Superlattices and Microstructures, Vol. 31, Nos 2–4, 2002
doi:10.1006/spmi.2002.1038
Available online at http://www.idealibrary.com on IDEAL®

3-D atomistic nanoelectronic modeling on high performance clusters: multimillion atom simulations

GERHARD KLIMECK†, FABIANO OYAFUSO, R. CHRIS BOWEN
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, MS 169-315 Pasadena, CA 91109-8099, U.S.A.

TIMOTHY B. BOYKIN
University of Alabama in Huntsville, AL, U.S.A.

THOMAS A. CWIK, EDITH HUANG, EDWARD S. VINYARD
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, U.S.A.

(Received 11 March 2002)

Electronic device scaling is ultimately limited by atomic dimensions. The simulation of electronic structure and electron transport on these length scales must be fundamentally quantum mechanical. This leads to computational models that account for fundamental physical interactions using an atomistic basis and tax even the largest available supercomputer when simulating measurable devices. The prototype development of a software tool that enables this class of simulation is presented. Realistically sized structures contain one million to tens of millions of atoms that need to be represented with an appropriate basis. The resulting sparse complex Hamiltonian matrix is of the order to tens of millions. A custom matrix–vector multiplication algorithm that is coupled to a Lanczos and/or Rayleigh–Ritz eigenvalue solver has been developed and ported to a Beowulf cluster as well as an Origin 2000. First benchmarking results of these algorithms as well as the first results of quantum dot simulations are reported.

Key words: quantum dot, nanoelectronics, sparse matrix–vector multiplication.

1. Introduction

The goal of reducing payload in future space missions while increasing mission capability demands miniaturization of measurement, analytical, and communication systems. The ultimate scaling limit of individual semiconductor devices are atomic dimensions. The enabling technology for miniaturization of space mission electronics has been the miniaturization of semiconductor devices (Fig. 1A adopted from the SIA Roadmap [1]) of the past 40 years. The development has surpassed every expectation and overcome (so far) every predicted technological obstacle. It has become evident that not technology but the *atomic dimensions* of the underlying crystalline lattice and the countable *electron number* (Fig. 1B) ultimately limit [2] this scaling trend. A variety of novel nano-scaled detector [3] and computation [4] schemes based on quantum

†Author to whom correspondence should be addressed. E-mail: gekco@ieee.org

0749–6036/02/020171 + 09 $35.00/0

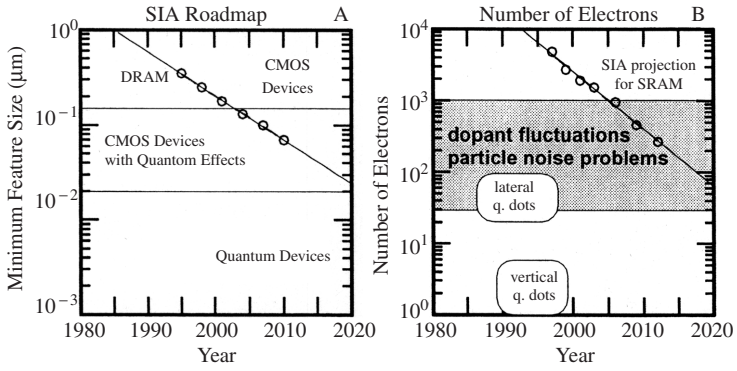

Fig. 1. A, Minimum 2D feature size as projected on the SIA roadmap [1]. Layer thickness of 0.01 μm in the next generation devices are not captured in this graph. B, Number of electrons under a a CMOS SRAM gate [2]. Dopant fluctuation and particle noise fluctuations may make reliable circuit design impossible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

dots have been proposed and/or demonstrated. The work presented here is aimed at providing a simulation tool that enables the fundamental modeling of electron devices on the length scales of a few nanometers. The problem size and the choice of atomistic basis sets results in numerical representations that require the usage of supercomputers [5].

Simulation for device design and characterization. Physics-based device simulation has penetrated the mainstream semiconductor industry device design and characterization process [6] in the last 1 or 2 years. This penetration can be attributed to the increased experimental costs of nano-scale layer and feature characterizations and the introduction of new materials. The modeling–fabrication–characterization triangle that has existed in circuit design for the past decade has now established itself for the underlying semiconductor device design as well. This coupled process will enable the device design for the next device generations if the physics-based simulation tools can deliver the needed accuracy. As the electronic device sizes shrink further towards the tens of nanometer size scale (deca-nano) in all three dimensions, new physical phenomena will emerge, that previously could be safely ignored. These phenomena are based on the quantum mechanical nature of electrons which is typically ignored by or patched into existing commercial device simulators. Such an approach will probably suffice for the next two device generations with empirical model calibration and limited scaling projection capability. In order to correctly model physically observed effects such as electron tunneling, state quantization and charge quantization more sophisticated models need to be developed and incorporated into device simulators that can handle realistically sized systems.

