Atomistic nanoelectronic device engineering with sustained performances up to 1.44 PFlop/s

Mathieu Luisier*, Timothy B. Boykin†, and Gerhard Klimeck*, and Wolfgang Fichtner‡
*Network for Computational Nanotechnology, Purdue University, West Lafayette, IN 47907, USA
†ECE Dept., University of Alabama in Huntsville, Huntsville, AL 35899, USA
‡Integrated Systems Laboratory, ETH Zürich, 8092 Zürich, Switzerland

***Abstract*—We present a multi-dimensional, atomistic, quantum transport simulation approach to investigate the performances of realistic nanoscale transistors for various geometries and material systems. The central computation consists in solving the Schrödinger equation with open boundary conditions several thousand times. To do that, a Wave Function approach is used since it can be relatively easily parallelized. To further improve the computational efficiency, three additional levels of parallelization are identified, the work load is optimally balanced between the CPUs, computational interleaving is applied where possible, and a mixed precision scheme is introduced. Using two different device types, a high electron mobility and a band-to-band tunneling transistor, sustained performances up to 1.28 PFlop/s in double precision (55% of the peak performance) and 1.44 PFlop/s in mixed precision are reached on 221,400 cores on the CRAY-XT5 Jaguar at Oak Ridge National Lab.**

I. Introduction

Electronic devices have seen a tremendous increase of their functionalities since the invention of the radio transistor in the 1950's. For example, laptops, desktops, GPS, cameras, and cell phones have kept improving from one generation to the other due to software progresses, but also due to better hardware characteristics. As predicted in 1965 by Gordon Moore, Intel's co-founder [1], the number of transistors that can be placed on an integrated circuit (IC) has doubled almost every two years, reaching nowadays a value larger than 1 billion. At the same time, the performance of each single transistor has been constantly enhanced. The combination of these two factors has led to more powerful ICs with better computational capabilities and has allowed for the development of complex and portable electronic devices.

To double the number of transistors per area, their lateral dimensions are reduced by about 70% each time a new technology node (TN) is introduced. We are currently at the 32nm TN [2] and transistors with a total length that does not exceed 112 nm are massively produced by semiconductor companies. For many years now the transistor structure has remained the same, a planar silicon metal-oxide-semiconductor field-effect-transistor (MOSFET) and the transition from one TN to the other could be achieved by "simply" using the transistor design from the previous generation, proportionally reducing its dimensions, and eventually optimizing its performances. Till very recently, the scaling of the transistor size alone automatically gave a performance boost of 25%.

However, at the horizon of 2018-2020, it will no more be possible to further scale the planar Si MOSFET and keep improving its performances due to material, structural, and power efficiency problems [3]. New device concepts need to be embraced that can overcome the limitations of the Si MOSFET and enable Moore's scaling law to continue. Among possible candidates to become the next generation switch, multi-gate devices [4], [5], III-V high electron mobility transistors (HEMTs) [6] and MOSFETs [7], graphene-based transistors [8], and band-to-band tunneling transistors (TFETs) [9] are often cited. The active region of all these devices is usually composed of a countable number of atoms and their behavior is dominated by quantum mechanical effects.

The development of new transistors has always been supported by computer aided design (CAD) tools that can predict the device characteristics before their fabrication. As the gate length of MOSFETs was longer than 100nm and their cross section larger than 15nm, the classical drift-diffusion (DD) approach [10], [11], [12] was accurate enough to give good predictions about transistor performances. The major concern about the DD model is that it does not capture energy quantization, quantum mechanical tunneling, the wave nature of electrons and holes, and the atomistic granularity of the devices, which are all essential at the nanometer scale. Hence, it is not only necessary to find a replacement for the Si MOSFET, but also for the simulation approach that will facilitate the discovery and the emergence of novel nanoelectronic devices.

A direct and self-consistent solution of the single-electron Schrödinger equation with open boundary conditions is more accurate than DD and fulfills the quantum mechanical requirement, but demands more computational power. However, the continuous increase of the CPU performances in the recent years represents a fantastic opportunity to re-think transistor simulation at the nanometer scale and go beyond standard approaches. In this context, we have developed OMEN, a next generation, multi-dimensional CAD tool based on quantum mechanical concepts and dedicated to the simulation of nanoelectronic devices [13]. Due to multiple parallelization levels, it can benefit from the largest available supercomputers to

SC11, November 12-18, 2011, Seattle, Washington, USA

investigate nanoscale transistors with an atomistic resolution. To ensure a good representation of the semiconductor properties while minimizing the computational burden, the empirical tight-binding method [14] has been chosen to describe the bandstructure of the simulated devices. Note that quantum transport based on the Kohn-Sham Density Functional Theory [15] is still computationally too intensive to be applied to realistic device structures and is restricted to molecules.

The implementation of OMEN started in 2005 at the ETH Zürich and has continued since 2008 at Purdue University. Since then, it has been used to study a broad range of transistor types. Some of them are schematized in Fig. 1. For example, nanowire [16], ultra-thin-body [17], graphene nanoribbon [18], high electron mobility [19], and band-to-band tunneling [20] transistors have been simulated in the ballistic limit of transport. It is also possible to treat dissipative (non-coherent) transport in OMEN by accounting for electron-phonon scattering [21]. More recently, the simulation capabilities have been extended to treat thermal transport and heat dissipation in nanoscale devices with out-of-equilibrium confined phonons.

In all these applications, the time-to-solution as well as the potential to treat large structures are key issues that must be properly addressed. The Non-equilibrium Green's Function (NEGF) formalism is very popular among the device modeling community to solve the Schrödinger equation with open boundary conditions [22], but it is computationally intensive, difficult to parallelize, and does not yet allow to simulate large transistors in an atomistic basis, especially in 2-D and 3-D. This is exactly the opposite of the Wave Function (WF) formalism [13] implemented in OMEN, which reduces the Schrödinger equation into a sparse linear system of equations in the case of ballistic or coherent transport (no inelastic scattering such as electron-phonon interactions).

We will therefore show in this paper that by combining the WF approach to other computational methods that are well-established in other research areas (load balance, computational interleaving, communication-computation overlap, mixed precision scheme), OMEN can simulate realistically extended devices in a reduced time and reach a sustained performance of 1.44 (1.28) PFlop/s in mixed (double) precision on 221,400 cores on the CRAY-XT5 Jaguar at Oak Ridge National Laboratory.

As benchmarks, two very recent Si MOSFET alternatives are selected, a multi quantum well $In_{0.53}Ga_{0.47}As$-InAs-$In_{0.53}Ga_{0.47}As$ high electron mobility transistor (HEMT) and an InAs band-to-band tunneling field-effect transistor (TFET). Both devices are based on different physics and electron transport concepts, they are simulated with different bandstructure models, but in both cases a sustained performance larger than 1.2 PFlop/s during more than 1 hour is measured in double precision, including the initialization phase, the central computation, and the generation of the output files. Such performances and capabilities open the doors for true nanoelectronic device engineering. Atomistic transistor design and optimization is now feasible in tens of minutes per device on a peta-scale machine.

The paper is organized as follows: in Section II, an overview of the simulation approach is given, followed by its implementation in Section III. The simulation setup for the HEMT and TFET structures is presented in Section IV as well as the accuracy of the mixed precision scheme. The performances of OMEN are finally described and analyzed in Section V before the paper is concluded in Section VI.

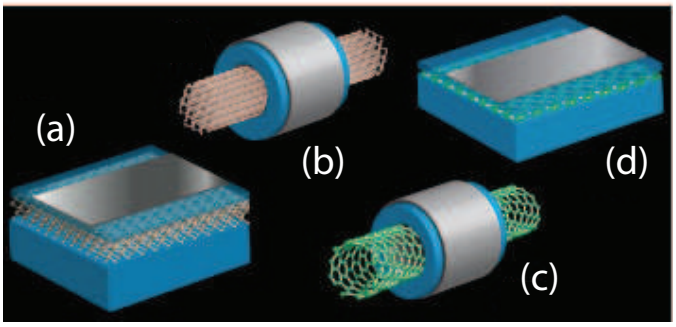

Fig. 1. Example of device structures that OMEN can simulate: (a) Single-Gate and Double-Gate Ultra-Thin-Body field-effect transistor (FET) made of Si, Ge, or III-V semiconductors, (b) Gate-All-Around Nanowire FET, (c) Coaxially-Gated Carbon Nanotube FET, and (d) Graphene Nanoribbon FET.

