3-2010

Atomistic Modeling of Realistically Extended Semiconductor Devices with NEMO and OMEN

Gerhard Klimeck
NCN, Purdue University, gekco@purdue.edu

Mathieu Luisier
Purdue University

Follow this and additional works at: http://docs.lib.purdue.edu/nanopub

Part of the Nanoscience and Nanotechnology Commons

http://dx.doi.org/10.1109/MCSE.2010.32

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
CYBER-ENABLED NANOTECHNOLOGY
Atomistic Modeling of Realistically Extended Semiconductor Devices with NEMO and OMEN

Researchers have continually developed the Nanoelectronic Modeling (NEMO) toolset over the past 15 years to provide insight into nanoscale semiconductor devices that are dominated by quantum mechanical effects. The ability to represent realistically large devices on an atomistic basis has been the key element in matching experimental data and guiding experiments. The resulting insights led to the creation of OMEN, a new simulation engine.

In 1965, Moore’s law predicted an exponential cost reduction with an exponential increase in the number of components per integrated circuit. Indeed, over the past 45 years, we’ve obtained an amazing increase in computational capabilities through the dramatic size reduction of individual transistor components. Despite predictions of insurmountable technical difficulties, sheer economic drivers have now created a global US$260 billion semiconductor industry. There is, however, a fundamental limit that we can’t overcome: atoms aren’t divisible, so downscaling must stop in the realm of countable atoms.

Some material layers in commercial devices have now reached the thickness limit of a few atoms. Lateral dimensions are now at 20 to 30 nanometers (80 to 120 atoms) and device geometries are no longer flat planar, but rather 3D objects. New materials have entered the device design realm to reduce leakage currents through thin layers and to deform the active transistor material through strain engineering, which improves transistor characteristics. The problems are large enough that silicon industry is considering new transistor materials, such as carbon and even the eternal materials of the future: Gallium arsenide (GaAs) and Indium arsenide (InAs). The design space’s size is suddenly exploding because the detailed local atom arrangements have become critical and new materials have been added in novel 3D geometries.

Modeling and simulation might offer ways to explore options before actual experimentation. However, it’s probably fair to say that we can’t use any of the typical commercial semiconductor device design tools to explore the 3D atomistically defined search space. Most tools are based on continuum material assumptions and therefore ignore the mere existence of the atomic granularity. Most tools also, at very best, patch quantum mechanical effects into the simulation concept through perturbative treatments. What we need, however, is a fundamentally quantum mechanical carrier transport model built on an atomistic material description.

A suite of tools that can model realistically extended nanoelectronic devices such as 3D quantum dots, ultra-thin-body transistors, nanowires, carbon nanotubes, and graphene sheets at an
atomistic resolution might help narrow the search and optimization space, reduce the cost of new technology developments, and reveal new device operation concepts (see Figure 1 for examples).

Device downscaling has also increased costs for experimental determination of transistor designs, as well as for overall manufacturing; constructing a new fabrication line is approaching US$10 billion. Quantitative modeling and simulation at the atomistic scale could enable the exploration of the design space for high-performance, fault-tolerant, manufacturable devices. Modeling and simulation might therefore be the key to keeping Moore’s law valid for a few more years.

**NEMO-1D: The First Industrial Quantum Transport Simulator**

The first analog and digital nanoelectronic devices to operate at room temperature with nanometer-scale material variations were resonant tunneling diodes (RTDs). Indeed, these devices require a quantum mechanical understanding and can’t be modeled with semiclassical approaches. The Texas Instruments Central Research Lab assembled a team of theorists, computational scientists, software engineers, and—last, but not least—experimentalists to create an industrial-strength modeling tool that could drive RTD design. The primary challenge was to increase the peak-to-valley current ratio to reduce the “off current” in possible digital and analog circuits. Theoretically, this amounted to understanding the valley current’s physical origin (see Figure 2).

In 1997, toward the project’s end, the modeling was guiding quantitative, predictive simulation that agreed with experiments (see Figure 2b), with the bottom line being that we can’t dramatically reduce the off current in high-performance, high-current-density devices in standard RTDs. Researchers eventually patented design alternatives for low-power memory cells. While general interest in RTDs has subsided since then, researchers achieved critical insights resulting from modeling carrier transport at the nanometer scale.

