Computational Challenges in Nanoscale Device Modeling

E. Polizzi, A. Sameh and H. Sun

Department of Computer Sciences, Purdue University
250 N. University Street, West-Lafayette, IN 47907-2066, USA
epolizzi@purdue.edu

ABSTRACT

The development of new simulation tools is critical for the exploration of quantum transport in nanoscale devices. Such simulation is commonly performed by solving self-consistently the transport problem using the Non-Equilibrium Green's Functions (NEGF) formalism and the Poisson's equation to account for the space charge e ects. The quest for ever higher levels of detail and realism in such simulations as the modeling of multidimensional devices with detailed band structure calculations with(or without) the inclusion of scattering e ects, requires huge computational e ort. Hence, the need for an active research e ort in developing novel numerical techniques and parallel algorithms that are ideally suited for high-end computing platforms. In this article, we will identify the identify the challenging numerical problems which arise from the NEGF/Poisson procedure and we will present new efficient parallel schemes for computing the problem.

Keywords: Nanoscale devices, Green's function, NEGF-Poisson, parallel numerical algorithms, linear systems, generalized eigenvalue problems

1 INTRODUCTION

The modeling and the numerical simulations of quantum transport is expected to play an essential role to design novel high speed and high functionality class of devices whose electron transport properties are mainly based on quantum e ects (as tunneling, interferences effects and electron con nements). Among all these new proposed devices, we nd: the double-gate MOSFETs, the silicon nanowire transistors, the resonant tunneling diodes (RTD), the electron waveguide devices, the spintronics devices, the molecular devices, the nanotubes, etc...

As showed in Fig. (1), the model commonly used to characterize the electron device, consists in solving self-consistently for a given voltage characteristics at the Source, Drain or Gates contacts, the transport problem for the electrons using the NEGF formalism, and the electrostatics problem using the Poisson's equation (Hartree approximation). Once the potential pro le inside the device is known, the current density (or conductance, etc ...) can be computed. In order to plot for example the I-V curves, the previous self-consistent procedure is then repeated for all the required inputs voltage characteristics.

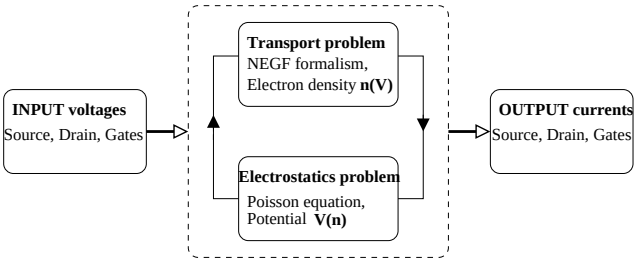

Figure 1: Current-voltage characteristics obtained using a self-consistent process between the calculation of the electron density and the electrostatics potential.

2 THE QUANTUM TRANSPORT

2.1 NEGF Formalism

The NEGF formalism is a general and powerful formalism which can be used to study any type of nanoscale devices using various physical models. Moreover, it allows to consider in the transport problem both the interactions with the reservoirs (open system and non-equilibrium transport), and the scattering e ects (such as electron-phonons, electron-photons, etc...) [1]. For a given energy E, the expression of the Green's function $G(E, \mathbf{x}, \mathbf{x}')$ inside the device ($\mathbf{x}, \mathbf{x}' \in$) is given by

$$\Big(E \quad H(V) \quad \Sigma_c(E) \quad \Sigma_s(E)\Big) G(E, \mathbf{x}, \mathbf{x}') = \quad (\mathbf{x} \quad \mathbf{x}'), \tag{1}$$

where E is the energy, H is the Hamiltonian operator which depends on the potential and it contains all the information about the electronic band-structure of the system, $\Sigma_c(E)$ is the self-energy function which de nes the interaction of the device with its external environment (such as electrons reservoirs), and $\Sigma_s(E)$ is the self-energy function which de nes the interaction of the device with its internal environment (see Fig. 2).

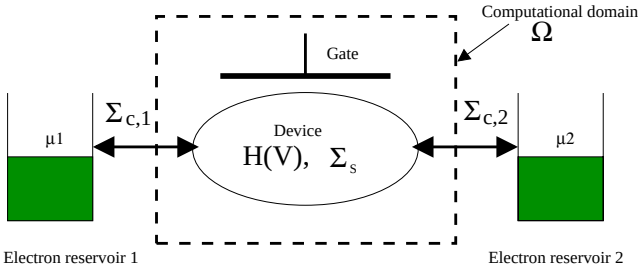

Figure 2: A device driven out of equilibrium by two contacts with di erent Fermi levels $_1$ and $_2$ (we note Σ_c the sum of these two self-energy functions $\Sigma_{c,1}$ and $\Sigma_{c,2}$).

