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

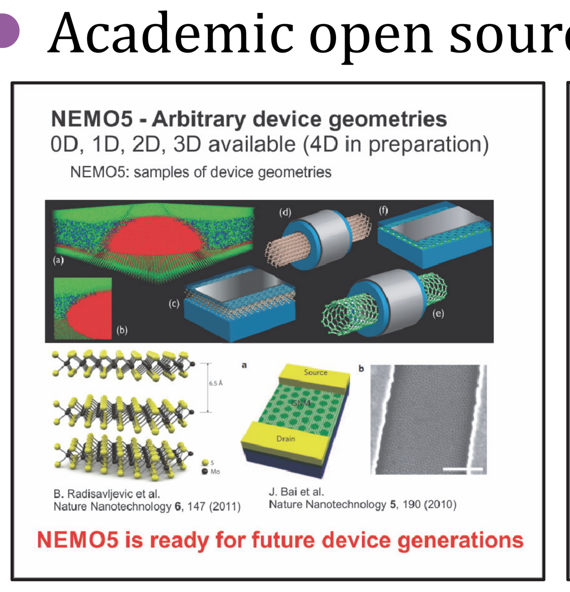
Xinchen Guo, Kuang-Chung Wang, James Charles, Junzhe Geng, Daniel Mejia, Daniel Valencia, Daniel Lemus, James E. Fonseca, Gerhard Klimeck, and Tillmann Kubis

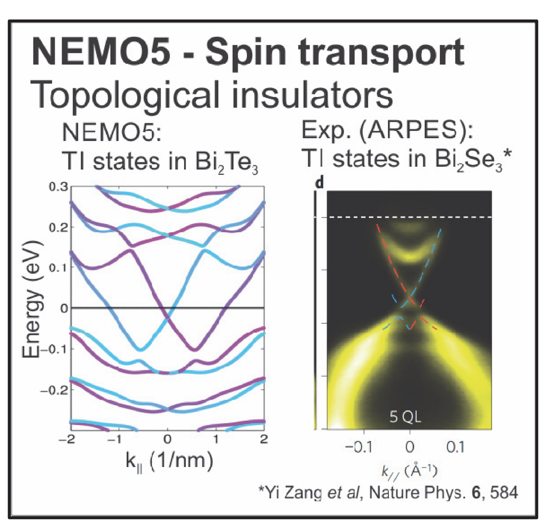
A. Motivation

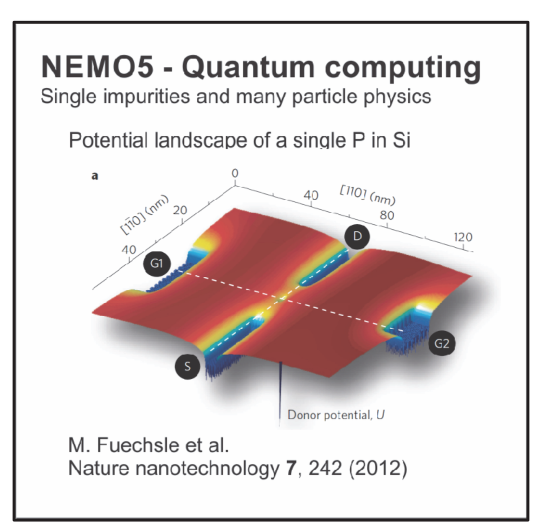
- Enable discovery of the characters of 2D materials
- Accelerate simulations with Xeon Phi co-processors

C. NEMO5

- Multipurpose and multiscale semiconductor device simulation tool
- Scale on large HPC clusters (~180k cores on Tianhe-2)
- Funded by NSF and semiconductor industry (Intel, Samsung, TSMC, Lumileds)
- Academic open source







D. Numerical Challenges

- Quantum mechanical effects are covered by nonequilibrium Green's function (NEGF) method
- 6 core functions to solve NEGF

$$g^r_{L,L}=[E-D_L-t_{L,L+1}g^r_{L+1,L+1}t_{L+1,L}]^{-1}. \quad (59)$$

$$G_{1,1}=[E-D_1-t_{1,0}g^{\ell}_{0,0}t_{0,1}-t_{1,2}g^r_{2,2}t_{2,1}]^{-1}. \quad (60)$$

$$G_{L,L}=g^r_{L,L}+g^r_{L,L}t_{L,L-1}G_{L-1,L-1}t_{L-1,L}g^r_{L,L}, \quad (61)$$

$$G_{L,L}=g^r_{L,L}+g^r_{L,L}(-t_{L,L-1})G_{L-1,L}, \quad (62)$$

$$G_{L-1,L}=G_{L-1,L-1}(-t_{L-1,L})g^r_{L,L}. \quad (63)$$

$$G_{L,1}=g^r_{L,L}(-t_{L,L-1})G_{L-1,1}. \quad (64)$$

- Eq. (61) is used to benchmark performance improvement
- Eq. (61) contains 1 addition, 1 conjugate transpose and 4 multiplications of 3 squared input matrices

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.