Our modeling agenda. Following our 1D nanoelectronic modeling work [7, 8] (NEMO 1-D) we are developing an atomistic-based, nanoelectronic modeling tool (NEMO 3-D). The following section discusses some of the system size issues ad quantum mechanical modeling capabilities that need to be included into such a simulator.

2. Quantum dot applications and modeling requirements

Quantum devices are not shown on the SIA roadmap for lithography because it focuses on pure silicon device scaling. In particular Fig. 1A only shows the lateral feature size, while layer thicknesses are already on the order of 0.01 μm. Various quantum dot implementations in different material systems as well as silicon have been examined since the late 1980s (Fig. 1B), and several designs have shown room temperature

operation. Pyramidal self-assembled quantum dot arrays in particular are promising candidates to be used in quantum well lasers and detectors [3] within a few years.

What is a quantum dot? A quantum dot (QD) can be described as a solid state structure in which a (small) number of electrons are isolated from the surrounding environment. This is achieved by 'placing' an electrical insulator around a (semi-)conductor. If the central region is small and clean enough effects due to state and charge quantization can be measured macroscopically. QDs can therefore be viewed as *artificial atoms*. They represent the ultimate limit of scaled solid state devices. The large parameter space of material systems, shapes, and doping profiles allows for a detailed engineering of the electrical and optical properties of QDs. In particular, the fine tuning of optical transition energies applies to JPL's immediate interest in far infrared detectors and emitters.

Near term quantum dot applications. Currently several research groups are incorporating self-assembled InAs QDs into AlGaAs quantum wells in optical detector and laser structures [3]. The reduced degree of freedom reduces the scattering of the confined electrons and therefore increases the state lifetimes. This increase promises better device performances such as reduced threshold currents, decreased linewidths, reduced dark currents [9] and higher operating temperatures. These are systems which apply to JPL's immediate interest in far infrared detectors and emitters.

Size and atom number estimates of realistic systems. The modeling of an individual self-assembled InAs QD of 30 nm diameter and 5 nm height embedded in GaAs of buffer width 5 nm roughly requires a simulation domain of $40 \times 40 \times 15$ nm^3 containing approximately 1 million atoms. A horizontal array of four such dots separated by 20 nm roughly requires a simulation domain of $90 \times 90 \times 15$ nm^3 which corresponds to 5.2 million atoms. A $70 \times 70 \times 70$ nm^3 cube of silicon which might be needed to simulate ultra-scaled CMOS device contains about 15 million atoms. There are about 43 atoms in 1 nm^3 as a rule of thumb.

3. Basis representation

Researchers have explored a variety of different approaches to represent matter in a nano-scaled system. All these approaches fall into two major categories: atomistic and nonatomistic. The nonatomistic 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. Such approaches typically deal on the lowest level with effective masses and band edges. The popular $k \cdot p$ approach belongs in this category. These approaches do not contain any crystalline information and are fundamentally not well suited for the atomistic representation of nano-scale features such as interfaces and disorder. Atomistic approaches attempt to include the electronic wavefunction of each atom in some approximation. The critical question is what basis set to use for the representation of the electronic wavefunction. There are two schools of thought: (1) plane waves [5], and (2) local orbitals [10]. A list of pros and cons for each method could be presented; instead we just state here that we consider both methods complementary to each other and that we choose the local orbital basis (tight-binding) approximation for its ability to model finite structures (not infinitely extended) and for its past success [7, 8] in quantum mechanical modeling of electron transport.

