NEMO5
NanoElectronics MOdeling

Jim Fonseca
Harshad Sahasrabudhe
Evan Wilson
Mehdi Salmani
Gerhard Klimeck

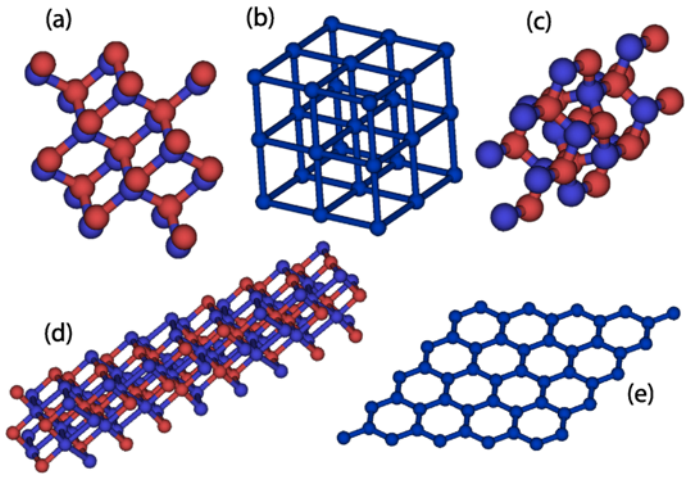
PURDUE UNIVERSITY

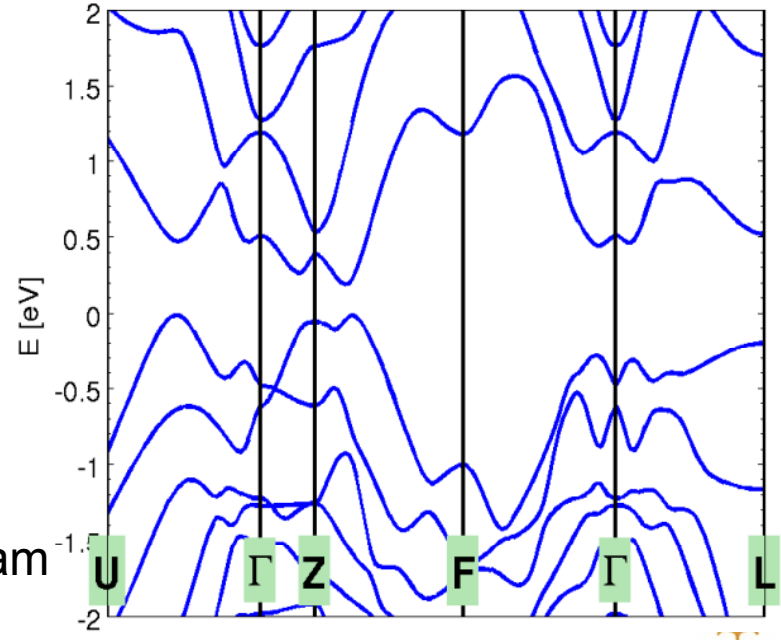
Outline

- NEMO5 Overview
- CPU Scaling
- ITRS Work-International Technology Roadmap for Semiconductors
- GPU Development
- Future Work

NEMO5 – A Multiscale Simulation Tool for Nanoelectronic Modelling

- **Multiscale modeling**
  - Quantum/semiclassical
- **General simulation structures**
  - 1D, 2D, 3D structures
  - Heterostructures, arbitrary shapes, multiple contacts
  - Various crystal structures
  - Metals
- **Hamiltonian basis**
  - Atomistic tight-binding basis
    - (sp3s*, sp3d5s*_SO, ...)
  - Effective-mass approximation
    - (multi-valley, nonparabolicity)

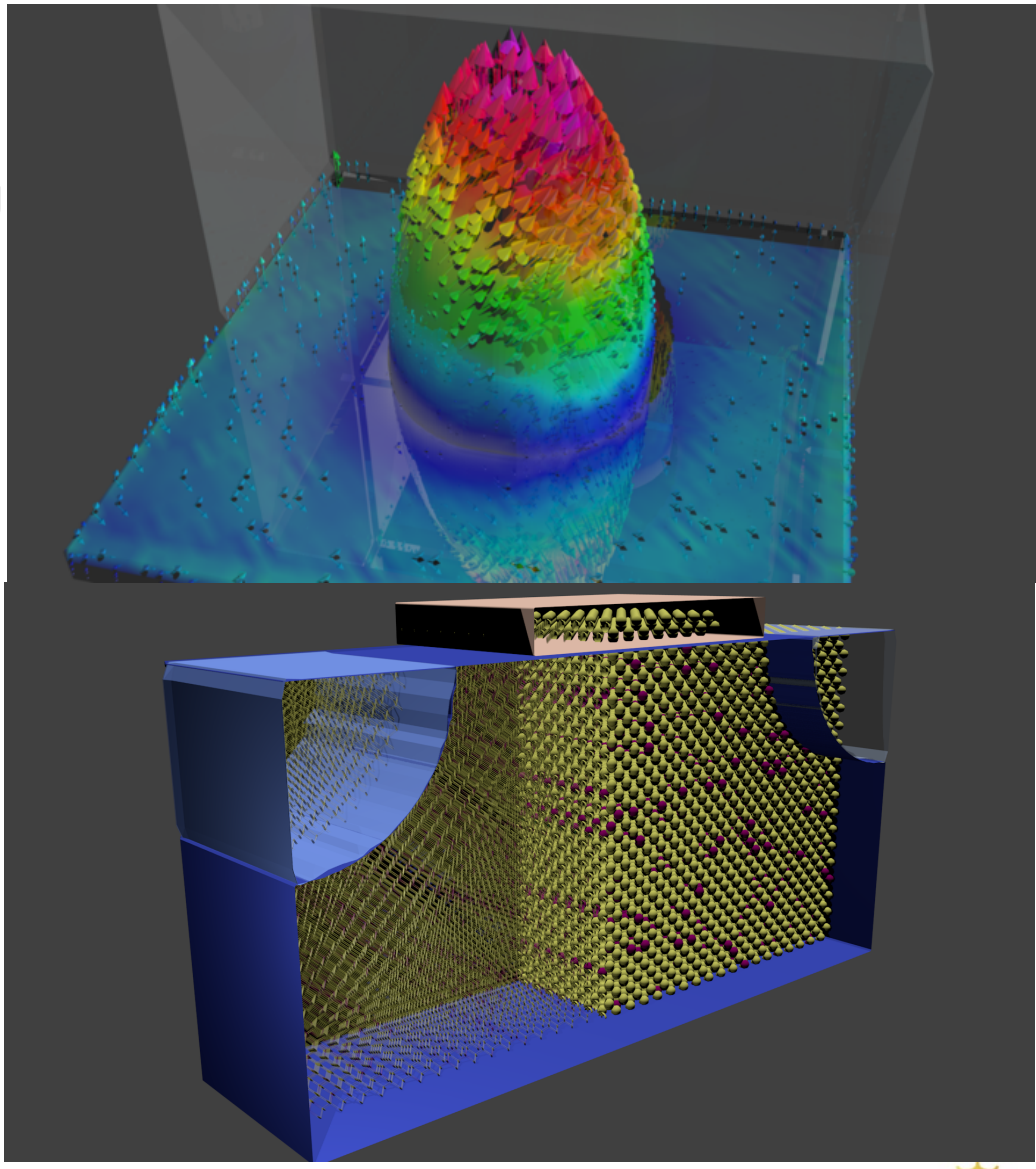




Bi_2Te_3
E-k diagram

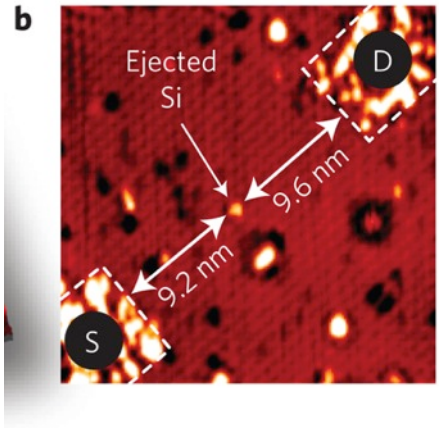
NEMO5 – A multiscale simulation tool for nanoelectronic modelling

