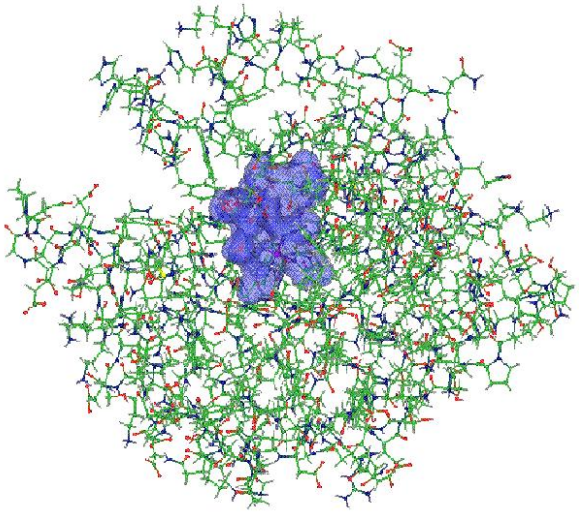
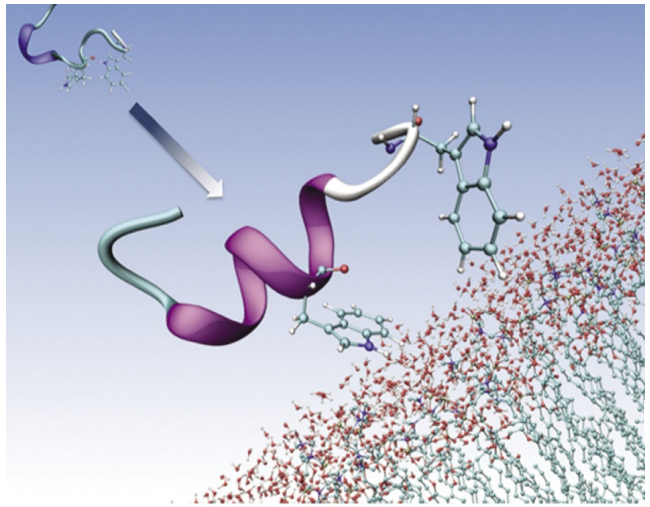
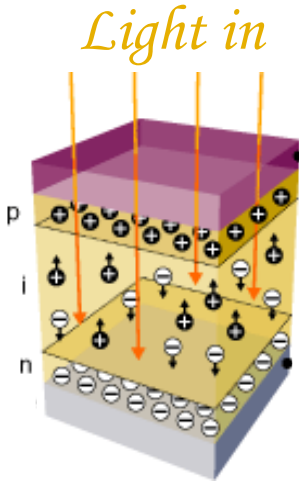
OpenAtom: On the fly ab initio molecular dynamics on the ground state surface with instantaneous GW-BSE level spectra

PIs: G.J. Martyna, IBM; S. Ismail-Beigi, Yale; L. Kale, UIUC;
Team: Q. Li, IBM, M. Kim, Yale; S. Mandal, Yale;
E. Bohm, UIUC; N. Jain, UIUC; M. Robson, UIUC;
E. Mikida, UIUC; P. Jindal, UIUC; T. Wicky, UIUC.

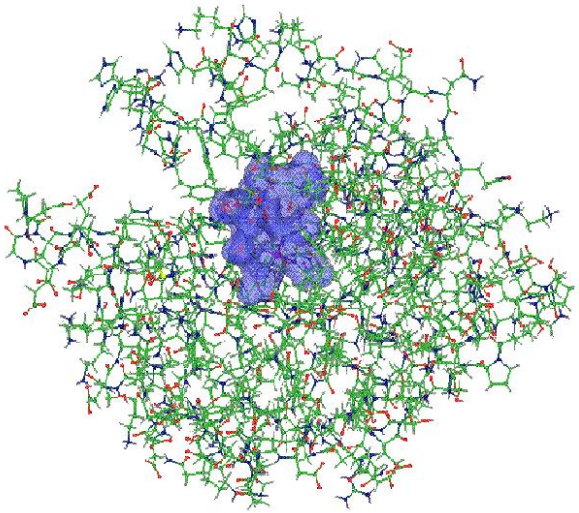


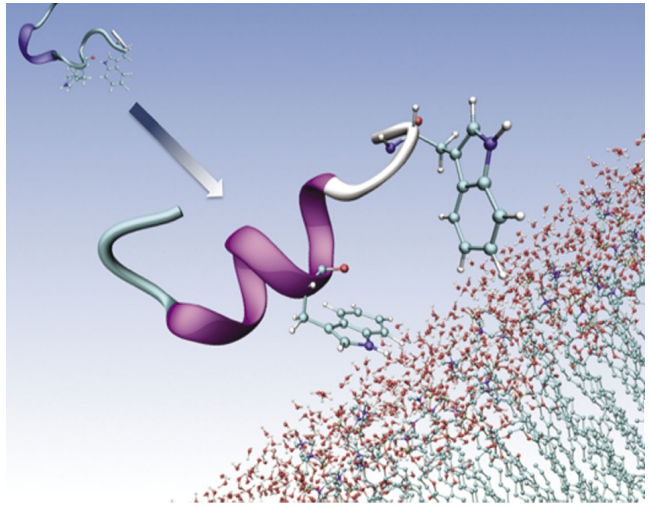


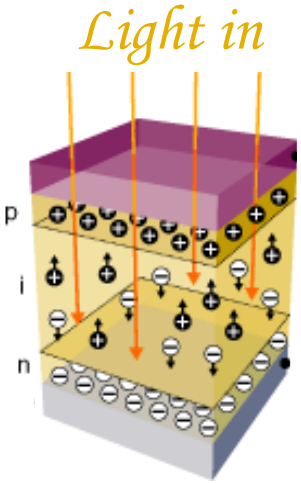


Outline:

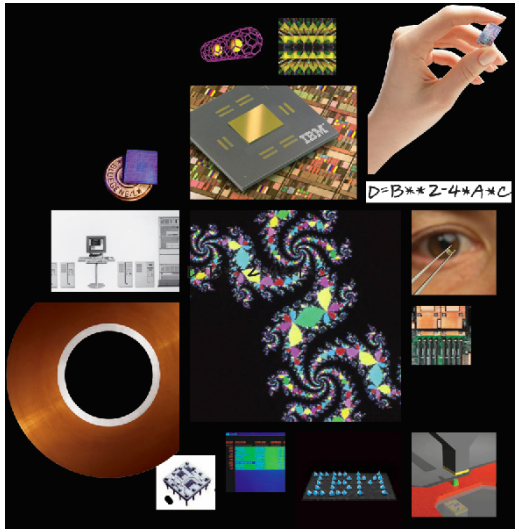
I. Introduction to the OpenAtom project
II. Statistical sampling of complex systems on the ground state surface
III. Optimization under charm++
IV. GW-BSE – a new charm++ application.







IBM : A History of Multidisciplinary Research



Exploratory work: *Yorktown, Almaden & Zurich*

2012 Excimer Laser Surgey (**Nation Medal of Technology**)

2009 Nano MRI

2008 World's First Petaflop Supercomputer

2006 Francis Allen: **1st Female Turing Award winner & 1st Female IBM Fellow**

2005 Cell Architecture

2004 Blue Gene (**National Medal of Technology**)

2003 Carbon Nanotube Transistors

1997 Copper Interconnect Wiring

1994 Silicon Germanium (SiGe)

1987 High-Temperature Superconductivity (**Nobel Prize**)

1986 Scanning Tunneling Microscope (**Nobel Prize**)

1980 RISC

1971 Speech Recognition

1970 Relational Database

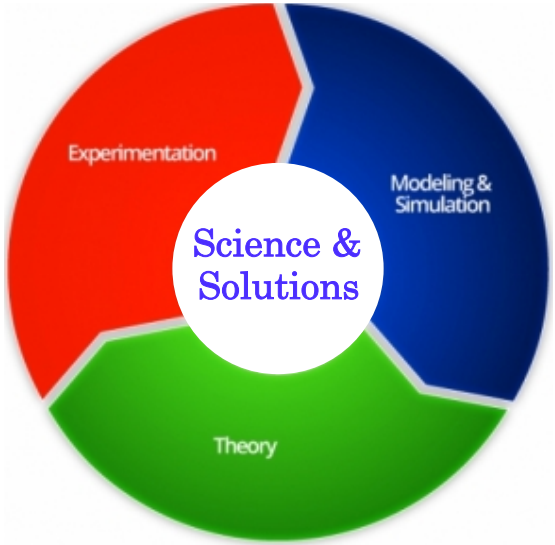
1967 Fractals

1966 One-Device Memory Cell

1965 FFT (Cooley and Tukey).

1957 FORTRAN (Program lang)

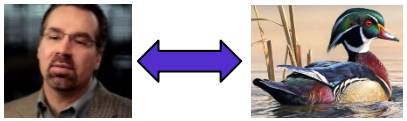
1956 RAMAC (comput. w. mag. disk)



"Treasure wild ducks"
James Watson, Jr.
(from Kierkegaard)

www-03.ibm.com/ibm/history/ibm100/us/en

IBM's Watson Cognitive Computer

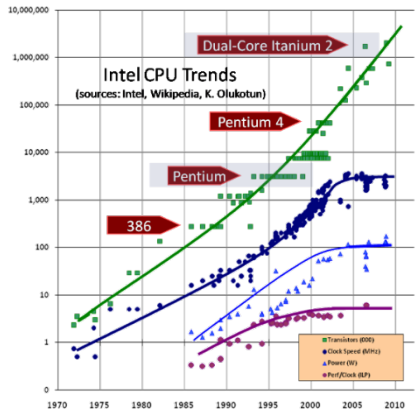


D. Ferrucci Wild Duck

Where is IT today and where is it going?

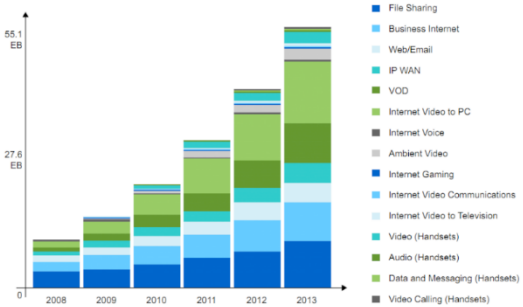
IT industry has driven giant productivity gains in the last 40 years – are we done?

Power: A Cause for Concern



- Power is limiting progress in the IT world – CMOS has reached fundamental limits.
- Data-centric: communication is expensive.
- Dark silicon/on-chip power savings will not deliver enough savings - Microsoft Project Natick – underwater data centers!!!!!! (projectnatick.com)
- Sensor revolution requires on-board low-power computing to preprocess data avoiding power cost of wireless communication.

Wireless Communication and Radar: New demands for high performance

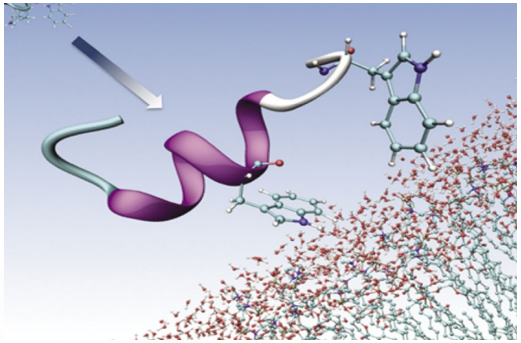


- The spectrum is congested and/or contested.
- Higher performance electronics for agile radars and communications are needed to move forward - RF FPGA's

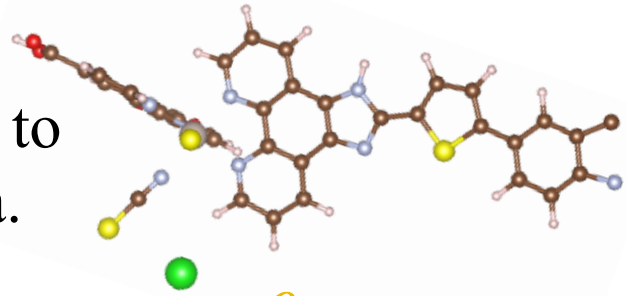
New efficient methodology and implementations required for progress!

Philosophy: Statistical Sampling of Complex Environments is Key to Understanding many Physical Systems across Science and Solutions

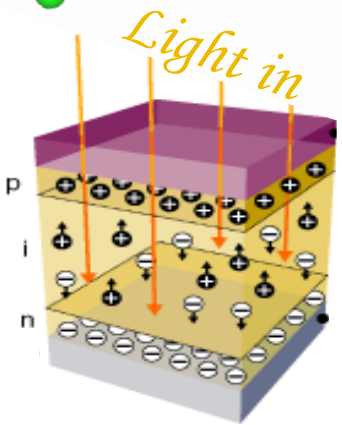
- Biological function is enabled by fluctuations in both the environment and the biomolecules.



- Pollutant detection requires sampling complex aqueous systems and then exporting the results to a GW/GW-BSE app for computation of spectra.

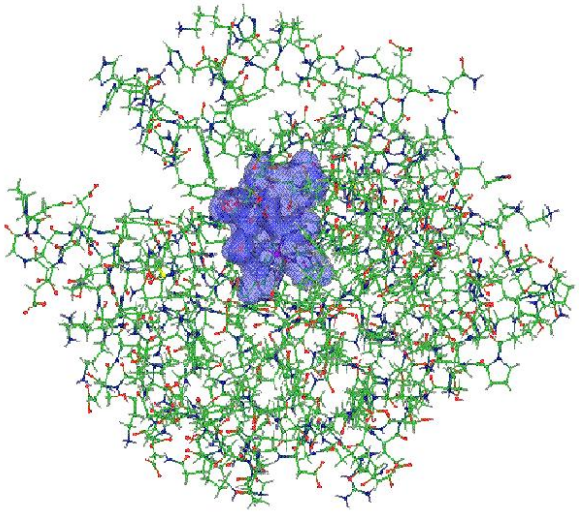


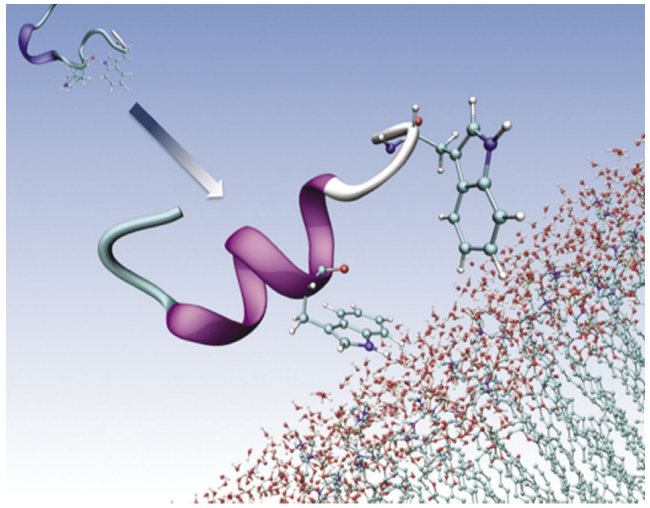
- Understanding chemical reactions in dense arrays requires non-trivial sampling of the full system due to complex many-body reaction paths.

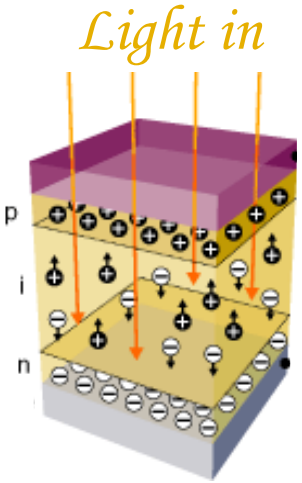


Simulating materials with atomic detail on the Ground State Born-Oppenheimer Surface: Reaching Long Time Scales via Parallel Software and Novel Physics Based Methodology

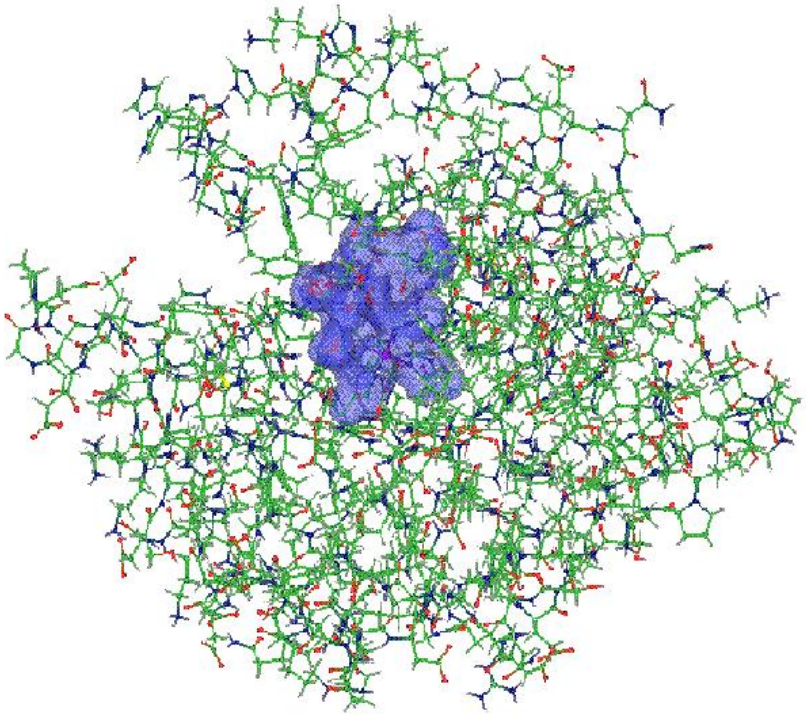
PI: Glenn Martyna, IBM TJ Watson Research Center,
Honorary Professor of Physics, University of Edinburgh,
2016 IPAM Senior Fellow
Postdoc: Qi Li, IBM TJ Watson Research Center

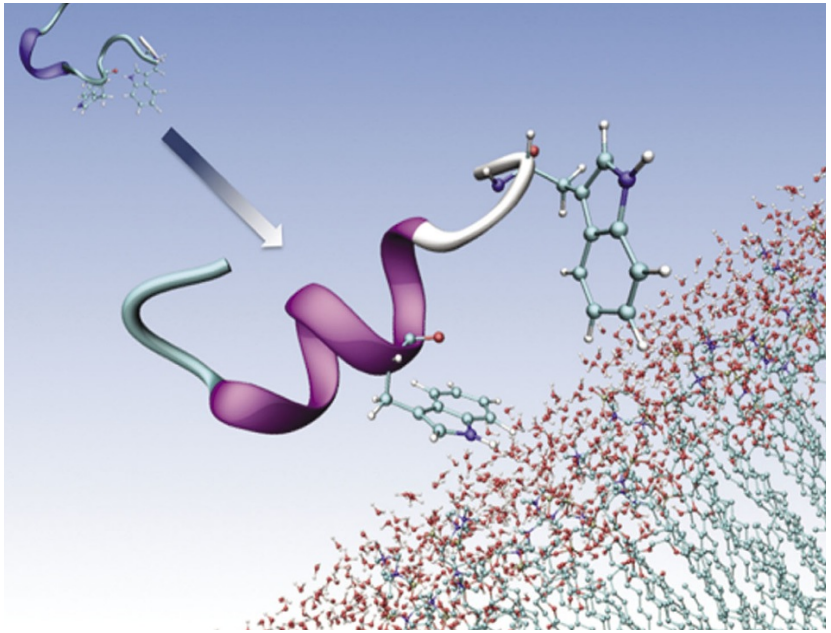


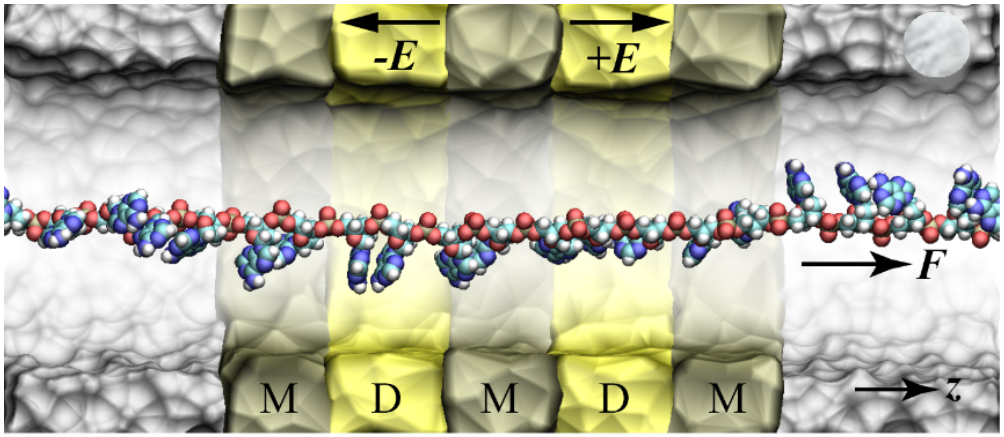


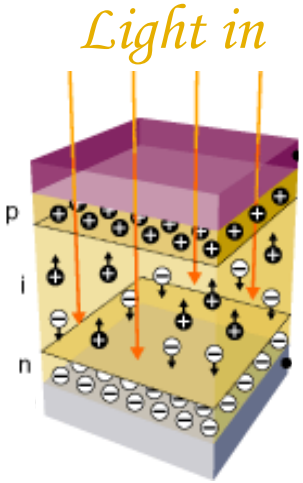


Goal : The accurate treatment of complex heterogeneous systems to gain physical insight.









Where we were at the start of the project from previous collaboration with Kale group:

OpenAtom

Charm++ implementation of the Car Parinello Ab initio Molecular Dynamics based on KS-DFT within Generalized Gradient Approx.

Features include:

- Order $N^2 \log(N)$ Euler Exponential Spline (EES) method for norm conserving non-local pseudopotentials.
- Order $N \log(N)$ EES method for local pseudopotential and Ewald interactions.
- High parallel scaling on BlueGene/L and BlueGene/P (10k procs).
- Roughed in path Integrals, k-point sampling, LSDA and parallel tempering sampling.
- Parallel 3D-FFTs handwritten by scientists.

Great for main group systems, achieve nanosecond time scales- a breakthrough in its day (just a few years ago)!

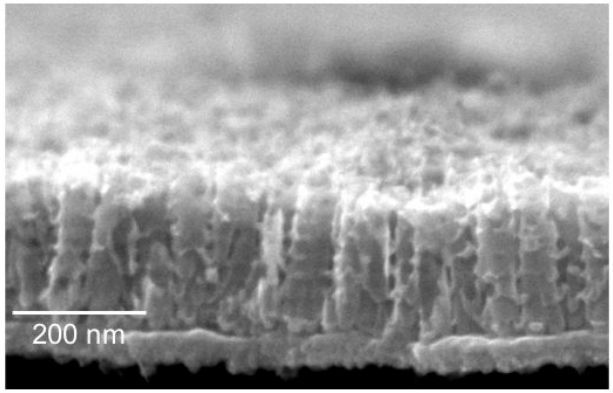
IBM Journal of Research and Development 52 (1.2), 159-175 (2008).

