nanoHUB.org

Network for Computational Nanotechnology (NCN)

NEMO5
NanoElectronics MOdeling

Jim Fonseca

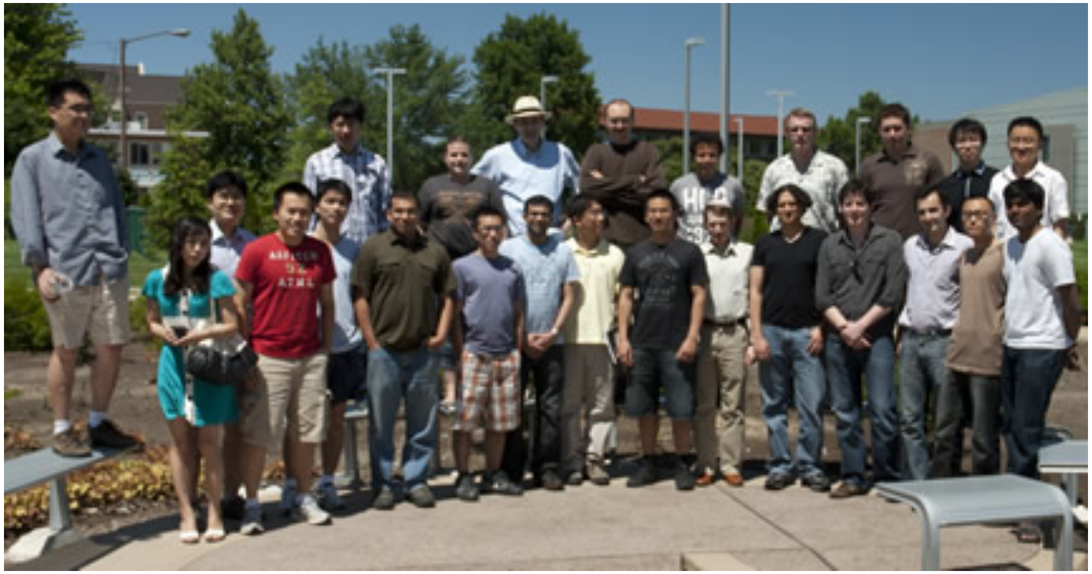
NCSA NEIS-P2 Symposium

May 22, UIUC

PURDUE UNIVERSITY

Klimeck Group

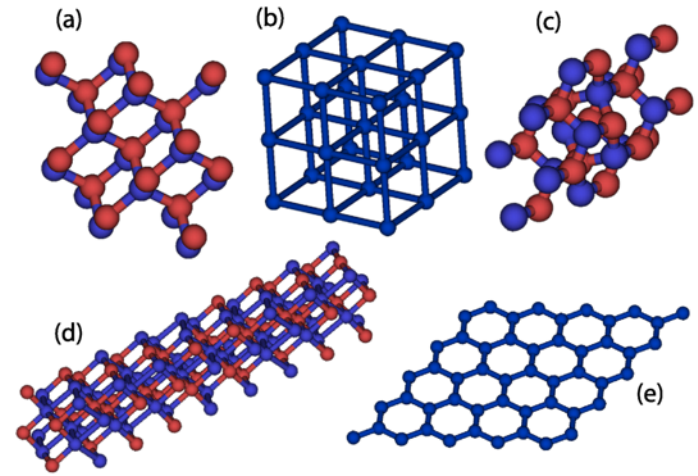
- PI: Gerhard Klimeck
- 3 Research Faculty: Tillmann Kubis, Michael Povolotskyi, Rajib Rahman
- 2 Postdocs: Bozidar Novakovic, Arvind Ajoy
- Students: Kaspar Haume, Yu He, Ganesh Hegde, Hesam Ilatikhameneh, Zhengping Jiang, **SungGeun Kim**, **Daniel Lemus**, Saumitra Mehrotra**,** Daniel Mejia, Samik Mukherjee, **Mehdi Salmani**, Daniel Valencia, Matthias Tan, Yaohua Tan, Evan Wilson, Junzhe Geng, Yuling Hsueh, Kai Miao, Seung Hyun Park, Ahmed Reza, Parijat Sengupta, Saima Sharmin, **Archana Tankasala**, Yu Wang, Pengyu Long, Fan Chen, James Charles

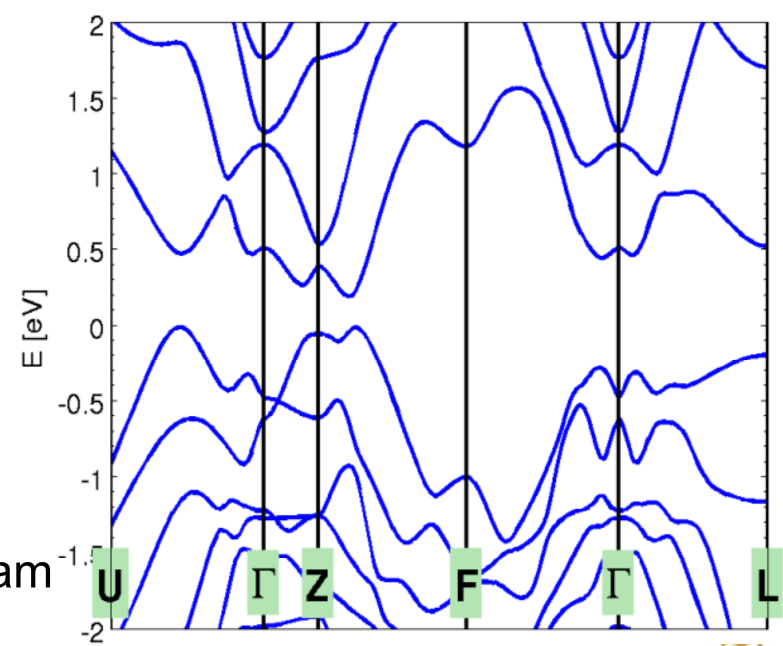


- Basic science in ultra-scaled physics oriented devices such as single atom transistors
- Engineering nanotransistors at the atomistic scale; we are working very closely with industry
- Deployment of apps in nanoHUB that are powered by NEMO5 and are being used so far by over 12,000 users.