II. Overview of the Simulation Approach

A. Formulation of the Quantum Transport Problem

OMEN is a multi-dimensional quantum transport simulator based on different flavors of the nearest-neighbor empirical tight-binding model (single s-orbital, single p_z-orbital, sp^3s^*, and $sp^3d^5s^*$). Its algorithms have been described in several publications before [13], [23], [24], [25], [26] and will only be briefly summarized here so that the latest improvements can be more easily tracked and understood. The starting point is the solution of the Schrödinger equation $H(\mathbf{r})\psi(E,\mathbf{r}) = E\psi(E,\mathbf{r})$ with open boundary conditions (OBCs) where the Hamiltonian H describes the interactions between the atoms composing the simulation domain and $\psi(E,\mathbf{r})$ is the electron wave function at position $\mathbf{r}$ and energy E. In the tight-binding formalism, $\psi(E,\mathbf{r})$ is expanded in terms of localized orbital functions $\phi^\sigma(\mathbf{r}-\mathbf{R}_{ijk})$ of type $\sigma = \{s,p,d,s*\}$ on an atom situated at $\mathbf{R}_{ijk}$=$\{x_i,y_j,z_k\}$

$$\psi(E,\mathbf{r}) \quad = \quad \sum_\sigma \sum_{ijk} C^\sigma_{ijk}(E)\phi^\sigma(\mathbf{r}-\mathbf{R}_{ijk}). \qquad (1)$$

The expansion coefficients $C^\sigma_{ijk}(E)$ are the unknown quantities that need to be determined. Note that if one direction, for example z, is assumed periodic, as in 2-D devices, the z dependence is replaced by a phase factor $e^{-ik_z z}$

$$\psi(E,\mathbf{r}) = \sum_\sigma \sum_{ij} \sum_{k_z} C^\sigma_{ij}(E,k_z)\phi^\sigma(\mathbf{r}_{xy}-\mathbf{R}_{ij})e^{-ik_z z}, \quad (2)$$

where k_z is the electron momentum along the z direction. As the devices considered in this work are two-dimensional, the ansatz in Eq. (2) is retained and the $C^\sigma_{ij}(E,k_z)$ are calculated by inserting it into the Schrödinger equation and recalling the nearest-neighbor Slater-Koster table from Ref. [14]

$$(\mathbf{E}-\mathbf{H}(k_z)-\mathbf{V}-\mathbf{\Sigma}(E,k_z))\cdot\mathbf{C}(E,k_z) = \mathbf{S}(E,k_z). \quad (3)$$

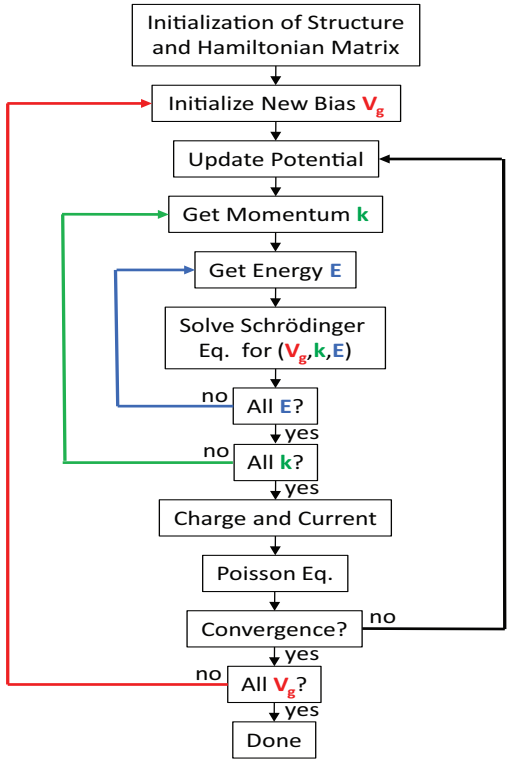

Fig. 2. OMEN simulation flow chart. The loops on the left-hand-side (bias points, momentum, and energy) are parallelized while the loop on the right-hand-side (Poisson) cannot. The floating point performances are measured by inserting the function PAPI_flops before the first task and after the last one.

Algorithm 1 OMEN Algorithm with four levels of parallelization (lines 4, 9, 14, and 16)

```
1:  Construction of the atomistic simulation domain
2:  Generation of the tight-binding Hamiltonian matrix H
3:  Initialization of the FEM Poisson environment
4:  for every bias point V_gs/V_ds (in parallel) do
5:      Get source and drain Fermi levels
6:      Get initial guess for electrostatic potential V
7:      repeat
8:          Update electrostatic potential V → H
9:          for every momentum point k_z (in parallel) do
10:             Update momentum H → H(k_z)
11:             Compute contact bandstructure
12:             Generate energy grid E(k_z)
13:             Balance work load through CPUs
14:             for every energy point E (in parallel) do
15:                 Compute open boundary Σ and S
16:                 Solve (in parallel) Schrödinger Eq.
                    (E − H(k_z) − Σ) · C(E, k_z) = S
17:                 Extract DOS and TE
18:             end for
19:         end for
20:         Compute charge ρ and current J_d
21:         Solve FEM Poisson equation ρ → V
22:     until self-consistent convergence of ρ and V
23: end for
```

Equation (3) is the Wave Function form of the Schrödinger equation. The vector $\mathbf{C}(E,k_z)$ is of size $N_A \times N_{orb}$, where N_A is the number of atoms composing the simulation domain and N_{orb} the number of orbitals describing the bandstructure of each atom. For example, N_{orb}=10 in the sp^3d^5s* tight-binding model without spin orbit coupling, 1 s-orbital, 3 p-orbitals, 1 excited s-orbital called $s*$, and 5 d-orbitals. Each single element $C_{ij}^{\sigma}(E,k_z)$ is stored in $\mathbf{C}(E,k_z)$. The diagonal matrix $\mathbf{E}$ contains the electron/hole energy E, $\mathbf{H}(k_z)$ is the device Hamiltonian matrix expressed in a tight-binding basis. It includes the on-site energy and the coupling elements between different orbitals (σ_1,σ_2) and atoms (n,m), $H_{nm}^{\sigma_1\sigma_2}$, while the entries of the diagonal matrix $\mathbf{V}$ are the electrostatic potential at each atomic site $V(x_i,y_j)$. The open boundary conditions, i. e. the coupling of the simulation domain with its environment, are cast into the matrix $\mathbf{\Sigma}(E,k_z)$ and the vector $\mathbf{S}(E,k_z)$. Both quantities are calculated using an approach developed specifically for OMEN and much more efficient than other techniques [13]. Since $\mathbf{S}(E,k_z)$ contains as many vectors as injection channels, $\mathbf{C}(E,k_z)$ has multiple entries.

At this stage it is important to note that the Schrödinger equation in the Wave Function formalism takes the form of a sparse linear system of equations $Ax=b$ with the matrix A=$(\mathbf{E}-\mathbf{H}(k_z)-\mathbf{V}-\mathbf{\Sigma}(E,k_z))$ and the multiple right-hand side vector b=$\mathbf{S}(E,k_z)$. It can therefore be solved using parallel sparse linear solvers, as explained in Section II-C, which is computationally less expensive than the NEGF formalism

$$(\mathbf{E}-\mathbf{H}(k_z)-\mathbf{V}-\mathbf{\Sigma}(E,k_z))\cdot\mathbf{G}^R(E,k_z)=\mathbf{I}. \quad (4)$$

In effect, in Eq. 4, the retarded Green's Function $\mathbf{G}^R(E,k_z)$ is a full and square matrix, not a vector, with the same size as the Hamiltonian $\mathbf{H}(k_z)$, $N_A \times N_{orb}$, while $\mathbf{I}$ represents the identity matrix. Only the diagonal blocks and the first block column of $\mathbf{G}^R(E,k_z)$ are required for ballistic transport so that partial matrix inversions can be used to treat Eq. (4) [27].