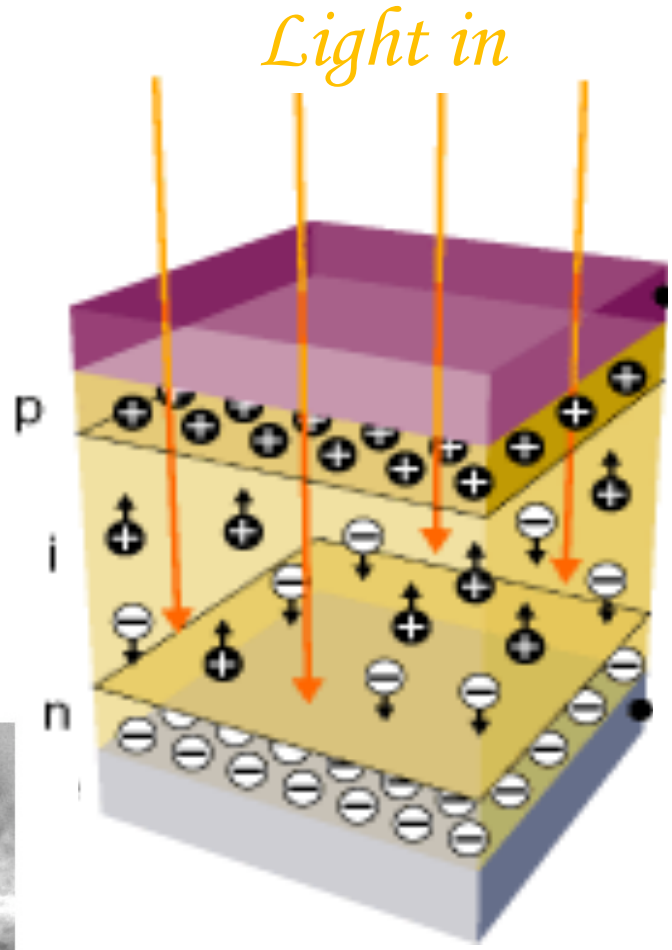
Transparent Conducting Electrodes (TCEs) for thin film amorphous silicon solar cells

Conventional TCEs:
- Indium Tin Oxide (ITO)
- Zinc Oxide (ZnO)

Performance:
- Transparency 95%
- Sheet resistance 10Ω

Manufacturing:

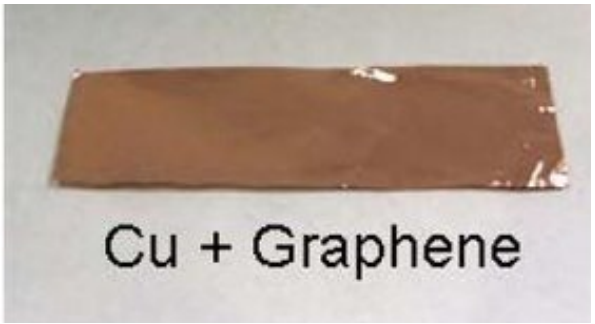




Graphene TCEs:
- 1 – 8 atomic layers

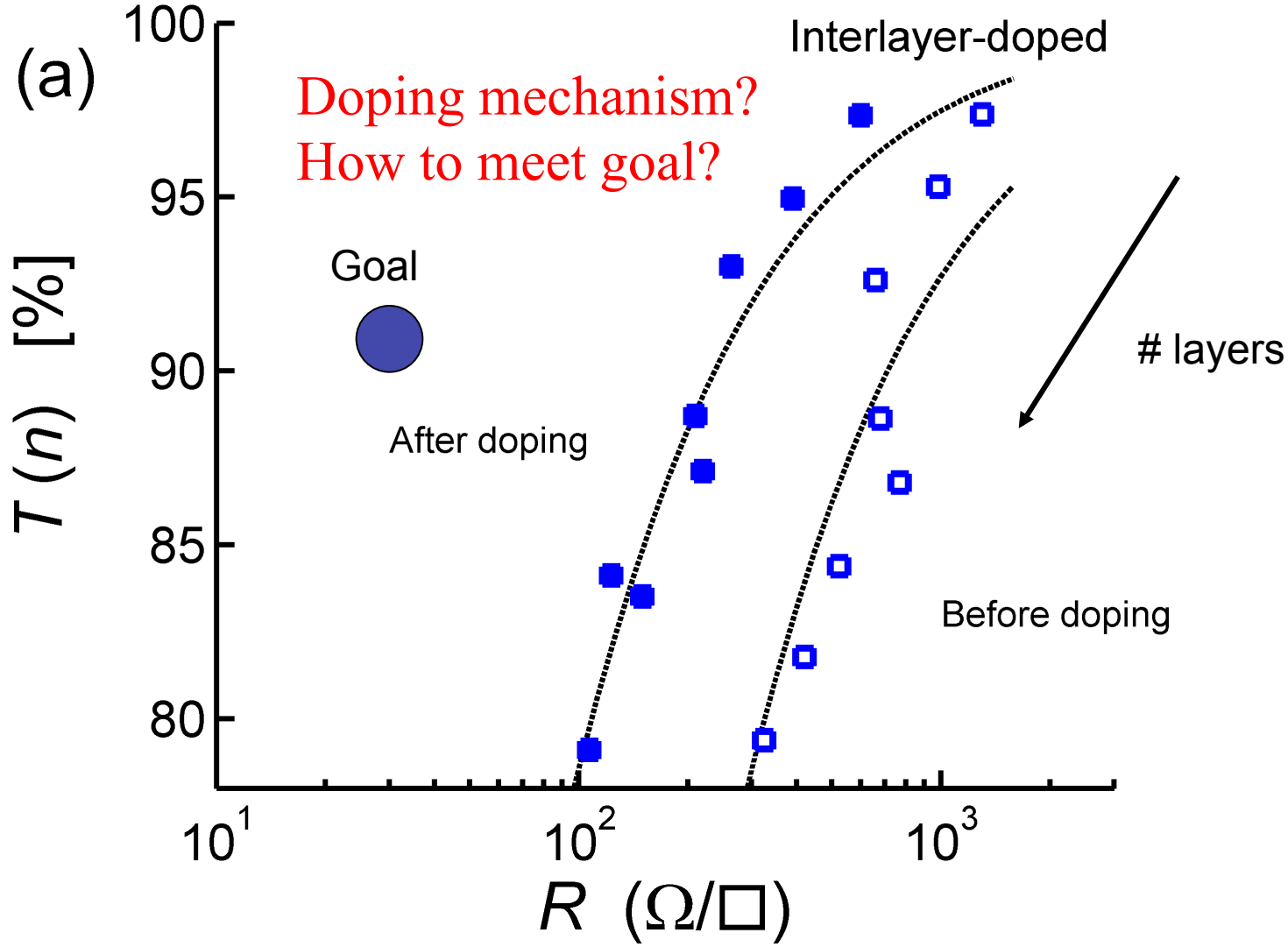
Performance:
- Transparency 85%
- Sheet resistance 100Ω

Manufacturing:
cm X cm size sheets



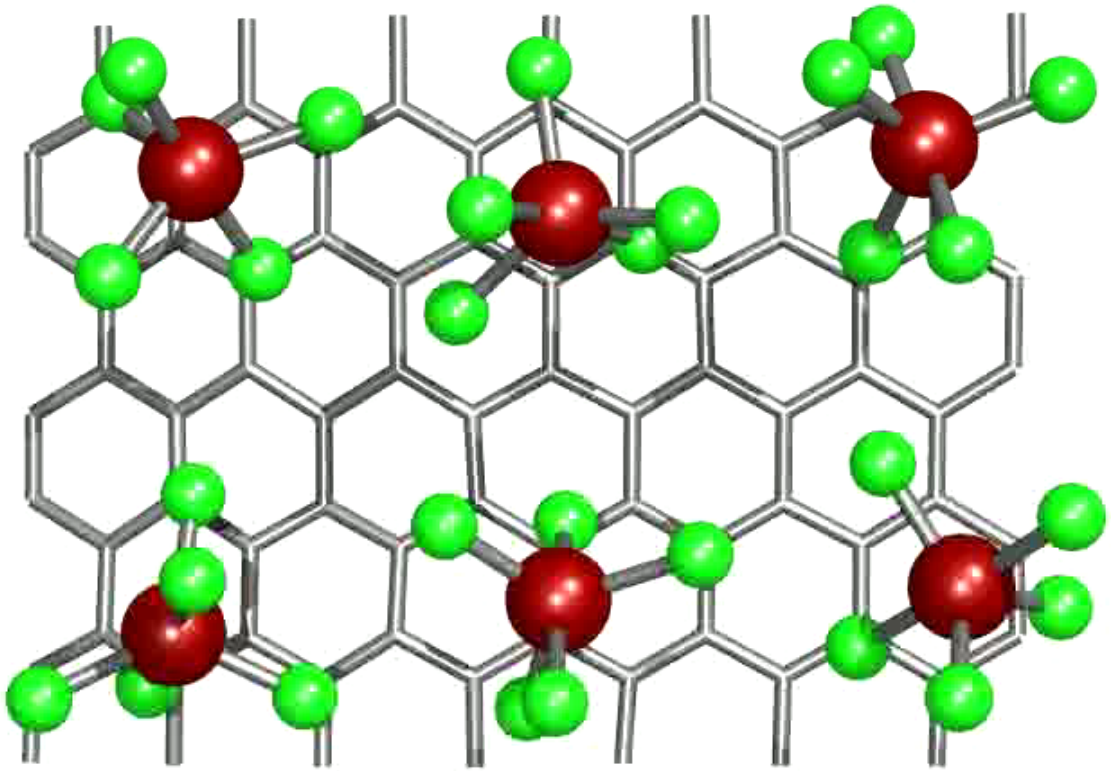
Science, 324, p. 1312 (2009).

Engineering goal



Experimental data: G. Tulevski (IBM), A. Kasry (EGNC), A. Boll (IBM)
ACS nano 4 (7), 3839-3844 (2010)

$$6SbCl_5 + C_{84} \rightarrow C_{84}^{2+}(2SbCl_6^-) + SbCl_3 + 3SbCl_5$$

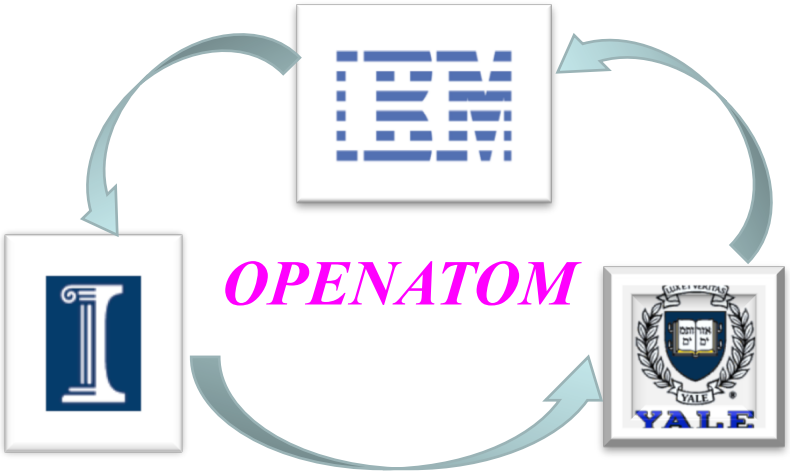


ACS nano 5 (4), 3096-3103 (2011)

Project Goals: Improve sampling, accuracy, applicability and parallel performance of OpenAtom to achieve breakthrough performance

Transition Metals: Plane Augmented Wave method, LSDA, k-point sampling.
Reactive Chem: Hybrid functionals (beyond GGA) – Exact exchange (HF).
Nuclear Quantum Effects: Path Integral Molecular Dynamics.
Sampling Rough Energy Landscapes: Parallel tempering (PT).
Metric Factors : Improve baseline CPAIMD with phase space metrics (PSM).
Extension to Analytics: Use power of OpenAtom in Discovery Projects.

Addressing complex systems and sampling problems requires significant collaborative development!!!!



Project Goals: Improve sampling, accuracy, applicability and parallel performance of OpenAtom

UIUC tasks:

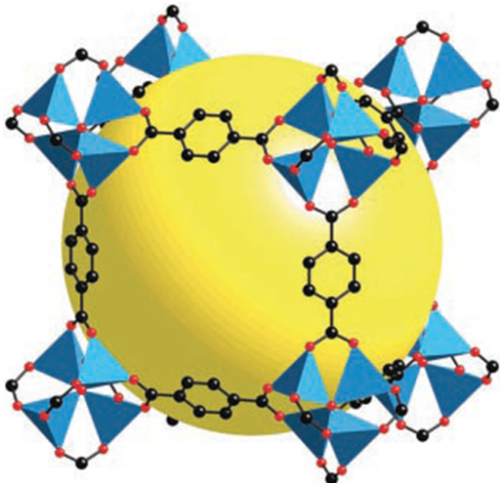
- 3D-FFT library – beyond GJM.
- Flow of control refactor.
- Bringing Tempering, Path integrals, k-points, LSDA to life – "**Übers**".
- Parallelization of "**Übers**".
- GPGPUs for orthogonality.
- Parallelization of "Advanced methods" (PAW, HF, PSM, …).

IBM tasks:

- Derive order $N^2 \log(N)$ PAW.
- Implement Grimme van der Waals.
- Derive reduced order HF exch.
- Derive selection rule for parallel tempering on sampled potential surfaces – penalty method.
- Derive improved CPAIMD via PSM.
- Write toy codes and scientific papers.

Joint work:

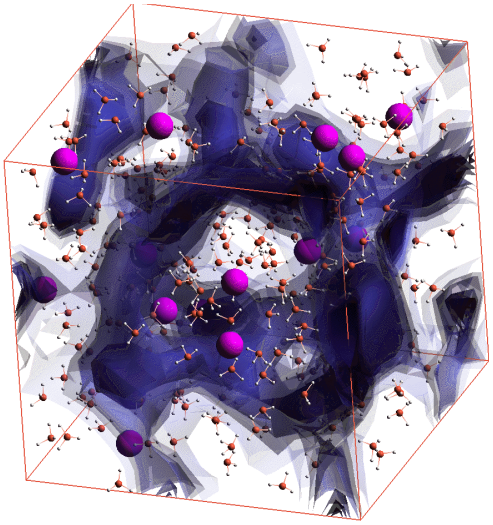
- Framework new methods.
- Implement new methods.
- Develop test suite for new methods – implement in OpenAtom's Jenkins app.
- Apply OpenAtom to important systems across S&T .e.g. Metal Organic Framework with Yale & UIUC.

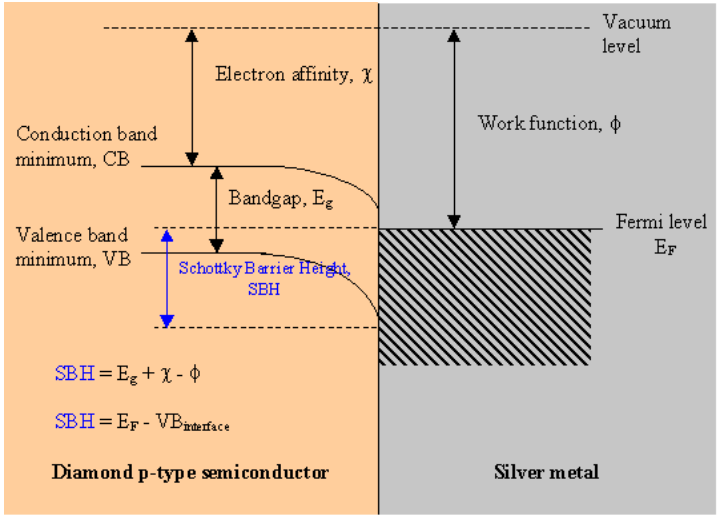


Hero System

Reduced order Hartree-Fock Exchange for Extended States – in 3 minutes or less

Challenge: Reduced order methods for HF exchange are all formulated for localized states. Metal-insulator transitions, Metal-Semiconductor-Metal junctions → "no-go".





Solution: A collaboratively developed **r**-space outer product formulation motivates a new **r**-space/**g**-space decomposition to reduce HF exch. computational complexity by $N^{2/3}$, $N^3 \rightarrow N^{7/3}$ for the plane-wave (pw) basis.

The proof – in a nutshell (2 slides)

$$E_x = -1/2 \int D(\boldsymbol{h})\, d\boldsymbol{r} \int D(\boldsymbol{h})\, d\boldsymbol{r}' \sum_{\boldsymbol{m}} f(\boldsymbol{r},\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'+\boldsymbol{m}\boldsymbol{h}|$$

$$\rho(\boldsymbol{r},\boldsymbol{r}') = \sum_{i \in occ} \varphi_i^*(\boldsymbol{r}')\varphi_i(\boldsymbol{r})$$

$$f(\boldsymbol{r},\boldsymbol{r}') = |\rho(\boldsymbol{r},\boldsymbol{r}')|^2$$

HF exch. under PBC: sum over periodic images, ***m***.

Outer product of orbitals on the discrete pw mesh ~ N^3.

$$1/r = \mathrm{erf}(\alpha r)/r + \mathrm{erfc}(\alpha r)/r$$

Following Ewald - insert long/short-range decomp. of Coulomb interaction.

$$E_x(\mathrm{short}) = -1/2 \int D(\boldsymbol{h})\, d\boldsymbol{r} \int D(\boldsymbol{h})\, d\boldsymbol{r}' \; f(\boldsymbol{r},\boldsymbol{r}')\,\mathrm{erfc}(\alpha|\boldsymbol{r}-\boldsymbol{r}'|)/|\boldsymbol{r}-\boldsymbol{r}'|$$

1st image sufficient

$$E_x(\mathrm{long}) = -1/2 \int D(\boldsymbol{h})\, d\boldsymbol{r} \int D(\boldsymbol{h})\, d\boldsymbol{r}' \; \sum_{\boldsymbol{m}} f(\boldsymbol{r},\boldsymbol{r}')\,\mathrm{erf}(\alpha|\boldsymbol{r}-\boldsymbol{r}'+\boldsymbol{m}\boldsymbol{h}|)/|\boldsymbol{r}-\boldsymbol{r}'+\boldsymbol{m}\boldsymbol{h}|$$

The proof – in a nutshell (2 slides)

Simplify* and introduce matrices on coarse, f_c, and fine f_f, meshes:

$$E_x(\text{short}) = -\frac{1}{2}\iint d\boldsymbol{R}\,d\boldsymbol{u}\;\; f_f\left(\boldsymbol{R}+\boldsymbol{u}/2,\boldsymbol{R}-\boldsymbol{u}/2\right)\operatorname{erfc}(\alpha|\boldsymbol{u}|)/|\boldsymbol{u}|\;\theta_H\left(U_c-|\boldsymbol{u}|\right)$$

$$E_x(\text{long}) = -\frac{1}{2V}\sum_{\boldsymbol{g}\neq\boldsymbol{0},\,|\boldsymbol{g}|<G_c} 4\pi/|\boldsymbol{g}|^2\;\exp\left(-|\boldsymbol{g}|^2/4\alpha^2\right)\; f_c\;(\boldsymbol{g},\boldsymbol{g}) + \pi/2V\alpha^2\;\; f_c\;(0,0)$$

$$f_c\;(\boldsymbol{g},\boldsymbol{g}') = \int D(\boldsymbol{h})\,d\boldsymbol{r}\int D(\boldsymbol{h})\,d\boldsymbol{r}'\;\;\exp(i\boldsymbol{g}\cdot r)\exp(-i\boldsymbol{g}'\cdot r')\; f_c\;(\boldsymbol{r},\boldsymbol{r}')$$ FFT's

$f_f\;(\boldsymbol{R}+\boldsymbol{u}/2,\boldsymbol{R}-\boldsymbol{u}/2)$ Create sparse-matrix on fine mesh: $N^2\;U_c^3 \sim N^2\alpha$

$f_c\;(\boldsymbol{r},\boldsymbol{r}')$ Create dense-matrix on coarse mesh: $N^3 G_c^6 \sim N^3$

$\alpha \sim N^{-1/9}$ Choose α to equalize scaling: $N^{7/3}$ (add log N for FFT's)

$$*\sum_{\boldsymbol{m}} p(\mathbf{r}+\mathbf{m}\boldsymbol{h}) = 1/\det(\boldsymbol{h})\;\;\sum_{\boldsymbol{g}} p\;(\boldsymbol{g})\;\boldsymbol{e}^{i\boldsymbol{g}\cdot\boldsymbol{r}},\;\; V=\det(\boldsymbol{h}),\;\; \boldsymbol{g}=\boldsymbol{g}\,\boldsymbol{h}$$

$$1,\;\boldsymbol{g}=\{i\,k\,l\},\;\; p(\boldsymbol{g}) = \int_{-\infty}^{\infty} d\boldsymbol{r}'\;\boldsymbol{e}^{-i\boldsymbol{g}\cdot\boldsymbol{r}'}\;p(\boldsymbol{r}')$$

Reduced order Hartree-Fock Exchange: Scaling

Complexity of short-range part governed by cutoff radius in **u**-space through erfc($\alpha|\mathbf{u}|$), $|\mathbf{u}|<U_c \sim \alpha^{-1}$

$$N^2 U_c^3 \sim N^2 \alpha^{-3}$$

Complexity of long-range part governed by cutoff radius in **g**-space through $\exp(-|\boldsymbol{g}|^2/4\alpha^2)$, $|\boldsymbol{g}|<G_c \sim \alpha$ [1]

$$N(VG_c^3)^2 \sim N^3 \alpha^6$$

Equating gives

$$\alpha \sim N^{-1/9}$$

which in turn yields the desired scaling of the short/long-range parts

$$\sim N^{2+1/3} = N^{7/3}$$

[1] A plane wave basis with cutoff G_c contains $\sim VG_c^3$ basis functions where

Reduced order Hartree-Fock Exchange: Accuracy

- Short-range HF exchange requires treating an integrable singularity (Coulomb)
- The plane wave mesh is equally spaced in Cartesian coordinates.
- Develop a method to treat integrable singularities on simple meshes.

