg, and S. Rogge, "Gate-induced quantum-confinement transition of a single dopant atom in a silicon FinFET," *Nature Physics*, vol. 4, no. 8, pp. 656–661, 2008.

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

[16] S. Lee, H. Ryu, G. Klimeck, H. Campbell, S. Mahapatra, M. Y. Simmons, and L. C. L. Hollenberg, "Equilibrium bandstructure of a phosphorus δ-doped layer in silicon using a tight-binding approach," *IEEE Proceedings of NANO 2010*, 2010.

[17] H. Ryu, S. Lee, B. Weber, S. Mahapatra, M. Simmons, L. Hollenberg, and G. Klimeck, "Quantum transport in ultra-scaled phosphorous-doped silicon nanowires," in *Silicon Nanoelectronics Workshop (SNW)*, pp. 1 –2, Jun. 2010.

[18] W. H. Lim, F. A. Zwanenburg, H. Huebl, M. Möttönen, K. W. Chan, A. Morello, and A. S. Dzurak, "Observation of the single-electron regime in a highly tunable silicon quantum dot," *Applied Physics Letters*, vol. 95, no. 24, p. 242102, 2009.

[19] A. Morello, J. J. Pla, F. A. Zwanenburg, K. W. Chan, K. Y. Tan, H. Huebl, M. Mottonen, C. D. Nugroho, C. Yang, J. A. van Donkelaar, A. D. C. Alves, D. N. Jamieson, C. C. Escott, L. C. L. Hollenberg, R. G. Clark, and A. S. Dzurak, "Single-shot readout of an electron spin in silicon," *Nature*, vol. 467, pp. 687–691, Oct. 2010.

[20] C. H. Yang, W. H. Lim, F. A. Zwanenburg, and A. S. Dzurak, "Dynamically controlled charge sensing of a few-electron silicon quantum dot," *AIP Advances*, vol. 1, p. 042111, 2011.

[21] H. W. Liu, T. Fujisawa, Y. Ono, H. Inokawa, A. Fujiwara, K. Takashina, and Y. Hirayama, "Pauli-spin-blockade transport through a silicon double quantum dot," *Physical Review B*, vol. 77, p. 073310, Feb. 2008.

[22] H. Liu, T. Fujisawa, H. Inokawa, Y. Ono, A. Fujiwara, and Y. Hirayama, "A gate-defined silicon quantum dot molecule," *Applied Physics Letters*, vol. 92, no. 22, p. 222104, 2008.

[23] B. Koiller, X. Hu, and S. Das Sarma, "Exchange in silicon-based quantum computer architecture," *Physical Review Letters*, vol. 88, p. 027903, Dec. 2001.

[24] S. J. Angus, A. J. Ferguson, A. S. Dzurak, and R. G. Clark, "Gate-defined quantum dots in intrinsic silicon," *Nano Letters*, vol. 7, no. 7, pp. 2051–2055, 2007.

[25] S. J. Angus, A. J. Ferguson, A. S. Dzurak, and R. G. Clark, "A silicon radio-frequency single electron transistor," *Applied Physics Letters*, vol. 92, no. 11, p. 112103, 2008.

[26] W. H. Lim, C. H. Yang, F. A. Zwanenburg, and A. S. Dzurak, "Spin filling of valleyorbit states in a silicon quantum dot," *Nanotechnology*, vol. 22, no. 33, p. 335704, 2011.

[27] W. H. Lim, H. Huebl, L. H. W. van Beveren, S. Rubanov, P. G. Spizzirri, S. J. Angus, R. G. Clark, and A. S. Dzurak, "Electrostatically defined few-electron double quantum dot in silicon," *Applied Physics Letters*, vol. 94, no. 17, p. 173502, 2009.

[28] G. Klimeck, F. Oyafuso, T. B. Bokyin, R. C. Bowen, and P. von Allmen, "Development of a nanoelectronic $3-\mathrm{D}$ (NEMO $3-\mathrm{D}$) simulator for multimillion atom simulations and its application to alloyed quantum dots," *Computer Modeling in Engineering and Science*, vol. 3, no. 5, pp. 601–642, 2002.

[29] G. Klimeck, S. Ahmed, H. Bae, N. Kharche, S. Clark, B. Haley, S. Lee, M. Naumov, H. Ryu, F. Saied, M. Prada, M. Korkusinski, T. Boykin, and R. Rahman, "Atomistic simulation of realistically sized nanodevices using NEMO $3-\mathrm{D}$ part I: Models and benchmarks," *IEEE Transactions on Electron Devices*, vol. 54, pp. 2079–2089, Sep. 2007.

[30] G. Klimeck, S. Ahmed, N. Kharche, M. Korkusinski, M. Usman, M. Prada, and T. Boykin, "Atomistic simulation of realistically sized nanodevices using NEMO $3-\mathrm{D}$ part II: Applications," *IEEE Transactions on Electron Devices*, vol. 54, pp. 2090–2099, Sep. 2007.

[31] G. Tettamanzi, A. Paul, G. Lansbergen, J. Verduijn, S. Lee, N. Collaert, S. Biesemans, G. Klimeck, and S. Rogge, "Thermionic emission as a tool to study transport in undoped $n-$ FinFETs," *IEEE Electron Device Letters*, vol. 31, pp. 150–152, Feb. 2010.

[32] G. Tettamanzi, A. Paul, S. Lee, S. Mehrotra, N. Collaert, S. Biesemans, G. Klimeck, and S. Rogge, "Interface trap density metrology of state-of-the-art undoped Si $n-$FinFETs," *IEEE Electron Device Letters*, vol. 32, pp. 440–442, Apr. 2011.

[33] A. Paul, G. C. Tettamanzi, S. Lee, S. Mehrotra, N. Colleart, S. Biesemans, S. Rogge, and G. Klimeck, "Interface trap density metrology from sub-threshold transport in highly scaled undoped Si $n-$ FinFETs," *arXiv:1102.0140*, Feb. 2011.

[34] http://nanohub.org

[35] A. G. Akkala, S. Steiger, J. M. D. Sellier, S. Lee, M. Povolotskyi, T. C. Kubis, H. Park, S. Agarwal, and G. Klimeck, "1d heterostructure tool," *https://nanohub.org/resources/5203*, Sep. 2008.

[36] J. M. Jancu, R. Scholz, F. Beltram, and F. Bassani, "Empirical $sp^3d^5s^*$ tight-binding calculation for cubic semiconductors: General method and material parameters," *Physical Review B*, vol. 57, pp. 6493–6507, Mar. 1998.

[37] 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. 2-3, pp. 77–88, 2000.

[38] T. B. Boykin, G. Klimeck, and F. Oyafuso, “Valence band effective-mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization,” *Physical Review B*, vol. 69, p. 115201, Mar. 2004.

[39] R. Rahman, G. P. Lansbergen, S. H. Park, J. Verduijn, G. Klimeck, S. Rogge, and L. C. L. Hollenberg, “Orbital stark effect and quantum confinement transition of donors in silicon,” *Physical Review B*, vol. 80, p. 165314, Oct. 2009.

[40] R. Rahman, C. J. Wellard, F. R. Bradbury, M. Prada, J. H. Cole, G. Klimeck, and L. C. L. Hollenberg, “High precision quantum control of single donor spins in silicon,” *Physical Review Letters*, vol. 99, p. 036403, Jul. 2007.

[41] T. B. Boykin, G. Klimeck, P. von Allmen, S. Lee, and F. Oyafuso, “Valley splitting in $V-$shaped quantum wells,” *Journal of Applied Physics*, vol. 97, no. 11, p. 113702, 2005.

[42] T. B. Boykin, G. Klimeck, M. A. Eriksson, M. Friesen, S. N. Coppersmith, P. von Allmen, F. Oyafuso, and S. Lee, “Valley splitting in strained silicon quantum wells,” *Applied Physics Letters*, vol. 84, no. 1, pp. 115–117, 2004.

[43] T. B. Boykin, G. Klimeck, M. Friesen, S. N. Coppersmith, P. von Allmen, F. Oyafuso, and S. Lee, “Valley splitting in low-density quantum-confined heterostructures studied using tight-binding models,” *Physical Review B*, vol. 70, p. 165325, Oct. 2004.

[44] T. B. Boykin, N. Kharche, and G. Klimeck, “Valley splitting in finite barrier quantum wells,” *Physical Review B*, vol. 77, p. 245320, Jun. 2008.

[45] N. Kharche, S. Kim, T. B. Boykin, and G. Klimeck, “Valley degeneracies in (111) silicon quantum wells,” *Applied Physics Letters*, vol. 94, no. 4, p. 042101, 2009.

[46] N. Kharche, M. Prada, T. B. Boykin, and G. Klimeck, “Valley splitting in strained silicon quantum wells modeled with 2° miscuts, step disorder, and alloy disorder,” *Applied Physics Letters*, vol. 90, no. 9, p. 092109, 2007.

[47] S. Lee, H. Ryu, Z. Jiang, and G. Klimeck, “Million atom electronic structure and device calculations on peta-scale computers,” in *13th International Workshop on Computational Electronics, 2009 (IWCE ’09)*, May 2009.

[48] M. Naumov, S. Lee, B. Haley, H. Bae, S. Clark, R. Rahman, H. Ryu, F. Saied, and G. Klimeck, “Eigenvalue solvers for atomistic simulations of electronic structures with $NEMO-3D$,” *Journal of Computational Electronics*, vol. 7, pp. 297–300, 2008. 10.1007/s10825-008-0223-5.

[49] H. Bae, B. Haley, R. Hoon, G. Klimeck, S. Lee, and M. Luisier, “A nanoelectronics simulator for petascale computing: From NEMO to OMEN,” in *TeraGrid 2008*, pp. 1–2, Jun. 2008.

[50] S. Ahmed, N. Kharche, R. Rahman, M. Usman, S. Lee, H. Ryu, H. Bae, S. Clark, B. Haley, M. Naumov, F. Saied, M. Korkusinski, R. Kennel, M. McLennan, T. B. Boykin, and G. Klimeck, *Multimillion Atom Simulations with* NEMO3D. Springer New York, 2009.

[51] H. Ryu, S. Lee, and G. Klimeck, "A study of temperature-dependent properties of n-type δ-doped Si band-structures in equilibrium," in *13th International Workshop on Computational Electronics, 2009 (IWCE '09)*, May 2009.

[52] T. B. Boykin, G. Klimeck, R. C. Bowen, and F. Oyafuso, "Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory," *Physical Review B*, vol. 66, p. 125207, Sep. 2002.

[53] T. B. Boykin, N. Kharche, and G. Klimeck, "Brillouin-zone unfolding of perfect supercells having nonequivalent primitive cells illustrated with a Si/Ge tight-binding parameterization," *Physical Review B*, vol. 76, p. 035310, Jul. 2007.

[54] M. Usman, H. Ryu, I. Woo, D. Ebert, and G. Klimeck, "Moving toward nano-tcad through multimillion-atom quantum-dot simulations matching experimental data," *IEEE Transactions on Nanotechnology*, vol. 8, pp. 330–344, May 2009.

[55] G. H. Golub and C. F. Van Loan, *Matrix Computations*. The Johns Hopkins University Press, 3rd ed., Oct. 1996.

[56] K. Maschhoff and D. Sorensen, "P_ARPACK, an efficient portable large scale eigenvalue package for distributed memory parallel architectures," in *Applied Parallel Computing Industrial Computation and Optimization* (J. Wasniewski, J. Dongarra, K. Madsen, and D. Olesen, eds.), vol. 1184 of *Lecture Notes in Computer Science*, pp. 478–486, Springer Berlin / Heidelberg, 1996.

[57] R. Jabr, M. Hamad, and Y. Mohanna, "Newton-Raphson solution of Poisson's equation in a *pn* diode," *International Journal of Electrical Engineering Education*, vol. 44, no. 1, pp. 23–33, 2007.

[58] http://top500.org

[59] http://www.nccs.gov/computing resources/jaguar

[60] http://www.nics.tennessee.edu/computing resources/kraken

[61] http://www.tacc.utexas.edu/resources/hpc

[62] http://www.sdsc.edu/us/resources/trestles

[63] http://www.rcac.purdue.edu

[64] http://www.cs.sandia.gov/CRF/aztec1.html

[65] G. L. G. Sleijpen and H. A. V. d. Vorst, "A Jacobi $-$ Davidson iteration method for linear eigenvalue problems," *SIAM Review*, vol. 42, no. 2, pp. pp. 267–293, 2000.

[66] J. J. Sakurai, *Modern Quantum Mechanics*. Addison Wesley, rev sub ed., Sep. 1993.

[67] L. Landau and E. Lifshitz, *Quantum mechanics: non-relativistic theory.* Teoreticheskaia fizika (Izd. 3-e) (Landau, L. D, 1908-1968), Butterworth-Heinemann, 1981.

[68] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide.* Society for Industrial and Applied Mathematics, 2000.

[69] K. Leung and K. B. Whaley, "Electron-hole interactions in silicon nanocrystals," *Physical Review B*, vol. 56, pp. 7455–7468, Sep. 1997.

[70] http://www.mpi forum.org

[71] http://www.netlib.org

[72] F. J. Rueß, B. Weber, K. E. J. Goh, O. Klochan, A. R. Hamilton, and M. Y. Simmons, "One-dimensional conduction properties of highly phosphorus-doped planar nanowires patterned by scanning probe microscopy," *Physical Review B*, vol. 76, p. 085403, Aug. 2007.

[73] L. C. L. Hollenberg, A. D. Greentree, A. G. Fowler, and C. J. Wellard, "Two-dimensional architectures for donor-based quantum computing," *Physical Review B*, vol. 74, p. 045311, Jul. 2006.

[74] L. Oberbeck, N. J. Curson, M. Y. Simmons, R. Brenner, A. R. Hamilton, S. R. Schofield, and R. G. Clark, "Encapsulation of phosphorus dopants in silicon for the fabrication of a quantum computer," *Applied Physics Letters*, vol. 81, no. 17, pp. 3197–3199, 2002.

[75] K. E. J. Goh, L. Oberbeck, M. Y. Simmons, A. R. Hamilton, and R. G. Clark, "Effect of encapsulation temperature on Si : P delta-doped layers," *Applied Physics Letters*, vol. 85, no. 21, pp. 4953–4955, 2004.

[76] K. E. J. Goh, Y. Augarten, L. Oberbeck, and M. Y. Simmons, "Enhancing electron transport in Si : P delta-doped devices by rapid thermal anneal," *Applied Physics Letters*, vol. 93, no. 14, p. 142105, 2008.

[77] K. E. J. Goh and M. Y. Simmons, "Impact of Si growth rate on coherent electron transport in Si : P delta-doped devices," *Applied Physics Letters*, vol. 95, no. 14, p. 142104, 2009.

[78] H. F. Wilson, O. Warschkow, N. A. Marks, S. R. Schofield, N. J. Curson, P. V. Smith, M. W. Radny, D. R. McKenzie, and M. Y. Simmons, "Phosphine dissociation on the Si(001) surface," *Physical Review Letters*, vol. 93, p. 226102, Nov. 2004.

[79] H. F. Wilson, O. Warschkow, N. A. Marks, N. J. Curson, S. R. Schofield, T. C. G. Reusch, M. W. Radny, P. V. Smith, D. R. McKenzie, and M. Y. Simmons, "Thermal dissociation and desorption of PH_3 on Si(001): A reinterpretation of spectroscopic data," *Physical Review B*, vol. 74, p. 195310, Nov. 2006.

[80] S. R. Schofield, N. J. Curson, M. Y. Simmons, F. J. Rueß, T. Hallam, L. Oberbeck, and R. G. Clark, "Atomically precise placement of single dopants in Si," *Physical Review Letters*, vol. 91, p. 136104, Sep. 2003.

[81] G. Tempel, F. Koch, H. P. Zeindl, and I. Eisele, “Electronic states and transport properties of an n-type δ-function doping layer in p-type Si,” *Journal of Physics Colloques*, vol. 48, no. C5, pp. C5–259–C5–262, 1987.