- **Various physical models**
  - Ohmic and Schottky contacts
  - Simple and fast phonon scattering model
  - Rigorous phonon model under development
  - Strain models
    - VFF, Keating
  - Magnetic fields
- **Solves**
  - Atomistic strain
  - Electronic band structures
  - Schrodinger, Poisson
  - Charge density, Potential
  - Current



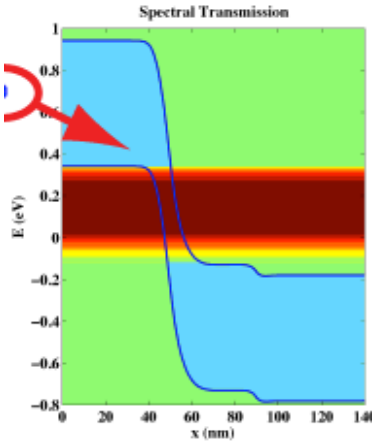
Why atomistic tight binding?

Single atom transistor



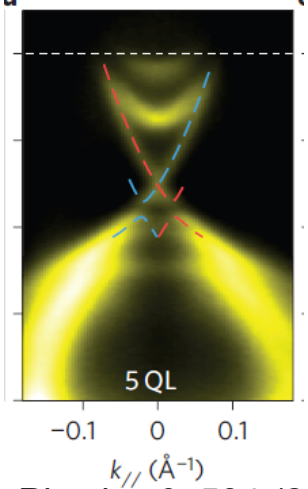
Nature Nanotechnology **7**, 242 (2012)

Band-to-band tunneling



IEEE Elec. Dev. Lett. **30,** 602 (2009)

Topological insulators



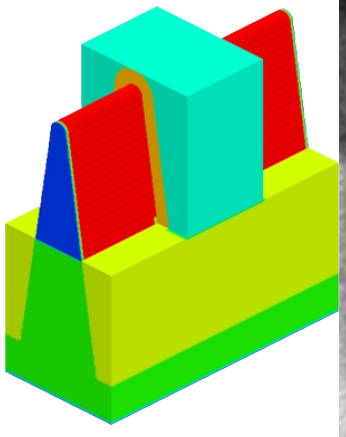
Nature Physics **6**, 584 (2010)

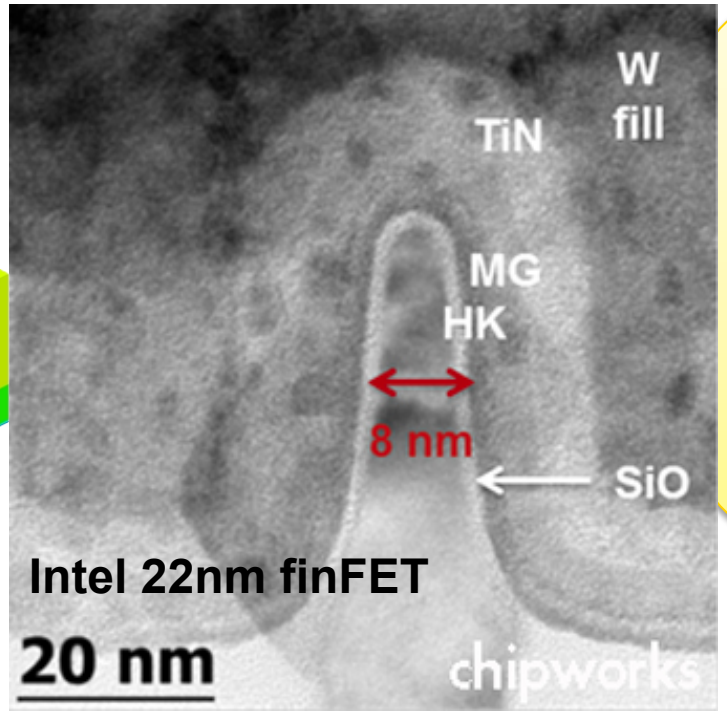
Countable device atoms suggest atomistic descriptions
Modern device concepts, e.g.

- Band to band tunneling
- Exotic materials (Topological insulators, MoS_2, etc.)
- Band/Valley mixing etc.

require multi band representations

Why non-equilibrium Green's functions?





State of the art semiconductor devices

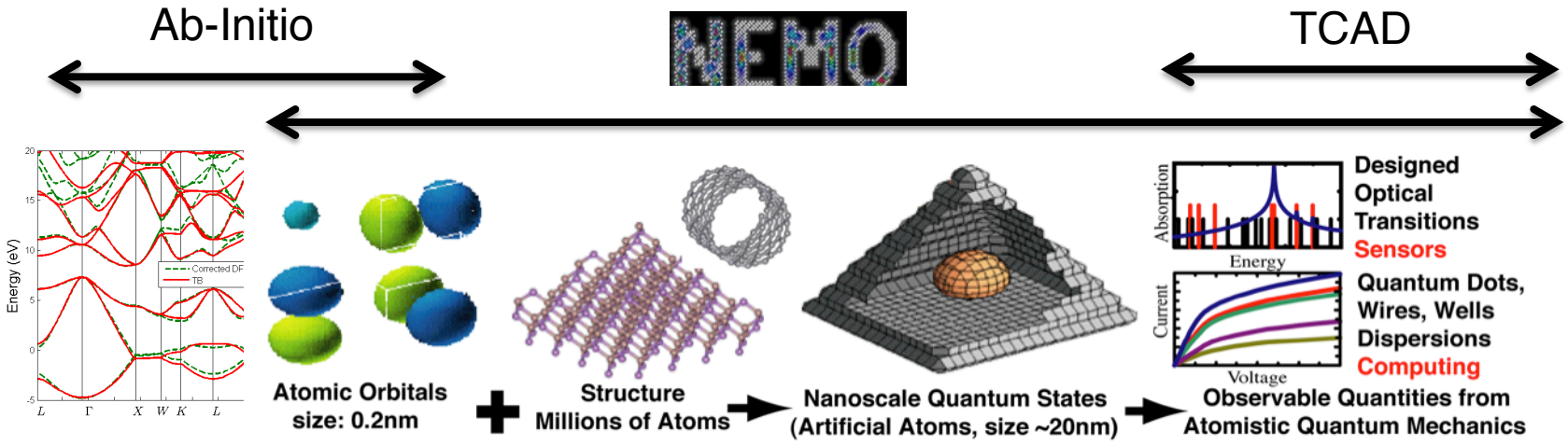
- ➢ utilize or suffer from quantum effects (tunneling, confinement, interference,…)
- ➢ are run in real world conditions (finite temperatures, varying device quality…)