Consider integration in 1 spatial dimension for simplicity:

$$I=\sum_{i=0}^{n-1} I_{cell}(i)\ ,\ I_{cell}(i)=\int_{i\Delta+a}^{(i+1)\Delta+a} dx\, f(x)k(x$$

Introduce a 1st order interpolation of *f(x)* in each cell

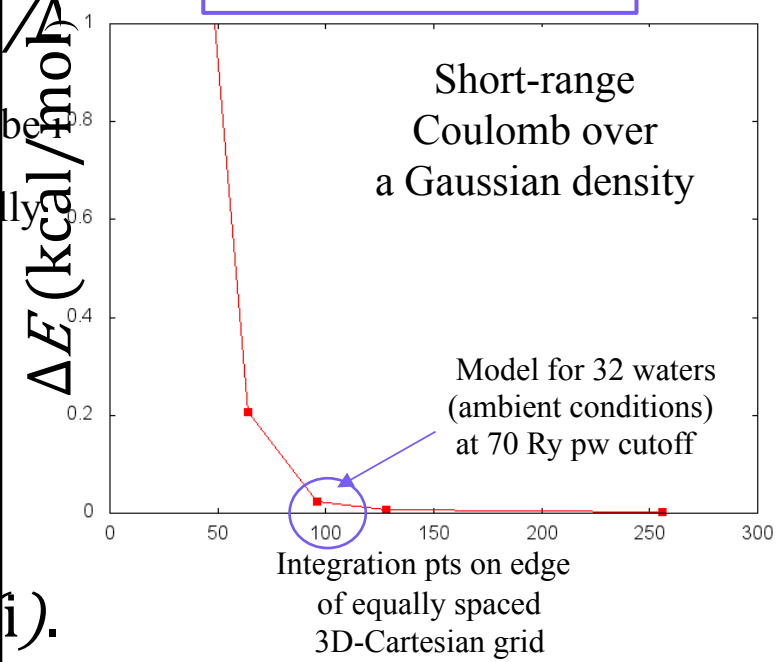
$$f(x)=[(x(i+1)-x)/\Delta]f(i)+[(x-x(i))/\Delta$$

where $f(i)=f(i\Delta+a)$ and $x(i)=(i\Delta+a)$. This can be into $I_{cell}(i)$ and the integral over x performed analytically. Combing terms, the 1st order approximation to I, is

$$I^{(1)}(\Delta)=\Delta\sum_{i=0}^{n} f(i)K(i)$$

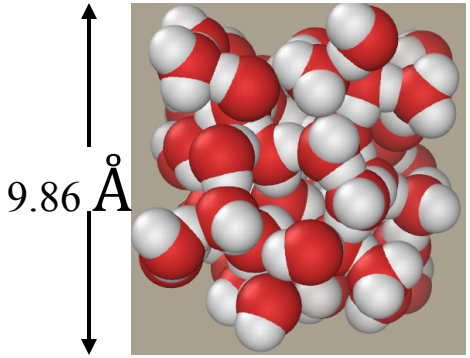
where for k(x)=1 the result is trapezoidal rule integration and *K(i)* is the "grid" function of *k(x)*. This is employed only in cells where *k(x)* is rapidly varying – otherwise trap. rule integration of product function, f(x)k(x), is used, $K(i)\rightarrow k(i)$.

3D Generalization



Reduced order Hartree-Fock Exchange: Sizing

System of interest : 32 water molecules under ambient conditions in a L=9.86 Å on edge cell with a 70 Ry pw cutoff sets the fine pw grid to be ≈100 points on edge.



Take: (1) the fine grid u-space cutoff: $U{\downarrow}c = 3.5/\alpha$
(2) the sparse grid g-space cutoff: $G{\downarrow}c = 7\alpha$

Choose: α=21.5/L

Estimate: New method saves ≈70x compared a N^3 computation (.i.e. **g**-space treatment on the fine pw grid, lim $\alpha \rightarrow \infty$).

Validate: Cutoff choices and timing estimates need testing in the "real world" where overhead, truncation error etc. matter.

Progress towards project goals

UIUC tasks:

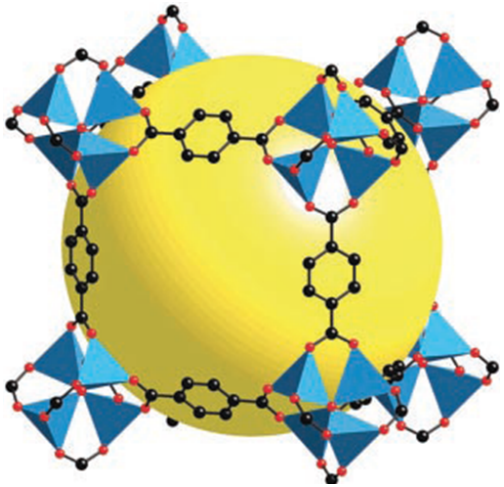
- 3D-FFT library – beyond GJM.
- Flow of control refactor.
- Bring Tempering, Path integrals, k-points, LSDA to life – "**Übers**".
- Parallelization of "**Übers**".
- GPGPUs for orthogonality.
- Parallelization of "Advanced methods" (PAW, HF, PSM, …).

IBM tasks:

- Derive $N^2 \log(N)$ scaling PAW.
- Implement Grimme van der Waals.
- Derive reduced order HF exch.
- Derive selection rule for parallel tempering with sampled. potential surfaces – penalty method.
- Derive improved CPAIMD via PSM.
- Write toy codes and scientific papers.

Joint work:

- Framework new methods.
- Implement new methods.
- Develop test suite for new methods – implement in OpenAtom's Jenkins app.
- Apply OpenAtom to important systems across S&T, e.g. Metal Organic Framework with Yale & UIUC.
- Develop analytics application for discovery.



Hero System

accomplished ongoing TBD

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

OpenAtom Ground State Software Overview

PPL Contributors: Eric Bohm, Nikhil Jain, Prateek Jindal, Eric Mikida, Michael Robson

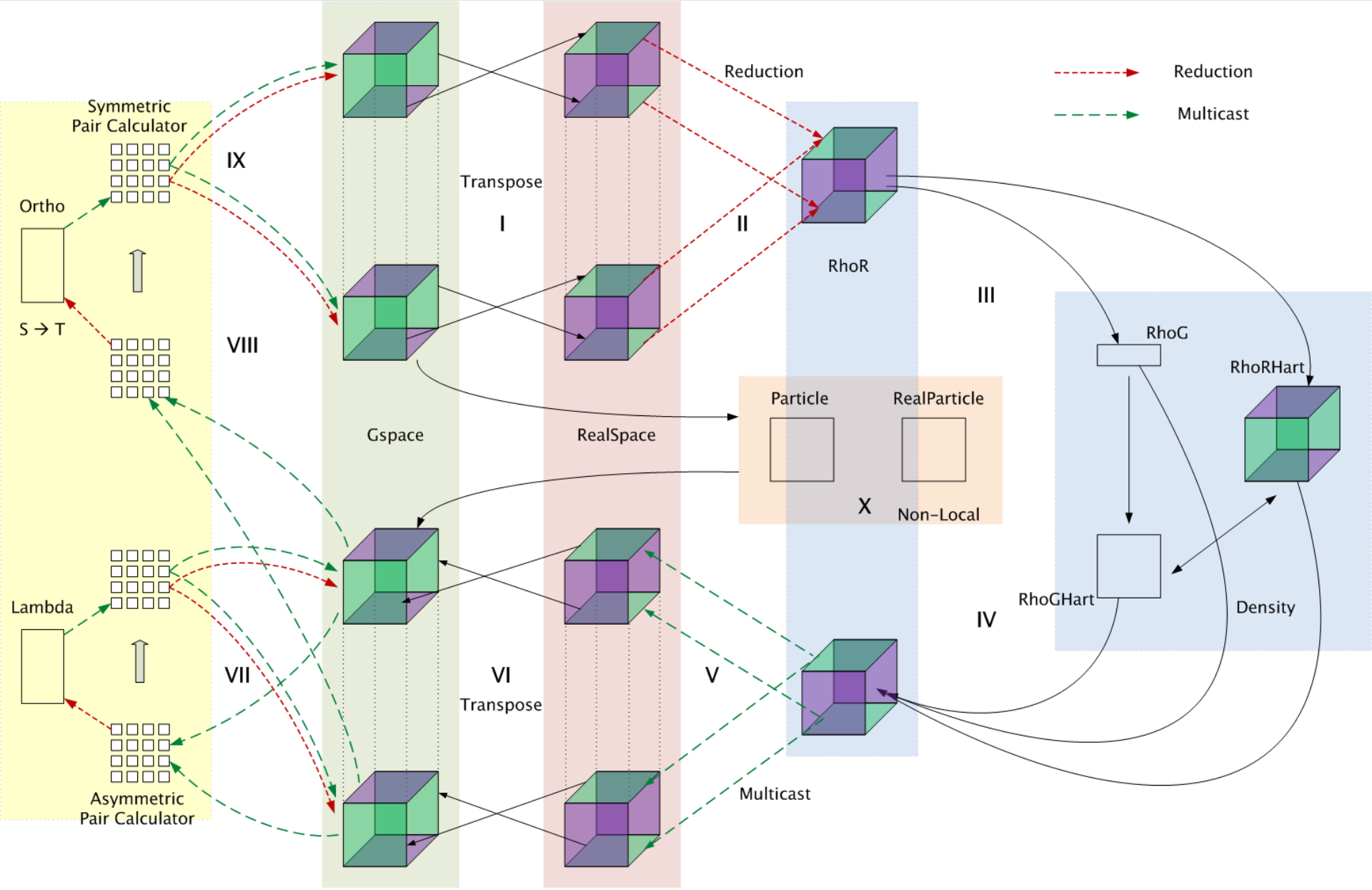
Software Infrastructure

- GIT (Gerrit) based repository:
  - http://charm.cs.illinois.edu/gerrit/openatom
  - Or https://github.com/ericbohm/OpenAtom/
- Test system datasets available in git
  - Make test - Basic feature verification
  - Make full_test - Extensive use case verification
- *Jenkins* testing
  - Release branch in nightly Charm++ testing
  - Release branch in Charm++ continuous integration testing

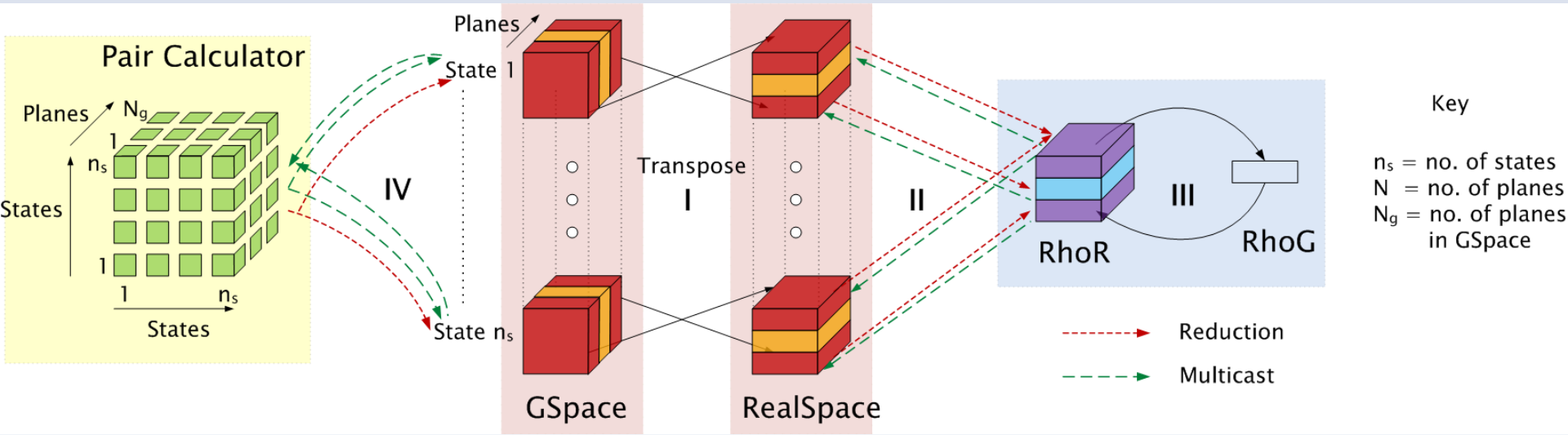
Ground State Feature Status

Feature	Minimization Status	Dynamics Status	Test Integration
CPAIMD Dynamics	NA	Production	Automated
Path Integrals	Production	Production	Automated
K-Points	Production	Needs Verification	Manual
Spin Orbitals	Production	Production	Manual
Tempering	NA	Production	Manual
Born Oppenheimer Dynamics	NA	Production	Automated
Band Generation	Needs Verification	Needs Verification	Manual

Data Structures and control flow in OpenAtom



Object Decomposition

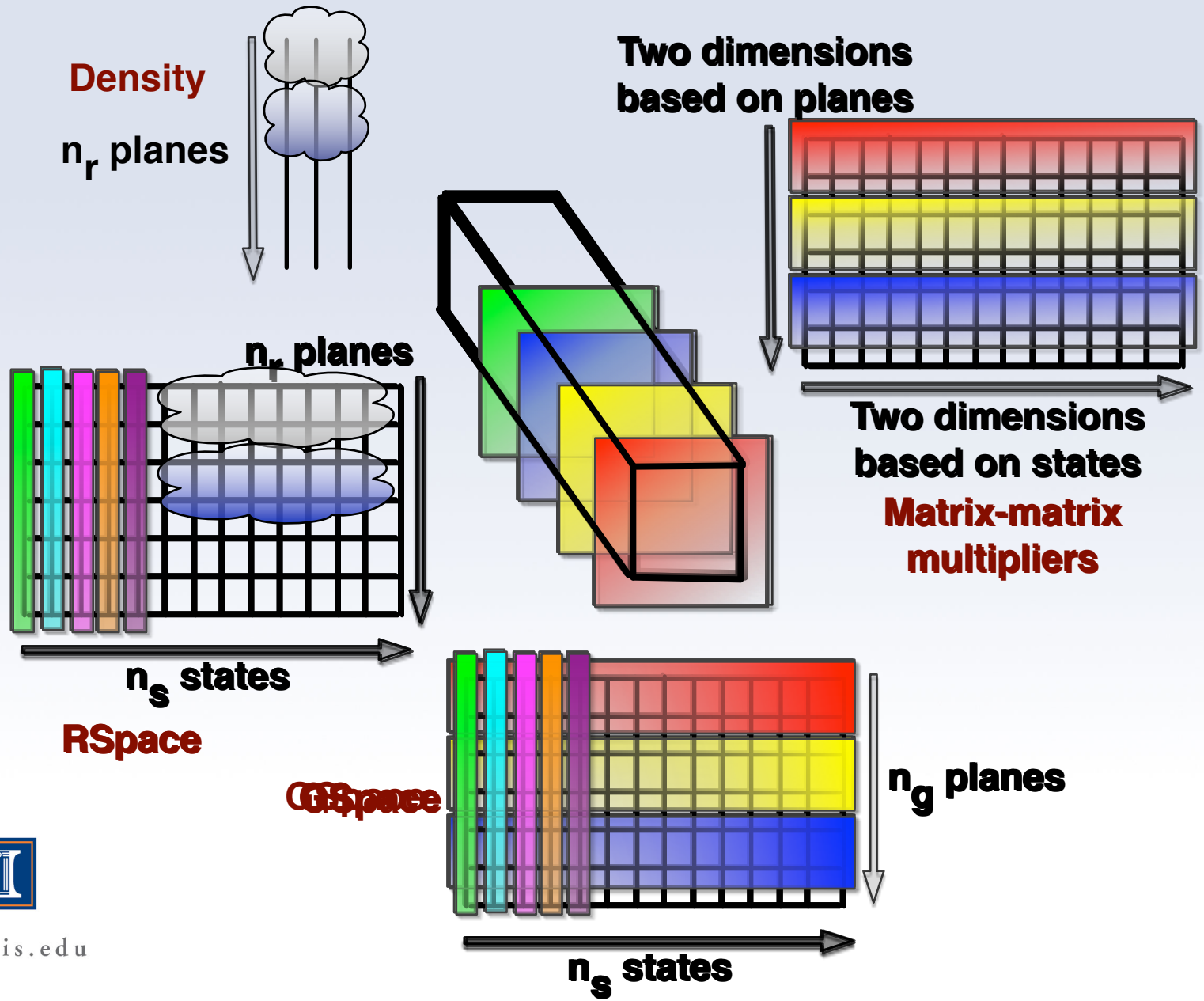


Showing a subset of object collections

Nikhil Jain

OBJECT PLACEMENT

Topology aware mapping



illinois.edu

Adapting to different systems

- Separate the logical operations and machine-specific operations. Example:
  - Logical operation: get an ordered list of nodes
  - Machine specific: Hilbert curve traversal, blocked traversal, plane-traversal

- Density FFTs: require use of full bisection bandwidth – spread throughout the allocation.

- Matrix-matrix multiplies (pair calculators): place near the GSpace planes, but load balance is important.

System utilization without mapping

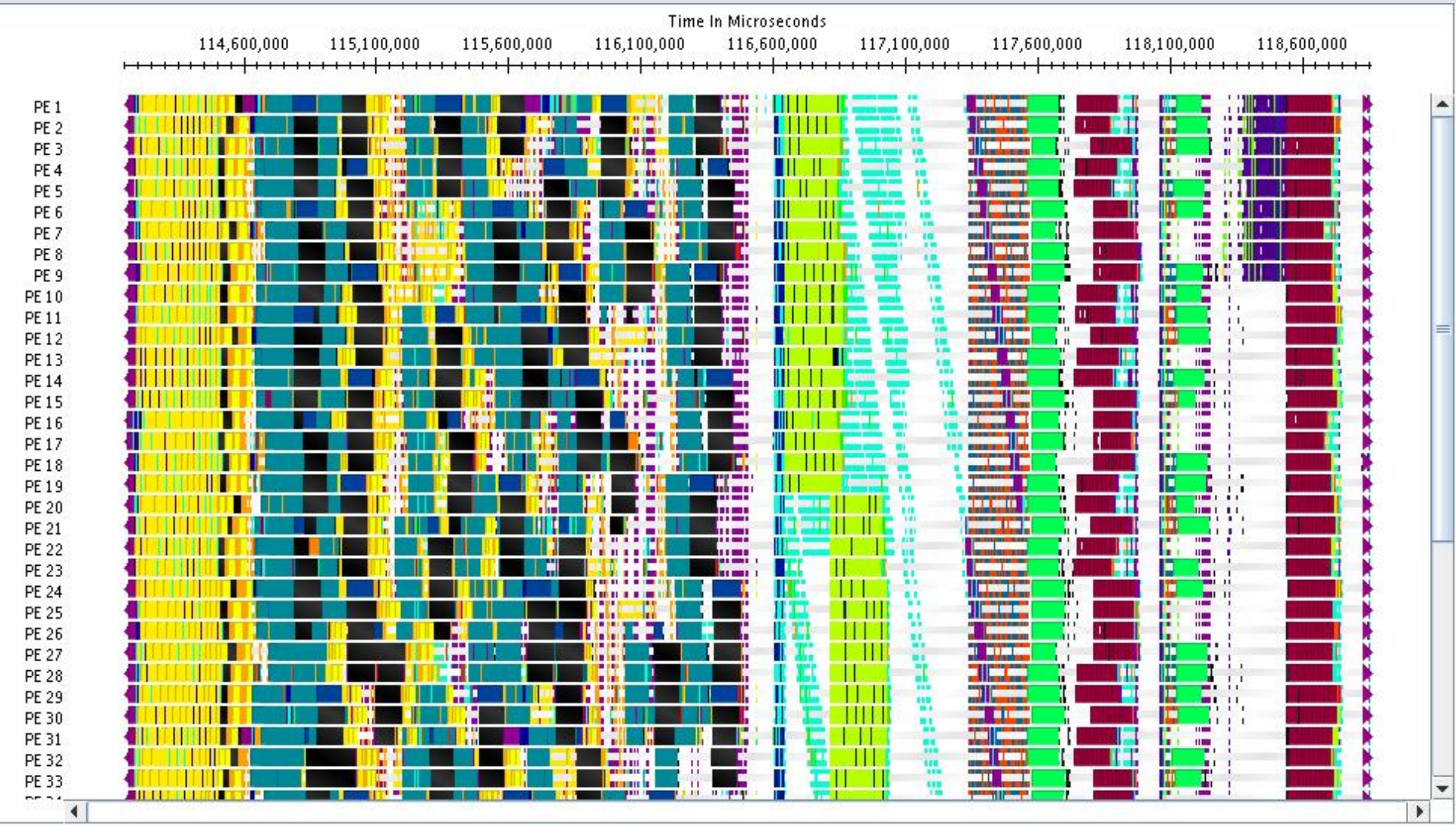
(barriers introduced for clarity)

States: Many G -> R FFTs

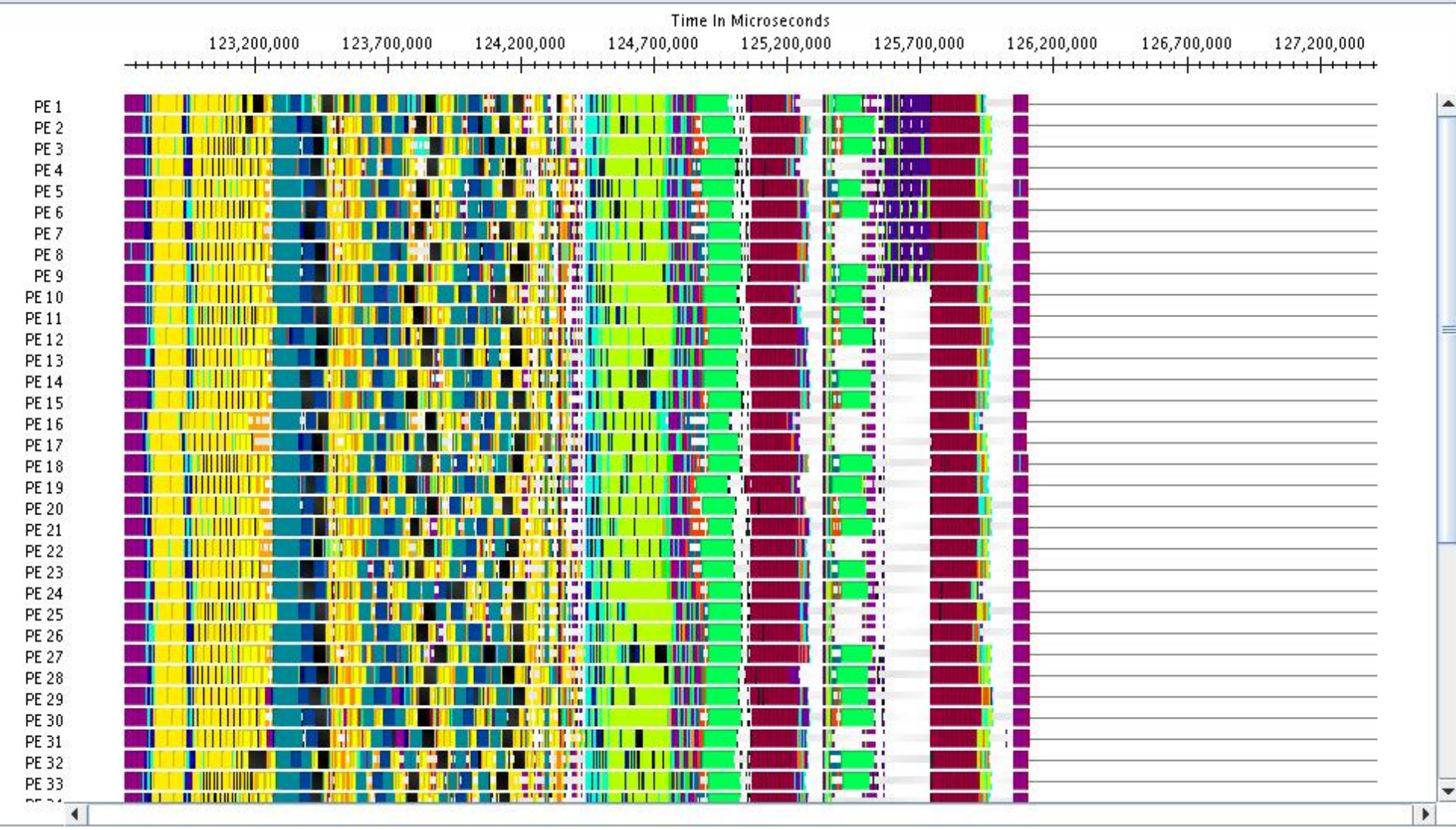
Density: G -> R -> G Non-local G-> R -> G

