1-1-2002

Development of a Nanoelectronic 3-D (NEMO 3-D) Simulator for Multimillion Atom Simulations and Its Application to Alloyed Quantum Dots

Gerhard Klimeck
Purdue University - Main Campus, gekco@purdue.edu

Fabiano Oyafuso
Jet Propulsion Laboratory

Timothy B. Boykin
University of Alabama, Huntsville

R. Chris Bowen
Jet Propulsion Laboratory

Paul von Allmen
Motorola Labs, Solid State Research Center

Klimeck, Gerhard; Oyafuso, Fabiano; Boykin, Timothy B.; Bowen, R. Chris; and von Allmen, Paul, "Development of a Nanoelectronic 3-D (NEMO 3-D) Simulator for Multimillion Atom Simulations and Its Application to Alloyed Quantum Dots" (2002). *Other Nanotechnology Publications.* Paper 149.
http://docs.lib.purdue.edu/nanodocs/149

CMES, vol.3, no.5, pp.601-642, 2002

Development of a Nanoelectronic 3-D (NEMO 3-D) Simulator for Multimillion Atom Simulations and Its Application to Alloyed Quantum Dots

Gerhard Klimeck[1][2], Fabiano Oyafuso[2], Timothy B. Boykin[3], R. Chris Bowen[2], Paul von Allmen[4]

Abstract: Material layers with a thickness of a few nanometers are common-place in today's semiconductor devices. Before long, device fabrication methods will reach a point at which the other two device dimensions are scaled down to few tens of nanometers. The total atom count in such deca-nano devices is reduced to a few million. Only a small finite number of "free" electrons will operate such nano-scale devices due to quantized electron energies and electron charge. This work demonstrates that the simulation of electronic structure and electron transport on these length scales must not only be fundamentally quantum mechanical, but it must also include the atomic granularity of the device. Various elements of the theoretical, numerical, and software foundation of the prototype development of a Nanoelectronic Modeling tool (NEMO 3-D) which enables this class of device simulation on Beowulf cluster computers are presented. The electronic system is represented in a sparse complex Hamiltonian matrix of the order of hundreds of millions. A custom parallel matrix vector multiply algorithm that is coupled to a Lanczos and/or Rayleigh-Ritz eigenvalue solver has been developed. Benchmarks of the parallel electronic structure and the parallel strain calculation performed on various Beowulf cluster computers and a SGI Origin 2000 are presented. The Beowulf cluster benchmarks show that the competition for memory access on dual CPU PC boards renders the utility of one of the CPUs useless, if the memory usage per node is about 1-2 GB. A new strain treatment for the sp^3s^* and $sp^3d^5s^*$ tight-binding models is developed and parameterized for bulk material properties of GaAs and InAs. The utility of the new tool is demonstrated by an atomistic analysis of the effects of disorder in alloys. In particular bulk $\mathrm{In_xGa_{1-x}As}$ and $\mathrm{In_{0.6}Ga_{0.4}As}$ quantum dots are examined. The quantum dot simulations show that the random atom configurations in the alloy, without any size or shape variations can lead to optical transition energy variations of several meV. The electron and hole wave functions show significant spatial variations due to spatial disorder indicating variations in electron and hole localization.

keyword: quantum dot, alloy, nanoelectronic, sparse matrix-vector multiplication, tight-binding, optical transition, simulation.

1 Introduction

Ongoing miniaturization of semiconductor devices has given rise to a multitude of applications unfathomed a few decades ago. Although the reduction in minimum feature size of semiconductor devices has thus far exceeded every expectation and overcome every predicted technological obstacle, it will nevertheless be ultimately limited by the *atomic granularity* of the underlying crystalline lattice and the *small number of "free" electrons*. Before long, device fabrication methods will reach a point at which both quantum mechanical effects and effects induced by the atomistic granularity of the underlying medium (Fig. 1) need to be considered in the device design.

Quantum dots represent one incarnation of semiconductor devices at the end of the roadmap. Quantum dots can be characterized roughly as well-conducting, low energy regions surrounded on a nanometer scale by ”insulating” materials. The self-capacitance of the spatial confinement region is reduced with decreasing sizes. A situation can arise, in which the capacitive energy as-

[1] gekco@jpl.nasa.gov
http://hpc.jpl.nasa.gov/PEP/gekco
[2] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109
[3] Department of Electrical and Computer Engineering and LICOS, The University of Alabama in Huntsville,
Huntsville, AL 35899
[4] Motorola Labs,
Solid State Research Center,
7700 S.River Pkwy.,
Tempe, AZ 85284

sociated with adding a single electron to the system is larger than the thermal energy, and charge quantization occurs. State quantization can occur if the central region is "clean" enough[5] and if the region's dimensions are roughly on the length scale of an electron wavelength. Quantum dot implementations in various material systems (including silicon) have been examined since the late 1980's (Fig. 1b), and several designs have succeeded at room temperature operation. In particular pyramidal self-assembled quantum dot arrays appear to be promising candidates for use in quantum well lasers and detectors [Liu, Gao, McCaffrey (2001)] within a few years.

Although simulation has proven, especially in recent years, to be an important (and cost-effective) component of device design [6], existing commercial device simulators typically ignore or "patch in" the quantum mechanical and atomistic effects that must be included in the next generation of electronic devices. This document describes the development of an atomistic simulation tool, NEMO-3D, that incorporates quantum mechanical and atomistic effects by expanding the valence electron wave function in terms of a set of localized orbitals for each atom in the simulation. NEMO-3D, an extension of the successful 1D Nanoelectronic modeling tool (NEMO) [Bowen, et al. (1997a); Klimeck, et al. (1997); Bowen, et al. (1997b)], models the electronic structure of extended systems of atoms on the length scale of tens of nanometers.

Section 2 of this document elaborates on our excitement about Nanoelectronic device modeling as it bridges gap between the "large" size, classical semiconductor device models and the molecular level modeling. Theoretical, numerical, and software methods used in NEMO 3-D ,such as the theoretical background underlying the sp^3s^* and $sp^3d^5s^*$ tight-binding models; the strain computation used to determine the atomic spatial configuration; sparse matrix eigenvalue solvers and object oriented I/O; are described in detail in Section 3.

Any atomistic, 3-D, nano-scale simulation of a physically realistic semiconductor heterostructure-based system must include a very large number of atoms. For example, modeling an individual, self-assembled InAs quantum dot of 30nm diameter and 5nm height embedded in GaAs of buffer width 5nm requires a simulation domain of $40 \times 40 \times 15\text{nm}^3$, containing approximately one million atoms. A horizontal array of four such dots separated by 20nm requires a simulation domain of $90 \times 90 \times 15\text{nm}^3$, 5.2 million atoms. A $70 \times 70 \times 70\text{nm}^3$ cube of Silicon contained in an ultra-scaled CMOS device contains about 15 million atoms. The memory and computation time required to model these realistic systems, necessitates usage of parallel computers. Section 4 discusses the specific parallel implementation and parallel performance of NEMO 3-D.

The tight-binding model employed by NEMO-3D is semi-empirical in nature. Since the employed basis set is not complete in a mathematical sense, the parameters that enter the model do not correspond precisely to actual orbital overlaps. Instead, a genetic algorithm package is used to establish a set of parameters that represents a large number of physical data of the bulk binary system well. Section 5 presents the parameterization of the tight-binding models in detail.

Finally, Sections 6 and 7 discuss effects of disorder on "identical" alloyed quantum dots (i.e. quantum dots that differ only in the distribution of their constituent atoms) is presented. Significant variations in the spatial distribution of hole eigenfunctions and a spread of several meV in transition energies are demonstrated.

2 Nanoelectronic Modeling: A Problem of Conflicting Scales

Nano-scale device technology is currently a heavily investigated research field. Nanoelectronic device modeling in particular is the intriguing area where the two worlds of micrometer-scale carrier transport simulations (engineering) and nanometer-scale electronic structure calculations (solid-state physics) collide. Effects that could be traditionally safely ignored (for reasons of computational complexity) in the semiconductor device engineering world such as quantum effects and material granularity are the key ingredients in the other world. By the same token, electronic structure calculations typically do not address issues regarding carrier transport and carrier interactions with their environment for reasons of computational complexity as well. Nanoelectronic device modeling must address all of these issues at once.

[5] Clean refers to a small number of unintentional impurities and crystal defects.

[6] Physics-based device simulation tools have typically only been used to improve individual device performance after careful calibration of the simulation parameters

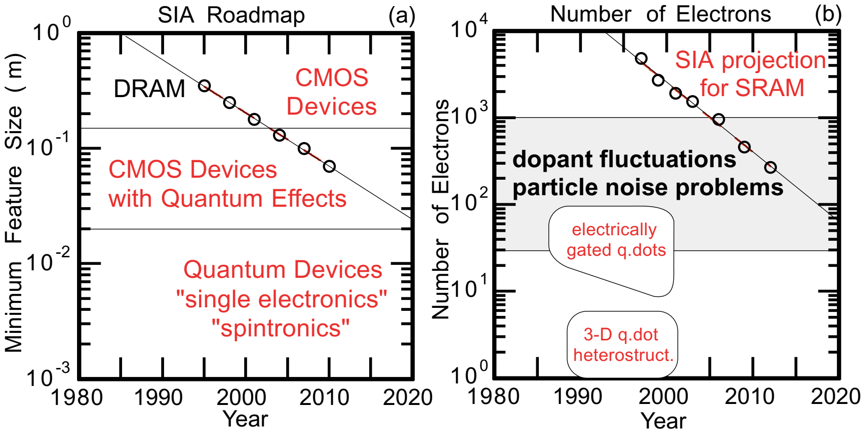

Figure 1 : (a) Minimum 2D feature size as projected on the SIA roadmap. Layer thickness od 0.01μm in the next generation devices are not captured in this graph. (b) Number of electrons under a CMOS SRAM gate. Dopant fluctuation and particle noise fluctuations may make reliable circuit design impossible, since each device may vary from the next significantly.

2.1 Top-Down Approaches

Traditionally, industrial semiconductor device research has approached nano-scale dimensions from micrometer dimensions. The object of this miniaturization is to make the current state-of-the-art devices operate faster, use less power, and perform at the same level of reliability. Commercial simulators of industrial Silicon based semiconductor devices are based on drift diffusion models, which treat electrons and holes in their respective bands as electron gases. The concept of individual electrons never explicitly appears since the electron gas is described by its density alone. Furthermore, the underlying matter is approximated by a so-called jellium with atoms represented by a uniform positive background. Effects due to interactions with impurities, phonons and other particles are included via mobility models, interaction rates, and other effective potentials.

More sophisticated and computationally much more demanding models solve the Boltzmann Transport Equation (BTE) within a Monte Carlo framework. Electrons and holes are treated as semi-classical particles moving like billiard balls in the six-dimensional phase space and interacting with their environment through adequately weighted random scattering events. The most comprehensive and commercially available simulator of this kind is DAMOCLES [7] built at IBM but other BTE based simulators have also recently appeared on the market[8][9].

The hydrodynamic approximation to the BTE has recently given rise to a class of models that is a step between the drift diffusion approach and the full-fledged BTE solver. Whereas the drift diffusion approach essentially only considers the zeroth order moment of the BTE, the hydrodynamic model extends the approximation to the first and second order moments. This treatment of higher order moments yields familiar momentum and energy conservation equations for an ideal fluid with additional terms for the electric and possibly the magnetic field. The hydrodynamic method enjoys considerable popularity since it describes hot carrier transport better than drift diffusion models yet it is significantly faster than the Monte Carlo BTE method.

An industry has evolved dedicated to the development and maintenance of such semiconductor device simulators[10] [11]. However, quantum mechanical effects such as tunneling and state quantization are not explicitly included in these models. Current efforts in the traditional

[7] See Damocles at http://www.research.ibm.com/ DAMOCLES or search for Damocles on IBM web-site, in http://www.ibm.com

[8] See MocaSim at http://www.silvaco.com/ products/ vwf/ mocasim/ mocasim_br.html or search for MocaSim on Silvaco website, in http://www.silvaco.com

[9] Search for DESSIS on the ISE, Integrated Systems Engineering web-site at http://www.ise.com

[10] See Medici at http://www.avanticorp.com/ Avant!/ SolutionsProducts/ Products/ Item/ 1,1500,192,00.html or search for Medici on the Avant! website, in http://www.avanticorp.com

[11] See Atlas at http://www.silvaco.com/ products/ vwf/ atlas/ atlas.html or search for Atlas on Silvaco web-site, in http://www.silvaco.com

device simulation community mainly focus on including these quantum mechanical effects into existing device simulation models incurring the least possible computational expense and with the overriding requirement of preserving the overall framework of the existing simulation tools. However, the problem with such simulator extensions is that they depend heavily on empirical parameterization to operate well on existing devices. The use of these tools and its parameterizations is generally not accurate for the next generation of devices.

2.2 Bottom-up Approaches

While the industry oriented semiconductor device research community approaches nano-scale transport from the top down, the physics oriented solid state research community approaches the same regime from the bottom up. The models in the latter approach are fully quantum mechanical and can only be applied to relatively small systems with emphasis on high accuracy. The systems are often periodic with unit cells containing a few hundred to a few thousand atoms, and the main output is the electronic structure and the equilibrium atomic configuration with emphasis on surface and interface reconstruction and on impurity and defect levels. Charge transport is usually not included at the fundamental level, although some attempts are mentioned below.

In contrast to the methods discussed in Section 2.1, electronic structure calculations explicitly include the granularity of condensed matter and describe the atoms at various levels of sophistication. At a fundamental level, the electrons are described by a many body Schrödinger equation in which the Hamiltonian contains interaction potentials with the atoms as well as electron-electron interaction terms. In this approach, it has already been assumed that the electrons adiabatically follow the motion of the atoms. Effects beyond this approximation lead to electron-phonon interaction terms that are evaluated in subsequent steps. In most cases, even the full electron problem is intractable, and calculations involving more than a handful of atoms rely on the so-called single electron approximation. The single electron approximation circumvents the difficulties raised by the interaction between the electrons by introducing a local or sometimes a non-local potential into a one particle Schrödinger equation. Familiar implementations of this idea are the Hartree-Fock approximation [McWeeny (1992)] and density functional theory [Hohenberg, Kohn (1964); Kohn, Sham (1965); Jones, Gunnarson (1985)]. Alternate approaches using the full Hamiltonian that explicitly includes electron-electron interaction are based on methods such as quantum Monte Carlo [Needs, Foulkes, Mitas, Rajagopal (2001)].

Within the one-electron picture, it is in some cases possible to solve the all-electron problem, which means that all the electrons in the atoms are explicitly included in the self-consistent solution of the density dependent Schrödinger equation. The atom is then simply described by a Coulomb potential with the appropriate charge for the nucleus. However, in most situations only a restricted number of electrons in the atom participate in the chemical bonding and transport properties (valence electrons). Several methods have emerged where the core electrons are taken into account by modifying the Coulomb potential of the nucleus with an additional repulsive potential, which describes the interaction of the core electrons with the valence electrons. The resulting potential is termed "pseudopotential". A number of approaches that have been explored to build these crucial components of electronic structure calculations are described below.

Pseudopotentials are divided into several classes. Empirical pseudopotentials are fitted so that a set of calculated properties match experimental results. Such empirical pseudopotentials can be defined in real space by a parameterized function or directly in reciprocal space, which offers advantages for periodic systems and was one of the first avenues explored [Cohen, Bergstresser (1966)]. The real space pseudopotentials [Appelbaum, Hamann (1973); RamanaMurty, Atwater (1995)] offer the advantage that non-bulk systems such as interfaces and surfaces can be described more realistically.

First principles pseudopotentials do not require any fitting procedure, but they do require the knowledge of the eigenstates and eigenenergies for isolated atoms. A number of schemes have been devised, most of which strive to eliminate the nodes in the valence band electronic wave functions within the core region, to reduce the computational cost of the numerical solutions. These schemes, in turn, can be divided into two categories.

Norm conserving pseudopotentials are derived (through inversion of the Schrödinger equation) from pseudo-wave functions with the reassuring property that the associated integrated charge inside the core region is identical to the charge obtained with the exact eigenfunctions. The most famous example and the most widely used for benchmarks is the method by Bachelet, Hamann and Schlüter (1982). Another method that has gained considerable popularity in conjunction with a plane wave expansion for the numerical solution was later developed by Troullier and Martins (1991). Troullier and Martins method differs by the prescription used to build the pseudo wave functions

Norm-conserving pseudopotentials require a large plane wave cutoff for elements of the first row, oxygen, and a number of other elements because the pseudo-wavefunction cannot be made sufficiently smooth in the core region. Conversely, a real space method would require a very fine mesh. Vanderbilt (1990) recently introduced a successful ultra-soft pseudopotential for which the norm-conserving constraint is relaxed. The disadvantages of Vanderbilt's method are more complex coding and the need to solve a generalized eigenvalue problem, rather than a standard eigenvalue problem.

While this document has reviewed the description of atoms with pseudopotentials it should be mentioned that a number of important issues related to improvements to the density functional theory and to the development of efficient numerical methods, which both lay at the core of other current investigations in the field, have been omitted.

Finally, as already mentioned, although earlier most electronic structure calculations using pseudopotentials are restricted to systems much smaller than the quantum dots of interest in this work, it is worthwhile noting that, with a number of approximations, Canning, Wang, Williamson, Zunger (2000) have recently managed to extend some of their pseudopotential work to systems containing up to one million atoms. Zunger's method has been applied extensively (see for example []) to model quantum dot structures, however, without yet including transport calculations.

2.3 An Intermediary Approach

Whereas traditional semiconductor device simulators are insufficiently equipped to describe quantum effects at atomic dimensions, most ab-initio methods from condensed matter physics are still computationally too demanding for application to practical devices, even as small as quantum dots. A number of intermediary methods have therefore been developed in recent years. The methods can be divided into two major theory categories: atomistic and non-atomistic.

The non-atomistic approaches do not attempt to model each individual atom in the structure, but introduce a variety of different approximations that are usually based on a continuous, jellium-type description of matter. At the lowest order approximation, such approaches only retain effective masses and band edges from the full electronic band structure, and they have given rise to the well-known effective mass approximation, in which on the scale of atomic distances a slowly varying envelope function describes the carriers. That envelope function is the solution to a one-particle effective mass Schrödinger equation. The general $k \cdot p$ method leads to a straightforward extension of that approximation by including the coupling between multiple bands. The $k \cdot p$ method has given rise to the popular multi-band effective mass approximation [Schuurmanst Hooft (1985), vonAllmen (1992a)], in which an envelope function is associated with each band explicitly included in the calculation, and a set of coupled Schrödinger-like equations is solved. It should be noted that the limitation to slowly varying perturbations remains in the multi-band version [vonAllmen (1992b)]. The different materials are described by space-dependent parameters which are separately determined for each of the materials in the device.

One strength of the effective mass approximation is the capability to discretize realistically-sized systems without the tremendous computational expense of previously mentioned ab-initio methods. However, the approximation inherently does not contain any direct atomic level information, and is, therefore, not well suited for the representation of nano-scale features such as interfaces and disorder from a fundamental perspective. This limitation has sparked lively discussions concerning the validity of the near-zone center plane wave expansion $k \cdot p$ basis and the need to include each atom in the simulation [Fu, Wang, Zunger (1998a); Fu, Wang, Zunger (1998b); Efros, Rosen (1998)]. Despite its limitations the effective mass approximation has provided excellent agreement with measurements for a large number of experiments. Another interesting issue [Keating (1966); Pryor, Kim, Wang (1998)] of particular relevance to quantum dots relates to the most appropriate treatment of strain: should continuum or atomistic models be preferred? This work uses the atomistic valence force field method by Keating (1996).

Atomistic approaches attempt to work directly with the electronic wave function of each individual atom. Ab-initio methods overcome the shortcomings of the effective mass approximation, however, additional approximations must be introduced to reduce computational costs. As described briefly in the previous section, one of the critical questions is the choice of a basis set for the representation of the electronic wave function. Many approaches have been considered, ranging from traditional numerical methods, such as finite difference and finite elements, as well as plane wave expansions [Canning, Wang, Williamson, Zunger (2000)], to methods that exploit the natural properties of chemical bonding in condensed matter. Among these latter approaches, local orbital methods are particularly attractive. While the method of using atomic orbitals as a basis set has a long history in solid state physics, new basis sets with compact support have recently been developed [Sankey (1989)], and, together with specific energy minimization schemes, these new basis sets result in computational costs which increase linearly with the number of atoms in the system without much accuracy degradation [Ordejon, Drabold, Grumbach Martin (1993),Ordejon, Galli, Car (1993)]. However, even with such methods, only a few thousand atoms can be described with present day computational resources. NEMO 3-D uses an empirical tight-binding method [Vogl Hjalmarson, Dow (1983); Jancu, Scholz, Beltram, Bassani (1998)] that is conceptually related to the local orbital method and that combines the advantages of an atomic level description with the intrinsic accuracy of empirical methods. It has already demonstrated considerable success [Bowen, et al. (1997a); Klimeck, et al. (1997); Bowen, et al. (1997b)] in quantum mechanical modeling of electron transport as well as the electronic structure modeling of small quantum dots [Lee, Joensson, Klimeck (2001)].