This requires a consistent description of

- coherent quantum effects (tunneling, confinement, interferences,…)
- incoherent scattering (phonons, impurities, rough interfaces,…)

http://www.goldstandardsimulations.com/index.php/news/blog_search/simulation-analysis-of-the-intel-22nm-finfet/
http://www.chipworks.com/media/wpmu/uploads/blogs.dir/2/files/2012/08/Intel22nmPMOSfin.jpg

NEMO5 - Bridging the Scales From Ab-Initio to Realistic Devices



Goal:

- Device performance with realistic extent, heterostructures, fields, etc. for new / unknown materials

Problems:

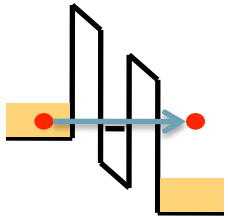
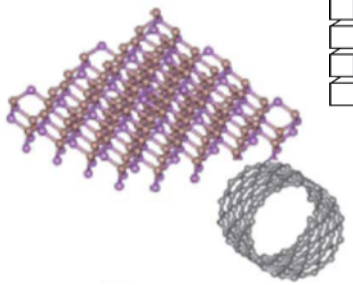
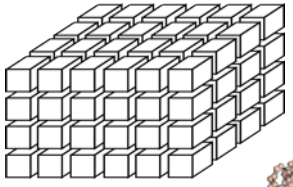
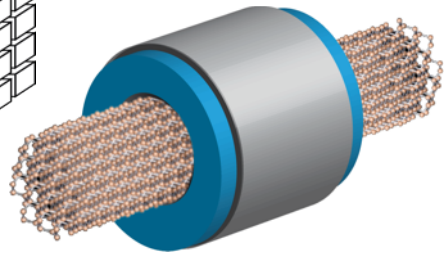
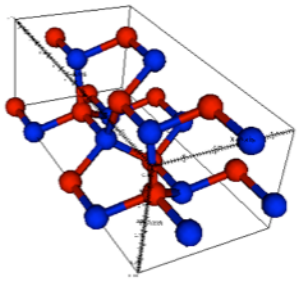
- Need ab-initio to explore new material properties
- Ab-initio cannot model non-equilibrium.
- TCAD does not contain any real material physics

Approach:

- Ab-initio:
  - Bulk constituents
  - Small ideal superlattices
- Map ab-initio to tight binding (binaries and superlattices)
- Current flow in ideal structures
- Study devices perturbed by:
  - Large applied biases
  - Disorder
  - Phonons

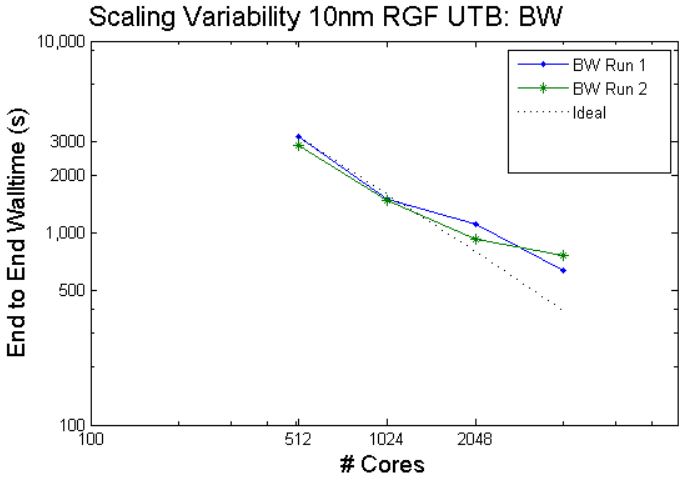
A Journey Through Nanoelectronics Tools NEMO and OMEN

	NEMO-1D	NEMO-3D	NEMO3Dpeta	OMEN	NEMO5
Transport	Yes	-	-	Yes	Yes
Dim.	1D	any	any	any	any
Atoms	~1,000	50 Million	100 Million	~140,000	100 Million
Crystal	[100] Cubic, ZB	[100] Cubic, ZB	[100], Cubic,ZB, WU	Any Any	Any Any
Strain	-	VFF	VFF	-	MVFF
Multi-physics	-				Spin, Classical
Parallel Comp.	3 levels 23,000 cores	1 level 80 cores	3 levels 30,000 cores	4 levels 220,000 co	4 levels 100,000 cores

Scaling

- NEGF Quantum Transport simulations
  - » Still undergoing capability improvements
  - » nearly ideal scaling up ~100 nodes
    - ✓ MPI overhead
  - » Implementation not previously optimized
  - » Variety of scaling issues resolved
- Custom profiling tool
  - » Tic tocs for timing and memory
  - » Web interface



OVERALL Time | OVERALL Memory | Process #0 | Process #1 | Process #2

Plot time for function with name

You are looking at information from the MPI process : 1
Show tic-tocs with time % at least:

Show tic-tocs with a diff peak memory that is at least this percentage of the max diff peak memory:

And title including: | Expand all | Collapse all | Download as CSV

Name	Calls count	Total time	% of total	% of total	Peak toc m...	Peak tic me...	System tic ...	System toc ...	Flops/s	diff_peak_memory
NEMO5_OVERALL	1	1483.48	100%		998.9	0	0	998.9	138.1	9.989e+02M
Module("")::Module	5	0.000311136	0%		64.46	64.46	64.46	64.46	0	0.000e+00K
Nemo("static")::close	1	2.20104	0.148%		998.9	998.8	998.8	998.9	0	1.240e+02K
Nemo("static")::init	1	0.233154	0.016%		54.63	0	0	54.63	0	5.463e+01M
Nemo("static")::init_materials	1	0.66888	0.045%		63.71	54.67	54.67	63.71	0.1095	9.035e+00M
NonlinearPoisson("")::NonlinearPoisson	1	0.0000529...	0%		63.74	63.74	63.74	63.74	0	0.000e+00K
Poisson("")::Poisson	8	0.000341177	0%		63.91	63.91	63.91	63.91	0	0.000e+00K
Poisson("Static")::create	8	0.000416756	0%		63.91	63.91	63.91	63.91	0	0.000e+00K
PoissonEquationInterface("")::PoissonEquationInterface	9	0.000421524	0%		63.91	63.91	63.91	63.91	0	0.000e+00K
Simulation("Propagation_Parallelizer")::init	1	0.00925684	0.001%		68.72	68.68	68.68	68.72	0	4.800e+01K
Simulation("Reinit")::init	1	0.238392	0.016%		72.21	68.73	68.73	72.21	0	3.480e+00M
Simulation("Transformation1")::init	1	0.005018	0%		67.68	67.67	67.67	67.68	0	1.200e+01K
Simulation("adaptive_grid_generator")::init	1	0.00143719	0%		68.73	68.72	68.72	68.73	0	1.200e+01K