Equation (3) must be solved for each electron/hole energy E and momentum k_z. Each equation can be solved independently from the others since there is no E or k_z coupling between them in the ballistic limit of transport, as assumed all along this paper. Once this is done, the electron (hole) density $n(x,y)$ ($p(x,y)$) and the current density J_d can be calculated using

$$n(x,y) = \sum_p \sum_{k_z} \int \mathrm{d}E\; Z_p(E,k_z,x,y) f(E-\mu_p) \quad (5)$$

$$J_d = \frac{e}{\hbar} \sum_p \sum_{k_z} \int \frac{\mathrm{d}E}{2\pi}\; T_p(E,k_z) f(E-\mu_p), \quad (6)$$

where e is the elementary charge constant, $\hbar$ the reduced Planck's constant, $f(E-\mu_p)$ the Fermi-Dirac distribution at contact p={Source, Drain} characterized by the Fermi level μ_p, $Z_p(E,k_z,\mathbf{r_{xy}})$ and $T_p(E,k_z)$ the density-of-states and electron transmission originating from contact p, respectively and derived from $\mathbf{C}(E,k_z)$ [28]. The hole density is simply obtained by replacing $f(E-\mu_p)$ by $(1-f(E-\mu_p))$ and adjusting the energy range E.

Finally, the solution of Eq. (3) directly depends on the electrostatic potential $V(x,y)$ which forms the diagonal entries of the matrix $\mathbf{V}$ and is related to the charge density $\rho(x,y)$=$p(x,y) - n(x,y)$ through the Poisson equation [29]

$$\Delta V(x,y) = -\frac{e}{\epsilon(x,y)}(p(x,y) - n(x,y) - N_{dop}(x,y)). \quad (7)$$

The doping concentrations are cast into $N_{dop}(x,y)$ and the permittivity of the different materials composing the simulation domain is represented by $\epsilon(x,y)$. The Poisson equation is solved with the finite element method [30] on a typically larger domain than the Schrödinger equation. The Schrödinger and Poisson equations are self-consistently iterated till convergence between $\rho(x,y)$ and $V(x,y)$ is achieved. The number of required iterations is labeled N_{poiss}.

B. Parallelization Scheme

The calculation of the charge and current densities in Eq. (5) and (6) depends on the electron/hole energy E and momentum k_z, but also on the externally applied gate V_{gs} and drain bias V_{ds} through the Poisson equation and the contact Fermi level $\mu_{S/D}$, respectively. The flow chart of OMEN and the basic algorithm are described in Fig. 2 and Algorithm 1.

Three levels of parallelization can be identified (from the highest to the lowest level): (i) the loop over the bias points V_{gs}/V_{ds}, which is embarrassingly parallel and consists of a parameter sweep, (ii) the loop over the momentum k_z, and (iii) the loop over the energy points E. Hence, starting from the total CPU pool, a different sub-group and sub-communicator is created for each bias point, then at the next level for each momentum, and finally, at the lowest level for each energy. Communication between the bias point sub-groups is not required, but a collective reduce operation is needed within each of them to calculate the charge and current densities in Eq. (5) and (6) (summation over k_z and integration over E).

While the number of bias points N_V and momentum N_{k_z} are input parameters, the number of energies N_E depends on the device characteristics and is different for each momentum, i. e. N_E=$N_E(k_z)$. Since $N_E(k_z)$ is not known "a priori", the same number of CPUs is assigned to each momentum sub-group. From there, an energy grid $E(k_z)$ based on the bandstructure of the device contacts and the position of the Fermi levels is generated. Then the total number of CPUs assigned to each bias point is redistributed according to $N_E(k_z)$ to optimally balance the work load among the different momentum sub-groups.

As compared to Ref. [24], the parallelization of OMEN has not evolved much. The generation of the energy vector has been accelerated by computing the bandstructure of the source and drain contacts in parallel instead of one after the other. The load balance scheme has been slightly improved by changing the criterion to redistribute the CPUs among the different momentum sub-groups, the goal being that each CPU treats more or less the same number of energy points. Finally, the construction of the atomistic simulation domain has been better parallelized by removing unnecessary and expensive communication operations.

C. Block Cyclic Reduction (BCR) of the Schrödinger Equation

OMEN contains a fourth level of parallelization, a 1-D spatial domain decomposition that is applied to the Schrödinger equation. Instead of being solved on a single CPU, Eq. (3), in its reduced form $Ax = b$, can be solved using a parallel sparse linear solver such as SuperLU_dist [31] or MUMPS [32]. Hence, if N_{CPU} cores are assigned to the solution of Eq. (3), each of them stores only the part of the matrix A it is dealing with, considerably reducing the memory consumption. The first CPU treats the n first lines of the matrix A, the second CPU the next n lines and so on, which corresponds to a 1-D spatial domain decomposition along the electron transport direction. Note that direct solvers are preferable to iterative ones here due to the multiple right-hand-side vector b.

Although the available parallel solver libraries greatly facilitate the solution of Eq. (3), they are intended to be as general as possible and cannot really benefit from the structure of the matrix A in order to reduce the computation time. In the nearest-neighbor tight-binding model, the matrix A possesses four important properties:

- A is block tri-diagonal and each block represents an atomic layer, i. e. each block contains all the atoms with the same coordinate along the electron transport direction x.
- the diagonal blocks of A contain the orbital on-site energies and are quasi-diagonal if the electrons flow along the <100> crystal axis as here, except the first and the last one.
- the off-diagonal blocks $A_{ii\pm1}$ connect one atomic layer to its previous and next neighbor, they are sparse, and A_{ii+1}=$A^{\dagger}_{i+1i}$
- the open boundary conditions, the matrix $\mathbf{\Sigma}(E,k_z)$ and the vector $\mathbf{S}(E,k_z)$, alter only the first and the last block of A.

Hence, Eq. (3) can be rewritten as

$$\underbrace{(\mathbf{E}-\mathbf{H}_{ii})}_{A_{ii}}\cdot\mathbf{C}_i \underbrace{-\mathbf{H}_{ii+1}}_{+A_{ii+1}}\cdot\mathbf{C}_{i+1} \underbrace{-\mathbf{H}_{ii-1}}_{+A_{ii-1}}\cdot\mathbf{C}_{i-1} = 0 \quad (8)$$

for each atomic layer i, except for the first, labeled "1", and the last, labeled "N", where it becomes

$$\underbrace{(\mathbf{E}-\mathbf{H}_{11}-\mathbf{\Sigma}_{11})}_{A_{11}}\cdot\mathbf{C}_1 - \mathbf{H}_{12}\cdot\mathbf{C}_2 = \mathbf{S}_1, \quad (9)$$

$$\underbrace{(\mathbf{E}-\mathbf{H}_{NN}-\mathbf{\Sigma}_{NN})}_{A_{NN}}\cdot\mathbf{C}_N - \mathbf{H}_{NN-1}\cdot\mathbf{C}_{N-1} = \mathbf{S}_N. \quad (10)$$

The matrices in Eq. (8) to (10) are of size $N_a \times N_{orb}$ where N_a is the number of atoms per atomic layer. The E, k_z, and σ indices are dropped for clarity. Based on this representation of the Schrödinger equation, it has been shown in Ref. [25], [26] that a parallel block cyclic reduction (BCR) [33] (also known in the physics community as "renormalization" [34]) of the matrix A gives better performances than SuperLU_dist and MUMPS while solving Eq. (3). The idea is to use Gaussian elimination to remove atomic layers until the first atomic layer

is only connected to the last one. For example, removing the atomic layer i "renormalizes" the blocks A_{i-1i-1} and A_{i+1i+1} as well as A_{i-1i+1} following

$$\begin{aligned} X_i &= A_{ii}^{-1} \cdot A_{ii-1} && (11)\\ Y_i &= A_{ii}^{-1} \cdot A_{ii+1} && (12)\\ \hat{A}_{i-1i-1} &= A_{i-1i-1} - A_{i-1i} \cdot X_i && (13)\\ \hat{A}_{i+1i+1} &= A_{i+1i+1} - A_{i+1i} \cdot Y_i && (14)\\ \hat{A}_{i-1i+1} &= -A_{i-1i} \cdot Y_i. && (15) \end{aligned}$$