The underlying idea of the empirical tight-binding method is the selection of a basis consisting of atomic orbitals (such as s, p, and d) which create a single electron Hamiltonian that represents the bulk electronic properties of the material. Interactions between different orbitals within an atom and between nearest neighbor atoms are treated as empirical fitting parameters. A variety of parameterizations of nearest neighbor and second-nearest neighbor tight-binding models have been published, including different orbital configurations [Vogl Hjalmarson, Dow; (1983); Boykin, Klimeck, Bowen, Lake (1997); Boykin (1997); Boykin, Gamble, Klimeck, Bowen (1999); Jancu, Scholz, Beltram, Bassani (1998); Klimeck, et al. (2000); Klimeck, Bowen, Boykin, Cwik (2000)]. NEMO 3-D typically uses an sp^3s^* or $sp^3d^5s^*$ model that consists of five or ten spin degenerate basis states, respectively.

For the modeling of quantum dots, three main methods have been used in recent years: $k\cdot p$ [Pryor (1998); Stier, Grundmann, Bimberg (1999)], pseudopotentials [Canning, Wang, Williamson, Zunger (2000)], and empirical tight-binding [Lee, Joensson, Klimeck (2001)]. It is fair to note that each of these methods grapples with the same intrinsic difficulty: the full description of about a million interacting atoms and all of their electrons. It should also be emphasized that for most semiconductor compounds, only fragmentary experimental data exists for the band gaps and effective masses and their dependence on stress and strain. While ab-initio pseudopotential calculations beyond density functional theory do in principle predict such properties, the computational cost is high for even simple properties such as the electronic band gap [Hybertsen, Louie (1993)]. It should also be noted that effective masses, which are a crucial element in the determination of correct electronic state quantization, are rarely listed as a result of first principles calculations. On the other hand, more empirical approaches such as $k\cdot p$ and tight-binding use "quality" bulk parameterizations and can achieve good experimental comparisons in quantum dot simulations. The question, however, remains whether these parameterizations are valid in presence of variations at the atomic scale. These on-going efforts can be viewed as complementary rather than mutually exclusive competitors, and each method can greatly benefit from insightful cross-fertilization.

The perspective taken in this work is that empirical tight-binding models link the physical content of the atomic level wave functions of the pseudopotential calculations to the jellium approach of $k\cdot p$, and are the method of choice for realistic modeling of transport in quantum dot structures. Finally, as will be discussed in further detail, it should be emphasized that the quality of the empirical tight-binding results depends strongly on a good parameterization of the bulk material properties.

2.4 Nanoelectronics with Transport

Nanoelectronic device simulation must ultimately include both, the sophisticated physics oriented electronic structure calculations and the engineering oriented transport simulations. Extensive scientific arguments have recently ensued regarding transport theory, basis representation, and practical implementation of a simulator capable describing a realistic device.

Starting from the field of molecular chemistry, Mujica, Kemp, Roitberg, Ratner (1996) applied tight-binding based approaches to the modeling of transport in molecular wires. Later, Derosa and Seminario (2001) modeled molecular charge transport using density functional theory and Green functions. Further significant advances in the understanding of the electronic structure in technologically relevant devices were recently achieved through ab initio simulation of MOS devices by Demkov and Sankey (1999). Ballistic transport through a thin dielectric barrier was evaluated using standard Green function techniques [Demkov, Liu, Zhang, Loechelt (2000)], Demkov, Zhang, Drabold (2001)] without scattering mechanisms.

Conversely, starting from the field of semiconductor device simulation, various efforts have been undertaken over the past eight years to develop quantum mechanics-based device simulators that incorporate scattering mechanisms at a fundamental level. The Nanoelectronic Modeling tool (NEMO 1-D) built at Texas Instruments / Raytheon from 1993-1997 is possibly the first large-scale device simulator based on the non-equilibrium Green function technique (NEGF) to meet the challenge. Its initial objective was to achieve a comprehensive simulation of the electron transport in resonant tunneling diodes (RTDs). NEGF is a powerful formalism capable of combining tight-binding band structure, self-consistent charging effects, electron-phonon interactions, and disorder effects with the important concept of charge transport from one electron reservoir to another. The concept of electron transport between reservoirs was pioneered in a simpler approach by Landauer (1970) and Büttiker (1986), and later expanded for the NEGF formalism by Caroli, Combescot, Nozieres, Saint-James (1971) Tunneling through silicon dioxide barriers, which is a classical problem of great technological interest for the development of thin dielectrics, was studied using tight-binding models within NEMO [Bowen, et al. (1997b)] as well as in a large 3-D cell model by Städele, Tuttle, Hess (2001). Other research groups [Ren (2000); Ren et al. (2000); Ren, Venugopal, Datta, Lundstrom (2001)] have since then started to develop NEGF-based simulators to model MOSFET devices in a 2-D simulation domain. These simulations are computationally extremely intensive, and fully exploit the computing power of realistically available parallel supercomputers and cluster computers.

Quantum mechanical simulations of electron transport through 3-D confined structures such as quantum dots have not yet reached the maturity of the 1-D and 2-D simulation capabilities mentioned above. Early efforts were rate equation based [Klimeck, Lake, Datta, Bryant (1994); Klimeck, Chen, Datta (1994); Chen et al. (1994)], where a simplified electronic structure was assumed. In the related area of molecular structures, detailed studies of charge transport have recently become a hot research topic where simulations are providing an improved understanding of experimental data [Damle, Ghosh, Datta (2001); Anantram, Govindan (1998)].

NEMO 3-D focuses on the atomistic electronic structure calculation of realistically sized quantum dots at this development stage. This work is a complement to quantum dot simulations [Williamson, Wang, Zunger (2000); Wang, Kim, Zunger (1999); Stopa (1996); Pryor (1998); Stier, Grundmann, Bimberg (1999), Sheng, Leburton (2001)] performed with other methods discussed in this section. NEMO 3-D currently does *not* include carrier transport. However, the Lanczos algorithm (see Section 3.6) has been tested successfully already for non-Hermitian matrices, introduced by open boundary conditions (see Section 3.3) and the code is structured such that transport simulations can be incorporated in the future without major re-writes of the software.

3 Theoretical, Numerical, and Software Methods

3.1 Tight Binding Formulation Without Strain

Quantum dots are characterized by confinement in all three spatial dimensions so that the Hamiltonian no longer commutes with *any* of the discrete translation operators. The wave vector is hence *not* a good quantum number in *any* direction. The most natural basis for representing such a highly confined wave function is, therefore, one consisting of atomic-like orbitals centered on each atom of the crystal. Solving for the electronic structure of a quantum dot requires detailed modeling of the local environment on an atomic scale, and, therefore, introduces material considerations into the calculation.

While quantum dots may be fabricated in any number of materials systems, from an electronic structure point of view, the treatment employed mainly depends on whether the bulk lattice constants of all materials are the same. When the bulk lattice constants are the same the system is said to be lattice-matched; when they are not, the system is said to be lattice-mismatched. Lattice-matched examples include GaAs/AlAs and its alloys $\mathrm{Ga_xAl_{1-x}As}$, as well as $\mathrm{In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.47}As}$. An InAs quantum dot surrounded by $\mathrm{Al_xGa_{1-x}As}$ and an InAs or AlAs layer in a high performance $\mathrm{In_{0.53}Ga_{0.47}As/InP}$ resonant tunneling diode are examples of lattice-mismatched devices. The treatment of the two cases is necessarily somewhat different, since a matrix element of the Hamiltonian between two orbitals centered on different atoms depends, in general, on the position of the atoms. In this work the two-center approximation is made, so that only the relative position of neighboring atoms is important. In a lattice-matched system, the atoms constitute a perfect crystal with uniform unit cells; in a lattice-mismatched system, the atomic positions vary and are only semi regular. In other words, in such a system one can roughly discern unit cells, but these cells vary somewhat in size, and the atomic positions within them vary. The Keating [Keating (1966)] valence force field model described later is employed in NEMO 3-D to determine the atomic positions.

For both types of materials systems, the atomic-like orbitals are assumed to be orthonormal, following Slater and Koster (1954). Bravais lattice points can describe a crystal in a lattice-matched system:

$$\mathbf{R}_{n1,n2,n3} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \tag{1}$$

where $\mathbf{a}_i$ are primitive direct lattice translation vectors and n_i are integers. If there is more than one atom per cell, as is the case with, for example, GaAs or Si, the atoms within a cell are indexed by μ, and the location of the μ^{th} atom within the cell located at Eq.(1) is given by $\mathbf{R}_{n1,n2,n3} + \mathbf{v}_\mu$, where $\mathbf{v}_\mu$ is the displacement relative to the cell origin. The wavefunction is normalized over a volume consisting of N_i cells in the $\mathbf{a}_i$ $(i = 1,2,3)$ direction, and the state is represented as a general expansion in terms of localized atomic-like orbitals:

$$|\Psi\rangle = \frac{1}{\sqrt{N_1N_2N_3}} \tag{2}$$

$$\sum_{n_1=1}^{N_1}\sum_{n_2=1}^{N_2}\sum_{n_3=1}^{N_3}\sum_{\alpha}\sum_{\mu} C_{n_1n_2n_3}^{(\alpha\mu)}|\alpha\mu;\mathbf{R}_{n1,n2,n3}+\mathbf{v}_\mu\rangle$$

In Eq.(2), α indexes the atomic-like orbitals centered on the μ atoms within each cell (n_1,n_2,n_3). The Schrödinger equation thus appears as a system of simultaneous equations given by:

$$<\alpha\mu;\mathbf{R}_{n1,n2,n3}+\mathbf{v}_\mu|H - E|\Psi> \ = \ 0 \tag{3}$$

In Eq.(3) the matrix elements between localized orbitals are expressed as tight-binding parameters with the additional limitation of interactions to nearest neighbors. The $\mathrm{sp^3s^*}$ model of Vogl et al. (1983), as well as the $\mathrm{sp^3d^5s^*}$ model of Jancu et. al.(1998), are employed within the two-center approximation, in which the matrix elements depend only upon the relative positions of the orbitals. The expressions for the matrix elements between these types of orbitals in the two-center approximation are given by Slater and Koster (1954) as functions of the relative atomic positions.

3.2 Tight Binding Formulation With Strain

In a lattice-mismatched system several additional complications arise. First, the "cells" are no longer regularly placed so that the $\mathbf{R}_{n1,n2,n3}$ are no longer representable in a form given by Eq.(1). In a lattice-mismatched quantum dot fabricated from zincblende crystal materials, the $\mathbf{R}_{n1,n2,n3}$ are best considered as giving the location of an anion-cation pair. Likewise, in Eq.(3), the displacements now depend on both the specific "cell" and atom type, and are more correctly written as $\mathbf{v}_\mu^{n_1n_2n_3}$. These complications, though important, are rather minor and are automatically accommodated since there is no assumption of a wave-vector in any dimension in Eq.(2).

The second complication affects the nearest neighbor parameters. As mentioned above, in the two-center approximation these nearest neighbor parameters depend upon the relative atomic positions. For example, the Hamiltonian matrix element between an s-orbital centered about an atom at the origin and a p_x-orbital centered about an atom located at $d = \ell\hat{\mathbf{x}} + m\hat{\mathbf{y}} + n\hat{\mathbf{z}}$, where d is the distance between the atoms and ℓ, m, and n are the direction cosines is:

$$E_{sx} = \ell V_{sp\sigma} \tag{4}$$

Since the bond angle between atoms is no longer uniform in a lattice-mismatched system, the direction cosines vary in magnitude for different pairs of nearest neighbor atoms, even in nominally zincblende or diamond structure materials. Furthermore, the two-center parameters such as $V_{sp\sigma}$ no longer take on their ideal values as distance d between the atoms in each pair is in general different from its ideal (bulk crystal) value. The two-center parameters are assumed to scale as:

$$V_{\alpha\beta\gamma} = \left(\frac{d_0}{d}\right)^{\eta_{\alpha\beta\gamma}} V^{(0)}_{\alpha\beta\gamma} \qquad (5)$$

where for the given pair of atoms d_0 is the ideal separation, d is the actual separation, and $V^{(0)}_{\alpha\beta\gamma}$ is the ideal parameter for the orbitals involved. The exponents are chosen to reproduce known bulk behavior under conditions such as hydrostatic pressure. ¿From the work of Harrison (1999)], it is expected that most of these exponents should be approximately 2.

Also the same-site parameters are, generally, changed from their bulk values. In a lattice-matched system, however, the changes are usually small. In the $sp^3d^5s^*$ model, there may be no change at all, since in this model it is often possible to use a single set of onsite parameters for a given atom type, independent of the material. For example, As has the same parameters in GaAs, AlAs, and InAs (see Table 3).

In a lattice-mismatched system, atom displacements affect the same-site parameters more strongly. To understand the reason for this shift, recall that the atomic-like orbitals are assumed to be orthogonal. They are, thus, not true atomic orbitals, but are more properly Löwdin functions [Loewdin (1950)], which are orthogonal yet transform under symmetry operations of the crystal, as would the atomic orbital whose label they bear. When atoms are displaced in a lattice-mismatched system, not only do the tight-binding parameters of Eq.(4) change, so, too, do the overlaps of the true atomic orbitals from which the Löwdin functions are constructed. While the overlaps do not appear in an orthogonal, empirical tight-binding approach such as the one employed here, a reasonable approximation is to assume that the overlap between two nearest neighbor orbitals is proportional to their Hamiltonian matrix element divided by the sum of the vacuum-referenced onsite energies of the orbitals [Harrison (1999)] With this approximation Löwdin's formula is used to first order in the orbital overlaps to obtain an onsite Hamiltonian matrix element, which includes the effect of the displacement of the nearest neighbor atoms:

$$E_{i\alpha} \approx E^{(0)}_{i\alpha} + \sum_{j\beta} C_{i\alpha,j\beta} \frac{\left(E^{(0)}_{(i\alpha,j\beta)}\right)^2 - \left(E_{(i\alpha,j\beta)}\right)^2}{E^{(0)}_{i\alpha} + E^{(0)}_{j\beta}} \qquad (6)$$

where $E^{(0)}_{i\alpha}$ is the vacuum-referenced ideal same-site Löwdin orbital parameter for an α-orbital on the ith atom, $E_{i\alpha}$ is the shifted vacuum-referenced corresponding same-site Löwdin orbital parameter, $E^{(0)}_{(i\alpha,j\beta)}$ $\left(E_{(i\alpha,j\beta)}\right)$ the ideal (lattice-mismatched) nearest neighbor parameter between an α-orbital on the ith atom and a β-orbital on the jth atom, and $C_{i,\alpha,j\beta}$ is a proportionality constant fit to properly reproduce bulk strain behavior. The sum covers all orbitals β and atoms j that are nearest neighbors of the atom i. The difference in squared matrix elements effectively removes the onsite shift implicit in the ideal onsite parameter, and replaces it with the lattice-mismatched shift. Parameterizations of InAs and GaAs, including the strain-induced shift of the on-site elements, are discussed in Section 5.2.

3.3 Electronic Structure Boundary Conditions

The finite simulation domain that is represented in the electronic structure calculation as a sparse matrix must be terminated by physically meaningful boundary conditions. There are currently 2 kinds of boundary conditions implemented in NEMO 3-D : periodic and closed system. Periodic boundary conditions which satisfy Bloch's theorem allow for a study of the bulk properties of alloys as long as the periodicity of the domain is much larger than the largest feature size within the domain. Closed system boundary conditions terminate the bonds of the surface atoms abruptly. The dangling bonds are "passivated" with fixed potentials to avoid the inclusion of surface states in the energy range of interest. The thickness of an isolating GaAs buffer around a InAs quantum dot does influence the energy of the confined states, and the buffer size must be chosen adequately large.

Another desirable boundary condition developed in the NEMO 1-D code is the open boundary through which particles can be injected from reservoirs and through which particles can escape to reservoirs. The boundary conditions developed [Klimeck, et al. (1995); Lake, Klimeck, Bowen, Jovanovic(1997)] for NEMO 1-D were the key to the success in the transport simulations through realistically sized resonant tunneling diodes [Bowen, et al. (1997a); Klimeck, et al. (1997)] and MOS devices [Bowen, et al. (1997b)]. These boundary conditions change the character of the Hamiltonian matrix from Hermitian to non-Hermitian, and the imaginary part of the quasi-bound state eigen-energies now corresponds to the lifetime of the state in the confinement. To enable the simulation of charge transport in NEMO 3-D, an open boundary condition for the 3-D system is currently under development.

3.4 Atomistic Strain Calculation

An accurate calculation of the electronic structure within the tight-binding model necessitates an accurate representation of the positions of each atom. The atom positions in strained materials are shifted from the ideal bulk positions to minimize the overall strain energy of the system. NEMO 3-D uses a valence force field (VFF) model [Keating (1966); Pryor, Kim, Wang (1998)] in which the total strain energy, expressed as a local nearest neighbor functional of atomic positions, is minimized. The local strain energy at atom i is given by:

$$E_i = \frac{3}{16}\sum_j \left[\frac{\alpha_{ij}}{2d_{ij}^2}\cdot\left(R_{ij}^2 - d_{ij}^2\right)^2 + \sum_{k>j}^{n} \frac{\sqrt{\beta_{ij}\beta_{ik}}}{d_{ij}d_{ik}}\left(\mathbf{R}_{ij}\cdot\mathbf{R}_{ik} - \mathbf{d}_{ij}\cdot\mathbf{d}_{ik}\right)^2\right] \qquad (7)$$

where the sum is over neighbors j of atom i. Here, $\mathbf{d}_{ij}$ and $\mathbf{R}_{ij}$ are the equilibrium and actual distances between atoms i and j, respectively. Eq. 7 is included as Eq. 14 in reference [] except for some corrected coefficients. The local parameters α_{ij} and β_{ij} represent the force constants for bond-length and bond-angle distortions in bulk zinc-blende materials, respectively, and, in the absence of Coulomb corrections, are related to the bulk elastic moduli by:

$$\begin{aligned} C_{11}+2C_{12} &= \frac{\sqrt{3}}{4d_{ij}}\left(3\alpha_{ij}+\beta_{ij}\right) \qquad (8)\\ C_{11}-C_{12} &= \frac{\sqrt{3}}{d_{ij}}\beta_{ij}\\ C_{44} &= \frac{\sqrt{3}}{4d_{ij}}\frac{4\alpha_{ij}\beta_{ij}}{\alpha_{ij}+\beta_{ij}} \end{aligned}$$

In zinc-blende materials, however, these relations are modified by the inclusion of Coulomb effects due to the unequal charge distribution between the anion and cation sublattices. In this paper, α's and β's obtained by Martin (1970) to account for the Coulomb correction are used. The total strain energy is computed as the sum of the local strain energies over all atoms.

3.5 Atomistic Strain Boundary Conditions

Several boundary conditions for the strain calculation are currently implemented in NEMO 3-D. To model systems of finite extent, three boundary conditions are available: 1) the hard wall condition in which all outer shell atoms are fixed to user determined lattice constants, 2) the soft wall condition in which no atom position is fixed, and 3) the softwall boundary condition in which one atom position in the system is fixed.

To enable the simulation of bulk systems, periodic boundary conditions have been implemented. In this case the dimensions of the fundamental domain and, therefore, the separations between neighboring boundary atoms are not known a priori. Thus, the crystal is allowed to "breathe" such that the strain energy is also minimized with respect to the period in each direction in which periodic boundary conditions are applied.

3.6 Eigenvalue Solution

One simulation objective is to solve the eigenvalue problem for low lying electron and hole states near the bandedge. The nearest neighbor tight-binding Hamiltonian can be represented in a sparse matrix. A one million atom system represented in the sp^{3}d^{5}s* basis establishes a matrix size of 20 million × 20 million. A "direct solver", in which the entire column space is worked on is completely unfeasible for a variety of reasons, especially due to the full matrix storage requirement of $(20\times10^6)^2\times16$ bytes$=$6400TB. A variety of sparse matrix eigenvalue and eigenvector algorithms have been developed, some of which are available publicly[12]. Most of these eigenvalue/vector algorithms are some form of a Krylov/Lanczos/Arnoldi subspace approach [Gloub, Van Loan (1989)]. These methods approximate the solution on a small subspace which is increased until a desired tolerance is achieved. One the major advantage is that only require memory of the order of the length of several eigenvectors is required. At the lowest level of the algorithm, trial vectors are repeatedly multiplied by the

[12] See ARPACK at http://www.caam.rice.edu/ software/ ARPACK/ index.html

matrix of interest. Storage of the matrix is not mandatory if the matrix can be reconstructed on the fly during the matrix-vector multiply process. The performance of these algorithms operating on large systems, therefore, strongly depends on the efficient implementation of a matrix-vector multiply algorithm for the problem at hand.

The Lanczos-based solver technology of non-Hermitian matrices developed [Bowen, Frensley Klimeck, Lake (1995)] for NEMO 1-D was applied for NEMO 3-D as well. Early in the development of NEMO 3-D , the Lanczos-eigenvalue solver prototype with was compared ARPACK. For a system of about 100,000 atoms it was found that our custom solver was significantly faster[13] than ARPACK. Therefore, parallelization of our custom solver was implemented to attack large-scale problems.

The folded-spectrum method [Wang, Zunger (1994)], which is based on a minimization of the squared target matrix, has been proposed, implemented, and heavily used by Zunger et al. Before the matrix is squared it is shifted to the energy range of interest, i.e. close to the expected eigenenergies. The overall algorithm is then based on a conjugate gradient minimization of a trial vector. This method also relies heavily on a matrix-vector multiply algorithm and it has been implemented in NEMO 3-D.