Density to States States R-> G

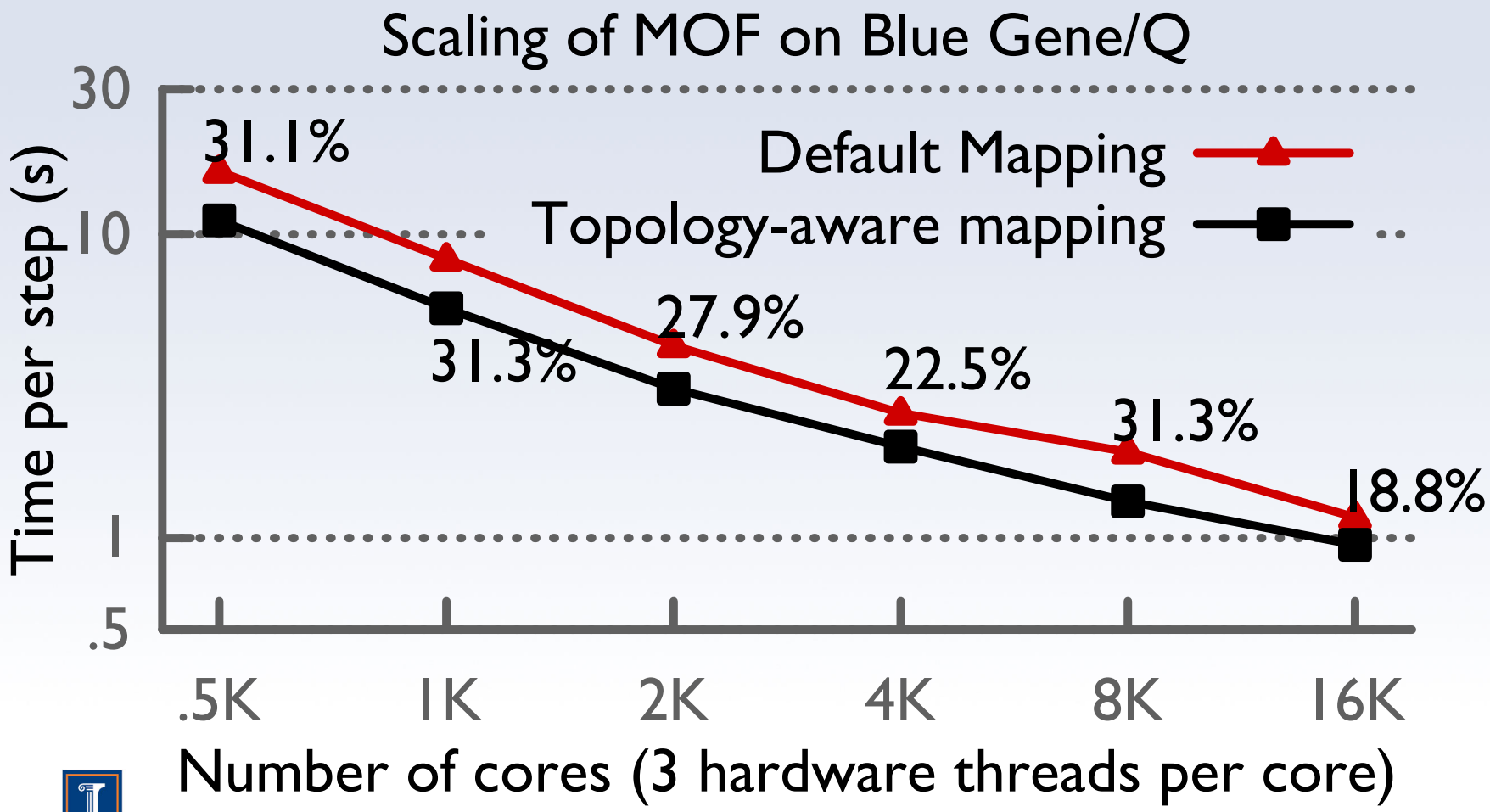
Force correction Ortho-normalizaton



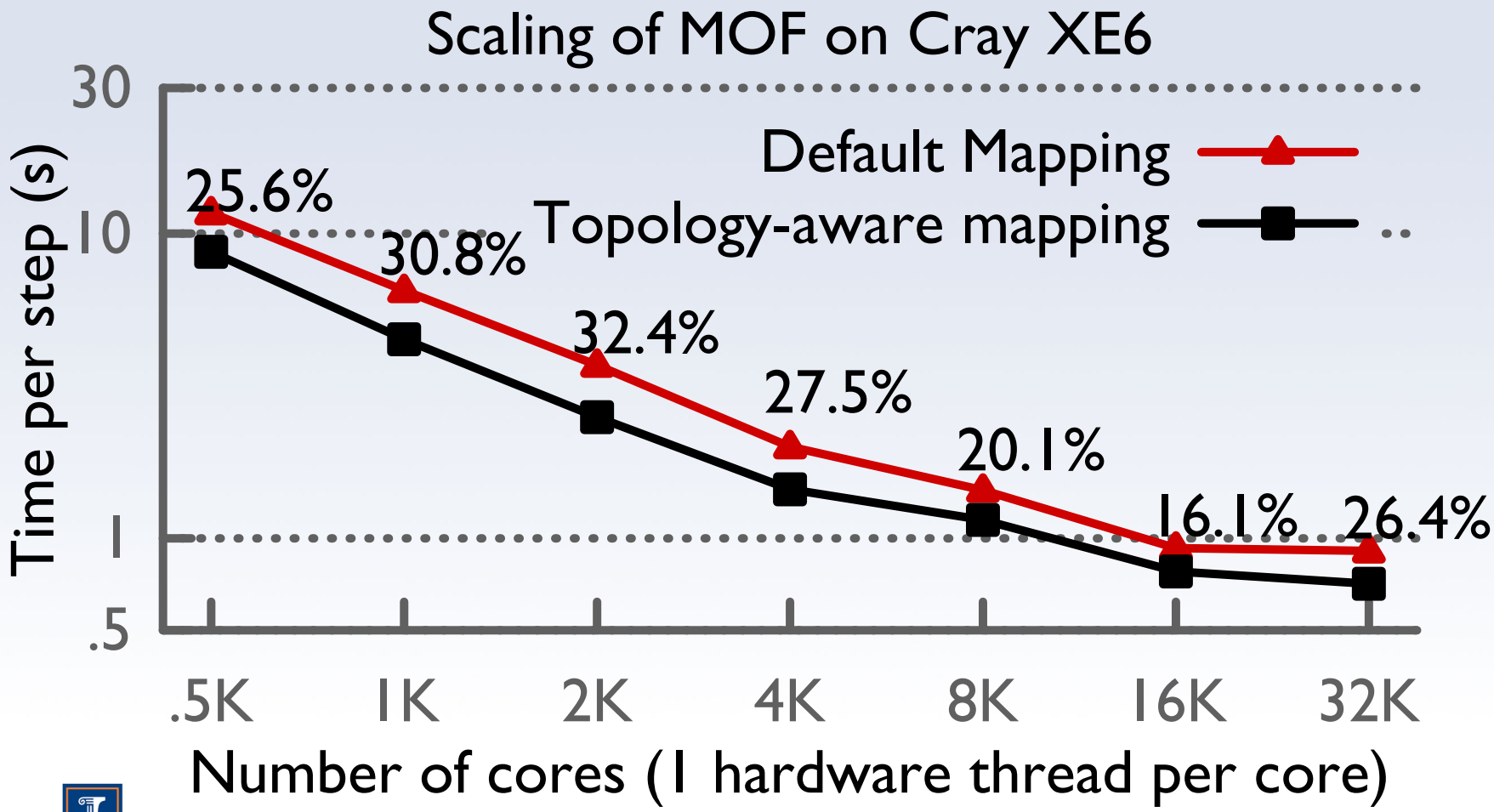
System utilization with mapping



Impact of mapping on Blue Gene/Q: up to 30% improvement



Impact of mapping on Blue Waters: up to 32% improvement

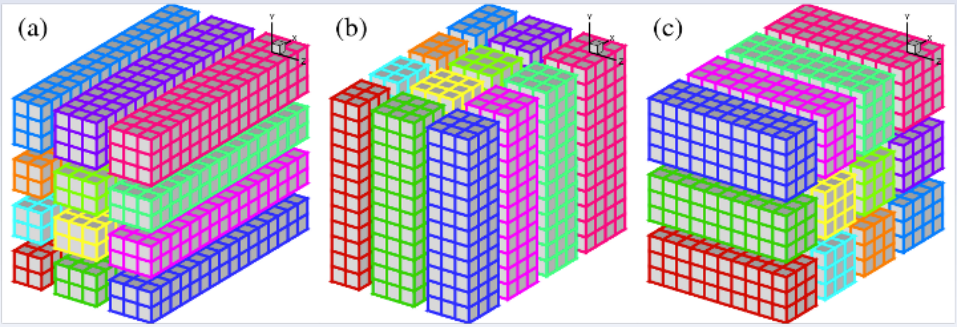


Nikhil Jain

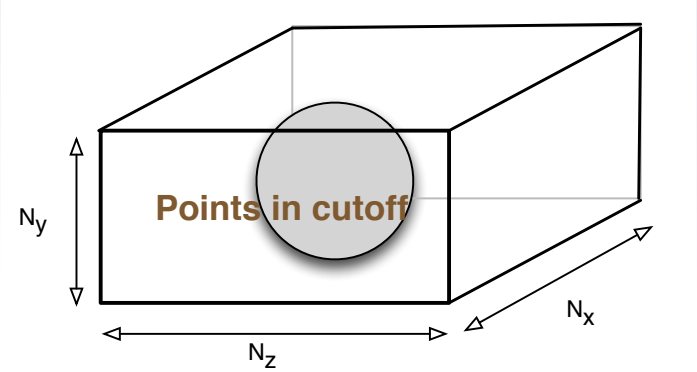
CHARM FFT

Optimizing FFTs in the density

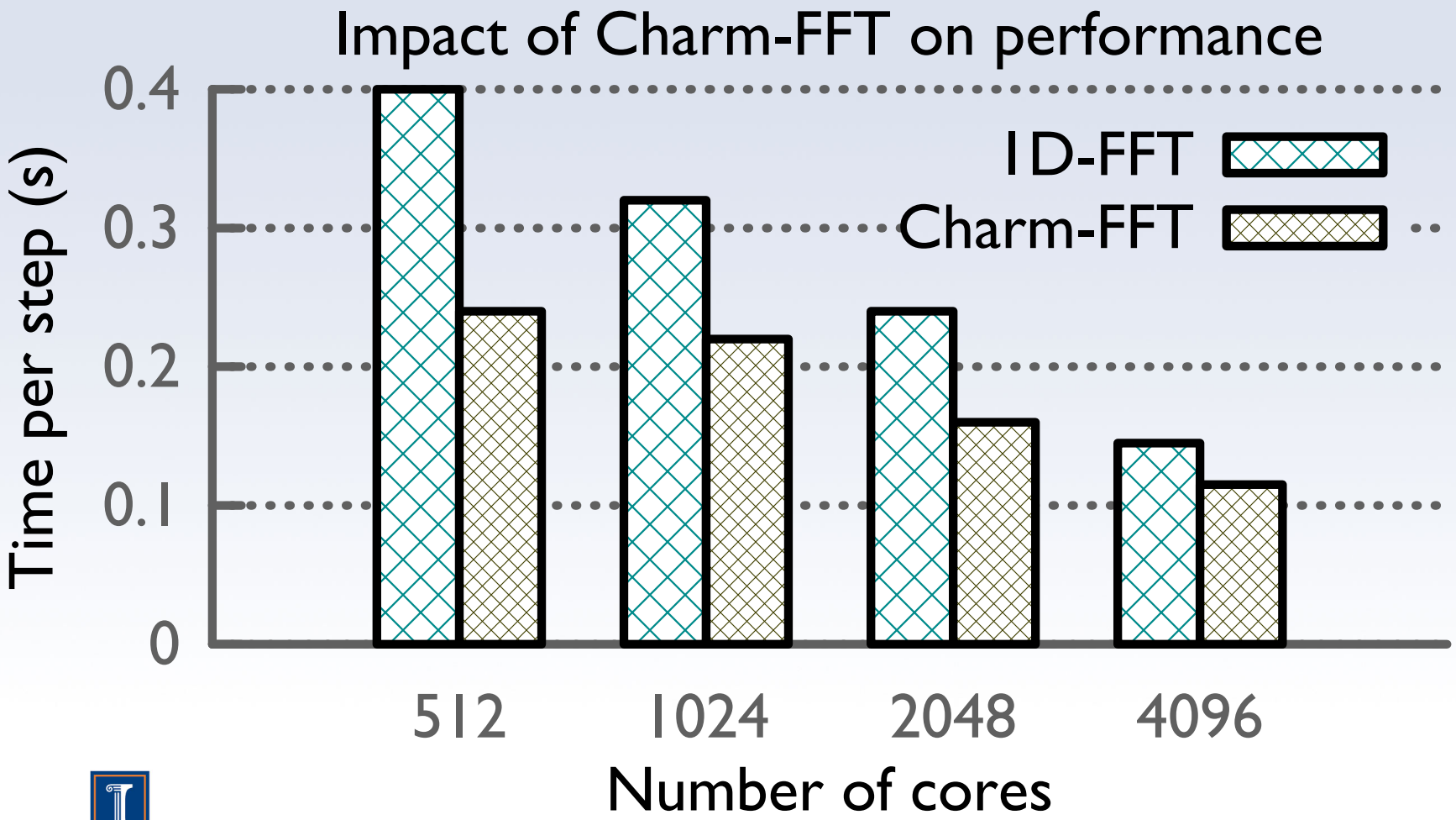
- Charm-FFT implemented to help improve the performance of the density phase.



- **Features:**
  - Concurrent FFTs
  - 2D decomposition
  - Overlaps with other work
  - Cutoff aware
  - Mapped by the application



Simulating 32 molecules of Water on Blue Waters



Eric Bohm

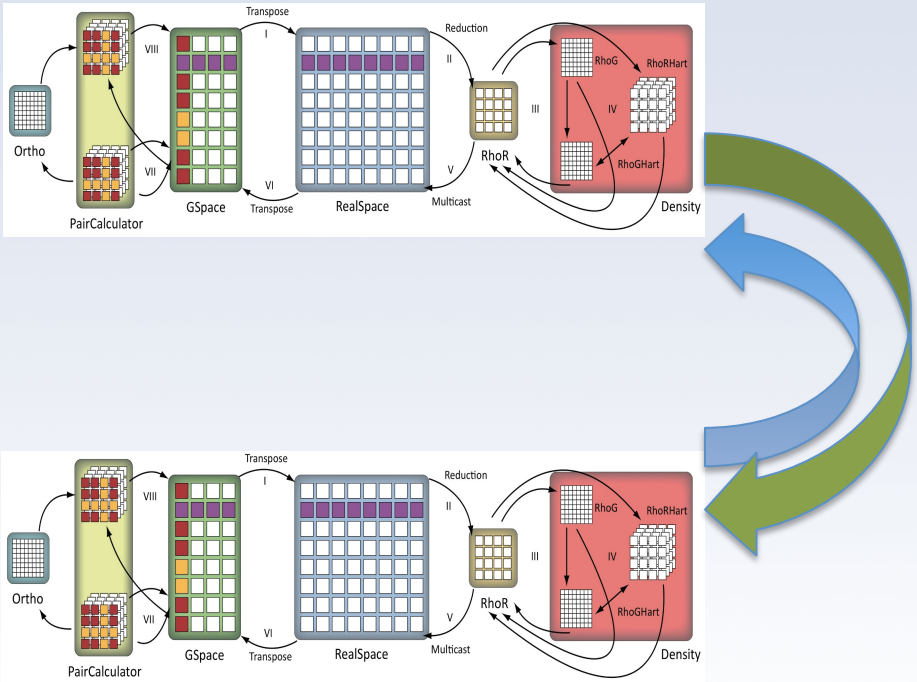
MULTI-INSTANCE METHODS

Multi Instance Methods

- Retain all existing code with minimal changes
- Any feature available for CP minimization or dynamics automatically available for multi-instance use
- Add Master Index of objects
  - Uber[temper][bead][k-point][spin]
  - Objects in any instance can be referenced by any object
  - Support simulations with many kinds of multi instance physics
  - Instance Controller
  - Temper Controller
    - Sum energies across Tempers and Beads
    - Switch Energies and Temperatures
  - Bead Controller
    - Intrapolymer force evaluation and integration

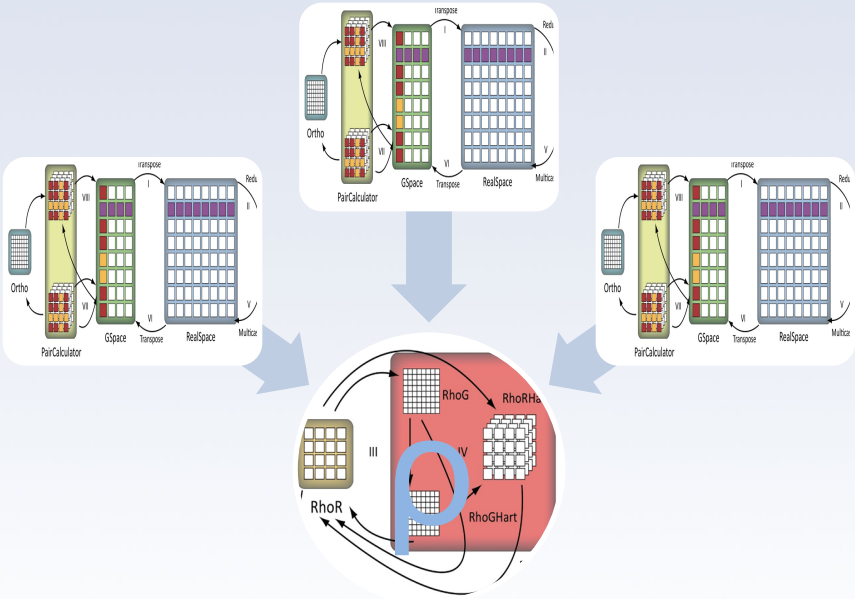
Spin Orbitals (LSDA)

- Each Spin shares : atom and energy chares
- Electron density from down passed to up
  - VKS computed for each spin
  - Returns to standard flow of control
- Independent I/O for state data
- Independent placement for instance chares



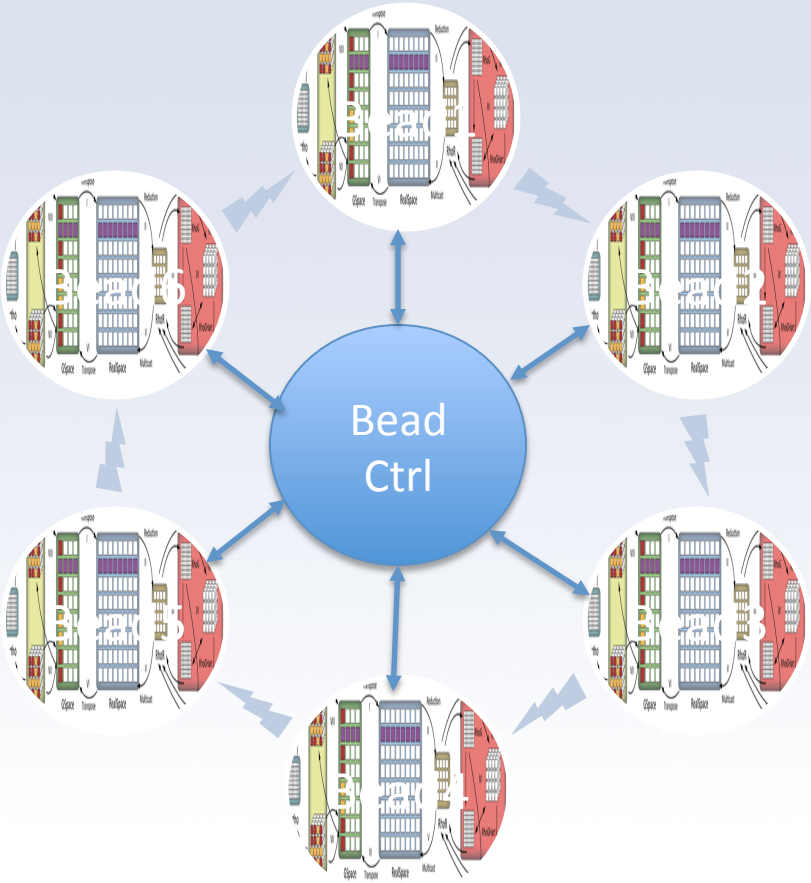
K-Points

- Each k-point shares:
  - electron density, atoms, energy chares
- Electron density = sum over KP electron states
- Wave functions outside the first Brillouin zone forces use of complex (e.g., ZGEMM)
  - Instead of the “doublepack” optimization used at the Γ point
- Independent I/O for state data
- Independent placement for electron state instance chares