F. Benchmark Environment

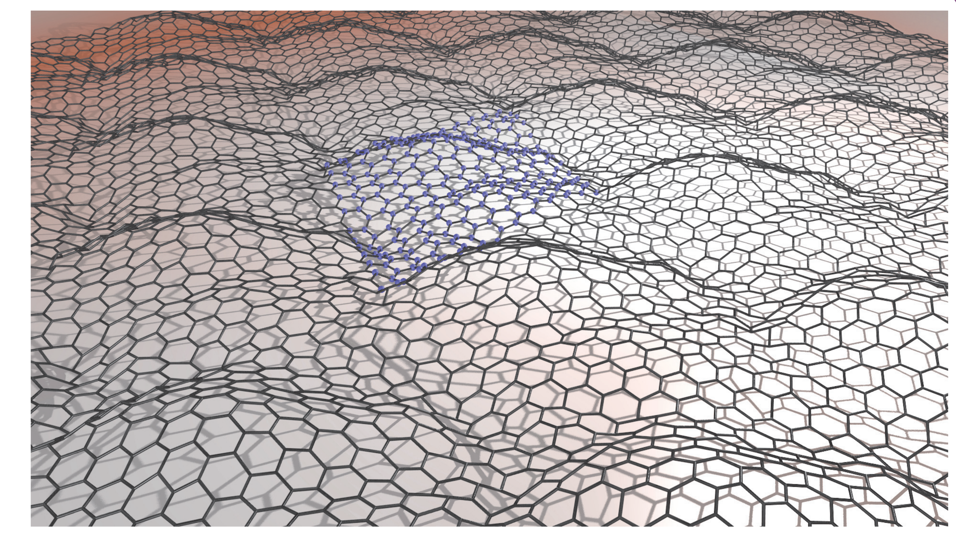
- Conte cluster at Purdue Rosen Center for Advanced Computing
- 2 Xeon E5-2670 CPUs and 2 Xeon Phi 5110P co-processors per computing node
- Intel compiler 14 + hStreams 1.0
- Intel Manycore Platform Software Stack (MPSS) 3.6.1

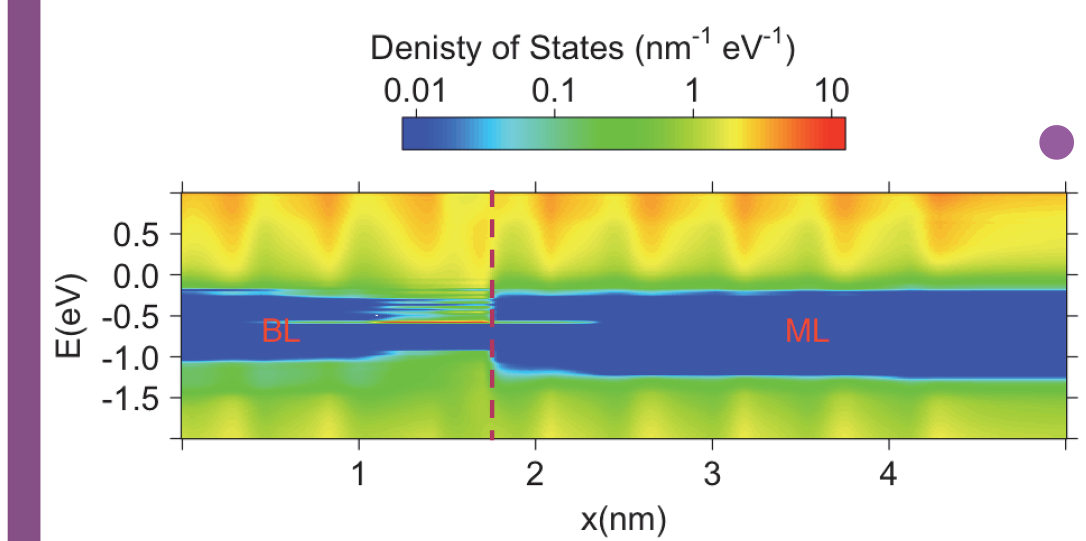
H. Discussions

- Easy storage of intermediate results on Xeon Phi with hStreams framework results in good speedup for complex equations
- Virtual Xeon Phis enable solving more matrices on a Phi, i.e. having less threads for one matrix, which increases overall performance for small matrices
- This model can be extended to a hybrid cluster with CPU nodes and self-hosted KNL nodes, having CPUs solve software logics and KNLs solve heavy FP calculations (offload over network)

