NEMO5, Xeon Phi and hStreams: Physics of Ultrascaled 2D Nanotransistors

Xinchen Guo, Kuang-Chung Wang, James Charles, Junzhe Geng, Daniel Mejia, Daniel Valencia, Daniel Lemus, James E. Fonseca, Gerhard Klimeck, and Tillmann Kubis
School of Electrical and Computer Engineering and Network for Computational Nanotechnology
Purdue University
West Lafayette, Indiana, USA
xg@purdue.edu

***Abstract*—The detailed electrical and thermal properties of future ultrascaled transistors are critically dependent on quantum effects. This is particularly true for the promising class of 2D-material based transistors. NEMO5 – a multipurpose, multiscale semiconductor device simulation tool is applied on these devices in many academic and industrial groups (including Intel, Samsung, TSMC etc.). Harnessing the power of modern HPC hardware in world's largest supercomputing centers, NEMO5 is able to unveil more fine details of nanotransistors. This work discusses how a physical Xeon Phi is partitioned into several virtual Xeon Phis using the hStreams library and put into a MPI parallelization scheme.**

Keywords—Xeon Phi; hStreams; NEMO5; Nanotransistor

I. INTRODUCTION

With the reproducible fabrication of layered graphene systems in 2004, a new field of material and device research started – the field of 2D materials [1] - [9]. Such materials, e.g. layers of transition metal dichalcogenides (TMDs) combine exciting thermal [7], electrical [8] and opto-electronic [9] properties with the possibility to design and fabricate ultimately thin (i.e. single or few atom layer thick) nanodevices. More efficient-than ever light emitting and absorbing devices, wearable nanoelectronics, and smallest possible nanotransistors are only a few examples. Given that the properties of such devices critically depend on unintuitive quantum mechanical features (such as tunneling, confinement and particle interference effects) theoretical modeling is essential.

In contrast to well-established 3D semiconductor materials (such as silicon, germanium, etc.), electrons in 2D materials face a unique physical mechanism: sheets of 2D material are hardly ever perfectly smooth, but they host corrugations of various lengths and intensities. Electrons that propagate in 2D devices interact with such corrugations and lose significant amounts of their momentum. This results in pronounced device resistances and degradation of device efficiency. 2D nanodevice simulations with NEMO5 [10] - [12] are essential to illustrate these effects and show ways for avoidance.

The use of nanoHUB.org computational resources operated by the Network for Computational Nanotechnology, funded by the US National Science Foundation under Grant Nos. EEC-0228390, EEC-1227110, EEC-0228390, EEC-0634750, OCI-0438246, OCI-0832623, and OCI-0721680, is gratefully acknowledged. NEMO5 developments were critically supported by an NSF Peta-Apps award OCI-0749140 and by Intel Corp. This work is supported in part by Intel Parallel Computing Center at Purdue and NSF EFRI-1433510. This work was supported in part by funding from the Semiconductor Research Corporation's Global Research Collaboration (GRC) (2653.001), and Member Specific Research Intel (MSR-Intel) (2434.001). Additional funding was provided by the Semiconductor Research Corporation membership in the Network for Computational Nanotechnology. This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (award number ACI 1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This work is also part of the "Accelerating Nano-scale Transistor Innovation with NEMO5 on Blue Waters" PRAC allocation support by the National Science Foundation (award number OCI-0832623).

II. NEMO5: PHYSICS AND IMPLEMENTATION

A. Physical Model of 2D Systems

NEMO5 is used to solve 2D device electrons in atomistic resolution. All quantum mechanical effects that are common in nanometer length scales are covered within the nonequilibrium Green's function method (NEGF) [13], [14]. Real-world effects, such as finite temperature (lattice vibrations) [15], corrugation (see Fig. 1) and device imperfections (discretized with about 200 atoms per cross line, with about 120,000 atoms in total) are included. The electrostatics of electrons in the field of applied gate and contact voltages are solved self-consistently within the Poisson equation.