[82] I. Eisele, “Delta-type doping profiles in silicon,” *Applied Surface Science*, vol. 36, no. 1-4, pp. 39 – 51, 1989.

[83] I. Eisele, “Quantized states in delta-doped Si layers,” *Superlattices and Microstructures*, vol. 6, no. 1, pp. 123 – 128, 1989.

[84] L. M. Gaggero-Sager, M. E. Mora-Ramos, and D. A. Contreras-Solorio, “Thomas-fermi approximation in p-type δ-doped quantum wells of GaAs and Si,” *Phys. Rev. B*, vol. 57, pp. 6286–6289, Mar. 1998.

[85] L. Gaggero-Sager, S. Vlaev, and G. Monsivais, “A tight binding calculation of δ-doped quantum wells in Si,” *Comp. Mat. Sci.*, vol. 20, no. 2, pp. 177 – 180, 2001.

[86] A. L. Rosa, L. M. R. Scolfaro, R. Enderlein, G. M. Sipahi, and J. R. Leite, “p-type δ-doping quantum wells and superlattices in Si: Self-consistent hole potentials and band structures,” *Phys. Rev. B*, vol. 58, pp. 15675–15687, Dec. 1998.

[87] L. M. R. Scolfaro, D. Beliaev, R. Enderlein, and J. R. Leite, “Electronic structure of n-type δ-doping multiple layers and superlattices in silicon,” *Phys. Rev. B*, vol. 50, pp. 8699–8705, Sep. 1994.

[88] G. Qian, Y.-C. Chang, and J. R. Tucker, “Theoretical study of phosphorous δ-doped silicon for quantum computing,” *Physical Review B*, vol. 71, no. 4, p. 045309, 2005.

[89] X. Cartoixà and Y. C. Chang, “Fermi-level oscillation in n-type δ-doped Si : A self-consistent tight-binding approach,” *Physical Review B*, vol. 72, p. 125330, Sep. 2005.

[90] D. J. Carter, O. Warschkow, N. A. Marks, and D. R. McKenzie, “Electronic structure models of phosphorus δ-doped silicon,” *Physical Review B*, vol. 79, p. 033204, Jan. 2009.

[91] D. J. Carter, N. A. Marks, O. Warschkow, and D. R. McKenzie, “Phosphorus δ-doped silicon: mixed-atom pseudopotentials and dopant disorder effects,” *Nanotechnology*, vol. 22, no. 6, p. 065701, 2011.

[92] R. Rahman, S. H. Park, J. H. Cole, A. D. Greentree, R. P. Muller, G. Klimeck, and L. C. L. Hollenberg, “Atomistic simulations of adiabatic coherent electron transport in triple donor systems,” *Physical Review B*, vol. 80, p. 035302, Jul. 2009.

[93] R. Rahman, S. H. Park, T. B. Boykin, G. Klimeck, S. Rogge, and L. C. L. Hollenberg, “Gate-induced g-factor control and dimensional transition for donors in multivalley semiconductors,” *Physical Review B*, vol. 80, p. 155301, Oct. 2009.

[94] N. Kharche, M. Luisier, T. Boykin, and G. Klimeck, “Electronic structure and transmission characteristics of SiGe nanowires,” *Journal of Computational Electronics*, vol. 7, pp. 350–354, 2008.

[95] E. Gawlinski, T. Dzurak, and R. A. Tahir-Kheli, "Direct and exchange-correlation carrier interaction effects in a resonant tunnel diode," *Journal of Applied Physics*, vol. 72, no. 8, pp. 3562–3569, 1992.

[96] R. G. Parr and W. Yang. Oxford University Press, 1994.

[97] E. Wigner, "On the interaction of electrons in metals," *Physical Review*, vol. 46, pp. 1002–1011, Dec. 1934.

[98] T. Frauenheim, G. Seifert, M. Elstner, Z. Hajnal, G. Jungnickel, D. Porezag, S. Suhai, and R. Scholz, *A self-consistent charge density-functional based tight-binding method for predictive materials simulations in physics, chemistry and biology*, pp. 41–62. Wiley-VCH Verlag GmbH & Co. KGaA, 2005.

[99] T. B. Boykin, N. Kharche, G. Klimeck, and M. Korkusinski, "Approximate bandstructures of semiconductor alloys from tight-binding supercell calculations," *Journal of Physics: Condensed Matter*, vol. 19, no. 3, p. 036203, 2007.

[100] D. C. Tsui, "Observation of surface bound state and two-dimensional energy band by electron tunneling," *Physical Review Letters*, vol. 24, pp. 303–306, Feb. 1970.

[101] M. Zachau, F. Koch, K. Ploog, P. Roentgen, and H. Beneking, "Schottky-barrier tunneling spectroscopy for the electronic subbands of a δ-doping layer," *Solid State Communications*, vol. 59, no. 8, pp. 591 – 594, 1986.

[102] C. Tahan, M. Friesen, and R. Joynt, "Decoherence of electron spin qubits in Si-based quantum computers," *Physical Review B*, vol. 66, p. 035314, Jul. 2002.

[103] A. L. Saraiva, M. J. Calderon, X. Hu, S. D. Sarma, and B. Koiller, "Intervalley coupling for silicon electronic spin qubits: Insights from an effective mass study," *arXiv:1006.3338*, Jun. 2010.

[104] M. Friesen and S. N. Coppersmith, "Theory of valley-orbit coupling in a Si/SiGe quantum dot," *Physical Review B*, vol. 81, p. 115324, Mar. 2010.

[105] S. Goswami, K. A. Slinker, M. Friesen, L. M. McGuire, J. L. Truitt, C. Tahan, L. J. Klein, J. O. Chu, P. M. Mooney, D. W. van der Weide, R. Joynt, S. N. Coppersmith, and M. A. Eriksson, "Controllable valley splitting in silicon quantum devices," *Natures Physics*, vol. 3, pp. 41–45, Jan. 2007.

[106] R. Rahman, J. Verduijn, N. Kharche, G. P. Lansbergen, G. Klimeck, L. C. L. Hollenberg, and S. Rogge, "Engineered valley-orbit splittings in quantum-confined nanostructures in silicon," *Physical Review B*, vol. 83, p. 195323, May 2011.

[107] Z. Jiang, N. Kharche, T. Boykin, and G. Klimeck, "Effects of interface disorder on valley splitting in SiGe/Si/SiGe quantum wells," *arXiv:1110.4097*, Oct. 2011.

[108] J. M. Elzerman, R. Hanson, L. H. W. van Beveren, L. M. K. Vandersypen, and L. P. Kouwenhoven, "Excited-state spectroscopy on a nearly closed quantum dot via charge detection," *Applied Physics Letters*, vol. 84, no. 23, pp. 4617–4619, 2004.

[109] J. C. Slater and G. F. Koster, "Simplified lcao method for the periodic potential problem," *Physical Review*, vol. 94, pp. 1498–1524, Jun. 1954.

[110] W. A. Harrison, *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond*. Dover Publications, 1989.

[111] S. Froyen and W. A. Harrison, "Elementary prediction of linear combination of atomic orbitals matrix elements," *Physical Review B*, vol. 20, pp. 2420–2422, Sep. 1979.

[112] P. Vogl, H. P. Hjalmarson, and J. D. Dow, "A semi-empirical tight-binding theory of the electronic structure of semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 44, no. 5, pp. 365 – 378, 1983.

[113] G. Klimeck, R. C. Bowen, T. B. Boykin, and T. A. Cwik, "sp^3s^* tight-binding parameters for transport simulations in compound semiconductors," *Superlattices and Microstructures*, vol. 27, no. 5-6, pp. 519 – 524, 2000.

[114] M. Dresselhaus, G. Dresselhaus, and A. Jorio, *Group theory: application to the physics of condensed matter*. Springer-Verlag, 2008.

[115] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C (2nd ed.): the art of scientific computing*. New York, NY, USA: Cambridge University Press, 1992.

[116] R. F. Pierret, *Modular Series on Solid State Devices - Advanced Semiconductor Fundamentals*. Addison-Wesley Publishing Company, 1987.

[117] R.-H. Xie, G. W. Bryant, S. Lee, and W. Jaskólski, "Electron-hole correlations and optical excitonic gaps in quantum-dot quantum wells: Tight-binding approach," *Physical Review B*, vol. 65, p. 235306, May 2002.

[118] S. Lee, L. Jönsson, J. W. Wilkins, G. W. Bryant, and G. Klimeck, "Electron-hole correlations in semiconductor quantum dots with tight-binding wave functions," *Physical Review B*, vol. 63, p. 195318, Apr. 2001.

[119] K. Leung, S. Pokrant, and K. B. Whaley, "Exciton fine structure in cdse nanoclusters," *Physical Review B*, vol. 57, pp. 12291–12301, May 1998.

[120] S. Lee, J. Kim, L. Jönsson, J. W. Wilkins, G. W. Bryant, and G. Klimeck, "Many-body levels of optically excited and multiply charged inas nanocrystals modeled by semiempirical tight binding," *Physical Review B*, vol. 66, p. 235307, Dec. 2002.

[121] A. Szabó and N. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*. Dover Publications, 1996.

[122] J. Vleck, *The theory of electric and magnetic susceptibilities*. International series of monographs on physics, Oxford University Press, 1952.

[123] D. Mattis, *The theory of magnetism: an introduction to the study of cooperative phenomena*. Harper's physics series, Harper & Row, 1965.

[124] J. R. Petta, A. C. Johnson, J. M. Taylor, E. A. Laird, A. Yacoby, M. D. Lukin, C. M. Marcus, M. P. Hanson, and A. C. Gossard, "Coherent manipulation of coupled electron spins in semiconductor quantum dots," *Science*, vol. 309, no. 5744, pp. 2180–2184, 2005.

[125] J. C. Slater, "Atomic shielding constants," *Physical Review*, vol. 36, pp. 57–64, Jul. 1930.

[126] K. Ohno, "Some remarks on the pariser-parr-pople method," *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, vol. 2, pp. 219–227, 1964. 10.1007/BF00528281.

[127] A. Franceschetti, H. Fu, L. W. Wang, and A. Zunger, "Many-body pseudopotential theory of excitons in inp and cdse quantum dots," *Physical Review B*, vol. 60, pp. 1819–1829, Jul. 1999.

APPENDICES

A. CONSTRUCTION OF ATOMISTIC TIGHT-BINDING HAMILTONIAN

This chapter describes the Hamiltonian constructor in an atomistic tight-binding basis is implemented in `NEMO3D-peta`. Actual construction is performed in `material.cpp` with published parameter sets stored in `params_tb_[bandmodel].par`. Calculation of each matrix element is based on the two-center integral approximation by Slater and Koster (Table 1 in Ref. [109].

A.1 Introduction

In the tight-binding model for describing electronic structures, localized atomic orbitals similar to what we see from analytical solutions of Schrödinger equation for single atom are used as basis functions to reasonably describe the valence electron distribution. From analytical solutions for atomic orbitals

$$\psi_{n,l,m}(r,\theta,\phi) = R_{n,l}(r)Y_{l,m}(\theta,\phi) \tag{A.1}$$

where n is principal quantum number, l is angular momentum quantum number $(0,\ 1,\ \cdots,\ n-1)$ and m is magnetic quantum number $(-l,\ -l+1,\ \cdots,\ l-1,\ l)$ and computed orbitals are orthonormal. Graphical form of atomic orbitals are shown in Fig. A.1.

Depending on the angular momentum number l, it is possible to categorize the orbitals in $s-$, $p-$, $d-$ and $f-$ like functions. The term *like* is used since in tight-binding, a definite form of orbitals are not taken into account. Instead, the hopping matrix elements between on-site and its neighbors are important and those elements are determined *empirically* by looking at the crystal symmetry and adjusting material

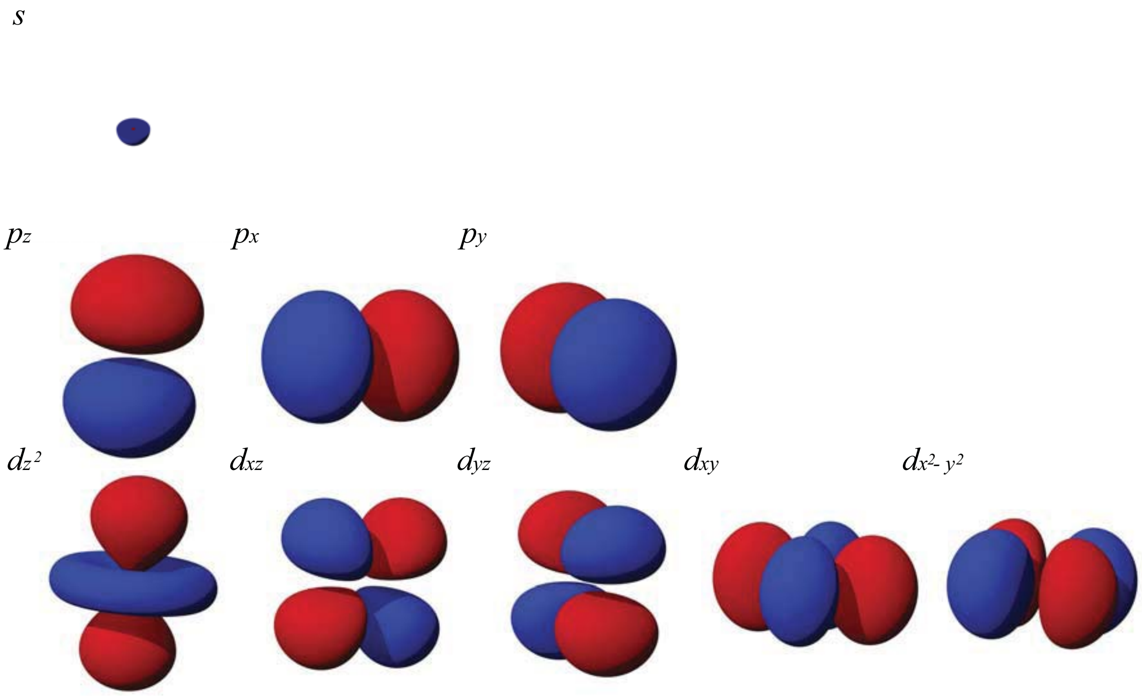


Fig. A.1. Graphical representation of first few s, p and d atomic orbitals adopted from `http://en.wikipedia.org/wiki/Atomic_orbital`.

dependent parameters. Now, the Hamiltonian becomes a block matrix $n \times n$, where n is number of orbitals per atom.

In most semiconductor materials of interest, the valence electrons occupy s, p and d orbital states. For example, silicon has electron configuration of $1s^2 2s^2 2p^6 \mathbf{3s^2 3p^2}$ and indium has $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 \mathbf{4d^{10} 5s^2 5p^1}$ in valence states. Therefore, the interatomic coupling term between s, p and d orbitals will be discussed in particular.

Let's assume two atoms are aligned along x axis with distance d and consider s and p orbitals constitute the basis of each atom.[1] Regardless of the details of each orbital there will be six different types of bonding $s - s$, $s - px$, $s - p(y, z)$, $px - px$, $py(z) - py(z)$ and $px - p(y, z)$. $s - p(y, z)$ and $px - p(y, z)$ bonding is neglected since it

[1] Note that following procedure not only applicable to s and p type orbital bonding but also d type orbital interactions. However, including d orbitals in the basis set requires more complicated analysis.

has least significance in creating bonds. The types of $s-p$ bonding between two atoms are defined and shown in Fig. A.2. The orbital bonding aligned along x is defined as σ bonding and the bonding between orbitals perpendicular to the directional axis is called π bonding.