3.7 Software Methods

The NEMO 3-D project leverages some of the software technology developed in the original NEMO 1-D project [Blanks, et al. (1997); Klimeck et al. (1997)] as well as improvements of NEMO 1-D undertaken at JPL[14] [Klimeck (2002)]. NEMO 1-D contains roughly 250,000 lines of C, FORTRAN and F90 code. Data management is performed in an object oriented fashion in C, without using C++. On the lowest level, FORTRAN and F90 are used to perform small matrix operations such as matrix inversions and matrix-vector multiplication. The language hybrid structure was introduced to utilize fast FORTRAN and F90 compilers that were available on the SGI, HP, and Sun development machines in the early stages of NEMO 1-D. At that time identical algorithms written in FORTRAN and C showed that FORTRAN could outperform C by about a factor of 4. On today's Intel cluster based computers such a speed discrepancy may not really exist anymore in part due to the advancements in C compilers and the lack of competition for fast FORTRAN compilers.

One major software component in NEMO 1-D is the representation of materials in a tight-binding basis including various orbitals and nearest neighbor counts. Adding a new tight-binding model amounts to adding a new Hamiltonian constructor. Bulk band structure and charge transport calculations are almost independent of the underlying Hamiltonian details and form a higher level building block by themselves. This modular design enables the introduction of more advanced tight-binding models as they become available, without interfering with higher level algorithms. The $sp^3d^5s^*$ model has been added at JPL recently within this architecture.

A hierarchically higher software block in NEMO 1-D accesses the bulk bandstructure routines through a script-based database module. The ASCII database can be modified outside the NEMO 1-D core to contain arbitrary tight-binding input parameters as well as a variety of different database entries. The relatively simple database access to bulk bandstructure has enabled a straight-forward integration of NEMO into a genetic algorithm based optimization tool. This tool is used for tight-binding parameter optimization as discussed on Section 5.1. The material parameter database is also accessed in the new NEMO 3-D code.

Most research oriented simulators must be fed a wash list of parameters, some of which are dependent on others, some of which may be superfluous, or some of which may cause crashes unless some other options have been set. Often these dependencies require an expert user increasing the initial barrier to simulator usage. The NEMO 1-D input has been structured hierarchically such that the user can provide information in automated dependent blocks. Information is, therefore, requested from the user as a progressively dependent input. Such input presentation is customary in a properly implemented in a graphical user interface (GUI).

Such well organized user input presentation is relatively simply incorporated with a static GUI in software whose input is well specified. Research software under rapid development, however, tends to change its requirements frequently. Rapid changes force a static GUI to always

[13] We speculate that this is in part due to our utilization of the Hermiticity of H.

[14] JPL Technical Report, ”NEMO Benchmarks on SUN, HP, SGI, and Intel Pentium II”. http://hpc.jpl.nasa.gov/ PEP/ gekco/ parallel/ benchmark.html

lag behind the actual theory software that it operates. Such static design also creates a maintenance nightmare, since new options must be added at two places independently, in the theory code and the GUI. Such issues are addressed in NEMO 1-D and NEMO 3-D in a way that is at least novel in the electronic device simulator and electronic structure simulator field. The input groups are formulated as hierarchical C data structures that are used by the theory code as well as the GUI. The input structures are formatted by translator functions into user-friendly and storage-friendly representations, such as windows and html-like text, respectively. With the translators in place GUI options are generated dynamically from the data structures that are determined by the requirements of the theory code. The theory programmer can add more options and data structures as needed, without concern for the representation of that information to the user or the transfer of it in and out the simulator. With the design of the data structure translators the development of the GUI and the theory code are essentially decoupled, and GUI, theory, and numerical developers can work on their respective blocks of code independently.

The input/output design has been presented in some detail in reference []. In NEMO 3-D this approach has been generalized significantly. The architecture of the threading of the various input/output options and data structures has been implemented in NEMO 3-D as an object oriented, table-based inheritance. Options that require more input are associated with the creation function of that child data structure. As the user input is translated into the content of the data structure, new creation functions are put on the stack of non-entered user input. User input is requested until the stack of required user input is empty. This object-oriented input completely precludes “if ... then ... else” input parsing in NEMO 3-D.

To tackle the data management on the various cluster computers in the High Performance Computing (HPC) group at JPL a Tcl/Tk client-server based interface was built. This interface works with NEMO 3-D and other completely independent simulators such as genetic algorithm-based optimization tools entitled GENES (Genetically Engineered Nanostructured Devices)[Klimeck, Salazar-Lazaro, Stoica, Cwik (1999) and EHWPack (Evolvable Hardware Package) [Keymeulen et al. (2000)]. To improve the generality of this approach and to enable a web-based treatment of the overall device simulation on a remote computing cluster a JAVA / XML based approach [15] is currently developed.

4 Numerical Implementations and Parallel Performance

4.1 Hardware and Software Specifications

The performance of the parallelized eigenvalue solver and strain minimization algorithm implemented in NEMO 3-D is benchmarked on four different parallel computers. Three of these computers are commodity PC clusters (Beowulf) of various generations, and the fourth one is a shared memory SGI Origin 2000. The three Beowulf clusters (P450, P800, and P933) are based on Intel Pentium III processors running at 450MHz, 800MHz, and 933 MHz in various memory, CPU, and network configurations. Details are shown in Table 1. The P800 has two networking systems that can operate simultaneously: 1) the standard 100Mbps Ethernet, and 2) the advanced, low latency, high bandwidth (and high breakdown experience) 1.8Gbps Myricon network[16]. Most of the benchmarks discussed here are based on the P800 performance. The other machines are used to analyze issues of memory latency and speed increase with increased clock and communication speed. *Hyglac*, the grandfather of Beowulf clusters was built in the High Performance Computing (HPC) Group at JPL by Thomas Sterling et al. in 1997 and it won the the Gordon Bell prize for lowest Cost/Performance at Supercomputing 1997. *Hyglac* is based on a cluster of 16 200MHz Pentium Pro processors with 128MB RAM each. JPL’s HPC group continued to push on Beowulf computers and is currently focused on the use of high-speed networks with real world MPI applications and large memory usage.

All of the parallel algorithms discussed in this paper are implemented with the message passing interface (MPI) [Gropp, Lusk, Skjellum (1997); Gropp, Lusk (1997)]. The SGI has its own proprietary implementation of MPI which utilizes the fast SGI interconnect as well as the shared memory within one 4-CPU board.

Various MPI/MPICH [Groupp, Lusk, Skjellum (1997); Gropp Lusk (1997)] releases have been installed on the hardware in Table 1 throughout the last three years. On the dual CPU Beowulf, the shared memory versus distributed memory configurations of MPICH have been

[15] See WIGLAF at http://ess.jpl.nasa.gov/ subpages/ reports/ 01report/ WIGLAF/ WIGLAF-01.htm

[16] See Myricom, in http://www.myricom.com

Table 1 : Specifications of the parallel computers used in this work.

Name	CPU	$\frac{\text{Clock}}{\text{MHz}}$	$\frac{\text{RAM}}{\text{node GB}}$	$\frac{\text{Bus}}{\text{MHz}}$	$\frac{\text{CPUs}}{\text{node}}$	Nodes	CPUs	$\frac{RAM}{GB}$	Network	Purchase	Motherboard
SGI	R12000	300	2		4	32	128	64		1998	SGI Origin 2000
P450	PIII	450	0.512	100	1	32	32	16	100Mbps	1999	Shuttle Intel 440BX chipset
P800	PIII	800	2	133	2	32	64	64	100Mbps, 1.8Gbps	2000	Supermicro 370DLE, Intel LE chipset
P933	PIII	933	1	133	2	32	64	32	100Mbps	2001	Supermicro 370DL3, Intel LE chipset

examined for their relative performance. Small performance increases due to the shared memory / reduced communication cost have been found in the electronic structure calculation. Even if the shared memory option is turned off, the communication from one CPU to the other on the same board is faster than to a CPU off-board. Apparently the network card relays the communication back to the on-board CPU without actually sending the message to the switch. A disadvantage of the shared memory implementation is the a priori determination of a maximum message buffer size as an environment variable before the software is executed. The simulation will fail if the simulation exceeds that maximum communication buffer size. Due to this static handicap and the minimal performance increase, the non-shared memory model is typically chosen.

Parallelization efficiency using OpenMP has been explored in the early stages of the development process as an enhancement to MPI. The objective is to communicate from CPU board to CPU board with MPI and within a board with OpenMP and shared memory. In the example algorithms that have been explored the creation and destruction of threads using OpenMP were found to cause a significantly large overhead such that the parallel efficiency was unsatisfactory. For that reason the combined MPI and OpenMP approach was abandoned. OpenMP was not pursued as an overall parallel communication scheme across the cluster, since no reliable cluster-based OpenMP compilers were available.

4.2 Parallel Implementation of Sparse Matrix-Vector Multiplication

The numerically most intensive step in the iterative eigenvalue solution discussed in Section 3.6 is the sparse matrix-vector multiplication of the matrix H and the trial vector $|\Psi_n>$. For example, the matrix-vector multiplication of the tight-binding Hamiltonian in a 1 million atom system with 4 neighbors per atom in a 10 orbital, explicit spin basis ($sp^3d^5s^*$) requires roughly 5 million full 20×20 complex matrix-vector multiplications. This corresponds to $5\times10^6\times400=2\times10^9$ complex multiplications or roughly 8×10^9 double precision multiplications and 4×10^9 additions. The single matrix-vector multiplication step can, therefore, be estimated as $8\times10^9+4\times10^9=12$ Gflop. In the sp^3s^* basis used in the benchmarks shown in Section 4.4 the operation count is reduced by a factor of 4 to about 3 Gflop. These estimates exclude overhead for the sparse matrix reconstruction, memory alignment, and construction of the fully assembled target vector $|\Psi_{n+1}>$. With an expected iteration count in the Lanczos algorithm of 2×5000, a total number of operations of 30 Tflop and 120 Tflop are anticipated for the sp^3s^* and $sp^3d^5s^*$ model, respectively. With a single CPU operating at 0.5 Gflops, such computations continue through 0.7 and 2.8 days, respectively. Actually, 0.5 Gflops appears to be a high estimate for sustained computational throughput on the latest 2 GHz Pentium 4 chips. Three years ago, when this project was initiated, peak performance was about a factor of 5 slower. The reduction in wall clock time for the completion of such a computation is highly desirable. This is particularly true for systems in excess of ten million atoms.

The 3 to 12 Gflop needed to perform a single matrix-vector multiplication correspond to 3 or 12 seconds on a single 0.5 Gflop machine. This load is large enough to warrant parallelization on multiple CPUs. For implementation on a distributed memory platform, data must be partitioned across processors to facilitate this fundamental operation. For good load balance, the device is partitioned into approximately equally sized sets of atoms, which are mapped to individual processors. Because only nearest neighbor interactions are modeled, a naive partition of the device by parallel slices creates a mapping such that any atom must communicate with neighbors that are, at most, one processor away.

This scheme, shown in Figure 2a), lends itself to a 1D chain network topology, and results in a block-tridiagonal Hamiltonian for non-periodic boundary conditions in which where each block corresponds to a pair of processors, and each processor holds the column of blocks associated with its atoms (Figure 2b). The gray squares in the corners symbolize fill-in regions due to periodic boundary conditions. Communication cost, roughly proportional to the boundary separating these sets, scales only with surface area ($O(n^{2/3})$) rather than with volume ($O(n)$), where n is the number of atoms. In a matrix-vector multiplication, both the sparse Hamiltonian and the dense vector are partitioned among processors in an intuitive way; each processor p, holds unique copies of both the nonzero matrix elements of the sparse Hamiltonian associated with the orbitals of the atoms mapped to processor p and also the components of the dense vector associated with atomic orbitals mapped to p. The matrix-vector multiplication is performed in a column-wise fashion as shown in Fig. 2b). That is, processor j computes:

$$y_{i,j} \quad = \quad H_{i,j}x_j \qquad\qquad (i = j, j \pm 1) \tag{9}$$

where $H_{i,j}$ is the block of the Hamiltonian associated with nodes i and j, and x_j are the components of x stored locally on node j. There are three results generated by the multiplication on processor j: the diagonal components $y_{j,j}$, which are needed locally by processor j; and two off-diagonal components $y_{j-1,j}$ and $y_{j+1,j}$, which need to be communicated to processors $j-1$ and $j+1$, respectively. Within the same scheme processors $j-1$ and $j+1$ share one of their off-diagonal results with processor j. This scheme lends itself to a two-step communication process.

In the first step or the two-step process all even numbered CPUs, $2n$, communicate to the CPU "to the right", $2n+1$. All odd numbered CPUs, $2n+1$, issue a communication command to CPUs, $2n$. This communication is issued with the MPI command `MPI_SendReceive`, which can be implemented in the underlying MPI library as a full duplexing operation. That means that once the communication channel is established, which can take a significant time on a standard 100 Mbps Ethernet, the information packages can be exchanged in both directions simultaneously. In the second communication step all even numbered CPUs, $2n$, communicate to the "left", $2n-1$. Simultaneously all odd CPUs $2n-1$ communicate to the even CPUs $2n$. Within this communication scheme collisions between messages do not occur and messages do not accumulate on one CPU while other CPUs wait for the completion of the communication[17].

The message size can be reduced by a compression scheme, since most of the off-diagonal blocks are zero. The sparse structure of the blocks depends on the particular crystal structure in question. In practice a sufficient fraction of zero rows exists such that compressing the matrix-vector multiplication by removing structurally guaranteed zeros is worthwhile despite the additional level of indirection required to track the non-zero structure.

The 1-D decomposition scheme performs well when the ratio of the number of atoms on the surface of the slab to the total number of atoms in the slab is small. As the number of CPUs in the parallel computation increases, for a given problem size, the surface to volume atom ratio increases to a limit of one, and the communication to computation ratio increases as well. Spatial decomposition schemes more elaborate than the 1-D scheme presented here can be implemented. One example is the 3-D decomposition in small cubes. Such schemes would probably enable the efficient participation of more CPUs in the computation; however such schemes come with immediately increased communication overhead, as six, since each CPU must exchange data with six rather then two ”surrounding” CPUs. Sections 4.4-4.8 explore the scaling of the simple 1-D topology parallel algorithms

[17] Only if periodic boundary conditions are applied with an odd number of CPUs in the MPI run one needs three communication cycles due to a conflict at the first and the last CPU communication.

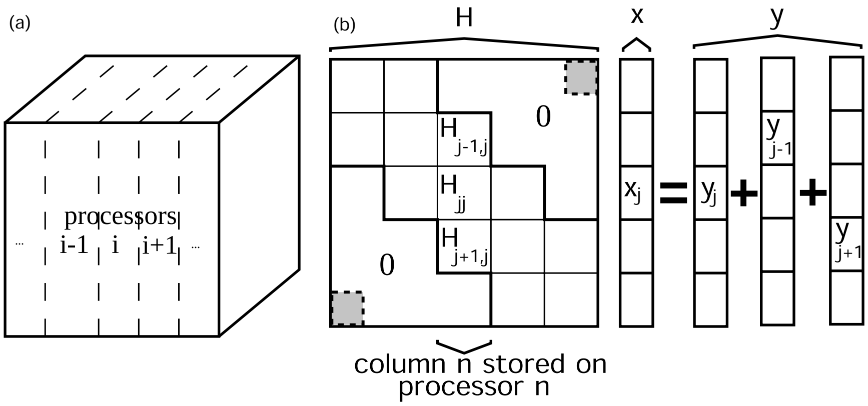

Figure 2 : (a) The device is decomposed into slabs (layers of atoms) which are directly mapped to individual processors. The gray blocks in the corner indicate the optional filling due to periodic boundary conditions. (b) Example matrix-vector multiplication on 5 processors performed in a column-wise fashion, so that the j^{th} block column and section x_j are stored on processor j. The nearest neighbor model with non-periodic boundary conditions guarantees that the Hamiltonian is block-tridiagonal, so that communication is performed only with nearest neighbor processors.

and show reasonable scaling for the mid-size clusters that are available at the High Performance Computing Group at JPL.

4.3 Hamiltonian Storage and Memory Usage Reduction

The first NEMO 3-D prototypes were focused on the generality of the tight-binding orbitals and explored the reduction of the memory requirements to simulate realistically sized structures of several million atoms. The memory requirement for storing the sparse matrix tight-binding Hamiltonian for a 1 million atom system in a 10 spin-degenerate orbital basis can be estimated as 10^6 $atoms \times 5$ $diagonals \times (20 \times 20$ $basis) \times 16 bytes/2(forHermiticity) = 16$ GB. Additional memory storage is needed for atom positions, eigenvectors, etc; therefore the 16 GB available in the P450 is inadequate.

If the system of interest is unstrained, as is the case for free standing quantum dots [Lee, Joensson, Klimeck (2001)], the memory requirement is reduced dramatically, since only a few uniquely different neighbor interactions need to be stored. The overall Hamiltonian can be generated from the replication of the few unique elements. Since immediate interest was focused on solid-state implementations on a bulk substrate, such simplifications were not in the immediate development path and they have not yet been implemented in NEMO 3-D . However, such a scheme was pursued in the NEMO 1-D transport code where the memory storage was arranged such that the Hamiltonian matrix elements fit completely into cache memory. This scheme allowed the rapid computation of the transport kernel [Bowen et al. (1997)] using the recursive Green function algorithm which scales linearly with the order N of lattice sites. The resulting computation time for a single energy pass through the whole Hamiltonian is so small, that the parallelization of the computation of a single transport kernel element cannot be parallelized efficiently [Klimeck (2002)].

The individual tight-binding Hamiltonian construction can be formulated as a table look-up operation, which is not, in principle, time consuming, except for the scaling of the nearest neighbor coupling elements due to strain (Eqs. 5 and 6). Therefore, the first implementation of the matrix-vector multiplication does not store the Hamiltonian, but re-computes the Hamiltonian on the fly in each multiplication step.

Hamiltonian storage became more feasible for million atom size systems when P800 with its 64 GB of total memory came on-line in the year 2000. The first Hamiltonian storage implementation stores the entire block of size $basis \times basis$ for each atom and its neighbor interactions. This storage scheme preserves the generality of the code and the independent choice of number of orbitals. Timing experiments similar to those presented in Section 4.4 show that the speed increase due to Hamiltonian storage is surprisingly small on the Beowulf systems, but is significant on the SGI. The low speed increase on the Beowulf may be associated with memory latency issues of the Pentium architecture. A further reduction in memory usage is, therefore, desirable.

A more detailed analysis of the sp^{3}s* and sp^{3}d^{5}s* Hamiltonian blocks provides insight into the memory allocation actually needed to store the Hamiltonian. The diagonal blocks are only filled on their diagonal and on a small number of off-diagonal sites. These off-diagonals are in general complex and describe the spin-orbit coupling of the spin-up and the spin-down Hamiltonian blocks. The off-diagonal blocks of the Hamiltonian can be separated into a smaller spin-up and spin-down components which are identical and real. This symmetry can be exploited to reduce the Hamiltonian storage requirement by a factor of 8 for both the sp^{3}s* and the sp^{3}d^{5}s* models. A priori knowledge on which matrix elements are real and which are complex can be utilized to increase the speed of the custom matrix-vector multiplication. A speed increase due to the compact storage scheme of slightly over 5 compared to the original storage scheme has been observed. This custom storage and matrix-vector multiplication scheme is used in the benchmarks in this paper when the Hamiltonian is stored. The utilization of C data management and the simple explicit access to real and imaginary elements of complex numbers leads to significantly faster small matrix-vector multiply algorithms in C compared to FROTRAN or F90.

4.4 Lanczos Scaling with CPU Number

This section describes the performance analysis of 30 Lanczos iterations on P800 in a variety of load distribution and memory storage schemes as a function of utilized CPUs. The execution time for seven different systems consisting of 1/4 to 16 million atoms for a Hamiltonian matrix that is reconstructed at each matrix-vector-multiplication step is shown in Figure 3a). The sp^{3}s* model is used in these simulations, resulting in 10×10 Hamiltonian matrix sub-blocks. In the 1 million atom system case, the problem is equivalent to a matrix of $10^7 \times 10^7$, and the myricom communication path is utilized. The nearest neighbor CPU communication limitation (discussed in Section 4.2) limits the 1/4, 1/2, 1, and 2 million atom systems to a maximum number of parallel processes to 32, 40, 51, and 63, respectively. The 4, 8, and 16 million atom systems cannot run on a single CPU, because the single CPU RAM on P800 would be exceeded. Even without Hamiltonian storage, these larger systems require at least 2, 10, and 16 CPUs, respectively, to avoid swapping.

Since P800 consists of 32 dual CPU nodes, a variety of loading schemes are possible in the distribution of MPI processes to the various CPUs. Figure 3a) explores two schemes: 1) dashed lines with crosses - one process per node (1 CPU idle), and 2) solid lines with circles - two processes per node (both CPUs active). Although the single process per node distribution incurs an increased cost in communication off the node, the overall computation time is slightly less when compared to the 2 processes per node case, for system sizes 1/4 - 4 million atoms. Larger systems (8 and 16 million atoms) produce a significantly better performance with the 1 process per node configuration. It appears more efficient to leave one CPU idle and utilize all the memory on board, rather than use all the CPUs and share the memory between two CPUs on the same board. This behavior can be associated with a memory latency / competition problem, and it is examined further below.

The green dashed lines in Figure 3a) indicate perfect scaling for the 1 and 4 million atom system sizes. An increasing deviation from ideal scaling is observed with an increased number of CPUs. However, the computation time is still reduced when the number of CPUs is increased. Figure 3b) shows the efficiency computed as the ratio of ideal time and actual time (1 and 4 million atom systems in red and blue, respectively). A serial to parallel code ratio of 1.6% can be extracted if the 1 million atom, two processes per node efficiency curve is fitted to Amdahl's law. This ratio indicates a high degree of parallelism in the code.