B. Numerical Implementation

The main computational load for the considered problem lies in the recursive solution of the NEGF equations. The impact of Xeon Phi on the overall NEMO5 simulation time,

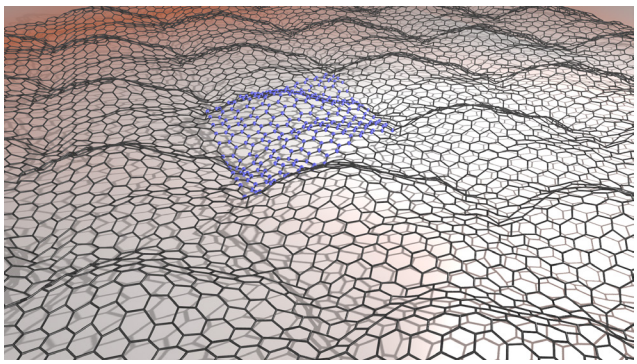

Fig. 1. Schematic of corrugation effects in graphene. Blue highlighted area depicts the central unit cell that gets repeated in the simulation domain. Large unitcells are required to avoid artificial correlation effects.

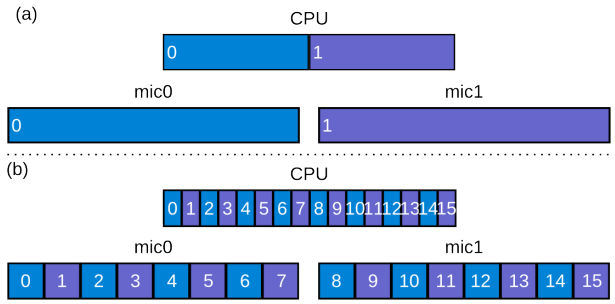

Fig. 2. Schematic of Xeon Phi logic domain partitioning for the case of 2 (a) and 16 (b) MPI processes.

one core NEGF equation (Eq. (61) of [16]) is used to benchmark offloading performance in the 2D quantum transport framework. This equation contains one addition, one conjugate transpose and four multiplications of three squared input matrices. All matrices are treated as dense.

Xeon Phi can work in native mode, compiler assisted offloading mode and automatic offloading mode. hStreams [17] is a variant of compiler assisted offloading mode. Users explicitly move data and schedule computing tasks on Xeon Phi. NEMO5 code paradigm allows a straightforward usage of Xeon Phi with hStreams control infrastructure. Alternative ways to control Xeon Phi usage include the "native mode", the "automatic offloading" and the "compiler assisted offloading mode" [18]. hStreams turned out to be NEMO5's best choice due to its library-like support for easy and complete task offload control.

The efficiency of Xeon Phis gets optimized with the appropriate number of logical domains on each physical Xeon Phi. Splitting a physical Xeon Phi requires specifying for each Xeon Phi core to which logical domain it belongs. This is done with the variable CPU_MASK. In NEMO5, the CPU_MASK variable is solved according to local communicator size and the respective MPI rank. Fig. 2 (a) and (b) schematically show 2 and 16 MPI process configurations, respectively.

The Conte cluster at the Purdue Rosen Center for Advanced Computing (RCAC) is used for performance benchmarking. Each computing node is equipped with two Intel Xeon E5-2670 CPUs and two Intel Xeon Phi 5110P co-processors. Intel compiler 14 and hStreams 1.0 are used on the hosts, while Intel Manycore Platform Software Stack (MPSS) 3.6.1 is installed on the Xeon Phis.