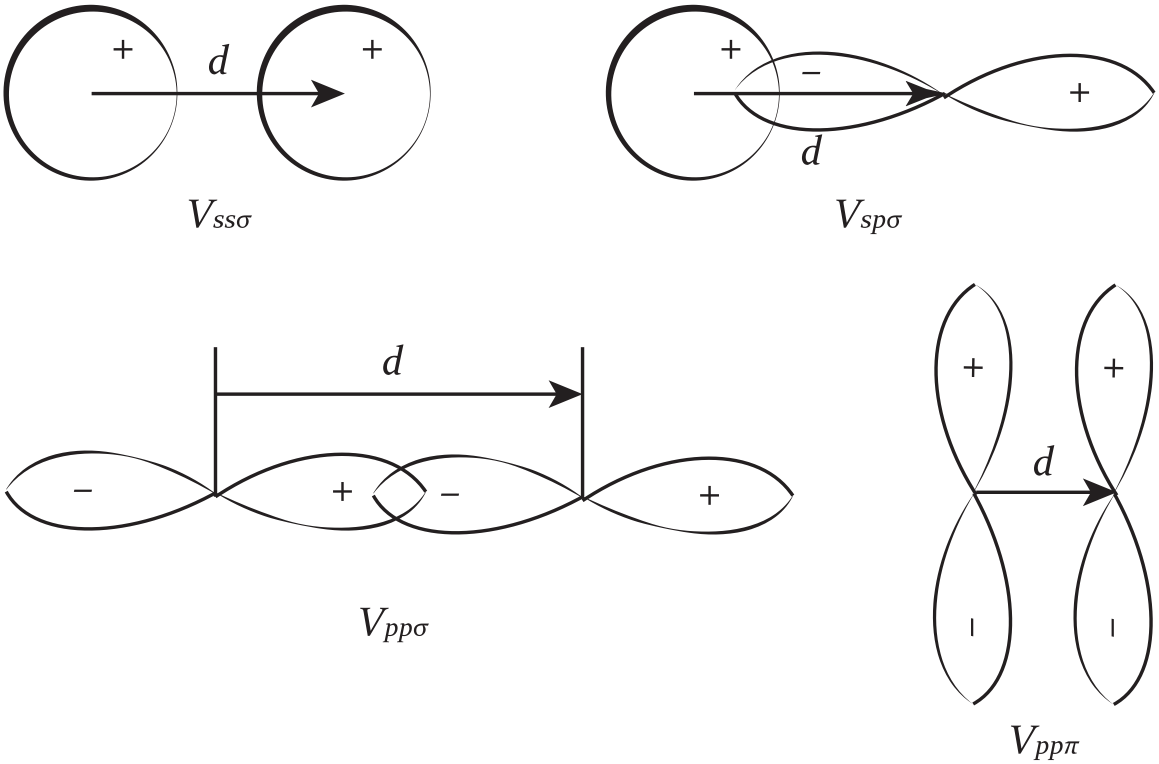


Fig. A.2. Types of bonding between two atoms involving s and p orbitals aligned along x axis and d apart.

There have been efforts to empirically determine the coupling matrix elements ($V_{q,q',r}$s in Fig. A.2) by looking at the similarity between the band energies from linear combination of atomic orbitals (LCAO) theory results and from the free electron as mentioned in Ref. [110] since experimental results showed that the coupling energies were proportional to d^{-2} where d is inter-atomic distance. If the inter-atomic matrix elements are considered only for every nearest neighbor, it can be analytically treated using following relationship [111].

$$V_{q,q',r} = \eta_{q,q',r}\frac{\hbar^2}{md^2} \tag{A.2}$$

where q and q' are type of orbitals at different centers and r denotes bond type (π, σ, δ), respectively. m is free electron mass and d is the distance between atoms. The coupling components of free electrons in variety of lattice structures are then mapped to LCAO bands computed at high symmetry points in the Brillouine zone (BZ). This simple rule turns out to be quite reasonable for first few bands that clearly exhibit s (symmetric) or p (anti-symmetric) nature (Fig. A.3) but inaccurate on describing other bands with mixed symmetry. Therefore, either increasing the number of basis (i.e. s^* or d like orbitals) [36, 112] or additional fitting process is performed to match exact band gap and effective mass information at high symmetry points using numerical optimization algorithms [37,110,113]. Based on the four types

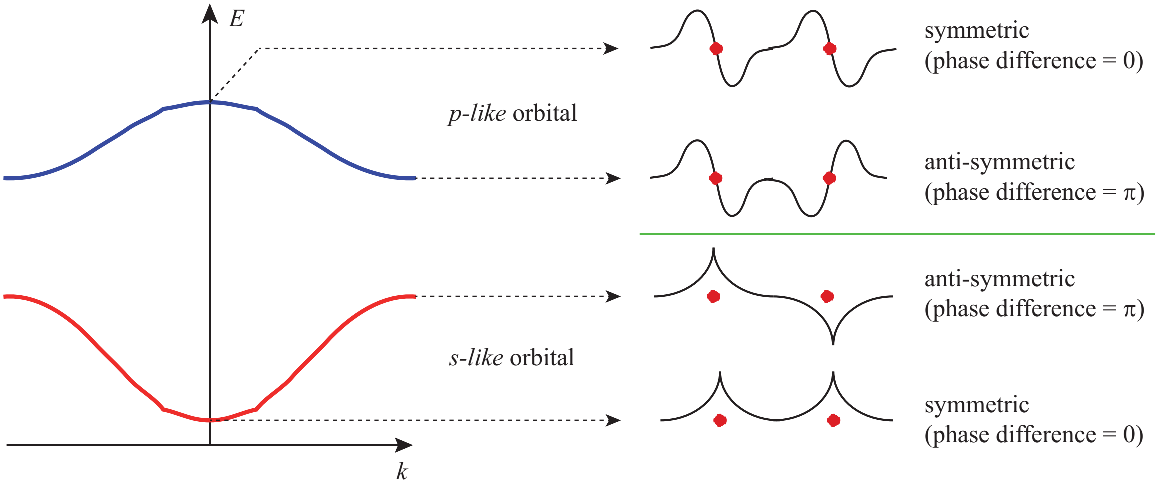


Fig. A.3. Graphical representation of $s-$like and $p-$like bands explained using the *nodal theorem*. The nodal theorem is a qualitative rule that relates between the number of nodes (points crossing zero) in a wavefunction and its energy. In general, more nodes in a wavefunction indicates that the state is higher in energy (less favorable). In the $s-$like band, symmetric wavefunction occupies lower energy since the anti-symmetric wavefunction has more nodes between the atoms. Since the anti-symmetric wavefunction has the phase difference of π, it stays at Brillouine zone boundary with higher in energy than $k = 0$ point. On the contrary, $p-$like orbital is energetically more stable in the anti-symmetric configuration according to the nodal theorem. As a result, the $p-$like band exhibits exactly the opposite behavior of $s-$like band.

of coupling in s, p orbital basis as shown in Fig. A.2, it is possible to define orbital interactions between adjacent atoms aligned in any directions in 3D (i.e. Zincblende, Wurtzite crystals) with some knowledge of vector algebra (Fig. A.4) using two-center approximation [109]. A complete table of two center approximation between s, p and d orbitals are shown in the Table A.1.[2]

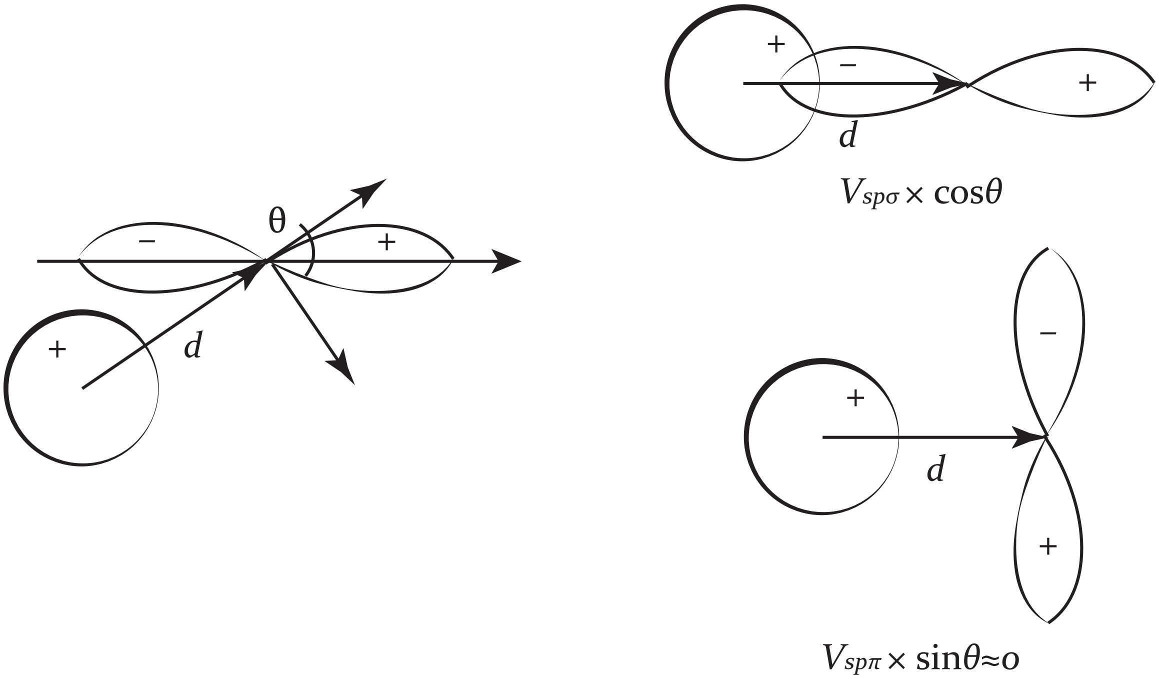

θ
−
+
+
d
+
−
+
d
$V_{sp\sigma} \times \cos\theta$
−
+
d
+
$V_{sp\pi} \times \sin\theta \approx o$

Fig. A.4. Bonding between two atoms aligned along x axis and d apart.

The reader should be able to follow what is described in the next two sections with references with the basic knowledge of how the atomistic tight-binding Hamiltonian matrix elements are constructed described above and the group theory applied to crystal structures for symmetry analysis [114].

[2] For polar materials in which cation and anion are present require parameter sets with polarity information. For instance, $V_{s^a p^c \sigma}$ and $V_{s^c p^a \sigma}$ should be distinguished. Required parameter sets are well documented in other tight-binding references [109, 112].

Table A.1

Table for two center approximation regarding s, p and d type orbitals adopted from Ref. [109]. l, m and n denotes directional cosines $l = d_x/|\vec{d}|$, $m = d_y/|\vec{d}|$ and $n = d_z/|\vec{d}|$, between two atoms, respectively. These formulae applies to all tight-binding Hamiltonian construction

$E_{s,s}$	$V_{ss\sigma}$
$E_{s,x}$	$lV_{sp\sigma}$
$E_{x,x}$	$l^2V_{pp\sigma} + (1 - l^2)V_{pp\pi}$
$E_{x,y}$	$lmV_{pp\sigma} - lmV_{pp\pi}$
$E_{x,z}$	$lnV_{pp\sigma} - lnV_{pp\pi}$
$E_{s,xy}$	$\sqrt{3}lmV_{sd\sigma}$
E_{s,x^2-y^2}	$\frac{1}{2}\sqrt{3}(l^2 - m^2)V_{sd\sigma}$
$E_{s,3z^2-r^2}$	$[n^2 - \frac{1}{2}(l^2 + m^2)]V_{sd\sigma}$
$E_{x,xy}$	$\sqrt{3}l^2mV_{pd\sigma} + m(1 - 2l^2)V_{pd\pi}$
$E_{x,yz}$	$\sqrt{3}lmnV_{pd\sigma} - 2lmnV_{pd\pi}$
$E_{x,zx}$	$\sqrt{3}l^2nV_{pd\sigma} - n(1 - 2l^2)V_{pd\pi}$
E_{x,x^2-y^2}	$\frac{1}{2}\sqrt{3}l(l^2 - m^2)V_{pd\sigma} + l(1 - l^2 + m^2)V_{pd\pi}$
E_{y,x^2-y^2}	$\frac{1}{2}\sqrt{3}m(l^2 - m^2)V_{pd\sigma} - m(1 + l^2 - m^2)V_{pd\pi}$
E_{z,x^2-y^2}	$\frac{1}{2}\sqrt{3}n(l^2 - m^2)V_{pd\sigma} - n(l^2 + m^2)V_{pd\pi}$
$E_{x,3z^2-r^2}$	$l[n^2 - \frac{1}{2}(l^2 + m^2)]V_{pd\sigma} - \sqrt{3}ln^2V_{pd\pi}$
$E_{y,3z^2-r^2}$	$m[n^2 - \frac{1}{2}(l^2 + m^2)]V_{pd\sigma} - \sqrt{3}mn^2V_{pd\pi}$
$E_{z,3z^2-r^2}$	$n[n^2 - \frac{1}{2}(l^2 + m^2)]V_{pd\sigma} + \sqrt{3}n^2(l^2 + m^2)V_{pd\pi}$
$E_{xy,xy}$	$3l^2m^2V_{dd\sigma} + (l^2 + m^2 - 4l^2m^2)V_{dd\pi} + (n^2 + l^2m^2)V_{dd\delta}$
$E_{xy,yz}$	$3lm^2nV_{dd\sigma} + ln(1 - 4m^2)V_{dd\pi} + ln(m^2 - 1)V_{dd\delta}$
$E_{xy,zx}$	$3l^2mnV_{dd\sigma} + mn(1 - 4l^2)V_{dd\pi} + mn(l^2 - 1)V_{dd\delta}$
E_{xy,x^2-y^2}	$\frac{3}{2}lm(l^2 - m^2)V_{dd\sigma} + 2lm(m^2 - l^2)V_{dd\pi} + \frac{1}{2}lm(l^2 - m^2)V_{dd\delta}$
E_{yz,x^2-y^2}	$\frac{3}{2}mn(l^2 - m^2)V_{dd\sigma} - mn[1 + 2(l^2 - m^2)]V_{dd\pi} + mn[1 + \frac{1}{2}(l^2 - m^2)]V_{dd\delta}$
E_{zx,x^2-y^2}	$\frac{3}{2}nl(l^2 - m^2)V_{dd\sigma} + nl[1 - 2(l^2 - m^2)]V_{dd\pi} - nl[1 - \frac{1}{2}(l^2 - m^2)]V_{dd\delta}$
$E_{xy,3z^2-r^2}$	$\sqrt{3}lm[n^2 - \frac{1}{2}(l^2 + m^2)]V_{dd\sigma} - 2\sqrt{3}lmn^2V_{dd\pi} + \frac{1}{2}\sqrt{3}lm(1 + n^2)V_{dd\delta}$
$E_{yz,3z^2-r^2}$	$\sqrt{3}mn[n^2 - \frac{1}{2}(l^2 + m^2)]V_{dd\sigma} + \sqrt{3}mn(l^2 + m^2 - n^2)V_{dd\pi} - \frac{1}{2}\sqrt{3}mn(l^2 + m^2)V_{dd\delta}$
$E_{zx,3z^2-r^2}$	$\sqrt{3}ln[n^2 - \frac{1}{2}(l^2 + m^2)]V_{dd\sigma} + \sqrt{3}ln(l^2 + m^2 - n^2)V_{dd\pi} - \frac{1}{2}\sqrt{3}ln(l^2 + m^2)V_{dd\delta}$
$E_{x^2-y^2,x^2-y^2}$	$\frac{3}{4}(l^2 - m^2)^2V_{dd\sigma} + [l^2 + m^2 - (l^2 - m^2)^2]V_{dd\pi} + [n^2 + \frac{1}{4}(l^2 - m^2)^2]V_{dd\delta}$
$E_{x^2-y^2,3z^2-r^2}$	$\frac{1}{2}\sqrt{3}(l^2 - m^2)[n^2 - \frac{1}{2}(l^2 + m^2)]V + dd\sigma + \sqrt{3}n^2(m^2 - l^2)V_{dd\pi}$ $+\frac{1}{4}\sqrt{3}(1 + n^2)(l^2 - m^2)V_{dd\delta}$
$E_{3z^2-r^2,3z^2-r^2}$	$[n^2 - \frac{1}{2}(l^2 + m^2)]^2V_{dd\sigma} + 3n^2(l^2 + m^2)V_{dd\pi} + \frac{3}{4}(l^2 + m^2)^2V_{dd\delta}$

A.2 Hamiltonian of sp^3s^* model

Although Harrison's simple rule matches reasonably well with simple sp^3 orbital basis, it fails to describe the bulk bandstructure of indirect band gap semiconductors, such as, Si, Ge, AlAs, or GaP. Vogl et al. introduced an additional state, the *excited* s (s^*) orbital, to couple with $p-$ like anti-bonding states near X or L points in the BZ. This s^* orbital suppresses energy level at off Γ points of the conduction band to create an indirect band gap [112].