The reconstruction of the Hamiltonian matrix at each matrix-vector-multiplication step saves memory, but does require additional computation time. The performance of the matrix-vector-multiplication step can be improved through Hamiltonian matrix storage and the utilization of the sp^{3}s* and sp^{3}d^{5}s* Hamiltonian submatrix symmetries (see Section 4.3). Analogous to Figure 3a), Figure 3c) shows the parallel performance in the case of Hamiltonian storage similar.

With the increased storage requirements, the minimum number of CPUs required for the swap-free matrix-vector multiplication for systems containing 1, 2, 4, and 8 mil-

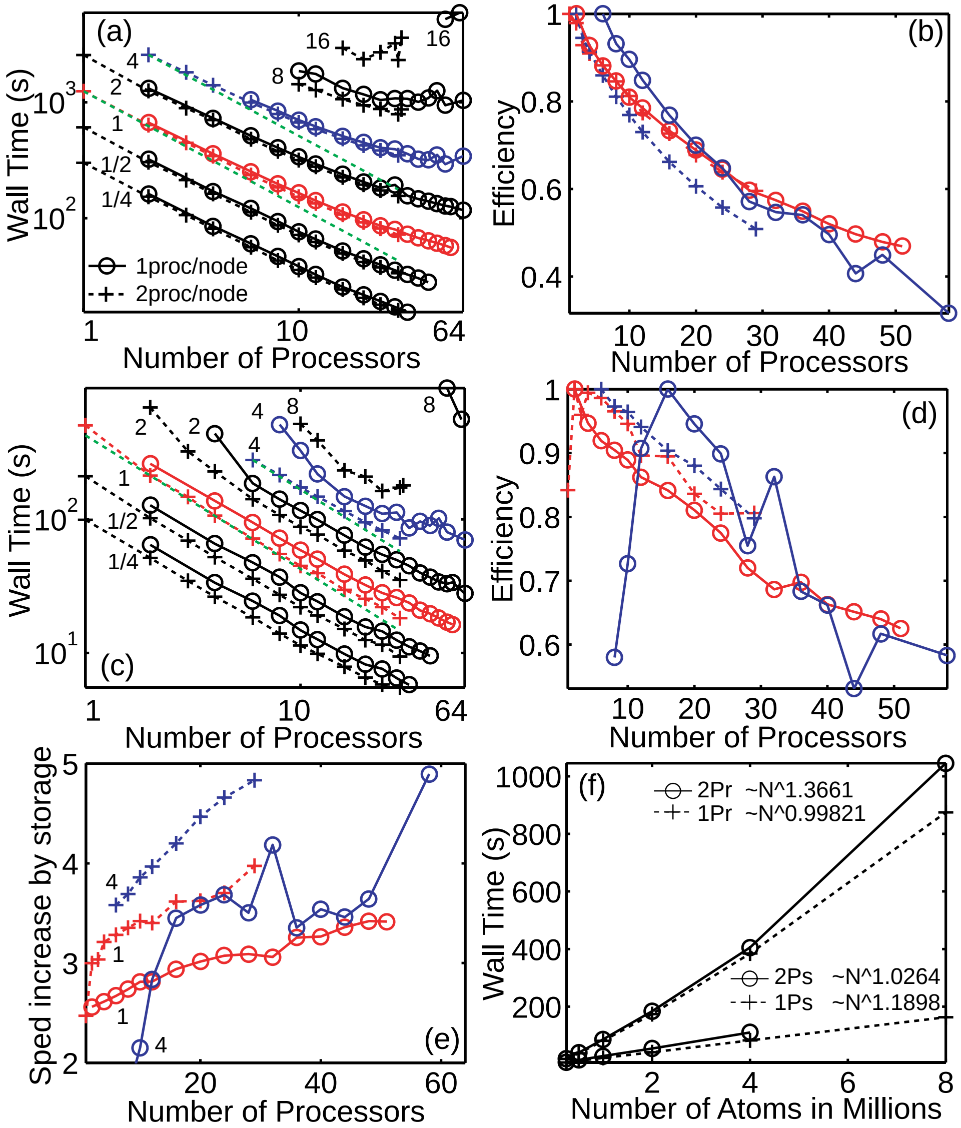

Figure 3 : (a) Execution time of 30 Lanczos iterations P800. The green dashed lines illustrate ideal scaling. Solid line: 2 processes per node (2Px), dashed line 1 process per node (1Px). First row: recomputed Hamiltonian (x=r), Second row: stored Hamiltonian (x=s). (b) Efficiency as defined as the ratio of actual compute time to ideal compute time. (c) and (d) similar to (a) and (b) except the Hamiltonian in *not* recomputed in each step, but stored in the first step. (e) Speed-up due to Hamiltonian storage for 1 and 4 million atom systems. (f) Execution time on 24 processors as a function of system size.

lion atoms increases to 2, 4, 6, and 16 CPUs from 1, 1, 2, and 10 CPUs. The 16 million atom system no longer fits onto P800. With the increased memory requirement, the distribution of processes onto different compute nodes becomes much more critical, even for smaller problem sizes. This result indicates clearly that the 2 CPUs on each motherboard compete for memory access at a significant performance cost. It appears to be more efficient to place a single process on each node for system sizes that are larger than about 4 million atoms when the Hamiltonian is stored, compared to 8 million atoms when the Hamiltonian is reconstructed. The 8 million atom simulation incurs dramatic performance losses if run on 2 processes per node, similar to the 16 million atom case without Hamiltonian storage shown in Figure 3a).

Figure 3d) shows a greater parallel efficiency of the stored Hamiltonian algorithm versus the recomputed Hamiltonian algorithm of Figure 3b). However, the point of ideal performance increases from 1 CPU since the problems no longer fit onto a single CPU. Comparing the ideal scaling indicated by the green lines in Figure 3a) and (c) shows that the stored Hamiltonian algorithm scales better with an increasing number of CPUs. This observation contradicts the expectation that a more CPU intensive calculation such as the slower recomputed Hamiltonian algorithm should scale better than a lower intensity job such as the faster stored Hamiltonian algorithm. At this time an explanation why the stored Hamiltonian algorithm scales better than the recomputed Hamiltonian algorithm is not available.

Figure 3e) shows the speed increase due to Hamiltonian storage for a system of 1 and 4 million atoms derived from the data shown in Figure 3a) and (c). Both system sizes show a greater speed increase when one process rather than two resides on a node. The speed increase due to storage is not constant, but increases with an increasing number of CPUs. The total memory used per CPU decreases with an increasing number of participating CPUs. This memory reduction reduces the competition for memory access and the speed increase curves increase with increasing number of CPUs. Competition for memory between the 2 processes on a single node with 2 CPUs is again visible.

With an estimate of 3 Gflop for a single matrix-vector multiplication in a 1 million atom system (see Section 4.2), the execution time of about 1247 seconds for 30 iterations in Figure 3a) on a single CPU, a operation rate of 0.07 Gflops is obtained. Using 24 CPUs and 81 seconds the operation count is 1.1 Gflops. For the largest achievable 16 million atom system running on 20 CPUs for 2355 seconds a 0.61 Gflops rating can be achieved. These operation counts exclude the operations needed to reconstruct the Hamiltonian on the fly. Hamiltonian storage roughly triples or quadruples these Gflops ratings. Figure 3 shows that the Lanczos algorithm performs well enough to enable the simulation of 8 and 16 million atom systems on reasonably sized Beowulf clusters. The sustained Gflop results are well within the expectations of a realistic application.

4.5 Lanczos Scaling With System Size

The preceding Section 4.4 presented the scaling of the Lanczos algorithm as a function of employed number of CPUs for different system sizes on the P800 cluster. This section discusses a subset of the same data as a function of system size for a fixed number of 24 CPUs. Four different data sets are considered based on the cross-product combination of 1 or 2 processes per node (symbol 1Px and 2Px, respectively) and stored or recomputed Hamiltonian (x=s and x=r, respectively).

Figure 3f) shows a plot of wall clock time as a function of the number of atoms in the simulation domain, N. The curves appear to be almost linear in N. Through linear regression the curves can be fitted to:

$$\begin{aligned}
T(2Pr) &= 18.568+59.825N^{1.3661},\ R=0.99976\\
T(1Pr) &= 1.064+20.342N^{0.99821},\ R=0.99997\\
T(2Ps) &= 1.5484+26.139N^{1.0264},\ R=0.99997\\
T(1Ps) &= 5.8046+73.154N^{1.1898},\ R=0.99999
\end{aligned}$$

The fitted exponentials range from $N^{0.998}$ to $N^{1.366}$ with a high regression value $R > 0.999$.

The total computation time not only depends on the time consumed on matrix-vector multiplication, but also on the number of iterations needed for convergence within the Lanczos algorithm. Experience shows that the number of iteration needed to obtain a certain number of bound eigen-states in a quantum dot system depends weakly on the system size. Typical iteration counts are of the order of 1000 to 5000. The Lanczos solver presented in this work, therefore, scales roughly linearly with the system size.

4.6 Lanczos Performance with different Network Speeds

The Myricom 1.8 Gbps networking system isutilized in the simulations shown in Figure 3. The Myricom network can be directly compared to a standard 100 Mbps Ethernet on P800, since both networks are installed independently. For the benchmarks shown in Figure 3 virtually identical results are obtained, if the simulation is performed on the significantly slower Ethernet network. This result indicates that the algorithm is not communication limited.

4.7 Examination of Memory Latency by Comparison of Different Machines

Section 4.4 showed that the dual CPU P800 machine suffers from performance degradation due to memory access in the computation of large systems or a stored Hamiltonian. This section examines this performance bottleneck further by comparing the Pentium-based cluster machines with the SGI machine (see Table 1 for the machine specifications). Figure 4a) compares execution times of 30 Lanczos iterations on P800 (red), P450 (blue), and SGI (black) with (dashed line) and without (solid line) storage of a 2 million atom Hamiltonian. The P450 outperforms the SGI without Hamiltonian storage by a factor of 1.6 to 1.9. The fast, yet expensive memory of the SGI produces a more dramatic speed increase compared to P450 and the two machines have roughly the same performance on this problem. Figure 4b) shows that the speed increase for SGI reaches a factor of about 9 while it reaches a factor of 5.5 on P450. P800 only achieves speed increase factors of about 3 to 4 due to Hamiltonian storage, depending on the node load configuration; however, P800 still outperforms the significantly more expensive (and 2 years older) SGI by a factor of approximately 2.

The memory latency problem can also be examined by comparison of execution times of the same executable and the same communication network type (100Mbps) on the P450, P800, and P933 machine when the number of CPU cycles is plotted as a function of employed number of parallel CPUs. The number of cycles is estimated as the total wall time multiplied by the frequency rating of the CPU in MHz. Figure 4c) shows such a plot for a system of 1 million atoms. If the Hamiltonian is recomputed on the fly and the required memory is small all three machines require almost identical number of cycles to compute the 30 Lanczos iterations and the curves lay on top of each other. By contrast, if the memory usage is increased due to the Hamiltonian storage, P450 requires fewer CPU cycles to compute the same problem as the machines with a high frequency rating. The additional cycles are spent waiting for the memory to arrive at the fast CPUs, which perform the computation faster than the memory delivery takes place.

4.8 Parallel Strain Algorithm Performance

The minimization of the total strain energy is numerically significantly less taxing than the electronic structure calculation. The strain computation was therefore not immediately parallelized. However, simulating system sizes of 1 million atoms or more, shows that the serial strain computation becomes computationally as taxing as the parallel electronic structure calculation that it precedes. The mechanical strain calculation has therefore been parallelized as well. This strain parallelization combined with the parallel electronic structure calculation enabled some of the alloy simulations shown in this paper as well as the bulk alloy simulations shown previously [Oyafuso, Klimeck, Bowen, Boykin (2002)].

Data are distributed in the same manner as in the electronic calculation: the simulation domain is decomposed into slabs such that atomic information associated with atoms within a slab is held by only one processor (see Figure 2a)). Message passing then takes place only between neighboring processors and the message size is proportional to the surface area of each slab, since the locality of the strain energy requires only that positions of atoms on the boundary be passed. Since the gradient of the strain energy in Eq.(7) is just as computationally inexpensive to determine as the total strain itself, a conjugate-gradient-based method that uses the derivative with respect to atomic configuration and periodicity to perform the line search[18] is used to minimize the strain energy. The parallelization of the algorithm occurs on two levels. First, the conjugate-gradient-based minimization involves computation of various inner products through a sum reduction and broadcast. Second, the function (and gradient) call to determine the local strain energy at an atomic site requires information about neighboring atoms that may lie on neighboring processors. Only position information of atoms on neighboring

[18] See the for example macopt in http://wol.ra.phy.cam.ac.uk/ mackay/c/ macopt.html

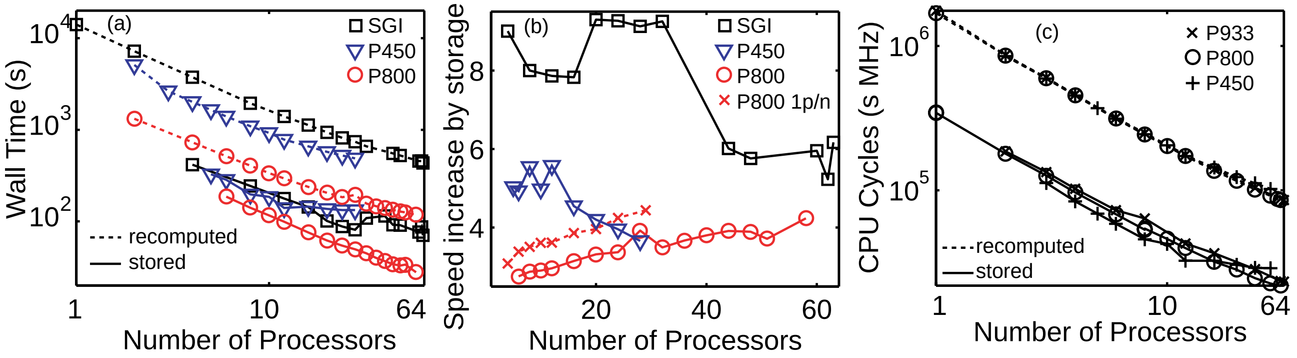

Figure 4 : (a) Execution time of 30 Lanczos iterations on three different machines: P800 (red), P450 (blue), and SGI (black) with (dashed line) and without (solid line) storage of a 2 million atom Hamiltonian. (b) Corresponding speed increase due to Hamiltonian storage (solid lines). Dashed line corresponds to P800 with 1 process per node. (c) Number of CPU cycles (wall time times CPU frequency) for the P450, P800, and P933 machine with and without stored Hamiltonian.

processors that are on the boundary is sent.

Figure 5 shows scaling results for the wall time required to achieve convergence for a system of size $32 \times 30 \times 32$ nm consisting of approximately one million atoms. The simulation was run on two different hardware configurations P800 connected by the 2 Gbps Myrinet (solid line with stars) and P933 connected by standard 100 Mbps ethernet (dashed line with circles). No shared memory was used in either case. On a single processor, there are no communication costs, and P933 outperforms P800 by about a ratio of the clock cycles of 800/933. As the number of processors is increased, however, the ratio of communication cost to computational cost increases; the communication expense is proportional to the surface area of the slabs which remains fixed while the computational cost is proportional to the slab volumes and thus inversely proportional to the number of processors. This reduction in efficiency with processor number is most evident for the slow 100 Mbps network. Using Ethernet the execution time is more than a factor of two greater than using Myricom 1.8 Gbps network.

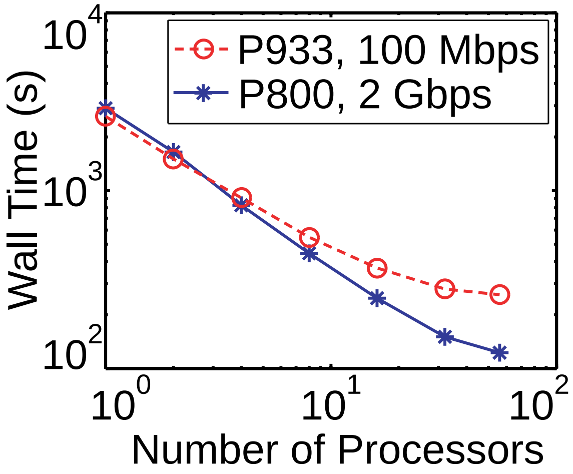

Figure 5 : Wall clock time to compute the strain in a 1 million atom system on P933 with its 100Mbps network (dashed line with circles) and on P800 with its 1.8Gbps network (solid line with stars).

For the mechanical strain calculation a significant improvement of the scaling with increasing number of CPUs with the usage of a faster, low latency network is observed. This result differs from the electronic structure calculation discussed in Section 4.4. In that case no speed increase of improved performance with increasing number of CPUs was observed (and therefore not shown in a graph). This discrepancy is a result of the larger computational demand in the electronic structure calculation. The mechanical strain calculation deals only with three real numbers (the displacements from some ideal position) for each atom and with the relative distance to its surrounding four neighbors. The electronic structure calculation by contrast deals with 10×10 and 20×20 complex matrices for each atom and its four neighbors.

5 Bulk Material Parameterizations and Properties

5.1 Genetic Algorithm-Based Fitting

Electronic structure calculations in the lowest conduction and the highest valence band require a good parameterization of the band gaps, effective masses and band-anisotropies (for the holes). One of the drawbacks of the empirical tight-binding models is that there is no simple relation between these physical observables and the orbital energies. The analytical formulas that have been developed in the past [Boykin, Klimeck, Bowen, Lake (1997); Boykin (1997); Boykin, Gamble, Klimeck, Bowen (1999)] serve as a guide for the general capability of a particular model and show that the optimization space is not smooth. The fitting of the parameters using these formulas has led to dramatic improvements in the simulation capabilities of high performance resonant tunneling diodes [Bowen, et al. (1997a); Klimeck, et al. (1997)], although the process remained tedious at best.

A very nice and diligent parameterization of the $sp^3d^5s^*$ model has been published by Jancu Scholz, Beltram, Bassani (1998)]. A large number of the technically relevant III-V materials as well as elemental semiconductors have been parameterized in their work. They have also optimized orbital-dependent distance scaling exponents η to fit strain-dependent quantities such as deformation potentials. To enhance the performance of the model with strain in a layered superlattice configuration Jancu et *al.* have developed a method where the d orbital on-site energy is shifted as a function of strain. For the general 3-D electronic structure case that is subject to this work, a more general treatment of the on-site energies as a function of strain must be included. In the NEMO 3-D implementation of the tight-binding model, all on-site energies can be shifted due to strain in an arbitrary 3-D configuration.

To automate the fitting of the orbital tight-binding parameters to the desired bulk material properties [Madelung (1996); Landolt-Bornstein (1982), Jancu, Scholz, Beltram, Bassani (1998)] a genetic algorithm (GA) based software package. The details of this algorithm and several improved material parameterizations are described elsewhere [Klimeck et al. (2000); Klimeck, Bowen, Boykin, Cwik (2000)] was developed. The general idea of the GA is the stochastic exploration of a parameter space with a large set of individuals, which represent different parameter configurations. The individuals are measured against a certain desired fitness function and ranked. Some of the individuals (for example 10% of the worst performers) are thrown out of the gene pool and replaced by new individuals that are derived from better performers by cross-over and mutation operations. The parameterizations used in this work have been obtained using this GA approach, starting from earlier parameterizations [Boykin, Klimeck, Bowen, Lake (1997); Boykin (1997); Boykin, Gamble, Klimeck, Bowen (1999); Jancu, Scholz, Beltram, Bassani (1998)].

The following sections present the parameterization data, and the resulting unstrained and strained bulk-properties.

5.2 Parameter Tables and Bulk Properties

Table 2 lists the parameters that enter the sp^3s^* model used in this paper. The parameterization for InAs was obtained from a GA, while the GaAs data was originally delivered by Boykin to the NEMO 1-D project. No effort has been made in this parameterization to fit the off-diagonal or the diagonal matrix element strain corrections. All off-diagonal matrix elements are scaled with the ideal exponent [Harrison (1999)] of $\eta = 2$ and the diagonal correction is set to zero.

An explicit InAs valence band offset vs. GaAs of 0.22795 is used in this parameterization. The $sp^3d^5s^*$ parameterization in contrast is based on common atom potentials and has the valence band offset built into the parameter set.

Table 3 shows the complete parameterization of GaAs and InAs in our $sp^3d^5s^*$ tight-binding model including the off-diagonal and diagonal strain scaling parameters. In this model a good fit based on common atomic potentials of the As in the GaAs and InAs has been obtained. A valence band offset of the unstrained materials of 0.2259eV is built into the parameter set. The $sp^3d^5s^*$ model is rich enough in its physical content to enable the fitting of GaAs, InAs, and AlAs with comon As potentials and built-in valence band offsets. Common atom potentials and built-in valence band offsets cannot be achieved in the sp^3s^* model, unless some of the fitting requirements are severely relaxed.

Table 4 summarizes the major unstrained bulk material properties that have been targeted in the sp^3s^* and $sp^3d^5s^*$ parameterization for GaAs and InAs. The target parameters are taken from various experimental and theoretical references [Madelung (1996); Landolt-Bornstein (1982), Jancu, Scholz, Beltram, Bassani (1998)]. The major parameters of interest are associated with the lowest conduction and the two highest valence bands. In Table 4 these properties are separated by a horizontal line

Table 3 : $sp^3d^5s^*$ tight-binding model parameters for GaAs and InAs. All energies are in units of eV, the lattice constant is in units of nm and the strain parameters η and C are unitless.

