Atomistic Simulation of Realistically Sized Nanodevices Using NEMO 3-D: Part I – Models and Benchmarks

Gerhard Klimeck[1,2], Shaikh Ahmed[1], Neerav Kharche[1], Hansang Bae[1], Steve Clark[1], Benjamin Haley[1], Sunhee Lee[1], Maxim Naumov[1], Hoon Ryu[1], Faisal Saied[1], Marta Prada[1], Marek Korkusinski[3], and Timothy B. Boykin[4]

***Abstract*—Device physics and material science meet at the atomic scale of novel nanostructured semiconductors and the distinction between new device or new material is blurred. Not only are quantum mechanical effects in the electronic states of the device important, but also the granular, atomistic representation of the underlying material. Approaches based on a continuum representation of the underlying material typically used by device engineers and physicists become invalid. *Ab-initio* methods used by material scientists typically do not represent the bandgaps and masses precisely enough for device design or they do not scale to realistically large device sizes. The plethora of geometry, material, and doping configurations in semiconductor devics at the nanoscale suggests that a general nanoelectronic modeling tool is needed. The Nanoelectronic Modeling tool (NEMO 3-D) has been developed to address these needs. Based on the atomistic valence-force field (VFF) and a variety of nearest-neighbor tight-binding models (*s*, *sp^3s^**, *$sp^3d^5s^*$*) NEMO 3-D enables the computation of strain and electronic structure for over 64 and 52 million atoms, corresponding to volumes of (110nm)^3 and (101nm)^3, respectively. The physical problem may involve very large-scale computations and NEMO 3-D has been designed and optimized to be scalable from single CPUs to large numbers of processors on commodity clusters and supercomputers. NEMO 3-D has been released with an open source license in 2003 and is continually developed by the Network for Computational Nanotechnology (NCN). A web-based online interactive version for educational purposes is freely available on the NCN portal www.nanoHUB.org. In this article, theoretical models, essential algorithmic and computational components that have been used in the development and successful deployment of NEMO 3-D are discussed**.

***Index Terms*—Atomistic simulation, NEMO 3-D, Nanostructures, Strain, Piezoelectricity, Valley splitting, Quantum computation, Tight binding, Keating model**.

I. Introduction

EMERGENCE of nanodevices. The rapid progress in nanofabrication technologies has led to the emergence of new classes of nanodevices and structures which are expected to bring about fundamental and revolutionary changes in electronic, photonic, biotechnology, information processing and computation, and medicine industries. These devices demonstrate new capabilities and functionalities where the *quantum nature* of charge carriers play an important role in determining the overall device properties and performance. The device sizes have already reached the level of tens of nanometers. In this regime, the *atomistic granularity* of constituent materials cannot be neglected: effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance. In effect the formerly disjoint fields of semiconductor devices and materials science meet at the atomic scale.

A critical facet of the nanodevices development is the creation of simulation tools that can quantitatively explain or even predict experiments. In particular it would be very desirable to explore the design space before or in conjunction with the (typically time consuming and expensive) experiments. A general tool that is applicable over a large set of materials and geometries is highly desirable. But just the tool development itself is not enough. The tool needs to be deployed to the user community so it can be made more reliable, flexible, and accurate. The main goal of this paper is to describe the theoretical models and the essential algorithmic and computational components that have been used in the development and successful deployment of NEMO 3-D on nanoHUB.org. Of particular importance, presented are some of the new capabilities that have been recently added to NEMO 3-D to make it one of the premier simulation tools for design and analysis of realistically-sized nanoelectronic devices, and therefore to make it a valid tool for the computational nanotechnology community. These recent advances include algorithmic refinements, performance analysis to identify the best computational strategies, and memory saving measures. Demonstrated is the effective scalability of NEMO 3-D code on the BlueGene, an Intel Woodcrest cluster, the Cray XT3 and other Linux clusters. The largest electronic structure calculation with *52 million atoms* involved a Hamiltonian matrix over one *billion* complex degrees of freedom. Compared is the performance with a stored Hamiltonian vs. re-computing the matrix each time it is needed. Through a set of

[1] School of Electrical and Computer Engineering and Network for Computational Nanotechnology, Purdue University, West Lafayette, IN 47907, USA. Tel: (765) 494 9212, Fax: (765) 494 6441, E-mail: gekco@purdue.edu

[2] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109.

[3] Institute for Microstructural Sciences, National Research Council of Canada, 1200 Montreal Road, Ottawa, Ontario K1A 0R6.

[4] *Electrical and Computer Engineering Dept., The University of Alabama in Huntsville, Huntsville, AL 35899.*

end-to-end calculations, it is shown how the eigenvalues vary as a function of the size of the domain. We describe the state-of-the-art algorithms that have been incorporated in the code, including a very effective Lanczos eigenvalue solver, and present a comparison of the different solvers. While such system sizes of tens of millions of atoms appear at first sight huge and wasteful, we claim here that some physical problems require such large scale analysis. We recently demonstrated [1] that the analysis of valley splitting in strained Si quantum wells grown on strained SiGe required atomistic analysis of 10 million atoms to match experimental data. The insight that disorder in the SiGe buffer increases valley splitting in the Si quantum well would probably not be predictable in a continuum effective mass model.

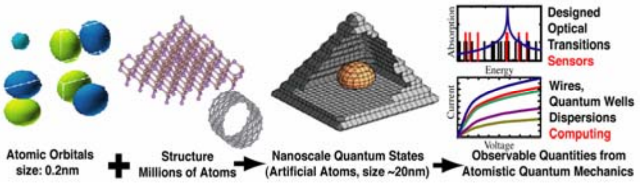

Fig. 1. NEMO 3-D modeling agenda: maps electronic properties of individual atoms into realistic structures containing millions of atoms, computation of nanoscale quantum dots that maps into real applications.

II. Modeling And Simulation Challenges

The theoretical knowledge of the electronic structure of nanoscale semiconductor devices is the first and essential step towards the interpretation and the understanding of the experimental data and reliable device design at the nanometer scale. Following is a list of the modeling and simulation challenges in the design and analysis of realistically-sized engineered nanodevices.