Table A.2
Onsite (top) and coupling (bottom) matrix in sp^3s^* basis. The matrix is split into 10×10 block matrices for to distinguish coupling matrices from onsite matrices. in Ref. [112]

	$\|s_a\rangle$	$\|p_a^x\rangle$	$\|p_a^y\rangle$	$\|p_a^z\rangle$	$\|s_a^*\rangle$		$\|s_c\rangle$	$\|p_c^x\rangle$	$\|p_c^y\rangle$	$\|p_c^z\rangle$	$\|s_c^*\rangle$
$\langle s_a\|$	ϵ_{s_a}	0	0	0	0	$\langle s_c\|$	ϵ_{s_c}	0	0	0	0
$\langle p_a^x\|$	0	ϵ_{p_a}	0	0	0	$\langle p_c^x\|$	0	ϵ_{p_c}	0	0	0
$\langle p_a^y\|$	0	0	ϵ_{p_a}	0	0	$\langle p_c^y\|$	0	0	ϵ_{p_c}	0	0
$\langle p_a^z\|$	0	0	0	ϵ_{p_a}	0	$\langle p_c^z\|$	0	0	0	ϵ_{p_c}	0
$\langle s_a^*\|$	0	0	0	0	$\epsilon_{s_a^*}$	$\langle s_c^*\|$	0	0	0	0	$\epsilon_{s_c^*}$
	$\|s_c\rangle$	$\|p_c^x\rangle$	$\|p_c^y\rangle$	$\|p_c^z\rangle$	$\|s_c^*\rangle$		$\|s_a\rangle$	$\|p_a^x\rangle$	$\|p_a^y\rangle$	$\|p_a^z\rangle$	$\|s_a^*\rangle$
$\langle s_a\|$	$E_{s_as_c}$	$E_{s_ap_c}$	$E_{s_ap_c}$	$E_{s_ap_c}$	0	$\langle s_c\|$	$E_{s_as_c}$	$-E_{p_as_c}$	$-E_{p_as_c}$	$-E_{p_as_c}$	0
$\langle p_a^x\|$	$-E_{p_as_c}$	$E_{p_a^xp_c^x}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^y}$	$-E_{p_as_c^*}$	$\langle p_c^x\|$	$E_{s_ap_c}$	$E_{p_a^xp_c^x}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^y}$	$E_{s_ap_c^*}$
$\langle p_a^y\|$	$-E_{p_as_c}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^x}$	$E_{p_a^xp_c^y}$	$-E_{p_as_c^*}$	$\langle p_c^y\|$	$E_{s_ap_c}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^x}$	$E_{p_a^xp_c^y}$	$E_{s_ap_c^*}$
$\langle p_a^z\|$	$-E_{p_as_c}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^x}$	$-E_{p_as_c^*}$	$\langle p_c^z\|$	$E_{s_ap_c}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^y}$	$E_{p_a^xp_c^x}$	$E_{s_ap_c^*}$
$\langle s_a^*\|$	0	$E_{s_a^*p_c}$	$E_{s_a^*p_c}$	$E_{s_a^*p_c}$	$E_{s_a^*s_c^*}$	$\langle s_c^*\|$	0	$-E_{p_as_c^*}$	$-E_{p_as_c^*}$	$-E_{p_as_c^*}$	$E_{s_a^*s_c^*}$

The s^* state is also considered to be a spherical orbital, just like the s state. But here, $s - s^*$ terms is omitted. The block Hamiltonian matrix for zincblende crystal can be analytically written as shown in Table A.2. For clarity, the block matrix is split into onsite matrix and coupling matrix and each matrix assumes heteropolar couplings (cation-anion pair).[3] [4]

A.3 Hamiltonian of $sp^3d^5s^*$ model

In the previous section, s^* orbital is introduced by Vogl et al. to couple with anti-symmetric p orbitals to accurately describe the off Γ valley behavior. However, extensive analysis revealed that X and L valleys exhibit d orbital like symmetry. Introducing s^* orbital fails to predict correct transverse masses and at X and L points and second conduction band matches poorly with experiment [36]. Therefore, the sp^3s^* is limited to band gap application (optical properties) at off-Γ valleys. To

[3] In III-V materials, the number of coupling parameters is increased to account for the polarity. For instance, $V_{sp\sigma}$ should now be split into $V_{s_c p_a \sigma}$ and $V_{s_a p_c \sigma}$. However, in homopolar materials, it is safe to equate those two parameters and construct Hamiltonian.

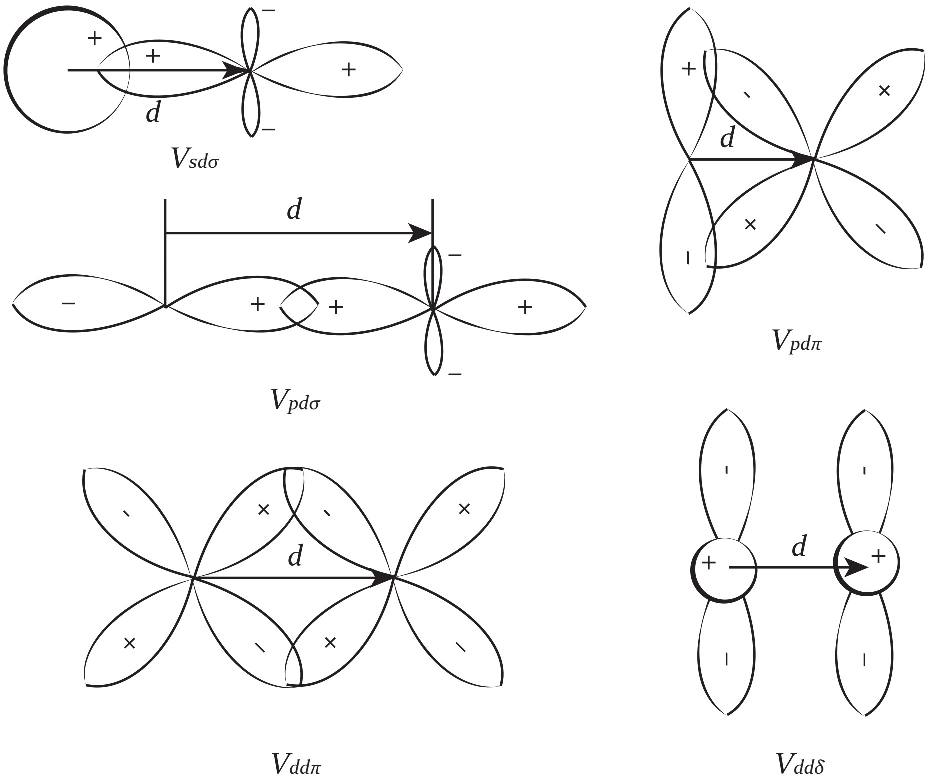
[4] Note that the coupling matrix elements in Table A.2 and Ref. [112] are different from the hopping terms discussed previously. To make the hamiltonian compatible with $sp^3d^5s^*$ Hamiltonian discussed in the next section, it is required to convert the parameters in Vogl's notation to the hopping parameters. The conversion formula is provided in the following equations.

$$\begin{aligned}
E_{s_a s_c} &= 4V_{s_a s_c \sigma},\ E_{s_a^* s_c^*} = 4V_{s_a^* s_c^* \sigma} \\
E_{s_a p_c} &= 4\left(\frac{1}{\sqrt{3}}\right)V_{s_a p_c \sigma},\ E_{s_a^* p_c} = 4\left(\frac{1}{\sqrt{3}}\right)V_{s_a^* p_c \sigma} \\
E_{p_a s_c} &= 4\left(\frac{1}{\sqrt{3}}\right)V_{s_c p_a \sigma},\ E_{p_a^* s_c} = 4\left(\frac{1}{\sqrt{3}}\right)V_{s_c^* p_a \sigma} \\
E_{p_a^x p_c^x} &= 4\left(\frac{1}{3}V_{p_a p_c \sigma} + \frac{2}{3}V_{p_a p_c \pi}\right) \\
E_{p_a^x p_c^y} &= 4\left(\frac{1}{3}V_{p_a p_c \sigma} - \frac{1}{3}V_{p_a p_c \pi}\right)
\end{aligned} \tag{A.3}$$

We can now consistently use the notation in Ref. [109].

$$\begin{aligned}
V_{s_a s_c \sigma} &= \frac{1}{4}E_{s_a s_c},\ V_{s_a^* s_c^* \sigma} = \frac{1}{4}E_{s_a^* s_c^*} \\
V_{s_a p_c \sigma} &= \left(\frac{\sqrt{3}}{4}\right)E_{s_a p_c},\ V_{s_a^* p_c \sigma} = \left(\frac{\sqrt{3}}{4}\right)E_{s_a^* p_c} \\
V_{s_c p_a \sigma} &= \left(\frac{\sqrt{3}}{4}\right)E_{p_a s_c},\ V_{s_c^* p_a \sigma} = \left(\frac{\sqrt{3}}{4}\right)E_{p_a^* s_c} \\
V_{p_a p_c \pi} &= E_{p_a^x p_c^x} - E_{p_a^x p_c^y} \\
V_{p_a p_c \sigma} &= \frac{1}{4}E_{p_a^x p_c^x} + \frac{1}{2}E_{p_a^x p_c^y}
\end{aligned} \tag{A.4}$$

incorporate correct $d-$type symmetry in the bandstructure, Jancu et al. introduced five additional d states ($xy,\ yz,\ zx,\ x^2-y^2,\ 3z^2-r^2$) for better agreement with experimentally known bandstructure. Fig. A.5 graphically shows additional orbital interactions needed to compute the coupling matrix components.[5] In addition, the interaction by s^* orbital was limited to s^*-p in sp^3s^* model, but in the $sp^3d^5s^*$ model, it turns out that s^* coupling to other orbitals should not be ignored for more accurate results.



$V_{sd\sigma}$

$V_{pd\sigma}$

$V_{pd\pi}$

$V_{dd\pi}$

$V_{dd\delta}$

Fig. A.5. Two-center orbital interactions involving d orbitals.

Again, the details of how each orbital interactions are derived are written in Ref. [36]. Actual construction is identical to previous Hamiltonian; two-center approximation (Table A.1) is used for any type of crystals.

[5] Like $V_{sp\pi}=0$, all other combinations of interactions are ignored.

B. ITERATIVE METHOD FOR SOLVING POISSON EQUATION

In this chapter, an efficient way to solve a non-linear Poisson equation is introduced based on Ref. [57]. The Poisson solver described below is implemented in `poisson_solver.cpp & poisson_solver.hpp` and corresponding Poisson matrix is constructed in `FDMPoisson.cpp & FDMPoisson.hpp` in `NEMO3D-peta`.

B.1 Poisson equation

To obtain the electrostatic potential profile of any device it is clear to solve Poisson equation with given charge profile and boundary conditions. Poisson equation can be written as follows.

$$\begin{aligned}\nabla \cdot (\epsilon(r)\nabla V(r)) &= -q(p(r) - n(r) + N^{+}(r) - N^{-}(r)) \\ \nabla\epsilon(r) \cdot \nabla V(r) + \epsilon(r)\nabla^2 V(r) &= -q(p(r) - n(r) + N^{+}(r) - N^{-}(r))\end{aligned} \tag{B.1}$$

where $\epsilon(r)$ and $V(r)$ is the dielectric constant times electric permittivity $K_s(r)\epsilon_0$[1] and potential at position r, respectively. $p(r)$ and $n(r)$ are position dependent positive and negative *mobile* charge density and $N^(+)(r)$ and $N^(-)(r)$ are positive and negative *stationary* charge density[2], respectively. Since Poisson equation is based on the finite difference method (FDM), discretization of (B.1) is required for implementation. In 1D example, the discretization with grid size Δ can be written as,

$$\begin{aligned}\nabla_x f(x = n\Delta) &= \frac{f(x = (n+1)\Delta) - f(x = n\Delta)}{\Delta} \\ \nabla_x^2 f(x = n\Delta) &= \frac{\nabla_x f(x = (n+1)\Delta) - \nabla_x f(x = n\Delta)}{\Delta} \\ &= \frac{f(x = (n+1)\Delta) - 2f(x = n\Delta) + f(x = (n-1)\Delta)}{\Delta^2}\end{aligned} \tag{B.2}$$

[1] Usually the dielectric constant is uniform, however, in many semiconductor problems, such as heterojunction devices, the dielectric constant is material dependent value

[2] For example, ionized dopants

Utilizing (B.2), it is possible to discretize (B.1). Again, in 1D grid:

$$\begin{aligned} \left(\frac{\epsilon_{n+1} - \epsilon_n}{\Delta}\right)\left(\frac{V_{n+1} - V_n}{\Delta}\right) + \epsilon_n \frac{V_{n+1} - 2V_n + V_{n-1}}{\Delta^2} &= -q(p_n - n_n + N_n^+ - N_n^-) \\ \frac{\epsilon_{n+1}}{\Delta^2}V_{n+1} - \frac{\epsilon_{n+1} + \epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-) \end{aligned} \tag{B.3}$$

where the subscript n stands for n-th grid point ($x = n\Delta$). From the equation above, for every grid point n, it is coupled only to its next nearest neighbor. As a result, the second order differential equation is transformed in to a linear algebraic equation $AV = q$. It is trivial to expand this discretization scheme in 3D.

B.1.1 Boundary conditions

In general, there are two types of boundary conditions (BCs) in Poisson equation. One is fixed potential BC and the other is zero field BC. In any kind of device, fixed potential BC is imposed wherever there are metal contacts (electrodes) attached and zero field BC elsewhere. The solution of Poisson equation is trivial if all boundaries are treated as zero field.[3]

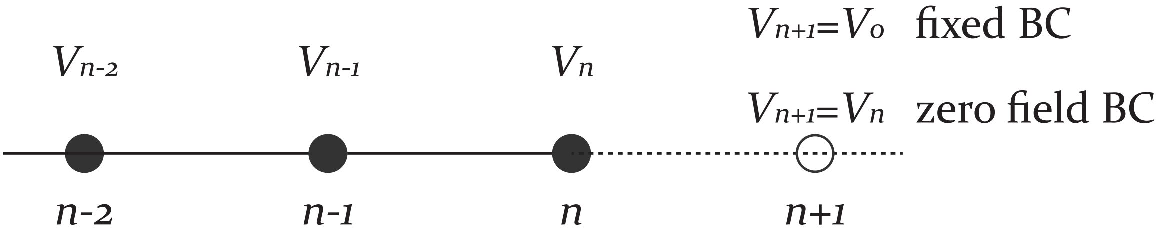


Fig. B.1. Boundary condition set up in FDM. Imaginary grid can be set as either fixed value (V_0) or zero field (V_n) depending on the device geometry.

[3] Mathematically, the matrix A becomes singular.

Let's assume there is an imaginary grid point $(n+1)$ (metal contact) outside the domain with fixed potential V_0 (Fig. B.1). Using (B.3), the discretized equation at grid point n can be written as,

$$\begin{aligned}
\frac{\epsilon_{n+1}}{\Delta^2}V_{n+1} - \frac{\epsilon_{n+1}+\epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-) \\
\frac{\epsilon_{n+1}}{\Delta^2}V_0 - \frac{\epsilon_{n+1}+\epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-) \qquad \text{(B.4)} \\
-\frac{\epsilon_{n+1}+\epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-) - \frac{\epsilon_{n+1}}{\Delta^2}V_0
\end{aligned}$$

it is clearly shown that the fixed potential is now included in the right-hand side of the matrix equation.