The basic idea of the tight-binding method is that one selects a basis consisting of atomic orbitals (such as s, p, and d) to create a single electron Hamiltonian that represents the bulk electronic properties of the material. The interactions between the 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 [11–16]. Our simulator typically uses an sp^3s^* or $sp^3d^5s^*$ model that consists of five or 10 spin degenerate basis

states, respectively. Each atom is therefore represented by a 10×10 or 20×20 matrix. We have limited ourselves to nearest-neighbor interactions to enable a simpler interaction representation in the presence of strain. This nearest-neighbor model restricts the number of nonzeros per row to a small value independent of device size and yields a sparse, block banded Hamiltonian with $O(n)$ nonzeros, where n is the number of atoms. The strain that arises near interfaces of materials with different lattice constants yields a different, position-dependent coupling between each neighboring pairs of atoms. For the case of a zincblende lattice, each atom has four neighbors. Therefore, the storage requirement for 1 million atoms can be estimated as $10^6\ atoms \times 5\ diagonals \times (20 \times 20\ basis) \times 16\ bytes/2(for\ Hermiticity) = 16$ GB. Algorithms that use more of the symmetry of the Hamiltonian are considered for future development. Currently we have the option to store the Hamiltonian or to recompute it on the fly.

QDs are characterized by confinement in all three spatial dimensions, so that the Hamiltonian no longer commutes with *any* of the (discrete) translation operators. The wavevector is hence *not* a conserved quantity in *any* dimension. The most appropriate basis for representing such a highly confined wavefunction is, therefore, one consisting of atomic-like orbitals centered on each atom of the crystal. Following Slater and Koster [18], we take the atomic-like orbitals to be orthonormal. We consider a crystal whose Bravais lattice points are given by

$$\mathbf{R}_{n1n2n3} = n_1\mathbf{a}_1 + n_1\mathbf{a}_2 + n_1\mathbf{a}_3 \tag{1}$$

where the $\mathbf{a}_i$ are primitive direct-lattice translation vectors and the n_i are integers. If there is more than one atom per cell (as is the case with, for example, GaAs or Si) we index the atoms within a cell by μ and the location of the μth atom within the cell located at 1 is given by $\mathbf{R}_{n1n2n3} + \mathbf{v}_\mu$ where $\mathbf{v}_\mu$ is the displacement relative to the cell origin. We normalize the wavefunction over a volume consisting of N_i cells in the $\mathbf{a}_i (i = 1, 2, 3)$ direction and write the state as a general expansion in terms of localized atomic-like orbitals:

$$|\Psi\rangle = \frac{1}{\sqrt{N_1N_2N_3}} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} \sum_{n_3=1}^{N_3} \sum_{\alpha} \sum_{\mu} C_{n_1n_2n_3}^{(\alpha\mu)} |\alpha; \mu; \mathbf{R}_{n1n2n3} + \mathbf{v}_\mu\rangle. \tag{2}$$

In eqn (2) α indexes the atomic-like orbitals centered on the μ atoms within each cell $(n_1n_2n_3)$. The Schrödinger equation thus appears as a system of simultaneous equations given by:

$$\langle\alpha; \mu; R_{n1n2nc} + \mathbf{v}_\mu|(\hat{\mathcal{H}} - E\hat{1})|\Psi\rangle = 0. \tag{3}$$

In eqn (3) we express the matrix elements between localized orbitals as tight-binding parameters, in our case limiting the interactions to the nearest neighbor.

4. Numerical details

4.1. Parallel implementation of sparse matrix–vector product

The goal of the simulation is to solve the eigenvalue problem for low lying electron and hole states near the band edge. Two algorithms, a Rayleigh–Ritz minimization algorithm and a Lanczos method, have been parallelized to solve this problem. At the heart of each is a sparse matrix-vector multiplication. For implementation on a distributed memory platform, data must be partitioned across processors so as 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 Fig. 2, lends itself to a 1D chain network topology and results in a block-tridiagonal Hamiltonian, where each block corresponds to a pair of processors, and each processor holds the column of blocks associated with its atoms. Communication

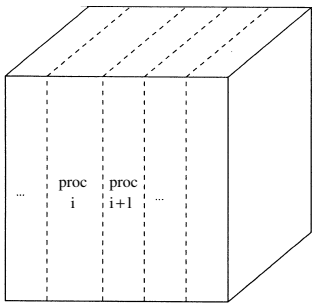

Fig. 2. The device is decomposed into slabs (layers of atoms) which are directly mapped to individual processors.

costs, 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 columnwise fashion as shown in Fig. 3. That is, processor j computes

$$y_{i,j} = H_{i,j}x_j \qquad (i = j, j \pm 1) \tag{4}$$

where $H_{i,j}$ is the block of Hamiltonian associated with nodes i and j, and x_j are the components of x stored locally on node j. The results of the matrix–vector multiplication of the off-diagonal blocks ($i \neq j$) with the local portion of the dense vector are sent via MPI calls to neighboring processors in two steps. First, all but the rightmost processor send data to the right (and receive data from the right, possibly in full duplexing depending on the actual MPI implementation), subsequently followed by communication to the left neighbors. In addition, not all rows of the off-diagonal blocks are in general nonzero, although the sparse structure of these blocks depends on the particular crystal structure in question. In practice, however, a sufficient fraction of zero rows exists that compressing the matrix–vector multiplication by removing structurally guaranteed zeros is worthwhile despite the additional level of data pointer indirection required to keep track of the nonzero structure.