Revising ITRS Projections

SOLUTIONS TO SCALING ISSUES FOR ULTRASCALED MOSFETS THROUGH PREDICTIVE MODELING

Mehdi Salmani Jelodar, SungGeun Kim, Kwok Ng, Gerhard Klimeck

PURDUE UNIVERSITY

msalmani@purdue.edu

ITRS roadmap predicts semiconductor industry standards

ITRS Quick facts

- Worldwide joint effort
- Since 1998

→ Ensures cost-effective advancements in ICs

→ More than 1000 engineers/ scientists worldwide

→ Most successful roadmap

ITRS Groups

- 16 chapters
- **Process integration and Device and Structures (PIDS)**
- System and drivers
- Lithography
- Test
- Packaging, …

"so that Moore's Law proposition can continue…"

ITRS identifies technological **challenges and needs** for the semiconductor industry over the **next 15 years**

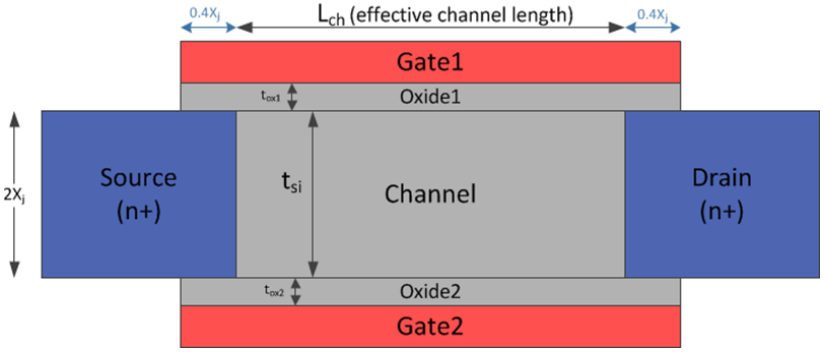
PIDS results

PIDS Scaling Summary

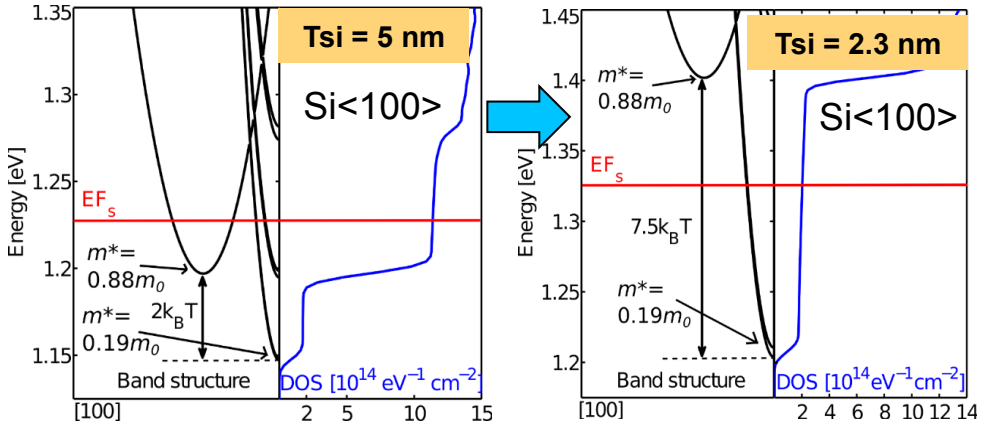
→ Projecting MOSFET scaling geometry such as Leff, V_{DD} and EOT for next 15 years

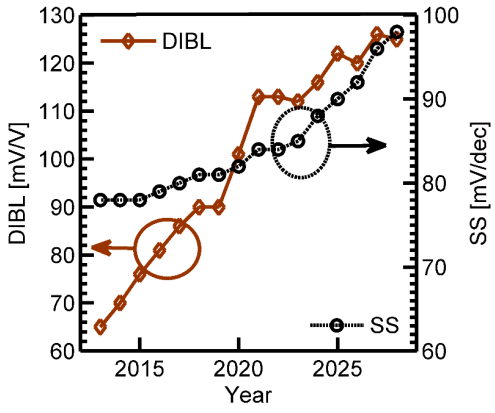
→ DIBL and SS for devices with Leff reduction increase (degrade)

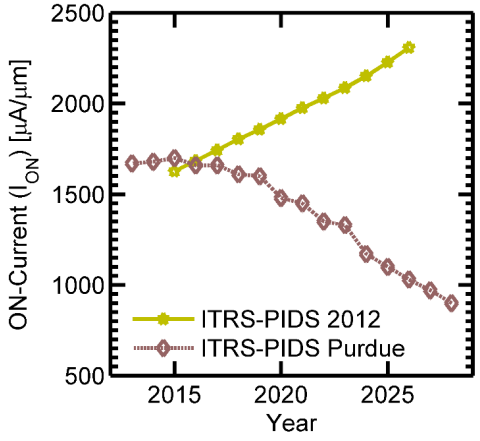
→ Device speeds with calculated current by quantum transport TCAD increase by ideal (8% increase per year) up to 2023.



Scaling Impacts on Performance





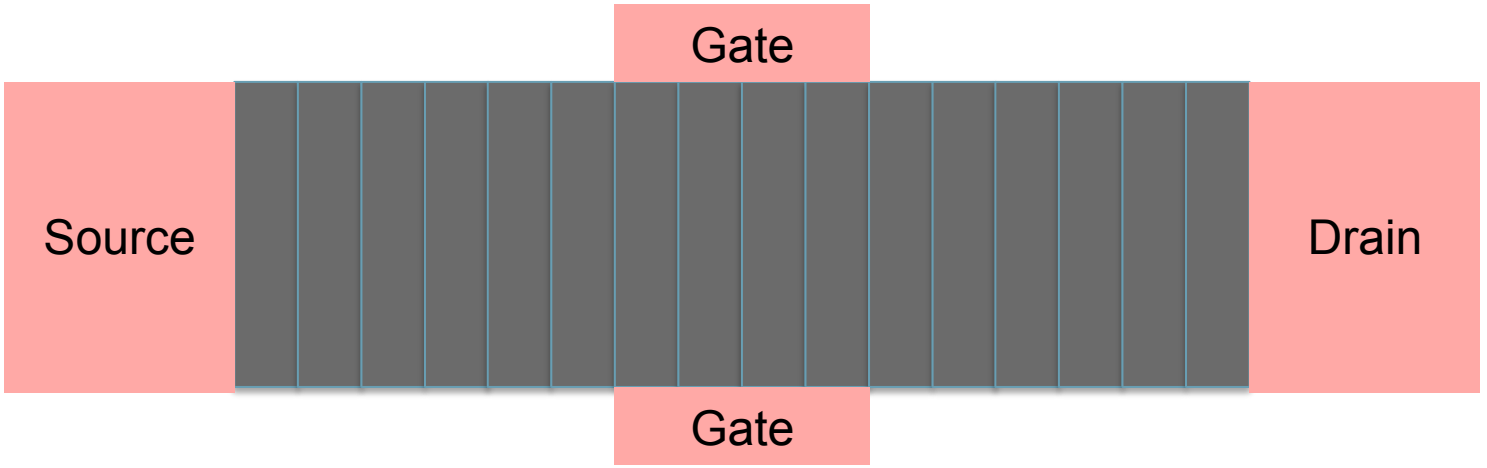


Non-Equilibrium Green's Functions (NEGF)