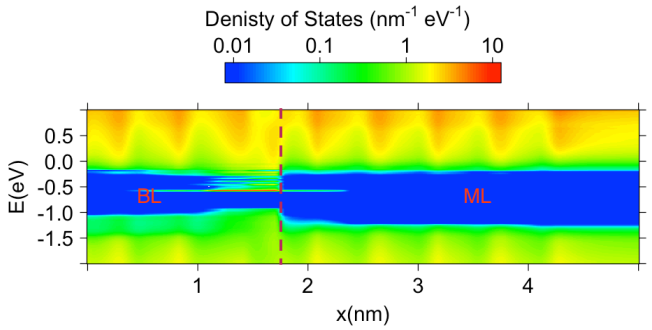

Fig. 3. Contour graph of the energy and spatially resolved density of states at the interline between the mono- and bilayer MoS_2 used in experiments of [1]

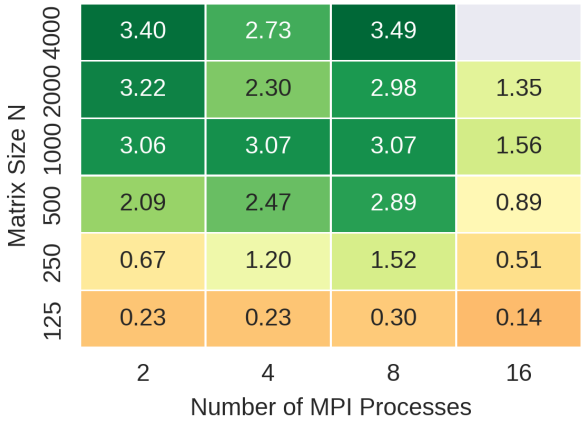

Fig. 4. Heat map of speedups when NEGF in NEMO5 runs with Xeon Phis instead of pure CPU calculations.

III. Results and Discussions

A. Physical Results

Most electronic devices consist of regions with different material properties. One possible configuration involves interlines between single and double layer MoS_2 structures [1]. Critical for device performances are information about the amount of trapped electrons at the interline between these layers and their impact on the current carrying electrons. Fig. 3 illustrates the electronic density of states at the interline between these 2D systems. Trapped electrons can be seen as peaks in the density of states in the middle of the band gap (blue portion of the contour graph, around the position of 1.7nm). This modeling result agrees well with experimental findings [1].

B. Speedup with hStreams

Fig. 4 shows the achieved speedup when using hStreams on Xeon Phi compared with CPU usage only. When the matrix size is 4000 by 4000, Xeon Phis run out-of-memory when 16 MPI processes are used (greyed out in Fig. 4). NEGF calculations can benefit from hStreams and virtual Xeon Phis with input matrices equal or larger than 500 by 500. When each Xeon Phi hosts 8 virtual Phis (i.e. 16 MPI processes per node), speedup decreases dramatically. This may be caused by OS and data movement overhead.

Currently, the CPU_MASK is calculated without cores reserved for Xeon Phi OS. It is imaginable that reserving enough resources for the OS can be critical when multiple hStreams processes are running simultaneously. Also, moving data asynchronously may reduce congestions as Knights Corner Xeon Phi only has one DMA engine. Future NEMO5 development will assess those points carefully.

Acknowledgment

We acknowledge the Rosen Center for Advanced Computing at Purdue University for the use of their computing resources and technical support.

REFERENCES

[1] C. Zhang, Y. Chen, J.-K. Huang, X. Wu, L.-J. Li, W. Yao, J. Tersoff, and C.-K. Shih, "Visualizing band offsets and edge states in bilayer–monolayer transition metal dichalcogenides lateral heterojunction," Nat. Commun., vol. 7, p. 10349, Jan. 2016.

[2] Á. Szabó, R. Rhyner, and M. Luisier, "Ab initio simulation of single- and few-layer MoS 2 transistors: Effect of electron-phonon scattering," Phys. Rev. B, vol. 92, no. 3, p. 035435, Jul. 2015.

[3] C.-C. Wu, D. Jariwala, V. K. Sangwan, T. J. Marks, M. C. Hersam, and L. J. Lauhon, "Elucidating the Photoresponse of Ultrathin MoS 2 Field-Effect Transistors by Scanning Photocurrent Microscopy," J. Phys. Chem. Lett., vol. 4, no. 15, pp. 2508–2513, Aug. 2013.