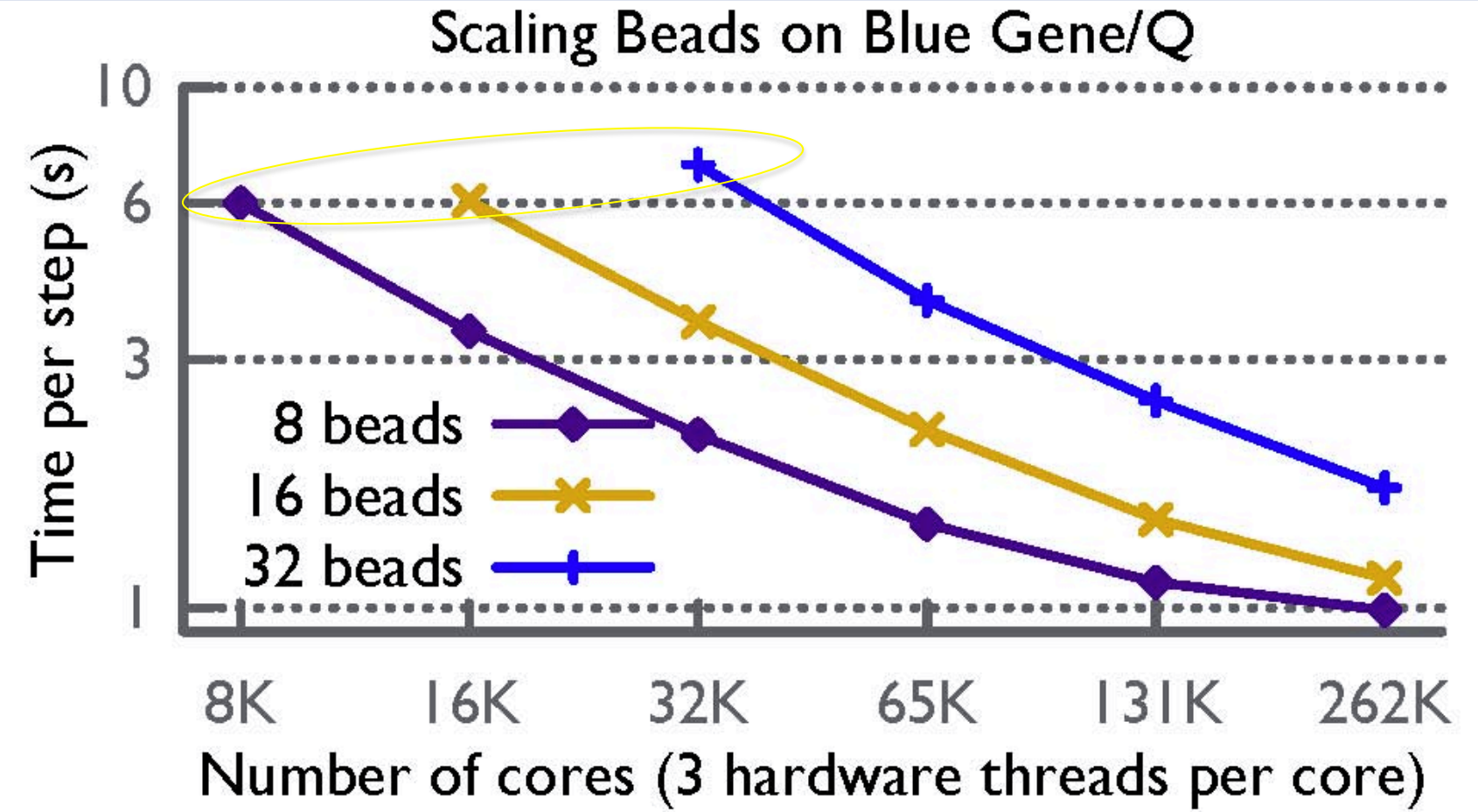


Path Integral Beads

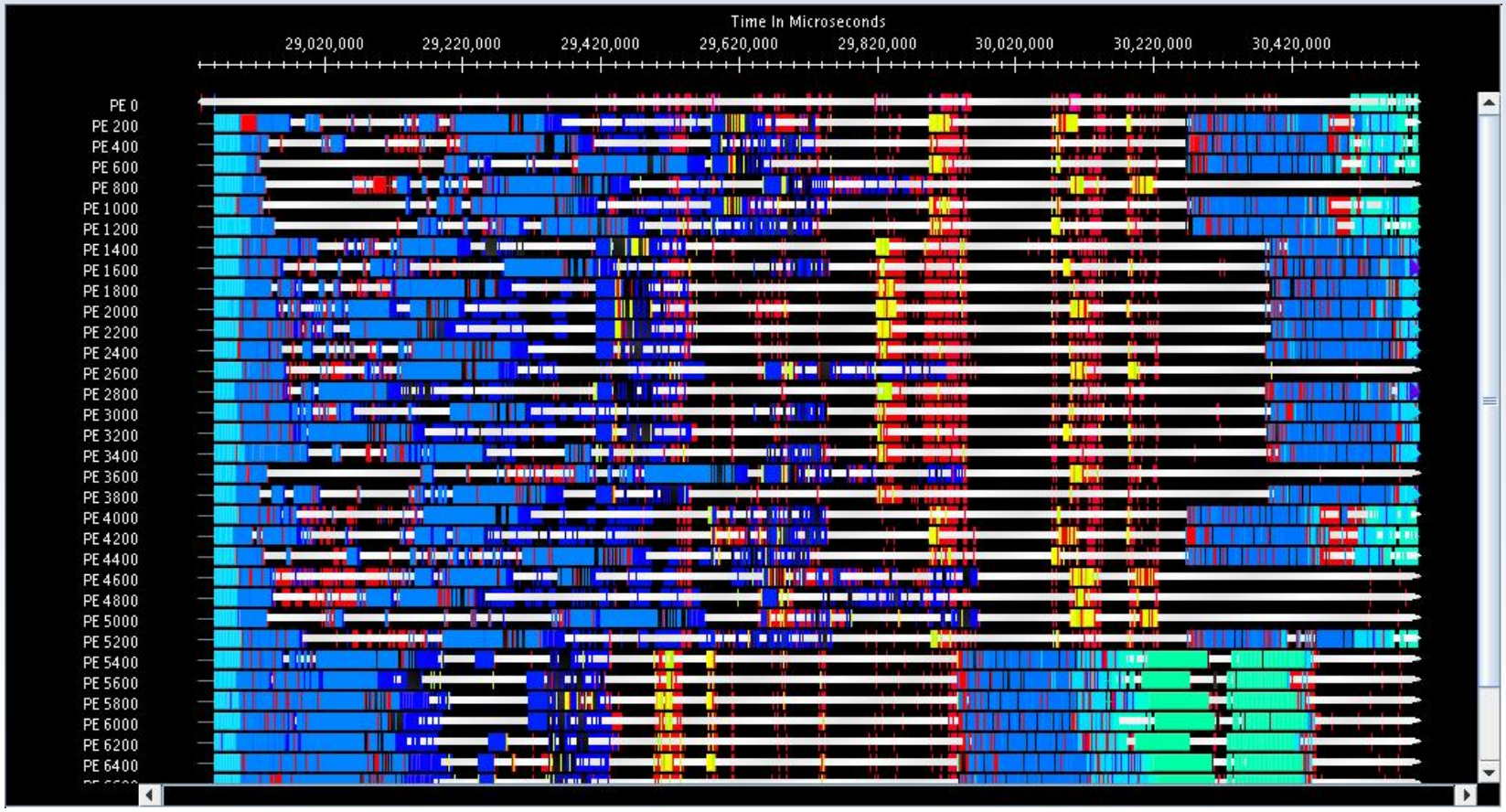
- Path Integral Bead replica contains independent instances of all phases of CPAIMD
  - May contain k-point and spin ensembles
- Intrapolymer force evaluation in PIBeadAtoms
  - Interacts with each Bead instance's AtomsCompute
  - Supplements CPAIMD nucleic force integration phase
  - Computation Parallelized across NumAtoms and NumBeads
- Independent I/O for state and coordinate data
- Independent placement for instance chares



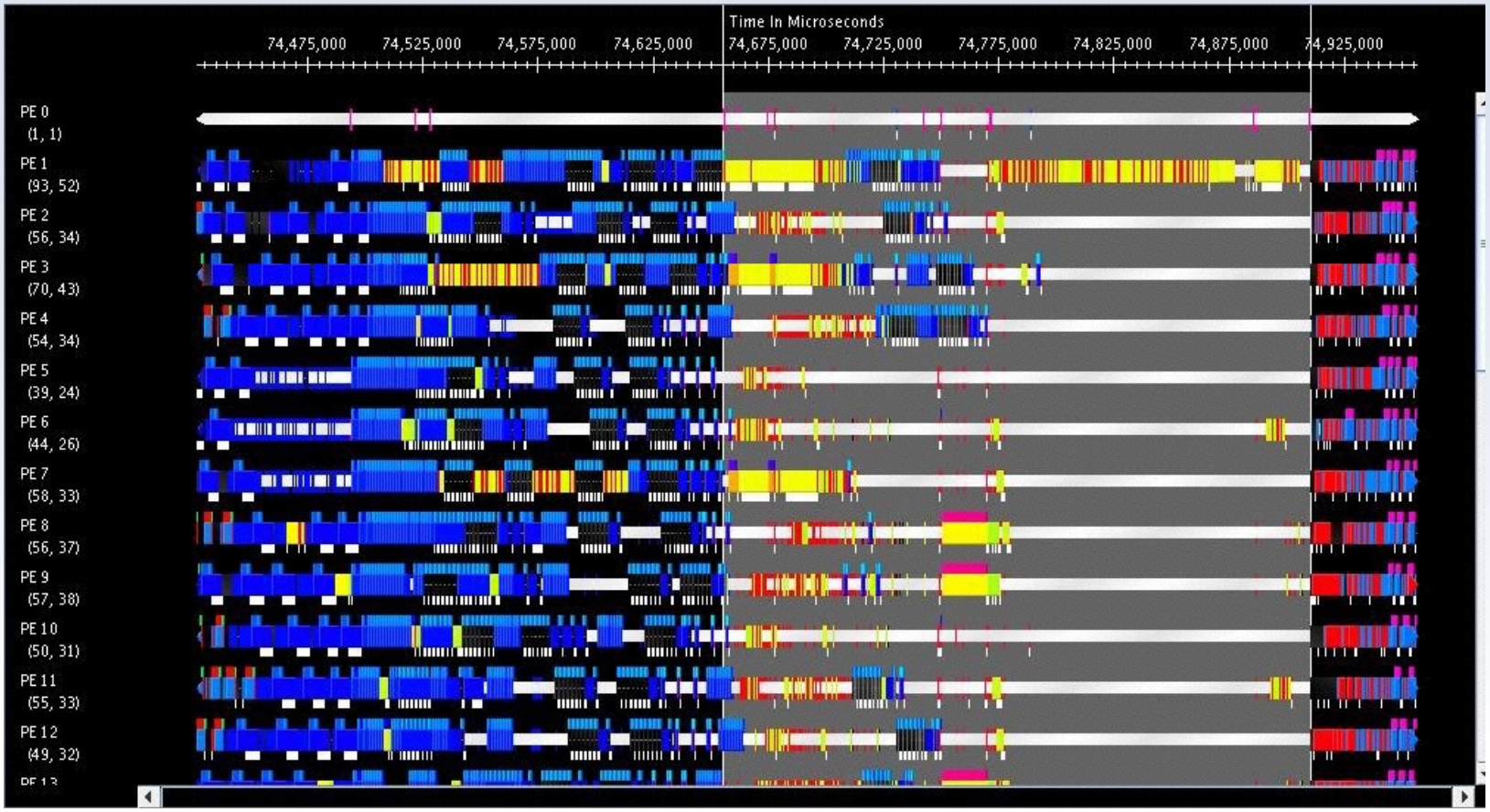
Path Integral Performance



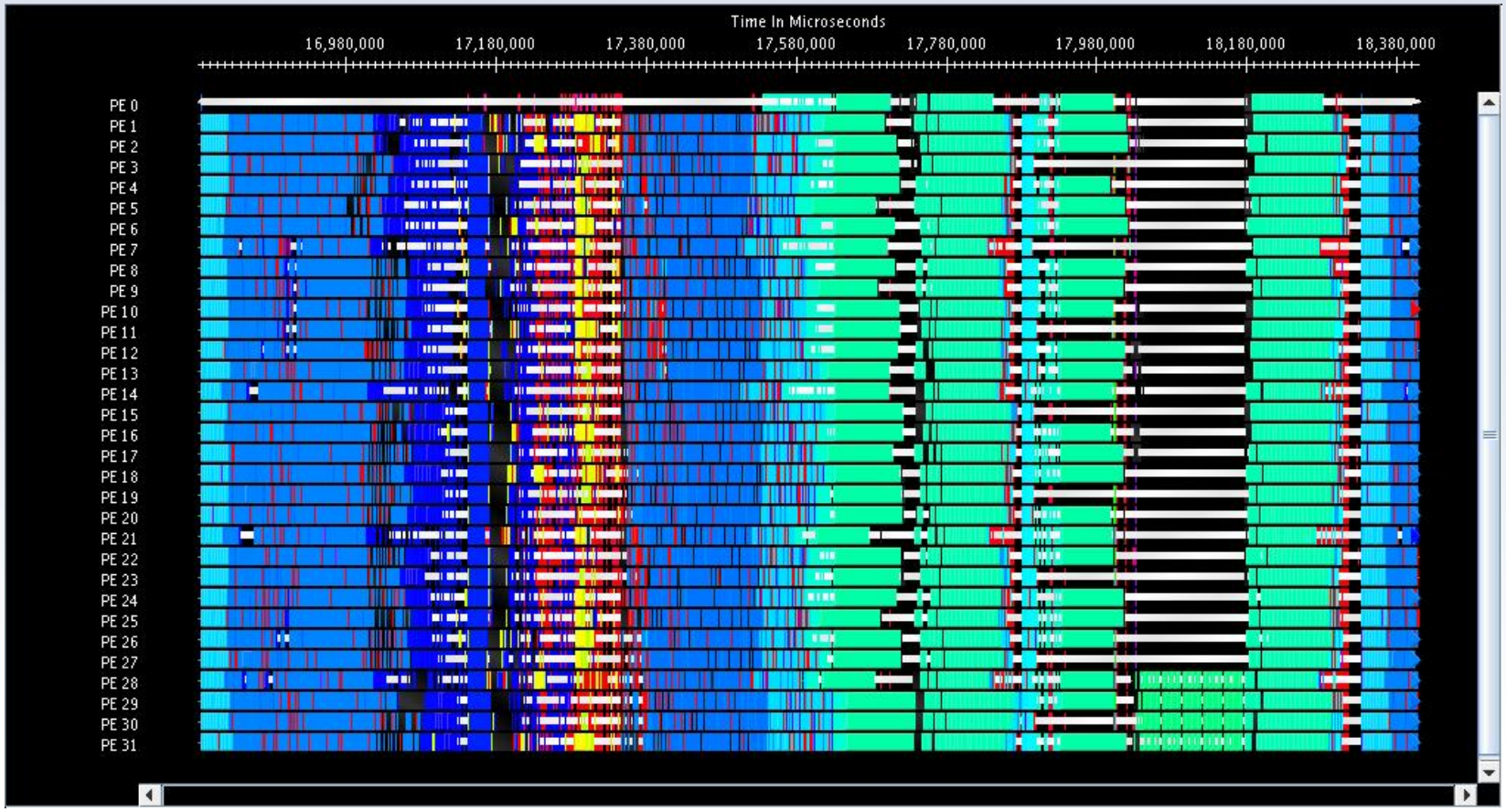
Multi Instance Challenge: cross bead interference



Cross FFT Interference

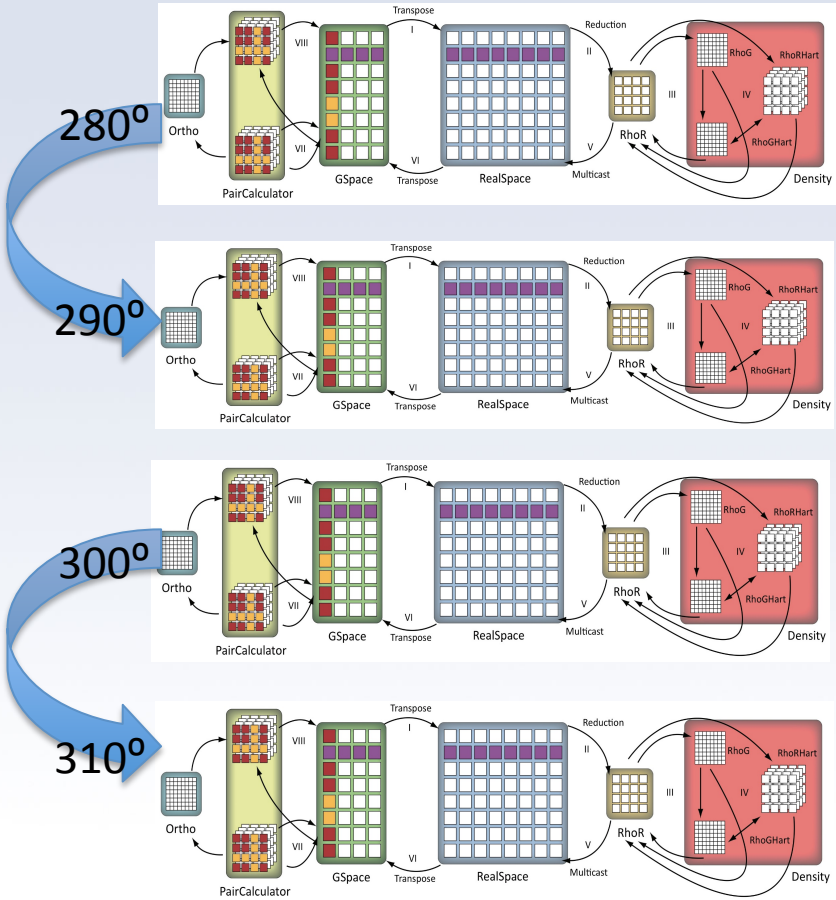


After Scheduling Improvements

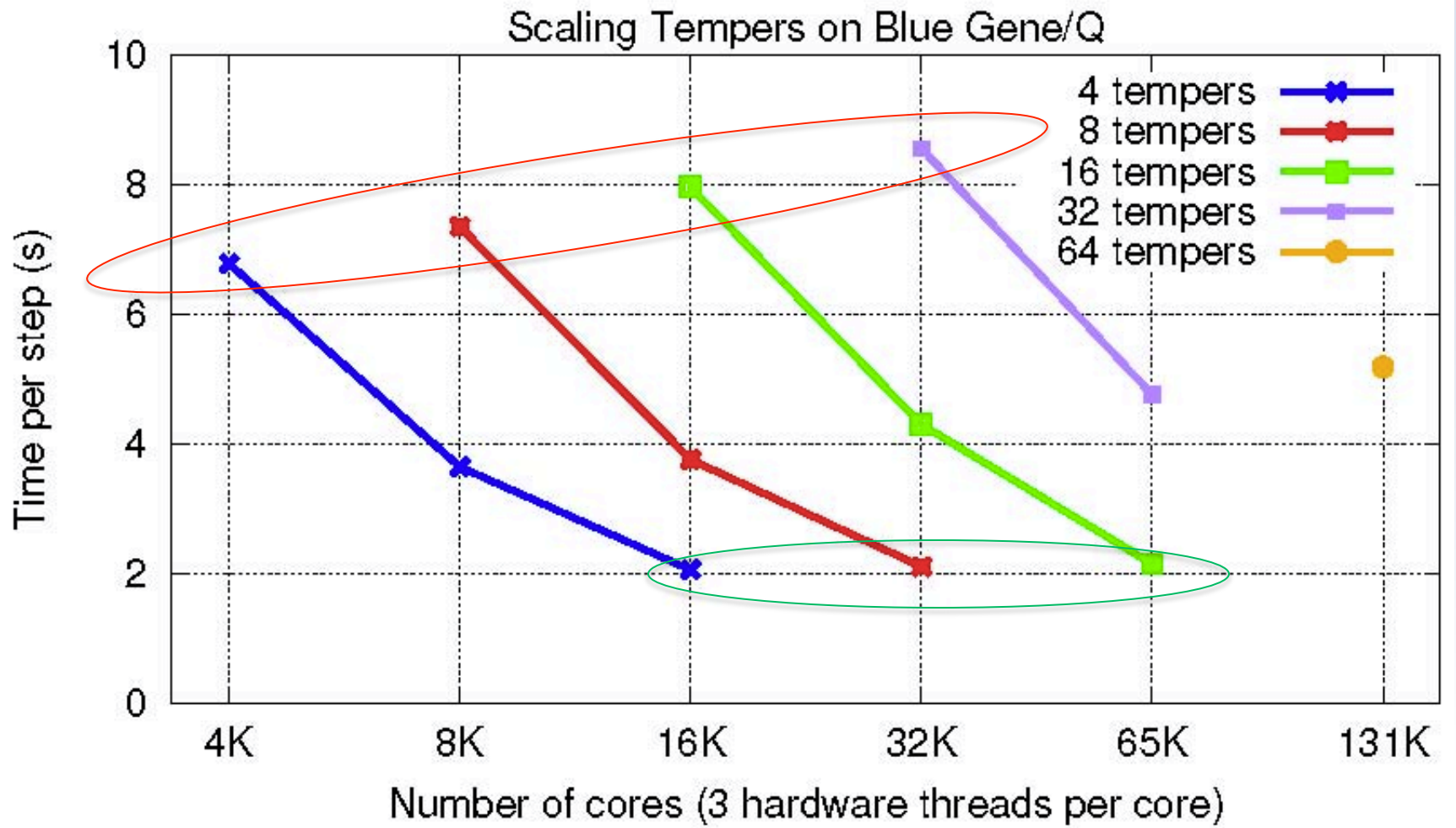


Tempers

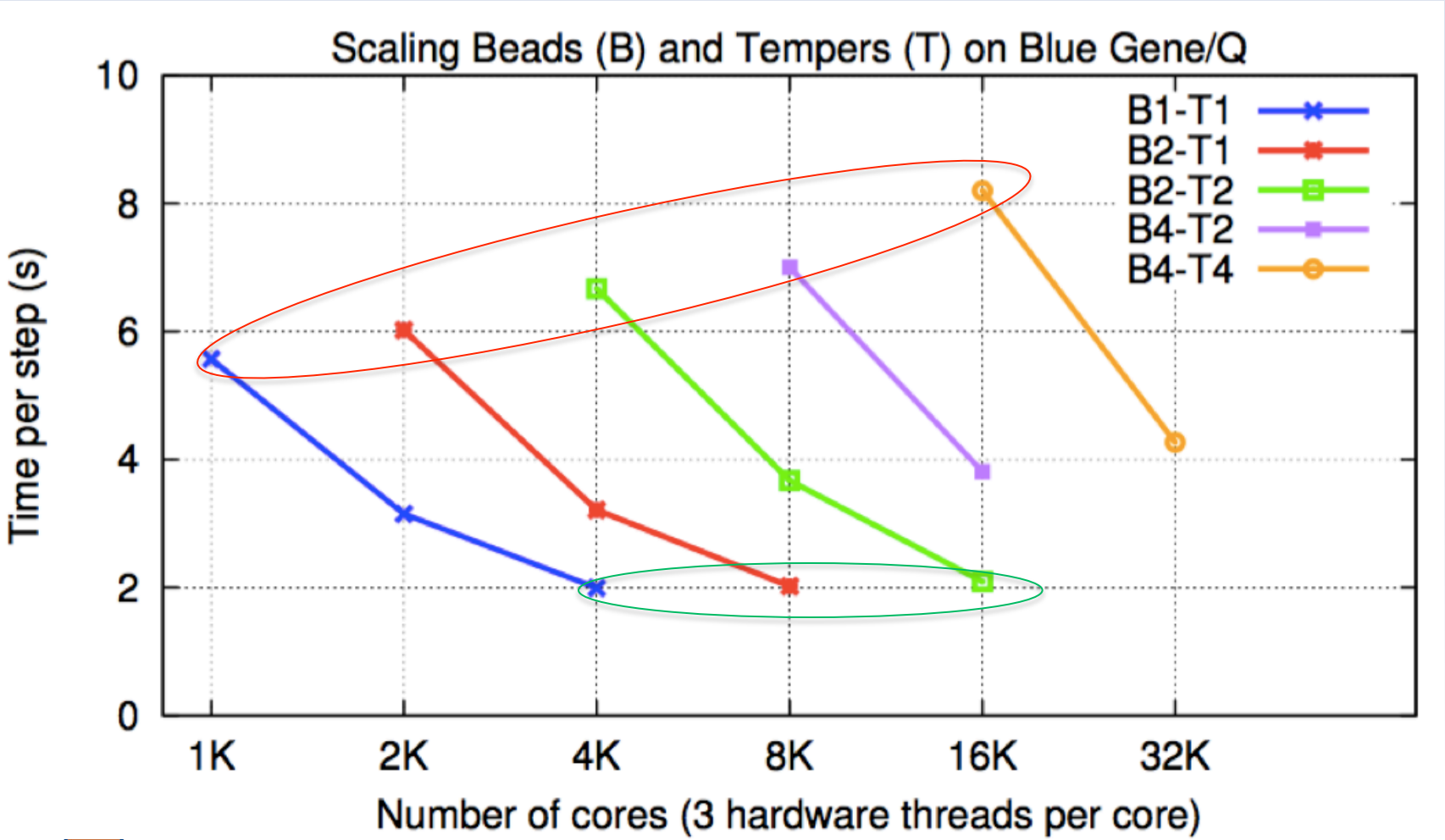
- Contains independent instances of all phases of CPAIMD
- Each temper may contain Beads, K-points, and Spin instances
- Temper controller manages random neighbor shuffle to exchange temperatures across temper replicas
- Independent I/O for state and coordinate data
- Independent placement for instance chares



