Parallelized *Ab Initio* Quantum Transport Simulation of Nanoscale Bismuthene Devices

M. Matić, T. Župančić, and M. Poljak*
Computational Nanoelectronics Group
Micro and Nano Electronics Laboratory, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, HR 10000, Croatia
*E-mail: mirko.poljak@fer.hr

**Abstract**—We describe our parallel *ab initio* quantum transport solver implemented in C programming language, with bismuthene nanoribbon (BiNRs) simulations used for the demonstration of its performance. The inputs are Hamiltonians obtained from *ab initio* density functional theory (DFT), which are wannierized into a localized basis to increase Hamiltonian matrix sparsity and to reduce the computational load without the loss of bandstructure accuracy. Numerical matrix operations are parallelized for cluster computation and optimized using Intel Message Passing Interface (MPI) and Intel oneAPI Math Kernel Library (MKL). We demonstrate that an acceleration of about 45× is achieved through parallelization on 64 Xenon Silver CPU cores compared to a single-core execution. Finally, we investigate the electronic, transport and device properties of ultra-scaled bismuthene nanodevices.

**Keywords**—quantum transport, Green’s functions, bismuthene nanoribbon, parallel computing, density functional theory (DFT), maximally-localized Wannier functions (MLWF), Message Passing Interface (MPI), Math Kernel Library (MKL)

I. INTRODUCTION

In the last 60 years, semiconductor industry growth has been driven by transistor miniaturization and introduction of performance boosters such as metal gates, high-κ dielectrics, and strain. The current state-of-the-art transistor technology is fin field-effect transistor (FinFET) with a nanoscale active region. Fabrication and device optimization at such scale are expensive and complex due to strong quantum mechanical effects which a strong suppression is difficult and, hence, new materials and structures are sought to enable further device scaling and to improve device performance [1].

The two-dimensional (2D) materials (2DMs) are potential candidates for future transistors [2]. Due to the dangling-bond-free surface and atomic thickness, the 2DMs have much higher immunity to SCEs and better transport properties than the traditional transistor materials and structures. Currently, the main downside and limiting factor of 2DM applications in electron devices is the high source and drain (S/D) contact resistance that reaches up to several ~kΩ·μm [3], [4]. Numerical studies have shown that there are more than ~1800 2DMs [5] as candidates for future FETs. Only a fraction of these materials was previously experimentally demonstrated, among which the group-V monoelemental 2DMs such as phosphorene, arsenene, antimonene, and bismuthene have attracted much research interest recently [6]–[9]. Electronic and transport properties of 2DMs, and device performance of 2DM FETs, can be tuned by configuring the material into various nanostructures such as nanoribbons [10]–[12], nanosheets [13] and nanowires [14].

With high experimentation cost and strong quantum mechanical effects, advanced physical and numerical modeling is needed for the development of future transistors based on 2DMs [14]. The first simulation step is calculating the electronic bandstructure, where the state-of-the-art approach is using *ab initio* density functional theory (DFT) [15] whose accuracy greatly depends on the exchange-correlation (XC) functional. Although DFT gives accurate results, it is computationally demanding which limits these simulations to periodic crystals or nanostructures containing only a few dozen atoms. The output of DFT is a Hamiltonian that describes the electronic structure of the simulated material. The DFT Hamiltonian is represented with a large dense matrix localized in energy. However, for quantum transport calculations localization in space is preferred as it leads to sparser matrices. Maximally-localized Wannier functions (MLWFs) [16] can be used to transform the DFT Hamiltonian into a spatially localized basis. The MLWF Hamiltonian matrix is sparse while keeping the high accuracy of the bandstructure from the DFT simulations. Finally, non-equilibrium Green function (NEGF) formalism is employed for quantum transport simulations of arbitrary nanodevices. Depending on computational resources, simulation of devices consisting of thousands of atoms can be possible using NEGF with MLWF Hamiltonian [17]–[19].

In this work, the NEGF solver is implemented in C programming language and the solver is parallelized using Intel Message Parsing Interface (MPI) and linear algebra calculations are optimized for Intel CPUs using Intel OneAPI Math Kernel Library (MKL). Acceleration through parallelization and matrix operation optimization is studied using bismuthene nanoribbons (BiNRs) of various sizes as an example. In addition, we studied the electronic and transport properties of BiNRs, and device
performance of ultra-scaled BiNR FETs using ab initio NEGF simulations.

II. AB INITIO QUANTUM TRANSPORT SIMULATIONS

Simulation of nanodevices with atomic resolution must properly account for the atomistic nature of the system, and quantum mechanics that governs device operation. The Hamiltonian is obtained using *ab initio* DFT calculations, and then transformed into a localized basis using MLWFs. Finally, the MLWF Hamiltonian is used as an input for our in-house NEGF solver.