(1) *Full three-dimensional atomistic representation*: The lack of *spatial symmetry* in the overall geometry of the nanodevices usually requires explicit three-dimensional representation. For example, Stranski-Krastanov growth techniques tend to produce self-assembled InGaAs/GaAs quantum dots [2][3][4][5] with cylindrical-like shape symmetry, e.g. disks, truncated cones, domes, or pyramids [6]. These geometries are generally not perfect geometric objects, since they are subject to interface interdiffusion, and discretization on an atomic lattice. There is no such thing as a round disk on a crystal lattice! The underlying crystal symmetry imposes immediate restrictions on the realistic geometry and influences the quantum mechanics. Continuum methods such as effective mass [7] and *k.p* [8][9] typically ignore such crystal symmetry and atomistic resolution. The required simulation domain sizes of ~1M atoms prevent the usage of *ab initio* methods. Empirical methods which eliminate enough unnecessary details of core electrons, but are finely tuned to describe the atomistically dependent behavior of valence and conduction electrons are needed. The current state-of-the-art leaves 2 choices: 1) pseudopotentials [10] and 2) Tight Binding [11]. Both methods have their advantages and disadvantages. Pseudopotentials use plane waves as a fundamental basis choice. Realistic nanostructures contain high frequency features such as alloy-disorder or hetero-interfaces. That means that the basis needs to be adjusted (by an expert) for every different device, which limit the potential impact for non-expert users. Numerical implementations of pseudopotential calculations typically require a Fourier transform between real and momentum space which demand full matrix manipulations and full transposes. This typically requires high bandwidth communication capability (i.e. extremely expensive) parallel machines, which limit the practical dissemination of the software to end users with limited compute resources. Tight-binding is a local basis representation, which naturally deals with finite device sizes, alloy-disorder and hetero-interfaces and it results in very sparse matrices. The requirements of storage and processor communication are therefore minimal compared to pseudopotentials and actual implementations perform extremely well on cheap clusters [11]. Tight-binding has the disadvantage that it is based on empirical fitting and the community continues to raise the issue on the fundamental applicability of tight-binding. The NEMO team has spent a significant effort to expand and document the tight-binding capabilities with respect to handling of strain [12], electromagnetic fields [13], and Coulomb matrix elements [14] and fit them to well known and accepted bulk parameters [11][15][16]. With tight-binding the NEMO team was able early on to match experimentally verified, high-bias current-voltage curves of resonant tunneling [17][18] that could not get modeled by ether effective mass (due to the lack of physics) or pseudopotential methods (due to the lack of open boundary conditions). We continue to learn about the tight-binding method capabilities and are in the process of benchmarking it against more fundamental *ab-initio* approaches and pseudopotential approaches. Our current Si/Ge parameterization is described in references [19][20]. Figure 1 depicts a range of phenomena that represent new challenges presented by new trends in nanoelectronics and lays out the NEMO 3-D modeling agenda.

(2) *Atomistic strain*: Strain that originates from the assembly of lattice-mismatched semiconductors strongly modifies the energy spectrum of the system. In the case of the InAs/GaAs quantum dots, this mismatch is around 7% and leads to a strong *long-range* strain field within and very wide reaching (typically ~ 25 nm) around each quantum dot [21]. Si/Ge core/shell structured nanowires are another example of strain dominated atom arrangements [22] and Si-based quantum well quantum computing architectures rely on strain for state separation [23]. The strain can be atomistically inhomogeneous, involving not only biaxial components but also non negligible shear components. Strain strongly influences the core and barrier material band structures, modifies the energy

band gaps and lifts the heavy hole-light hole degeneracy at the zone center. In the nanoscale regime, the classical harmonic linear/continuum elasticity model for strain is inadequate and device simulations must include the fundamental quantum character of charge carriers and the long-distance atomistic strain effects with proper boundary conditions on equal footing [24][25].

(3) *Piezoelectric field*: A variety of advanced materials such as GaAs, InAs, GaN of interest are piezoelectric. Any spatial distortions in nanostructures made of these materials will create significant piezoelectric fields, which will significantly modify the electrostatic potential landscape. Recent spectroscopic analyses of self-assembled QDs demonstrate polarized transitions between confined hole and electron levels [6]. While the continuum models (effective mass or *k.p*) can reliably predict aspects of the single-particle energy states, they fail to capture the observed non-degeneracy and optical polarization anisotropy of the excited energy states in the (001) plane. These methods fail because they use a confinement potential which is assumed to have only the *shape symmetry* of the nanostructure and they ignore the underlying crystal symmetry. However, experimentally noticeable is the fact that the true symmetry is lower than the assumed continuum symmetry because of (a) underlying crystalline symmetry, (b) atomistic strain relaxation and (c) piezoelectric field. For example, in the case of pyramid shaped quantum dots with square bases, continuum models treat the underlying material in C_{4v} symmetry while the atomistic representation lowers the crystal symmetry to C_{2v}. Piezoelectric potential originating from the non-zero shear component of the strain field must be taken into account to properly model the associated symmetry breaking and the introduction of a global shift in the energy spectra of the system..

III. NEMO 3-D Simulation Package

(A) *Basic Features*

NEMO 3-D [11][26][27][28][29] bridges the gap between the large size, classical semiconductor device models and the molecular level modeling. This package currently allows calculating single-particle electronic states and optical response of various semiconductor structures including bulk materials, quantum dots, quantum wires, quantum wells and nanocrystals. NEMO 3-D includes spin in its fundamental atomistic tight binding representation. Spin is therefore not added in as an afterthought into the theory, but spin-spin interactions are naturally included in the Hamiltonian. Effects of interaction with external electromagnetic fields are also included [11][30][13]. This paper focuses on the design and performance of NEMO 3-D illustrated on the case of InAs quantum dots embedded in a GaAs barrier material. A schematic view of the sample is presented in Figure 2. The quantum dot is positioned on a 0.6 nm thick wetting layer (dark region). The simulation of strain is carried out in the large computational box D_{strain}, while the electronic structure computation is restricted to the smaller domain D_{elec}. In part-II of this paper it has been shown that under the assumptions of realistic boundary conditions, strain is long-ranged and penetrates around 25 nm into the dot substrate thus stressing the need for using large substrate thickness in the simulations. NEMO 3-D enables the computation of strain and electronic structure in an atomistic basis for over 64 and 52 million atoms, corresponding to volumes of $(110\text{nm})^3$ and $(101\text{nm})^3$, respectively. These volumes can be spread out arbitrarily over thin layer geometry. For example, if a thin layer of 15 nm height is considered, the corresponding widths in the *x-y* plane correspond to 298 nm for strain calculations and 262 nm for electronic structure calculations. No other atomistic tool can currently handle such volumes needed for realistic device simulations. NEMO 3-D runs on serial and parallel platforms, local cluster computers as well as the NSF Teragrid.

(B) *Components and Models*

The NEMO 3-D program flow consists of four main components.

(1) *Geometry construction*. The first part is the geometry constructor, whose purpose is to represent the treated nanostructure in atomistic detail in the memory of the computer. Each atom is assigned three single-precision numbers representing its coordinates, stored is also its type (atomic number in short integer), information whether the atom is on the surface or in the interior of the sample (important later on in electronic calculations), what kind of computation it will take part of (strain only or strain and electronic), and what its nearest neighbor relation in a unit cell is. The arrays holding this structural information are initialized for all atoms on all CPUs, i.e., the complete information on the structure is available on each CPU. By default most of this information can be stored in short integer arrays or as single bit arrays, which does not require significant memory. This serial memory allocation of the atom positions, however, becomes significant for very large systems which must be treated in parallel. A compile option exists in the code to use a parallelized atom position storage scheme, which limits some output capabilities, but provides significant memory savings.