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)

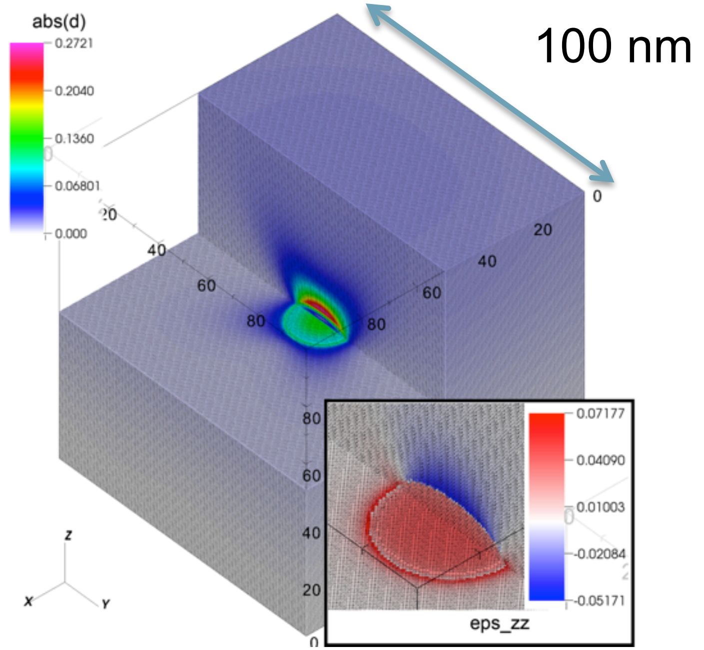


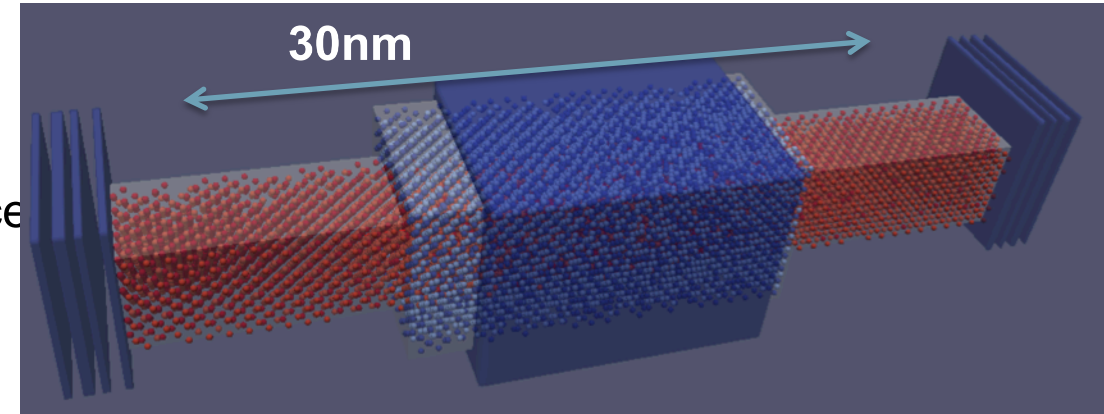


Bi_2Te_3
E-k diagram

NEMO5 – A multiscale simulation tool for nanoelectronic modelling

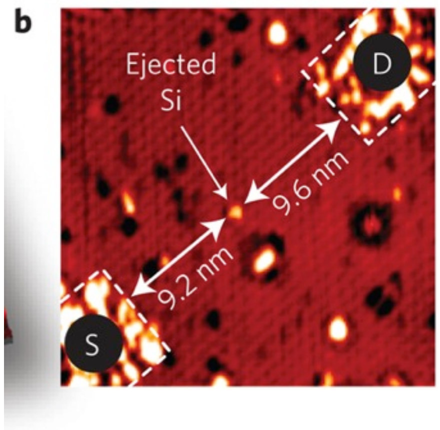
- **Various physical models**
  - Ohmic and Schottky contacts
  - Simple and fast phonon scattering model
  - Rigorous phonon model under development
  - Strain models
    - VFF, Keating
  - Magnetic field under test
- **Solves**
  - Atomistic strain
  - Electronic band structures
  - Charge density
  - Potential
  - Current
- **4-level MPI parallelization**
  - bias, energy, momentum, space





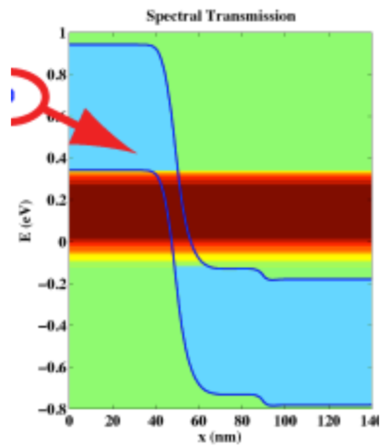
Why atomistic tight binding?

Single atom transistor



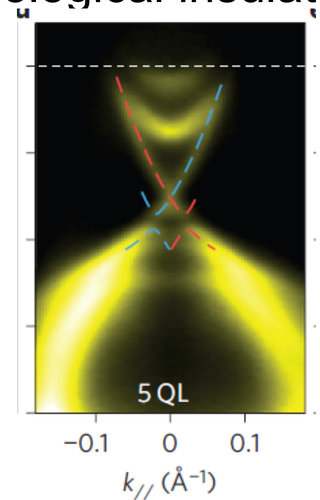
Nature Nanotechnology **7**, 242 (2012)

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.

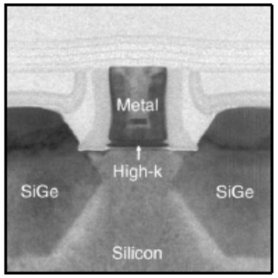
- Band to band tunneling
- Exotic materials (Topological insulators, MoS_2, etc.)
- Band/Valley mixing etc.

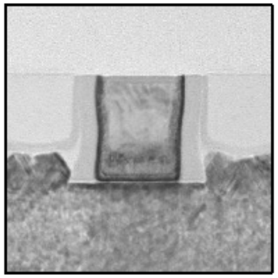
require multi band representations

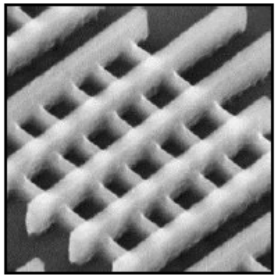
Why non-equilibrium Green's functions?