A. Density functional theory

DFT is an *ab initio* method used for the simulation of materials with atomic and orbital resolution. In DFT, geometry and electronic density of the studied material are iteratively optimized until the forces on atoms and the total energy converge below specified values, i.e. when the solution is considered close enough to the system’s true ground state. Atoms are moved to minimize the forces on atoms in geometry optimization, while electronic density optimization relies on geometry and Kohn-Sham scheme [15] to iteratively solve a set of single-particle wavefunctions based only on three space variables. One of the main components of the Kohn-Sham scheme is the exchange-correlation (XC) functional. The most frequently used XC-functional is the generalized gradient approximation (GGA) parametrized by Perdew, Burke, and Ernzerhof (PBE) [20]. The XC-functional are included in the pseudopotential input file for each element.

For DFT calculations presented in this work we use Quantum Espresso (QE). QE uses plane waves to expand the single-particle Kohn-Sham wavefunctions. The most important inputs are the primitive unit cell of the material with definitions of position and types of each atom, cutoff energy for plane waves due to computational limits, ionic force and electron convergence thresholds, and wavevectors (k-points) in the first Brillouin Zone (BZ) for which calculations are performed. The output of DFT calculations is the Hamiltonian of the simulated system.

The DFT calculations of BiNRs, which are quasi-one-dimensional (1D) nanostructures, starts with the identification of 2D bismuthene primitive unit cell from Materials Cloud [21]. Vacuum is added in confined direction of the 2D material to eliminate any unwanted interactions. The structure is relaxed and the primitive unit cell atom coordinates are extracted and used to construct the primitive unit cell of the bismuthene nanoribbon. Nanoribbon edges are passivated by hydrogen atoms and vacuum is added in confined directions. For 2D material and nanoribbon, k-points are sampled using an equally-spaced Monkhorst-Pack grid [22] in transport directions and 1 k-point in confined directions. The cutoff energy is set to 60 Ry, whereas the convergence threshold is set to $10^{-3}$ eV/Å for the ionic force, and to $10^{-4}$ eV for energy. Spin-orbit Coupling (SOC) affects 2D bismuthene considerably [7], but in this work it is not included due to the high computational cost of including SOC in DFT calculations. Therefore, our investigation of BiNR and BiNR FET properties are approximate, serve to demonstrate our DFT-MLWF-NEGF simulation chain, and the usefulness of results depends on whether edge states can be turned off by SOC engineering.

B. Maximally-localized Wannier functions

QE uses plane-wave basis to expand single-particle Kohn-Sham wavefunctions, which enables highly accurate calculation of electronic state that are localized in energy, but are not localized in space. However, quantum transport simulations desire spatially localized states since they result in sparser Hamiltonians that reduce the computational burden of NEGF calculations.

Plane-waves in periodic crystal are Bloch waves, and resulting electronic states are Bloch states. Localized real-space representation of Bloch states was introduced by Wannier [23], and later a technique was developed by Marzari and Vanderbilt [16] to construct Wannier functions (WFs) with maximal localization criteria. Maximally-localized Wannier functions (MLWFs) Hamiltonians are represented with maximally sparse block-tridiagonal matrix which enables the simulation of nanodevices with realistic physical sizes.

Process of finding MLWFs starts with projecting the already localized trial orbitals (usually atomic orbitals) on the Bloch manifold results in functions that are smooth in the wave-vector space but are not orthonormal. The functions are orthonormalized by applying the Löwdin orthonormalization technique [24]. Finally, the problem is reduced to finding gauge transformation that minimizes the localization functional defined as quadratic spread sum of WFs. This method was implemented in Wannier90 program [25] that is used for Hamiltonian wannierization in this work.

Wannierization process for BiNRs consists of three steps. First, the DFT output is preprocessed into Wannier90 compatible form. In second step, initial projections and orbital overlaps are calculated. For BiNRs, we choose $p_x$, $p_y$, and $p_z$ atomic orbitals as trial orbitals. Finally, gauge transformation that minimizes the localization functional is calculated and MLWF Hamiltonian is obtained. The MLWF Hamiltonian presents a supercell along the nanoribbon width, and it is repeated along the nanoribbon length in order to construct the total Hamiltonian of the BiNR that enters the NEGF equations.