(2) *Strain.* The materials making up the QD nanostructure may differ in their lattice constants; for the InAs/GaAs system this difference is of the order of 7%. This lattice mismatch leads to the appearance of strain: atoms throughout the sample are displaced from their bulk positions. Knowledge of equilibrium atomic positions is crucial for the subsequent calculation of QD's electronic properties, which makes the computation of strain a necessary step in realistic simulations of these nanostructures.

NEMO 3-D computes strain field using an atomistic valence force field (VFF) method [31] with the Keating Potential. In this approach, the total elastic energy of the sample is computed as a sum of bond-stretching and bond-bending contributions from each atom. The local strain energy at atom *i* is given by a phenomenological formula

$$E_i = \frac{3}{8}\sum_j \left[\frac{\alpha_{ij}}{2d_{ij}^2}\left(R_{ij}^2 - d_{ij}^2\right)^2 + \sum_{k>j}^{n} \frac{\sqrt{\beta_{ij}\beta_{ik}}}{d_{ij}d_{ik}}\left(\vec{R}_{ij}\cdot\vec{R}_{ik} - \vec{d}_{ij}\cdot\vec{d}_{ik}\right)^2 \right], \quad (1)$$

where the sum is carried out over the n nearest neighbors j of atom i, $\vec{d}_{ij}$ and $\vec{R}_{ij}$ are the bulk and actual (distorted) distances between neighbor atoms, respectively, and α_{ij} and β_{ij} are empirical material-dependent elastic parameters. The equilibrium atomic positions are found by minimizing the total elastic energy of the system. Several other strain potentials [24] [25] are also implemented in NEMO 3-D. While they modify some of the strain details they roughly have the same computational efficiency.

(3) *Electronic structure.* The single-particle energies and wave functions are calculated using an empirical first-nearest-neighbor tight-binding model. The underlying idea of this approach is the selection of a basis consisting of atomic orbitals (such as s, p, d, and s^*) centered on each atom. These orbitals are further treated as a basis set for the Hamiltonian, which assumes the following form:

$$\hat{H} = \sum_i \varepsilon_i^{(\nu)} c_{i,\nu}^{+} c_{i,\nu} + \sum_{i,\nu,\mu} t_i^{(\nu\mu)} c_{i,\nu}^{+} c_{i,\mu} + \sum_{i,j,\nu,\mu} t_{ij}^{(\nu\mu)} c_{i,\nu}^{+} c_{j,\mu}, \quad (2)$$

where $c_{i,\nu}^{+}$ ($c_{i,\nu}$) is the creation (annihilation) operator of an electron on the orbital ν localized on atom i. In the above equation, the first term describes the onsite orbital terms, found on the diagonal of the Hamiltonian matrix. The second term describes coupling between different orbitals localized on the same atom (only the spin-orbit coupling between p-orbitals), and the third term describes coupling between different orbitals on different atoms. The restriction in the summation of the last term is that the atoms i and j be nearest neighbors.

The characteristic parameters ε and t are treated as empirical fitting parameters for each constituent material and bond type. They are usually expressed in terms of energy constants of σ and π bonds between the atomic orbitals. For example, for a simple cubic lattice, the interaction between the s orbital localized on the atom i at origin and the orbital p_x localized on the atom j with coordinate $\vec{d}_{ij} = a\hat{x}$ with respect to the atom i would simply be expressed as $t_{ij}^{(s,p_x)} = V_{sp\sigma}$. Most of the systems under consideration, however, crystallize in the zinc-blende lattice, which means that the distance between the nearest neighbors is described by a 3-D vector $\vec{d}_{ij} = l\hat{x} + m\hat{y} + n\hat{z}$, with l, m, n being the directional cosines. These cosines rescale the interaction constants, so that the element describing the interaction of the orbitals s and p_x is $t_{ij}^{(s,p_x)} = lV_{sp\sigma}$. The parameterization of all bonds using analytical forms of directional cosines for various tight-binding models is given in Ref. [32]. NEMO 3-D provides the user with choices og the $sp^3d^5s^*$, sp^3s^*, and single s-orbital models with and without spin, in zincblende, wurzite, and simple cubic lattices.

Additional complications arise in strained structures, where the atomic positions deviate from the ideal (bulk) crystal lattice [33]. The presence of strain leads to distortions not only of bond directions, but also bond lengths. In this case, the discussed interaction constant $t_{ij}^{(s,p_x)} = l' V_{sp\sigma}\left(\frac{d}{d_0}\right)^{\eta(sp\sigma)}$, where the new directional cosine l' can be obtained analytically from the relaxed atom positions, but the bond-stretch exponent $\eta(sp\sigma)$ needs to be fitted to the experimental data. The energy constants parameterizing the on-site interaction change as well due to bond renormalization [11][12].

The 20-band nearest-neighbor tight-binding model is thus parameterized by 34 energy constants and 33 strain parameters, which need to be established by fitting the computed electronic properties of materials to those measured experimentally. This is done by considering bulk semiconductor crystals (such as GaAs or InAs) under strain. The summation in the Hamiltonian for these systems is done over the primitive crystallographic unit cell only. The model makes it possible to compute the band structure of the semiconductor throughout the entire Brillouin zone. For the purpose of the fitting procedure, however, only the band energies and effective masses at high symmetry points are targeted, and the tight-binding parameters are adjusted until a set of values closely reproducing these target values is found. Search for optimal parameterization is done using a genetic algorithm, described in detail in Refs. [11][23]. Once it is known for each material constituting the QD, a full atomistic calculation of the single-particle energy spectrum is carried out on samples composed of millions of atoms. No further material properties are adjusted for the nanostructure, once they are defined as basic bulk material properties.

(4) *Post processing of QD eigenstates* . From the single-particle eigenstates various physical properties can be calculated in NEMO 3-D such as optical matrix elements [34], Coulomb and exchange matrix elements [14], approximate single cell bandstructures from supercell bandstructure [36][37][38].

(C) *Algorithmic and Numerical Aspects*

(1) *Parallel implementation*. The complexity and generality of physical models in NEMO 3-D can place high demands on computational resources. For example, in the 20-band electronic calculation the discrete Hamiltonian matrix is of order 20 times the number of atoms. Thus, in a computation with 20 million atoms, the matrix is of order 400 million. Computations of that size can be handled because of the parallelized design of the package. NEMO 3-D is implemented in ANSI C, C++ with MPI used for message-passing, which ensures its portability to all major high-performance

computing platforms, and allows for an efficient use of distributed memory and parallel execution mechanisms.