- Observables (current, charge density etc.) require retarded Green's function

$$G^R = (E - H_0 - \Sigma^R)^{-1}$$

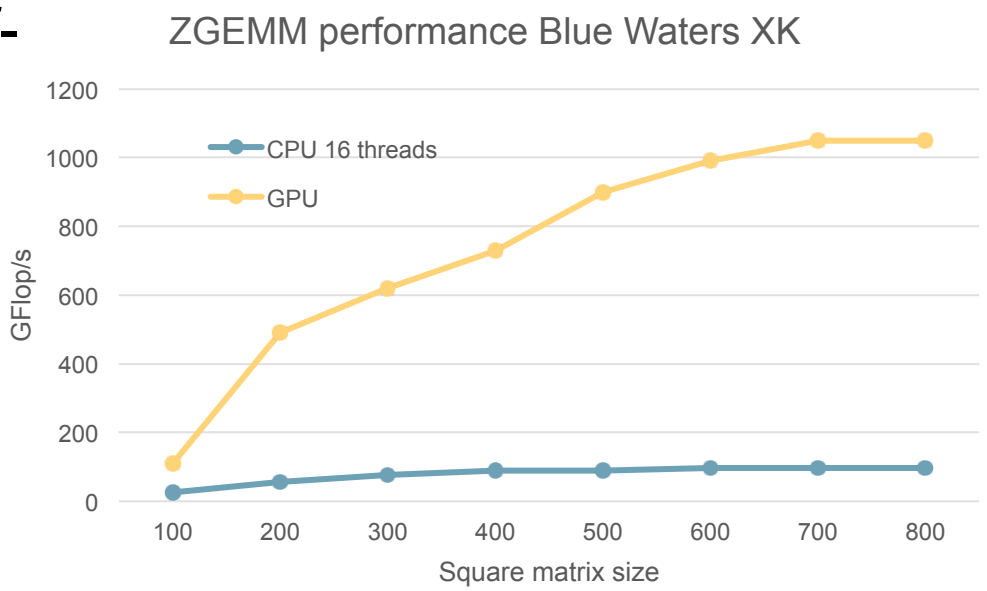
- Typical Tight binding Hamiltonian size ~ 10 Million x 10 Million
- Σ^R – Boundary conditions from contacts (source and drain)
- Naïve inversion: (**RAM ~N^2, Time ~N^3**)
- Device is split into slabs: Invert Hamiltonian for the slabs.

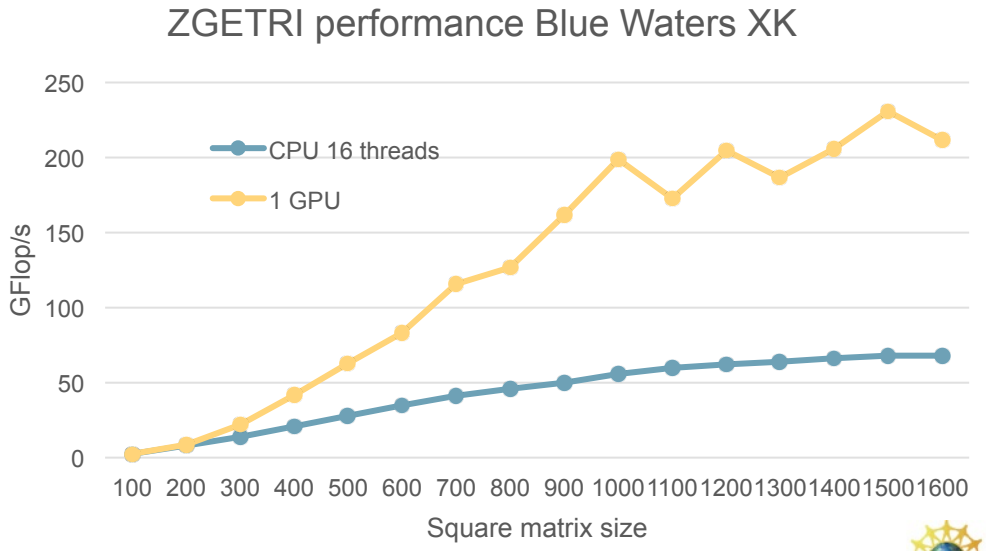


Why GPU acceleration?

- Sancho Rubio method for Self-Energy of contacts
- Computation per iteration
  - » 6 dense matrix multiplications
  - » 1 dense matrix inversion

- RGF approach for solving NEGF in device
- Computation per iteration
  - » 4 Sparse-Dense matrix multiplication
  - » 4 Dense-Dense matrix multiplication
  - » 1 Dense matrix inversion

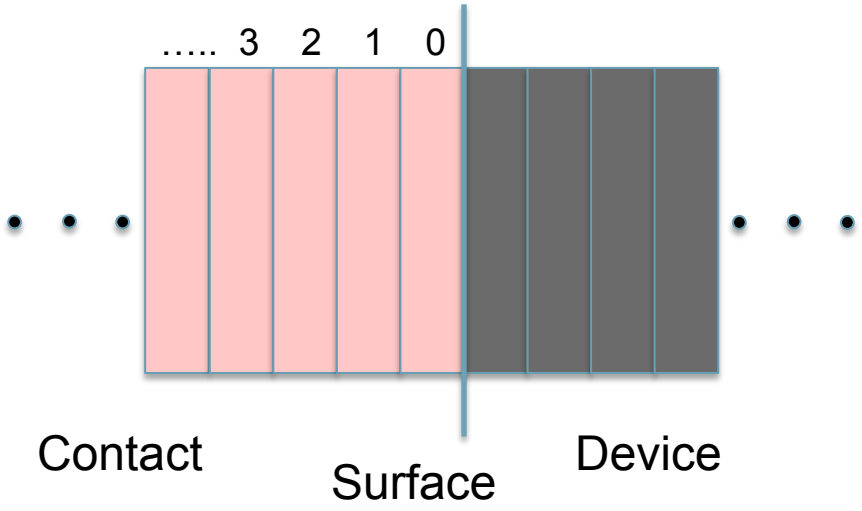




Sancho Rubio

- Sancho Rubio method for Self-Energy of contacts
  - » Accounts for bulk effects
  - » *Semi-infinite* contact divided into layers
  - » Only adjacent layers connected
  - » Alternate layers iteratively eliminated

Uses cuBLAS, cuBLAS-XT, MAGMA libraries



- Matrix algebra per iteration
  - » 6 dense matrix multiplications
  - » 1 dense matrix inversion
- Ideal for offloading to GPUs.
- Input: 3 sparse matrices. Output: 1 dense matrix
- Storage: 7 dense matrices

Recursive Green's function (RGF)

- RGF approach for solving NEGF in device
- Comprises of
  - » Sparse-Dense matrix multiplication
  - » Dense-Dense matrix multiplication
  - » Dense matrix inversion

Uses cuBLAS, cuSPARSE, MAGMA libraries

- Device Hamiltonian in the matrix form:

D-ε	*t*	
t	*D-ε*	*t*
	t	*D-ε*

D: Diagonal block (complex)

t: Coupling block (complex)