These identities involve one matrix inversion and five matrix multiplications and lead to the suppression of the atomic layer i from the matrix A. To take full advantage of the quasi-diagonal structure of the A_{ii} blocks and the sparsity of the $A_{ii\pm1}$, it is important to carefully choose the removal orders. Hence, three renormalization (or cyclic reduction) stages are performed

1) all the blocks with an even index i are removed. This requires the inversion of a quasi-diagonal matrix (analytic) and the multiplication of diagonal and sparse matrices.
2) half of the blocks with an odd index i (3, 7, 11, $\cdots$) are removed. A sparse matrix must be inverted and the multiplication of sparse and full matrices must be performed.
3) the remaining half blocks with an odd index i are removed. All the involved matrices are full. The elimination process is repeated till only $\hat{A}_{11}$, $\hat{A}_{NN}$, $\hat{A}_{1N}$, and $\hat{A}_{N1}$ are different from 0.

After the last renormalization step, the 2 block $\times$ 2 block remaining system of equations is solved, b_1 and b_N are computed, and the rest of the solution vector b=$\mathbf{C}(E, k_z)$ is constructed according to the recursion

$$b_i = -X_i \cdot b_{i-p} - Y_i \cdot b_{i+q}, \quad (16)$$

where the atomic layer i was connected to the layers $i-p$ and $i+q$ before being removed.

A 1-D domain decomposition can also be applied to the BCR algorithm. Each CPU stores only one part of the matrix A, removes all the local atomic layers till only its first and last blocks are connected, and finally exchanges information with the neighbor CPUs to remove the remaining layers.

Apart from its good scalability, the main advantage of the BCR method over SuperLU_dist and MUMPS is that it allows for the introduction of computational interleaving. To solve $Ax = b$, a linear solver requires the entire matrix A and therefore the open boundary condition matrix $\mathbf{\Sigma}$ which modifies its first and last blocks. In the BCR approach, the first and last blocks are removed at the end so that the computation of the OBCs can be done at the same time as the renormalization of A. This is of great interest since the calculation of the OBCs is expensive and cannot be efficiently parallelized beyond two CPUs. Hence, if N_{CPU} cores have to solve Eq. (3) in parallel with SuperLU_dist or MUMPS, N_{CPU}-2 CPUs first remain idle while the OBCs are computed while these N_{CPU}-2 CPUs can start renormalizing A right from the beginning with the BCR algorithm, improving both the time-to-solution and the performances.

The computational interleaving of the OBCs and the BCR method were introduced in Ref. [24]. To improve their efficiency, the number of CPUs that deal with the OBCs can now be either 1 (a single CPU treats both the source and drain contacts) or 2 as before (1 CPU for the source and 1 CPU for the drain). This change is driven by the fact that for long 2-D devices, the time to compute the OBCs is much faster than the time to renormalize the matrix A and a better balance is obtained if one additional CPU works on reducing A and one less on the OBCs. Another improvement has been made to the BCR algorithm by overlapping computation and communication where possible.

III. Code Implementation

A. General Overview

The OMEN code is written in C++ and implements the quantum transport approach described in Section II. The purpose of using an object-oriented language is to make the solution of the Schrödinger equation as transparent as possible to different solvers and to the choice of real or complex arithmetic. In effect, to be as general as possible, OMEN takes only the Hamiltonian matrix $\mathbf{H}$ as an input parameter and then solves the Schrödinger equation either in the NEGF or in the WF formalism with MUMPS, SuperLU_dist, our BCR algorithm, or any other direct sparse linear solver. This is achieved by class inheritance from a general class called "Solver". More than 50% of OMEN is based on template classes so that the Schödinger equation can either be solved in real arithmetic as for 3-D devices without spin-orbit (SO) coupling or in complex arithmetic as in 1-D and 2-D structures or in the presence of SO.

OMEN has four natural levels of parallelism as shown in Fig. 2 and in Algorithm 1 (bias points, momentum, energy, and 1-D spatial domain decomposition), all using MPI [35] and a hierarchical organization of communicators. Starting from the global communicator MPI_COMM_WORLD, each parallelization level is characterized by sub-communicators derived from the higher level root communicator. The command MPI_Comm_split or the combination MPI_Group_incl/MPI_Comm_create are employed to create all the sub-communicators. The communicators at the bias point and energy level are generated at the beginning of each simulation while new communicators at the momentum level are created at the beginning of each Poisson iteration to optimally balance the work load. This means that the number of CPUs working on each bias point as well as solving the Schrödinger equation at a given E and k_z are fixed and input parameters, but the number of CPUs associated to each momentum is a function of the work load. The most recent improvement made to the parallelization of OMEN concerns the creation of the sub-communicators, each CPU creating only the communicator it belongs to and not all the communicators existing at a given parallelization level.

All the algebraic operations including full matrices rely on the high-performance BLAS/LAPACK libraries [36], [37],

especially zgemm, zgetrf, zgetrs, and zaxpy. We have developed our own sparse matrix multiplication routines based on a compressed sparse row (CSR) and compressed sparse column (CSC) format. However, when a sparse matrix multiplies a full matrix, the routine zaxpy from BLAS is used to operate on a matrix row or column instead of single elements. The inversion of sparse matrices, as required by the BCR algorithm, is done with the Umfpack library [38]. Finally, the sparse linear system of equations resulting from the application of the Newton-Raphson scheme to the Poisson equation is solved in parallel with the Aztec library [39].

B. Mixed Precision Scheme

Most of the simulation time in OMEN is consumed by the solution of the Schrödinger equation, line 16 in Algorithm 1. More precisely, the third (and last) renormalization stage of the BCR algorithm described in Section II, which involves only one quarter of the blocks (or atomic layers) composing the matrix A, takes about 70% of the total time, irrespective of the number of CPUs solving the system $Ax = b$, because all the matrices are full. To reduce the computational time, the full matrix multiplications in Eq. (11) to (15) could be executed in single instead of double precision, i. e. by using cgemm instead of zgemm. However, this approach does not offer the required accuracy and cannot be used in that form.

It turns out that Eq. (11) and (12) must always be computed in double precision, otherwise rounding errors are propagated through the repeated usage of X_i and Y_i, both in Eq. (13) to (15) as well as in the recursion (16). This means that only Eq. (13) to (15) can be computed using cgemm .

There is a second constrain about the utilization of single precision matrix multiplication: as mentioned earlier, the Schrödinger and Poisson equations must be self-consistently solved N_{poiss} times before the charge density $\rho(x, y)$ and the electrostatic potential $V(x, y)$ converge to their final value. Single precision operations can only be used for the N_{poiss}-1 first iterations, in the best case, and double precision matrix multiplications must still be used for at least one last iteration to obtain accurate results. The difficulty is to determine the right moment to switch from single to double precision. This can be done either by defining a criterion related to the convergence of $\rho(x, y)$ and $V(x, y)$ or by setting N_{poiss} to a fixed, large enough value so that converged results are obtained. The latter approach has been chosen in this paper.