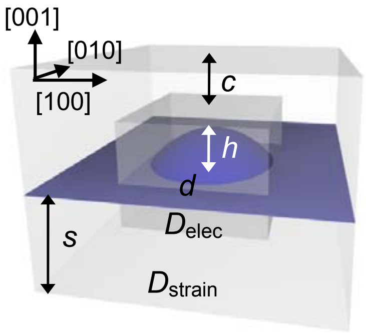

Fig. 2. Simulated dome shaped InAs/GaAs quantum dot. Two simulation domains are shown. D_{elec}: central domain for electronic structure calculation, and D_{strain}: larger/outer domain for strain calculation. In the figure: *s* is the substrate height, *c* is the cap layer thickness, *h* is the dot height, *d* is the dot diameter.

Although the strain and electronic parts of the computation are algorithmically different, the key element in both is the sparse matrix-vector multiplication. This allows the use of the same memory distribution model in both phases. The computational domain is divided into vertical slabs. All atoms from the same slab are assigned to a single CPU, so if all nearest neighbors of an atom belong to its slab, no inter-CPU communication is necessary. The interatomic couplings are then fully contained in one of the diagonal blocks of the matrix. On the other hand, if an atom is positioned on the interface between slabs, it will couple to atoms belonging both to its own and the neighboring slab. This coupling is described by the off-diagonal blocks of the matrix. Its proper handling requires inter-CPU communication. However, due to the first-nearest-neighbor character of the strain and electronic models, the messages need to be passed only between pairs of CPUs corresponding to adjacent domains – even if the slabs are one atomic layer thick. Full duplex communication patterns are implemented such that all inter-processor communications can be performed in 2 steps [11].

(2) *Core Algorithms and Memory requirements.* In the strain computation, the positions of the atoms are computed to minimize the total elastic strain energy. The total elastic energy in the VFF approach has only one, global minimum, and its functional form in atomic coordinates is quartic. The conjugate gradient algorithm in this case is well-behaved and stable. This is done using the Conjugate Gradient minimization algorithm. The total elastic energy is never stored in its matrix form, but the interatomic couplings are computed on the fly. Therefore the only data structures allocated in this phase are the vectors necessary for the conjugate gradient. The implementation used in NEMO 3-D requires six vectors, each of the total size of 3 × number of atoms (to store atomic coordinates, gradients, and intermediate data), however all those vectors are divided into slabs and distributed among CPUs as discussed above. The final atom position vectors are by default stored on all the CPU for some technical output details. They can be distributed to the various CPUs at compile time resulting in reduced output capabilities.

The electronic computation involves a very large eigenvector computation (matrices of order of hundreds of millions or even billion). The algorithms/solvers available in NEMO 3-D include the PARPACK library [39], a custom implementation of the Lanczos method [40], the spectrum folding method [41] and the Tracemin [42]. The research group is also working on implementation of Lanczos with deflation, Block Lanczos and Jacobi-Davidson [43] methods.

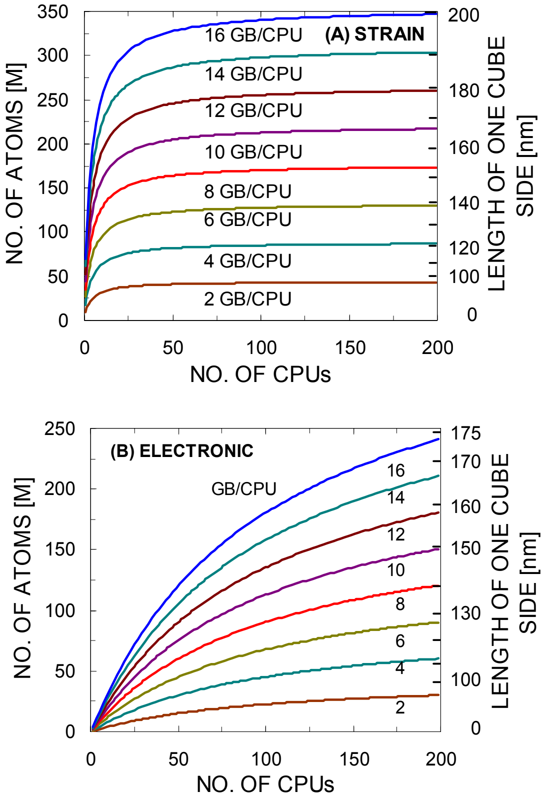

Fig. 3. Number of atoms that can be treated, as a function of the number of CPUs for different amounts of memory per CPU. The plot on top is for the strain calculation, and the one on the bottom is for the electronic structure calculation. The vertical axis on the right side of each plot gives the equivalent length in nm of one side of the cube that would contain the given number of atoms.

The Lanczos algorithm employed here is not restarted, and the Lanczos vectors are not reorthogonalized. Moreover, the spectrum of the matrix has a gap, which lies in the interior of the spectrum. Typically, a small set of eigenvalues is sought, immediately above and below the gap. The corresponding eigenstates are electron and hole wave functions, assuming effectively nonzero values only inside and in the immediate vicinity of the dot. Also, in the absence of the external

TABLE I

Performance on *P* processors: time (in seconds), number of matrix vector multiplications (# mvs), memory (mem.) and number of correct eigenvalues times their multiplicity (# eigs) for Lanczos, Tracemin and PARPACK *k* eigenvalue solvers in NEMO 3-D software package.

	LANCZOS				TRACEMIN				PARPACK			
P	*time*	*# mvs*	*mem.*	*# eigs*	*Time*	*# mvs*	*mem.*	*# eigs*	*time*	*# mvs*	*mem.*	*# eigs*
1	197.2	7000	55.15	14	14298.0	390000	227.15	14	2155.9	11700	177.83	11
2	121.9	7000	28.11	14	8341.7	400000	114.28	14	579	8300	89.58	11
3	89.0	7000	18.95	14	5480.0	390000	76.20	14	412	8900	59.85	11
4	75.3	7000	14.49	14	4346.9	400000	57.66	14	912.8	26900	45.35	12

TABLE II

Spectrum of the eigenvalues around 0 (with correct multiplicity 2) and eigenvalue multiplicity obtained by the Lanczos, Tracemin and PARPACK eigenvalue solvers. *Number* of searched eigenvalues was kept constant for these three methods.

Eigenvalues	LANCZOS	TRACEMIN	PARPACK
-1.72338200E-01	1		1
-1.66029400E-01	1		1
-1.59010400E-01	1		1
-1.47522100E-01	1		1
-1.38917800E-01	1		1
-1.17807000E-01	1		1
-1.01703700E-01	1	2	2
-7.80348200E-02	1	2	2
-5.17194400E-02	1	2	1
-2.81959000E-03	1	2	
3.93045300E-02	1	2	
7.66237500E-02	1	2	
1.16104700E-01	1	2	
1.57112300E+00	1		

TABLE III

Specifications for the HPC platforms used in the performance comparisons.

