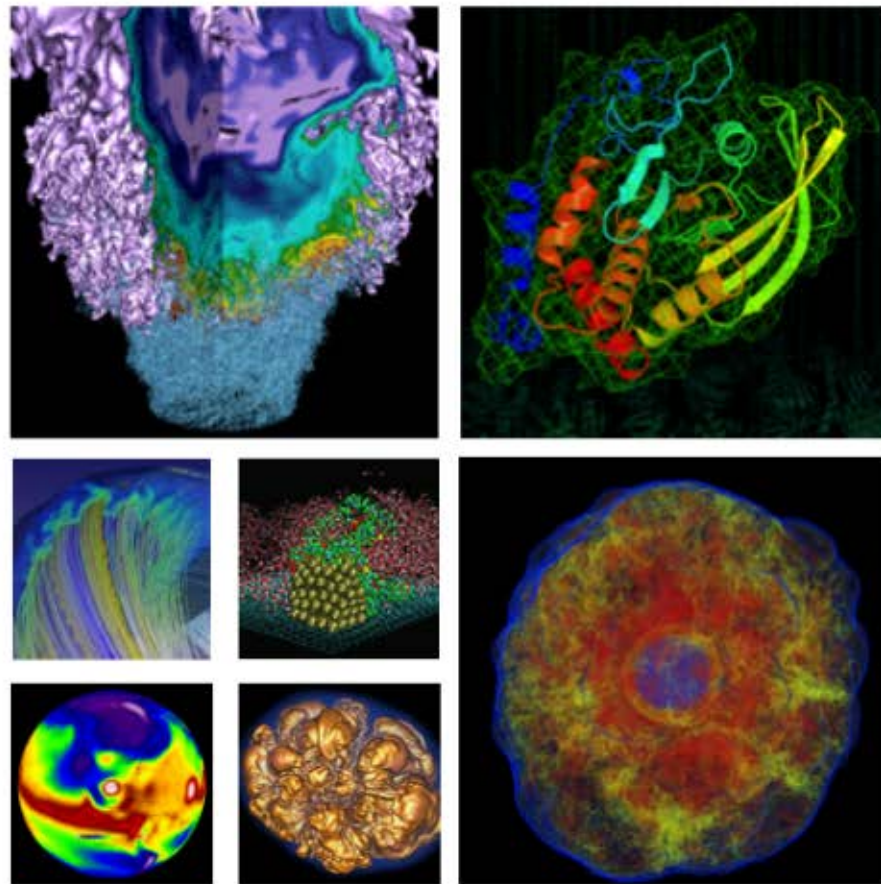


Scalable Computational Tools for Discovery and Design -- Excited State Phenomena in Energy Materials



Jack Deslippe (Team Rep)



Jim Chelikowsky
UT Austin



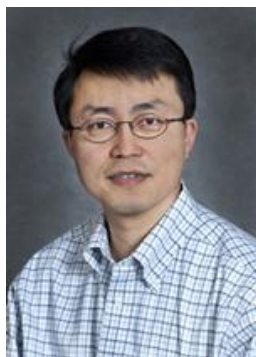
Jeff Neaton
LBNL / UC
Berkeley



Steven Louie
LBNL / UC
Berkeley



Yousef Saad
U. of Minnesota



Chao Yang
LBNL



Andrew Canning
LBNL



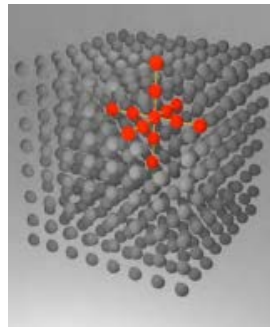
Alex Demkov UT
Austin



Jack Deslippe
LBNL

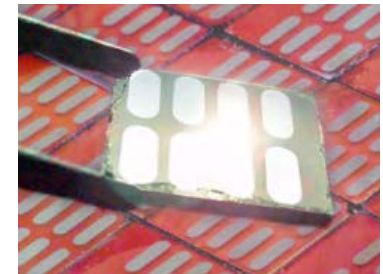
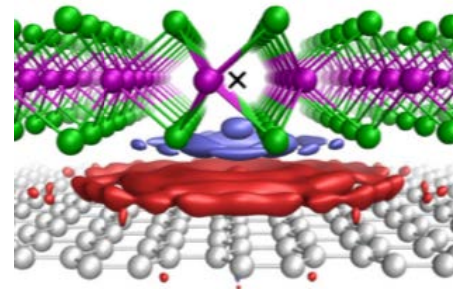
PARSEC

- **Massively parallel real-space DFT code.**
- **Capable of studying systems of 10K atoms.**
- **Implements spectrum slicing approach for parallel eigenstate generation.**



BerkeleyGW

- **Massively parallel excited state (both one- and two-particle) code.**
- **Computes quasiparticle and optical properties of materials of interest to the DOE.**



Who am I?