4.2. Performance

In this section, we discuss the performance of NEMO 3-D using the parallelized Lanczos algorithm on two platforms, a Beowulf commodity cluster of Pentium III's and a shared memory SGI Origin 2000. Figure 4A displays the performance of 30 iterations of the Lanczos solver on a Beowulf system consisting of 32 nodes with two 933 MHz Pentium III CPUs and 1 GB of RAM per node. Different processors communicate using a nonshared memory MPI implementation and over a slow 100 base-T ethernet connection. Curves corresponding to five different problem sizes are shown. Dashed curves indicating a linear (ideal) scaling are also shown for reference. The two largest problems require enough memory that they require more than one processor to avoid swapping. The efficiency, the ratio of speedup to ideal speedup, as a function of number of processors decreases with increasing network size. This decrease corresponds to a fraction of unparallelized code of approximately 1.6%. However, Fig. 4B also shows that for the regime of interest the efficiency is independent of problem size. This result suggests that while communication bandwidth is not a limiting factor for this problem, there may be some inherent load imbalance in the computation. This issue is under continued investigation.

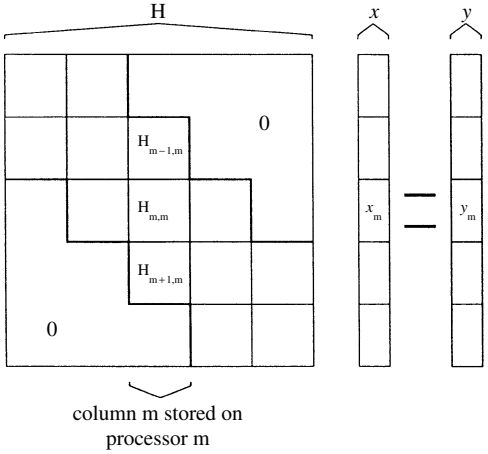

Fig. 3. Example matrix–vector multiplication on five processors performed in a columnwise fashion, so that the mth block column and section x_m are stored on processor m. The nearest-neighbor model with nonperiodic boundary conditions guarantees that the Hamiltonian is block-tridiagonal, so that communication is performed only with nearest-neighbor processors.

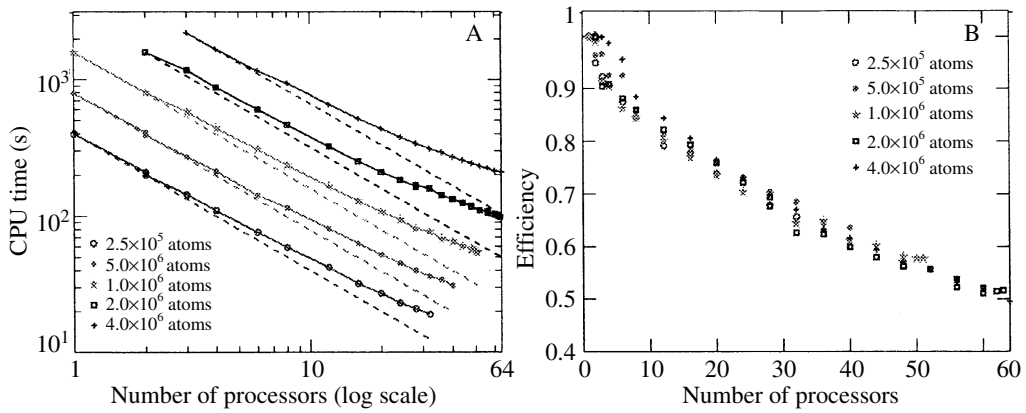

Fig. 4. A, Execution time of 30 Lanczos iterations on a 933 MHz Pentium III Beowulf cluster (solid line with symbols). The dashed lines illustrate ideal scaling. The nearest-neighbor CPU communication limits the 0, 25, 0.5, 1, and 2 million atom simulations to 32, 40, 51, and 63 CPUs, respectively. B, Efficiency as defined as the ratio of actual speed-up to ideal speed-up.