Platform	*Type*	*CPU*	*Interconnect*	*Location*
PU/Xeon64	Linux Cluster	Xeon x86-64 3.2GHz	Gigabit Ethernet	RCAC
PU/Xeon32	Linux Cluster	Xeon 3.06GHz	Gigabit Ethernet	RCAC
PU/Woodcrest	Linux Cluster	Xeon x86-64 Dual Core 2.33GHz	Gigabit Ethernet	RCAC
PSC/XT3	Cray XT3	Opteron x86-64 2.6GHz	Native	PSC
NCSA/Altix	SGI Altix	Itanium2 IA-64 1.6GHz	SGI NUMAlink	NCSA
RPI/BGL	BlueGene/L	PowerPC 440 0.7 GHz	Native	RPI

TABLE IV

Iteration counts for the Lanczos computation as a function of system size.

NUMBER OF ATOMS (MILLIONS)	0.89	1.99	3.92	6.80	16.18	21.07	52.57
Number of Lanczos iterations	5,061	5,121	6,141	7,921	9,621	10,401	14,691

magnetic field the eigenvalues are repeated, which reflects the spin degeneracy of electronic states. The advantage of Lanczos algorithm is that it is fast, while the disadvantage is that it does not find the multiplicity and can potentially miss eigenvalues. Some comparisons have shown that the Lanczos method is faster by a factor of 10 for the NEMO 3-D matrix than PARPACK. Tracemin algorithm finds the correct spectrum of degenerate eigenvalues, but is slower than Lanczos. PARPACK has been found to be less reliable for this problem, taking more time than Lanczos and missing some of the eigenvalues and their multiplicity. Tables I and II give a comparison of Lanczos, PARPACK and Tracemin (the *number* of eigenvalues searched was kept constant). The majority of the memory allocated in the electronic calculation in Lanczos is taken up by the Hamiltonian matrix. This matrix is very large, but typically very sparse; this property is explicitly accounted for in the memory allocation scheme. All matrix entries are, in general, complex, and are stored in single precision. The code has an option to not store the Hamiltonian matrix, but to recompute it, each time it needs to be applied to a vector. In the Lanczos method, this is required once in each iteration. The PARPACK and Tracemin algorithms require the allocation of a significant number of vectors as a workspace, which is comparable to or larger than the Hamiltonian matrix. This additional memory need may require a matrix recomputed for memory savings.

Figure 3 shows the memory requirements for the two main phases of the code (strain and electronic structure calculations). It shows how the number of atoms that can be treated grows as a function of the number of CPUs, for a fixed amount of memory per CPU. The number of atoms can be intuitively characterized by the length of one side of a cube that would contain that many atoms. This length is shown in Figure 3, on the vertical axis on the right side of each plot. This figure shows that for a given amount of memory per CPU in the strain calculation (shown in the left plot), the number of atoms that can be handled levels off after a certain CPU count, whereas for the electronic structure calculation (shown in the right plot), the number of atoms that can be treated in NEMO 3-D continues to grow for larger CPU counts. The unfavorable memory scaling in the strain calculation is due to the allocation of all the atom positions on a single CPU. Distribution of this memory is possible at compile time but has limited output capability. The strain calculations have so far never been memory limited. NEMO 3-D is typically size limited in the electronic structure calculation.

(3) *Scaling*. Out of the two phases of NEMO 3-D, the strain calculation is algorithmically and computationally simpler. The Lanczos diagonalization of the Hamiltonian matrix, on the other hand, is much more challenging computationally.

To investigate the performance of NEMO 3-D package, computation was performed in a single dome shaped InAs quantum dot nanostructure embedded in a GaAs barrier material as shown in Figure 2. The HPC platform used in the performance studies are shown in Table III. These include three Linux clusters at the Rosen Center for Advanced Computing (RCAC) at Purdue with Intel processors (32 bit Xeon, 64 bit Xeon and dual core Woodcrest). The PU/Woodcrest cluster has two dual core chips per node. The other three platforms are a BlueGene at the Rensselaer Polytechnic Institute (RPI), the Cray XT3 at the Pittsburgh Supercomputing Center (PSC) and the SGI Altix at the National Center for Supercomputing Applications (NCSA). The processors on the Altix are Intel Itanium 2 processors, on the BlueGene they are IBM PowerPC's while the Cray XT3 has AMD Opterons. These three platforms have proprietary interconnects, that are higher performance than Gigabit Ethernet (GigE) for the three Linux clusters at Purdue. In the following, the terms processors and cores are used interchangeably.

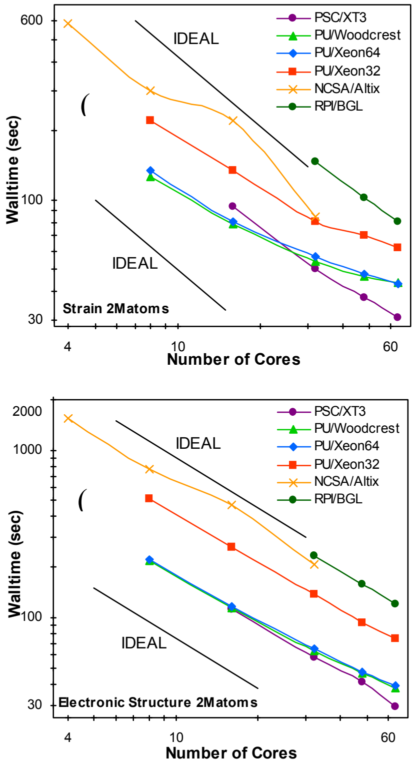

Fig. 4. Parallel performance of NEMO 3-D on some HPC platforms.

Figure 4 shows the performance of NEMO 3-D for each of the architectures. The wall clock times for 100 iterations for the energy minimization in the strain phase and 100 iterations of the Lanczos method for the electronic structure phase are shown as a function of the number of cores. The problem is a benchmark problem with 2 million atoms. Figure 4 shows that the PU/Woodcrest cluster and the PU/Xeon64 cluster are very close in performance for the same number of cores. These are both close to the performance of the Cray XT3 for lower core counts, while the XT3 performs better for higher core counts, due to its faster interconnect. The older cluster, PU/Xeon32, is slower by a factor of about 2-2.3 compared to the Woodcrest cluster. The BlueGene's slower performance is consistent with its lower clock speed, while the scalability reflects its efficient interconnect. The performance of the Altix is lower than expected.

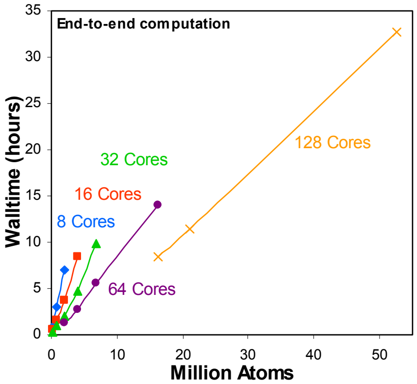

Fig. 5. Wall clock time vs. number of atoms for end-to-end computations of the electronic structure of a quantum dot, for various numbers of cores on the PU/Woodcrest cluster.