Device dimensions

2007	2009	2011
45 nm	32 nm	22 nm







http://newsroom.intel.com/docs/DOC-2035

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,…)
and
incoherent scattering (phonons, impurities, rough interfaces,…)

Numerical load of atomic NEGF

Reminder:
NEGF requires for the solution of four coupled differential equations

$G^R = (E - H_0 - \Sigma^R)^{-1}$
$\Sigma^R = G^R D^R + G^R D^< + G^< D^R$
$G^< = G^R \Sigma^< G^A$
$\Sigma^< = G^< D^<$

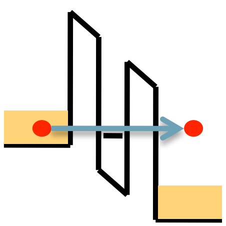
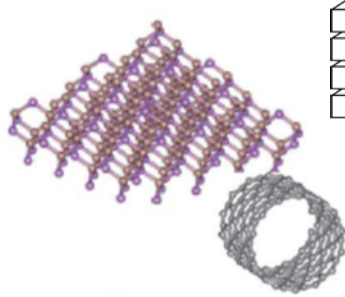
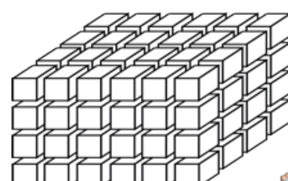
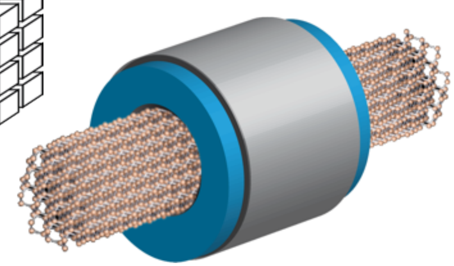
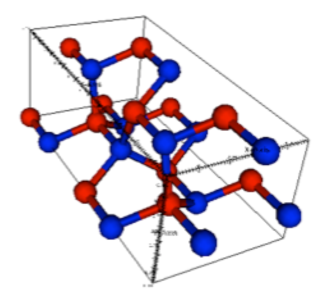
G's and Σ's are matrices in discretized propagation space
(**RAM ~N^2, Time ~N^3**)

Atomic device resolutions can yield very large N (e.g. **$N = 10^7$**)

Huge numerical load is often preventing atomistic device calculations …
…even on supercomputers

3.125%

	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

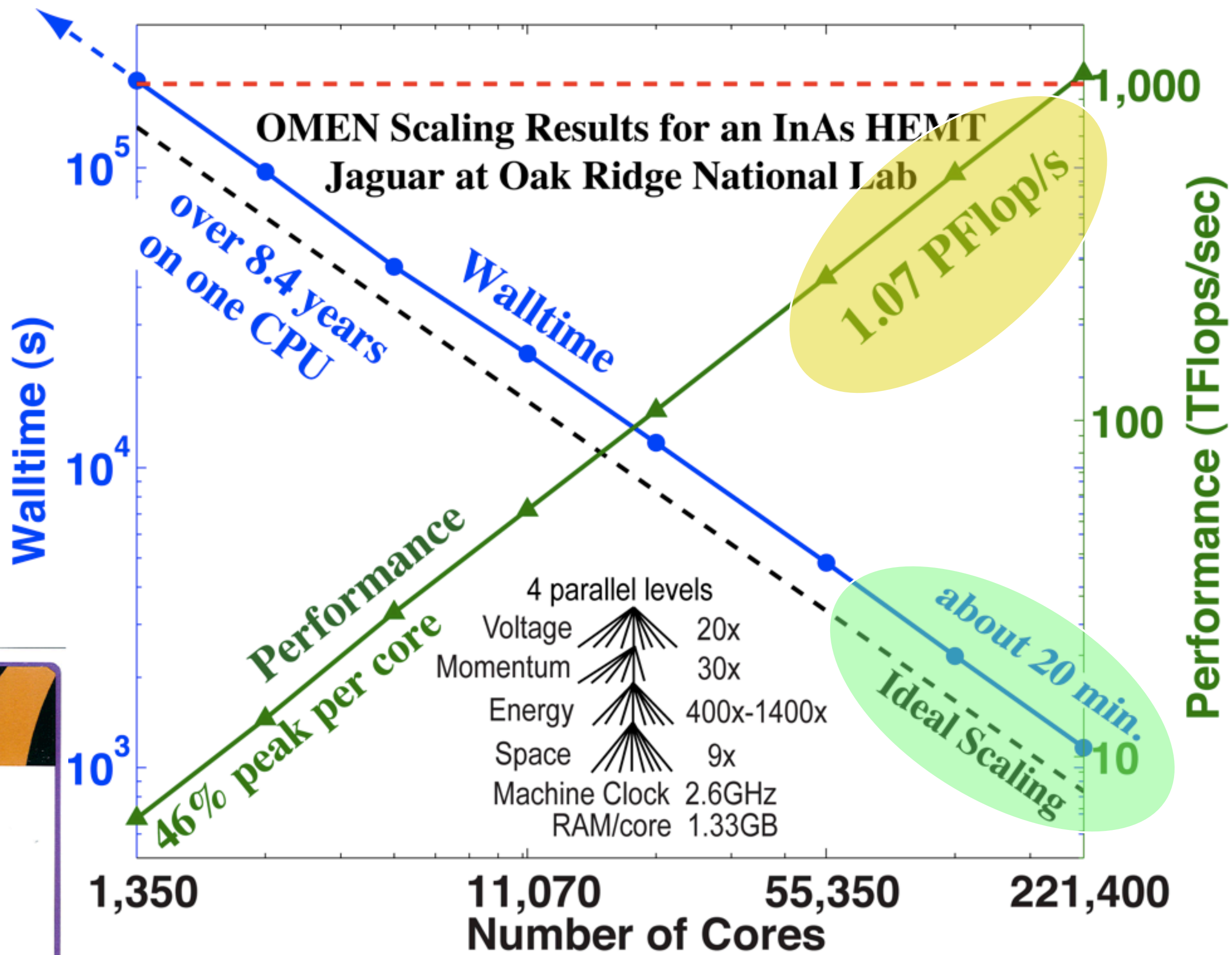
OMEN Scaling to 221,400 Cores
Engineering at the Peta-Scale