TB Parameter	GaAs	InAs	η_{strain}	GaAs	InAs	C_{strain}	GaAs	InAs
lattice	0.56532	0.60583						
$E(s_a)$	-5.50042	-5.50042						
$E(p_a)$	4.15107	4.15107						
$E(s_c)$	-0.24119	-0.58193						
$E(p_c)$	6.70776	6.97163						
$E(s_a^*)$	19.71059	19.71059						
$E(s_c^*)$	22.66352	19.94138						
$E(d_a)$	13.03169	13.03169						
$E(d_c)$	12.74846	13.30709						
$\Delta_a/3.0$	0.17234	0.17234						
$\Delta_c/3.0$	0.02179	0.13120				E_{shift}	27.0000	27.0000
$V(s,s)$	-1.64508	-1.69435	$ss^*\sigma$	0.00000	0.06080	$C(s,s)$	0.58696	0.53699
$V(s^*,s^*)$	-3.67720	-4.21045	$s^*s^*\sigma$	0.21266	0.00081	$C(s^*,s^*)$	0.48609	1.05899
$V(s_a^*,s_c)$	-2.20777	-2.42674	$ss\sigma$	2.06001	1.92494	$C(s_a^*,s_c)$	0.88921	0.46356
$V(s_a,s_c^*)$	-1.31491	-1.15987	$sp\sigma$	1.38498	1.57003	$C(s_a,s_c^*)$	0.77095	1.94509
$V(s_a,p_c)$	2.66493	2.59823	$pp\sigma$	2.68497	2.06151	$C(s_a,p_c)$	0.75979	1.86392
$V(s_c,p_a)$	2.96032	2.80936	$pp\pi$	1.31405	1.60247	$C(s_c,p_a)$	1.45891	3.00000
$V(s_a^*,p_c)$	1.97650	2.06766	$sd\sigma$	1.89889	1.76566	$C(s_a^*,p_c)$	0.81079	0.40772
$V(s_c^*,p_a)$	1.02755	0.93734	$s^*p\sigma$	1.39930	1.79877	$C(s_c^*,p_a)$	1.21202	2.99993
$V(s_a,d_c)$	-2.58357	-2.26837	$pd\sigma$	1.81235	2.38382	$C(s_a,d_c)$	1.07015	0.00000
$V(s_c,d_a)$	-2.32059	-2.29309	$pd\pi$	2.37964	2.45560	$C(s_c,d_a)$	0.38053	0.07982
$V(s_a^*,d_c)$	-0.62820	-0.89937	C_{diag}	2.93686	2.34322	$C(s_a^*,d_c)$	1.03256	0.00000
$V(s_c^*,d_a)$	0.13324	-0.48899	$dd\sigma$	1.72443	2.32291	$C(s_c^*,d_a)$	1.31726	0.75515
$V(p,p,\sigma)$	4.15080	4.31064	$dd\pi$	1.97253	1.61589	$C(p,p)$	0.00000	1.97354
$V(p,p,\pi)$	-1.42744	-1.28895	$dd\delta$	1.89672	2.34131			
$V(p_a,d_c,\sigma)$	-1.87428	-1.73141	$s^*d\sigma$	1.78540	2.02387	$C(p_a,d_c)$	1.61350	0.00000
$V(p_c,d_a,\sigma)$	-1.88964	-1.97842				$C(p_c,d_a)$	0.00000	0.00000
$V(p_a,d_c,\pi)$	2.52926	2.18886						
$V(p_c,d_a,\pi)$	2.54913	2.45602						
$V(d,d,\sigma)$	-1.26996	-1.58461				$C(d,d)$	1.26262	0.10541
$V(d,d,\pi)$	2.50536	2.71793						
$V(d,d,\delta)$	-0.85174	-0.50509						

Table 2 : sp^{3}s* tight-binding parameters for GaAs and InAs. All energies are in units of eV and the lattice constant is in units of nm. For this parameterization all relevant off-diagonal stain scaling parameters are set to $\eta = 2$ and all diagonal strain scaling parameters are set to $C = 0$.

Parameter	GaAs	InAs
lattice/(nm)	0.56660	0.60583
$E(s_a)$	-8.51070	-9.60889
$E(p_a)$	0.954046	0.739114
$E(s_c)$	-2.77475	-2.55269
$E(p_c)$	3.43405	3.71931
$E(s_a^*)$	8.45405	7.40911
$E(s_c^*)$	6.58405	6.73931
$V(s,s)$	-6.45130	-5.40520
$V(x,x)$	1.95460	1.83980
$V(x,y)$	4.77000	4.46930
$V(s_a,p_c)$	4.68000	3.03540
$V(s_c,p_a)$	7.70000	6.33890
$V(s_a^*,p_c)$	4.85000	3.37440
$V(p_a,s_c^*)$	7.01000	3.90970
Δ_a	0.42000	0.42000
Δ_c	0.17400	0.39300
E_v^{offset}	0.00000	0.22795

from parameters that are outside these central bands of interest. The upper and lower band edges as well as the minimum point in the [111] direction k_L are included in the optimization target with a relatively small weight. These properties are included in the optimization to preserve an ”overall” good shape of the bands outside the major interest. If they are not included, upper and lower bands will distort significantly to aid the desired perfect properties of the central bands. This distortion can lead to undesired band crossings on and off the zone center.

Also included (yet not shown in the table) is another restriction on the GaAs and InAs parameters to alloy “well” within the virtual crystal approximation (VCA). It has been found that parameter sets that represent the individual GaAs and InAs quite well can result in a $\mathrm{In_xGa_{1-x}As}$ alloy representation that has completely wrong behavior of the bands as a function of x (dramatic non-linear bowing). Typically a target that linearly interpolates the central conduction and valence band edges for $\mathrm{In_xGa_{1-x}As}$ from GaAs and InAs as a function of x is included. Bowing is not built into these VCA parameters, but establishes itself in the 3-D disordered system (see reference [] for an $\mathrm{Al_xGa_{1-x}As}$ example and Section 6.2 for a discussion on $\mathrm{In_xGa_{1-x}As}$).

Compared to the sp^{3}s* model, the sp^{3}d^{5}s* model generally provides better fits to the hole effective masses and the electron effective masses at Γ and L. The failure of the sp^{3}s* model to properly reproduce the transverse effective mass on the Δ line towards X is well understood [Klimeck, et al. (2000)]. The sp^{3}d^{5}s* model does allow the proper modeling of the effective masses in that part of the Brillouin zone.

Figure 6 shows the bulk dispersion of GaAs (left column) and InAs (right column) computed from the tight-binding parameters listed in Tables 2 and 3 without strain. The dispersion corresponding to the sp^{3}s* model is plotted in a dashed line and compared to the results from the sp^{3}d^{5}s* model in a solid line. The first row in Figure 6 shows the bands in a relatively large energy range including the lowest valence band in the models as well as several excited conduction bands. The second row in Figure 6 zooms in on the central bands of interest. The sp^{3}s* and sp^{3}d^{5}s* model agree reasonably well with each other at the Gamma point in their energies as well as their curvatures of the central bands of interest. Off the zone center the deviation between the two models become significant. Some of the band energies are hard to probe experimentally and are only known from other theoretical models [Madelung (1996); Landolt-Bornstein (1982); Jancu, Scholz Beltram, Bassani (1998)]. However the conduction band energies at X and L and their corresponding masses are well known, and the sp^{3}s* model does fail to deliver a good fit. The sp^{3}s* model generally appears to deviate strongly from the sp^{3}d^{5}s* model in the [111] direction even for the central bands of interest.

5.3 Band Edges as a Function of Strain

The deformation of atomic positions from their ideal values in a relaxed semiconductor crystal modifies the interaction between atomic neighbors and therefore the electronic bandstructure. The ability to form strained structures without defects opens a new design space exploited by many commercially relevant devices, including, for example, InGaAsP-based laser diodes operating at 1.55μm. Although good qualitative results have been obtained for the strain-dependence of the effects of interest in these devices [Silver, Oreilly (1995)], very precise measurements of all the empirical parameters that influence strain are still lacking. The baseline strain parameterization to which the calculation is compared and fitted to has been presented by Van de Walle (1989). Van de

Table 4 : Optimization targets and optimized results for the sp ^{3}s* and s3d5s* model for GaAs and InAs.

Property	GaAs Target	sps	% dev	spds	% dev	InAs Target	sps	% dev	spds	% dev
E_g^Γ	1.4240	1.4173	0.4676	1.4242	0.0150	0.3700	0.3740	1.0679	0.3699	0.0232
E_c^Γ	1.4240	1.4173	0.4675	1.4212	0.1996	0.5957	0.6019	1.0406	0.5942	0.2496
V_{hh}	0.0000	0.0000	0.0000	-0.0031	0.3055	0.2257	0.2279	0.2247	0.2243	0.1401
$m_c^*[001]$	0.0670	0.0679	1.3195	0.0662	1.1353	0.0239	0.0245	2.3030	0.0235	1.5417
$m_{lh}^*[001]$	-0.087	-0.0699	19.7125	-0.0830	4.6849	-0.0273	-0.0282	3.2117	-0.0281	2.9541
$m_{lh}^*[011]$	-0.080	-0.0661	17.7381	-0.0759	5.6414	-0.0264	-0.0275	4.3227	-0.0273	3.3575
$m_{lh}^*[111]$	-0.078	-0.0498	36.6755	-0.0740	5.8547	-0.0261	-0.0207	20.7535	-0.0270	3.5783
$m_{hh}^*[001]$	-0.403	-0.4436	10.0710	-0.3751	6.9198	-0.3448	-0.4410	27.9049	-0.3516	1.9617
$m_{hh}^*[011]$	-0.660	-0.7103	7.6270	-0.6538	0.9421	-0.6391	-0.7159	12.0144	-0.5634	11.8389
$m_{hh}^*[111]$	-0.813	-0.8726	7.3332	-0.8352	2.7329	-0.8764	-0.8972	2.3757	-0.6982	20.3385
$m_{so}^*[001]$	-0.150	-0.1447	3.5239	-0.1629	8.6134					
$E_c^X - E_c^\Gamma$	0.4760	0.4742	0.3753	0.4760	0.0099	1.9100	1.9008	0.4829	1.9131	0.1626
$m_X^*[long]$	1.3000	1.2552	3.4436	1.3138	1.0596					
$m_X^*[trans]$	0.2300	4.1920	1722.61	0.1740	24.3358					
k_X	0.9000	0.8550	5.0000	0.8860	1.5556					
$E_c^L - E_c^\Gamma$	0.2840	0.5339	88.0001	0.2825	0.5201	1.1600	1.3394	15.4628	1.1589	0.0915
$m_L^*[long]$	1.9000	2.9849	57.0979	1.7125	9.8685					
$m_L^*[trans]$	0.0754	1.1972	1487.74	0.0971	28.7342					
k_L	1.0000	1.0000	0.0000	1.0000	0.0000	1.0000	1.0000	0.0000	1.0000	0.0000
Δ_{so}	0.3400	0.3636	6.9545	0.3265	3.9792	0.3800	0.4150	9.2117	0.3932	3.4644
E_{6v}^Γ	-13.10	-12.703	3.0321	-12.370	5.5723	-12.300	-12.535	1.9149	-13.417	9.0847
E_{so}^Δ	-0.340	-0.3636	6.9545	-0.3265	3.9792	-0.3800	-0.4150	9.2117	-0.3932	3.4644
E_{6c}^Γ	1.4240	1.4173	0.4676	1.4242	0.0150	0.3700	0.3740	1.0679	0.3699	0.0232
E_{7c}^Γ	4.5300	4.3557	3.8468	3.4544	23.7449	4.3900	4.3314	1.3342	3.7402	14.8027
E_{8c}^Γ	4.7160	4.5861	2.7546	3.5785	24.1198	4.6300	4.7294	2.1474	4.0466	12.6013
E_{6v}^X	-2.880	-2.8013	2.7321	-2.2302	22.5631	-2.4000	-2.5311	5.4607	-2.3486	2.1434
E_{7v}^X	-2.800	-2.6699	4.6481	-2.0470	26.8927	-2.4000	-2.4383	1.5958	-2.2525	6.1441
E_{6c}^X	1.9800	1.9278	2.6376	1.9199	3.0363	2.5000	3.2599	30.3955	2.6286	5.1458
E_{7c}^X	2.3200	2.1101	9.0469	2.1298	8.1972					
E_{4v}^L						-10.920	-11.111	1.7551	-10.580	3.1117
E_{5v}^L						-6.2300	-6.9272	11.1903	-5.8611	5.9221
E_{6v}^L	-1.420	-1.5793	11.2172	-1.1169	21.3432	-1.2000	-1.4897	24.1448	-1.3048	8.7339
E_{7v}^L	-1.200	-1.2766	6.3866	-0.8975	25.2051	-0.9000	-1.1221	24.6830	-1.0129	12.5407
E_{6c}^L	1.8500	1.9513	5.4736	1.7067	7.7440	1.5000	1.7133	14.2213	1.5289	1.9235
E_{7c}^L	5.4700	3.1464	42.4786	3.9357	28.0501	5.4000	4.3284	19.8452	4.1758	22.6708

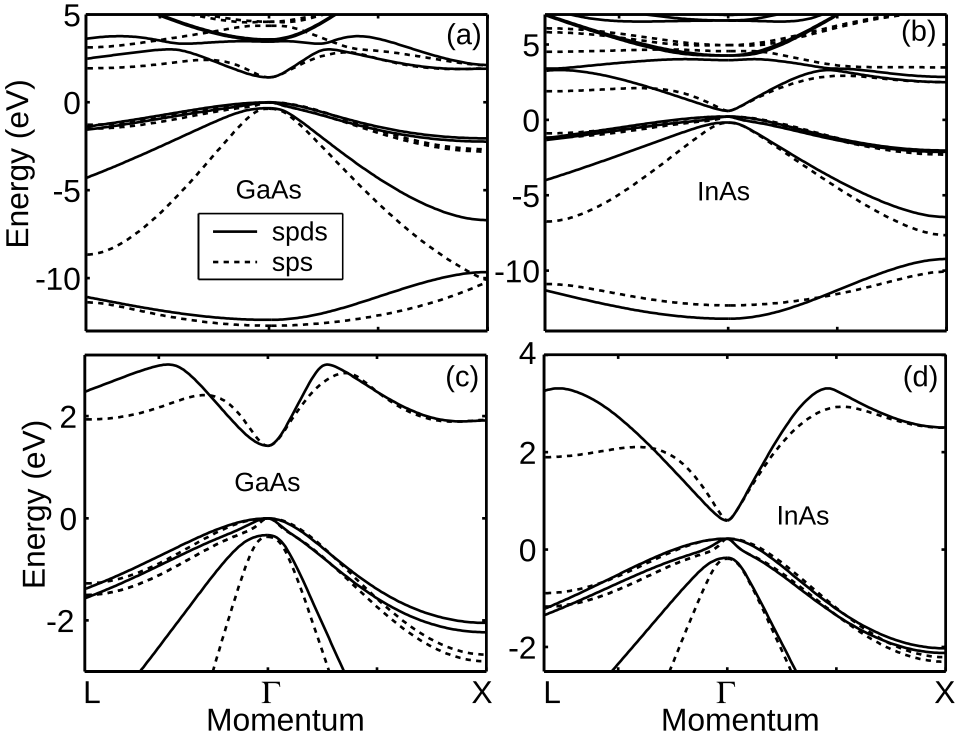

Figure 6 : E(k) dispersion for GaAs (left column) and InAs (right column) computed with the sp ^{3}s* (dashed line) and sp^{3}d^{5}s* (solid line model).

Walle's parameterization is not purely empirically based, but partially dependent on a $k{\cdot}p$ expansion following Pollak and Cardona (1968). For this work, Van de Walle's parameters have been slightly modified to represent room temperature bandgaps.

Figure 7 shows the conduction and valence band edges for GaAs (left column) and InAs (right column) as a function of hydrostatic strain (top row) and bi-axial strain (bottom row). Three parameterizations are compared in each graph: 1) reference data by Van de Walle (dashed line), 2) data computed from the sp^{3}d^{5}s* model (circles), and 3) data computed from the sp^{3}s* model (solid line). The test application in this paper is the modeling of a strained InGaAs system grown on top of a GaAs substrate. Since InAs has a larger lattice constant than GaAs one needs to model effects on InAs as it is compressed towards the GaAs lattice constant (7% negative strain). GaAs bonds, by contrast, are expected to be stretched towards the InAs bondlength at interfaces (positive strain). Since the InGaAs quantum dots grown on GaAs which are considered in the next two sectins 6 and 7 are significantly larger in their width than their height, one can expect the strain in the dot to be mostly bi-axial. However some hydrostatic strain distributions can be expected as well, due to the finite extent of the InAs quantum dots inside the GaAs buffer. The z-directional strain component in the bi-axial strain case is computed as $\varepsilon_{zz}=2\frac{c_{12}}{c_{11}}\varepsilon_{xx}$ and $\varepsilon_{yy}=\varepsilon_{xx}$.

Both tight-binding models follow the trends set by the Van de Walle reference reasonably well. Generally speaking the sp^{3}d^{5}s* model performs better than the sp^{3}s* model (which actually was not optimized for its strain performance). It has been particularly hard to improve the under-prediction of the InAs band gap (Figure 7d)) for large compressive bi-axial strain. The reasonably good fit has been obtained by compromising the fit of the conduction band under hydrostatic compressive strain. In contrast, the InAs valence bands has not not posed any problem at all to be fit to the Van de Walle data.

5.4 *Effective Masses as a Function of Strain*

Previous nanoelectronic transport simulations have shown that it is essential [Bowen, et al. (1997a); Klimeck, et al. (1997); Bowen, et al. (1997b)] to properly model the band edges and effective masses in the

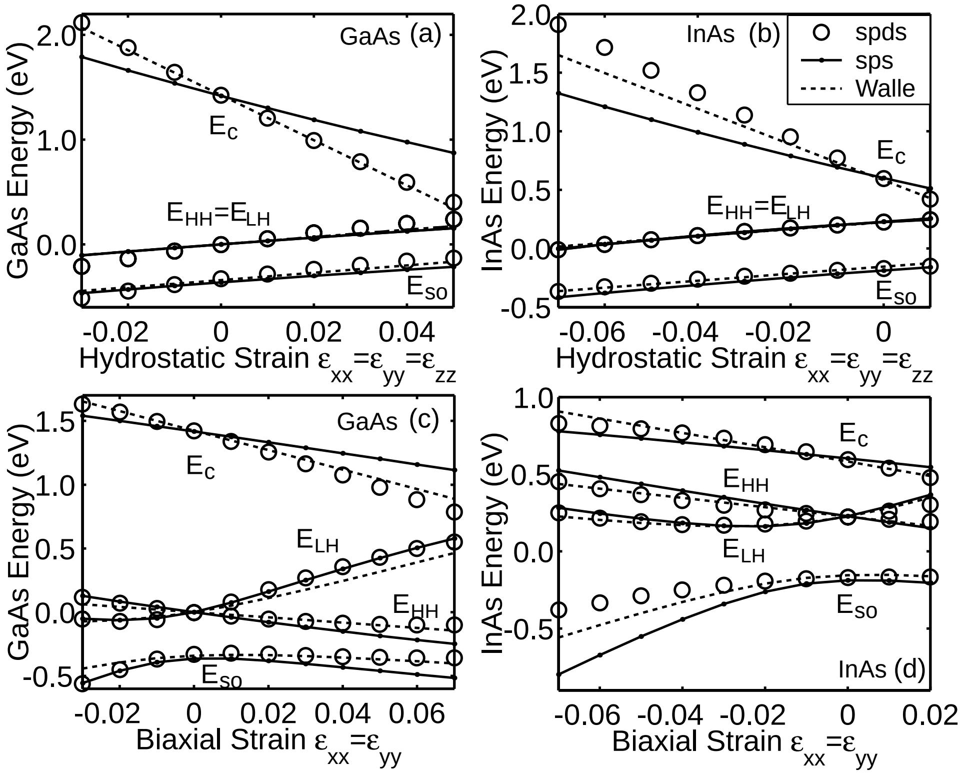

Figure 7 : GaAs and InAs conduction band, heavy-hole, light-hole and split-off band edges as a function of hydrostatic and bi-axial strain. Dashed line from parameterization of Van de Walle (1989). Circles from sp^{3}d^{5}s* model and solid line from sp^{3}s* model.