In addition to the performance for the benchmark cases, with 100 iterations in the strain and electronic structure cases, end-to-end runs on the PU/Woodcrest cluster are carried out next (Figure 5). This involves iterating to convergence and computing the eigenstates in the desired range (conduction band and valence band). For each problem size, measured in millions of atoms, the end-to-end cases were run to completion, for one choice of number of cores. The iteration counts for the Lanczos computation are given in Table IV.

The numerical experiment is designed to demonstrate NEMO 3-D's ability to extract targeted interior eigenvalues and vectors out of virtually identical systems of increasing size. A single dome shaped InAs quantum dot embedded in GaAs is considered. The GaAs buffer is increased in size to increase the dimension of the system while not affecting confined states in the QD. It is verified [44] that the eigenvectors retain the expected symmetry of the nanostructure.

(D) *Visualization.* The quantum dot simulation data of NEMO 3-D contains multivariate wave functions and strain profiles of the device structure. For effective 3-D visualizations of these results, a hardware-accelerated direct volume rendering system [45] has been developed, which is combined with a graphical user interface based on *Rappture*[1]. This visualization system uses data set with *OPEN-DX*[2] format that are directly generated from NEMO 3-D. Figure 6 shows the wave functions of electron on the first 4 eigenstates in conduction band of quantum dot which has 268800 atoms in the electronic domain.

(E) *Release and Deployment of NEMO 3-D Package.* NEMO 3-D was developed on Linux clusters at the Jet Propulsion Lab (JPL) and was released with an open source license in 2003. The originally released source appears to be no longer hosted at openchannelfoundation.org web site. As NEMO 3-D is undergoing further developments by the NCN we are planning future releases of the NEMO 3-D source through nanoHUB.org. NEMO 3-D has been ported to different high performance computing (HPC) platforms such as the NSF's TeraGrid (the Itanium2 Linux cluster at NCSA), Pittsburgh's Alpha cluster, SGI Altix, IBM p690, and various Linux clusters at Purdue University and JPL.

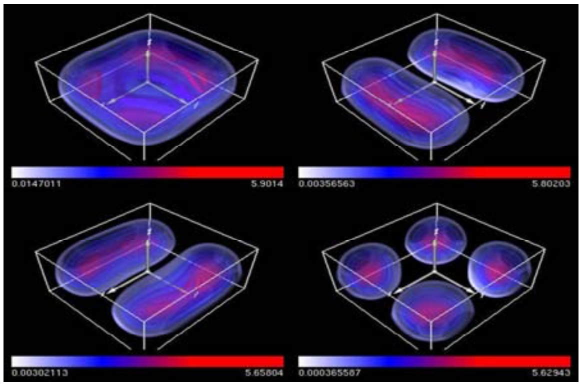

Fig. 6. Wave functions of electron on the first 4 states in conduction band.

The NEMO 3-D project is now part of a wider initiative, the NSF Network for Computational Nanotechnology (NCN). The main goal of this initiative is to support the National Nanotechnology Initiative through research, simulation tools, and education and outreach. Deployment of these services to the science and engineering community is carried out via web-based services, accessible through the nanoHUB portal www.nanoHUB.org. The educational outreach of NCN is realized by enabling access to multimedia tutorials, which demonstrate state-of-the-art nanodevice modeling techniques, and by providing space for relevant debates and scientific events (cyber-infrastructure). The second purpose of NCN is to provide a comprehensive suite of nano simulation tools, which include electronic structure and transport simulators of molecular, biological, nanomechanical and nanoelectronic systems. Access to these tools is granted to users via the web browsers, without the necessity of any local installation by the remote users. The definition of specific sample layout and parameters is done using a dedicated Graphical User Interface (GUI) in the remote desktop (VNC) technology. The necessary computational resources are further assigned to the simulation dynamically by the web-enabled middleware, which

[1] *Rappture* is a toolkit supporting rapid application infrastructure, which is developed by Network for Computational Nanotechnology, Purdue University.

[2] *OPEN-DX* is a package of open source visualization software based on IBM's Visualization Data Explorer.

automatically allocates the necessary amount of CPU time and memory. The end user, therefore, has access not only to the code, a user interface, and the computational resources necessary to run it but also to the scientific and engineering community responsible for its maintenance.

Recently, a prototype graphical user interface (GUI) based on the *Rappture* package (www.rappture.org) is incorporated within the NEMO 3-D package and a web-based online *interactive* version (Quantum Dot Lab) for educational purposes is freely available on www.nanohub.org [46]. The currently deployed educational version is restricted to a single *s* orbital basis (single band effective mass) model and runs in seconds. Quantum Dot Lab was deployed in November 2005 and usage during the past year increased to 924 users conducting 6127 simulation runs (Figure 7). Users can generate and freely rotate 3-D wavefunctions interactively powered by a remote visualization service.

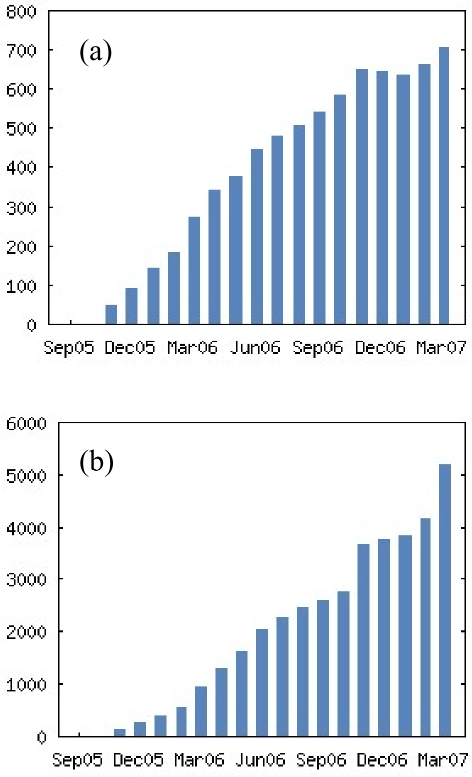

Fig. 7. (a) Number of annual users who have run at least one simulation. (b) Annualized simulation runs executed by nanoHUB users.