[4] V. K. Sangwan, H. N. Arnold, D. Jariwala, T. J. Marks, L. J. Lauhon, and M. C. Hersam, "Low-Frequency Electronic Noise in Single-Layer MoS 2 Transistors," Nano Lett., vol. 13, no. 9, pp. 4351–4355, Sep. 2013.

[5] D. Jariwala, V. K. Sangwan, D. J. Late, J. E. Johns, V. P. Dravid, T. J. Marks, L. J. Lauhon, and M. C. Hersam, "Band-like transport in high mobility unencapsulated single-layer MoS2 transistors," Appl. Phys. Lett., vol. 102, no. 17, p. 173107, 2013.

[6] D. Jariwala, V. K. Sangwan, L. J. Lauhon, T. J. Marks, and M. C. Hersam, "Emerging Device Applications for Semiconducting Two-Dimensional Transition Metal Dichalcogenides," ACS Nano, vol. 8, no. 2, pp. 1102–1120, Feb. 2014.

[7] M.-J. Lee, J.-H. Ahn, J. H. Sung, H. Heo, S. G. Jeon, W. Lee, J. Y. Song, K.-H. Hong, B. Choi, S.-H. Lee, and M.-H. Jo, "Thermoelectric materials by using two-dimensional materials with negative correlation between electrical and thermal conductivity," Nat. Commun., vol. 7, p. 12011, Jun. 2016.

[8] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer MoS2 transistors," Nat. Nanotechnol., vol. 6, no. 3, pp. 147–50, Mar. 2011.

[9] O. Lopez-Sanchez, D. Lembke, M. Kayci, A. Radenovic, and A. Kis, "Ultrasensitive photodetectors based on monolayer MoS2," Nat. Nanotechnol., vol. 8, no. 7, pp. 497–501, Jun. 2013.

[10] S. Steiger, M. Povolotskyi, H. H. Park, T. Kubis, and G. Klimeck, "NEMO5: A parallel multiscale nanoelectronics modeling tool," IEEE Trans. Nanotechnol., vol. 10, no. 6, pp. 1464–1474, 2011.

[11] S. Steiger, M. Povolotskyi, H.-H. Park, T. Kubis, G. Hegde, B. Haley, M. Rodwell, and G. Klimeck, "The nanoelectronic modeling tool NEMO 5: Capabilities, validation, and application to Sb-heterostructures," in 69th Device Research Conference, 2011, pp. 23–26.

[12] Y. He, T. Kubis, M. Povolotskyi, J. Fonseca, and G. Klimeck, "Quantum transport in NEMO5: Algorithm improvements and high performance implementation," in 2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2014, pp. 361–364.

[13] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, "Modeling of nanoscale devices," Proc. IEEE, vol. 96, no. 9, pp. 1511–1550, 2008.

[14] S. Datta, "Nanoscale device modeling: the Green's function method," Superlattices Microstruct., vol. 28, no. 4, pp. 253–278, Oct. 2000.

[15] J. Charles, P. Sarangapani, R. Golizadeh-Mojarad, R. Andrawis, D. Lemus, X. Guo, D. Mejia, J. E. Fonseca, M. Povolotskyi, T. Kubis, and G. Klimeck, "Incoherent transport in NEMO5: realistic and efficient scattering on phonons," J. Comput. Electron., pp. 1–7, Jun. 2016.

[16] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices," J. Appl. Phys., vol. 81, no. 12, p. 7845, 1997.

[17] C. Newburn, G. Bansal, M. Wood, L. Crivelli, and J. Planas, "Heterogeneous streaming," IPDPSW, AsHES, 2016.

[18] V. Kumar, "Intel® MKL programming models for Intel® Xeon Phi," 2013. [Online]. Available: https://software.intel.com/en-us/articles/intel-mkl-programming-models-for-intel-xeon-phi.