Temper Performance



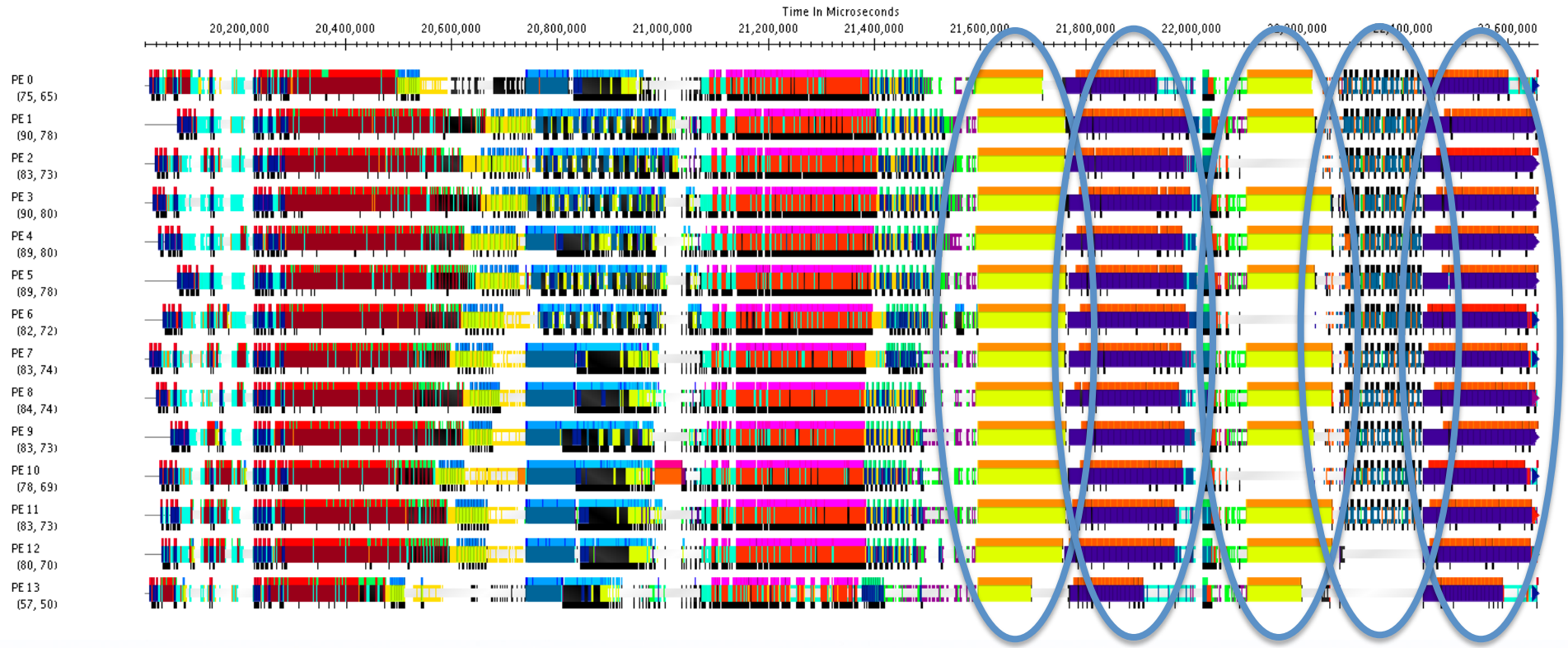
Combined Performance



Michael Robson

GPGPU

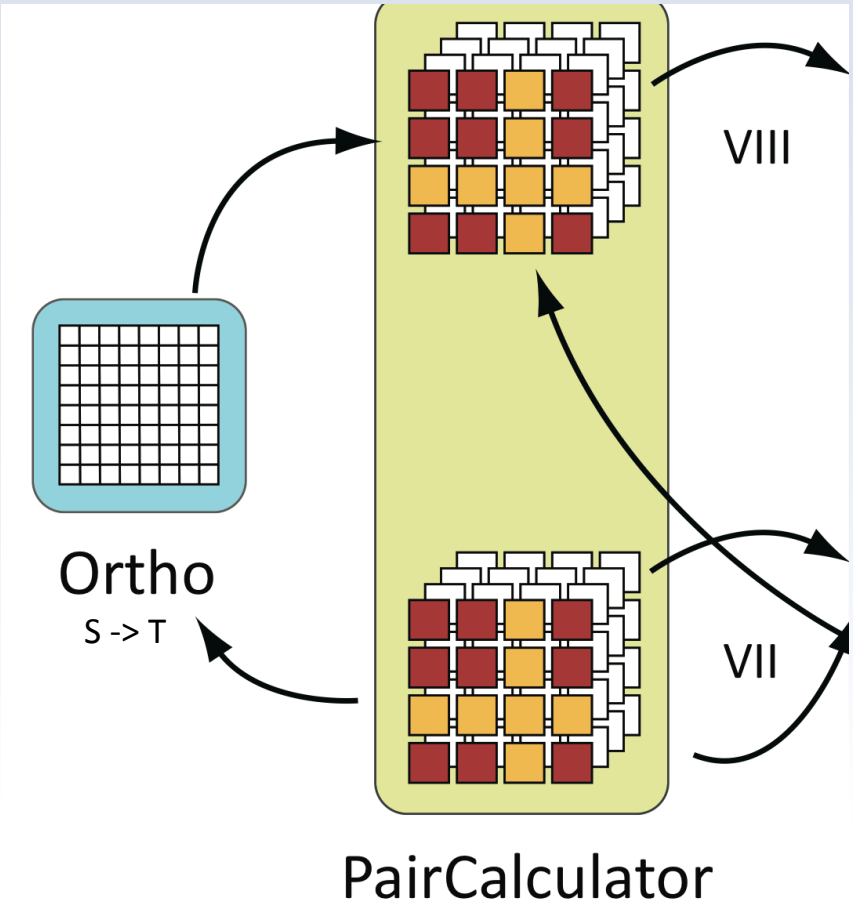
Opportunities for GPUs



PC PC PC ORTHO PC

Ortho Integration

- $T = S^{-1/2}$
- Kernel developed by summer intern
- Changes from distributed to centralized
- Gather pieces of S from PCs



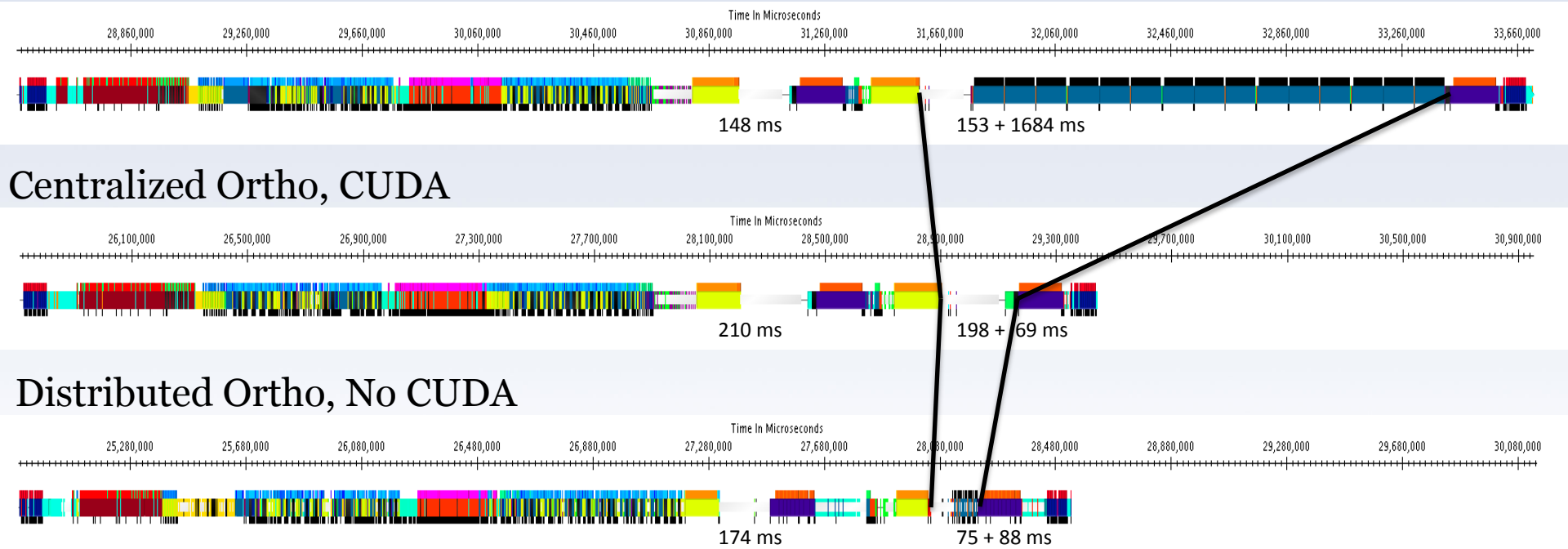
Opportunities for GPUs

- Orthonormalization (Ortho)
  - Inverse Square Root of Matrix
- Pair Calculator (PC)
  - Matrix Matrix Multiplication
- Both make good GPU targets

Results for MOF System

1 iteration, 1PE, 64 XK Nodes of Blue Waters

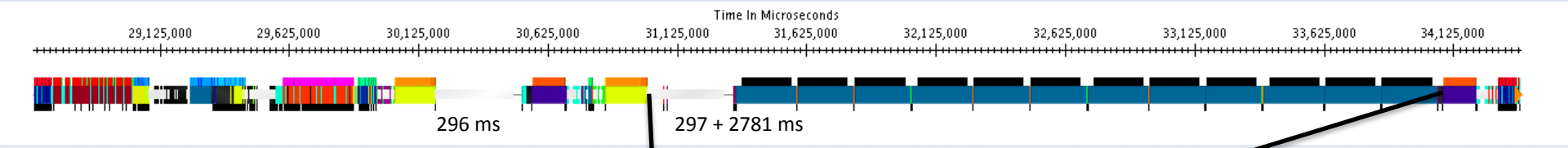
Centralized Ortho, No CUDA



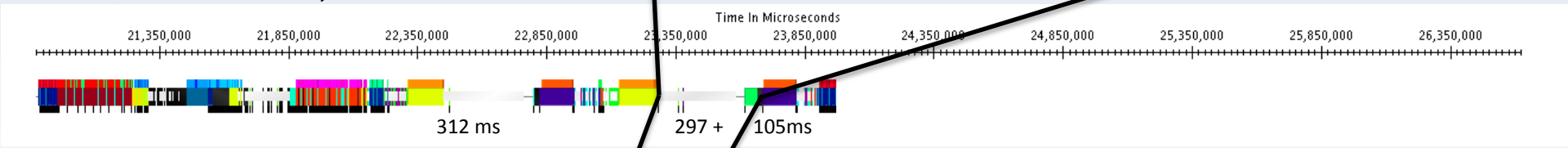
Results for 256 Water Molecules

1 iteration, 1 PE, 64 XK Nodes of Blue Waters

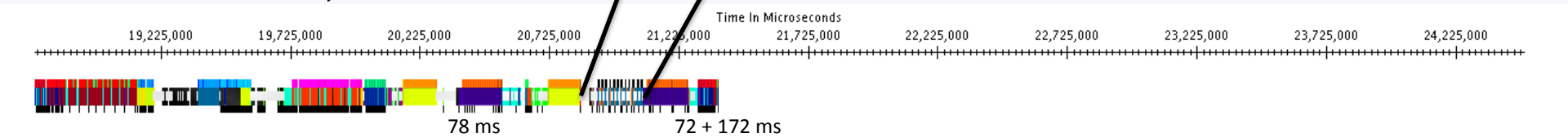
Centralized Ortho, No CUDA



Centralized Ortho, CUDA



Distributed Ortho, No CUDA



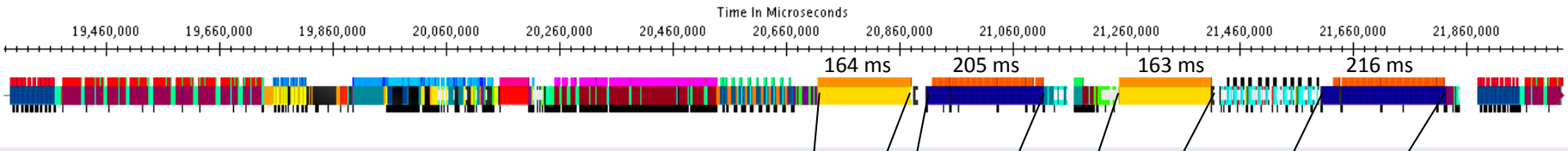
Pair Calculator Results

- Baseline CUBLAS implementation
  - Offload forward and backward multiply
  - CUDA Streams
  - Directly allocate memory
  - Synchronize on data move
  - Fine because we don’t have other work

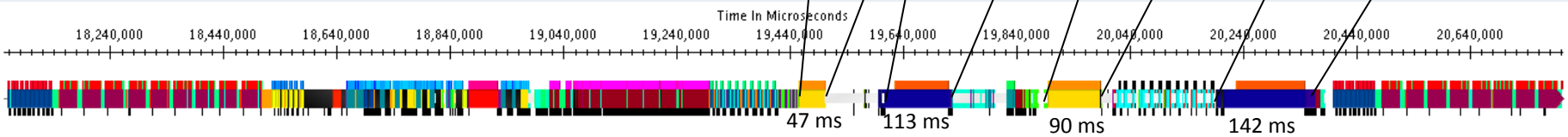
Results for 256 Water Molecules

1 iterations, 1 PE, 64 XK Nodes of Blue Waters

Baseline



CUBLAS



	Forward	Speedup	Backward	Speedup	Per Iter	Speedup
Baseline	327 ms		421 ms		2.582796 s	
CUBLAS	137 ms	2.39x	255 ms	1.65x	2.346571 s	1.10x

Ground State Future Work

- Test automation and feature verification
- Multi-Instance Performance Tuning
- CharmFFT integration for State and Non-local
- Fast Hartree Fock
- PAW
- GPU
  - GPU Manager integration
  - Backward path PC
  - CharmFFT
  - Communication optimization (NVLink, etc)
- Auto-tuning controls in PICS

Towards high scalable GW calculations

Subhasish Mandal[1], Minjung Kim[1], Eric Mikida[2], Eric Bohm[2], Prateek Jindal[2], Nikhil Jain[2], Laxmikant V. Kale[2], Glenn Martyna[3], & Sohrab Ismail-Beigi[1]

[1]Dept. Of Applied Physics, Yale University

[2]Department of Computer Science, University of Illinois at Urbana–Champaign

[3]IBM T. J. Watson research Center

Towards high scalable GW calculations

OpenAtom + GW

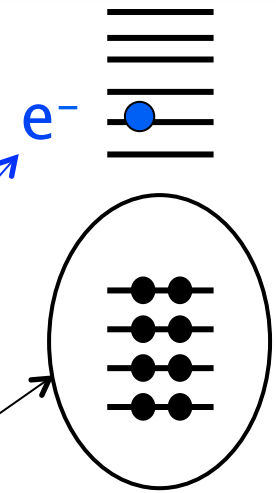
University of Illinois
IBM Watson Research
Yale University

git clone http://charm.cs.illinois.edu/gerrit/openatom.git

Outline

1. Introduction & Motivation of GW
2. Stages of GW calculation
3. Static Polarizibility : Methods & Scaling
4. Computing self-energy $\Sigma(\omega)$
5. Summary

GW-BSE: what is it about?



DFT is a ground-state theory for electrons

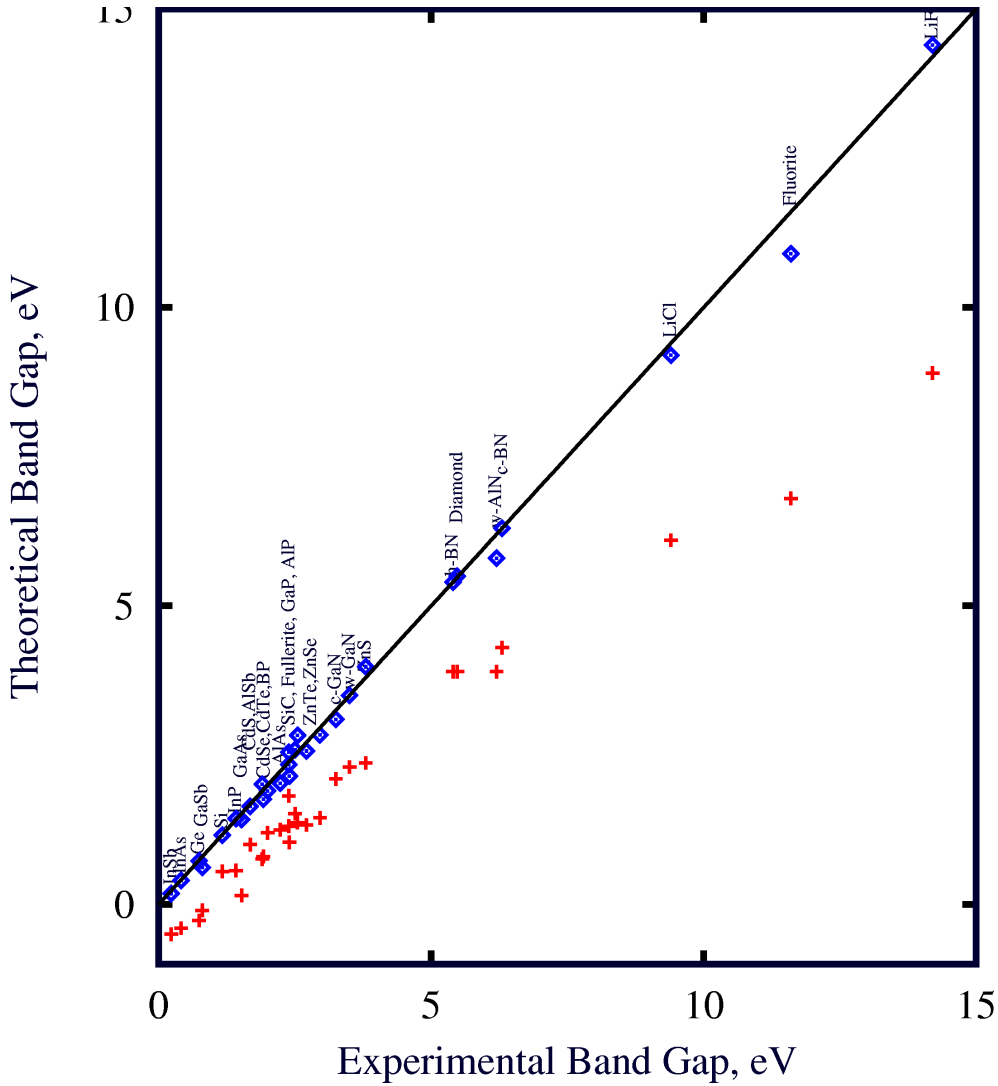
But many processes involve exciting electrons:

- Transport of electrons in a material or across an interface: dynamically adding an electron

→ The *other electrons* respond to this and modify energy of added electron

"GW" method solves this problem

Successes of GW



Energy gaps (eV)

Material	DFT-LDA	GW*	Expt.
Diamond	3.9	5.6	5.48
Si	0.5	1.3	1.17
LiCl	6.0	9.1	9.4

* Hybertsen & Louie, *Phys. Rev. B* (1986)

Motivation (Cont.)

GW-BSE is computationally challenging:

- Huge number of FFTs
- Huge memory footprints
- Large and dense matrix multiplications

Theoretical scaling		Ex. 50-75 atoms (GaN)		
DFT:	N^3	DFT:	1	cpu x hours
GW:	N^4	GW:	91	cpu x hours
BSE:	N^6	BSE:	2	cpu x hours

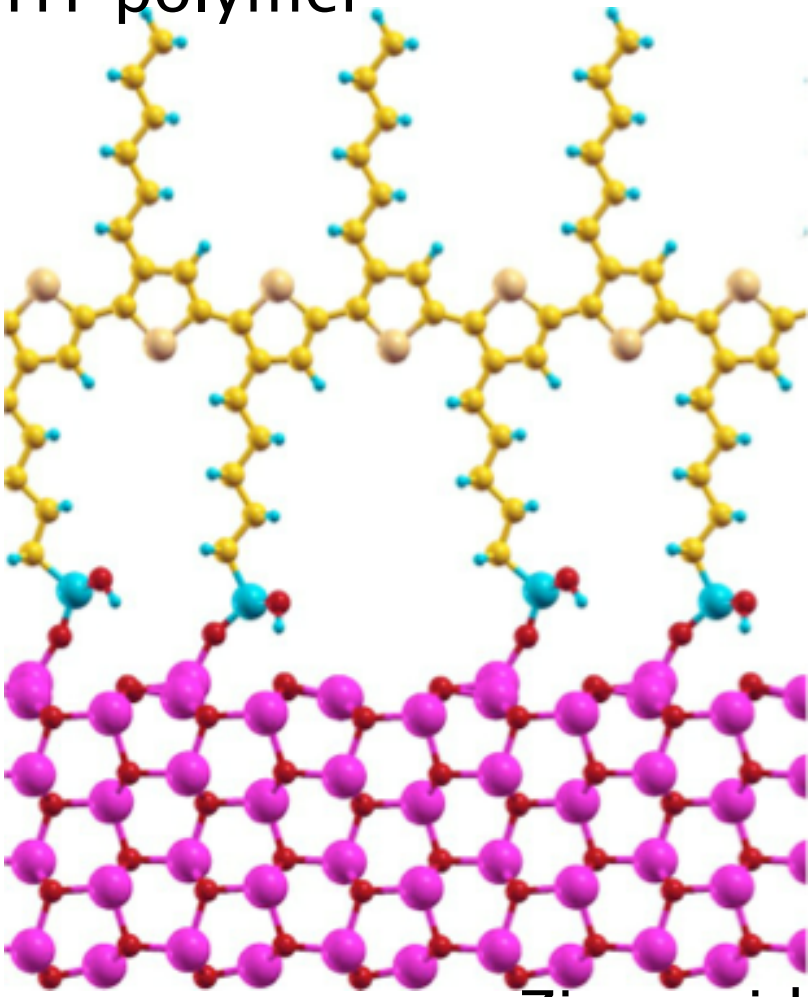
Goal:

Developing highly scalable GW-BSE software to tackle these challenges

Motivation (Cont.)