The zero field BC can be similarly derived from (B.3) by assuming the imaginary $(n+1)$-th grid has the same potential data as V_n to force zero field at the boundary.

$$\begin{aligned}
\frac{\epsilon_{n+1}}{\Delta^2}V_{n+1} - \frac{\epsilon_{n+1}+\epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-) \\
\frac{\epsilon_{n+1}}{\Delta^2}V_n - \frac{\epsilon_{n+1}+\epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-) \qquad \text{(B.5)} \\
-\frac{\epsilon_n}{\Delta^2}V_n + \frac{\epsilon_n}{\Delta^2}V_{n-1} &= -q(p_n - n_n + N_n^+ - N_n^-)
\end{aligned}$$

Forcing zero-field at the boundary simply modifies the diagonal matrix element.

In summary, Poisson equation can easily be discretized using finite difference method. There are two types of boundary condition, one is fixed BC and is added to the right-hand side of the equation and the other is zero field BC which modifies the diagonal term of the Matrix A. In any cases, proper BCs should be chosen to correctly compute the potential profile in the device.

B.2 Newton-Raphson's method for solving Poisson equation

The Newton-Raphson's method is a simple and popular way to solve nonlinear differential equations. It is an iterative procedure to obtain the solution x from successive equations [115].

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$f(x_n) + f^{'}(x_n)\Delta x_n = 0,\ x_{n+1} = x_n + \Delta x_n$$

This iteration is carried out until mse(x_n) < tolerance. It is also possible to make it as a matrix version [57].

$$\begin{aligned} f_n(\vec{x}^{\ i}) = 0,\ n = 1, \cdots, m,\ \vec{x} \in \mathbb{R}^{n\times 1} \\ f_n(\vec{x}^{\ i}) + \sum_{j=1}^{m} \frac{\partial f_n}{\partial x_j}\bigg|_{\vec{x}^i} \Delta x_j^i = 0 \\ \vec{x}^{\ i+1} = \vec{x}^{\ i} + \Delta x^i \end{aligned} \tag{B.6}$$

Poisson equation can be interpreted as a non-linear vector function and it is also possible to use similar approach. Poisson equation in matrix version is written as $A\vec{V} = \vec{\rho}(\vec{V})$, where A is discretized version of differential operator, $\vec{V}$ is desired potential at each grid point, and $\vec{\rho}(\vec{V})$ is the right-hand side of the equation. Physically, the right-hand side of the equation represents charge which is in fact a non-linear function of potential, making ρ a function of $\vec{V}$. The boundary condition mentioned in the previous section is by default included both in A and $\vec{\rho}(\vec{V})$. Now the solution becomes solving the equation $f(\vec{V}) = A\vec{V} - \vec{\rho}(\vec{V}) = 0$. For every grid point $n = 1, \cdots, m$, we can convert the equation to the iterative form of Newton's equation using (B.6).

$$\begin{aligned} f_n(\vec{V}^i) = \frac{\epsilon_{n+1}}{\Delta^2} V_{n+1}^i - \frac{\epsilon_{n+1} + \epsilon_n}{\Delta^2} V_n^i + \frac{\epsilon_n}{\Delta^2} V_{n-1}^i + \rho_n(V_n^i)(= 0) \\ \sum_{j=1}^{m} \frac{\partial f_n}{\partial V_j^i}\bigg|_{\vec{V}} \Delta V_j^i = \frac{\epsilon_{n+1}}{\Delta^2} \Delta V_{n+1}^i - \frac{\epsilon_{n+1} + \epsilon_n}{\Delta^2} \Delta V_n^i + \frac{\epsilon_n}{\Delta^2} \Delta V_{n-1}^i + \frac{\partial \vec{\rho}(V_n^i)}{\partial V_n} \Delta V_n^i \\ \sum_{j=1}^{m} \frac{\partial f_n}{\partial V_j^i}\bigg|_{\vec{V}} \Delta V_j^i = -f_n(\vec{V}^i) \\ \left\{ A + \mathrm{diag}\left(\frac{\partial \rho_n(V_n^i)}{\partial V_n} \right) \right\} \Delta \vec{V}^i = -(A\vec{V}^i - \rho_n(\vec{V}^i)) \\ \vec{V}_n^{i+1} = \vec{V}_n^i + \vec{\Delta} V_n^i,\ \text{until mse}(\Delta \vec{V}^i) < \text{tolerance} \end{aligned} \tag{B.7}$$

To compute $\Delta\vec{V}^i$, not only A matrix with boundary conditions, but also the *mobile* charge profile ($\rho(\vec{V}^i)$) and its derivative ($\partial\vec{\rho}(\vec{V}^i)/\partial\vec{V}^i$) at each grid point should be updated with respect to $\vec{V}^i$ at every iteration.[4]

[4] For fixed charge, such as ionized dopants are not independent of the potential profile, rather it is treated as constant.

If the charge profile is computed semi-classically (Thomas-Fermi approximation) [116] then at iteration i, the charge profile and its derivative at grid point j will become

$$\begin{aligned}
n_j^i &= (N_C)_j \mathcal{F}_{1/2}\left(\frac{E_F - ((E_C)_j + V_j^i)}{kT/q}\right) \\
p_j^i &= (N_V)_j \mathcal{F}_{1/2}\left(\frac{((E_V)_j + V_j^i) - E_F}{kT/q}\right) \\
\frac{\partial n_j^i}{\partial V_j^i} &= -\frac{(N_C)_j}{kT/q}\mathcal{F}_{-1/2}\left(\frac{E_F - ((E_C)_j + V_j^i)}{kT/q}\right) \\
\frac{\partial p_j^i}{\partial V_j^i} &= \frac{(N_V)_j}{kT/q}\mathcal{F}_{-1/2}\left(\frac{((E_V)_j + V_j^i) - E_F}{kT/q}\right)
\end{aligned} \tag{B.8}$$

On the other hand, if the charge is quantum mechanical, the charge is set as the initial guess of Newton-Raphson iteration. Let's consider electron charge ($n_0 \equiv n_q(V_q)$) for convenience, where $n_q(V_q)$ is quantum charge with previously converged potential, V_q. First, compute $\vec{E}_F$ profile for every grid j. Note, this is not the physical Fermi level, rather it should be considered as point-wise .

$$\begin{aligned}
(n_Q)_j \equiv n_j^0 &= N_C^j \mathcal{F}_{1/2}\left(\frac{(E_F)_j - ((E_C)_j + (V_Q)_j)}{kT/q}\right) \\
(E_F)_j &= \frac{kT}{q}\mathcal{F}_{1/2}^{-1}\left(\frac{n_j^0}{N_C^j}\right) + ((E_C)_j + (V_Q)_j)
\end{aligned} \tag{B.9}$$

Now, E_F in (B.8) simply is replaced by $(E_F)_j$ for Newton-Raphson iteration.

In summary, Poisson equation implemented in `NEMO3D-peta` is based on finite difference method. Two different boundary conditions, fixed BC and zero field BC can be chosen depending on the device type and geometry. Poisson equation can be solved by iterative process called Newton-Raphson method.

C. CALCULATION OF DENSITY OF STATES FROM ARBITRARY BANDSTRUCTURE DATA

C.1 Introduction

Most device physicists and engineers learn how to derive the analytical expression for DOS under parabolic $E-k$ relations as shown in Fig. C.1 [15]. In reality, however, bands are not perfectly parabolic; non-parabolicity such as warping and anti-crossings between sub-bands is very common in semiconductor bandstructures. In this case, the analytical model learned from basic solid-state physics course does not apply any more and numerical treatment should be considered. In this chapter, a simple way of computing the DOS from arbitrary dispersion relation is introduced.

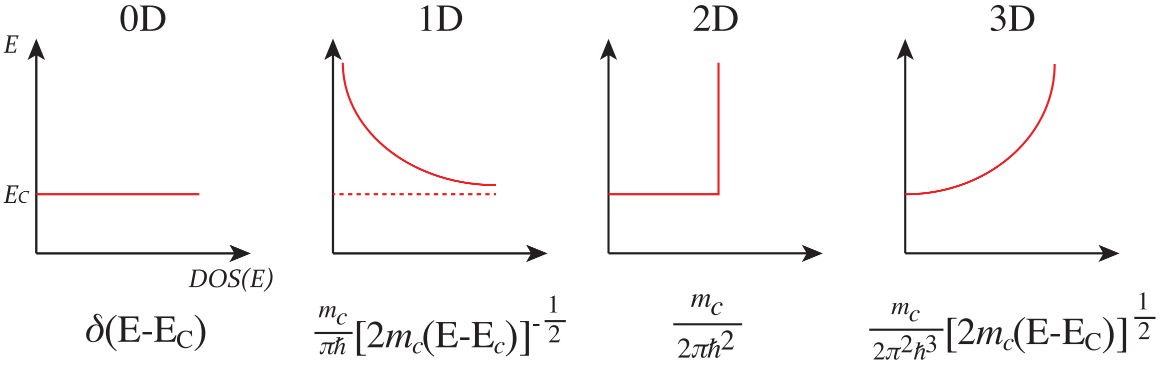


Fig. C.1. 0/1/2/3D density of states expression for parabolic band $E = \hbar^2k^2/2m_c$. Spin degeneracy is not included in the analytical expression.

C.2 Derivation

In many cases, $E-k$ dispersion is stored in discretized format (k_x, k_y, k_z, E) with equal spacing in each direction (Δk_x, Δk_y, Δk_z). A full Brillouine zone (BZ) data is expected but sometimes only a part of Brillouine zone (BZ) is simulated due to the symmetry in the k-space.

First, let's consider the full BZ representation and consider only 1D case since 2D and 3D are easily derived from 1D relationship. Total number of states up to energy E can be written as

$$N(E) = s \sum_{n} \sum_{k_i} \Theta(E - \epsilon^{n}_{k_i})$$

where, n is the sub-band index and k_i is discretized k points from dispersion data. s is spin degeneracy term and if spin factor is included in the raw data $s = 1$ and $s = 2$ otherwise. k_i can be expressed as

$$k_i = \frac{2\pi}{L_x} i, \;\; i \in \mathbb{Z}$$

where L_x is the length of the supercell. The DOS is the energy derivative of $N(E)$.

$$D(E) = \frac{d}{dE} N(E) = s \sum_{n} \sum_{k_i} \delta(E - \epsilon^{n}_{k_i}) \tag{C.1}$$

In an infinite bulk, $N_x \to \infty$, Eqn. (C.1) becomes an integral form as k becomes a continuous function and it is now changed to integral form.

$$\begin{aligned} D(E) &= s \sum_{n} \int_{-0.5}^{0.5} \frac{a_x}{2\pi} dx \; \delta(E - \epsilon^{n}_{k_i}) \\ &= s \sum_{n} \sum_{i} \frac{a_x}{2\pi} \frac{1}{N_x} \; \delta(E - \epsilon^{n}_{i}) \end{aligned} \tag{C.2}$$

where N_x is the number of discretized k points. Eqn. (C.2) can easily expanded to 2D and 3D form in the following way

$$D(E) = \frac{s}{N_x N_y N_z} \left(\frac{a_x}{2\pi} \cdot \frac{a_y}{2\pi} \cdot \frac{a_z}{2\pi} \right) \sum_{i} \delta(E - \epsilon_i) \tag{C.3}$$

Note since we are simply counting all available states, the sub-band index is dropped.

Next, If the dispersion data is only computed for one part of BZ as shown in Fig. C.2, DOS is multiplied by some degeneracy factor to account for the full BZ. This is due to the fact that the bandstructure always has symmetry along family of directions $< hkl >$. However, the degeneracy factor should be carefully determined at

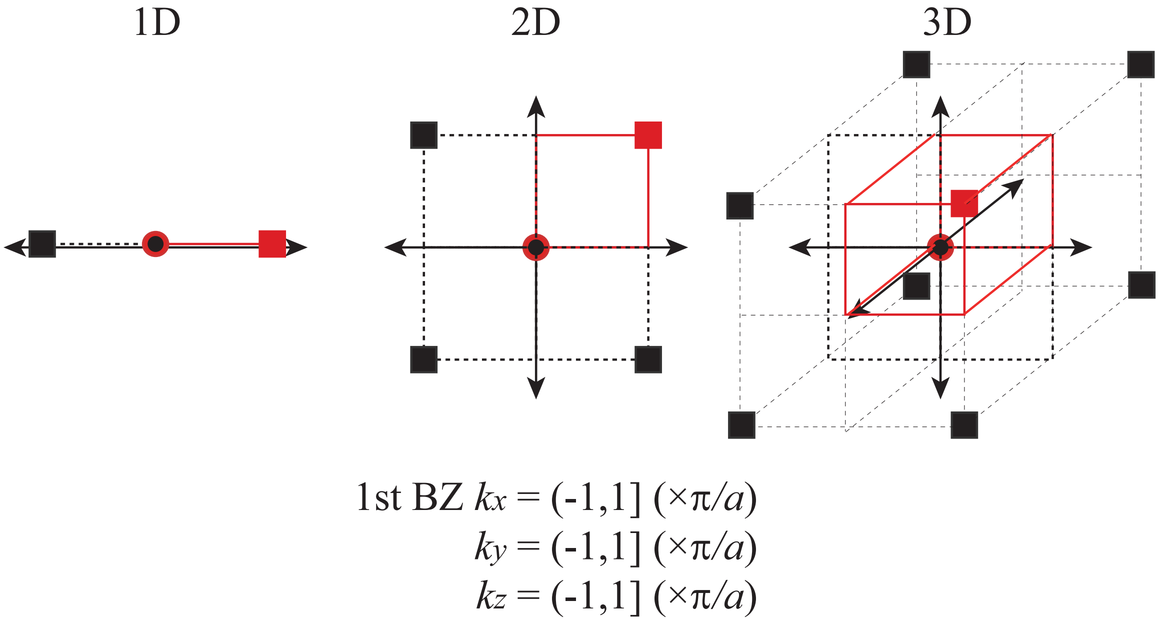


Fig. C.2. Symmetry in the first BZ in 1/2/3D. Red line(box) represents partial BZ data in most simulation cases. Black dashed area indicates where the bands are symmetric with respect to red area. symmetry direction is marked as red and black squares. In DOS calculations with given $E-k$ relation, it must be multiplied by 2, 2^2, and 2^3 for 1D, 2D, and 3D cases, respectively to account for the whole DOS in the full BZ. However, the multiplicity must be adjusted around the boundaries to avoid overestimation of DOS.

zone boundaries to avoid DOS overestimation. Let's give a 1D example and suppose only $E-k$ data we have is $k_x = [0\ 1](\times\pi/a_x)$. Then from Fig. C.2, it is natural to multiply by 2 to computed DOS and it is equivalent to adding DOS at $k_x = [0\ 1]$ and $k_x = [-1\ 0]$. But actual first BZ is $k_x = (-1\ 1]$ and we immediately realize that $k_x = 0$ and $k_x = 1$ are counted twice, leading to overestimation of DOS. Therefore, it is required to count the states at $k_x = 0, 1$ only once $(2 \times 1/2 = 1)$. This is same for 2D/3D cases and simple checking routine can be generally summarized as following.

1. m (multiplicity factor) $= 2$(1D) / 2^2(2D) / 2^3(3D)

2. For state i, examine each component of $\vec{k}$ to count how many components lie at the boundary $(0,\ 1)$. Denote this number c.