Result:

- Highly efficient parallel algorithm, stressing the most advanced resources available today

Impact

- Move from nano-science to nanodevice engineering in minutes
- Unprecedented insight into atomistic device simulation

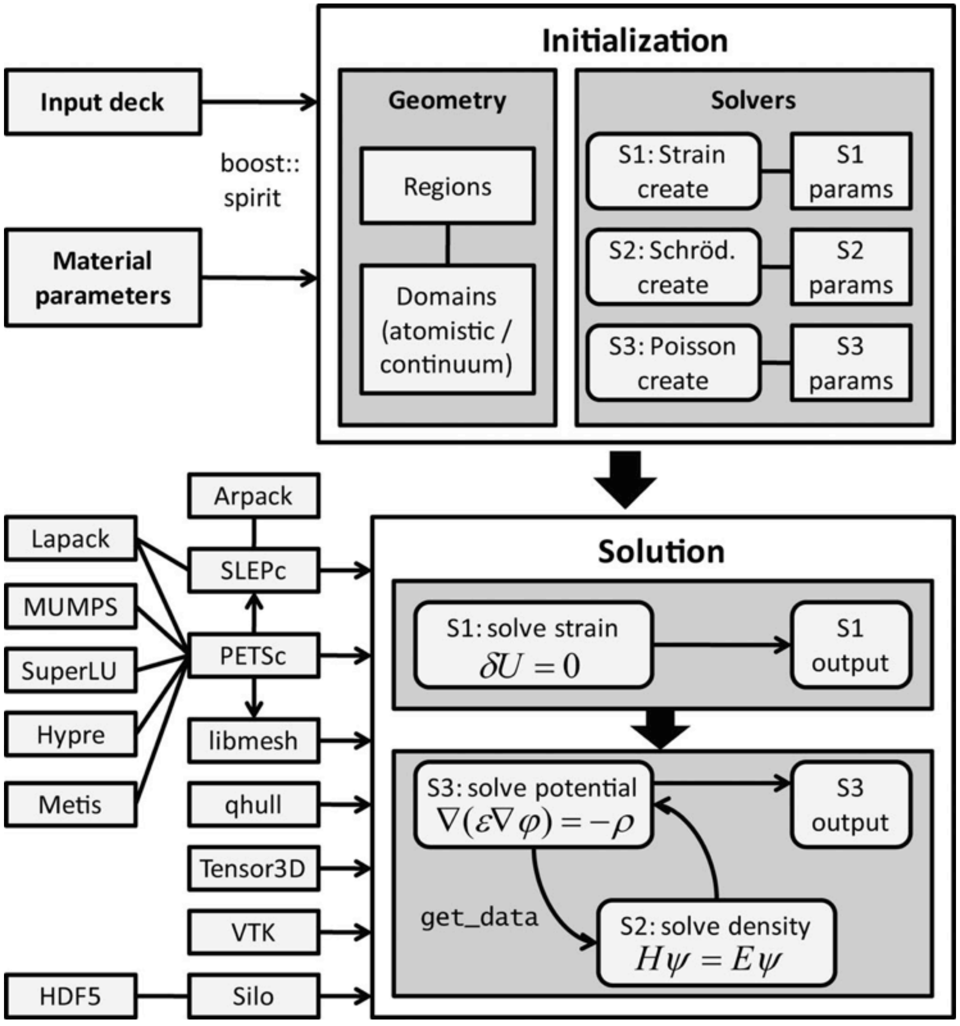


PRAC Planned Work

- GPU Goals
  - Vector-matrix multiplier
  - Lanczos eigenvalue solver
  - Schrodinger (Hermitian matrix algorithms)
  - Low rank approximation (non-Hermitian matrix algorithms)
- Heterogeneous implementation
- Load balancer
- Use PETSc GPU capability

NEMO5 Infrastructure

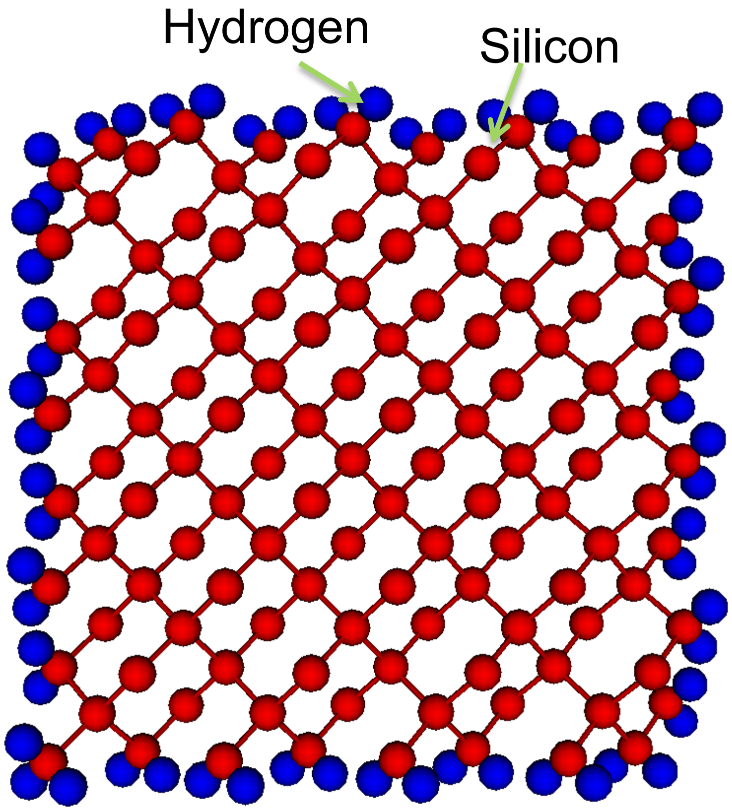
- Building required libraries
  - » Libmesh, SLEPc, etc.
- PETSc
  - » Portable, Extensible Toolkit for Scientific Computation
  - » Data structure and routines for PDEs
- We use two builds of PETSc
  - » Double
  - » Complex
- Could not use installed version of PETSc
- Also need petsc-dev