IV. Simulation Results: HEMT and TFET

A. Structure Definition

To demonstrate the performances of OMEN, two realistic 2-D nanoelectronic devices are simulated, a single-gate III-V high electron mobility transistor (HEMT) [6] and a double-gate band-to-band tunneling field-effect transistor (TFET) [40]. Both transistor structures are schematized in Fig. 3. Currently, many experimental groups are trying to fabricate nanoscale III-V HEMTs and TFETs and compare their performances to the standard Si MOSFET. What is expected is mainly a reduction of the IC power consumption. Being able to simulate such devices could accelerate the innovation of

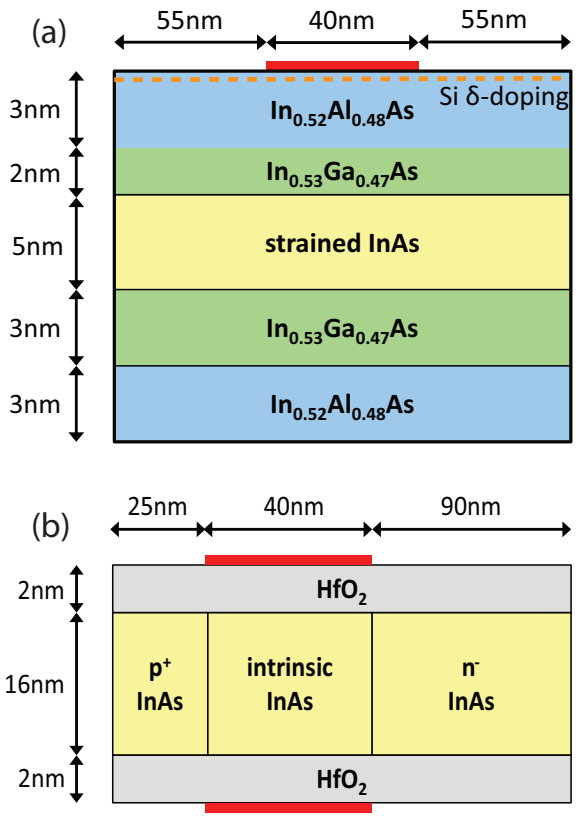

Fig. 3. Schematic view of the devices considered in this paper. (a) A single-gate, multi quantum well high electron mobility transistor (HEMT). The channel is composed of a strained InAs layer and two $In_{0.53}Ga_{0.47}As$ layers. The two surrounding $In_{0.52}Al_{0.48}As$ layers act as insulator. Doping is realized through a Si delta-doped layer with a donor concentration N_D=3e12 cm^{-2}. (b) A double-gate band-to-band tunneling transistor (TFET). A pure InAs channel is surrounded by two HfO_2 oxide layers. The source is p-doped with an acceptor concentration N_A=4e19 cm^{-3} and the drain n-doped with a donor concentration N_D=4e18 cm^{-3}.

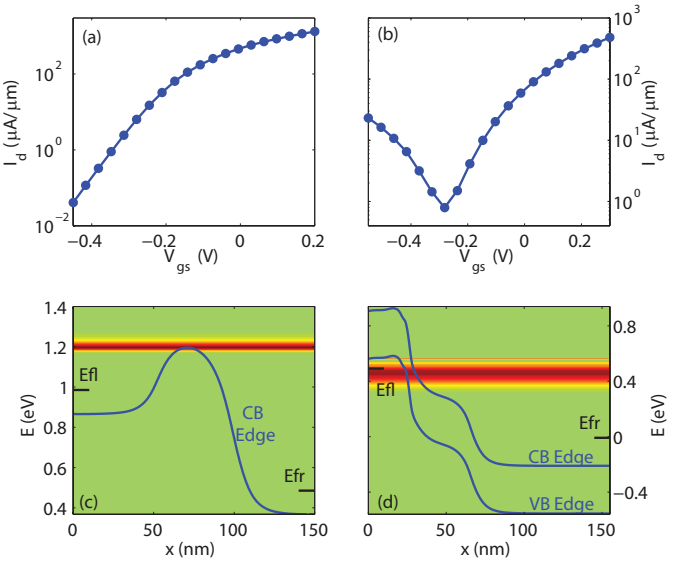

Fig. 4. Transfer characteristics I_d-V_{gs} at V_{ds}=0.5 V (upper subplots) and spectral current density (lower subplots) of the HEMT (a and c) and TFET (b and d) structures schematized in Fig. 3. In subplots (c) and (d), red indicates high current concentrations while green means "no current". The position of the conduction (CB) and valence (VB) band edges are given as well as the left (Efl) and right (Efr) contact Fermi levels.

these new technologies. Some technical details about them are summarized below:

- III-V HEMT: the active region of this transistor is composed of a 5nm biaxially stressed InAs quantum well surrounded by a 2nm (top) and a 3nm (down) $In_{0.53}Ga_{0.47}As$ quantum well. The insulator layer separating the gate contact (length L_g=40 nm) from the channel as well as the bottom substrate layer are made of $In_{0.52}Al_{0.48}As$. Doping is realized though a Si delta-doped layer of concentration N_D=3e12 cm^{-2}. The structure measures 150nm in length and contains N_A=55,226 atoms with N_{orb}=10 orbitals per atom ($sp^3d^5s^*$ tight-binding model without spin-orbit coupling) so that the size of the Hamiltonian matrix in Eq. (3) N=$N_A \times N_{orb}$=552,260. The tight-binding parameters are taken from Ref. [41]. A total of N_V=20 bias points are simulated with N_{k_z}=30 momentum per bias and a range of $N_E(k_z)$=[420,1470] energy points per momentum.
- InAs TFET: the second device is made of a single material, a 16nm relaxed InAs quantum well. The source and drain extensions are asymmetric in size and doping. The L_s=25nm long source is p-doped with an acceptor concentration N_A=4e19 cm^{-3} while the drain, of length L_d=90nm, is n-type with a donor concentration N_D=4e18 cm^{-3}. The double-gate contact measures L_g=40nm. The InAs TFET is composed of N_A=54,272 atoms, the number of orbitals per atom N_{orb}=10, as for the III-V HEMT, but the bandstructure model is different. To ensure a good representation of both the conduction and valence bands of InAs, as required in TFETs, the nearest-neighbor sp^3s^* tight-binding model with spin-orbit coupling is used with the parameters from Ref. [42]. The number of bias points, N_V=20, and momentum, N_{k_z}=30, are the same as for the HEMT. However, the number of energies $N_E(k_z)$ ranges from 160 to 1800.

Equation (3) must be solved $N_V \times N_{k_z} \times N_{E,mean} \times N_{poiss}$ times to simulate the HEMT and TFET, where $N_{E,mean}$ is the average number of energy points per momentum. This represents more than 1 million solutions of the Schrödinger equation per device simulation.

B. Device Simulation

The transfer characteristics (current as function of the applied gate bias V_{gs}) at a drain-to-source voltage V_{ds}=0.5 V and the energy-resolved spectral current of the HEMT and TFET are plotted in Fig. 4. In the lower subplots, the red regions indicate high current concentrations, green means no current. It can be seen that the HEMT and TFET exhibit significantly different current characteristics, both in terms of magnitude and overall shape. These trends are consistent with available experimental data [6], [9].

The fundamental difference between the two devices is illustrated in Fig. 4 (c) and (d). "Hot" electrons flowing on the top of a potential barrier carry the current through the HEMT while "cold" electrons tunneling through a potential barrier are responsible for the current in the TFET. Hence, in the HEMT case, everything happens within the conduction band of the channel, but in the TFET case, electrons located in the valence band of the source contact move into the conduction band of the drain contact. A proper treatment of TFETs requires therefore the inclusion of electrons, holes, and quantum mechanical tunneling, as implemented in OMEN.