C. NEGF formalism for quantum transport

The NEGF formalism is used to directly solve the Schrödinger's equation with open boundary conditions (OBCs). The central quantity of the NEGF formalism is the device retarded Green's function, $G^R$, which is defined as

$$G^R(E) = [(E + i\eta) - H - \Sigma_i^L(E) - \Sigma^R_d(E)]^{-1}, \quad (1)$$

where $H$ is a total device Hamiltonian matrix, while $\Sigma$ matrices represent source (S) and drain (D) retarded contact self-energy matrices, respectively. The term $i\eta$ is the infinitesimal positive imaginary convergence constant added to the energy term, which is needed to obtain a unique causal retarded function solution from the two
possible solutions. Main outputs of NEGF formalism are transmission and density of states (DOS) that are used for calculating the current-voltage characteristics of BiNR FETs. Using the DFT-MLWF-NEGF results, we employ the ballistic self-consistent top-of-the-barrier (ToB) model for BiNR FET simulations [26]–[28] and study the on-state performance of these devices obtained for a supply voltage of 0.7 V.

III. PARALLEL IMPLEMENTATION WITH MPI

A. Cluster description and setup

Computations are performed on a high-performance cluster composed of one main and eight compute nodes. Nodes run Ubuntu 20.04 Focal Fossa, with desktop environment installed only on the main node, used as a user login node. Each compute node contains an Intel Xeon Silver CPU (8 core), a single Nvidia Quadro P4000 GPU (1792 CUDA cores, 8 GB GDDR5, FP64: 166 GFLOPS) and 32 GB of RAM. Primary storage is located on the main node and dynamically mounted on compute nodes. The cluster is connected over a segregated network, configured exclusively for the purposes of cluster operation, and designed to reduce communication overhead. Software installation and cluster management is done using the Ansible automation system. Software utilized for computation acceleration that is installed on the cluster is the Intel oneAPI Math Kernel Library (MKL), which includes numerical linear algebra libraries BLAS (Basic Linear Algebra Subroutines) and LAPACK (Linear Algebra Package). Message Passing Interface (MPI) is utilized for parallelizing program execution over compute nodes.

B. Contact self-energy matrices

Self-energy matrices Σ in Eq. (1) can be interpreted as a modification of the Hamiltonian to account for the OBCs. The Σ matrices are energy-dependent and non-Hermitian, where the anti-Hermitian part of Σ defines the level broadening matrix Γ, which is needed for the calculation of transmission. The real part of Σ causes energy level shifts, and the imaginary part gives the electrons a finite lifetime.

The Σ matrices are calculated using surface Green's function (GF) which itself is calculated using a recursive method. For each energy in the defined energy spectrum, the contact GF is calculated layer by layer in the direction of contact extension until difference between the two consecutive layers is smaller than a specified tolerance. After convergence is achieved, the contact with a finite lifetime in their tag parameter. Messages are sent after all individual computations are completed, thus eliminating the need for periodical probing or an intermediary MPI process.

Listing 1 shows the C code for the transmission calculation using the NEGF formalism over energy points in the pre-defined spectrum. Parallelization is enabled by using mpi_rank and mpi_size functions which return rank from MPI_Comm_rank and world size from MPI_Comm_size, respectively. Intel MKL functions used in this part of the code are mkl_zomatadd, mkl_zomatcopy, and cblas_zgemm, for matrix additions, copying and multiplication, respectively. As an addition to these MKL functions, mm_zgeadd and mm_invert are written from scratch and used for matrix addition and inversion. Here, mm_zgeadd is essentially mkl_zomatadd, but the output is written over the second matrix input. On the other hand, mm_invert computes the inverse of the input matrix, and the result overwrites the data in memory where the input matrix was stored. Calculation of contact self-energy matrices is performed using our C function sgf_iter_sr_v2 for the calculation of surface GFs using the Sancho Rubio method. After the calculation of surface GFs, our C function set_sigma is used for the creation of S/D contact self-energy matrices (Σ).

IV. RESULTS AND DISCUSSION

Bismuthene is a group-V 2D monolayer material with a hexagonal crystal lattice. Patterning bismuthene into quasi-1D nanoribbons enables the tuning of electronic and transport properties in nanoscale bismuthene-based devices. The simulation chain from DFT over quantum transport to BiNR FET simulation based on the ToB model is reported for BiNR widths (W) of 0.86 nm, 1.29 nm, 1.73 nm, 2.18 nm, and 2.61 nm. For the ToB device study, we consider an n-type BiNR FET with an intrinsic channel and highly doped S/D regions with the Fermi level in the source region set to approx. 0.02 eV above the conduction band minimum (CBM) on the source side of the device, in order to preserve charge neutrality condition in S/D regions doped to 0.001 molar fraction of Bi atoms areal density [4]. Gate oxide is a 1 nm-thick SiO2 with εr = 3.9. For simplicity, gate work function is set automatically so that the flat band voltage equals 0 V, and that all BiNR FETs have the same threshold voltage of 0.25 V.