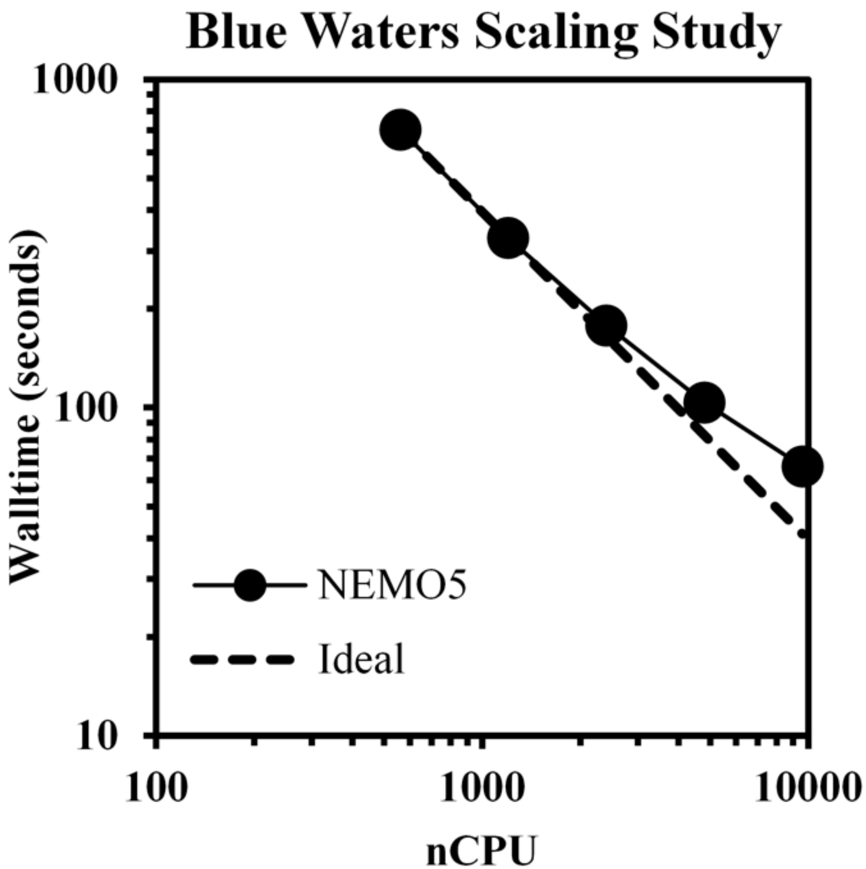
NEMO5 with GPUs

- PETSc
  - » PETSc has some GPU support
  - » PETSc API presents abstraction from CUDA calls and will be used directly in NEMO5
- Segmentation fault occurred upon initializing PETSc
  - » …
  - » **Solved**: The function causing the problem was removed from PETSc and a functional NEMO5 was built with PETSc 3.3
- PETSc 3.3 could not be configured with CUDA support
  - » …
  - » **Solved (May 8th)**: Developer version of PETSc was built with CUDA support
- Current obstacle: Undefined references result when building NEMO5 with developer version of PETSc

Blue Waters Scaling Study



Silicon Nanowire 2nm x 2nm



- Electronics bandstructure calculation for 2 nm x 2 nm silicon nanowire for 9600 k points
- Scaling up to 9600 cores

ITRS (International Technology Roadmap for Semiconductors)

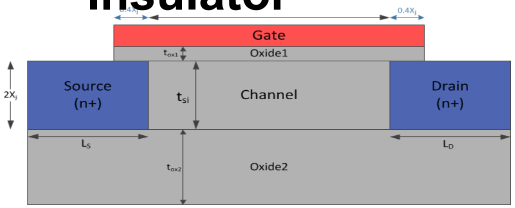
Objective:

- Prediction of next 15 years of technology road map for double gate and Silicon-on-Insulator transistors
- Capturing quantum mechanical effects

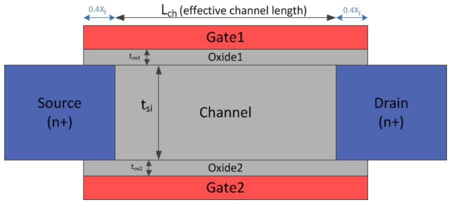
Approach:

- Full-band (tight binding) NEGF
- Series resistance by post-processing
- Scattering with Backscattering method
- Rigorous electron-phonon scattering (for few cases)

Silicon-on-Insulator



Double Gate



Results/Impacts:

- A table for 3 nodes is demonstrated below (for SOI devices) at the end the project it should be extended for next 15 year of scaling SOI and DG devices s
- Tables will be available for all related industry and academia

Year	L_g (nm)	L_{eff} (nm)	V_{DD} (V)	T_{Body} (nm)	T_{OX} (nm)	R_{SD}	ITRS-I_{ON}	N5-I_{ON} (μA/μm)	W/ Scatt	W/ R_{SD} and Scatt	Qinj /cm2	V_{inj}/ (1e7cm/ sec)	N5-SS/DIBL
2013	20.0	16	0.86	4	0.8	298	1475	3890	2200	1475	1e13	2.6	84
2017	14.0	11.2	0.8	2.8	0.7	208	1717	4130	2050	1375	9e12	3.2	83
2020	10.6	8.48	0.75	2.2	0.6	153	1942	4980	2200	1475	8e12	3.7	79

Scaled up to 10,000 cores / 0.5M CPU-hour used and 5-10M CPU-hour is required

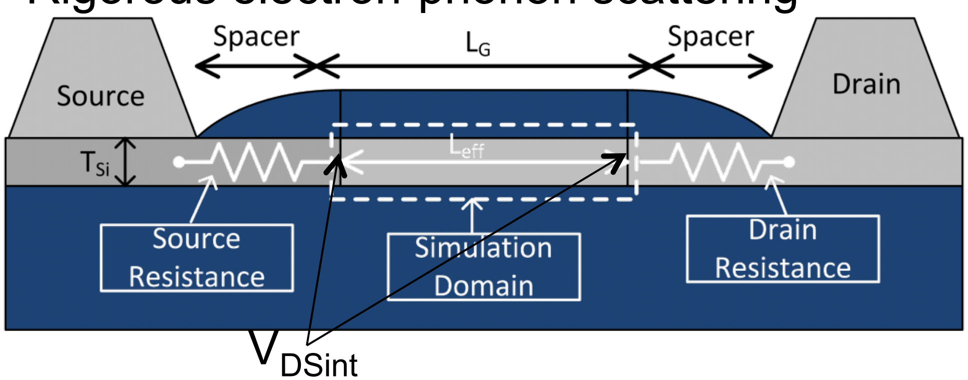
Channel Thickness Effects on ETSOI MOSFETs