The expression of the Hamiltonian $H(V)$ will depend on the levels of sophistication of the model used to de ne the electronic structure. A full quantum mechanical treatment of the system would require a many-body approach which appears numerically intractable. Hence, the need to work within the single electron picture where di erent levels of approximation can be derived from the way how the interaction of the single particle with the other electrons is taken into account. At the lowest order approximation, H can be de ned by the e ective mass Schrodinger operator

$$H = \frac{h^2}{2}\nabla \quad \frac{1}{m^*(\mathbf{x})}\nabla \qquad qV(\mathbf{x}), \tag{2}$$

with m^* the e ective mass which depends on the material. However one should notice that in order to handle the electronic structure with much more fundamental accuracy, it should be necessary to account for the direct informations at the atomic levels. Both ab-initio approach (obtained with rst principle calculations) and semi-empiral approach (obtained with tting parameters), can be used for such purposes.

The NEGF formalism allows to describe any type of interactions with the surroundings using the concept of self-energy functions. The self-energy function $\Sigma_s(E)$, are related to the scatterings for the electron transport inside the device as: electron-phonons, electron-photons, etc... The derivation of this term would increase dramatically the computational cost of the simulations, since $\Sigma_s(E)$ depends on all the Green's functions for all the energies. Since this problem is still under investigation, they will not be described in this article. Therefore, we will assume in the following that Σ_s is equal to zero, and the electron transport is then ballistic.

2.2 Self-energy functions and open boundary conditions

The Green's function can be considered as the wave function at $\mathbf{x}$ resulting for a unit excitation applied at $\mathbf{x}'$. We are only interested by the retarded Green's function which represents the response of the system of an impulse excitation within the device (then $\mathbf{x}' \in$). In the particular (but commonly used) case where the electron reservoirs can be considered as multidimensionnal semi-in nite leads, the potential is assumed invariant by translation along the transport direction and the solutions of the Green's function are planes waves. We can then construct the Green's function inside the leads and calculate the self-energy function using for example, the embedding approach described in [2].

Otherwise, an elegant formalism, the QTBM (Quantum Transmitting Boundary Method) proposed in [3], leads to express the boundary expressions independantly of the unknown re ection coefficients of the plane waves in the leads. When the incoming waves are not taken into account (since we are looking for the retarded Green's function), these boundary conditions are equivalent to the self-energy functions Σ_c [4]. For example, the self-energy matrix obtained for a 1D problem (one device and two leads) within the e ective mass approach (2), and using a Finite Element Discretization (FEM), is given by:

$$[\Sigma_c(E)]_{ii'} = \frac{i}{2}\sum_{j=1}^{2} \quad _{i\in\gamma_j} \quad _{i'\in\gamma_j} k_j(E), \tag{3}$$

where k_j denotes the wave vector associated to the lead j, and $_1$, $_2$ are the boundaries of the domain .

3 IDENTIFY THE CHALLENGING NUMERICAL PROBLEMS

In the following, we propose to list the main numerical problems encountered in the NEGF/Poisson procedure, and to give the limitation of the algorithms currently used.

3.1 Linear systems on the Green's function

The Green's function can be expanded in terms of some set of basis functions, such as

$$G(E,\mathbf{x},\mathbf{x}') = \sum_{i,i'} G_{i,i'}(E)\omega_i(\mathbf{x})\omega_{i'}(\mathbf{x}'). \tag{4}$$

Using a nite element basis ω_i, this expansion becomes an approximation. After discretization, the variational form of (2) leads to the following expression of the ballistic Green's function in the matrix notation:

$$[G(E)] = \Big(E[S] \quad [H(V)] \quad [\Sigma_c(E)]\Big)^{-1}, \tag{5}$$

where the matrix elements on the overlap matrix $[S]$ are given by

$$[S]_{ii'} = \int_\Omega \omega_i(\mathbf{x})\omega_{i'}(\mathbf{x})\ d\mathbf{x}. \tag{6}$$

The expression and the size of the matrices $[H]$ and $[\Sigma_c]$ depend on the model used to describe the electronic structure of the device. For examples for the 3D simulation of semiconductor nanodevices, switching between an e ective mass approach for the electron transport [5], to an atomistic tight-binding sp3d5s* approach [6], will increase the size of the matrix by a factor $\sim O(10^4)$ (from $\sim O(10^5)$ to $\sim O(10^9)$). In the general cases, $[H]$ is hermitian and sparse ($[H]$ is real symetric using FEM), $[\Sigma_c]$ is complex symetric and very sparse (involving only few non-zero elements compare to $[H]$).