Would love to do GW-BSE on this photovoltaic system…

P3HT polymer



Zinc oxide nanowire

But with available GW-BSE methods

it would take "forever"

i.e. use up all my supercomputer allocation time

GW

Physicist's favorite : $\mathbf{H}\psi_{n} = E_{n}\,\psi_{n}$

self-energy

Dyson equation for many-body problem:

$$[-\nabla^2/2 + V_H + V_{ion} + \Sigma(E_n)]\ \psi_n = E_n\,\psi_n$$

***GW* Equation:**

$$\Sigma(E) = \int dE'\, G(E-E')\,W(E')$$

Hedin, Phy Rev, **139**, 1965; Hybertsen, Louie, PRB, **34**,1986

Stages of GW calculation

Stage 1 : Run DFT calc. on structure → output : E_i and $\psi_i(r)$

Stage 2.1 : compute Polarizability matrix $P(r,r') = \frac{\partial n(r)}{\partial V(r')}$

Stage 2.2 : double FFT rows and columns → $P(G,G')$

Stage 3 : compute and invert dielectric screening function

$$\epsilon = I - \sqrt{V_{coul}} * P * \sqrt{V_{coul}} \rightarrow \epsilon^{-1}$$

Stage 4 : "plasmon-pole" method → dynamic screening $\rightarrow \epsilon^{-1}(\omega)$

Stage 5 : put together E_i , $\psi_i(r)$ and $\epsilon^{-1}(\omega)$→ self-energy $\Sigma(\omega)$

Hardest parts

Stage 1 : Run DFT calc. on structure → output : E_i and $\psi_i(r)$

Most expensive

Stage 2.1 : compute Polarizability matrix $P(r,r') = \frac{\partial n(r)}{\partial V(r')}$

Stage 2.2 : double FFT rows and columns → *P(G,G')*

Stage 3 : compute and invert dielectric screening function

$$\epsilon = I - \sqrt{V_{coul}} * P * \sqrt{V_{coul}} \rightarrow \epsilon^{-1}$$

Stage 4 : "plasmon-pole" method → dynamic screening $\rightarrow \epsilon^{-1}(\omega)$

Stage 5 : put together E_i , $\psi_i(r)$ and $\epsilon^{-1}(\omega)$→ self-energy $\Sigma(\omega)$

Static Polarizability

$$P(G,G') = \sum_{v,c} \boxed{\langle c | e^{-iG\bullet r} | v \rangle} \langle v | e^{iG'\bullet r} | c \rangle \frac{2}{\varepsilon_v - \varepsilon_c}$$

$$FFT\ [\psi_c^*(r)\psi_v(r)]$$

<u>Standard G-space approach</u>:

- Directly compute *P* in G space
- A huge number of FFTs ($N_v N_c$)

N_v : # occupied states
N_c : # unoccupied states

<u>R-space approach</u>: **Much fewer number of FFTs**

Static Polarizability

New R-space approach:

1) Calculate static polarizability in real space

$$P(r,r') = \sum_{v,c} \psi_c^*(r)\psi_v(r)\psi_v^*(r')\psi_c(r') \frac{2}{\varepsilon_v - \varepsilon_c}$$

2) FFTs for columns

$$P(r,r') \longrightarrow P(G,r')$$ # FFTs: N_r

3) FFTs for rows

$$P(G,r') \longrightarrow P(G,G')$$ # FFTs: N_r

N_r: # R grid

$N_r \approx 4Nc$

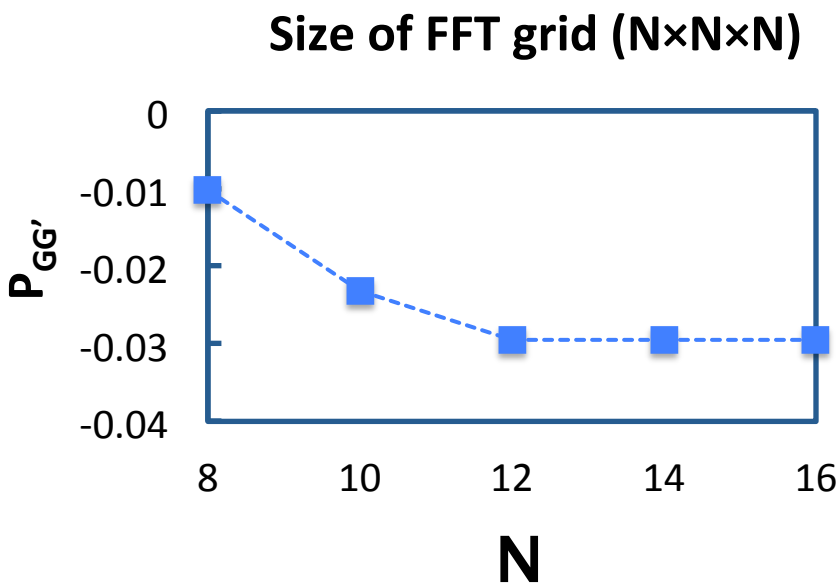
	G-space approach	R-space approach
#FFT	N_vN_c	$8N_c$
N_v=500 N_c=1,500	750,000	12,000

N_v : # occupied states
N_c : # unoccupied states

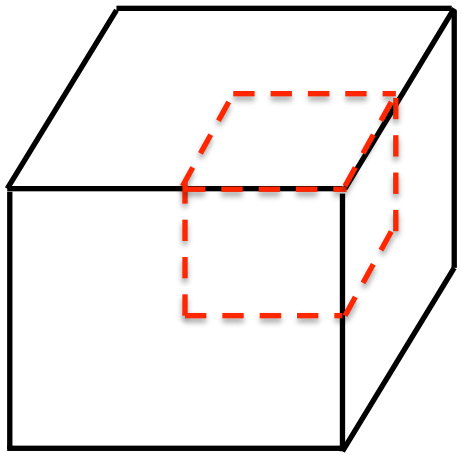
Static Polarizability

What R grid should we use?:

$$P(r,r') = \sum_{v,c} \psi_c^*(r)\psi_v(r)\psi_v^*(r')\psi_c(r') \frac{2}{\varepsilon_v - \varepsilon_c}$$



3D FFT box



reduced to 1/8 !!

- Si crystal
- Dense FFT grid: 24×24×24
- P and ε^{-1} converge at: 12×12×12

Scaling of Polarizibility

$$P(r,r') = \sum_{v,c} \underbrace{\psi_c^*(r)\psi_v(r)}_{f\downarrow\uparrow = \psi\downarrow c\uparrow * \times \psi\downarrow v\uparrow} \psi_v^*(r')\psi_c(r') \frac{2}{\varepsilon_v - \varepsilon_c}$$

Test system

- Si bulk
- Total number of atom : 54
- 108 occupied states
- 892 unoccupied states
- 3 k-points
- **289,008 total f vectors**
- Each state and f vector ~0.5 MB

Performance

* Results on 1024 nodes of Vesta, 32 threads per node

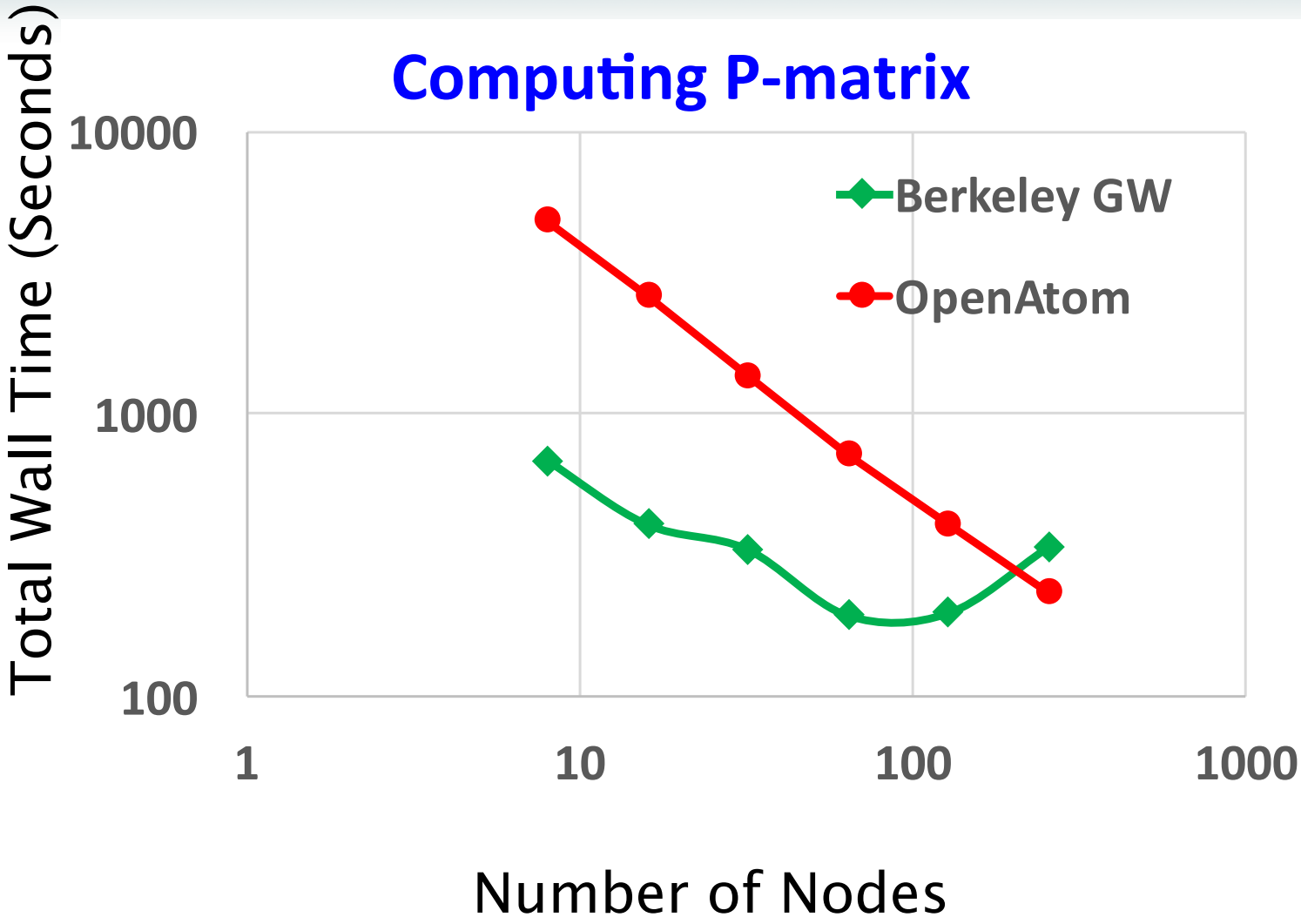
$$P(r,r') = \sum_{v,c} \psi_c^*(r)\psi_v(r)\psi_v^*(r')\psi_c(r')\frac{2}{\varepsilon_v - \varepsilon_c}$$

$$f\downarrow r\uparrow vc = \psi\downarrow r\uparrow v * \times$$

Calculation	Time (seconds)
	No SMP, BLAS 2 (1D decomposition)
Compute f vectors for a single unocc state	0.08
Total for 2,676 sets of f vectors	378.04
Total wall time	**416.22**
Overhead (comm, synchronization)	38.18

SMP: shared memory programming: takes advantage of node level parallelism

Performance



32 Threads per node on Vesta

GW-Self-Energy (Σ)

Stage 5 : Put together E_i , $\psi_i(r)$ and $\epsilon^{-1}(\omega)$ → self-energy $\Sigma(\omega)$

GW-Static Self-Energy (COHSEX)

Static self-energy approximation: a useful simplification

$$W(r,r',E) \rightarrow W(r,r',E=0)=W(r,r')$$

$$Hqp\uparrow = HKS + \Sigma\uparrow X + \Sigma\uparrow SEX + \Sigma\uparrow COH - V\downarrow xc$$

GW-Static Self-Energy (COHSEX)

Static self-energy approximation: a useful simplification

$$W(r,r',E) \rightarrow W(r,r',E=0)=W(r,r')$$

Band Gaps (eV)			
System	**Experiment**		**GW(full)**
Diamond	5.48		5.70
Si	1.17		1.29

*Kang & Hybertsen, Phy Rev B, **82**, 2010

GW-Bare Exchange

$$\langle nk|\Sigma_x|n'k\rangle^q = -\sum_{l=1}^{L}\sum_{g}\int dr \int dr'\; v(g)\; e^{-ig\cdot r} f_r^{nl} f_{r'}^{n'l}{}^{*}\; e^{ig\cdot r'}$$

Coulomb

$$\psi_r^n \times \psi_r^l{}^{*}$$

Strategy 1:
(Motivated from P(G,G'))

$$B_{r,r'} = \sum_{l=1}^{L} \psi_{r,l}\psi_{r',l}{}^{*}$$

Double FFT

$$\boxed{Dgg'}$$

$$n|\Sigma X\uparrow|n'\uparrow> = \sum g\uparrow \text{▒} vc\uparrow(g) D\downarrow g,g$$

$$n,n'\uparrow$$

Coded and tested

Strategy 2:

$$f\downarrow r\uparrow n,l$$

Single FFT

$$f\downarrow g\uparrow n,l$$

$$<n|\Sigma X\uparrow|n'\uparrow> = \sum g,l\uparrow \text{▒} vc\uparrow(g)$$

$$f\downarrow g\uparrow n'\uparrow,l * f\downarrow g\uparrow n,l\uparrow$$

Coded and tested

GW-Bare Exchange

GW In-house code

Strategy 1:

Real time required on Si calculation:
135.369 sec

FFT of rank 2 matrix

Strategy 2:

Real time required on Si calculation:
2.783 ssec

FFT of vector

Strategy 2 wins

Screened Exchange: followed with Strategy 2

Coulomb-Hole: followed with Strategy 2

Summary

- New R-space approach of Polarizibility
- Polarizibility is parallelized with Charm++ & take the advantage of SMP
- We have investigated various ways to calculate static self-energy.
- Strategy 1 is effective for P(G,G') But for COHSEX Strategy 2 wins

Acknowledgement

NCSA & ALCF@Argonne National Laboratory.

QUESTION ?

Static polarizability calculations

$$\varepsilon(G,G') = \delta_{G,G'} - \sqrt{V_G}\,P(G,G')\sqrt{V_{G'}}$$

The most time consuming part

$$P(G,G') = \sum_{v,c} \boxed{\langle c|e^{-iG\bullet r}|v\rangle}\langle v|e^{iG'\bullet r}|c\rangle \frac{2}{\varepsilon_v - \varepsilon_c}$$

$$FFT\ [\psi_c^*(r)\psi_v(r)]$$

<u>Standard G-space approach</u>:

- Directly compute *P* in G space
- A huge number of FFTs ($N_v N_c$)

Nv : # occupied states

Nc : # unoccupied states

<u>R-space approach</u>: **Much fewer number of FFTs**

GW-Static Self-Energy (COHSEX)

Static self-energy approximation: a useful simplification

$$W(r,r',E) \rightarrow W(r,r',E=0)=W(r,r')$$

$$\Sigma^{X}(r,r') = -\sum_{n}^{occ} \psi_n(r)\psi_n^*(r')v_c(r,r')$$

$$\Sigma^{SEX}(r,r') = -\sum_{n}^{occ} \psi_n(r)\psi_n^*(r')[W(r,r') - v$$

$$\Sigma^{COH}(r,r') = 1/2\ \delta(r-r')[W(r,r') - v_c(r,r')]$$

$$H_{qp}^{\uparrow} = H_{KS} + \Sigma^{X} + \Sigma^{SEX} + \Sigma^{COH} - V_{xc}$$

GW-Bare Exchange

$$\langle nk|\Sigma_x|n'k\rangle^q = -\sum_{l=1}^{L}\sum_{g}\int dr \int dr'\, v(g)\; e^{-ig\cdot r}\; f_r^{nl} f_{r'}^{n'l} * \; e^{ig\cdot r'}$$

Coulomb

$$\psi_r^n \times \psi_r^l *$$

Strategy 1: Motivated from P(G,G')

Sum over l & compute:

$$B_{r,r'} = \sum_{l=1}^{L} \psi_{r,l}\psi_{r',l}\,*$$

Double FFT

$$Dgg' = \int Br,r'\; e\uparrow -ig.r\; e\uparrow ig'.r\; \psi\downarrow n,r\uparrow * \downarrow\; \psi\downarrow n',r'\; drdr'\downarrow$$

$$<n|\Sigma X\uparrow\; |n'\uparrow> = \sum g\uparrow\; vc\uparrow\; (g) D\downarrow g,g'\uparrow n,n'\quad \uparrow$$

Strategy 1: **Coded and tested**

GW-Bare Exchange

Strategy 2:

$$\langle nk|\Sigma_x|n'k\rangle^q = -\sum_{l=1}^{L}\sum_{g}\int dr \int dr'\, v(g)\; e^{-ig\cdot r}\; f_r^{nl} f_{r'}^{n'l} * e^{ig\cdot r'}$$

Compute f_r (vector) and do FFT:

$$f_{g,l}^{n} = \int f_{r,l}\, e^{-ig.r}\, \psi_{n,r}$$

Multiply: $v_g\, f_g\, f_{g'}$ and take sum over g

$$\langle n|\Sigma X|n'\rangle = \sum_{g,l} v_c(g)\, f_{g,l}^{n'} * f_{g,l}^{n}$$

Strategy 2: **Coded and tested**

Density Functional Theory

For the ground-state of an interacting electron system

we solve a Schrodinger-like equation for electrons

$$\left[-\frac{\hbar^2\nabla^2}{2m_e} + V_{ion}(r) + \phi(r) + V_{xc}(r)\right]\psi_j(r) = \epsilon_j\,\psi_j(r)$$

Approximations needed for $V_{xc}(r)$: LDA, GGA, *etc.*

Tempting: use these electron energies ϵ_j
to describe processes where
electrons change energy
(absorb light, current flow, etc.)

Hohenberg & Kohn, *Phys. Rev.* (1964); Kohn and Sham, *Phys. Rev.* (1965).

Performance

Step 2 – FFT P to G-Space

- FFT each row of P locally: $P(r,r') \rightarrow P(G,r')$
- Transpose P
- FFT each row locally again
- Transpose P again: $P(G,r') \rightarrow P(G,G')$
- Total time: 13.93s

* Results on 1024 nodes of Vesta, 32 threads per node

GW-Bare Exchange

GW In-house code

Strategy 1:

FFT of rank 2 matrix

Real time required on Si calculation:
135.369 sec

Strategy 2:

FFT of vector

Real time required on Si calculation:
2.783 ssec

Strategy 2 wins

Screened Exchange: $\langle n|\Sigma^{sex}|n'\rangle = \sum_{l,L} \sum_{g,g'} f_{g,n,l} S_{g,g'} f_{g',n',l}$

$S_{g,g'} = \sqrt{v_g}\, [(\varepsilon^{-1})_{g}$

Coulomb-Hole: $\langle n|\Sigma^{COH}|n'\rangle = \sum_{l,L} \sum_{g,g'} S_{g,g'} Y_{n,n'}$

$Y_{n,n'}(g) = \int dr\, \psi'$

Summary:

GW-Static Self-Energy (COHSEX)

Static self-energy approximation: a useful simplification

$$W(r,r',E) \rightarrow W(r,r',E=0)=W(r,r')$$

$$\Sigma^{X}(r,r') = -\sum_{n}^{occ} \psi_n(r)\psi_n^*(r')v_c(r,r')$$

$$\Sigma^{SEX}(r,r') = -\sum_{n}^{occ} \psi_n(r)\psi_n^*(r')[W(r,r') - v_c(r,r$$

$$\Sigma^{COH}(r,r') = 1/2\ \delta(r-r')[W(r,r') - v_c(r,r')]$$

$$H_{qp}^{\uparrow} = H_{KS} + \Sigma^{X} + \Sigma^{SEX} + \Sigma^{COH} - V_{xc}$$

Band Gaps (eV)					
System	**Experiment**	**DFT-LDA**	**COHSEX**	**Corrected COHSEX***	**GW(full)**
Diamond	5.48	4.20	6.99	5.93	5.70
Si	1.17	0.49	1.70	1.18	1.29

*Kang & Hybertsen, Phy Rev B, **82**, 2010

Scaling of Polarizibility

Most expensive

Stage 2.1 : compute Polarizability matrix $P(r, r') = \frac{\partial n(r)}{\partial V(r')}$

Stage 2.2 : double FFT rows and columns → *P(G,G')*

GW-Bare Exchange

Strategy 1:

Computational load (on pen & paper)

Step 1: $B_{r,r'} \sim 10^5 \times N_L^3$

Step 3: $\sum_g v_g D_{g,g} \sim 2 \times N_L^3$

Step 2: $D_{g,g'} \sim 10^5 \times N_L^4$

N_L: # of occupied bands

Strategy 2:

Step 1: $f_g^{nl} f_{g'}^{n'l*} \sim 6 \times 10^3 \times N_L^3$

Step 2: $\Sigma_{l,g} v_g f_g^{n,l} f_{g'}^{n',l} \sim 1.7 N_L^4$

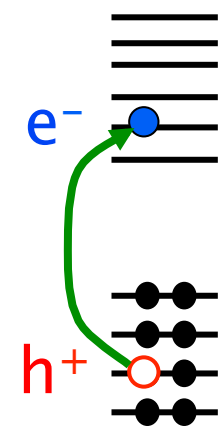
GW-BSE: what is it about?