Objective:

- Analysis of effects of body thinning in ETSOI
- Series Resistance and scattering effects in ETSOI

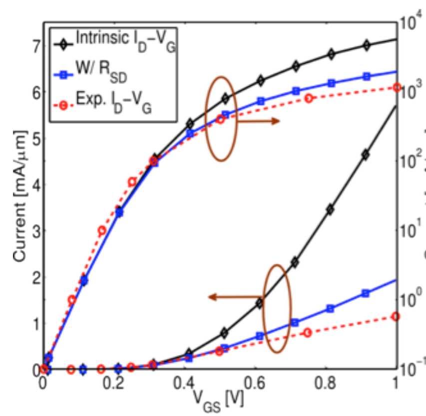
Approach:

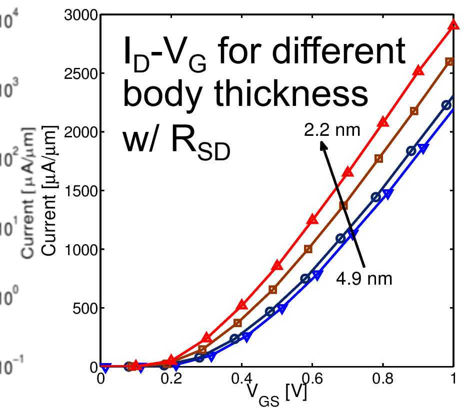
- Full-band (tight binding) NEGF with electron phonon scattering
- Silicon [100], Tsi = 5, 4.4, 3.3 and 2.2 nm, EOT = 0.7nm
- Series resistance by post-processing
- Scattering with Backscattering method
- Rigorous electron-phonon scattering

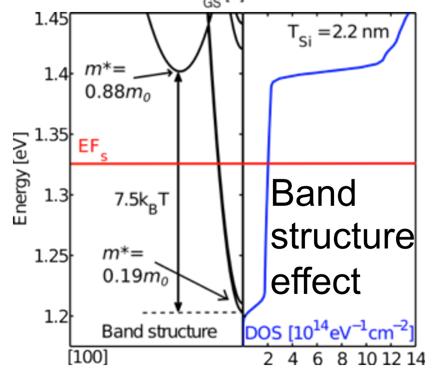


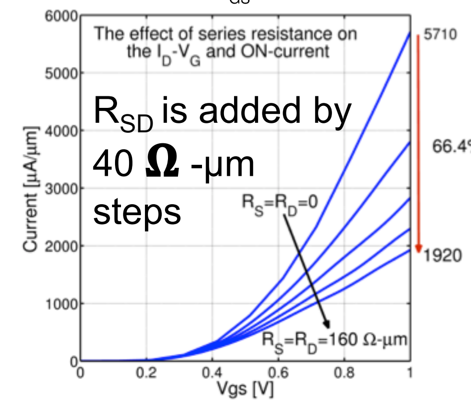
Scaled up to 48,000 cores / ~2M CPU-hour

Results:









- Ballistic ON-current keeps increasing with body thickness reduction (>5nm)
- Parasitic resistance effect is drastic
- Scattering rate increases by body thickness reduction

Impact : Trans. On Elec. Dev. (under prep.)

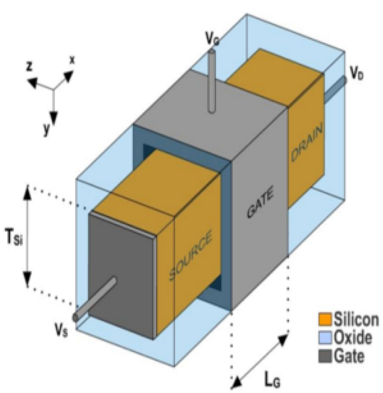
Guideline for Nanowire MOSFETs in the Tunneling Dominant Regime [Lg<12nm]

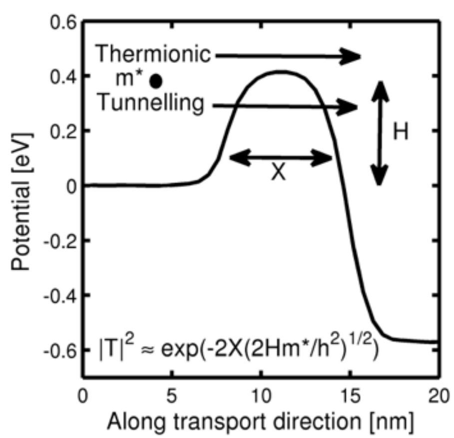
Objective:

- Analysis of nanowires below 12nm to see the tunneling effects and finding the optimum m*

Approach:

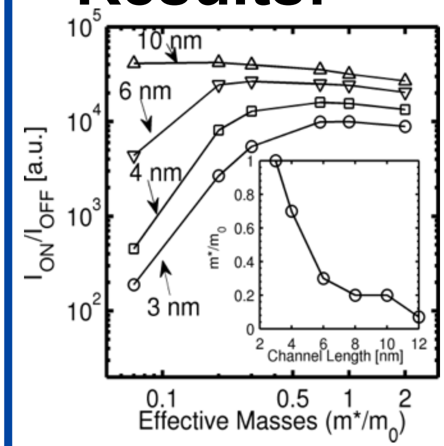
- Real space NEGF in effective mass regime
- Leff = 3, 4, 6, 8, 10 and 12 nm
- m^*/m_0 = 0.07, 0.2, 0.3, 0.7, 1.0
- Square cross-section (5x5nm)
- VDS=0.6 V and EOT = 0.4 nm

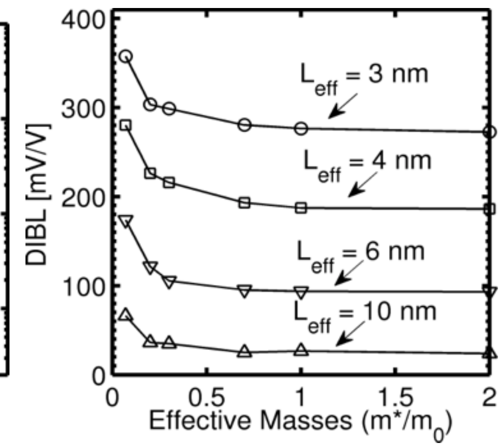




Scaled up to 6400 cores/ ~0.5M CPU-hour

Results:





1. Heavy mass materials:

a. Reduction of tunneling effects (better SS),

b. Improvement of DIBL (due to higher C_Q/C_G)

c. There is a transition point where high mobility materials starts to underperform (10nm and below)

2. There is an optimum effective mass (m*) for each given channel length.

3. Guidelines for identifying required m* for optimal performance for any given L_{eff} down to 3 nm. The optimal m* increases from 0.2 to $1.0m_0$ while L_{eff} reduces from 10 nm to 3 nm.