The complete NEMO 3-D package is available to selected members of the NCN community through the use of a nanoHUB workspace. A nanoHUB workspace presents a complete Linux workstation to the user within the context of a web browser. The workstation persists beyond the browser lifetime enabling to user to perform long duration simulations without requiring their constant attention. As shown in this paper the computational resources required to perform device scale simulations are considerable and beyond the reach of many researchers. With this requirement in mind NCN has joined forces with Teragrid [47] and the Open Science Grid [48] to seamlessly provide the necessary backend computational capacity to do scientifically significant computing. Computational resources necessary for large scale parallel computing are linked to nanoHUB through the Teragrid *Science Gateways* program. Access to a Teragrid allocation is provided for members of the NCN community. Development of a more comprehensive NEMO 3-D user interface continues. The more comprehensive interface will provide access to a broader audience and encourage the continued growth of the nanoHUB user base.

V. Conclusion

NEMO 3-D is introduced to the IEEE Nanoelectronics community as a versatile, open source electronic structure code that can handle device domains relevant for realistic large devices. Realistic devices containing millions of atoms can be computed with reasonably, easily available cluster computers. NEMO 3-D employs a VFF Keating model for strain and the 20-band sp^{3}d^{5}s* empirical tight-binding model for the electronic structure computation. It is released under an open source license and maintained by the NCN, an organization dedicated to develop and deploy advanced nanoelectronic modeling and simulation tools. NEMO 3-D is not limited to research computing alone; A first educational version including visualization capabilities has been released on nanoHUB.org and has been used by hundreds of users for thousands of simulations. In the next part, the use of NEMO 3-D is demonstrated in the modeling and calculation of single-particle electronic states of a large variety of relevant, realistically sized nanoelectronic devices.

Acknowledgment

Supported by the Indiana 21st Century Fund, Army Research Office, Office of Naval Research, Semiconductor Research Corporation, and National Science Foundation under Grant No. EEC-0228390. The work described in this publication was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology under a contract with the National Aeronautics and Space Administration. The development of the NEMO 3-D tool involved a large number of individuals at JPL and Purdue, whose work has been cited. Drs. R. Chris Bowen, Fabiano Oyafuso, and Seungwon Lee were key contributors in this large effort at JPL. The authors acknowledge an NSF Teragrid award DMR070032. Access to the Bluegene was made available through the auspices of the Computational Center for Nanotechnology Innovations (CCNI) at Rensselaer Polytechnic Institute. We would also like to thank the Rosen Center for Advanced Computing at Purdue for their support.

REFERENCES

[1] Neerav Kharche, Marta Prada, Timothy B. Boykin, and Gerhard Klimeck, "Valley-splitting in strained Silicon quantum wells modeled with 2 degree miscuts, step disorder, and alloy disorder", *Applied Phys. Lett*. Vol. 90, 092109, 2007.

[2] PM. Petroff, *Single Quantum Dots: Fundamentals, Applications, and New Concepts*, Peter Michler, Ed., Springer, Berlin, 2003.

[3] P. Michler, A. Kiraz, C. Becher, W. V. Schoenfeld, P. M. Petroff, Lidong Zhang, E. Hu, A. Imamolu1, “A Quantum Dot Single-Photon Turnstile Device”, *Science*, vol. 290, pp. 2282-2285, 2000.

[4] M. A. Reed, J. N. Randall, R. J. Aggarwal, R. J. Matyi, T. M. Moore, and A. E. Wetsel,” Observation of discrete electronic states in a zero-dimensional semiconductor nanostructure”, *Phys. Rev. Lett*. 60, 535, 1988.

[5] M. A. Reed, Quantum Dots, “Quantum Dots”, *Scientific American*, vol. 268, no. 1, p.118, 1993.

[6] Gabriel Bester and Alex Zunger, “Cylindrically shaped zinc-blende semiconductor quantum dots do not have cylindrical symmetry: Atomistic symmetry, atomic relaxation, and piezoelectric effects”, *Physical Review B*, vol. 71, 045318, 2005. Also, please see references therein.

[7] C. Pryor, J. Kim, L.W. Wang, A. J. Williamson, and A. Zunger, “Comparison of two methods for describing the strain profiles in quantum dots”, *J. Apl. Phys*., 83, 2548, 1998.

[8] M. Grundmann, O. Stier, and D. Bimberg, ”InAs/GaAs pyramidal quantum dots: Strain distribution, optical phonons, and electronic structure”, *Phys. Rev. B*, vol 52, 11969, 1995.

[9] Stier, M. Grundmann, and D. Bimberg, “Electronic and optical properties of strained quantum dots modeled by 8-band k · p theory”, *Phys. Rev. B*, vol. 59, pp. 5688, 1999.

[10] Canning, A. Wang, L.W., Williamson, A., Zunger, A, “Parallel Empirical Pseudopotential Electronic Structure Calculations for Million Atom Systems”, *J. of Comp. Physics* 160, 29, 2000.

[11] G. Klimeck, F. Oyafuso, T. Boykin, R. Bowen, and P. von Allmen, "Development of a Nanoelectronic 3-D (NEMO 3-D) Simulator for Multimillion Atom Simulations and Its Application to Alloyed Quantum Dots", *Computer Modeling in Engineering and Science*, vol. 3, pp. 601, 2002.

[12] Timothy B. Boykin, Gerhard Klimeck, R. Chris Bowen, and Fabiano Oyafuso, "Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory", *Phys. Rev. B*, 66, 125207, 2002.

[13] Timothy B. Boykin, R. Chris Bowen, and Gerhard Klimeck, "Electromagnetic coupling and gauge invariance in the empirical tight-binding method", *Physical Review B*, vol 63, 245314, 2001.

[14] Seungwon Lee, Jeungnim Kim, Lars Jönsson, John W. Wilkins, Garnett Bryant, and Gerhard Klimeck, "Many-body levels of multiply charged and laser-excited InAs nanocrystals modeled by empirical tight binding", *Phys. Rev. B*, 66, 235307, 2002.

[15] Gerhard Klimeck, R. Chris Bowen, Timothy B. Boykin, Carlos Salazar-Lazaro, Thomas A. Cwik, and Adrian Stoica, "Si tight-binding parameters from genetic algorithm fitting", *Superlattices and Microstructures*, vol. 27, No. 2/3, pp. 77-88, Mar 2000.

[16] Gerhard Klimeck, R. Chris Bowen, Timothy B. Boykin, and Thomas A. Cwik, "sp^{3}s* Tight-Binding Parameters for Transport Simulations in Compound Semiconductors", *Superlattices and Microstructures*, vol. 27, pp. 519-524, 2000.

[17] R. Chris Bowen, Gerhard Klimeck, Roger Lake, William R. Frensley and Ted Moise,"Quantitative Resonant Tunneling Diode Simulation", *J. of Appl. Phys*., vol 81, 3207, 1997.

[18] Gerhard Klimeck, Timothy B. Boykin, R. Chris Bowen, Roger Lake, Dan Blanks, Ted Moise, Y. C. Kao, and William R. Frensley, "Quantitative Simulation of Strained InP-Based Resonant Tunneling Diodes", in Proceedings of the 1997 55th *IEEE Device Research Conference Digest*, IEEE, NJ, p. 92, 1997.

