Atomistic Modeling of Realistically Extended Semiconductor Devices with NEMO/OMEN

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

Mathieu Luisier
Purdue University - Main Campus

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.
Atomistic Modeling of Realistically extended Semiconductor Devices with NEMO / OMEN

Gerhard Klimeck, Mathieu Luisier, Purdue University

Keywords: J.8.f, H.5.2.f

The Nanoelectronic Modeling (NEMO) Toolset has been continually developed over the past 15 years to provide insight into nano-scale semiconductor devices that are dominated by quantum mechanical effects. The ability to represent realistically large devices in an atomistic basis has been the key element to match experimental data and to guide experiments. Now these insights flow into the creation of a new simulation engine called OMEN. Critical modeling requirements, scaling on parallel computers, and sample scientific results are discussed.

SEMICONDUCTOR INDUSTRY IS FACING A CRISIS

Moore’s law has been driven by economic forces producing a 260 billion dollar annual revenue industry. These forces were strong enough to overcome any of the predicted roadblocks that presumably were to stop Moore’s law in the past 15 years. There is, however a fundamental limit in down-scaling that cannot be overcome: atoms will not be divisible and down-scaling must stop in the realm of countable atoms. Indeed some material layers in commercial devices have reached the thickness limit of a few atoms. Lateral dimensions are now at 20-30nm and device geometries are no longer flat planar but 3-D dimensional objects.

New materials enter the “device design” 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 the Silicon-based industry is considering new transistor materials such a carbon and even the eternal material of the future: GaAs and InAs. The size of the design space suddenly exploded since the detailed local arrangements of atoms become critical and new materials are added in novel 3D geometries.

Modeling and simulation may offer means to explore options before experiments. However, it is probably fair to say that none of the typical commercial semiconductor device design tools can be used to explore the 3D atomistically defined search space. Most tools are based on continuum material assumptions therefore ignore the mere existence of the atomic granularity and most tools at the very best patch in quantum mechanical effects into the simulation concept through perturbative treatments. However, a fundamentally quantum mechanical carrier transport model built onto an atomistic material description is required.

NEMO1D – THE FIRST INDUSTRIAL QUANTUM TRANSPORT SIMULATOR

The first analog and digital nanoelectronic devices operated at room temperature with material variations on the nanometer scale were resonant tunneling diodes (RTDs). Indeed these devices require a quantum mechanical understanding and cannot be modeled with semi-classical approaches. The Central Research Lab of Texas Instruments had 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 the design of RTDs. 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 physical origin of the valley current (Figure 1).
Towards the end of the project in 1997 the modeling was able to guide experiments quantitatively (Fig. 1b) [1] with a bottom line that the off current in high performance, high current density devices cannot be reduced dramatically in standard RTDs. Design alternatives for low power memory cells were eventually patented [2]. While general interest in RTDs has subsided since then, critical insights resulting from modeling carrier transport at the nanometer scale were achieved.

![Figure 1. Resonant Tunneling Diode - a) conceptual sketch of a double barrier structure under various bias conditions, leading to a current turn-on and turn-off with increasing bias. Controlling the peak and the valley (off) current and their relative size is critical to transistor action and the amount of energy consumed when the device is “off” (b) Experimental data from 12 different I-V curves overlapping theoretical prediction of NEMO 1-D.](image)

The NEMO1D team developed new boundary conditions that enabled the treatment of extended device contact regions in which strong scattering / thermalization of carriers and electrostatic control are critical, while quantum mechanically confined states still rule over the carrier injection into the central device. Most members of the RTD community had suspected that incoherent scattering in the central device region is the key element to the valley current. That effectively turned out to be incorrect. The critical element was actually to understand where the resonances are in energy and how they are coupled to the contacts. An atomistic representation of the device layers was needed to enable quantitative modeling of carrier transport in room temperature high performance, high current density RTDs.

**NEMO1D – SETTING THE PATH FOR GENERAL NANOELECTRONIC DEVICE SIMULATION**

NEMO1D set the model, user, and developer requirements for a variety of different simulation engines that followed.

**Model Requirements**

- Must not be based on continuum material description – must include atomistic granularity, 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 meV and effective masses to within a few percent – bandstructure engineering works in the domain of tuning bandedges to within a few tens of meV.
- Must have transferability from bulk to the nanometer scale
- Must be computationally light-weight 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 finite extent in realistic environments – precludes the assumption of infinite periodic structures surrounded by 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.
- Atomistic model should represent atomistic disorder explicitly – rather than a statistical average way – each device is indeed different from the next.