4. All of the required masses are shown to be engineered with Si.

Impact : Elec. Dev. Lett. (under prep.)

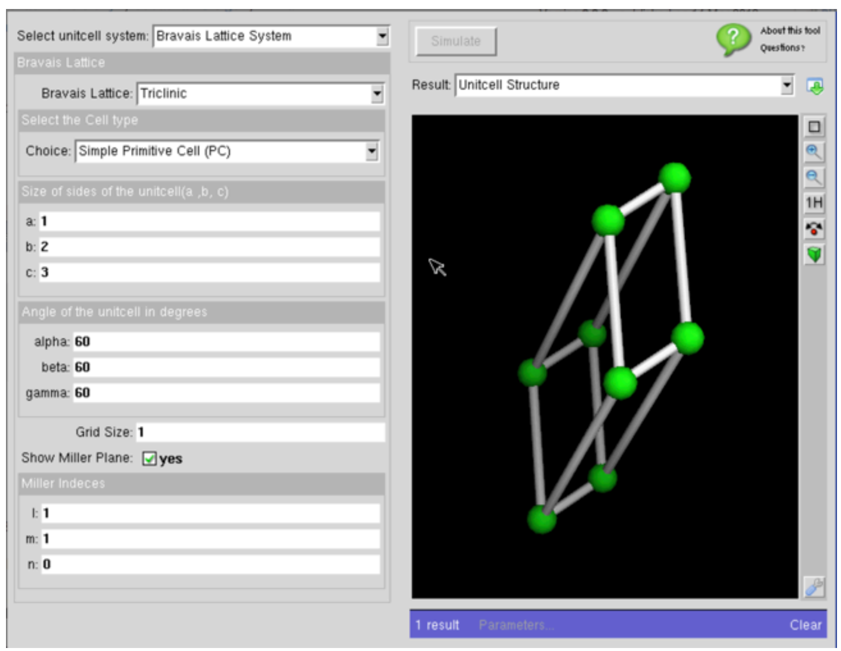
Future Directions

- 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

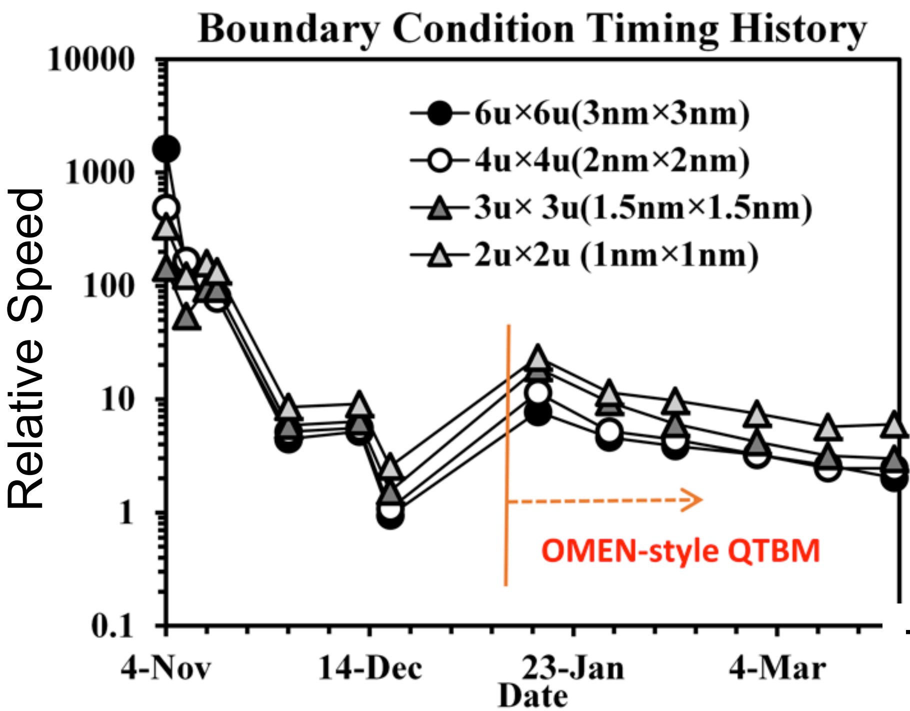
Thanks!

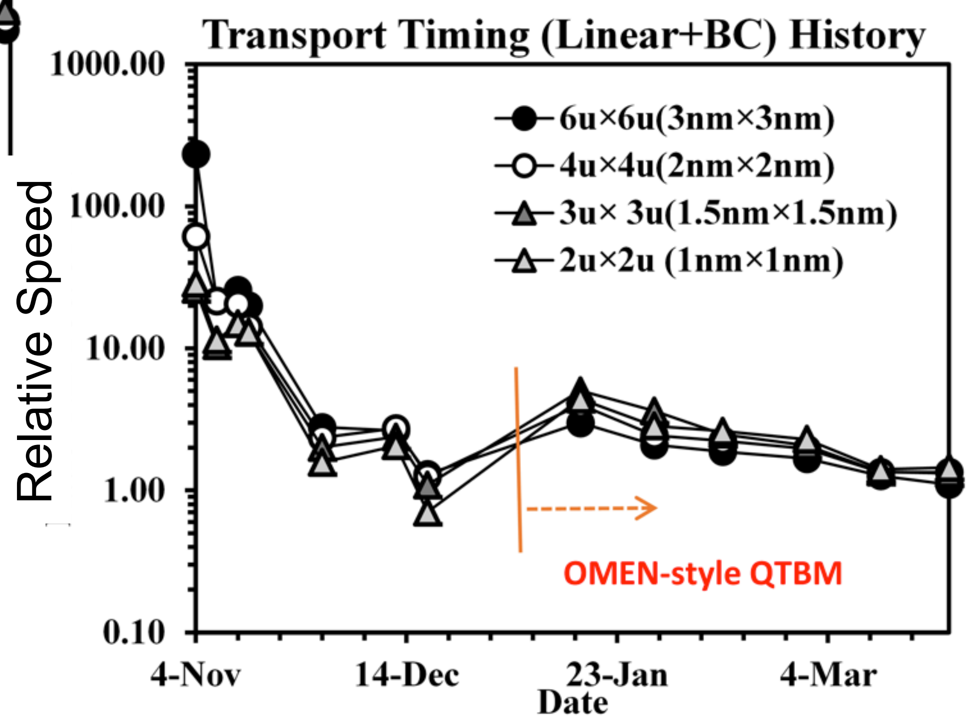
» https://engineering.purdue.edu/gekcogrp/software-projects/nemo5/