Figure 5 compares a 933 MHz Pentium III dual CPU Beowulf to a 450 MHz single CPU Pentium III Beowulf. If execution times are scaled by CPU clock frequency, the performances of the two systems are indistinguishable. This result reinforces our finding that for the problem sizes of interest, our solver is limited by CPU speed rather than by communication bandwidth.

NEMO 3-D has the option of reusing the Hamiltonian for more than one iteration or computing the matrix vector multiplications on the fly without explicitly storing the Hamiltonian. Clearly, storing the Hamiltonian for later re-use is preferable, but is not possible for sufficiently large problems. Figure 6 compares the performance of the 933 MHz cluster with a 128 CPU 300 MHz R12k SGI Origin 2000 with 512 MB of memory per processor. As before, dashed curves indicate ideal performance. Figure 6A simply shows that the SGI Origin 2000 scales similarly to the Beowulf cluster. Figure 6B, interestingly, shows that the benefit of storing the Hamiltonian is much greater for the Origin 2000 than it is for the Beowulf cluster. Indeed, while execution time on the Beowulf system is reduced only by a factor of 1.3, it is reduced roughly by a factor of 4 on

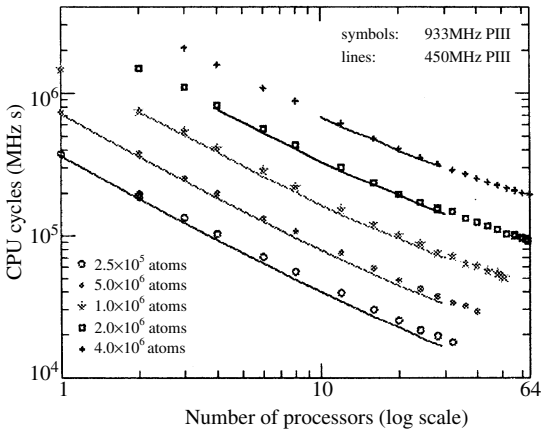

Fig. 5. Comparison of a 450 MHz Beowulf with 512 MB of memory per node (solid) with a 933 MHz Beowulf with 1 GB of memory per node (symbols) in terms of CPU cycles. Compute times are multiplied by the CPU cycles. Virtually identical curves for 450 MHz and 933 MHz machines indicate perfect clock rating scaling.

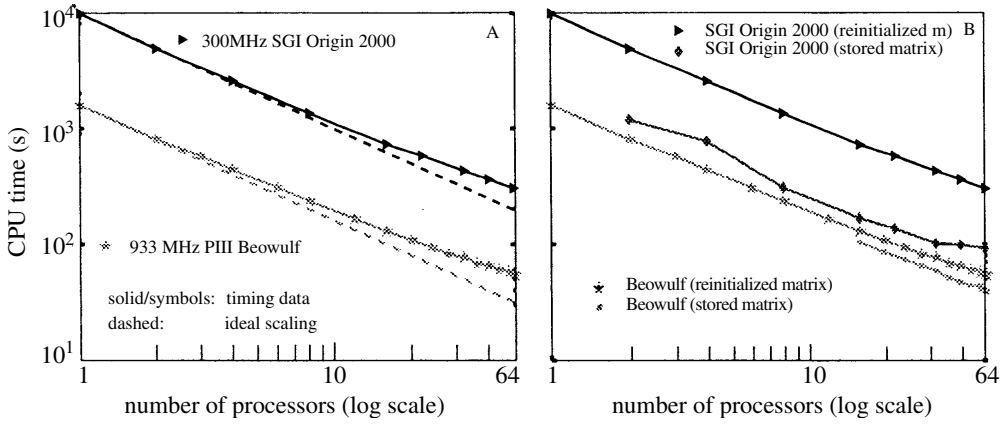

Fig. 6. Comparison of CPU times on a Beowulf cluster with an SGI Origin 2000 with A, ideal CPU time and B, stored versus recomputed Hamiltonian.

the SGI Origin 2000. The reason for this discrepancy is not completely clear, but is likely attributable to a difference in cache size and memory access speed.

5. Simulation of Alloyed Quantum Dots

Since we represent each individual atom in the QD system explicitly we can demonstrate this capability by simulating an $In_{0.6}Ga_{0.4}As$ alloyed QD system [19] in a GaAs matrix. The dome shaped QD has a diameter of 30 nm and a height of 5 nm. A 5 nm GaAs buffer surrounded a QD in the simulation. 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 of the dot is assumed to be fixed. We therefore try to answer the question: What is the minimal optical linewidth that can be expected for such an alloyed dot neglecting any experimental size variations? A side view of such an alloyed QD which is half the size of the system considered here without the surrounding GaAs is shown in Fig.7A.