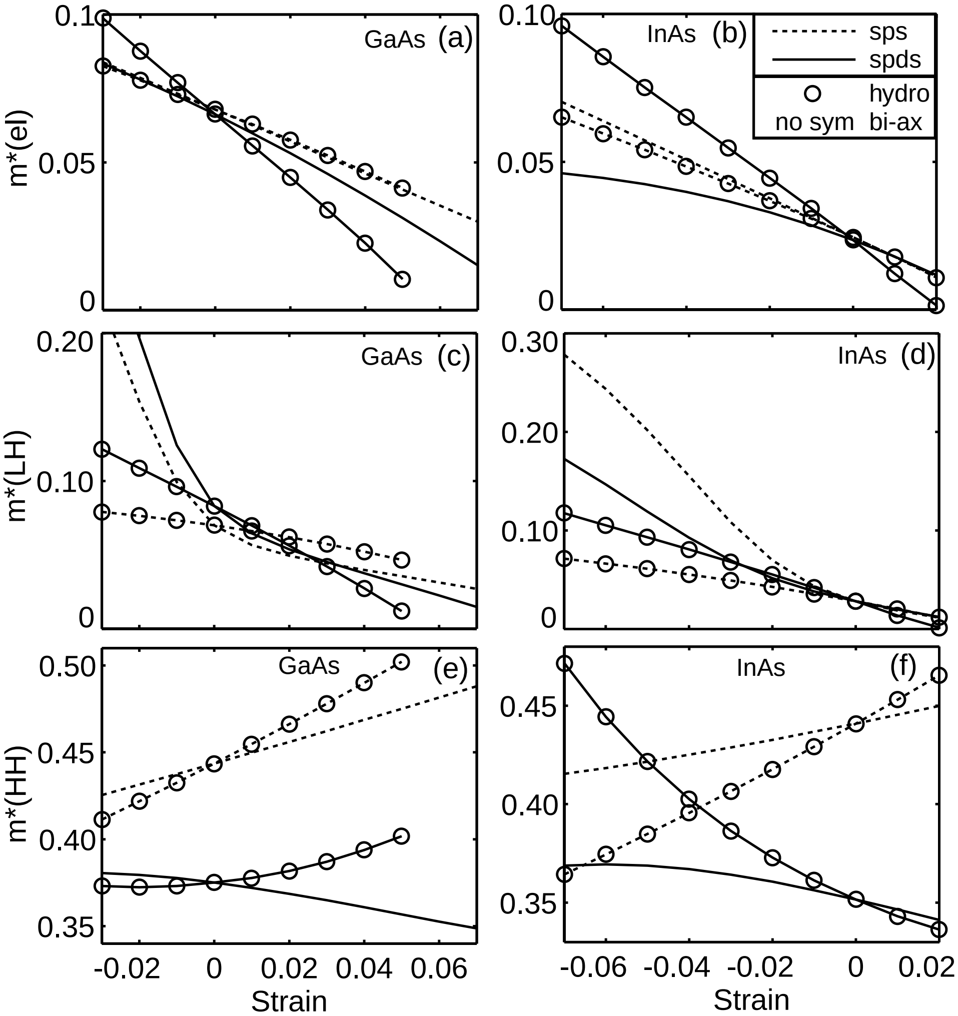

Figure 8 : Electron, heavy hole (HH) and light-hole (LH) effective masses in the [001] direction as a function of hydrostatic (circles) and bi-axial (no symbols) strain for the sp ^{3}s* and the sp^{3}d^{5}s* model. Left column GaAs, right column InAs. Negative strain numbers correspond to compressive strain.

heterostructure. In a single band model the dependence of eigen-energy of a confined state is inversely proportional to the effective mass. With the strong dependence of the band-edges on the strain shown in Figure 7 one can also expect a strong dependence of the effective masses on the strain. Figure 8 shows the electron (first row), light hole (second row), and heavy hole (third row) effective masses for GaAs (first column) and InAs (second column) as a function of hydrostatic strain (lines with circles) and bi-axial strain (lines without symbols) for the sp^3s^* (dashed line) and the $sp^3d^5s^*$ model (solid line) computed in the [001] direction. Negative strain values correspond to compressive strain. For the electron mass the sp^3s^* and the $sp^3d^5s^*$ model show roughly the same trends for GaAs as well as InAs. The GaAs mass drops towards the smaller InAs mass as GaAs is stretched towards InAs. In InAs the electron mass is increased towards the heavier GaAs mass as the material is compressed towards GaAs. The $sp^3d^5s^*$ model shows a larger difference between the effect of hydrostatic and bi-axial strain than the sp^3s^* model.

The change in the effective mass in InAs under compressive bi-axial strain is quite important. Under 7% bi-axial strain the effective mass approximately doubles. This increase in the effective mass lowers the confinement energies in the the quantum dots, effectively increasing the confinement. The spacing between the confined electron states will also be significantly reduced.

The light hole masses (Fig 8c,d)) show a similar linear dependence to *hydrostatic* strain as the electron masses for both band structure models. Under bi-axial compressive strain, however, the light hole mass increases dramatically towards the heavy hole mass. Both tight-binding models predict roughly the same behavior. In the case of thin InGaAs quantum dots strained on GaAs this implies that the light hole confinement is much stronger and the light hole state separation is much smaller than the unstrained LH effective mass would indicate. Note however that the LH band is significantly separated from the HH band due to strain as indicated in Figure 7d).

While the two tight-binding models show similar trends for the electron and light hole effective masses for GaAs and InAs under both pressure types, the two models show different trends for the heavy hole masses. In the case of GaAs under hydrostatic pressure the two models still predict the same trends for the HH mass. However, with increasing bi-axial strain the sp^3s^* model predicts an increase in the GaAs HH mass, while the $sp^3d^5s^*$ model shows the opposite trend. In the case of InAs the two models predict conflicting trends in both strain regimes. Note that both models have slightly different zero-strain origins as indicated in Table 4. The difference in the strain dependence trends for the HH mass in the two tight-binding models may result in different hole confinements and hole state separations predicted by the two models. Although the conflicting trends are somewhat disturbing and warrant further examination on their effects on confined hole masses, it is also important to note that the overall variation due to strain is small to within about 15%. Variations with strain in the electron and light hole masses are much more significant on the order of 100% and both models predict the same trend.

6 Application of NEMO 3-D to InGaAs Alloyed Systems

The previous Section 5 discusses the parameterization of GaAs and InAs in the sp^3s^* and $sp^3d^5s^*$ model. All the material properties in that section were computed on the basis of a single primitive fcc-based cell. This section 6 and section 7 focus on the properties of the alloy $\mathrm{In_xGa_{1-x}As}$ modeled by the constituents of GaAs and InAs in a 3-D chunk of material consisting of tens of thousands to over 6 million atoms. Two different systems are considered in detail: 1) bulk $\mathrm{In_xGa_{1-x}As}$ and its properties as a function of In concentration x, and 2) $\mathrm{In_{0.6}Ga_{0.4}As}$ dome shaped quantum dots embedded in GaAs. Within each system the strain properties are examined first, followed by an analysis of the electronic structure. Throughout this section the $sp^3d^5s^*$ model is used for all the electronic structure calculations.

6.1 *Strain Properties of Bulk* $\mathbf{In_xGa_{1-x}As}$

The starting point of many atomistic electronic structure calculations is a determination of the atomic configuration through a minimization of the total strain energy. The strain calculation discussed earlier is applied to a small, periodic $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ system consisting of approximately 13000 atoms. Figure 9 shows the mean bond lengths for such a small system. The curve in red (blue) corresponds to the mean In-As (Ga-As) bond length and is bounded by dotted curves that delimit the range of bond lengths that lie within one standard deviation of the mean. Clearly, as the material in question becomes less alloy-like (i.e more GaAs-like or InAs-like) the standard deviations approach zero.

The curve in green is the average of the Ga-As and In-As bond lengths weighted by the concentration of each cation and represents the mean bond length throughout the crystal. Note that this mean is strongly linear with a very slight upward bowing and is consistent with Vegard's law [Chen, Sher (1995)]. Also evident is the bimodal nature of the bond length distribution which demonstrates that on a local scale the crystalline structure around any particular cation retains to a large degree the character of the binary bulk material. The computed bond lengths show reasonable agreement with those determined from experiment [Mikkelsen, Boyce (1982)] (shown in black), but tend closer to the mean crystal value.

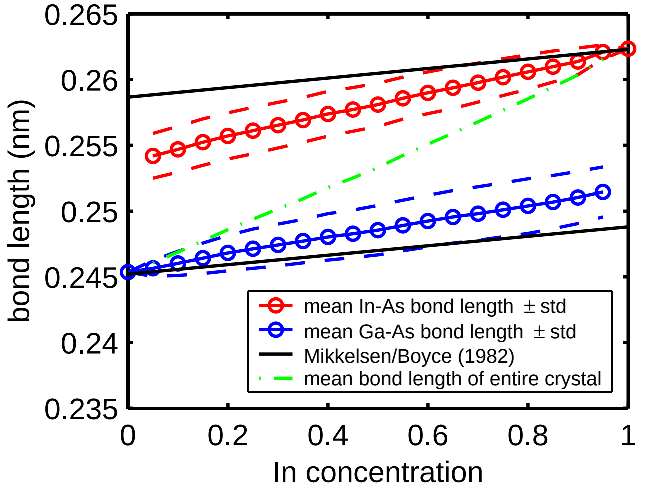

Figure 9 : In-As (red) and Ga-As (blue) bond length average (with an error margin of one standard deviation) as a function of In concentration x. Black line corresponds to experimental data reported by Mikkelsen and Boyce (1982). Dot-dashed green line corresponds to a VCA result representing the mean bond length of the entire crystal.

6.2 *Electronic Properties of Bulk* $\mathbf{In_xGa_{1-x}As}$

With the atomic configurations the electronic properties of the $In_{0.6}Ga_{0.4}As$ system can be obtained. Figure 10 compares the experimentally measured [Landolt-Bornstein (1982)] energy gap (shown in green) of $In_xGa_{1-x}As$ as a function of In concentration x with numerical results, obtained in two different ways. The red curve is the VCA result, obtained by diagonalizing the tight-binding Hamiltonian on a single unit cell with periodic boundary conditions, in which the cation-anion coupling potentials are determined by a strict average of the In-As and Ga-As coupling potentials. The lattice constant of the single cubic unit cell is determined by Vegard's law [Chen, Sher (1995)]. The resulting energy gap is mostly linear, but displays a very slight upward bowing. The blue curve is obtained by diagonalizing the full Hamiltonian of the alloyed system. The system size is sufficiently large that variations of the energy gap due to configurational noise (see analysis in Section 7) is not visible on the energy scale shown in the figure. The determined energy gap differs from the VCA result by a maximum of 60 meV and displays a slight downward bowing, although significantly less than that of the experimental result [Landolt-Bornstein (1982)]. The linear behavior in the VCA computed bandgap is included in the tight-binding parameter fitting as discussed in Section 5.2. The random cation disorder in the 3-D bulk system can, therefore, be attributed with the bowing. In similar $Al_xGa_{1-x}As$ simulations [Oyafuso, Klimeck, Bowen, Boykin (2002)] much better agreement between the 3-D simulation and the experimental results has been achieved. Some bowing might have to be built into the VCA based parameterization of GaAs and InAs to accommodate the larger bowing in the $In_xGa_{1-x}As$ system compare to the $Al_xGa_{1-x}As$.

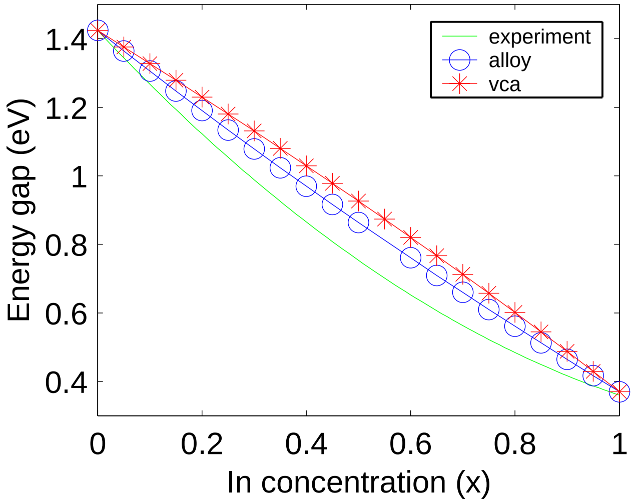

Figure 10 : Experimentally measured [Landolt-Bornstein (1982)] energy gap (solid line) of $In_xGa_{1-x}As$ as a function of In concentration x compared with results based on the 3-D random alloy simulation (circles) and a virtual crystal approximation (stars).

6.3 Strain in an Alloyed Quantum Dot

This section demonstrates an example of the strain calculation in an alloyed quantum dot using NEMO 3-D . The model problem is a dome-shaped $In_{0.6}Ga_{0.4}As$ QD of diameter 30 nm and height 5 nm enclosed in a GaAs box of size $74 \times 74 \times 28nm^3$. The entire structure, which contains roughly 6.3 million atoms, is allowed to expand freely to minimize the total strain energy (no fixed boundary conditions). The diagonal part of the local strain tensor is examined along the *x*-axis, which lies halfway between the top and bottom of the dome and parallel to its base. Figures 11a) and 11b) show the components ε_{xx} (blue), ε_{yy} (green), ε_{zz} (red), and Tr(ε) (black) of the local strain tensor of the primitive cell centered about the Ga and In cations respectively. Within the QD, the In-As bonds (see Figure 11b) are compressively strained roughly equally in the x and y directions (approximately 4.69% and 4.99% respectively). There is a very slight tensile strain in the z direction ($\sim$ 0.02 %). There are three competing effects that determine the sign and magnitude of this strain. First, there is a negative hydrostatic component due to the smaller lattice constant of the buffer. Second, the flatness of the dome means that close to the center of the QD, the strain field should approach that of quantum well in which the cubic cell is compressively strained laterally (i.e. in x and y) and stretched in z. Finally, the presence of nearby Ga cations provides an additional negative hydrostatic component to the strain. The combination of these three effects gives rise to a large biaxial compressive strain and a nearly vanishing strain component normal to the flat dome.

The Ga cations within the dome (see Figure 11a) are subject to only one of these effects, that of the biaxial strain of nearby In-As bonds. Interestingly, the Ga-As bond lengths are reduced laterally (-1.98% (x) and -1.90% (y)) from their bulk values. This reduction is likely an effort to match the very large z-component of the In-As bondlengths. The resulting average tension in the z direction is 2.14%.

Just outside the dome, along x, the Ga atoms suffer tensile strain in y and in z (although more so in z) to match the effective lattice constant on the boundary of the dome. This stretching results in compressive strain along x as indicated in Figure 11(a) by the negative value of ε_{xx} outside the QD.

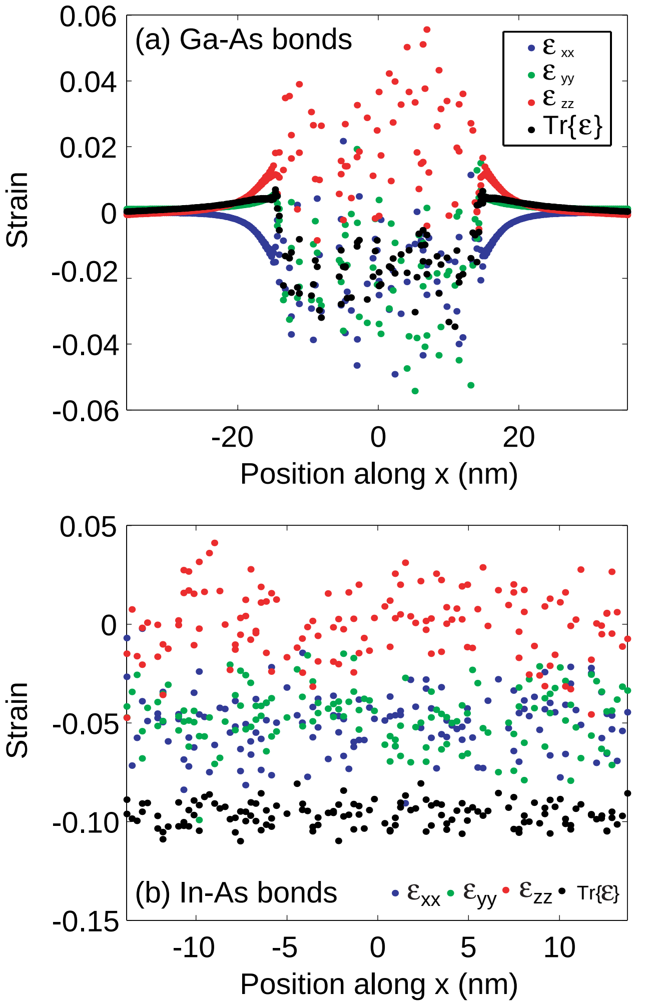

Figure 11 : Components of strain tensor for primitive cells centered around (a) Ga and (b) In cations along an axis midway from the top and bottom of the dome and parallel to its base.

6.4 Local Band Structure in an Alloyed Quantum Dot

Figure 12 shows the effect of the deformation of the primitive cells under strain on the local electron and hole band structure. Each point represents a “local” eigen-energy obtained by constructing a bulk solid from the primitive cell formed from the four As anions enclosing each cation. One sees that outside the QD, the tensile strain the GaAs cells experience reduces the conduction band edge slightly from its bulk value and splits the degenerate valence band (shown in black) into heavy hole (HH) and light hole (LH) bands.

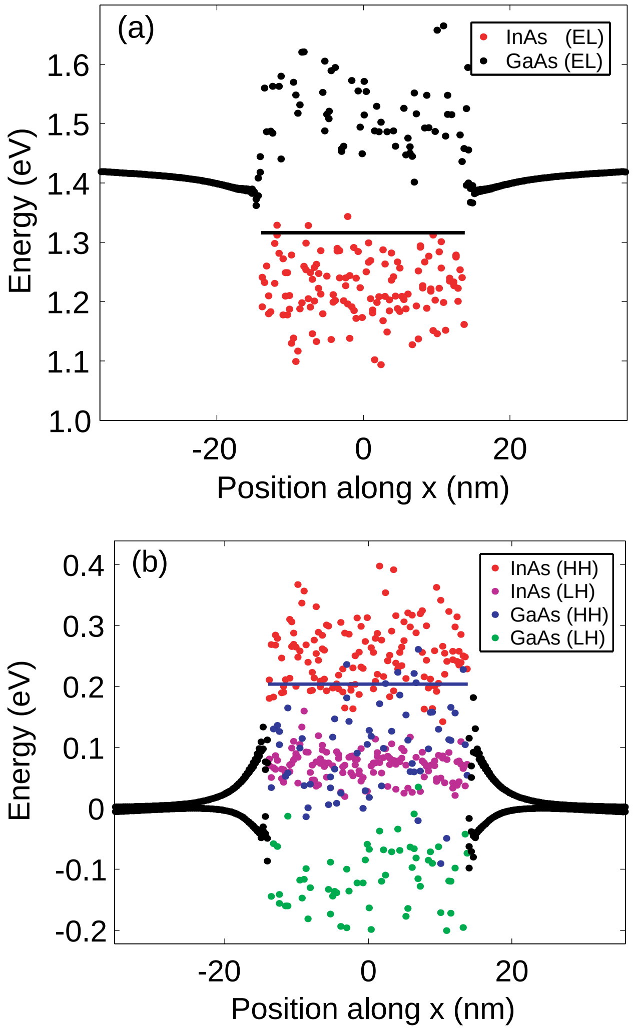

Figure 12 : “Local” conduction (a) and valence (b) band edges determined by imposing periodic boundary conditions on a primitive cell constructed from the four anions surrounding a given cation.

Within the QD, both GaAs and InAs cells are squeezed laterally resulting in a increase in local electron eigen-energies. The resulting mean electron band edge along the *x*-axis and within the QD is indicated by the thin solid line. Biaxial compressive strain also raises (lowers) the local HH (LH) eigen-energies within the QD for both InAs and GaAs cells, and, again, the average HH QD band edge is indicated by a thin solid line. Clearly, the random distribution induces a large variation in local potentials, which will shortly be seen to strongly affect shallow hole states.

6.5 Wave Functions in an Alloyed Quantum Dot

This section examines the effect of disorder on electron and hole eigenfunctions. Three different alloy configurations are examined for the same quantum dot size, shape, and number of included atoms – two different random alloy configurations that differ only by the random placement of the In and Ga atoms in the $In_{0.6}Ga_{0.4}As$, and a VCA-based configuration without spatial disorder. In the VCA representation all cations within the QD are of a fictitious type “$In_{0.6}Ga_{0.4}$” in which all tight-binding parameters (and the strain parameters) are linearly averaged between InAs and GaAs parameters [19]. This case corresponds most closely to a jellium description and is used as our baseline reference. The disordered wave functions are shown to be significantly different from each other as well as from the homogeneous VCA system wavefunctions. A detailed statistical analysis of the computed eigen energies as a result of the wave function variations is deferred to Section 7. The quantum dot and composition is identical to the discussion above. However, to reduce the computational expense, the GaAs buffer is reduced to a size of $74 \times 74 \times 15$ nm^3 and contains roughly 3.6 million atoms.

Figure 13 shows four different representations of the ground state electron wave function obtained for three different configurations. The first column shows results for a VCA implementation. The other two columns display results for two separate random distributions of In and Ga atoms within the QD. The first row depicts scatter plots of the probability density, where the red points mark atomic sites where the probability density exceeds one-third of the maximum value, and green and blue points mark higher values. Clearly there is not much differ-

[19] The anion As parameters are in general averaged as well in the VCA approximation. However in the $sp^3d^5s^*$ parameterization discussed in Section 5.2 all As parameters are already identical.