DFT is a ground-state theory for electrons

But many processes involve <u>exciting</u> electrons:

- Transport of electrons

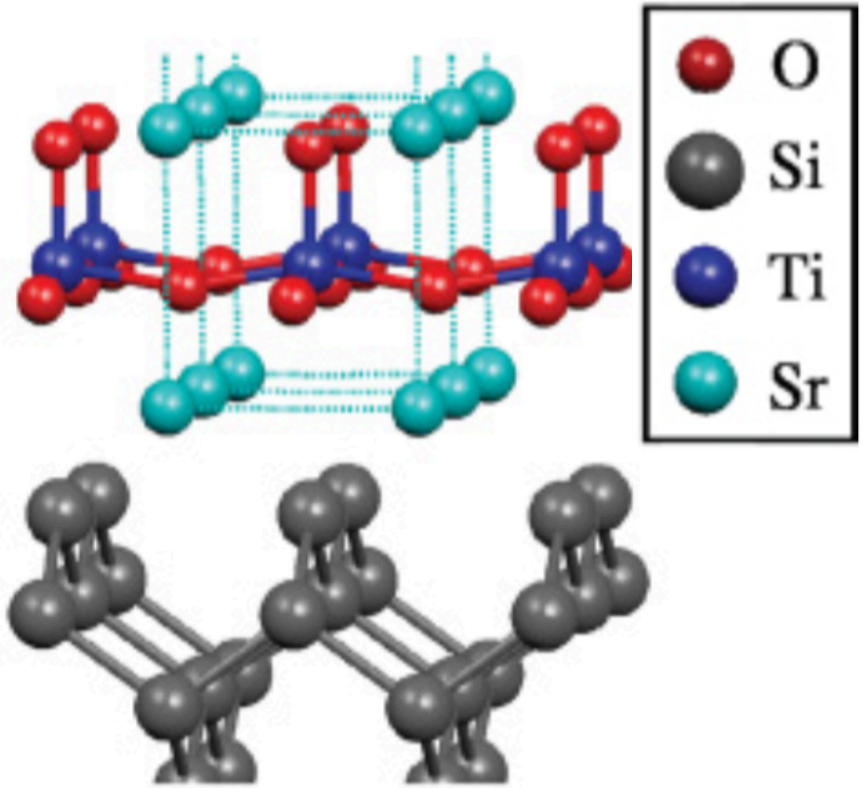
- Excited electrons: optical absorption promotes electron to higher energy



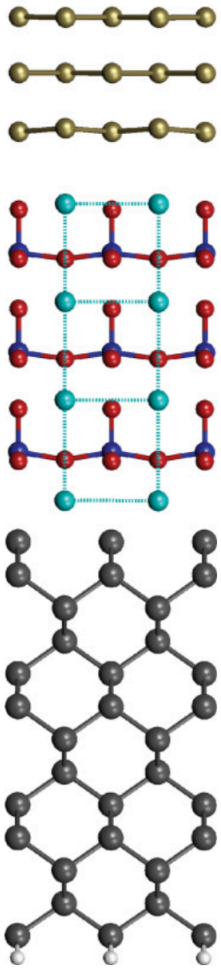
→ The missing electron (hole)
has + charge, *attracts* electron:
<u>modifies</u> excitation energy and absorption strength

"BSE" method solves this problem

Would love to do GW on this interfacial oxide/semiconductor system…

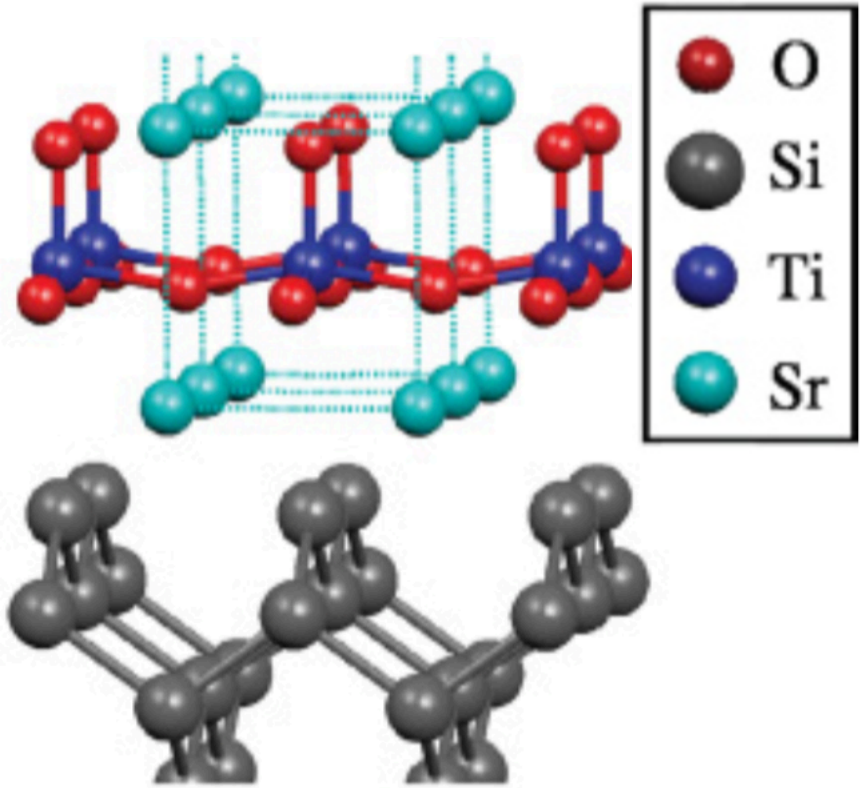


Si interface with $SrTiO_3$

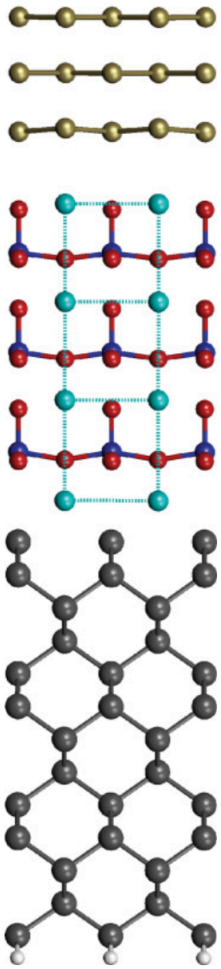


Full unit cell

Would love to do GW on this interfacial oxide/semiconductor system…



Si interface with $SrTiO_3$



Full unit cell

GW-Bare Exchange

Biggest Computational load (on pen & paper)

Strategy 1:

$$D_{g,g'} :\sim 10^5 \times N_L^4$$

N_L: # of occupied bands

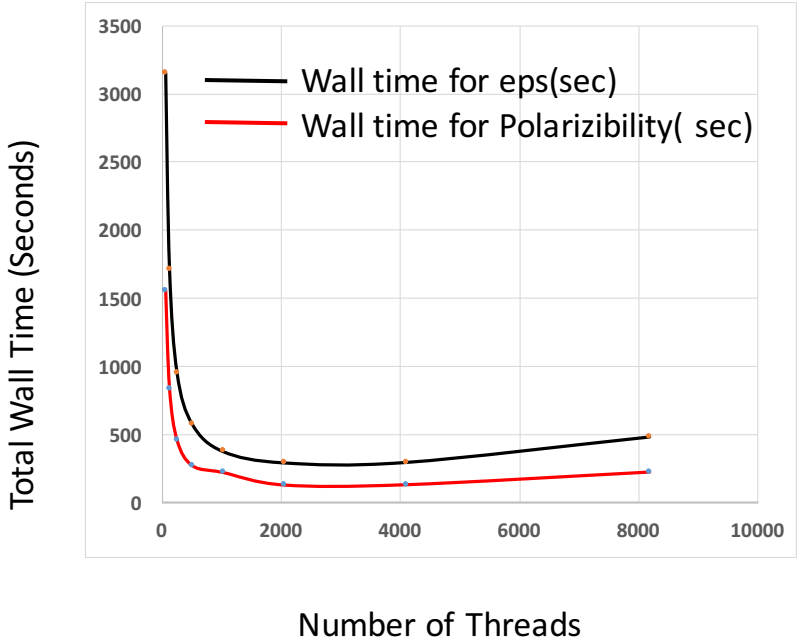
Strategy 2:

$$\Sigma_{l,g} v_g f_g^{n,l} f_{g'}^{n',l} :\sim 1.7 N_L^4$$

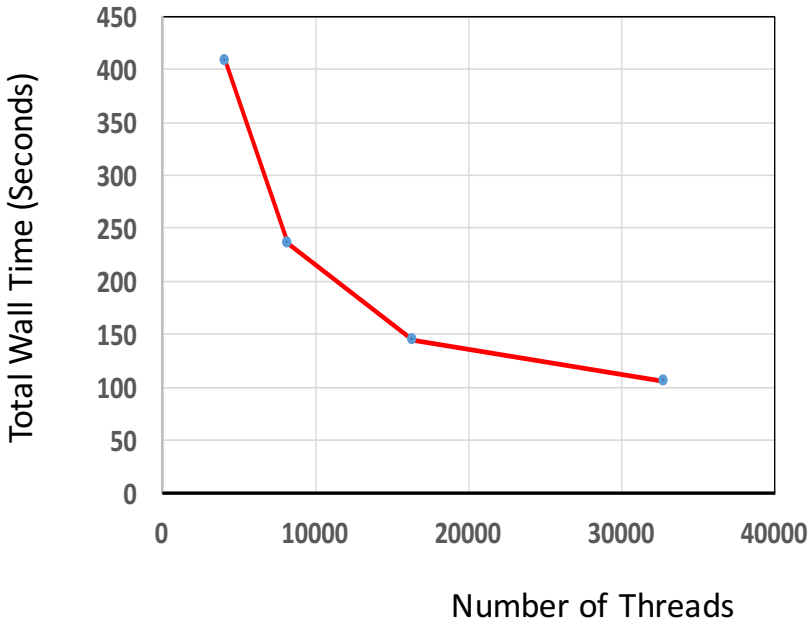
Performance

Step 1 – P in R-Space

Berkeley GW



OpenAtom

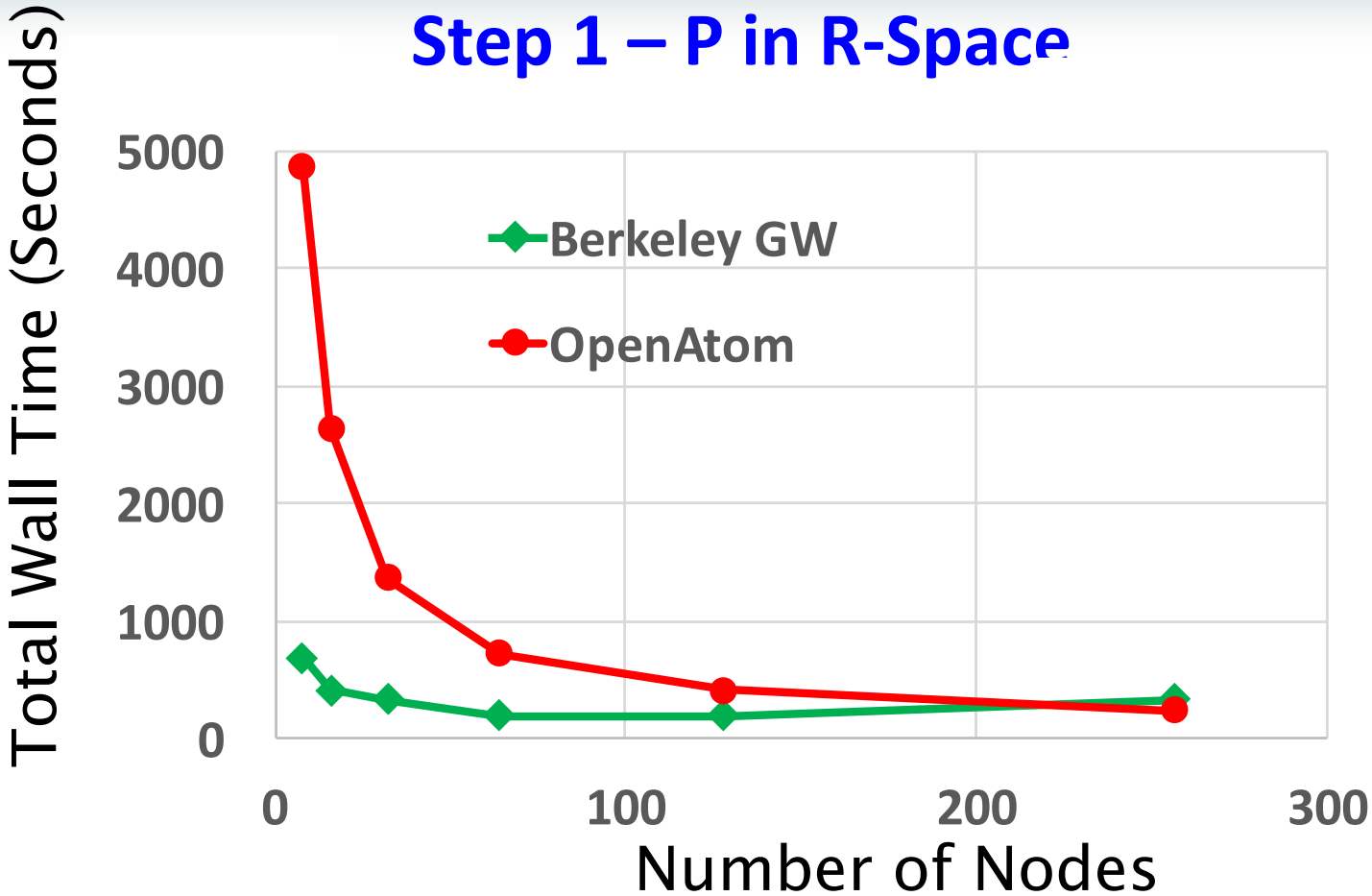


Step 2 – FFT P to G-Space

….Work in progress

Performance

Step 1 – P in R-Space



32 Threads per node on Vesta

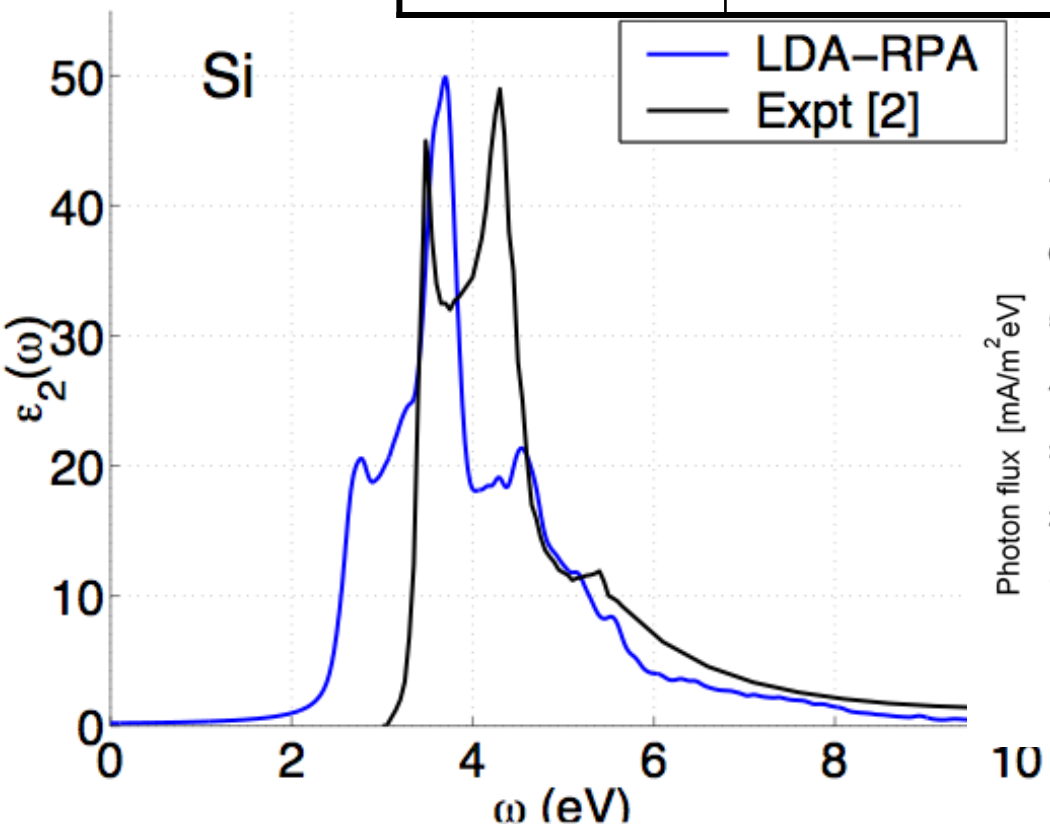
Step 2 – FFT P to G-Space

….Work in progress

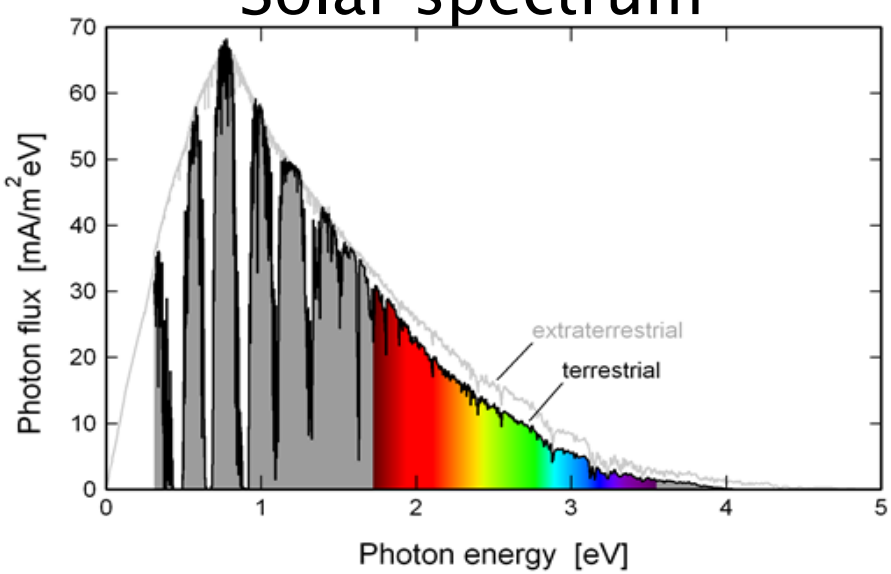
DFT: problems with excitations

Energy gaps (eV)

Material	DFT-LDA	Expt. [1]
Diamond	3.9	5.48
Si	0.5	1.17
LiCl	6.0	9.4



Solar spectrum



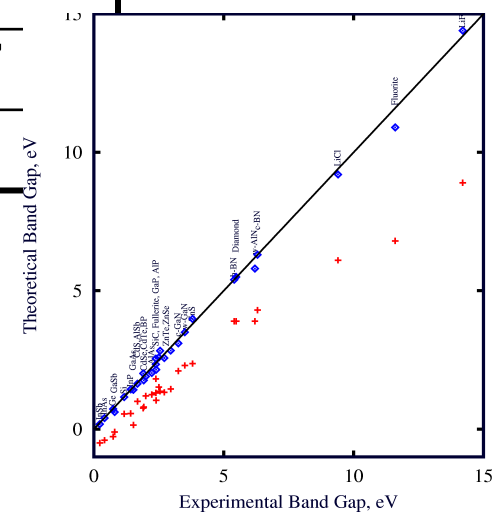
[1] Landolt-Bornstien, vol. III; Baldini & Bosacchi, *Phys. Stat. Solidi* (1970).

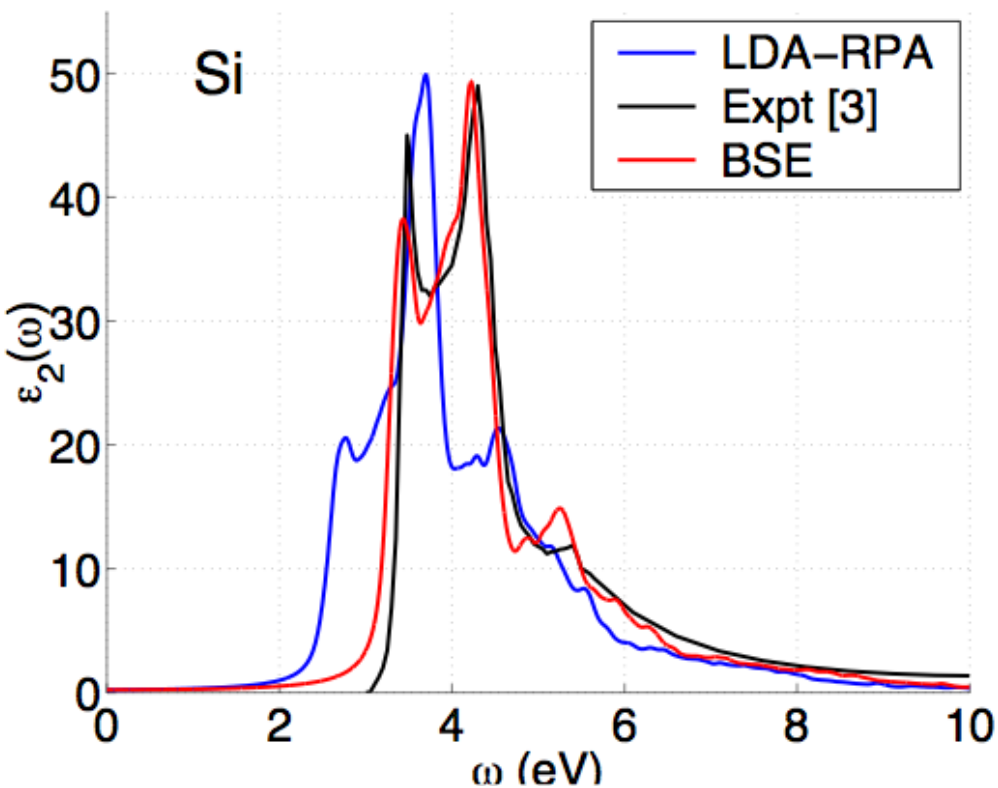
[2] Aspnes & Studna, *Phys. Rev. B* (1983)

Green's functions successes

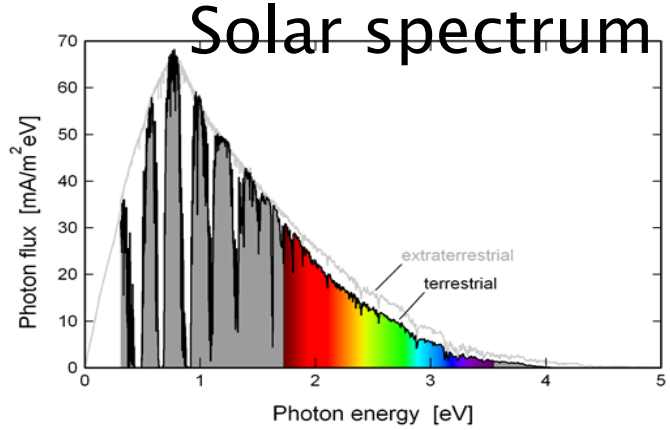
Energy gaps (eV)

Material	DFT-LDA	GW*	Expt.
Diamond	3.9	5.6	5.48
Si	0.5	1.3	1.17
LiCl	6.0	9.1	9.4





Solar spectrum



* Hybertsen & Louie, *Phys. Rev. B* (1986)