» www.nanoHUB.org



NEMO5 Other Work: Quantum Transmission

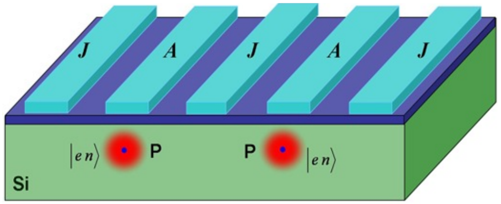




- BC/Transport timing slowly approaching OMEN's timing
- Timing for larger cross-section: almost the same level of OMEN

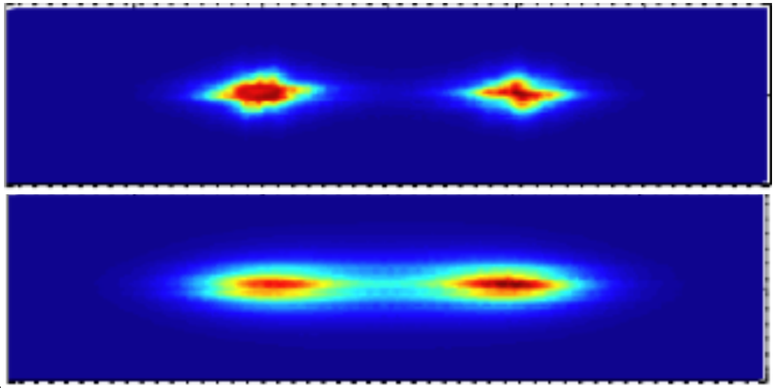
Qubits for Quantum Computing

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.

Computing Long Range Interactions…

…between electrons in a system of N atoms for R different charge distributions or wave-functions:

The interactions are Coulombic or long-range in nature decaying as r^{-1} where r is the distance between the two electrons.

The sum is only conditionally convergent.

Computational effort in simulating such a system involving all pair interactions is proportional to N^2R^2.

Massively parallel processing required.

The DFT Approach

The approaches using Fourier transforms techniques recasts the slowly and conditionally convergent series into:

- a term that converges rapidly in real space
- a term that converges rapidly in reciprocal space
- a constant term.

Algorithms like Ewald summation and Particle-Particle Particle-Mesh method scale as $O(N^{3/2})$ and $O(N \log N)$ respectively.

The complexity of methods using DFT techniques depends on performance of:

Real Space Computations : Pair interactions up to a cutoff distance
SIMD execution
Reciprocal Space Computations : FFTs
For many different distributions

Garland *et. Al*[1] showed that 2D FFT to simulate ultrasound propagation using cuFFT was found to be about 8 times faster than an optimized FFT on CPU.

Also, that implementing batched 2D FFT to effectively utilize the GPU hardware by assigning multiple FFTs to different thread blocks, the performance was almost 16 times faster than the CPU implementation.

Runtime of FFT routine running on GeForce 8800 GTX and its optimized CPU version on one core of a 2.4 GHz Q6600 GPU:

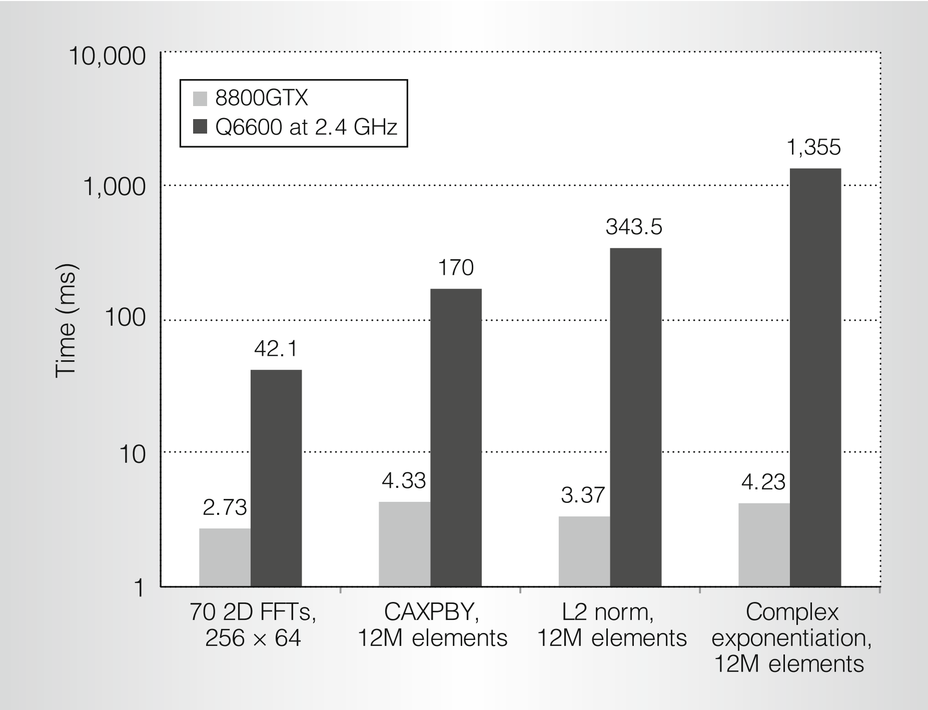


Figure 9. GPU speedup of individual computational routines.

[1]: M. Garland, S.L. Grand, J. Nickolls, IEEE, Parallel Computing Experiences with CUDA