These requirements guided the choice of basis sets in subsequent developments of the NEMO team as well as other researchers around the world. “Exact” bandgaps and masses, millions of atoms, and finite extent preclude most typical ab-initio models. In fact one can argue that for typical semiconductor devices these may not be needed anyhow since device simulation does not need to establish the existence and formation of bonds. If the bonds were varying in time, that would be a pretty unstable semiconductor device, plagued by noise. The chosen empirical tight binding approach meets the requirements and needs stated above.

![Image](image_url)

**Figure 2. Nanoelectronic device geometries.** (a) Multi-million atom NEMO 3-D simulation geometry of an InAs quantum dot on a GaAs substrate, capped by an alloy InAlAs strain reduction layer. The “As” atoms are not shown for clarity. (b) Zoom view of (a) from a different view angle. (c-f) 2-D and 3-D Device geometries for OMEN transport simulations. (c) Ultra-scaled thin body transistor. (d) Gate all-around nanowire. (e) Gate-all-around carbon nanotube. (f) Top-gated graphene sheet.

**User Requirements**

We envisioned various classes of users: 1) computational scientists, 2) experimentalists, 3) educators/students, and 4) NEMO developers.

**Computational scientists** typically push the model development and validation. They typically 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 gain fundamental insight. They are willing to move data and restart files all over the place, perform data analysis in ad-hoc interfaces.

**Experimentalists** generally know little about the very details of the model 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 should be seconds to minutes. The secondary execution times might be the lunch hour or a few simulations over night. Data management, visualization, and export into notebooks must be integrated into the overall tool experience. The software must not demand that the user has an intricate knowledge of the underlying theory. Computational science experiments that add more and more basis functions to test the validity of the model are inappropriate for this class of users.

**Educators and Students** need a rapid simulation turn-around as the Experimentalists, but they also need a yet simplified access method. Use in the class-room or wherever, without any license and installation issues. They need reuse of previous simulation results or set-up before class time and simple result reporting and interpretation.

**NEMO Developers** included theorists, computational scientist, algorithm developers, software
engineers, user interface designers. They needed to work together in various phases of the software project. These developers have quite different requirements in their usual work flow. At one end of the spectrum the code may need to change on a daily basis and the other end of the spectrum the code should be frozen for several months, while it is tested and rolled out to users. A dynamical I/O design that limits the exposure of new models to users and enables developers to add models rapidly is needed.

FROM NEMO 1-D TO OMEN - TECHNICAL APPROACHES

Model Selection

The fundamental transport methodology was identified as Non-Equilibrium Green Function (NEGF) approach and has been established as such in this field. The model requirements laid out above led to the selection of the use of empirical orthogonal tight binding (ETB) approaches. Different basis sets were implemented ranging from simple effective mass (single "s" orbital) to a suite of more comprehensive models such as sp\(^2\)s\(^*\) and sp\(^3\)d\(^*\) typically in nearest neighbor representations.

ETB is based on the symmetry-formalized interaction of valence electrons on neighboring atomic sites. ETB ignores the core electrons and is therefore not a total energy Hamiltonian. But ETB can be fit to match experimental and more fundamental theoretical bandstructure properties [3] when the model is improved to match general strain behaviors [4,4b].

NEMO 3-D development at NASA JPL

In 1998 NASA JPL had a strong experimental technology group interested in development of optical detectors and lasers. Self-assembled quantum dots were being grown and the need arose to guide the experimental 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 and the focal point of the work was to create an electronic structure simulator that can compute the confined conduction and valence band states of realistically extended quantum dot systems fully atomistically [4]. 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. NEMO3D was released as open source in 2003 and its development continued at Purdue University [4c].

OMEN development at ETH and Purdue

In the years 2004/05 it became clear that the compute power needed to perform atomistic transport simulations for extended devices would become available soon. Approaches to 3D transport in nanowires using a simple effective mass model were under way already and being deployed on nanoHUB.org. Mathieu Luisier developed during his Ph.D. work at ETH in collaboration with Purdue several completely new Matlab prototypes and then a C++ foundation for a general 3D quantum transport approach [4d]. At first the simulators did not include incoherent scattering and a wavefunction approach was used to reduce the compute time. Since 2008 Luisier is developing the OMEN transport tool at Purdue and the first implementations that include incoherent scattering were achieved in 2009 [15].