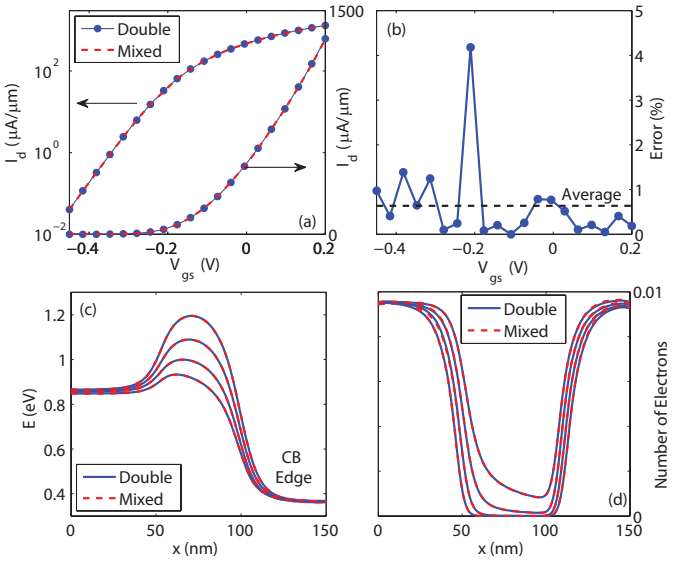

Fig. 5. Comparison of the HEMT simulation results obtained in double precision (blue lines with circles) and in mixed precision (dashed red lines). (a) Transfer characteristics I_d-V_{gs} at V_{ds}=0.5 V on a log and linear scale. (b) Percentage error of the mixed precision scheme for the current I_d as function of V_{gs} as well as the average error. (c) Conduction band edge at different V_{gs} ranging from -0.45 to -0.04 V. (d) Electron density along the HEMT channel at the same V_{gs} as in (c).

C. Accuracy of the Mixed Precision Scheme

Asserting the accuracy of the mixed precision scheme introduced in Section III-B is an important issue. It is only meaningful to accelerate the time-to-solution by sacrificing a part of the accuracy if the approximate results are close enough to the exact ones. Many tests have been conducted so far for many different structures and the main conclusions for the HEMT and TFET are that (i) the mixed precision scheme does not work well for TFETs because the hole bandstructure is too complex with multiple coupled bands in the same energy region to support rounding approximations, whereas the electron bandstructure is almost parabolic with one fundamental band, (ii) the average error on the HEMT current, which is the quantity of interest, is less than 1% as compared to the double precision scheme, and (iii) maximum errors up to 4% have been observed in the HEMT current. A comparison between the double (solid blue lines) and mixed (dashed red lines) precision schemes is given in Fig. 5 for the current defined in Eq. (6), the conduction band edge resulting from Eq. (7), and the charge density in Eq. (5). The relative error between the current calculated in mixed and double precision is also reported in subplot (b).

For engineering applications, an average error of less than 1% is acceptable, but we may need to find a way to improve the bias points for which the error is around 4%. Equation (3) must be solved almost 100,000 times per bias point,

	HEMT (TFlop/s)		TFET (TFlop/s)
num. of cores	double	mixed	double
2700	16.20	18.39	16.09
5535	32.79	37.34	32.94
11070	65.52	73.84	64.93
44280	260.99	295.93	259.75
110700	645.65	735.99	645.54
221400	1268.03	1439.27	1276.25

TABLE I
SUSTAINED PERFORMANCE OF OMEN ON THE CRAY-XT5 JAGUAR AT ORNL FOR THE SIMULATION OF A HEMT AND TFET DEVICE IN DOUBLE AND MIXED (ONLY FOR THE HEMT) PRECISION. THE THEORETICAL PEAK PERFORMANCE OF JAGUAR IS 2.3 PFLOP/S ON 221400 CORES.

over 1 million overall. However, the application of the mixed precision scheme is problematic for very few energy and momentum configurations, usually less than 100. The goal is therefore to find a proper criterion to identify these badly conditioned cases and to eliminate them without increasing the computational burden. A criterion based on the calculation of the residual $r{=}Ax{-}b$ for certain entries has been tested without success. Other approaches are currently under investigation.

The mixed precision scheme could be used for other applications where the accuracy is not critical. For example, device simulations including electron-phonon scattering are computationally very intensive [43], but a good initial guess of the electrostatic potential can considerably reduce the simulation time. By starting with the electrostatic potential resulting from a ballistic (coherent) simulation, the number of required self-consistent Schrödinger-Poisson iterations can be reduced by around 50% when electron-phonon scattering is turned-on. The electrostatic potential resulting from a mixed-precision ballistic run works very well as an initial guess for a device simulation with incoherent electron transport.

V. CODE PERFORMANCES

A. Simulation Configuration

The HEMT and TFET described in Section IV were simulated on the CRAY-XT5 Jaguar at Oak Ridge National Laboratory [44] on 2,700 up to 221,400 cores. The simulation settings are identical in all the computational experiments, which means that the same number of bias points (20), momentum (30), and energy (from 430 to 1470 per momentum for the HEMT and from 160 to 1800 per momentum for the TFET) are used in all cases. For convenience, the self-consistent calculation of the Schrödinger and Poisson equations is stopped after N_{poiss}=5 iterations for the HEMT and N_{poiss}=4 iterations for the TFET. When the HEMT is simulated in mixed precision, the 4 first Schrödinger-Poisson iterations are solved in single precision (only full matrix multiplications in the last stage of the BCR algorithm as explained earlier), the last iteration in double precision. This configuration has been used to produce the comparison between the mixed and double precision scheme in Fig. 5.

For all the results reported here, the four levels of parallelization of OMEN are turned on and organized as follows

1) the maximum number of CPUs that can work on a single bias point is set to 11,070. If the total number of CPUs is larger than 11,070, for example 44,280, then four bias points are treated in parallel. If the total number of CPUs is smaller than 11,070, then they will all be assigned to the same bias point and compute one bias after the other. Another possibility, not shown here, would be to always treat all the bias points in parallel, but this would not lead to different conclusions.
2) at the next lower parallelization level, all the momentum points are computed simultaneously. The number of CPUs attributed to each momentum sub-group is determined by the load balance scheme.
3) at the energy parallelization level, each CPU treats more than a single energy E, here 18-21 when 11,070 cores are assigned to each bias point, but all belonging to the same momentum k_z.
4) finally, at the lowest parallelization level, the Schrödinger equation Eq. (3) is solved on 9 cores using our BCR algorithm and computational interleaving. One CPU computes only the source and drain open boundary conditions, the 8 others renormalize and reduce the matrix A in the $Ax = b$ system of equations.

The time-to-solution, the sustained performance, and the number of floating point instructions as function of the number of CPUs are measured using the PAPI_flops function v3.6.2 [45]. The counters are initialized before line 1 in Algorithm 1 and read after line 23 so that the entire simulation flow is taken into account. As a sanity check, it has been verified that PAPI_flops returns the same results as PAPI_start_counters put before line 1 in Algorithm 1 combined with PAPI_stop_counters set after line 23.

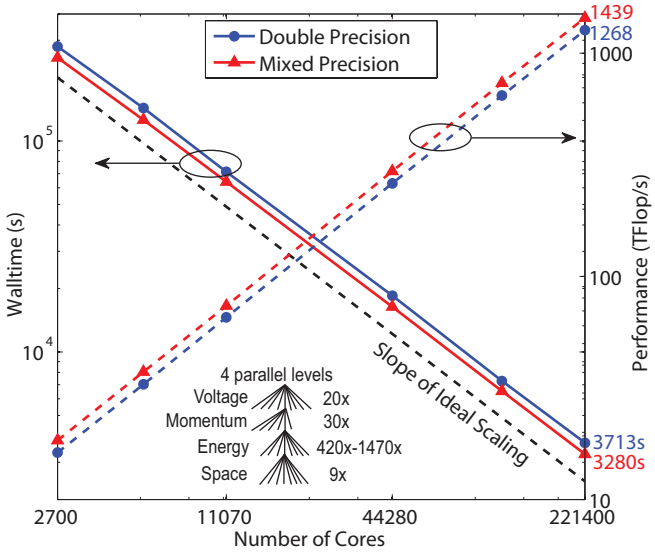

Fig. 6. Walltime and sustained performance of OMEN up to 221,400 cores for the simulation of the HEMT structure in Fig. 3 on the Cray-XT5 Jaguar at ORNL. The blue curves refer to double precision, the red curves to mixed precision calculations. A total of 20 bias points, 30 momentum, from 420 to 1470 energy points per momentum, and a 1-D spatial domain decomposition on 9 cores are used. For each bias points, the number of Poisson iterations is limited to 5. Since the computation of the bias points is embarrassingly parallel and the execution time on Jaguar limited, only 1 bias point is computed on 2,700 to 11,070 cores and the simulation time is multiplied by 20. It has been verified that this simplification does not affect the scaling behavior of OMEN.