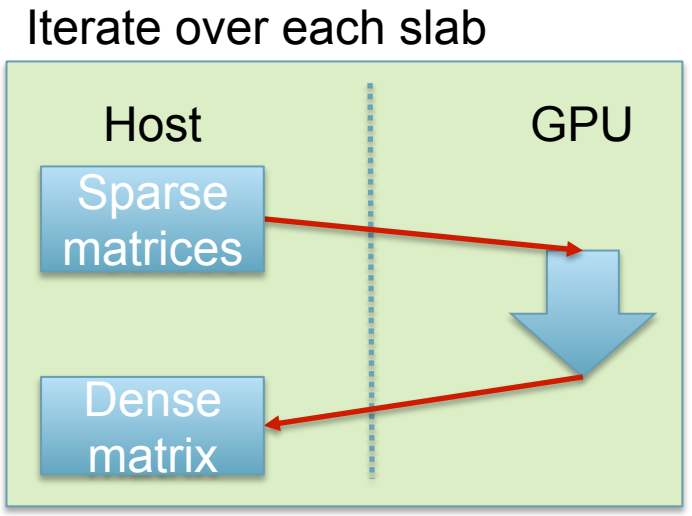
ε: Energy

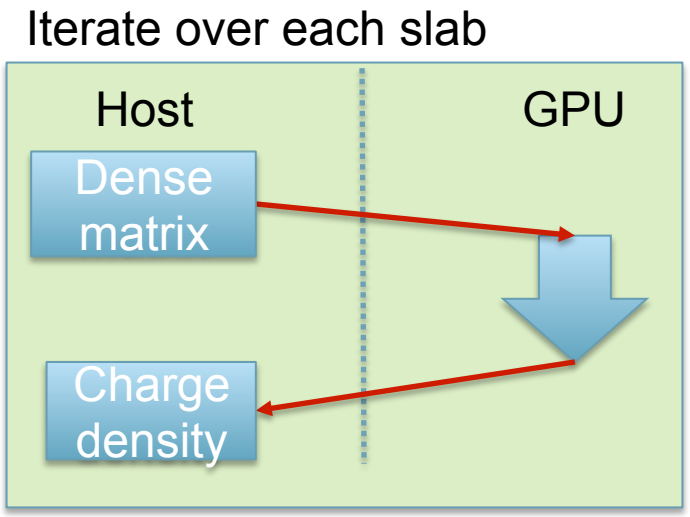
Diagonal and coupling blocks dependent on Energy and momentum (E, k) tuple

Recursive Green's function (RGF)

- Divided into 2 parts: Forward and Backward RGF

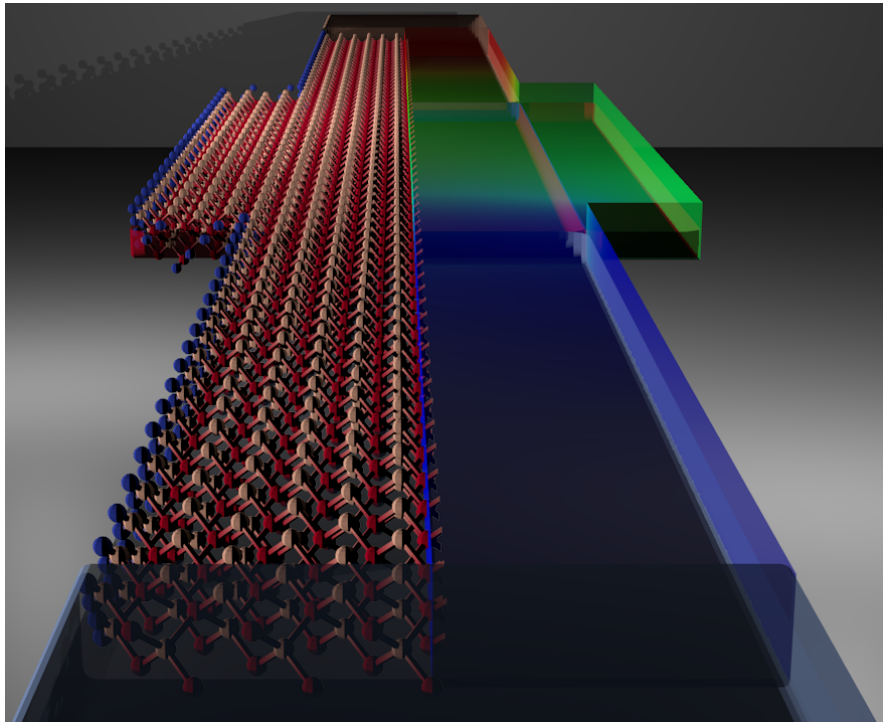
- Forward RGF
  - » Input is device Hamiltonian
  - » Each iteration generates one dense matrix
  - » Storage on host
  - » Result is current
- Backward RGF
  - » Each iteration uses one dense block
  - » Result is orbital resolved charge density
- Asynchronous data transfer using pinned memory