Input / Output Approaches - User Interface Designs

The development of rapidly evolving software and its rapid deployment to a user base has its own interesting requirements. Graphical user interfaces are built most easily when I/O is carved in stone and does not change. However, theory and algorithm developers have a need to add new models and algorithm parameters rapidly. For the three large software projects described here we have experimented with different I/O handling mechanisms.

NEMO 1-D uses a mixed static and dynamic GUI design approach. A GUI developer implemented a set of static Windows that are filled with dynamically defined datastructure objects. Device design and Material data windows can be dynamically expanded with new design descriptions and material
models. Simulation and algorithm parameters are structured in parent-child related dependency trees that can be dynamically configured at compile time. With this dynamic design we were able to decouple the static GUI work of the software engineer from the dynamic requirements of the algorithm and theory developers. All I/O is defined in C data structures and translators convert these into GUI components or batch input deck items. Another key element in NEMO 1-D I/O is the ability to define all material parameters in an externally scriptable interface.

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

OMEN uses a lex/yacc based I/O handling at the very core with tcl bindings to a meta language. Rappiture-based drivers allow for the creation of GUIs for rapid deployment on nanoHUB.org.

SAMPLE MODELING RESULTS AND IMPACT

NEMO 1-D

NEMO 1-D has set the standard for quantitative resonant tunneling diode simulation. Figure 1b indicates a prototypical comparison between NEMO 1-D simulations and experimental data. High performance, high current density RTDs were modeled in reference [1]. Simulations at low temperature demonstrate the capability to quantitatively model the phonon echo in the valley current [5].

NEMO 1-D can require a significant computation time and a tri-level parallelism around voltage points and a double integral over momentum, k, and energy E had been developed [6] on a relatively small Beowulf cluster. As a preparation for the peta-scale computing initiative we demonstrated [6a] in 2007 the parallel scalability of the NEMO 1-D code to 23,000 processors on a Cray XT4 (Fig. 3a)

![Figure 3. (a) Parallel scaling of NEMO 1-D for an end-to-end simulation of a hole-based resonant tunneling diode. Tri-level parallelism in current (l) and the double integral in momentum (r) and energy (E) results in very good scaling. (b) Demonstration of an end-to-end performance benchmark of NEMO 3-D for an InAs quantum dot embedded in 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 and a system size of 52 million atoms is demonstrated. (c) Parallel scaling of OMEN of an InAs HEMT device [10][6a] ](image)

NEMO 3-D

NEMO 3-D can compute the electronic structure in typical semiconductor systems in the III-V, and Si/Ge material systems. Closed and periodic boundary conditions can be applied in various dimensions such that 2-D and 1-D devices can be considered as well. End-to-end calculations of 52 million atom systems have been demonstrated [4c] (Fig. 3b). 52 million atoms correspond to a cubic simulation domain of roughly (101nm)$^3$ or a laterally extended domain of 230x230x20 nm$^3$ or a nanowire geometry of 50x50x425 nm$^3$. This capability enabled the modeling of realistic structures for embedded quantum dot stacks, strained quantum wells and disordered wires. Some of the NEMO 3-D usage cases have been the modeling of valley splitting in tilted Si quantum wells on disordered SiGe [7] and the metrology of single impurities in Si-FinFETs modeling of single impurities [8] where agreement of multimillion atom electronic structure simulations was achieved with experiment without any adjustments to previously published material parameters.
Here we highlight [9] the modeling of InAs self-assembled quantum dots grown on GaAs and selectively capped with In$_x$Ga$_{1-x}$As, where the Indium x is varied from 0 to 40% to achieve optical activity at the optical communication wavelength of 1.5μm. A nominally 20nm wide and 5nm tall dome shaped quantum dot embedded by the random alloy is visualized in Figure 1a. Fig. 4a shows three experimental curves under different growth conditions where each point on the curve represents a different device. A set of 9 million atom NEMO3D simulations where the nominal quantum dot size as given in the experiments, no material parameters are modified and only the Indium fraction is varied virtually lays on top of the experimental data. The interesting part is that the data shows a non-linear behavior in the experiment and the simulation. Careful analysis shows that two critical atomistic details are important to achieve such quantitative agreement: 1) the bonds in the InGaAs alloy are distributed bi-modal retaining their In-As and Ga-As bond lengths rather than building an average (InGa)-As bond length and that the quantum dot changes its shape [9]. The requirement to model the bi-modal bond-length distribution precludes the usage of a continuum material representation.