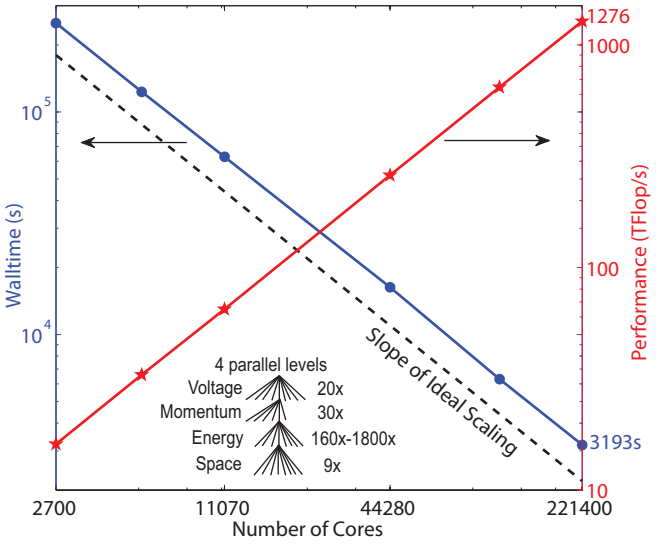

Fig. 7. Same as in Fig. 6, but for the TFET structure in Fig. 3 and in double precision only. The number of bias points and momentum is the same as for the HEMT structure, but the number of energies per momentum is more scattered and goes from 160 to 1800. Here, only 4 Poisson iterations per bias point are considered since the convergence of the Poisson equation is faster for TFETs than for HEMTs.

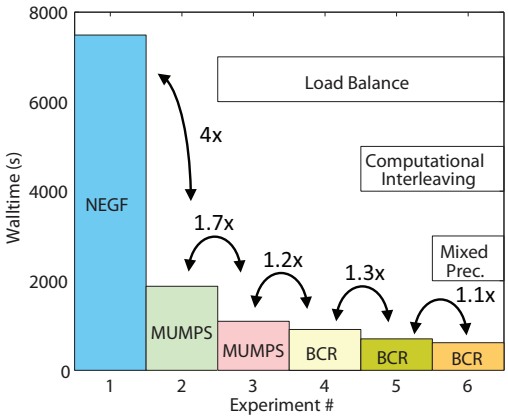

Fig. 8. Evolution of the HEMT simulation time as function of the numerical methods and parallel algorithms that are used. The time to solve the Schrödinger equation for all the momentum k_z and energies E and the Poisson equation at one given bias, on 11,070 cores, is reported. Three solvers are compared (NEGF, MUMPS, and BCR) and combined with three schemes (load balance, computational interleaving, and mixed precision). The speed-up from one numerical experiment to the other is also reported. Note that the time for the first experiment using NEGF is just a "best-case" estimation based on the parallel performances presented in Ref. [46] since OMEN does not provide any parallelization of the NEGF formalism.

B. Sustained Performances on Jaguar

Figures 6 and 7 report the time-to-solution and the sustained performance of OMEN for the simulation of the realistic HEMT and TFET described above. The sustained performances are also summarized in Table I. These are strong scaling experiments since the amount of work to perform remains the same, but the number of CPUs increases.

As significant achievement, it can be mentioned that (i) OMEN reaches a *sustained* performance of 1.44 PFlop/s during about *1 hour* on 221,400 cores for the HEMT structure simulated in mixed precision, (ii) it reaches a *sustained* performance of 1.27 and 1.28 PFlop/s in double precision for the HEMT and TFET, respectively (55% of the peak performance), (iii) the time-to-solution scales almost perfectly from 2,700 to 221,400 cores. In fact, the scaling can be expected to remain as good as shown here down to 9 cores, the number of CPUs required to solve Eq. (3). Then, the increase of the simulation time will slightly saturate since the time to solve the Schödinger equation on 9 cores is "only" 7.9 times faster than on a single core. Hence, a transfer characteristics that can be computed in less than 1 hour on 221,400 cores would last more than 20 years on a single CPU, more realistically around 9 days on a small cluster with 1,000 CPUs.

C. Evolution of the Simulation Approach

OMEN offers unprecedented simulation capabilities to investigate nanoelectronic devices. Many factors have contributed to the improvement of its performances from 190 TFlop/s on 60k cores in 2008 up to 1.44 PFlop/s on 221,400 cores today. Better hardware (faster CPUs), better software (compiler, high performance libraries), and above all better algorithms have enabled the simulation of larger device structures in a shorter amount of time.

Figure 8 shows the evolution of the simulation time for the HEMT structure for one self-consistent Schrödinger-Poisson iteration on 11,070 cores as function of the introduction of new algorithms. The real breakthrough is the replacement of the NEGF formalism, still used by more than 90% of the researchers in the field of quantum transport modeling, even in the ballistic limit of transport, by the more efficient Wave Function approach (speed up of 4×). This explains why OMEN is more powerful than probably any other quantum transport simulator. At the beginning, the parallel library MUMPS was used to solve the Schrödinger equation, but it was found that the BCR algorithm of Section II is slightly faster in 2-D structures although it was originally developed for 3-D nanowires and requires 60% more floating point intructions (speed up of 1.2×). This naturally leads to better sustained performances. Another advantage of the BCR method over MUMPS is that it allows for a computational interleaving of the open boundary calculations and the solution of the Schödinger equation, further reducing the simulation time (1.3×). Other performance boosters include the load balance (1.7×) and the mixed precision schemes (1.1×).

The combination of all these algorithmic innovations have helped reduce the time-to-solution by more than one order of magnitude (12×), as compared to the simulation approach using the NEGF formalism and no parallelization tricks, without increasing the number of CPUs.

VI. Conclusion

We have improved the numerical algorithms and parallel scheme of our multi-dimensional atomistic quantum transport simulator OMEN to reach *sustained* petascale performances on 221,400 cores of the CRAY-XT5 Jaguar at ORNL. Using two very different realistic nanoelectronic devices, a high electron mobility transistor and a band-to-band tunneling field-effect

transistor, it was demonstrated that the time-to-simulation scales almost perfectly from 2,700 up to 221,400 cores and that a sustained performance up to 1.44 PFlop/s can be reached.

These achievements open new perspectives for the computer aided design of nanoscale transistors. More efficient collaborations with experimental groups will be possible due to the recent progresses made to OMEN. Experimentalists are usually disappointed by the time that device simulations take, but in the future, based on the availability of more and more petascale machines, it will be possible to simulate new transistor designs within a couple of hours only, including the time spent in the waiting queue. We are already offering a simplified version of OMEN with reduced simulation capabilities through the web platform *nanohub.org* [47], and expect to extend it soon to serve more than the current $\sim$500 users.

ACKNOWLEDGEMENT

This work was partially supported by NSF grant EEC-0228390 that funds the Network for Computational Nanotechnology, by NSF PetaApps grant number 0749140, by the Nanoelectronics Research Initiative through the Midwest Institute for Nanoelectronics Discovery, and by NSF through TeraGrid resources provided by the National Institute for Computational Sciences (NICS). This research also used resources of the National Center for Computational Sciences (NCCS) at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits", Electronics **38**, 114-117 (1965).

[2] P. Packan et al., 'High Performance 32nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors", Proc. of the International Electron Devices Meeting (IEDM) 2009, paper 28.4 (2009).

[3] W. Haensch et al., "Silicon CMOS devices beyond scaling", IBM J. Res. & Dev. **50**, 339-361 (2006).

[4] B. Doris et al., "Extreme scaling with ultra-thin Si channel MOSFETs", IEDM Tech. Dig. **2002**, 267-270 (2002).

[5] S. D. Suk et al., "Investigation of nanowire size dependency on TSNWFET", IEDM Tech. Dig. **2007**, 891-894 (2007).