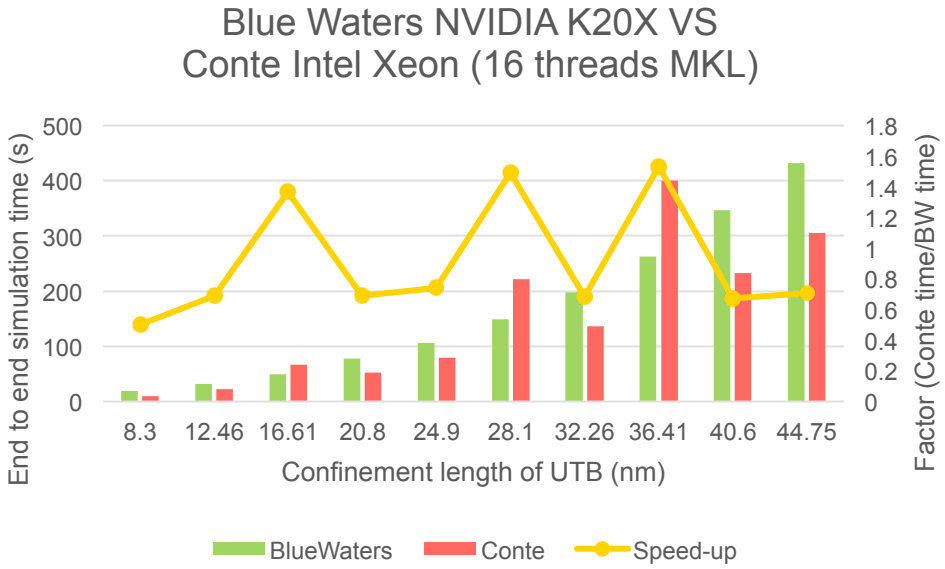


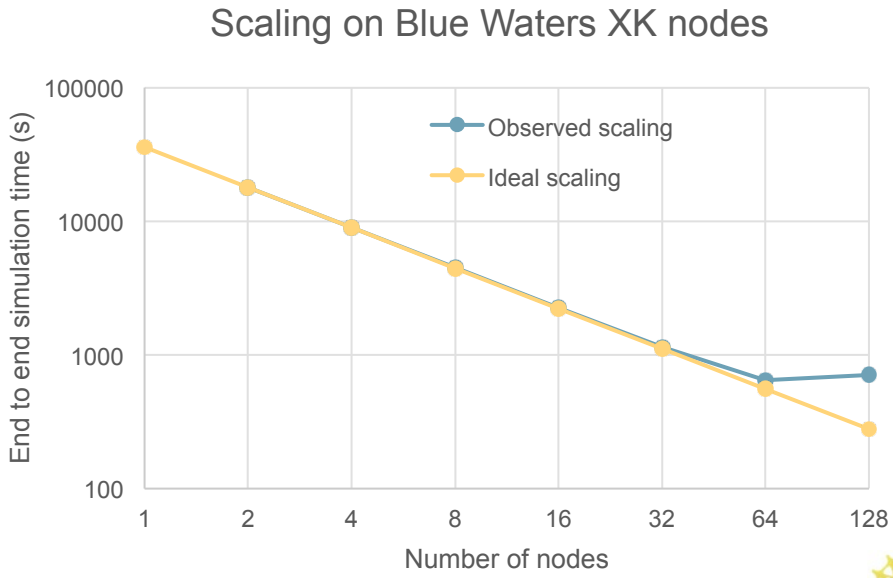


Scaling Results

- Tested scaling of Ultra Thin Body (UTB) transistors.
- End to end simulation comparison, Blue Waters vs Conte



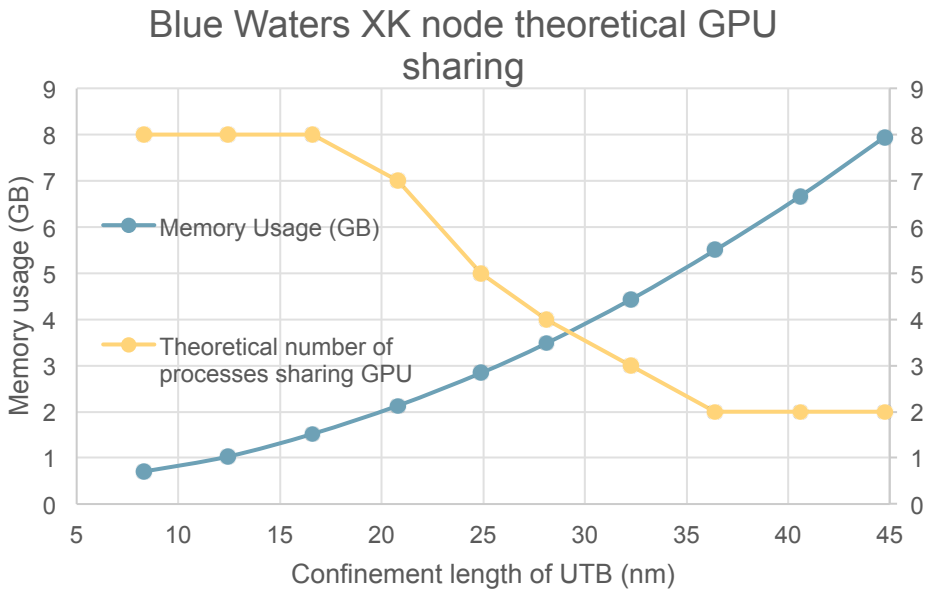




Future work

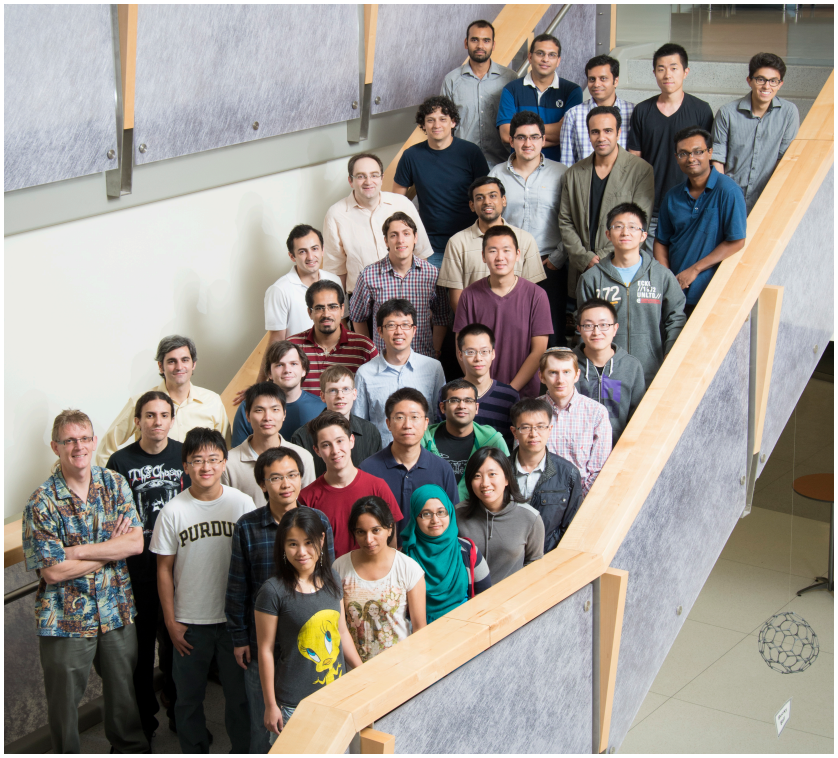
- Sancho Rubio algorithm
  - » Matrix sizes exceed 11000x11000 for large problems
  - » Need to test heterogeneous multiplication in cuBLAS-XT
  - » CUDA 6.0 available in September 2014 on BlueWaters
- Multiple processes on a XK node
  - » Share GPU using CUDA MPS for small problems
  - » Investigate MAGMA memory error with CUDA MPS

Why Blue Waters?



iNEMO Group

- PI: Gerhard Klimeck
- 3 Research Faculty: Tillmann Kubis, Michael Povolotskyi, Rajib Rahman
- Research Scientist: **Jim Fonseca**
- 2 Postdocs: Bozidar Novakovic, Jun Huang
- Students: Kyle Aitken, Tarek Ameen, Yamini Bansal, Jose Bermeo, James Charles, Chin-Yi Chen, Fan Chen, Sicong Chen, Yuanchu Chen, Rifat Ferdous, Jun Zhe Geng, Yu He, Yuling Hsueh, Jun Huang, Hesameddin Ilatikhameneh, Zhengping Jiang, Daniel Lemus, Pengyu Long, Saumitra Mehrotra, Daniel Mejia Padilla, Kai Miao, Samik Mukherjee, Ahmed kamal Reza, Santiago Rubiano, **Harshad Sahasrabudhe**, **Mehdi Salmani Jelodar**, Prasad Sarangapani, Saima Sharmin, Yaohua Tan, Yui Hong Tan, Archana Tankasala, Daniel Valencia Hoyos, Yu Wang, **Evan Wilson**