3. $m' = m/(2)^c$ will be the multiplicity of the state i.
4. $DOS_i \leftarrow DOS_i \times m'$, $DOS+ = DOS_i$

C.3 DOS refinement

As seen in Eqn. (C.3), each state is represented by delta function. In principle, it is required that one needs *infinite* number of $E - k$ data to obtain an ideal DOS plot. In reality, however, only minimum amount of data is computed to minimize the computation time, resulting in very noisy DOS plots. To overcome this problem, one can think of replacing the delta function representation of each state by much smoother function, such as, normalized Gaussian or Lorentzian.

$$f_G(E-\epsilon,\gamma) = \frac{1}{\sqrt{2\pi\gamma^2}} e^{-\frac{(E-\epsilon)^2}{2\gamma^2}}$$
$$f_L(E-\epsilon,\gamma) = \frac{\gamma}{(E-\epsilon)^2 + (\gamma/2)^2}$$

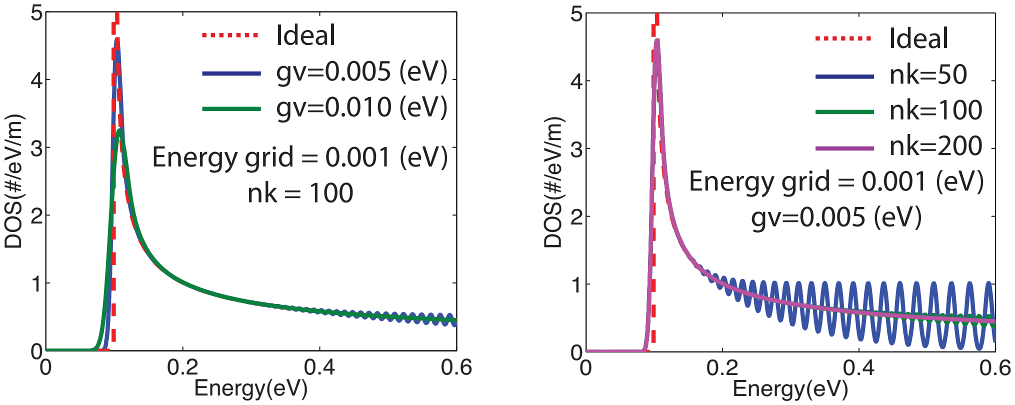
where γ is the broadening factor for each function. Thus DOS expression (Eqn. (C.3)) is rewritten as

$$D(E) = \frac{s}{N_x N_y N_z}\left(\frac{a_x}{2\pi}\cdot\frac{a_y}{2\pi}\cdot\frac{a_z}{2\pi}\right)\sum_i f_{G \text{ or } L}(E-\epsilon_i,\gamma) \tag{C.4}$$

These functions act as a low pass filter to each state and it results in much smoother DOS plot. Since, we are ignoring high frequency components, DOS with high peaks (i.e. 1D DOS plot at singular point or 2D DOS right at the step rise) cannot be properly represented. Despite of shortcomings, since the area of broadened function is equal to one and the computed results are in general not too far off from actual DOS values. For better shapes, γ can be adjusted depending on the density of the k grid.

C.4 Comparison results

Numerical comparisons were performed on a parabolic dispersion $E = k^2$ on 1/2/3D cases, respectively. First, 1D DOS profile was calculated and compared using analytical expression and numerical integration as shown in Fig.C.3. Since each state is broadened, a finite and smooth rise in DOS instead of infinite value ($1/\sqrt{E-0.1}$) is observed at the lowest energy (0.1 in the figure). Also, wiggles are observed in higher energy points since the sample points are not dense enough. The wiggles can be successfully eliminated by increasing the broadening factor or setting more dense k points. However, the broadening factor should be carefully adjusted as well as number of sample points, especially at points where a steep rise of DOS is observed since most of the inaccuracy stems from getting rid of high frequency components.

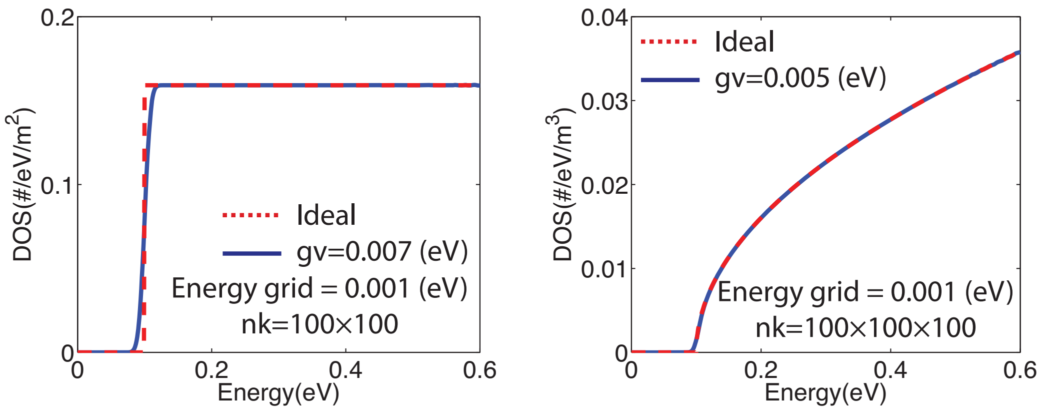


(a)

(b)

Fig. C.3. 1D DOS comparison for $E = k_x^2{+}0.1$. (a) The effect of broadening factor (gv) with fixed sample point nk. (b) The effect of sample points. More wiggles appear in smaller number of sample points due to the under sampling at high energy region.

Wiggles in DOS are also observed in 2D and 3D cases, when you do not have sufficient number of sample points or not enough broadening in each state. The number of sample points in 2D and 3D dispersion relations is limited since sample points increase by N_k^2 and N_k^3, respectively. Therefore, adjusting broadening parameter can be critical to show better DOS plot close to actual ones. Comparison of DOS in 2D and 3D cases is shown in Fig. C.4.



(a) 2D DOS

(b) 3D DOS

Fig. C.4. 2D and 3D DOS comparison for parabolic bands. Wiggles come from the broadened representation (no high frequency components) of each state. User can adjust the broadening factor to obtain DOS closer to the actual one.

C.5 Matlab code for DOS comparison

```
function bulk_DOS_test(dim,gv,nk)

% bulk_DOS\_test(dim,gv,nk)

% Comparison numerical result vs. analytical solution of E=k^2

% dim: dimension (1/2/3) of the parabolic band E=k^2.

% gv: Broadening factor in (eV). Set to (0~10e-3)(eV) depending on simulation conditions.

% nk: Number of k-points between the interval k=[0 1].

%     Note: setting this value too large might cause memory error,

%     especially in 2D and 3D cases.

% Spin degeneracy is set to 2


close all;


if dim==3

    % 3D

    kx=0:1/nk:1;lkx=length(kx);

    ky=0:1/nk:1;lky=length(ky);
```

```
kz=0:1/nk:1;lkz=length(kz);

[kxx kyy kzz]=meshgrid(kx,ky,kz);



E=(kxx.^2+kyy.^2+kzz.^2)+0.1;

Egrid=0.0:1e-3:0.6;


E=reshape(E,[lkx*lky*lkz 1]);

kxx=reshape(kxx,[lkx*lky*lkz 1]);

kyy=reshape(kyy,[lkx*lky*lkz 1]);

kzz=reshape(kzz,[lkx*lky*lkz 1]);


DOS=zeros(size(Egrid));


for eidx=1:length(E)

    m_factor=8;


    if abs(kxx(eidx,1)+0)<1e-3 || abs(kxx(eidx,1)-1)<1e-3

        m_factor = m_factor/2;

    end


    if abs(kyy(eidx,1)+0)<1e-3 || abs(kyy(eidx,1)-1)<1e-3

        m_factor = m_factor/2;

    end


    if abs(kzz(eidx,1)+0)<1e-3 || abs(kzz(eidx,1)-1)<1e-3

        m_factor = m_factor/2;

    end


    DOS = DOS + 2/sqrt(2*pi*gv*gv)*exp(-(E(eidx,1)-Egrid).^2 ...

                             /(2*gv*gv))*m_factor/(nk*nk*nk);

end


figure(1);
```

```
    set(gca,'Fontsize',24);
    plot(Egrid,DOS/(8*pi*pi*pi),'b');hold on;

    X=Egrid;
    Y=1/(2*pi*pi)*sqrt(X-0.1);
    plot(X,Y,'r')
    axis([0 0.6 0 max(DOS)/(8*pi*pi*pi)*1.2])

    xlabel('Energy(eV)');
    ylabel('DOS(#/eV/m^{3})');
end

if dim==2
    %% 2D
    kx=0:1/nk:1;
    ky=kx;
    [kxx kyy ]=meshgrid(kx,ky);

    lkx=length(kx);
    lky=length(ky);

    E=(kxx.^2+kyy.^2)+0.1;
    Egrid=0.0:1e-3:0.6;

    E=reshape(E,[lkx*lky 1]);
    kxx=reshape(kxx,[lkx*lky 1]);
    kyy=reshape(kyy,[lkx*lky 1]);


    DOS=zeros(size(Egrid));

    for eidx=1:length(E)
        m_factor=4;

        if abs(kxx(eidx,1)+0)<1e-3 || abs(kxx(eidx,1)-1)<1e-3
```

```
            m_factor = m_factor/2;
        end

        if abs(kyy(eidx,1)+0)<1e-3 || abs(kyy(eidx,1)-1)<1e-3
            m_factor = m_factor/2;
        end


        DOS = DOS + 2/sqrt(2*pi*gv*gv)*exp(-(E(eidx,1)-Egrid).^2 ...
                                 /(2*gv*gv))*m_factor/(nk*nk);
    end

    figure(2);
    set(gca,’Fontsize’,24);

    plot(Egrid,DOS/(4*pi*pi),’b’);hold on;

    X=Egrid;
    for idx=1:length(X)
        if X(idx)<0.1
            Y(idx)=0;
        else
            Y(idx)=1/(2*pi);
        end
    end
    plot(X,Y,’r’)
    axis([0 0.6 0 max(DOS)/(2*pi*2*pi)*1.2])
    xlabel(’Energy(eV)’);
    ylabel(’DOS(#/eV/m^{2})’);
end

if dim==1
    %% 1D
    kx=0:1/nk:1;
    Egrid=-0.1:1e-3:0.6;
    kxx=kx';
    E=(kxx.^2)+0.1;



    DOS=zeros(size(Egrid));


    for eidx=1:length(E)
        m_factor=2;


        if abs(kxx(eidx,1)+0)<5e-6 || abs(kxx(eidx,1)-1)<5e-6
            m_factor = m_factor/2;
        end



        DOS = DOS + 2/sqrt(2*pi*gv*gv)*exp(-(E(eidx,1)-Egrid).^2 ...
                                /(2*gv*gv))*m_factor/nk;
    end


    figure(3);
    set(gca,'Fontsize',24);


    plot(Egrid,DOS/(2*pi),'b');hold on;



    X=1e-8:1e-3:1.0;
    Y=1/(1*pi)./sqrt(X-0.1);
    plot(X,Y,'r')
    axis([0 0.6 0 max(DOS)/(2*pi)*1.2])


    xlabel('Energy(eV)');
    ylabel('DOS(#/eV/m)');
end
```

D. COMPUTATION OF COULOMB AND EXCHANGE INTEGRAL IN TIGHT-BINDING

This chapter discusses about the construction of multi-electron Hamiltonian and describes about how to evaluate Coulomb and exchange integrals in tight-binding. It is implemented in `configuration_interaction.cpp`. Previous formulation and application to quantum dots can be found in the Refs. [69, 117–120].

D.1 The two-electron wavefunction

The time-independent single particle (electron) Schrödinger equation is written as

$$H_{\text{single}}\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{D.1}$$

where $H_{\text{single}} = H_0 + V_{ext}(\mathbf{r})$ and $H_0 = -\frac{\hbar^2}{2m}(\mathbf{p}+e\mathbf{A})^2$, respectively. The two-electron wavefunction is built upon two single-electron wave functions with satisfying following conditions. First, electrons are indistinguishable particles; the probability of wavefunction should be invariant when each electron exchanges its position; if $\Psi(\mathbf{r}_1, \mathbf{r}_2)$ is the two-electron wavefunction, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the positions of electron 1 and 2, respectively, then $|\Psi(\mathbf{r}_1, \mathbf{r}_2)|^2 = |\Psi(\mathbf{r}_2, \mathbf{r}_1)|^2$. If the wavefunction of each electron is defined as $\psi_1(\mathbf{r})$ and $\psi_2(\mathbf{r})$, respectively, there are two possibilities to satisfy the probability condition.

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) + \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\}$$
$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) - \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\}$$

Second, Pauli's exclusion principle should be satisfied; electrons cannot occupy the same position simultaneously, which the wavefunction should lead to $\Psi(\mathbf{r}_1, \mathbf{r}_2) = 0$, whenever $\mathbf{r}_1 = \mathbf{r}_2$. The only solution that satisfies both conditions would be the *anti-symmetric* wavefunction.

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) - \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\} \tag{D.2}$$

In general, the n-electron wavefunction that satisfies the indistinguishability and Pauli's exclusion principle can be constructed using *Slater determinant.*

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r_n}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\mathbf{r_1}) & \psi_2(\mathbf{r_1}) & \cdots & \psi_n(\mathbf{r_1}) \\ \psi_1(\mathbf{r_2}) & \psi_2(\mathbf{r_2}) & \cdots & \psi_n(\mathbf{r_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{r_n}) & \psi_2(\mathbf{r_n}) & \cdots & \psi_n(\mathbf{r_{n1}}) \end{vmatrix}$$

D.1.1 Considering Spin in Two-electron Wavefunction

In the absence of magnetic field or high electric field, all electronic states are doubly degenerate due to spin. Spin up (α) and spin down (β) states are independent from real space assumed to be orthonormal[1] and indistinguishable in spin space. Single electron wavefunction including spin ($\chi(\mathbf{r}, s)$) is represented as either $\psi(\mathbf{r})\alpha$ or $\psi(\mathbf{r})\beta$. Even with spin information included, the two-electron wavefunction must satisfy the anti-symmetric property. Since ψ and (α and β) are acting on different subspace, we can treat them independently to satisfy the following to construct an anti-symmetric wavefunction :

- spatial part of the wavefuction is symmetric and spin is anti-symmetric.
- spatial part of the wavefuction is anti-symmetric and spin is symmetric.

[1] $< \alpha|\alpha >= 1, < \beta|\beta >= 1$ and $< \alpha|\beta >= 0$

For instance, six possible two-electron wavefunctions can be constructed from two doubly degenerate single electron states ($\psi_1\alpha,\ \psi_1\beta,\ \psi_2\alpha,\ \psi_2\beta$).

$$\chi_1(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2) = \psi_1(\mathbf{r}_1)\psi_1(\mathbf{r}_2) \times \frac{1}{\sqrt{2}}\{\alpha_1\beta_2 - \alpha_2\beta_1\} \tag{D.3a}$$

$$\chi_2(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2) = \psi_2(\mathbf{r}_1)\psi_2(\mathbf{r}_2) \times \frac{1}{\sqrt{2}}\{\alpha_1\beta_2 - \alpha_2\beta_1\} \tag{D.3b}$$

$$\chi_3(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) - \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\} \times \alpha_1\alpha_2 \tag{D.3c}$$

$$\chi_4(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) - \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\} \times \beta_1\beta_2 \tag{D.3d}$$

$$\chi_5(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) - \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\} \times \frac{1}{\sqrt{2}}\{\alpha_1\beta_2 + \alpha_2\beta_1\} \tag{D.3e}$$

$$\chi_6(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2) = \frac{1}{\sqrt{2}}\{\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) + \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1)\} \times \frac{1}{\sqrt{2}}\{\alpha_1\beta_2 - \alpha_2\beta_1\} \tag{D.3f}$$

On the other hand, if the spin-orbit coupling is inherently considered in the basis set[2], the spin up and down components are mixed up and cannot be treated explicitly. In this case, each wavefunction can simply be treated as independent states regardless of spin degeneracy. For instance, consider two doubly degenerate states and denote them $\xi_1^a, \xi_1^b, \xi_2^a$ and ξ_2^b, respectively. Then the two electron wavefunctions will be the anti-symmetric combinations of four single-electron wavefunctions.

$$\chi_1(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\xi_1^a(\mathbf{r}_1)\xi_1^b(\mathbf{r}_2) - \xi_1^a(\mathbf{r}_2)\xi_1^b(\mathbf{r}_1)\} \tag{D.4a}$$

$$\chi_2(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\xi_1^a(\mathbf{r}_1)\xi_2^a(\mathbf{r}_2) - \xi_1^a(\mathbf{r}_2)\xi_2^a(\mathbf{r}_1)\} \tag{D.4b}$$

$$\chi_3(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\xi_1^a(\mathbf{r}_1)\xi_2^b(\mathbf{r}_2) - \xi_1^a(\mathbf{r}_2)\xi_2^b(\mathbf{r}_1)\} \tag{D.4c}$$

$$\chi_4(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\xi_1^b(\mathbf{r}_1)\xi_2^a(\mathbf{r}_2) - \xi_1^b(\mathbf{r}_2)\xi_2^a(\mathbf{r}_1)\} \tag{D.4d}$$

$$\chi_5(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\xi_1^b(\mathbf{r}_1)\xi_2^b(\mathbf{r}_2) - \xi_1^b(\mathbf{r}_2)\xi_2^b(\mathbf{r}_1)\} \tag{D.4e}$$

$$\chi_6(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\{\xi_2^a(\mathbf{r}_1)\xi_2^b(\mathbf{r}_2) - \xi_2^a(\mathbf{r}_2)\xi_2^b(\mathbf{r}_1)\} \tag{D.4f}$$

In general, total number of two-electron wavefunctions with n single electron wavefunctions increases as ${}_n\mathrm{C}_2(\sim O(n^2))$. For deeper understanding, the reader may refer to Ref. [121].