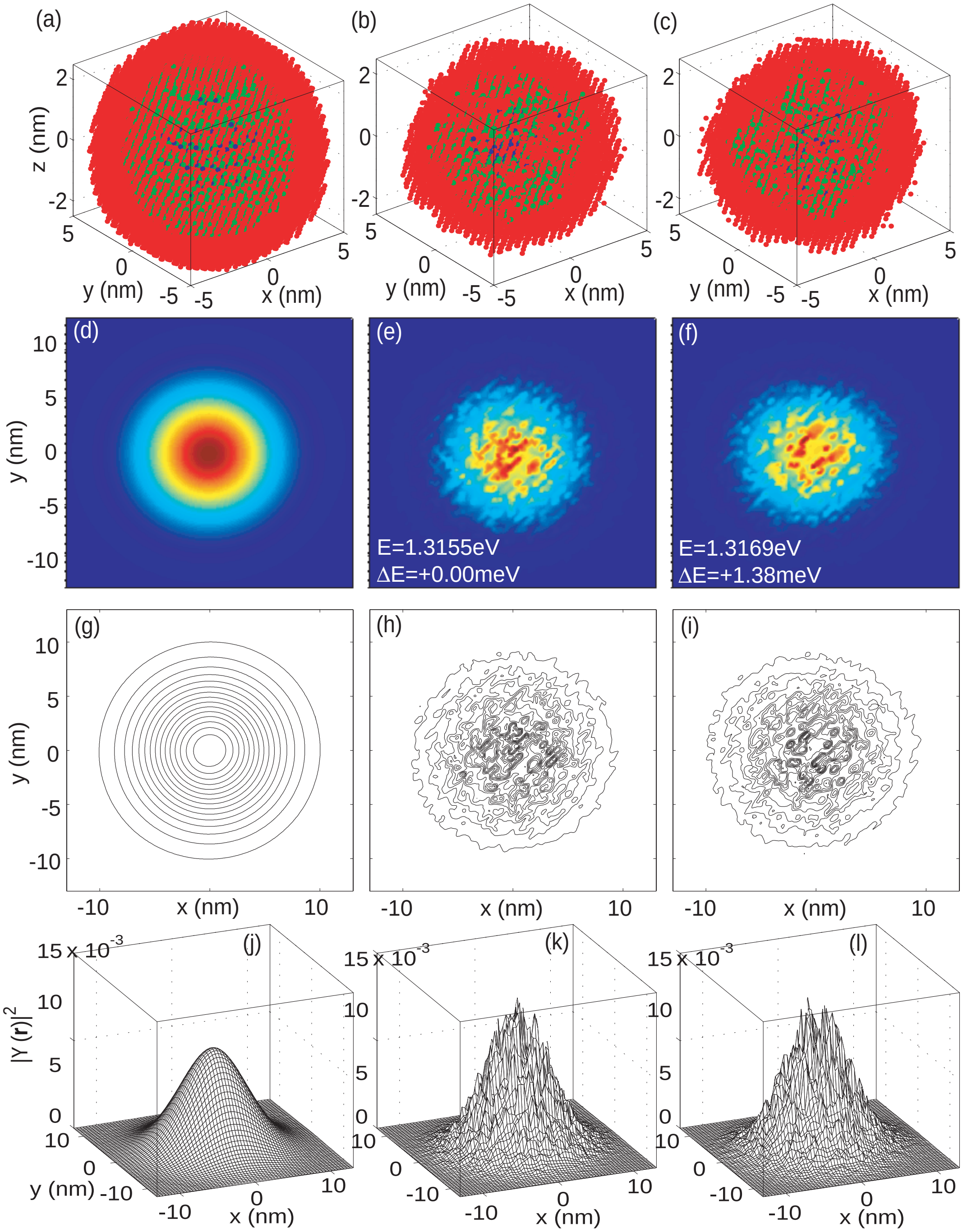

Figure 13 : Electron ground state wave functions without disorder - VCA (first column), and two different random alloy configurations (middle and right column). First row: scatter plot of wave function in 3-D. Second, third, and fourth row: colored contour plot, outlined contour and surface plot sliced through the middle of the quantum dot at a constant z, respectively.

ence between the three plots except that the VCA result is somewhat smoother. Also, the VCA plot is slightly larger indicating a slower decay as one moves away from the central axis of the QD. The next two rows are contour plots of a slice parallel to the base of the dome and mid-way up in height. Here, the impact of the disorder on the wave function is quite evident, although the s-like character of the wave function still closely resembles that of the homogeneous QD. Also, the difference between the two disordered QDs is not significant. The eigenenergies differ by about 1.38meV. The last row depicts a surface plot of the wave function (normalized to unity over the entire simulation domain) and shows that the homogenous result is a smoothed version of the disordered wave function with a lower maximum.

Figure 14 shows a set of hole wave functions analogous to the those shown in Figure 13 for the electron ground state. First note that the VCA scatter plot looks similar to the ground state VCA electron wavefunction, except that it is flatter. The stronger localization in the *z* direction reflects the greater confinement due to the larger hole mass relative to that of the electron. The larger hole mass also makes the wave function more susceptible to perturbations in the local potential. This effect is demonstrated in the three hole scatter plots, where the disorder strongly changes the appearance of the wave function. Note, also, that different placements of cations can produce noticeably different results as seen in the contour plots where the location of the wave function peaks vary by several nm. The greater localization in systems with disorder also manifests itself by the much larger peaks in the surface plots, where the probability density is, again, normalized to unity over the entire simulation domain in each of the three cases. The hole eigenvalues differ by -3.44meV compared to a difference of +1.38meV for the electrons. A more detailed statistical analysis of the distribution of eigenvalues is the topic of the following section 7.

Figure 15 shows the six lowest electron (rows 1 and 2) and hole (rows 3 and 4) states for a similar system with the same dome dimensions, but enclosed in a buffer of size $56 \times 56 \times 24$ nm^3. First, note that the electron states more closely resemble the states one would expect from a homogeneous QD. Also, the three lowest hole states correspond well to their electron counterparts, but higher energy states differ. There are two possible explanations. First, the Lanczos algorithm might not have yet converged on an intermediate eigenvalue. Second, the disorder in the system may rearrange the ordering of the eigenstates. Note that there exist several pairs of states (electron 2 and 3; electron 5 and 6; and hole 2 and 3) that stem from degenerate states in the homogeneous system, yet the disorder splits their eigenenergies by up to 1.4 meV. Since this splitting due to disorder is roughly the same order of magnitude as the separation of the excited states, it is conceivable that the disorder can rearrange the ordering of the eigenstates.

7 Statistical Analysis of Random Disorder in Alloyed Quantum Dots

7.1 Set-up of the Numerical Experiment

This section considers the same dome shaped $In_{0.6}Ga_{0.4}As$ quantum dot as the previous sections. Since the In and Ga ions inside the alloyed dot are randomly distributed, different alloy configurations exist and optical transition energies from one dot to the next may vary, even if the size and the shape of the dot are assumed to be fixed. This section seeks to answer the question: What is the minimal optical line width that can be expected for such an alloyed dot neglecting any experimental size variations? To enable the simulation of about 1000 different configurations the required simulation time was reduced by three additional approximations / simplifications: 1) the surrounding GaAs buffer is reduced to 5nm in each direction around the quantum dot. This results in a total simulation domain of approximately 1,000,000 atoms with about 718,000 atoms in the quantum dot itself, 2) the use of sp^{3}s* model instead of the sp^{3}d^{5}s* model (a reduction of the required compute time by about 4×), and 3) the computation of the eigenvalues without the corresponding eigenvectors (resulting in a reduction of compute time by exactly a factor of two). With these approximations and simplifications the wall clock time to obtain one set of eigenvalues for one particular alloy configuration took about 25 minutes on 31 processors of P933. 1000 different alloy samples therefore required approximately 420 hours or 17.4 days wall clock time or about 13,000 hours or 538 days single processor computing time. The mechanical strain is minimized using a valence force field method [Keating (1966); Pryor, Kim, Wang (1998)] as discussed in Section 3.4 for each alloy configuration. Changing the random seed

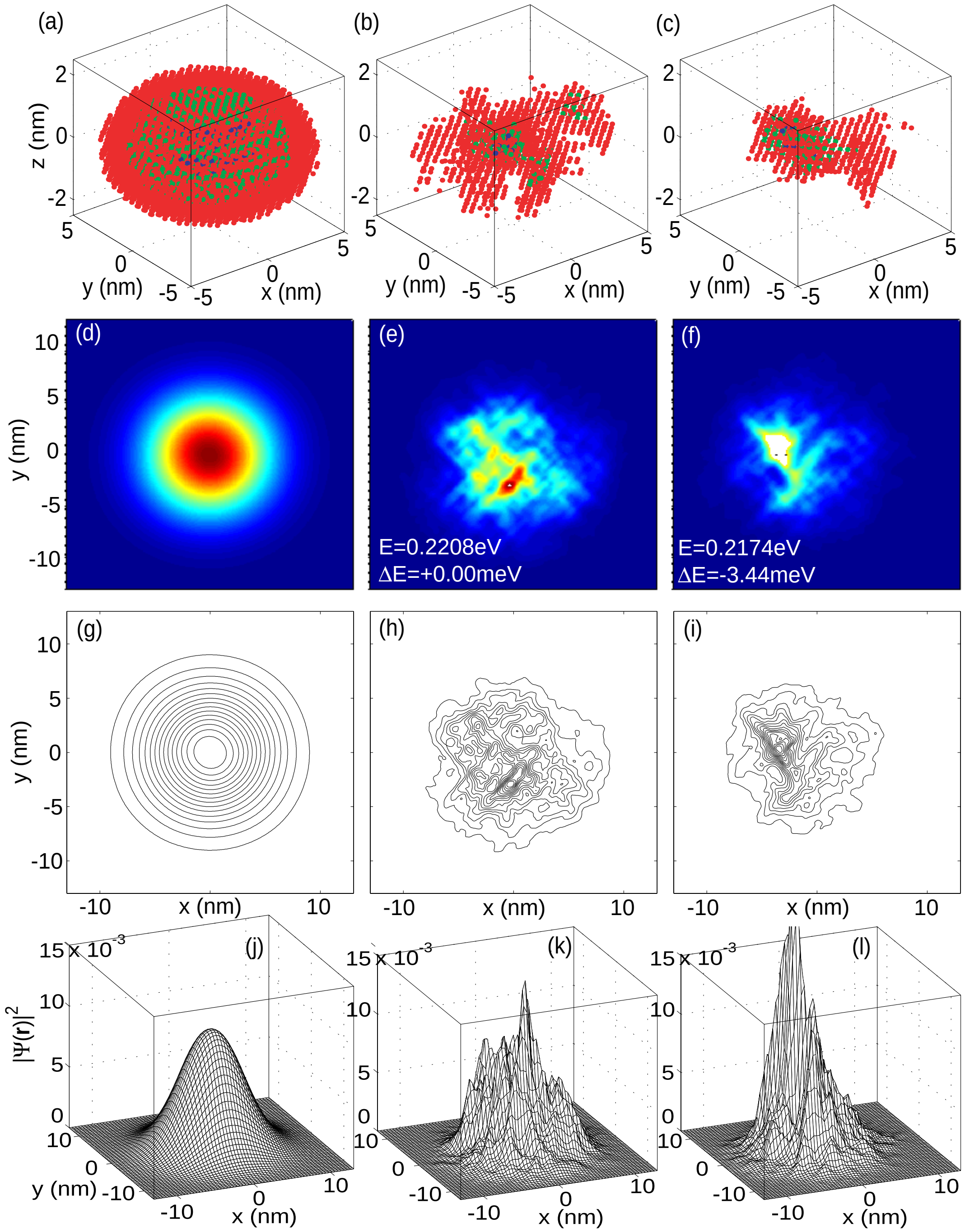

Figure 14 : Hole ground state wave functions without disorder - VCA (first column), and two different random alloy configurations (middle and right column). First row: scatter plot of wave function in 3-D. Second, third, and fourth row: colored contour plot, outlined contour and surface plot sliced through the middle of the quantum dot at a constant z, respectively.

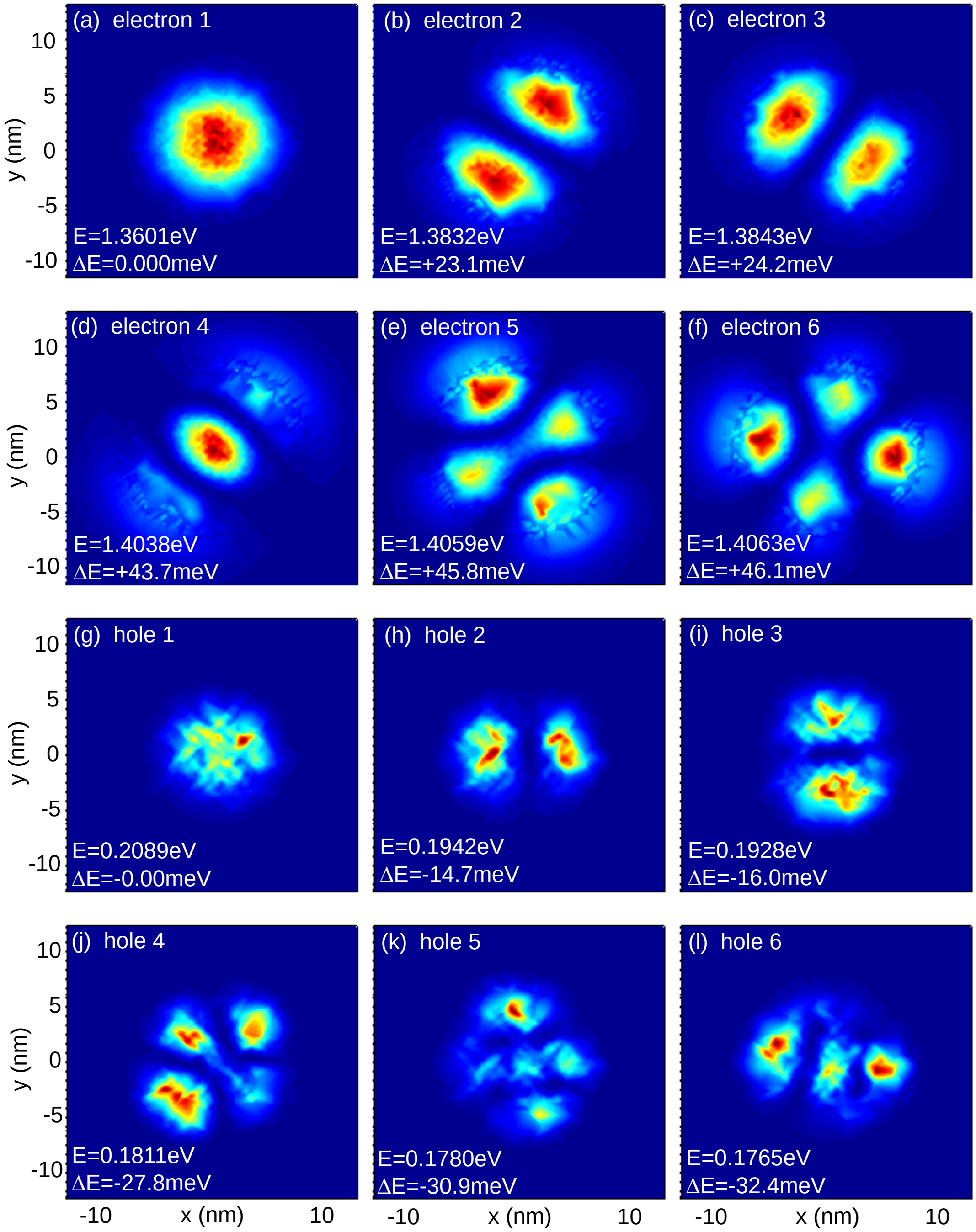

Figure 15 : Electron and hole wave functions with disorder.

of the random number generator generates the random alloy configurations. The reduced GaAs buffer size tends to increase the optical band-to-band transition energy by simultaneously raising the electron energy and lowering the hole energy. The use of the sp^{3}s* mode compromises some of the accuracy of the electronic structure due to strain (see discussions in Section 5). In particular one can expect that the optical band gap will be underpredicted since the bi-axially strained InAs band gap is under predicted in bulk (see Figure 7d). While the absolute energies are shifted from the experimental data, one can, however, still expect that the distribution and extent of the variations in the eigenvalues generated by the alloy disorder around the mean energies are independent of the mean, and therefore independent of the buffer size[20]

NEMO 3-D currently supports two disorder models. The first makes the simplifying assumption that neighboring cations are completely uncorrelated so that the species at a particular cation site is determined randomly according to the expected concentration x and independently of the configuration of the remainder of the supercell. This disorder is referred to as atomic granularity (AG) in this paper. The second model of compositional disorder, increases the granularity of the disorder from the atomic level to that of the cubic cell, so that all four cations within a unit cell are of the same species resulting in cell granularity (CG). Within the AG model the cation concentrations can be allowed to vary statistically or they can be pinned to a single value, enabling the simulation of pure configuration noise.

Previous work [Oyafuso, Klimeck, Bowen, Boykin (2002)] on an unstrained $\mathrm{Al_xGa_{1-x}As}$ bulk system showed that concentrational noise (concentration x varies statistically) dominates over the configurational noise (fixed concentration x) by at least one order of magnitude in the standard deviations in the conduction and valence band edge. The pure configurational noise is therefore not considered here anymore, since there is experimentally no exact control over the concentration x anyhow. Instead the two granularity models are examined in more detail.

7.2 Statistical Analysis of State Distributions

560 and 490 samples were evaluated for the atomic granularity (AG) and cell granularity (CG), respectively. Figure 16 provides a graphical analysis of some of the data obtained. The first row of Figure 16 shows histograms of the valence band, conduction band, and band-gap energies for the 560 atomic granularity samples using 30 samples per bin. The standard deviation of the valence band states ($\sigma_v^{AG} = 0.9meV$) is smaller than that of the conduction band states ($\sigma_c^{AG} = 1.4meV$). This difference arises from the heavier hole mass and the higher density of states as observed in bulk alloy simulations [Oyafuso, Klimeck, Bowen, Boykin (2002)]. The second row of Figure 16 compares the histograms of the first row (atomic granularity) to the corresponding histograms obtained in the model of cell granularity. Similar to the pure bulk alloy results [Oyafuso, Klimeck, Bowen, Boykin (2002)] the cell granularity results in standard deviations of the energies that are larger than the atomic granularity deviations($\sigma_c^{AG} = 2.8meV, \sigma_v^{AG} = 1.9meV$). For this particular dot size the difference between AG and CG is about a factor of two. Similar to the bulk simulations [Oyafuso, Klimeck, Bowen, Boykin (2002)] a change in the valence band state state energy average that is larger than the shift in the conduction band energy average can be observed ($\Delta E_v = E_v^{CG} - E_v^{AG} = 0.1938eV - 0.1862eV = 7.7meV, \Delta E_c = E_c^{CG} - E_c^{AG} = 1.2291eV - 1.2298eV = -0.8meV$). Again an overall reduction in the optical band gap is the result of the increased granularity ($\Delta E_G = E_G^{CG} - E_G^{AG} = 1.0352eV - 1.0436eV = -8.4meV$).

The band-gap deviation is roughly additive from the valence and conduction band deviations. This additive behavior indicates a correlation between the conduction and valence state energies. This correlation can be explored with a scatter plot of E_c versus E_c for all samples as shown in Figure 16g). A linear regression of the two scatter plots result in $E_c = 1.4685 - 1.2823E_v$ and $E_c = 1.405 - 1.1939E_v$, for the AG and CG distributions, respectively, with a good regression quality of $R \approx 0.8$. A relation between these strongly correlated energy values and the bulk band structure is discussed in the following Section 7.3.

The data can also be analyzed with respect to its dependence of the actual In concentration x in the $\mathrm{In_xGa_{1-x}As}$ alloy. Figure 16i) shows a histogram of the In distribution in the CG model. The expectation value is the desired 0.6, the standard deviation of 0.0016. This deviation is

[20] During the review process we have started to examine a possible GaAs buffer size dependence on the distribution function of the eigenenergies and have found that the dependence is not negligible. An increase in the GaAs buffer size decreases the spread in energy due. We are still in the process of exploring these data more carefully and plan to publish details of that study at a later time.

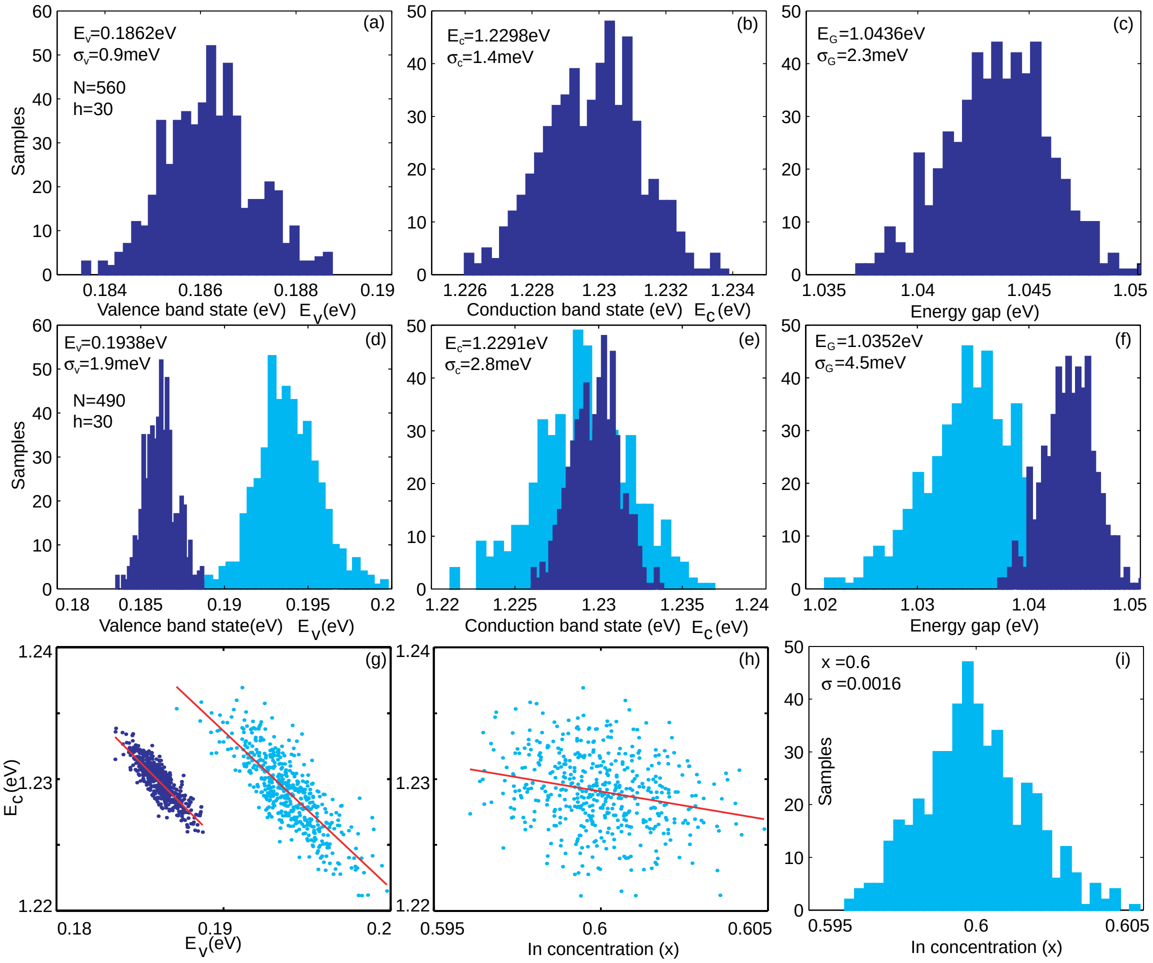

Figure 16 : First row: histogram distributions of 560 samples with 30 samples per bin for the valence band edge (a), conduction band edge (b), and band gap (c) using atomic granularity (AG). Second row: comparison of cell granularity (CG) disorder results to results in the first row based on 490 samples. (g) scatter plots of E_c versus E_v for all the samples shown in (a-b) and (d-e). Red solid lines indicate least mean square fit. (h) scatter plot of E_c as a function of actual In concentration for cell granularity. (i) histogram on actual In alloy concentration in the cell granularity.

purely determined by the statistical variation as a function of systems size [Oyafuso, Klimeck, Bowen, Boykin (2002)] (about 718,000 atoms in the dot in this case) and the random selection with expectation of 0.6. The statistical variation is certainly smaller than the experimental control on the alloy concentration and one can expect the experimental uncertainty in the alloy concentration to be significantly larger.