A. Electronic, transport and device properties of BiNRs

Dispersion obtained by DFT and after the wannierization process is compared in Fig. 1 for the...
1.29 nm-wide BiNR. The plot shows that the DFT-MLWF procedure is working well since an excellent match of bandstructure around the bandgap is achieved, which is the relevant energy range for device operation. In this case, we observe that the 1.29 nm-wide BiNR is a direct semiconductor, but there are two other valleys near the CBM along the transport $k$-direction.

The satellite valleys away from the Γ point directly reflect on the sharp increase of the transmission near the CBM as can be seen in Fig. 2a, where the red dash-dot-dot line shows the transmission for $W = 1.29$ nm. The transmission in the conduction and valence bands are reported in Fig. 2a and Fig. 2b, respectively, for BiNRs that are 0.86 nm, 1.29 nm, 1.73 nm, and 2.61 nm wide. Scaling down BiNR width increases the bandgap from 0.88 eV to 1.7 eV and decreases the transmission near the CBM, as can be seen in Fig. 2a, where the red dash-dot-dot line shows the transmission for $W = 1.29$ nm. On the other hand, width downscaling shifts down the second and lower subbands away from the valence band maximum, which leads to lower hole transmission in narrower BiNRs (Fig. 2b).

Finally, the ballistic ON-state current for $n$-channel BiNR FETs is calculated for all the considered BiNR widths and is plotted in Fig. 3. Scaling down BiNR width leads to a decrease of the ON-state current from 2.85 mA/μm for the 2.61 nm-wide BiNR FET down to only 0.35 mA/μm for the 0.86 nm-wide BiNR FET. Although the 0.86 nm-wide BiNR has the highest transmission near the CBM, ON-state current value is smallest of all studied BiNRs due to the wide bandgap. To further analyze reasons of this behavior, mobile charge density and average charge velocity must be studied, but this is beyond the scope of the current work.

B. Acceleration by parallel execution

In Fig. 4 we compare the time to solution using different number of cores for a Hamiltonian dimension ($N$) of $N = 780$ and Hamiltonian matrix size ($N_H$) of $N_H = 780 \times 780 = 608.400$, corresponding to the BiNR size of 2.61 nm and 7.44 nm for the width and length, respectively. The workload is equally divided on the smallest of all studied BiNRs due to the wide bandgap. To further analyze reasons of this behavior, mobile charge density and average charge velocity must be studied, but this is beyond the scope of the current work.

```
Listing 1. Part of the code for transmission calculation for each energy point parallelized using MPI and optimized using Intel MKL.
```
is given by the Amdahl’s law [30] with the following formula

$$S(N) = \frac{1}{1 - p + p/N}$$  \hspace{1cm} (2)

where $S$ is the theoretical speedup, $p$ is the proportion of the program execution time that can be parallelized, and $N$ is the number of cores used. Applying the formula to the observed execution speedup obtained for the 64-core case gives $p$ greater than 0.9933, demonstrating the high parallelization effectiveness of the implemented quantum transport solver.

Effects of increasing $N$, which represents the total number of elements in bismuthene nanodevice Hamiltonians, on the time to solution is shown in Fig. 5 for 16-, 32- and 64-core parallel execution. Increasing the Hamiltonian dimension from 300 to 780, i.e. increasing the number of atoms by 2.6× and the number of matrix elements by ~6.8×, increases the time to solution by 15.3× (from 8.7 s to 133.6 s), when all 64 cores are utilized. We observe that the NEGF solver scales with sub-cubic complexity in relation to the Hamiltonian dimension. For all Hamiltonian dimensions approximately the same relative speedup is observed for the 16-, 32- and 64-core execution, which confirms the parallelization efficiency of the code.

V. CONCLUSIONS

The implementation of our \textit{ab initio} quantum transport solver, parallelized using MPI and optimized using MKL, is described in detail and its performance is demonstrated on ultra-scaled 2D material nanoribbons based on bismuthene. The simulation chain from DFT, over wannierization for spatially localized Hamiltonians, to quantum transport based on NEGF is illustrated and applied for the study of electronic, transport and device properties of BiNRs with the widths under ~2.6 nm. While further studies including SOC are needed, BiNR FETs show promising current drivability for $W > 2.5$ nm.
Fig. 5. Time-to-solution vs. BiNR Hamiltonian size in log-log scale for three cases when our NEGF solver is running on 16, 32, and 64 cores.

Finally, we demonstrate a code-execution acceleration by parallelization of about ~45.8× by using 64 cores in comparison to the single-core execution. Our NEGF solver enables the simulation of relatively large nanostructures and acceptable time to solution.

ACKNOWLEDGMENTS

This work was supported by the Croatian Science Foundation under the project CONAN2D (Grant No. UIP-2019-04-3493).

REFERENCES