[2] $sp^3d^5s^*$ with spin-orbit coupling

D.2 The Two-electron Hamiltonian

The two-electron Hamiltonian is constructed by two non-interacting Hamiltonians of each electron and the Coulomb interaction energy between electrons.

$$\mathcal{H} = \mathcal{H}_1(\mathbf{r}_1, \mathbf{p}_1) + \mathcal{H}_2(\mathbf{r}_2, \mathbf{p}_2) + \frac{1}{|\mathbf{r}_{12}|} \tag{D.5}$$

Generally, Eqn. (D.3) or Eqn. (D.4) (χ_i) are set as the basis of the Hamiltonian, each matrix element can be obtained by evaluating. $H_{ij} = \langle \chi_i | H | \chi_j \rangle$ which are orthogonal and make the two-electron Hamiltonian a diagonal matrix.

$$\mathcal{H} = \begin{vmatrix} 2\epsilon_1 + J_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & 2\epsilon_2 + J_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \epsilon_1 + \epsilon_2 + J_{12} - K_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & \epsilon_1 + \epsilon_2 + J_{12} - K_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & \epsilon_1 + \epsilon_2 + J_{12} - K_{12} & 0 \\ 0 & 0 & 0 & 0 & 0 & \epsilon_1 + \epsilon_2 + J_{12} + K_{12} \end{vmatrix} \tag{D.6}$$

$$\begin{aligned} J_{ij} &= \left\langle \psi_i(\mathbf{r}_1)\psi_j(\mathbf{r}_2) \left| \frac{1}{|\mathbf{r}_{12}|} \right| \psi_i(\mathbf{r}_1)\psi_j(\mathbf{r}_2) \right\rangle = \int d\mathbf{r}_1 d\mathbf{r}_2 \psi_i^*(\mathbf{r}_1)\psi_j^*(\mathbf{r}_2) \frac{1}{|\mathbf{r}_{12}|} \psi_i(\mathbf{r}_1)\psi_j(\mathbf{r}_2) \\ K_{ij} &= \left\langle \psi_i(\mathbf{r}_1)\psi_j(\mathbf{r}_2) \left| \frac{1}{|\mathbf{r}_{12}|} \right| \psi_i(\mathbf{r}_2)\psi_j(\mathbf{r}_1) \right\rangle = \int d\mathbf{r}_1 d\mathbf{r}_2 \psi_i^*(\mathbf{r}_1)\psi_j^*(\mathbf{r}_2) \frac{1}{|\mathbf{r}_{12}|} \psi_i(\mathbf{r}_2)\psi_j(\mathbf{r}_1) \end{aligned} \tag{D.7}$$

where J_{ij} and K_{ij} are Coulomb and exchange integral, respectively. If $i = j$, then $J_{ii} = K_{ii}$ from the equation. Coulomb integral originates from the electrostatic repulsion between two wavefunction and it is spin independent while the exchange integral comes from the interaction between the two wavefunctions with *same* spin to avoid each other to satisfy Pauli's exclusion principle. In other words, the exchange integral is a pure quantum mechanical outcome of stabilizing the multi-electron distribution having same spin and therefore usually has negative sign ($K_{ij} < 0$). Since the Coulomb integral is analogous to the classical expression for electrostatic potential, the value has large in magnitude and the value decays slowly even if there is no significant overlap between the wavefunctions. On the other hand, the exchange integral has cross terms involved and decays rapidly depending on the overlap of the wavefunctions, which is always smaller than Coulomb integral up to orders of magnitude. The exchange integral, however, plays a key role in manipulating spin operations in multiple qubits. The energy levels of the two-electron Hamiltonian D.6 in a double quantum dot is represented in Fig. D.1(a).[3] Instead of the ground state level of each quantum dot (ϵ_1 and ϵ_2), the ground state of two-electron level becomes $\epsilon_S = \epsilon_1 + \epsilon_2 + J_{12} + K_{12}$ and the next excited state becomes a triplet state $\epsilon_S = \epsilon_1 + \epsilon_2 + J_{12} - K_{12}$ (Fig. D.1(b)). The *exchange splitting* is defined as the difference between triplet and singlet states $\mathbf{J} = \epsilon_T - \epsilon_S = -2K_{12}$, and clearly shown to be related only to the exchange integral. The exchange splitting is directly involved in the *Heisenberg exchange Hamiltonian* describing two-electron spin operations ($\mathcal{H}_{\text{Heis}} = -\mathbf{J}\mathbf{S}_1 \cdot \mathbf{S}_2$). [122,123] Since the exchange integral is direct function of the wavefunction overlap, the *J gate* can always be found on proposals and realizations of solid-state quantum computing devices involving multiple quantum dots. [1,4,27,124]

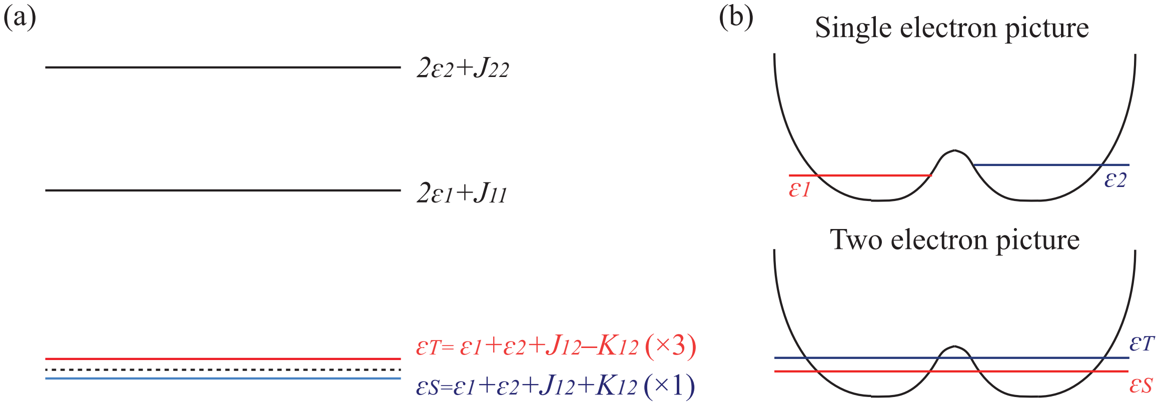


Fig. D.1. (a) Sketch of energy levels of two-electron states in the double quantum dot. ϵ_S is the ground state which involves the sum of two single electron energy levels plus the Coulomb and exchange effects. ϵ_T is the triplet excited state. (b) Comparison between the single electron and the two-electron states. In the two-electron picture, first two states should be ϵ_S and ϵ_T, instead of ϵ_1 and ϵ_2.

[3] In other two-electron systems other than weakly coupled double quantum dot, the position of each energy levels may vary. In the case of strongly confined single quantum dot with two-electrons, the ground state may become $2\epsilon_1 + J_{11}$ instead of ϵ_S.

As previously mentioned, Coulomb and exchange integral can also be used to compute exciton energies in quantum dots. [69, 117–120] The equations discussed in this section seems to be straightforward but it is not trivial when the Hamiltonian is constructed in tight-binding basis.

D.3 Integrals in tight-binding formalism

Regardless of the application, it is necessary to compute following equation for general purposes.

$$\langle ij|kl\rangle \equiv \int d\mathbf{r}_1 d\mathbf{r}_2 \psi_i^*(\mathbf{r}_1)\psi_j^*(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\psi_k(\mathbf{r}_1)\psi_l(\mathbf{r}_2) \tag{D.8}$$

where $\psi_{i,j,k,l}$ is the wavefunction computed from single electron Hamiltonian. If $i = k,\ j = l$, it will be Coulomb integral and $i = l,\ j = k$ will make exchange integral. Under tight-binding basis function each wavefunction can be written as the weighted summation over orbital components (b) including spin (s) at every atom site (n).

$$\begin{aligned} \psi(\mathbf{r}) &= \sum_{n,b,s} c(n,b,s)\phi_b^s(\mathbf{r}-\mathbf{R}_n) \\ &= \sum_{n,b,s} c(n,b,s)\phi_b(\mathbf{r}-\mathbf{R}_n)f(s) \end{aligned} \tag{D.9}$$

where $c(n,b,s)$ is the eigenvector component at site n, orbital b and spin s and $f(s) = \alpha$ for spin up and β for spin down. Assuming each orbital is real, Eqn. (D.8) is written as following

$$\begin{aligned} \langle ij|kl\rangle = &\sum_{n,b,s}\sum_{n',b',s'}\sum_{n'',b'',s''}\sum_{n''',b''',s'''} c_i^*(n,b,s)c_j^*(n',b',s')c_k(n'',b'',s'')c_l(n''',b''',s''') \\ &\times \Bigg[\int d\mathbf{r}_1 d\mathbf{r}_2\, \overbrace{\phi_b^i(\mathbf{r}_1-\mathbf{R}_n)f^i(s)}\, \underbrace{\phi_{b'}^j(\mathbf{r}_2-\mathbf{R}_{n'})f^j(s')}\, \frac{1}{|\mathbf{r}_{12}|} \\ &\times \overbrace{\phi_{b''}^k(\mathbf{r}_1-\mathbf{R}_{n''})f^k(s'')}\, \underbrace{\phi_{b'''}^l(\mathbf{r}_2-\mathbf{R}_{n'''})f^l(s''')}\Bigg] \end{aligned} \tag{D.10}$$

Note $f^i(s),\ f^k(s'')$ and $f^j(s'),\ f^l(s''')$ should retain same spin configuration, otherwise the integral is zero. Possible spin configurations are tabulated in Table D.3. Finally,

Table D.1
List of spin configurations allowed to perform the operation given in Eqn. (D.10). All other combinations are forbidden since the integral is zero due to the spin orthogonality.

$f^i(s)$	$f^j(s')$	$f^k(s'')$	$f^l(s''')$
UP	UP	UP	UP
UP	DOWN	UP	DOWN
DOWN	UP	DOWN	UP
DOWN	DOWN	DOWN	DOWN

the orbital integral component should be resolved to compute Eqn. (D.10).

$$\int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1 - \mathbf{R}_n)\phi_{b'}(\mathbf{r}_2 - \mathbf{R}_{n'})\frac{1}{|\mathbf{r}_{12}|}\phi_{b''}(\mathbf{r}_1 - \mathbf{R}_{n''})\phi_{b'''}(\mathbf{r}_2 - \mathbf{R}_{n'''}) \tag{D.11}$$

Although there are significant number of combinations of (b, b', b'', b''' and n, n', n'', n''') the integral is non-zero only on limited conditions[4]:

```
if $n == n''$ and $n' == n'''$
  if $n == n'$
    if $b == b''$ and $b' == b'''$
      $J_{b,b'} = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)$
    else if $b == b'$ and $b'' == b'''$  $(b \neq b'')$
      $K_{b,b''} = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b''}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b''}(\mathbf{r}_1)$
    else if $b == b'''$ and $b' == b''$  $(b \neq b')$
      $K_{b,b'} = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b'}(\mathbf{r}_1)$
    end
  else $/ * n0 \neq n1 * /$
    if $b == b''$ and $b' == b'''$
      Offsite Coulomb integral  $J_{off}(n, n', b, b')$
end; end; end
```

[4] The integral is zero whenever the integral is involved with more than two different orbitals or atom positions.

The final integral of $\langle ij|kl\rangle$ is computed under the spin rule in Table D.3.

```
if $n == n''$ and $n' == n'''$
   if $n == n'$
      if $b == b''$ and $b' == b'''$
         $\sum_n \sum_{b,b'} \sum_{s,s'} c_i^*(n,b,s)c_j^*(n,b',s')c_k(n,b,s)c_l(n,b',s')J_{b,b'}$
      else if $b == b'$ and $b'' == b'''$   $(b \neq b'')$
         $\sum_n \sum_{b,b',b\neq b'} \sum_{s,s'} c_i^*(n,b,s)c_j^*(n,b,s')c_k(n,b',s)c_l(n,b',s')K_{b,b'}$
      else if $b == b'''$ and $b' == b''$   $(b \neq b')$
         $\sum_n \sum_{b,b',b\neq b'} \sum_{s,s'} c_i^*(n,b,s)c_j^*(n,b',s')c_k(n,b',s)c_l(n,b,s')K_{b,b'}$
      end
   else $/*n0 \neq n1*/$
      if $b == b''$ and $b' == b'''$
         $\sum_{n,n',n\neq n'} \sum_{b,b'} \sum_{s,s'} c_i^*(n,b,s)c_j^*(n',b',s')c_k(n,b,s)c_l(n',b',s') \times J_{off}(n,n',b,b')$
end; end; end
```

Only terms left to complete the integral is to evaluate

$$J_{b,b'},\ K_{b,b',b\neq b'} \text{ and } J_{off}(n,n',b,b'),\ n \neq n'$$

which will be discussed in next sections.

D.4 Onsite orbital integrals

The onsite Coulomb and exchange integrals $J_{b,b'}$ and $K_{b,b'}$ is written for clarity.

$$J_{b,b'} = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2) \tag{D.12}$$

$$K_{b,b'} = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b'}(\mathbf{r}_1),\ b \neq b' \tag{D.13}$$

It is mentioned previously in Appendix A, the tight-binding basis functions do not have the analytic form. Instead, the basis functions are assumed to maintain the symmetry traits of each orbital ($s,\ p,\ d$). However, an analytical form for every orbital has to be assumed to compute the Coulomb and exchange integrals. According to the

Ref. [125], simple formula, or Slater's rule is proposed and applied to determine the radial components (thus the spatial extent of each orbital) of valence orbitals. The radial component of each Slater type orbital (STO) is expressed as

$$R(r) = r^{n^*-1} exp\left(-\frac{Z-s}{n^*}r\right)$$

where r is the radius of the orbital, n^* is the effective quantum number and $Z-s$ is screening constant. Depending on the atomic number (Z) and electrons filled in each shell, n^* and s can be determined by Slater's rule [125]. Parameters for widely used semiconductor materials for s, p, d and s^* orbitals are summarized in Table D.4.

The s, p, d and s^* orbitals used for evaluating Eqns. (D.12)∼(D.13) are used in the following form.[5]

$$\begin{aligned}
\phi_{s/s^*} &= r^n exp(-ar) \\
\phi_{p_x} &= r^n exp(-ar)\frac{x}{r} \\
\phi_{p_y} &= r^n exp(-ar)\frac{y}{r} \\
\phi_{p_z} &= r^n exp(-ar)\frac{z}{r} \\
\phi_{d_{xy}} &= r^n exp(-ar)\frac{xy}{r^2} \\
\phi_{d_{yz}} &= r^n exp(-ar)\frac{yz}{r^2} \\
\phi_{d_{zx}} &= r^n exp(-ar)\frac{zx}{r^2} \\
\phi_{d_{x^2-y^2}} &= r^n exp(-ar)\frac{x^2-y^2}{r^2} \\
\phi_{d_{z^2-3r^2}} &= r^n exp(-ar)\frac{z^2-3r^2}{r^2}
\end{aligned} \tag{D.14}$$

Unfortunately, the integral still cannot be computed since the integrals are multivariable integration and there is no deterministic numerical method to calculate the integral within reasonable time. For this reason, a probabilistic approach called Monte-Carlo (MC) method is applied.