Figure 16h) shows a scatter plot of the conduction band state energies computed in the CG model as a function of their corresponding In concentration. The scatter plot appears more like a round blob, indicating a weak correlation, and indeed the linear regression to $Ec = 1.4871 - 0.43015x$ is characterized by a small regression quality $R = 0.24$. A similarly weak correlation can be found for the valence band states: $E_v = 0.04574 + 0.24682x$ with $R = 0.21$. This weak correlation of E_c and E_v with x is in contrast to a strong correlation we have seen in our bulk simulations [Oyafuso, Klimeck, Bowen, Boykin (2002)].

7.3 Quantized Energy Comparison Against Bulk Data

The quantized single particle energies in the quantum dot are determined by a multitude of influences. The first order effects are based on the underlying semiconductor band structure, the confinement by the heterostructure interfaces, the composition of the material and the size of the dots. Effects due to disorder and electron-electron interactions have to be considered second order effects. The discussion in this section examines the correlation between the quantized E_c and E_v energies shown in Figure 16g) and verifies that the quantized eigen-energies shown in Figure 16 fit within the underlying semiconductor band structure. This comparison serves as an overall sanity check and as a characterization of the relative importance of hydrostatic and bi-axial strain contributions to the quantized states in the quantum dot.

Figure 17a) shows the bulk conduction and valence band edge of $\mathrm{In_xGa_{1-x}As}$ as a function of In concentration x in a VCA approximation of a single unit cell. The graphs are computed using the same GaAs and InAs sp^3s^* parameter set that was used for the statistical quantum dot analysis in the section above. Three different strain conditions are evaluated: 1) unstrained bulk (dashed line), 2) hydrostatically compressed ($\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{zz}$) to the GaAs lattice constant (dotted line), and 3) bi-axially compressed ($\varepsilon_{xx} = \varepsilon_{yy} \neq \varepsilon_{zz}$) to a GaAs substrate in the x-y plane (solid line). It is interesting to see that the hydrostatically compressed InGaAs has a very small dependence on the In concentration. The same graph also shows the scatter plot of the quantized conduction and valence band energies as a function of actual alloy composition.

The single cross symbols indicate the average bottom of the conduction and the average top of the valence band in a small region of the center of the quantum dot (see the discussion of the spatially varying local band structure in section 6.4). The electron and valence band states show confinement energies of roughly 96meV and 50meV, respectively. The difference in the confinement energies ie of course expected due to the difference in the heavy-hole and electron masses. The placement in energy of the local conduction and valence band edge as well as the quantized state energies in the quantum dot indicate that the states inside the quantum dot are influenced by bi-axial as well as hydrostatic strain components combined.

Similar to to Figure 16h) the linear regression is shown as a red line running roughly parallel to the bulk bi-axial strain line. Little trust can be given to the linear regression lines due to their small quality values of $R \approx 0.21$. However, the slopes do not conflict with the hypothesis that the electronic state is dominated by the bi-axial strain shifts with some hydro-static strain shift contributions.

Figure 17b) plots the data of (a) on a E_c vs. E_v coordinate system. Within this reference frame Figure 16g) already show a strong correlation between the quantized E_v and E_c energies with a trustworthy linear regression of slope -1.1939. The unstrained, hydrostatically strained and bi-axially strained bulk slopes are: -10.747, -3.5726, and -1.3186, respectively. Again one can infer a strong influence on the confined states by the bi-axial strain. The individual square symbols indicate the alloy composition of 60% In and the cross indicates the average local band structure value in the middle of the dot.

7.4 Comparison Against Experimental Data

The numerical experiment shows a mean optical transition energy of about 1.04eV and a standard deviation, or associated linewidth of approximately 2 to 5meV assuming a fixed quantum dot size and a narrow In concentration distribution. This corresponds to an experimentally reported [Leon, Fafard, Piva (1998)] linewidth of 34.6meV at an optical transition energy of about 1.09eV.

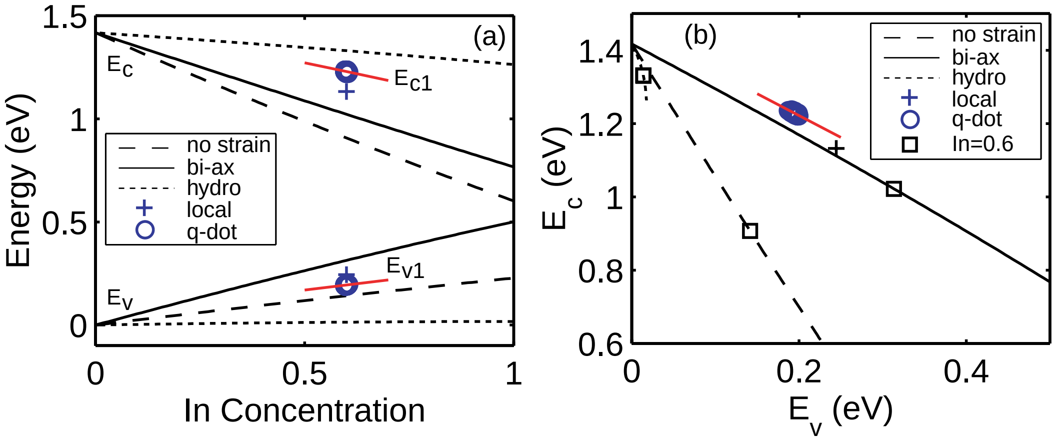

Figure 17 : (a) three sets of conduction and valence bands of $In_xGa_{1-x}As$ as a function of In concentration x. Dashed line: no strain, solid line: bi-axially strained to GaAs, dotted line: hydrostatically strained to GaAs. E_{c1} and E_{v1} quantum dot energy distributions (circles) similar to Figure 16h) with their linear regression fits (red lines). Locally averaged conduction and valence band edge inside a quantum dot is indicated with a plus sign. (b) conduction band energies plotted versus corresponding valence band edges from (a). Squares indicate an In concentration of 0.6.

The optical bandgap is underpredicted by about 50meV in the sp^3s^* model used in the alloy disorder study of Section 7. The $sp^3d^5s^*$ model used in Section 6 predicts an optical band gap of about 1.151eV, about 60meV too large compared to the experiment. Excitonic interactions will reduce the optical bandgap by about 10 to 30meV. With an experimentally observed optical line width of 34.6meV the $sp^3d^5s^*$ model is well within the experimental and theoretical errors. The experimental data do of course include quantum dot size variations, and the actual alloy concentration and alloy distribution are unknown. We plan to simulate larger sample space that does include quantum dot size variations and different alloy profiles [Sheng, Leburton (2001)] in the future. The major result of this simulation is the observation that there will be a significant optical line width variation due to alloy disorder alone, even if all the quantum dots were perfectly identical in size with a well-known alloy concentration.

8 Conclusion and Future Outlook

8.1 Summary

This paper presents the major theoretical, numerical, and software elements that have entered into the NEMO 3-D development over the past three years. The atomistic valence-force field method is used for the determination of atom positions in conjunction with the atomistic sp^3s^* and $sp^3d^5s^*$ tight-binding models to compute electronic structure in systems containing up to 16 million atoms. An eigenvalue solver that scales linearly with the number of atoms in the system has been demonstrated. Beowulf cluster computers are shown to be efficient computing engines for such electronic structure calculations. The electronic structure calculations require a significant RAM access by the CPU, and the Intel Pentium III benchmarks presented in this work show that dual CPU motherboards suffer from severe memory access problems. Faster computation completion can be obtained by leaving one of the CPUs on each board idle. This suggests that high memory use applications do not benefit at all from a dual Pentium III motherboard. Genetic algorithms are used to determine the empirical parameterization of the atomistic tight-binding models. The details of the new tight-binding model parameterizations for GaAs and InAs are discussed with respect to their unstrained and strained bulk properties. NEMO 3-D is used to study the effects of disorder in $In_xGa_{1-x}As$ bulk material and in $In_{0.6}Ga_{0.4}As$ quantum dots. The bulk properties are shown to be represented well within NEMO 3-D. The quantum dot simulations show significant distortions in the confined electron and hole wavefunctions introduced by random cation disorder. The distortion is more pronounced for the hole states than the electron states and it is not visible within a smooth virtual crystal approximation, which resembles non-atomistic methods. Over a thousand different alloy distributed quantum dots are simulated and a variation of the optical transition energy of several meV is observed.

8.2 Future Simulations

The strong variations in the electron and hole wavefunctions introduced by the alloy disorder beg for the evaluation of these effects on the optical matrix elements on these transitions. This is an issue that is planned for examination in the near future. The computation of the optical matrix elements in the incomplete tight-binding basis will follow the prescription given in references [].

Other possible studies include the comparison of the alloy disorder effects with a constant quantum dot size and shape with effects due to variations in size and shape. Also effects of strain fields due to neighboring quantum dots can now be simulated, since enough atoms can be included in the simulation domain.

A needed extension to NEMO 3-D is the inclusion of charge interactions in order to compute single electron charging energies and exciton binding energies. The typical way to calculate charge interaction matrix elements is based on the explicit usage of the wavefunction of the interacting states. However this is not as simple in the empirical tight-binding approach, since the actual orbital wavefunctions are unknown. However the computation of charging energies can be performed within two different frameworks: 1) projection of the tight-binding orbitals onto an explicit spatial orbital basis [Lee, Joensson, Klimeck (2001)], and 2) a modification to the operator representation in references.

8.3 Extension to Spintronic Simulations

In addition to charge transport and orbital wavefunction related relaxation times spin dependent transport and spin related relaxation times have recently received considerable attention. While important applications such as MRAM memory modules have reached the level of commercialization others such as spin based quantum computing have achieved outstanding visibility in the research community. It is therefore highly desirable to extend a nanoelectronic simulation tool such as NEMO 3-D to systems where spin effects are important.

Part of the spin related effects are already included in the tight-binding method described above through the spin-orbit coupling term in the Hamiltonian. A further step will be to include the direct coupling of the electron spin with an external magnetic field, which can originate directly from a source external to the device or from magnetic impurities embedded in the semiconductor. This interaction contributes a $\mathbf{S}\cdot\mathbf{H}$ term to the Hamiltonian and in a first approximation can be taken as an on-site diagonal term in the tight-binding Hamiltonian, where each spin band is now treated explicitly. Together with the vector potential contribution to the kinetic energy this addition should yield a Landé factor in good agreement with experiment and give a good description of charge transport for example in MRAMs.

The other spin-induced contribution to the Hamiltonian is the spin-spin interaction term proportional to $\sum \mathbf{S_i}\cdot\mathbf{S_j}$. Since a scattering event mediated by this interaction can change the spin of the electron, the self-energy and hence the single particle Green's function become non-diagonal. Despite this complication the general formalism of non-equilibrium Green's function still applies and transport properties including finite spin relaxation time can be obtained. This extension is particularly relevant for the simulation of devices probing the solid state implementation of quantum computing logic gates.

Acknowledgement: The work described in this publication was carried out at the Jet Propulsion Laboratory, California Institute of Technology under a contract with the National Aeronautics and Space Administration. The supercomputer used in this investigation was provided by funding from the NASA Offices of Earth Science, Aeronautics, and Space Science. We also would like to acknowledge very useful discussions and collaborations during the course of his project with Edith Huang, Edward S. Vinyard, Dr. Thomas A. Cwik, Dr. Charles Norton, Dr. Victor Decyk, and Dr. Dan Katz.

T. Wack is acknowledged for the review of the manuscript. We would also like to acknowledge the project support by Dr. Robert Ferraro, Dr. Benny Toomarian, and Dr. Leon Alkalai.

Reference

Anantram; Govindan (1998): Phys. Rev. B **58**, 4882.

Appelbaum, J.A.; Hamann, D.R. (1973): Phys. Rev. B **8**, 1777.

Bachelet, G.B.; Hamann, D.R.; Schlüter, M. (1982): Phys. Rev. B **26**, 4199.

Blanks etal. (1997): *NEMO: General Release of a New*

Comprehensive Quantum Device Simulator IEEE, New York.

Bowen, R.C.; et al. (1997a): J. Appl. Phys **81**, 3207.

Bowen, R.C.; etal. (1997): IEDM 1997, IEEE, New York, pp. 869–872.

Boykin, T. (1997): Phys. Rev. B **56**, 9613.

Boykin; Bowen; Klimeck (1999): Phys. Rev. B **23**, 15810.

Boykin; Bowen; Klimeck (2001): Phys. Rev. B **63**, 245314.

Boykin; Gamble; Klimeck; Bowen (1999): Phys. Rev. B **59**, 7301.

Boykin; Klimeck; Bowen; Lake (1997): Phys. Rev. B **56**, 4102.

Büttiker (1986): Phys. Rev. Lett. **57**, 1761.

Bowen; Frensley; Klimeck; Lake (1995): Phys. Rev. B **52**, 2754.

Canning, A. Wang, L.W.; Williamson, A.; Zunger, A, (2000) J of Comp. Physics **160**, 29.

Caroli; Combescot; Nozieres; Saint-James (1971): J. Phys. C **4**, 916.

Chen et al. (1994): Phys. Rev. B **50**, 8035.

Chen; Sher (1995): *Semiconductor Alloys* (Plenum Press, New York.

Cohen, M.L.; Bergstresser, T.K. (1966): Physical Review **141**, 789.

Damle; Ghosh; Datta (2001): Phys. Rev. B **6420**, 1403.

Demkov; Liu; Zhang; Loechelt (2000): J. of Vac. Sci. and Techn. B **18**, 2388.

Demkov; Sankey (1999): Phys. Rev. Lett. **83**, 2038.

Demkov; Zhang; Drabold (2001): Phys. Rev. B **6412**, 5306.

Derosa; Seminario (2001): J. of Phys. Chemistry B **105**, 471.

Didier Keymeulen et al. (2000): *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000), July 8-12, 2000, Las Vegas, Nevada USA*, edited by Whitley Darrell San Francisco.

Efros, A.L. and Rosen, M. (1998): Appl. Phys. Lett. **73**, 1155.

Fu, H.X.; Wang, L.W.; Zunger, A. (1998a) Phys. Rev. B **57**, 9971.

Fu, H.X.; Wang, L.W.; Zunger, A. (1998b) Appl. Phys. Lett. **73**, 1157.

Golub; Van Loan (1989): *Matrix Computations* (The Johns Hopkins Press, London.

Gropp; Lusk (1997): International Journal of Supercomputer Applications and High Performance Computing **11**, 103.

Gropp; Lusk; Skjellum (1997): *Using MPI: Portable Parallel Programming with the Message-Passing Interface* Maasachusetts Instituteof Technology, Boston.

Harrison, W.A (1999): *Elementary electronic structure* (World Scientific, Singapore ; River Edge, NJ.

Hohenberg, P.; Kohn, W. (1964): Phys. Rev. **136**, 864.

Hybertsen; Louie (1993): Phys. Rev. B **47**, 9973.

Jancu, J.M.; Scholz, R.; Beltram, F,; Bassani, F. (1998): Phys. Rev. B **57**, 6493.

Jones, R.O.; Gunnarsson, O. (1989): Rev. Mod. Phys. **61**, 689.

Keating, P. (1966): Phys. Rev. **145**.

Klimeck (2002): Journal of Computational Electronics in print.

Klimeck, Bowen, Boykin, Cwik (2000): Superl. and Microstr. **27**, 519.

Klimeck; Chen; Datta (1994): Phys. Rev. B. **50**, 2316.

Klimeck, G.; et al.(1997): 55th Annual Device Research Conference Digest, (IEEE, NJ, p. 92.

Klimeck, et al. (2000): Superl. and Microstr. **27**, 77.

Klimeck et al. (1995) Appl. Phys. Lett. **67**, 2539.

Klimeck et al. 91997): VLSI Design **8**, 79.

Klimeck; Lake; Datta; Bryant (1994): Phys. Rev. B. **50**, 5484.

Klimeck; Salazar-Lazaro; Stoica; Cwik (1999): *Materials in Space Science, Technology, and Exploration* Material Research Society, MRS Symposium Proceedings, Boston, Vol. 551, p. 149.

Kohn, W.; Sham, L.J. (1965): Phys. Rev. **140**, 1133.

Lake; Klimeck; Bowen; Jovanovic (1997): J. Appl. Phys. **81**, 7845.

Landauer (1970): Philos. Mag. **21**, 863.

Landolt-Bornstein (1982) *Numerical Data and Functions in Science and Technology*, Springer, Berlin, Vol. 22a.

Lee, Joensson, Klimeck (2001): Phys. Rev. B **63**, 195318.

Leon, Fafard, Piva (1998): Phys. Rev. B **58**, R4262.

Liu, H.C; Gao; McCaffrey (2001): , Appl. Phys. Lett. **78**, 79.

Löwdin (1950): J. Chem. **3**, 365.

Madelung (1996): *Semiconductors - Basic Data* Springer Verlag, Berlin, p. 317.

Martin (1970): Phys. Rev. B **1**, 4005.

McWeeny, R. (1992): *Methods of Molecular Quantum Mechanics* Academic Press, San Diego, p. 159.

Mikkelsen; Boyce (1982): Phys. Rev. Lett. **49**, 1412.

Mujica, Kemp, Roitberg, Ratner (1996): J. of Chem. Physics **104**, 7296.

Needs, R.J.; Foulkes, W.M.C.; Mitas, L.; Rajagopal, G. (2001): Rev. Mod. Phys. **73**, 33.

Ordejón, P.; Drabold, D.A.; Grumbach, M.P.; Martin, R.M. (1993): Phys. Rev. B **48**, 14646.

Ordejón, F.; Galli, G.; Car, R. (1993): Phys. Rev. B **47**, 9973.

Oyafuso; Klimeck; Bowen; Boykin (2002): Journal of Computational Electronics in print.

Pollak; Cardona (1968): Phys. Rev. **172**, 816.

Pryor (1998): Phys. Rev. B **57**, 7190.

Pryor, Kim, Wang (1998): J. of Appl. Phys. **83**, 2548.

RamanaMurty,M.V.; Atwater, H.A (1995): Phys. Rev. B **51**, 4889.

Ren (2001): Ph.D. thesis, Purdue University, School of Electrical and Computer Engineering.

Ren etal. (2001): *IEDM 2000* IEEE, New York, pp. 715–718.

Ren; Venugopal; Datta; Lundstrom (2001): *IEDM 2001* IEEE, New York.

Sankey, O; Niklewski, D.J. (1989): Phys. Rev. B **40**, 3979.

Schuurmans, M.F.H.; 't Hooft, G.W. (1985): Phys. Rev. B **31**, 8041.

Silver, Oreilly (1995): IEEE J. of Quant. Elec. **31**, 1193.

Sheng; Leburton (2001): Phys. Rev. B **6316**, 1301.

Slater; Koster (1954): Physical Review **94**, 1498.

Stadele; Tuttle; Hess (2001): J of Appl. Phys. **89**, 348.

Stier, Grundmann, Bimberg (1999): Phys. Rev. B **59**, 5688.

Stopa (1996): Phys. Rev. B **54**, 13767.

Troullier, N.; Martins, J.L. (1991): Phys. Rev. B **43**, 1993.

Vanderbilt, D. (1990): Phys. Rev. B **41**, 7892.

Van de Walle (1989): Phys. Rev. B **39**, 1871.

Vogl, P.; Hjalmarson, H.P.; Dow, J.D. (1983): J. Phys. Chem. Solids **44**, 365.

von Allmen, P. (1992a) Phys. Rev. B **46**, 15382.

von Allmen, P. (1992b): Phys. Rev. B **46**, 15377.

Wang, L.W.; Kim, J.N.; Zunger, A. (1999): Phys. Rev. B **59**, 5678.

Wang; Zunger (1994): J. of Chem. Physics **100**, 2394.

Williamson, A.J.; Wang, L.W.; Zunger, A. (2000) Phys. Rev. B **62**, 12963.