For a given energy, one can show that only few columns of the Green's functions are involved in the computation of the electron density. The electron density is then given by the diagonal elements of the matrix density

$$[\rho] = \int_{-\infty}^{+\infty} \frac{dE}{2\pi} \sum_{j=1}^{N} f_{FD}(E \quad _j)[\mathrm{A_j(E)}], \qquad (7)$$

where the electrons reservoirs are assumed to be non interacting systems at equilibrium, with a Fermi-Dirac distribution f_{FD} for the electrons. We then denote $_j$ the fermi level associated to the reservoir j. The spectral function $[\mathrm{A_j}]$ associated to the states lled by the reservoir j is given by

$$[\mathrm{A}_j(E)] = [G(E)]\ [\Gamma_j(E)]\ [G(E)]^{\dagger}. \qquad (8)$$

where Γ_j is de ned as the broadening function of the contact j, which is equal to two times the imaginary part of $[\Sigma_c]$ associated only to the contact j. Since Γ_j has the same very sparse structure of $[\Sigma_c]$, only few columns of the Green's function are then involved in the calculation of $[\mathrm{A_j}]$ (and then in the calculation of the electron density).

Finally, the NEGF formalism involves to solve by energy, the following linear system with multiple right hand side

$$[G(E)]^{-1}[X] = [B], \qquad (9)$$

where $[B]$ is a $(n \times c)$ matrix, with c ($c << n$) the number of columns of the Green's function that we are looking for by energy (which corresponds to the number of columns of Σ_c which contains non-zero elements). Therefore, each column of $[B]$ contains only one non zero element (equal to one) which is used to select the desired column of the Green's function written in $[X]$ (also a $(n \times c)$ matrix).

However, we note that when the contact can be considered as semi-in nite leads, it is computationally more efficient to consider the "wave function formalism" used to compute the electron density in [5]. Indeed, in this case, one can show that $[B]$ becomes a $(n \times m)$ matrix where m ($m < c$) is the number of excitation by energy (excitations which come from the contacts). In this case, $[B]$ also depends on the energy and the matrix $[X]$ gives the m solutions for the wave function for a given energy. The two approaches are analytically equivalent.

The linear systems are currently solved using an iterative scheme as QMR (Quasi Minimal Residual) algorithm with SSOR or ILUT preconditionner. Since the number of energy values in the system is usually of the order of $\sim O(1000)$, one needs to repeat this procedure for all the energies. However, this can be done independently on each processor using MPI directives, and the problem is then currently handled on a Linux cluster. One advantage of this parallel procedure is that the communications between for processors are minimal, but its major drawback is that each processor needs to solve many huge linear systems.

3.2 Linear system on the potential

After a FEM discretization of the Poisson equation on the electrostatics potential, we get one linear system to solve. The obtained matrix is real symetric, positive de nite and sparse. The size of this matrix can be very high (the number of the discretization points in the real space domain is currently $\sim O(10^{5-6})$ but can reach $\sim O(10^{10})$ for further application), but the linear system is solved only once by NEGF/Poisson iteration. However, this problem becomes very time consuming for some particular 3D devices as molecular [7] or nanotubes [8] devices which involve very huge domain for the electrostatics problem but only low dimensional domain for the transport problem. The PCG (Preconditioned Conjugate Gradient) algorithm with incomplete Cholesky factorization as preconditionner, is currently used in our applications, but it could become quickly ine ective to handle realistic atomistic devices.

3.3 Generalized eigenvalues problem for the subbands approach

Another modeling approach which can be applied to particular but often used devices, consists in the decomposition of the transport problem using subbands associated to the con ned directions for the electrons [9]. Using this method, one has now to solve also the m rst eigenpairs of many big generalized eigenvalues problems for a given potential, but the size of the Green's function matrix is drastically reduced (typically from N to $m * (N)^{1/3}$). Each of the $N^{1/3}$ eigenvalues problems (with a size of $N^{2/3}$) is currently solved using a LAPACK direct method, and independently on each processor with MPI directives [9], [13].

4 NOVEL PARALLEL ALGORITHMS

In this section, we present novel numerical techniques and parallel algorithms that are ideally suited for high-end computing platforms. In order to compute the linear systems both on the Green's function and on the

potential, we propose a parallel Spike algorithm based on an Implicit Block Jacobi [10], while the parallel trace minimization (TRACE-MIN) algorithm [11] is proposed to handle the generalized eigenvalues problems.