[5] Normalization of each orbital is not required in Monte-Carlo method which uses probability rather than normalized value.

Table D.2
Parameters obtained from Slater's rule for semiconductor materials.

Si	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	3	3	3	3.7
$\mathbf{Z-s}$	4.15	4.15	1	1.45
P	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	3	3	3	3.7
$\mathbf{Z-s}$	4.8	4.8	1	1.6
Ge	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	3.7	3.7	3.7	4
$\mathbf{Z-s}$	6.3	6.3	1.6	1.6
Al	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	3	3	3	3.7
$\mathbf{Z-s}$	3.5	3.5	1	1.3
Ga	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	3.7	3.7	3.7	4
$\mathbf{Z-s}$	5	5	1.3	1.3
In	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	4	4	4	4.2
$\mathbf{Z-s}$	5	5	1	1.3
As	**s**	**p**	**d**	$\mathbf{s^*}$
$\mathbf{n^*}$	3.7	3.7	3.7	4
$\mathbf{Z-s}$	6.3	6.3	1.6	1.6

D.4.1 Monte-Carlo method applied to multi-variable integrals

The Coulomb integral (Eqn. (D.12)) can be rewritten as following.

$$\begin{aligned} J_{b,b'} &= \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2) \\ &= \int d\mathbf{r}_1 d\mathbf{r}_2 |\phi_b(\mathbf{r}_1)|^2|\phi_{b'}(\mathbf{r}_2)|^2\frac{1}{|\mathbf{r}_{12}|} \\ &\equiv \int d\mathbf{r}_1 d\mathbf{r}_2 f(\mathbf{r}_1,\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|} \end{aligned} \tag{D.15}$$

where $f(\mathbf{r}_1, \mathbf{r}_2) = |\phi_b(\mathbf{r}_1)|^2|\phi_{b'}(\mathbf{r}_2)|^2$ is the probability density function. $J_{b,b'}$ is actually equivalent to computing the expectation value of $\frac{1}{|\mathbf{r}_{12}|}$ which can now be solved using *biased random walk*:

```
Initialize $J = 0$
Set up initial random value $(\mathbf{r}_1, \mathbf{r}_2)$ and compute $f(\mathbf{r}_1, \mathbf{r}_2)$
Start $N$ random walk
   Generate random displacement $(d\mathbf{r}_1, d\mathbf{r}_2)$
   Compute $\mathbf{r}'_1 = \mathbf{r}_1 + d\mathbf{r}_1,\ \mathbf{r}'_2 = \mathbf{r}_2 + d\mathbf{r}_2,\ f(\mathbf{r}_1, \mathbf{r}_2)$
   if $f(\mathbf{r}_1, \mathbf{r}_2) < f(\mathbf{r}'_1, \mathbf{r}'_2)$
      $J+ = \frac{1}{|\mathbf{r}'_{12}|}$
      $\mathbf{r}'_1 \to \mathbf{r}_1,\ \mathbf{r}'_2 \to \mathbf{r}_1,\ f(\mathbf{r}'_1, \mathbf{r}'_2) \to f(\mathbf{r}_1, \mathbf{r}_2)$
   else if $\mathrm{rand}(0,1) < \frac{f(\mathbf{r}'_1,\mathbf{r}'_2)}{f(\mathbf{r}_1,\mathbf{r}_2)}$
      $J+ = \frac{1}{|\mathbf{r}'_{12}|}$
      $\mathbf{r}'_1 \to \mathbf{r}_1,\ \mathbf{r}'_2 \to \mathbf{r}_1,\ f(\mathbf{r}'_1, \mathbf{r}'_2) \to f(\mathbf{r}_1, \mathbf{r}_2)$
   else
      $J+ = \frac{1}{|\mathbf{r}_{12}|}$
   end
end
```

Exactly same procedure is used for computing the exchange integral (Eqn. (D.13)) but it is converted in different way.

$$\begin{aligned} K_{b,b'} &= \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b'}(\mathbf{r}_1) \\ &= \int d\mathbf{r}_1 d\mathbf{r}_2 |\phi_b(\mathbf{r}_1)|^2|\phi_{b'}(\mathbf{r}_2)|^2 \frac{\phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b'}(\mathbf{r}_1)}{|\phi_b(\mathbf{r}_1)|^2|\phi_{b'}(\mathbf{r}_2)|^2} \\ &\equiv \int d\mathbf{r}_1 d\mathbf{r}_2 f(\mathbf{r}_1,\mathbf{r}_2)\frac{\phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b'}(\mathbf{r}_1)}{|\phi_b(\mathbf{r}_1)|^2|\phi_{b'}(\mathbf{r}_2)|^2} \end{aligned} \tag{D.16}$$

Similarly, $K_{b,b'}$ is equivalent to finding the expectation value of $\frac{\phi_b(\mathbf{r}_1)\phi_{b'}(\mathbf{r}_2)\frac{1}{|\mathbf{r}_{12}|}\phi_b(\mathbf{r}_2)\phi_{b'}(\mathbf{r}_1)}{|\phi_b(\mathbf{r}_1)|^2|\phi_{b'}(\mathbf{r}_2)|^2}$ and the integral is obtained through the biased random walk discussed previously.

The onsite Coulomb and exchange integrals between different orbital combinations are tabulated for higher level calculation discussed in the previous section.

D.5 Offsite Coulomb integral

J_{off} can also be computed using MC method but it is very time consuming to calculate all offsite integrals explicitly especially on large quantum dot systems. Ohno proposed an analytical formula on computing the off-site integrals using on-site or nearest neighbor integrals [126].

$$J_{off}(n,b,n',b') = \frac{1}{\sqrt{(1/J_{b,b'})^2 + 0.48|\mathbf{r}_{12}|^2}} \tag{D.17}$$

The Ohno formula is very effective in replacing MC method for offsite integrals, however, since it involves multiplication, it can only be used up to certain cut-off distance. If the two orbitals are far enough such that each can be seen as a point charge, simpler Coulomb formula is used.

$$J_{off}(n,b,n',b') = \frac{1}{|\mathbf{r}_{12}|} \tag{D.18}$$

D.5.1 Comparison: MC vs. Ohno vs. $1/r$

Fig. D.2 shows how Ohno formula compares to the MC integral on the off-site silicon $s-d$ orbitals with increasing distance. It shows that Ohno formula is gives

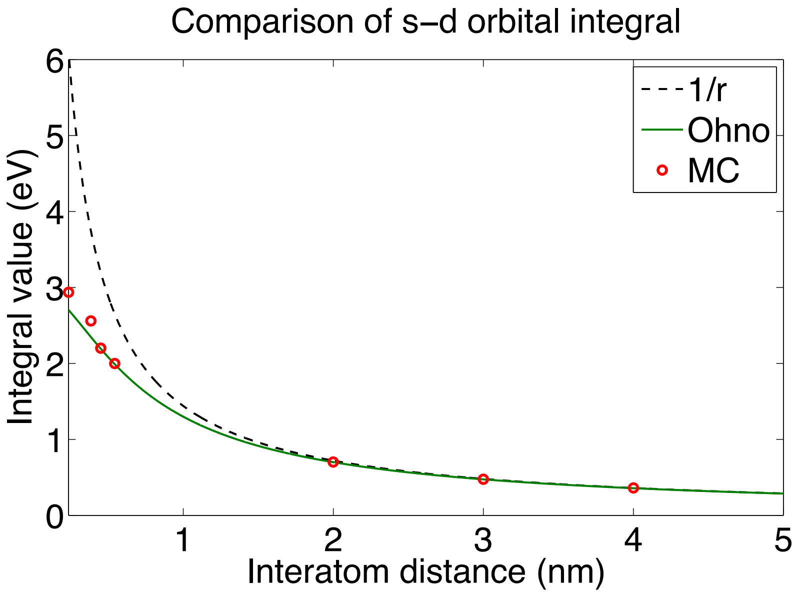


Fig. D.2. Comparison of off-site integral values of silicon $s - d$ orbitals obtained from MC method and Ohno formula (Eqn. (D.17)) as a function of distance. $J_{s-d} = 3.014$ (eV) is used obtained from the on-site MC integral. The Ohno formula agrees well with MC results around 0.5 (nm) distance.

close results to MC integral from 0.5 (nm) distance (cf. the distance between nearest neighbor is 0.235 (nm)) and is effective enough to replace MC integral for off-site Coulomb integral calculation.

Comparison between Ohno formula and $1/r$ potential in silicon is shown in Fig. D.3 on different $J_{b,b'}$ values. Regardless of the $J_{b,b'}$, Ohno formula eventually approaches simple Coulomb potential in a few nanometers, which indicates that the orbitals are far apart from each other to be considered as point charge.

In summary, the off-site Coulomb integral between orbitals required for computing $\langle ij|kl\rangle$ can be effectively approximated using Ohno formula and Coulomb potential for short and long range integrals, respectively. Simulation results shows that for short range integrals Ohno formula agrees well with MC results and MC integrals will eventually converge to $1/r$ potential. These two approximations helps to reduce

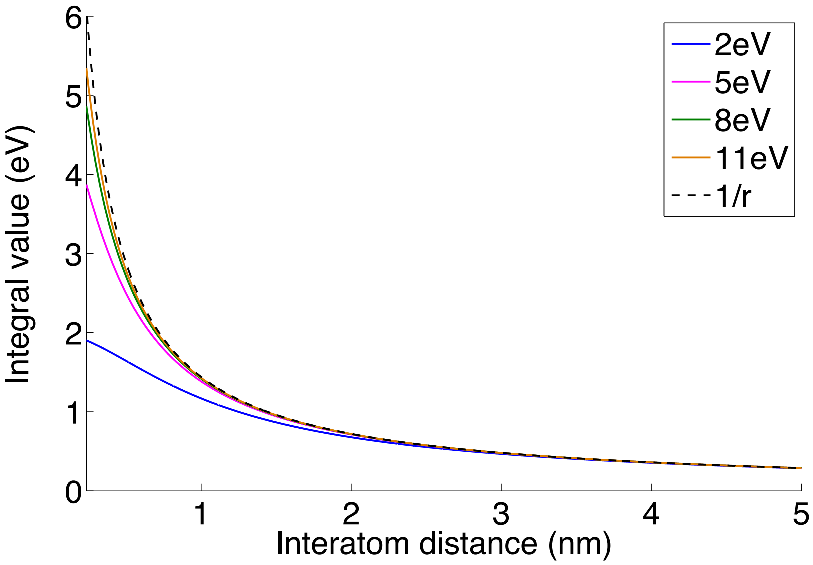
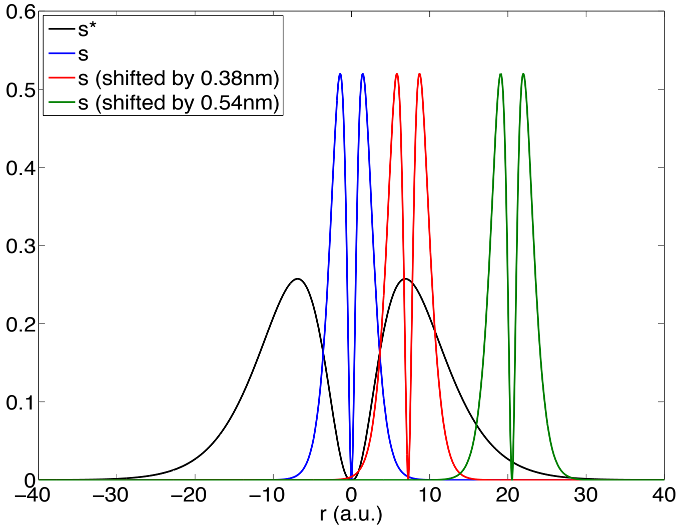


Fig. D.3. Comparison between Ohno formula and Coulomb potential $(1/r)$. Ohno formula eventually converges to the $1/r$ potential within a few nanometers.

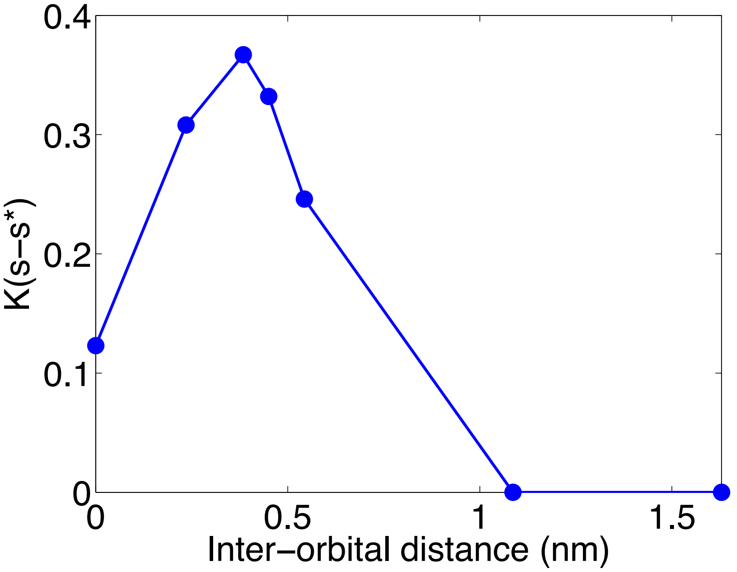
the computation time dramatically, especially on computing J/K integrals on large quantum dots.

D.6 Summary and Issues

To evaluate Coulomb and exchange integrals for constructing multi-electron Hamiltonian or computing exciton energy is non-trivial under atomistic tight-binding basis (Section D.3). The heart of the process is to compute the Coulomb and exchange integrals between localized orbitals with reasonable approximations. Slater type orbital adjusted using Slater's rule for different atoms used for constructing the analytical form of s, p, d, s^* orbitals and the two-center J/K integral is carried out with Monte-Carlo probabilistic method. For off-site Coulomb integral, Ohno formula is used for short range interaction ($\sim$ few nanometers) and long range interaction is replaced by Coulomb potential $(1/r)$.



(a)



(b)

Fig. D.4. (a) The overlap of orbitals between silicon $s - s^*$. The overlap is not maximized at the on-site. (b) The exchange integral between two orbitals are expected to increase and maximized when the orbitals are roughly 0.4 (nm) apart, which indicates that the off-site exchange integral between orbitals cannot be completely ignored.

There are a couple of issues not covered in this chapter. First, a careful examination of the dielectric screening which may significantly reduce the interaction range between off-site orbitals is not discussed [118,127]. Second, the off-site exchange term is neglected for computing $\langle ij|kl\rangle$ the procedure mentioned in Section D.3. However, depending on the extent of the orbitals constructed from Slater's rule, it is generally not true to always ignore the off-site exchange term as shown in Fig. D.4 which is heavily dependent on the overlap of the wavefunction. Considering the off-site exchange integral is expected to be small since exchange integrals are generally smaller than Coulomb integrals, but it is worthwhile to verify in the future.

VITA

VITA

Sun Hee Lee was born in Seoul, South Korea, in January 9th, 1977. He received the Bachelor of Engineering degree in Electrical and Computer Engineering from Seoul National University, Seoul, South Korea in 1999. He continued his study and received the Master of Engineering degree in Electrical and Computer Engineering from Seoul National University, Seoul, South Korea in 2001. From 2001 on, he worked for Samsung Electronics Co., Ltd., Suwon, South Korea until joining Purdue University in 2006 to pursue a Ph.D. degree in Electrical and Computer Engineering. After receiving his Ph.D. degree in December, 2011, Sun Hee Lee will be working at the Samsung Advanced Institute of Technology (SAIT), Kiheung, South Korea as a Research Scientist.