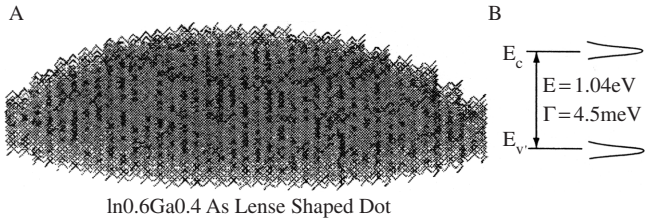

Fig. 7. A, Lens shaped $In_{0.6}Ga_{0.4}As$ QD visualized the molecular viewer Rasmol. The QD size is scaled down to include less than 10 000 atoms for visualization purposes. B, Lowest conduction and highest valence quantized states that are inhomogeneously broadened by alloy disorder.

In simulations of 490 random alloy configurations [17] we have obtained the single particle electron and hole ground state energies and the optical transition energy from their difference. The mechanical strain is minimized using a valence force field method [20, 21] which considers contributions to the total strain energy due to bond length changes as well as bond angle modification. The mechanical strain field is recomputed for each alloy configuration. For these particular simulations we used the sp^3s^* basis set where the matrix elements scale with respect to the equilibrium position with the ideal [14] exponent 2.

The simulation of 490 different alloyed dots shows [22] a mean optical transition energy of 1.04 eV and a standard deviation, or associated linewidth of 4.5 meV compared to experimentally reported [19] transition energy of 1.09 eV and a linewidth of 34.6 meV. The experimental data does of course include QD size variations as well. We plan to simulate larger samples that do include QD size variations in the future. The major result of this simulation is the observation that there will be a significant optical linewidth variation due to alloy disorder alone, even if all the QDs were perfectly identical in size.

Acknowledgements—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.

References

[1] Data taken from The National Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, California, USA, 1998, Classifications such as CMOS Devices with Quantum effects and Quantum Devices taken from private communication with Alan Seabaugh, Texas Instruments (1997).
[2] R. C. Brown, Best case estimates for number of electrons under an SRAM cell, personal communication, 1998, Texas Instruments.
[3] H. C. Liu *et al*., Appl. Phys. Lett. **78**, 79 (2001).
[4] A. O. Orlov *et al*., Appl. Phys. Lett. **78**, 1625 (2001).
[5] A. Canning, L. W. Wang, A. Williamson, and A. Zunger, J. Comput. Phys. **160**, 29 (2000).
[6] Physics-based device simulation tools have typically only been used to improve individual device performance after careful calibration of the simulation parameters.
[7] R. C. Bowen *et al*., J. Appl. Phys. **81**, 3207 (1997).
[8] G. Klimeck *et al*., *in the 1997 55th Annual Device Research Conference Digest* (IEEE, NJ, 1997) p. 92.
[9] V. Ryzhii *et al*., Japanese J. Appl. Phys. Lett. **39**, L1283 (2000).
[10] S. Lee *et al*., Phys. Rev. B **63**, 195318 (2001).
[11] P. Vogl, H. P. Hjalmarson, and J. D. Dow, J. Phys. Chem. Solids **44**, 365 (1983).
[12] T. B. Boykin, G. Klimeck, R. C. Bowen, and R. Lake, Phys. Rev. B **56**, 4102 (1997).

[13] T. B. Boykin, Phys. Rev. B **56**, 9613 (1997).
[14] J. M. Jancu, R. Scholz, F. Beltram, and F. Bassani, Phys. Rev. B **57**, 6493 (1998).
[15] G. Klimeck *et al*., Superlatt. Microstruct. **27**, 77 (2000).
[16] G. Klimeck, R. C. Bowen, T. B. Boykin, and T. A. Cwik, Superlatt. Micostruct. **27**, 519 (2000).
[17] G. Klimeck, F. Oyafuso, R. C. Bowen, and T. B. Boykin, J. Comput. Electron. (2001) (accepted for publication).
[18] J. C. Slater and G. F. Koster, Phys. Rev. **94**, 1498 (1954).
[19] R. Leon *et al*., Phys. Rev. B **58**, R4262 (1998).
[20] P. Keating, Phys. Rev. **145**, 637 (1966).
[21] C. Pryor *et al*., J. Appl. Phys. **83**, 2548 (1998).
[22] 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.