OMEN

OMEN solves the coupled Poisson and Schrödinger equations in an atomistic basis with open boundary conditions. Atoms can be arranged with each other through bonds in arbitrary 3-D geometries (Fig. 2c-f). The electron and current density must be computed in a 3-D spatial domain for different electron energies and momenta, which are accumulated in a double integral. This has to be performed for many bias points in a current-voltage characteristic. This offers the opportunity for 4 levels of parallelism. Indeed OMEN has been scaled to over 220,000 cores [6a] on the second fastest computer in the world. Such large-scale parallelism enables the reduction of compute time from about several years on a single CPU to around 15 minutes on a parallel machine. On a day-to-day basis OMEN researchers are using 2,000 to 8,000 cores on NSF and DOE sponsored parallel computers to explore nanoelectronic device concepts and optimizations.

OMEN does, however, still have serious limits that need to be overcome. The 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 80nm$^2$. When incoherent scattering is introduced, all electron degrees of freedom in energy and momentum are coupled and the computational cost increases by a factor of 200 to 1,000. The pure wavefunction approach needs to be replaced by the non-equilibrium Green function approach at a computational cost of about 10x and then one needs 20-100 iterations in the self-consistent Born approximation loop which couples all degrees of freedom. Finally the treatment of surfaces (open, passivated, and relaxed), the amorphous dielectrics and strongly polar bonded
semiconductors remain problems with the utilized empirical tight binding approach.

OMEN has been used for example to investigate the influence of interface roughness on the threshold voltage of triple-gate Si nanowires [11], to simulate the performances of n- and p- doped ultra-thin-body field-effect transistors with different crystal orientations [12], to reproduce the experimental data of realistically extended InAs high-electron mobility transistors [10], to study the properties of single- and double-gate ultra-thin body as well as gate-all-around nanowire InAs tunneling field-effect transistors [13], and to determine the limitations of graphene-based tunneling field-effect transistors [14]. More recently, the effect of electron-phonon scattering on the performances of Si nanowire transistors has been considered using OMEN [15].

Broad Impact on nanoHUB.org

The NEMO and OMEN tools are not just restricted in their use to an elite few, but they are used as engines for 5 nanoHUB tools: Quantum Dot Lab, Bandstructure Lab, OMENwire, 1Dhetero, and RTDnegf. These tools have served over 4,100 users with over 46,000 simulation runs. The execution time varies from tool to tool from a few seconds on a single virtual machine to several hours on a parallel computer with 256 cores. These tools are cited in the scientific literature over 40 times.

Acknowledgements

The NEMO / OMEN developments has been a process extending over 15 years which involved many professionals, post-docs, and students who have spent countless hours working on the codes. The list of people would be too long to provide here, but the critical citations are listed below as part of the acknowledgement. Computational resources of nanoHUB.org, the leadership computing facility at Oak Ridge National Lab, and the TeraGrid resources Ranger and Kraken have been used in this work.

The conclusion.

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

Vita:

Gerhard Klimeck is Director of the Network for Computational Nanotechnology (NCN) and professor of electrical and computer engineering at Purdue University. He 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. Prior to Purdue Klimeck was a Pricipal Member of the Technical Staff at the NASA Jet Propulsion Laboratory, Caltech where he developed NEMO 3-D and a of the Member Technical Staff at the Central Research Laboratory of Texas Instruments where he developed NEMO 1-D. Klimeck has a PhD in electrical engineering from Purdue and is a senior member of the IEEE as well as a member of the American Physical Society, Eta Kappa Nu, and Tau Beta Pi. Contact him at gekco@purdue.edu.

Mathieu Luisier received the degree of Dipl.-Ing (with honors) in electrical engineering in 2003 and the doctoral degree in 2007 from the ETH Zurich, Switzerland. His PhD thesis was about atomistic and full-band simulation of nanowire transistors. This work gave birth to a parallel nanowire device simulator called GreenSolver, renamed OMEN today and available on the nanohub. In February 2008 he joined the Network for Computational Nanotechnology, Purdue University, USA as a research assistant professor. His current research interests are focused on quantum transport in nanoscale devices, like multi-gate and tunneling FETs, on parallel numerical algorithms for large computer applications, and on the development of next generation computer aided design (CAD) tools. Contact him at mluisier@purdue.edu.

References

1. R. Chris Bowen, Gerhard Klimeck, Roger Lake, William R. Frensley and Ted Moise, "Quantitative Resonant