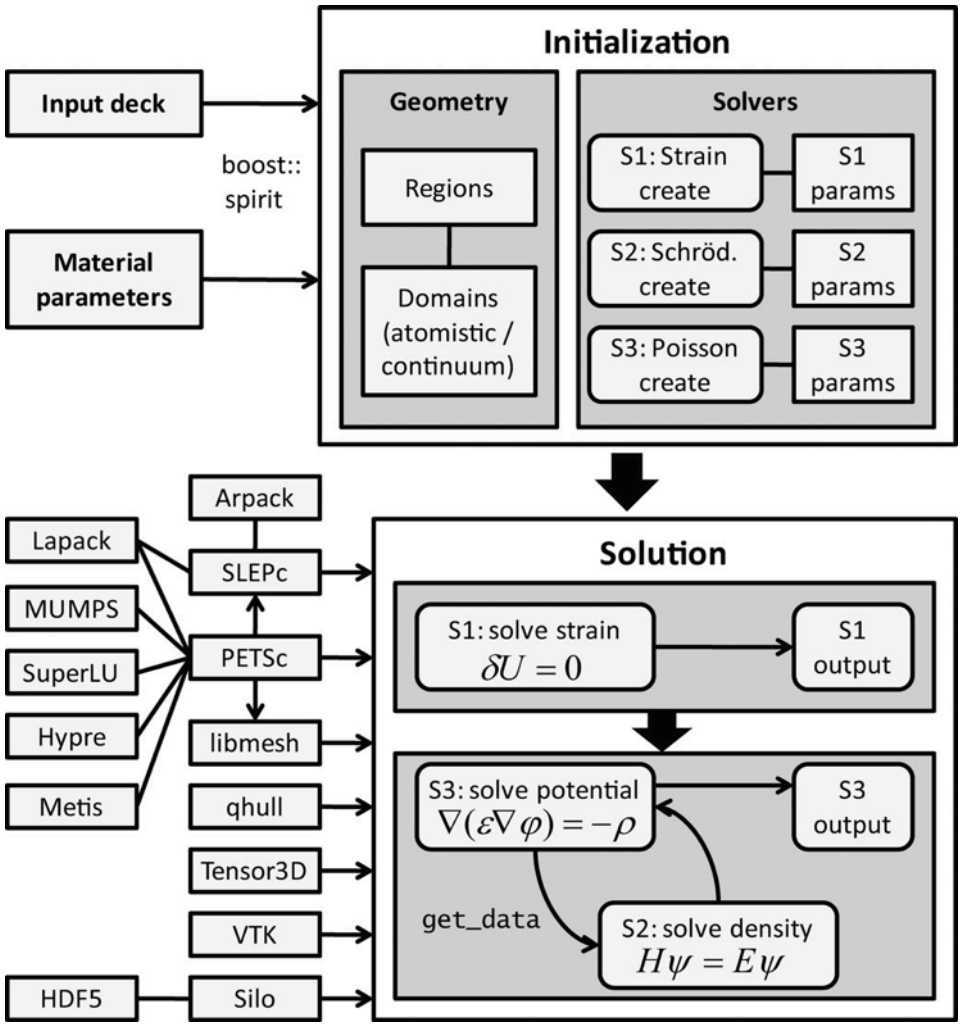
- Ryan Mokos
- PRAC
- GLCPC
- Intel, Samsung, Philips, TSMC

Backup Slides

PURDUE UNIVERSITY

NEMO5 Infrastructure

- Building required libraries
  - » Libmesh, SLEPc, etc.
- PETSc
  - » Portable, Extensible Toolkit for Scientific Computation
  - » Data structure and routines for PDEs
- We use two builds of PETSc
  - » Double
  - » Complex
- Could not use installed version of PETSc
- Also need petsc-dev



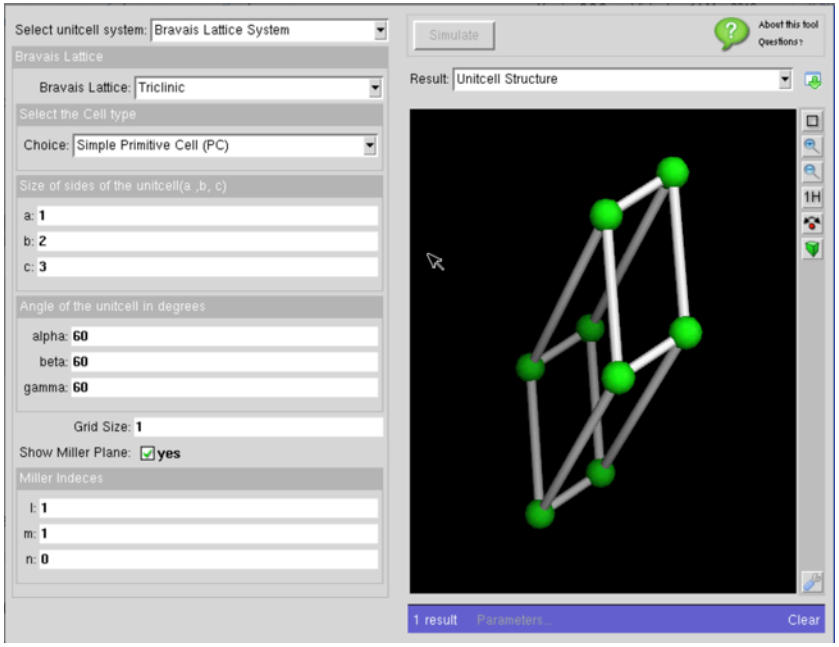
Future Directions

- GPU work
  - » Plans for GPU implementations
    - ✓ Previous plans
    - ✓ CuFFT
      - – Quantum computing
      - – 8x speedup for long range interactions
- OMEN plans
  - » Continue ITRS work
- NEMO plans
  - » New physics models
  - » Optimization
  - » Scalability
  - » GPUs/MICs
  - » Usability

Thanks!

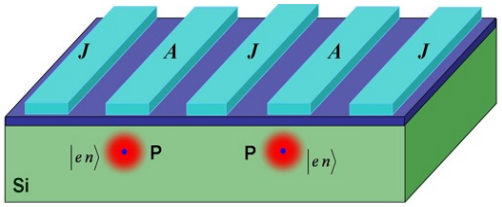
» https://engineering.purdue.edu/gekcogrp/software-projects/nemo5/

» www.nanoHUB.org



Qubits for Quantum Computing

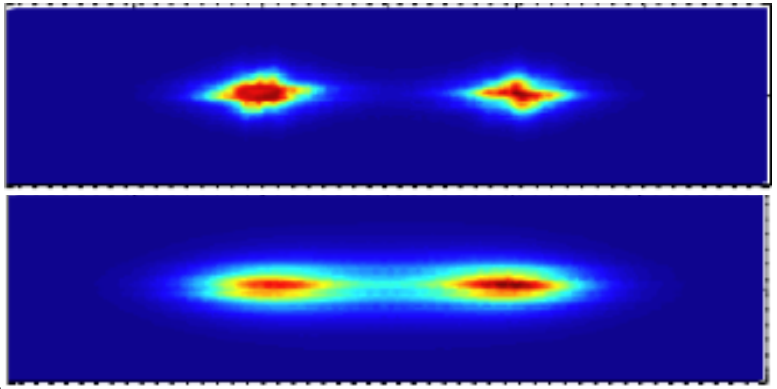
Kane Qubit
P Donor Qubits in Si



In **Quantum Mechanical Analysis** of such a system, the quantum state of an electron is described by a wave-function.

The wave-function is a probability distribution spread over a range of atoms.

Molecular states of the donor impurity system: for single electron



NEMO3D results, Rajib Rahman

Its interaction with any other particle in the system involves **integrating the interaction** over the whole domain.

Simulation of any few-electron systems requires computing the exchange and coulomb energies due to electron-electron interactions.