4.1 The SPIKE algorithm

Instead of using the previous parallel procedure which consisted in distributing the linear systems on the Green's function on a large number of processors (as done with a Linux cluster), we propose to make use of robust parallel algorithm to solve each linear system. Such strategy will allow us to consider nanoelectronics problems with high order of dimensions, but will require massive compute power, high scalability and fast communication, and the support for both the shared memory and distributed memory programming models.

The Spike algorithm is a powerful parallel banded linear solver which consits in preconditionning the original system with the bloc diagonals of the matrix to obtain a modi ed system with spikes. By permutation, one can form a small independant reduced system which can be solved directly or iteratively. The rest of the solution vector is retrieved directly. One can show that the algorithm enables multi-level of parallelism. In order to deal with the sparse matrices which appear in nanoelectronic modeling (using FEM for example), the Cuthill-McKee algorithm can be used to reorder the elements of these matrices to narrow banded matrices.

Using as direct solver, the Spike algorithm appears faster than ScaLAPACK [12] with only one level of parallelism. In its truncated version (only one small bloc of each spike is formed), the algorithm is suitable for diagonally dominant matrices allowing a speed factor up to 20 (with 16 processors on a Linux cluster) compare to LAPACK (one processor). We applied this scheme to compute the Green's functions of a 1D device de ned by the equations (2), (3) and (5), when the range of the energy makes the matrix diagonally dominant. In the most general case where the matrices are not diagonally dominant, the truncated Spike algorithm is expected to become a robust and efficient preconditioner which will enable very fast convergence for an iterative method (as BiCG-stab, QMR or Conjugate Gradient currently used). Moreover, the same preconditioner (updating time to time) could be used to solve the di erent linear systems along the energy, allowing a very high competitive procedure.

4.2 The TRACE-MIN algorithm

A novel parallel trace minimization solver is proposed for obtaining the smallest eigenpairs of those symmetric generalized eigenvalues problems that arise when considering the subbands decomposition approach of the transport problem.

The generalized eigenvalue problem is de ned by

$$Ax = \lambda Bx \tag{10}$$

where A and B are $n \times n$ real sparse symmetric matrices with B positive de nite (using FEM). Because of the large size of the problem, methods that rely only on operations like matrix-vector multiplications, inner products, vector updates, that utilize only high-speed memory are usually considered.

A variant of simultaneous iteration, called the trace minimization method, was proposed in [11] in attempt to avoid solving linear system $A\mathbf{x} = b$ repeatedly. Let X_k be the current approximation to the eigenvectors corresponding to the p smallest eigenvalues ($p << n$) where $X_k^T B X_k = I_p$. The idea of TRACE-MIN is to nd a correction term Δ_k that is B-orthogonal to X_k such that $tr(X_k \quad \Delta_k)^T A(X_k \quad \Delta_k) < tr(X_k^T AX)$. It follows that, for any B-orthogonal basis X_{k+1} of the new subspace $span\{X_k \quad \Delta_k\}$, we have $tr(X_{k+1}^T AX_{k+1}) < tr(X_k^T AX)$, i.e., $span\{X_k \quad \Delta_k\}$ gives rise to a better approximation of the desired eigenspace than space $\{X_k\}$. This trace reduction property can be maintained without solving any linear systems exactly. The trace minimization method can be implemented in the parallel machine which is expected to improve the computation time signi cantly for the simulation of the Silicon nanowire transistors devices [13].

REFERENCES

[1] S. Datta, Electronic transport in mesoscopic system, Cambridge University Press, 1995.

[2] J. E. Ingles eld, J. Phys. C 14, 3795, 1981.

[3] C. S. Lent and D. J. Kirkner, J. Appl. Phys. 67, 6353, 1990.

[4] E. Polizzi and S. Datta, IEEE nano, 2003.

[5] E. Polizzi and N. Ben Abdallah, Phys. Rev. B, 66, 245301, 2002.

[6] G. Klimeck, NEMO-3D, see http://www-hpc.jpl.nasa.gov/PEP/gekco/

[7] F. Zahid, M. Paulsson, E. Polizzi and S. Datta, unpublished, 2003.

[8] J. Guo, J. Wang, E. Polizzi, S. Datta and M. Lundstrom, accepted on IEEE Trans. on Nano., 2003.

[9] E. Polizzi and N. Ben Abdallah, submitted, 2003.

[10] A. Sameh and V. Sarin, Inter. J. of comput. uid dynamics, 12, 213, 1999.

[11] A. Sameh and Z. Tong, , J. Comput.& Appl. Math, 123, 155, 2000.

[12] C. Xu, Master's thesis, Purdue University 2003.

[13] J. Wang, E. Polizzi, and M. Lundstrom, IEDM2003, 2003.