[6] D. H. Kim and J. A. del Alamo, "30-nm InAs Pseudomorphic HEMTs on an InP Substrate With a Current-Gain Cutoff Frequency of 628 GHz", IEEE Elec. Dev.Lett. **29**, 830-833 (2008).

[7] Y. Q. Wu, W. K. Wang, O. Koybasi, D. N. Zakharov, E. A. Stach, S. Nakahara, J. C. M. Hwang and P. D. Ye, "0.8-V Supply Voltage Deep-Submicrometer Inversion-Mode In0.75Ga0.25As MOSFET", IEEE Elec. Dev. Lett. **30**, 700-702 (2009).

[8] X. Wang, Y. Ouyang, X. Li, H. Wang, J. Guo, H. Dai, "Room Temperature All Semiconducting sub-10nm Graphene Nanoribbon Field-Effect Transistors", Phys. Rev. Lett. **100**, 206803 (2008).

[9] J. Appenzeller, Y.-M. Lin, J. Knoch, and P. Avouris, "Band-to-band tunneling in carbon nanotube field-effect transistors," Phys. Rev. Lett. **93**, 196805 (2004).

[10] W. Fichtner, D. J. Rose, and R. E. Blank, "Semiconductor device simulation", IEEE Trans. on Elec. Dev. **30**, pp. 1018-1030 (1983).

[11] C. S. Rafferty, M. R. Pinto, and R. W. Dutton, "Iterative methods in semiconductor device simulation", IEEE Trans. on Elec. Devices **32**, pp. 2018-2027 (1985).

[12] S. Selberherr, A. Schutz, and H. W. Potzl, "MINIMOS-A two-dimensional MOS transistor analyzer", IEEE Trans. on Elec. Devices **27**, pp. 1540-1550 (1980).

[13] M. Luisier, G. Klimeck, A. Schenk, and W. Fichtner, "Atomistic Simulation of Nanowires in the $sp^3d^5s^*$ Tight-Binding Formalism: from Boundary Conditions to Strain Calculations, Phys. Rev. B, **74**, 205323 (2006).

[14] J. C. Slater and G. F. Koster, "Simplified LCAO Method for the Periodic Potential Problem", Phys. Rev. **94**, 1498-1524 (1954).

[15] W. Kohn, L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects", Physical Review **140**, A1133-A1138 (1965).

[16] M. Luisier, A. Schenk, and W. Fichtner, "Three-Dimensional Full-Band Simulations of Si Nanowire Transistors", Proc. of International Electron Devices Meeting (IEDM) 2006, 811 (2006).

[17] M. Luisier and G. Klimeck, "Full-band and atomistic simulation of n- and p-doped double-gate MOSFETs for the 22nm technology node", Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD) Hakone, Japan (2008).

[18] M. Luisier and G. Klimeck, "Performance analysis of statistical samples of graphene nanoribbon tunneling transistors with line edge roughness", App. Phys. Lett. **94**, 223505 (2009).

[19] N. Kharche, G. Klimeck, D.-H. Kim, J. A. del Alamo, and M. Luisier, "Performance Analysis of Ultra-Scaled InAs HEMTs", Proc. of the International Electron Devices Meeting (IEDM) 2009, pp. 456-459 (2009).

[20] P. M. Solomon, I. Lauer, A. Majumdar, J. T. Teherani, M. Luisier, J. Cai, and S. J. Koester, "Effect of Uniaxial Strain on the Drain Current of a Heterojunction Tunneling Field-Effect Transistor", IEEE Elec. Dev. Lett **32**, 464 (2011).

[21] M. Luisier and G. Klimeck, "Atomistic full-band simulations of silicon nanowire transistors: Effects of electron-phonon scattering", Phys. Rev. B **80**, 155430 (2009).

[22] S. Datta, "Electronic Transport in Mesoscopic Systems", Cambridge University Press (1995).

[23] M. Luisier and G. Klimeck, "A multi-level parallel simulation approach to electron transport in nano-scale transistors", Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, article 12 (2008).

[24] M. Luisier and G. Klimeck, "Numerical strategies towards peta-scale simulations of nanoelectronics devices", Parallel Computing **36**, 117-128 (2010).

[25] T. B. Boykin, M. Luisier, and G. Klimeck, "Multiband transmission calculations for nanowires using an optimized renormalization method", Phys. Rev. B **77**, 165318 (2008).

[26] M. Luisier, A. Schenk, W. Fichtner, T. B. Boykin, and G. Klimeck, "A parallel sparse linear solver for nearest-neighbor tight-binding problems", Proc. of the 14th international Euro-Par conference on Parallel Processing, 790-800 (2008).

[27] A. Svizhenko, M. P.Anantram, T. R. Govindan, R. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors", J. Appl. Phys. **91**, 2343-2354 (2002).

[28] M. Luisier and A. Schenk, "Atomistic Simulation of Nanowire Transistors", J. of Computational and Theoretical Nanoscience **5**, pp. 1031-1045 (2008).

[29] R. E. Bank, D. J. Rose, and W. Fichtner, "Numerical Methods for Semiconductor Device Simulation", IEEE Trans. Electron Dev. **30**, 1031 (1983).

[30] P. M. Gresho and R. L. Sani, "Incompressible Flow and the Finite Element Method: Isothermal Laminar Flow", John Wiley and Sons, New York (2000).

[31] X. S. Li and J. W. Demmel "SuperLU_DIST: A Scalable Distributed Memory Sparse Direct Solver for Unsymmetric Linear Systems", ACM Trans. on Math. Software **29**, 110 (2003).

[32] P. R. Amestoy, I. S. Duff, and J.-Y. L'Excellent, "Multifrontal parallel distributed symmetric and unsymmetric solvers" Comput. Methods in Appl. Mech. Eng. **184**, 501 (2000).

[33] R. A. Sweet, "A cyclic reduction algorithm for solving block tridiagonal systems of arbitrary dimension", SIAM J. Numer. Anal. **14**, 707 (1977).

[34] G. Grosso, S. Moroni, and G. P. Parravicini, "Electronic structure of the InAs-GaSb superlattice studied by the renormalization method", Phys. Rev. B **40**, 12328 (1989).

[35] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the MPI message passing interface standard", Parallel Computing **22**, 789 (1996).

[36] J. Dongarra, "Basic Linear Algebra Subprograms Technical Forum Standard", International Journal of High Performance Applications and Supercomputing, **16**, 1–111 (2002).

[37] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, "LAPACK User's Guide", Third Edition, SIAM, Philadelphia (1999).

[38] T. A. Davis, "A column pre-ordering strategy for the unsymmetric-pattern multifrontal method", ACM Trans. on Math. Software **30**, 165 (2004).

[39] R. S. Tuminaro, M. Heroux, S. A. Hutchinson, and J. N. Shadid, "Official Aztec User's Guide: Version 2.1" (1999).

[40] Q. Zhang, W. Zhao, and A. C. Seabaugh, "Low-subthreshold-swing transistors", IEEE Elec. Dev. Lett. **27**, 297-300 (2006).
[41] T. B. Boykin, G. Klimeck, R. Chris Bowen, and F. Oyafuso, "Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory", Phys. Rev. B **66**, 125207 (2002).
[42] G. Klimeck, R. C. Bowen, T. B. Boykin, and T. A. Cwik, "sp3s* Tight-Binding Parameters for Transport Simulations in Compound Semiconductors", Superlattices and Microstructures **27**, 519 (2000).
[43] M. Luisier, "A parallel implementation of electron-phonon scattering in nanoelectronic devices up to 95k cores", Proceedings of the 2010 ACM/IEEE Conference on Supercomputing, (2010).
[44] http://www.nccs.gov/computing-resources/jaguar/
[45] http://icl.cs.utk.edu/papi
[46] D. E. Petersen, S. Li, K. Stokbro, H. H. B. Soerensen, P. C. Hansen, S. Skelboe, and E. Darve, "A hybrid method for the parallel computation of Greens functions", J. of Comp. Phys. **228**, 5020 (2009).
[47] http://nanohub.org/resources/omenwire