The Nanoelectronic Modeling 1D (NEMO-1D) team developed new boundary conditions that enabled treatment of extended device contact regions in which strong scattering and thermalization of carriers and electrostatic control are critical, while quantum mechanically confined states still rule over the carrier injection into the central device. Most RTD community members had suspected that incoherent scattering in the central device region was the valley current’s key element. That effectively turned out to be incorrect. The critical element was actually to understand where the resonances are in energy and how they’re coupled to the contacts. To enable quantitative modeling of carrier transport in room-temperature high-performance, high-current-density RTDs, we needed an atomistic representation of the device layers.

**Setting the Path for General Nanoelectronic Device Simulation**

While the academic and industrial interest in RTDs has subsided, we’ve gained critical insights into carrier transport at the nanometer scale. As we now describe in more detail, the development of NEMO-1D has set the model, user, and
Among the requirements, models

- must not be based on continuum material description, but must include atomistic granularity and crystal symmetries;
- need basic mechanical atom–atom interactions (strain and phonons);
- need full-band electronic structure representation for the central valence and conduction bands (as opposed to expansions around central symmetry points like gamma);
- must represent bulk material properties, such as band gaps to within a few millielectron volts (meV) and effective masses to within a few percent, because band structure engineering works in the tuning band edges domain to within a few tens of meV;
- must have transferability from bulk to the nanometer scale;
- must be computationally lightweight enough to represent realistically extended devices (realistic structures imply tens of millions of atoms in quantum dot structures and around 100,000 atoms in ultra-scaled field effect transistors and nanowires);
- must model devices of a finite extent in realistic environments, which precludes the assumption of infinite periodic structures surrounded by a vacuum;
- must be able to include open boundary conditions, not just closed systems;
- must be able to represent atomistic disorder without additional computational burden or user interference—computation time and model tuning should not increase with degrees of disorder; and
- must have an atomistic model that can represent atomistic disorder explicitly—rather than in a statistical average way—as each device is indeed different from the next.

These requirements guided the choice of basis sets in subsequent developments, both with the NEMO team and other researchers around the world. “Exact” band gaps and masses, millions of atoms, and finite extent preclude most typical *ab initio* models. In fact, you could argue that for typical semiconductor devices, these might be unnecessary anyhow, because device simulation doesn’t need to establish the existence and formation of bonds. If the bonds varied in time, the semiconductor device would be unstable and plagued by noise. However, the empirical tight-binding approach that we selected meets the requirements and needs stated earlier.

**Model Requirements**

Among the requirements, models

**User Requirements**

We envisioned various classes of users, including computational scientists, experimentalists, educators and students, and NEMO developers.

*Computational scientists* typically push the model development and validation. They usually have no real compute-time requirements and do whatever it takes to solve the model. Simulations are often limited to a few cases to establish the existence of solutions or to gain fundamental insight. These users are willing to move data and restart files all over the place and perform data analysis in ad hoc interfaces.

*Experimentalists* generally know little about the model’s details, but have realistic problems to solve rapidly. They need to understand and develop device concepts, explore many different designs, and ask “What if?” questions rapidly. The primary exploration times are seconds to minutes, while secondary execution times might be the lunch hour or a few overnight simulations. With this in mind, developers must integrate data management, visualization, and export into notebooks into the overall tool experience. The software must not demand that these users intricately understand the underlying theory. Computational science experiments that add more and more basis functions to test the model’s validity are inappropriate for this class of users.

As with the experimentalists, *educators and students* need a rapid simulation turnaround, but
they also need an even simpler access method. Use might be anywhere in the world and in the classroom, without any license and installation issues. In fact, neither teachers nor students are typically allowed to install additional software on classroom or laboratory computers. These users also need reuse of previous simulation results or set-up before classtime and simple results reporting and interpretation.

**Developer Requirements**

NEMO developers included theorists, computational scientists, algorithm developers, software engineers, and user interface designers. They must be able to work together in various software project phases. These developers have quite different requirements in their usual workflow. At one end of the spectrum, they might need code to change on a daily basis; at the other end, they might need to freeze code for several months while they test it and roll it out to users. They need a dynamical I/O design that limits the exposure of new models to users and enables developers to add models rapidly.