[19] Timothy B. Boykin, Gerhard Klimeck, and Fabiano Oyafuso "Valence band effective mass expressions in the sp^{3}d5s* empirical tight-binding model applied to a new Si and Ge parameterization", *Phys. Rev. B*, 69, 115201, No 11, 2004.

[20] Timothy B. Boykin, Neerav Kharche, and Gerhard Klimeck, “Brillouin zone unfolding of perfect supercells composed of non-equivalent primitive cells”, *submitted to* Phys. Rev. B.

[21] S. Ahmed, M. Usman, C. Heitzinger, R. Rahman, A. Schliwa, and G. Klimeck, “Atomistic Simulation of Non-Degeneracy and Optical Polarization Anisotropy in Zincblende Quantum Dots”, *The 2nd Annual IEEE International Conference on Nano/Micro Engineered and Molecular Systems (IEEE-NEMS)*, Bangkok, Thailand, Jan 16-19, 2007.

[22] G. Liang, J. Xiang, N. Kharche, G. Klimeck, Charles M. Lieber and M. Lundstrom, “Performance Analysis of a Ge/Si Core/Shell Nanowire Field Effect Transistor”, *cond-mat/0611226*.

[23] M. A. Eriksson, M. Friesen, S. N. Coppersmith, R. Joynt, L. J. Klein, K. Slinker, C. Tahan, P. M. Mooney, J. O. Chu, and S. J. Koester, “Spin-based quantum dot quantum computing in Silicon”, *Quantum Information Processing*, 3, 133 (2004).

[24] A. J. Williamson, L. W. Wang, and Alex Zunger, “Theoretical interpretation of the experimental electronic structure of lens-shaped self-assembled InAs/GaAs quantum dots”, *Phys. Rev. B* 62, 12963 – 12977, 2000.

[25] Olga L. Lazarenkova, Paul von Allmen, Fabiano Oyafuso, Seungwon Lee, and Gerhard Klimeck, "Effect of anharmonicity of the strain energy on band offsets in semiconductor nanostructures", *Appl. Phys. Lett.* 85, 4193, 2004.

[26] Marek Korkusinski, Gerhard Klimeck, "Atomistic simulations of long-range strain and spatial asymmetry molecular states of seven quantum dots", *Journal of Physics*: Conference Series, vol. 38, pp. 75-78, 2006.

[27] Oyafuso, F., Klimeck, G., von Allmen, P., Boykin, T.B., and Bowen, R.C., ” Strain Effects in large-scale atomistic quantum dot simulations”, *Phys. Stat. Sol*. (b), 239, 71, 2003.

[28] Oyafuso, F., Klimeck, G., Bowen, R.C., Boykin, T.B., and von Allmen P., “Disorder Induced Broadening in Multimillion Atom Alloyed Quantum Dot Systems”, *Phys. Stat. Sol*. (c), 4, 1149, 2003.

[29] M. Korkusinski, F. Saied, H. Xu, S. Lee, M. Sayeed, S. Goasguen, and G. Klimeck, "Large Scale Simulations in Nanostructures with NEMO3-D on Linux Clusters", 2005 Linux Cluster Institute Conference, Raleigh, NC, April, 2005.

[30] M. Graf and P. Vogl, “Electromagnetic fields and dielectric response in empirical tight-binding theory,” *Phys. Rev. B*, 51, 4940, 1995.

[31] Keating, *P., “*Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure”, *Phys. Rev*. 145, 1966.

[32] J. C. Slater and G. F. Koster, “Simplified LCAO Method for the Periodic Potential Problem”, *Phys. Rev*. 94, 1498, 1954.

[33] Jancu, J.M., Scholz, R., Beltram, F., Bassani, F., "Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters“, *Phys. Rev. B* 57, 6493, 1998.

[34] Timothy B. Boykin, P. Vogl, “Dielectric response of molecules in empirical tight-binding theory”, *Phys. Rev. B*, vol. 65, 035202, 2001.

[35] Seungwon Lee, Jeungnim Kim, Lars Jönsson, John W. Wilkins, Garnett Bryant, and Gerhard Klimeck, "Many-body levels of multiply charged and laser-excited InAs nanocrystals modeled by empirical tight binding", *Phys. Rev. B* 66, 235307, 2002.

[36] Timothy B. Boykin and Gerhard Klimeck, “"Practical Application of Zone-Folding Concepts in Tight-Binding", *Physical Review B*, Vol. 71, 115215, 2005.

[37] Timothy B. Boykin, Neerav Kharche, Gerhard Klimeck, and Marek Korkusinski, “Approximate bandstructures of semiconductor alloys from tight-binding supercell calculations,” *J. Phys: Condensed Matter* 19, 036203, 2007.

[38] Timothy B. Boykin, Mathieu Luisier, Andreas Schenk, Neerav Kharche, and Gerhard Klimeck, “The electronic structure and transmission characteristics of disordered AlGaAs nanowires,” *IEEE Trans Nanotechnology*, 6, 43, 2007.

[39] K. Maschhoff and D. Sorensen, “A portable implementation of ARPACK for distributed memory parallel architectures”, *Copper Mountain Conference on Iterative Methods*, 1996.

[40] C. Lanczos, “An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators”, *Journal of Research of the National Bureau of Standards*, Vol 45, No 4, 1950.

[41] Wang, L.W., Zunger A, “Solving Schrödinger's equation around a desired energy: Application to silicon quantum dots”, *J. of Chem. Physics*, 100, 2394, 1994.

[42] A. Sameh and Z. Tong, “The trace minimization method for the symmetric generalized eigenvalue problem”, *J. Comp. and Appl. Math*, 123, pp. 155-175, 2000.

[43] G.L.G Sleijppen and H.A. Van der Vorst, "A Jacobi-Davidson Iteration Method for Linear Eigenvalue Problems", *SIAM Journal on Matrix Analysis and Applications*, Vol 17, No 2, pp. 401-425, 1996.

[44] Hansang Bae, Steve Clark, Ben Haley, Gerhard Klimeck, Marek Korkusinski, Sunhee Lee, Maxim Naumov, Hoon Ryu, and Faisal Saied, "Electronic structure computations of quantum dots with a billion degrees of freedom", Submitted to *Supercomputing* 07, Reno, NV, USA, November 2007.

[45] W. Qiao, M. Mclennan, R. Kennell, D. Ebert, and G. Klimeck. "Hub-based Simulation and Graphics Hardware Accelerated Visualization for Nanotechnology Applications." *IEEE Transactions on Visualization and Computer Graphics,* Vol. 12, Issue 5, pp1061 – 1068, Sept.-Oct. 2006.

[46] https://www.nanohub.org/simulation_tools/qdot_tool_information.

[47] TeraGrid at http://www.teragrid.org

[48] Open Science Grid at http://www.opensciencegrid.org.