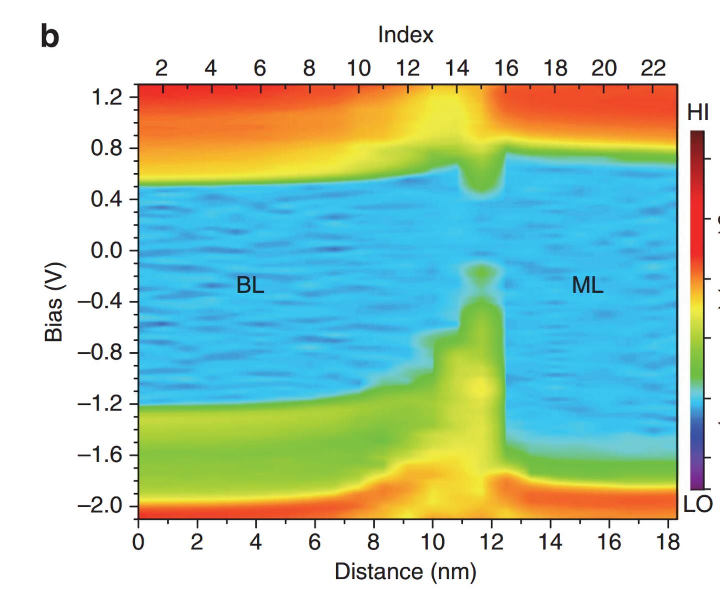
B. 2D Materials

- Emerge in 2004 after reproducible fabrication of layered graphene
- Exciting thermal, electrical and opto-electronic properties
- Ideal for ultra thin nanodevices





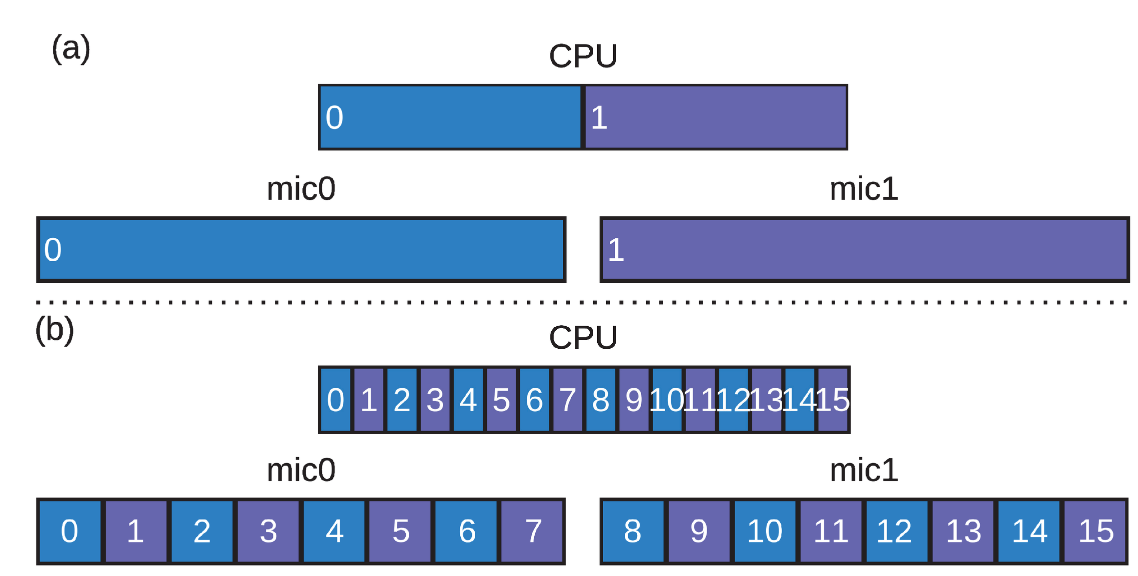
- Corrugations reduce electron mobility in 2D material
  - High electric resistance
  - Low heat conductance
- Corrugations can be modeled with quantum mechanical effects in NEMO5
- Electric and thermal properties can also be simulated and evaluated in NEMO5



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. 6, pp. 1–6, 2016.

E. Partition Xeon Phi with hStreams

- MPI processes calculate own CPU_MASK without overlapping with each other
- hStreams creates non-overlapped logical domain based on supplied CPU_MASK
- (a) and (b) shows 2 and 16 MPI processes configurations respectively



G. Results

- 2 Xeon Phis vs 2 CPUs
- Good speedup for matrices equal or larger than 500 by 500
- Good speedup for MPI processes equal or less than 8 per node
- Reserving cores for Xeon Phi OS and moving data asynchronously may improve performance of 16 MPI processes configuration

Matrix Size N \ Number of MPI Processes	2	4	8	16
4000	3.40	2.73	3.49	
2000	3.22	2.30	2.98	1.35
1000	3.06	3.07	3.07	1.56
500	2.09	2.47	2.89	0.89
250	0.67	1.20	1.52	0.51
125	0.23	0.23	0.30	0.14

I. Acknowledgement

- NEMO5: NSF OCI-0749140, SRC 2653, SRC 2434, and Intel Parallel Computing Center (IPCC) at Purdue
- nanoHUB.org: NSF EEC-0228390, EEC-1227110, EEC-0228390, EEC-0634750, OCI-0438246, OCI-0832623, and OCI-0721680
- Blue Waters: ACI-1238993 and OCI-0832623
- Purdue Rosen Center for Advanced Computing