**From NEMO-1D to OMEN**

Several researchers identified the fundamental transport methodology as the non-equilibrium Green's function (NEGF) approach and established it as such in this field. Given our model requirements, we used empirical orthogonal tight-binding (ETB) approaches to represent the semiconductor materials. We implemented different basis sets, ranging from simple effective mass (single “s” orbital) to a suite of more comprehensive models, such as sp3s* and sp3d5s*, typically in nearest-neighbor representations.

ETB is based on the symmetry-formalized interaction of valence electrons on neighboring atomic sites. It ignores the core electrons and is therefore not a total energy Hamiltonian. However, we can fit ETB to match experimental and more fundamental theoretical band structure properties when we improve the model to match general strain behaviors.

**NEMO-3D Development at NASA JPL**

In 1998, the NASA Jet Propulsion Laboratory had a strong experimental technology group interested in developing optical detectors and lasers. Experimentalists were growing self-assembled quantum dots and needed to guide the efforts with modeling and simulation. The advent of Beowulf-based cluster computing in the JPL High Performance Computing Group made dedicated parallel compute power available to engineers. Transport simulations through 3D-resolved structures with millions of atoms were completely unfeasible; the work’s focal point was to create an electronic structure simulator that could compute the confined conduction and valence band states of realistically extended quantum dot systems fully atomistically. Strain is a crucial element in the self-assembled quantum dot system and is modeled through a classical ball-on-a-spring model. The code is designed to be parallel with a variety of different memory and compute time trade-off capabilities. Million-atom electronic structure simulations were first demonstrated in the year 2000. In 2003, NEMO-3D was released as open source and Purdue University continued its development.

**OMEN Development at ETH and Purdue**

In 2004 and 2005, it became clear that the computing power needed to perform atomistic transport simulations for extended devices would soon be available. Approaches to 3D transport in nanowires using a simple effective mass model were underway already and being deployed on nanoHUB.org. During his PhD work at the Swiss Federal Institute of Technology (ETH), Mathieu Luisier worked with Purdue to develop several completely new Matlab prototypes and then a C++ foundation for a general 3D quantum transport approach. Initially, the simulators didn’t include incoherent scattering, and the compute time was reduced using a wavefunction approach. In 2008, Luisier began developing the OMEN transport tool at Purdue; the first implementations that included incoherent scattering were achieved in 2009.

**Input/Output Approaches and Interface Designs**

Developing rapidly evolving software and deploying it quickly to a user base has its own interesting requirements. For instance, it’s easier to build GUIs when the I/O is carved in stone and doesn’t change. However, theory and algorithm developers need to add new models and algorithm parameters rapidly. For the three large software projects described here, we experimented with different I/O handling mechanisms.

NEMO-1D uses a mixed static and dynamic GUI design approach. A GUI developer implemented a set of static windows filled with dynamically defined data structure objects. We can dynamically expand device design and material data windows with new design descriptions and
material models. We structured simulation and algorithm parameters in parent-child-related dependency trees that we can dynamically configure at compile time. This dynamic design helped us decouple the static GUI work of the software engineer from the dynamic requirements of the algorithm and theory developers. We defined all I/O in C data structures; translators convert these into GUI components or batch input deck items. Another key element in NEMO-1D I/O is the ability to define all material parameters in an externally scriptable interface using LEX (a lexical analyzer) and YACC (yet another compiler compiler).

NEMO-3D uses a variant of the NEMO-1D approach except that all I/O items are defined in XML. The C++ data structures and a Java GUI are created dynamically from the XML.

OMEN handles I/O at the core based on LEX and YACC, with Tcl bindings to a metalanguage. Rapperture-based drivers let us create GUIs for rapid deployment on nanoHUB.org.

Sample Modeling Results and Impact

We’ve extensively used NEMO-1D, NEMO-3D, OMEN, and related concepts to analyze nanoelectronic devices, as documented in over 200 peer-reviewed publications. With the deployment of these tool concepts on nanoHUB.org, more researchers are already using these tools as well. In the following, we highlight a few of the fundamental tool capabilities in terms of both physical content and numerical performance.

NEMO-1D

NEMO-1D set the standard for quantitative RTD simulation. Figure 2b shows a prototypical comparison between NEMO-1D simulations and experimental data. Researchers have modeled high-performance, high-current-density RTDs with quantitative agreement with experiments. Also, simulations at low temperature demonstrate NEMO-1D’s capability to quantitatively model the phonon echo in the valley current.

NEMO-1D can require a significant computation time, and we developed a trilevel parallelism around voltage points and a double integral over momentum $k$ and energy $E$ on a relatively small Beowulf cluster. In 2007, in preparation for the petascale computing initiative, we demonstrated the parallel scalability of the NEMO-1D code to 23,000 processors on a Cray XT4 (see Figure 3a).

NEMO-3D

NEMO-3D can compute the electronic structure in typical semiconductor systems in the Zincblende Group III-V semiconductor and Silicon-Germanium (Si/Ge) material systems. We can apply closed boundary conditions for calculated electron states. As the buffer is increased, the same states can still be computed; we’ve demonstrated a system size of 52 million atoms. (c) OMEN’s parallel scaling results for an InAs High-Electron Mobility Transistor (HEMT) device on ORNL’s Jaguar.

Figure 3. End-to-end performance results. (a) Parallel scaling of Nanoelectronic Modeling 1D (NEMO-1D) for a hole-based resonant tunneling diode on the Oak Ridge National Laboratory (ORNL) Cray XT3/4. Trilevel parallelism in current (I) and the double integral in momentum ($k$) and energy ($E$) results in very good scaling. (b) Demonstration of an end-to-end performance benchmark of NEMO-3D for an Indium arsenide (InAs) quantum dot embedded in Gallium arsenide (GaAs). The GaAs buffer is increased to increase the simulation domain with little effect on the central confined states. As the buffer is increased, the same states can still be computed; we’ve demonstrated a system size of 52 million atoms. (c) OMEN’s parallel scaling results for an InAs High-Electron Mobility Transistor (HEMT) device on ORNL’s Jaguar.
wires. Some of the NEMO-3D usage cases include the modeling of valley splitting in tilted Si quantum wells on disordered SiGe and the metrology of single impurities in Si-FinFETs (Silicon-fin field effect transistors) where we achieved agreement of multimillion-atom electronic structure simulations with experimental data without any adjustments to previously published material parameters. Here, we highlight the modeling of InAs self-assembled quantum dots grown on GaAs and selectively capped with $\text{In}_x\text{Ga}_{1-x}\text{As}$, where the Indium mol fraction $x$ is varied from 0 to 40 percent to achieve optical activity of the InAs quantum dots at the optical communication wavelength of 1.5 micrometers. Figure 2a and the insert of Figure 4a show a nominally 20-nm wide and 5-nm tall dome-shaped quantum dot embedded by the random alloy. Figure 4a shows three experimental curves under different growth conditions, where each point on the curve represents a different device. We

- represented a total volume of $60 \times 60 \times 66$ nm$^3$ with 9-million-atoms,
- chose the nominal quantum dot size as given in the experiment,
- didn’t modify any of the previously published atomistic material parameters, and
- varied only the Indium fraction.

To our surprise, the NEMO-3D simulations overlaid the experiments rather interesting nonlinear behavior. Careful analysis shows that two critical atomistic details are important to achieve such quantitative agreement. First, the InGaAs alloy’s bonds must be distributed bimodally and retain their In-As and Ga-As bond lengths rather than build an average (InGa)-As bond length. Second, the quantum dot must change its shape. The requirement to model the bimodal bond-length distribution precludes use of a continuum material representation. NEMO-3D provided a virtual microscope, which enabled us to understand non-trivial mechanical and electronic structure interactions that cannot be measured experimentally.

**OMEN**

OMEN solves the coupled Poisson and Schrödinger equations on an atomistic basis with open boundary conditions. Atoms can be arranged with each other through bonds in arbitrary 3D geometries (see Figures 1c through 1f). The electron and current density must be computed in a 3D spatial domain for different electron energies and momenta, which are accumulated in a double integral. This must be performed for many bias points in a current-voltage characteristic. Spatial decomposition, a double integral in energy and momentum, and multiple bias points offer the opportunity for four levels of parallelism. Indeed, OMEN has been scaled to more than 220,000 cores on the world’s second fastest computer. Such large-scale parallelism lets us reduce compute time from several years on a single CPU to approximately 15 minutes on a parallel machine.
OMEN researchers are using 2,000 to 8,000 cores on parallel computers—sponsored by the US National Science Foundation and Department of Energy—to explore nanoelectronic device concepts and optimizations.

OMEN does, however, have serious limits that must be overcome. Coherent transport calculations are limited by the cross section, in which transport is modeled atomistically. The limiting number of atoms in repeated cross-section cells is around 2,000 atoms and therefore limits the cross sections to around 80 nm². When we introduce incoherent scattering, all electron degrees of freedom in energy and momentum are coupled and the computational cost increases by a factor of 200 to 1,000. This increase in computational cost is required because the pure wavefunction approach must be replaced by the NEGF approach at a computational cost of about 10× and for each bias point we need 20 to 100 iterations in the self-consistent Born approximation loop to couple all degrees of freedom. Finally, the treatment of surfaces (open, passivated, and relaxed), the amorphous dielectrics, and the strongly polar-bonded semiconductors remain problems with the empirical tight-binding approach.

Researchers have used OMEN to

- investigate the influence of interface roughness on the threshold voltage of triple-gate Si nanowires,
- simulate the performances of n- and p-doped ultra-thin-body field-effect transistors with different crystal orientations,
- reproduce experimental data for realistically extended InAs high-electron mobility transistors,
- study the properties of single- and double-gate ultra-thin body and gate-all-around nanowire InAs tunneling field-effect transistors,
- determine the limitations of graphene-based tunneling FETs.

In addition, researchers have used OMEN to study the energy loss mechanisms through electron-phonon scattering and their impact on nanowire transistors.

**Broad Impact on nanoHUB.org**

Use of the NEMO and OMEN tools isn’t restricted to just an elite few; in fact, they’re used as engines for five nanoHUB tools—Quantum Dot Lab, Bandstructure Lab, OMENwire, OMENFET, 1Dhetero, and RTDnegf—that have served more than 4,000 users in more than 60,000 simulation runs. Tool execution time varies from a few seconds on a single virtual machine to several hours on a parallel computer with 256 cores. The tools are cited in the scientific literature more than 40 times.

The NEMO and OMEN tool suite brings together material and device-modeling capabilities at the atomic resolution to impact realistically large devices. The codes perform well on serial and parallel computers, deliver results that explain and guide experiments, and are in the hands of real users with problems to solve.

**Acknowledgments**

NEMO and OMEN developments have been ongoing for more than 15 years and have involved many professionals, postdocs, and students who’ve spent countless hours working on the codes. The list of people would be too long to provide here, but the following list of critical references is part of our acknowledgments. In our work, we used nanoHUB.org’s computational resources, Oak Ridge National Laboratory’s leadership computing, and the TeraGrid resources Ranger and Kraken. Work at the Network for Computational Nanotechnology is supported by grants from the US National Science Foundation, the Semiconductor Research Corporation, Nanoelectronics Research Initiative, Focus Center Research Program, the Army Research Office (ARO), the State of Indiana, and Purdue University.

**References**


Gerhard Klimeck is a professor of electrical and computer engineering at Purdue University, where he is director of the Network for Computational Nanotechnology and leads the development and deployment of Web-based simulation tools hosted on nanoHUB.org. His research interests include modeling nanoelectronic devices, parallel cluster computing, and genetic algorithms. He was previously a principal member of the technical staff at JPL (where he developed NEMO-3D) and a member of the technical staff at the Texas Instruments Central Research Laboratory (where he developed NEMO-1D). Klimeck has a PhD in electrical engineering from Purdue University. He is a senior member of the IEEE and a member of the American Physical Society, Eta Kappa Nu, and Tau Beta Pi. Contact him at gekco@purdue.edu.

Mathieu Luisier is a research assistant professor at the Network for Computational Nanotechnology, Purdue University. His research interests are in quantum transport in nanoscale devices, such as multigate and tunneling field effect transistors; parallel numerical algorithms for large computer applications; and the development of next-generation computer-aided design (CAD) tools. Luisier has a PhD in electrical engineering from the Swiss Federal Institute of Technology, Zurich; his thesis on atomistic and full-band simulation of nanowire transistors eventually evolved into OMEN. Contact him at mluisier@purdue.edu.