Jack Deslippe



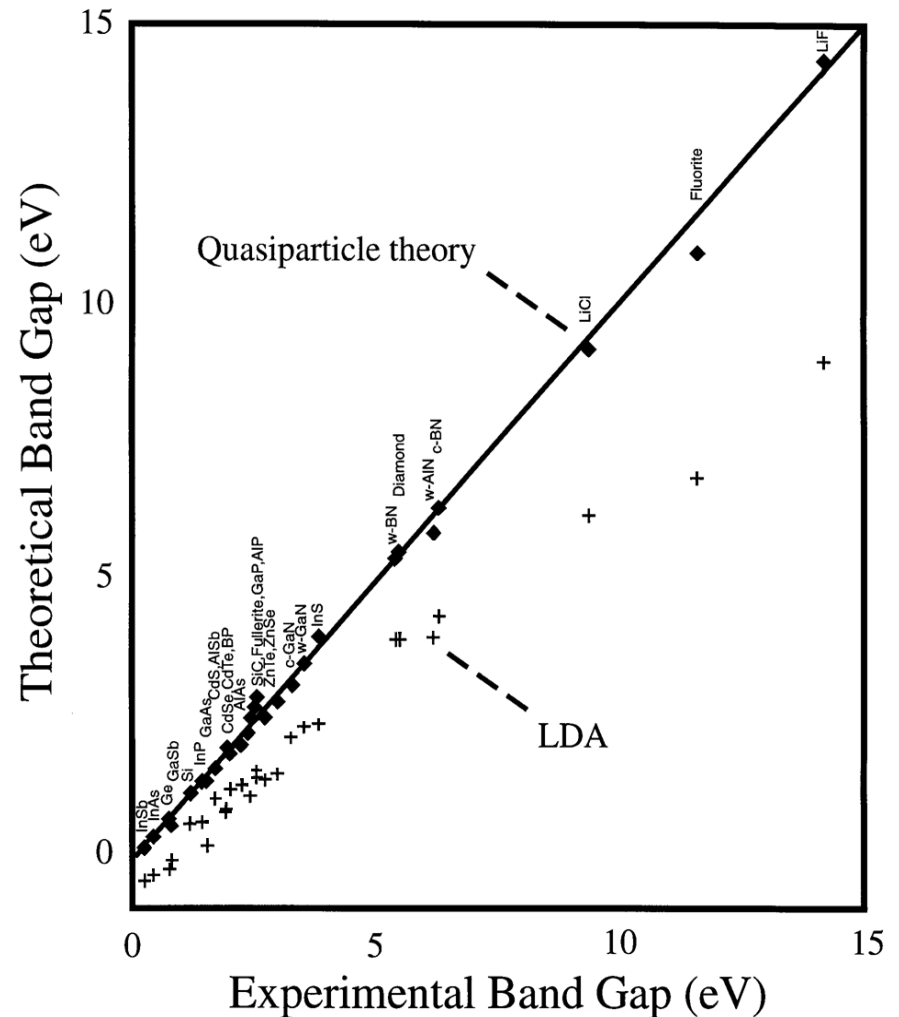
- **NERSC User Services Group (Materials Science / Chemistry Consultant)**
- **NESAP Lead (NERSC's exascale readiness program)**
- **Developer in BerkeleyGW project (SCIDAC Team Member)**
- **My Focus in this presentation will be on BerkeleyGW enhancements over the last couple of years.**



$$[E_{n\mathbf{k}} - H_0(\mathbf{r}) - V_H(\mathbf{r})] \psi_{n\mathbf{k}}(\mathbf{r}) - \int \Sigma(\mathbf{r}, \mathbf{r}', E_{n,\mathbf{k}}) \psi_{n\mathbf{k}}(\mathbf{r}') d\mathbf{r}' = 0$$

Materials:

- InSb, InAs
- Ge
- GaSb
- Si
- InP
- GaAs
- CdS
- AlSb, AlAs
- CdSe, CdTe
- BP
- SiC
- C₆₀
- GaP
- AlP
- ZnTe, ZnSe
- c-GaN, w-GaN
- InS
- w-BN, c-BN
- diamond
- w-AlN
- LiCl
- Fluorite
- LiF



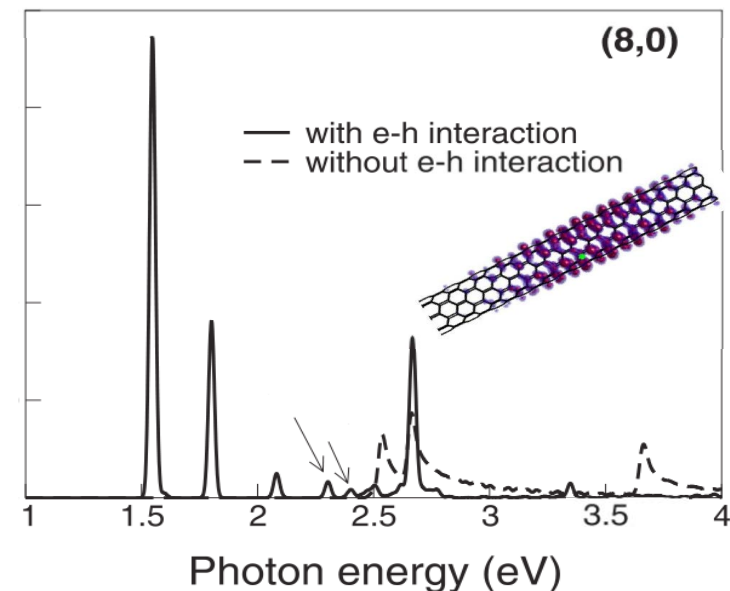
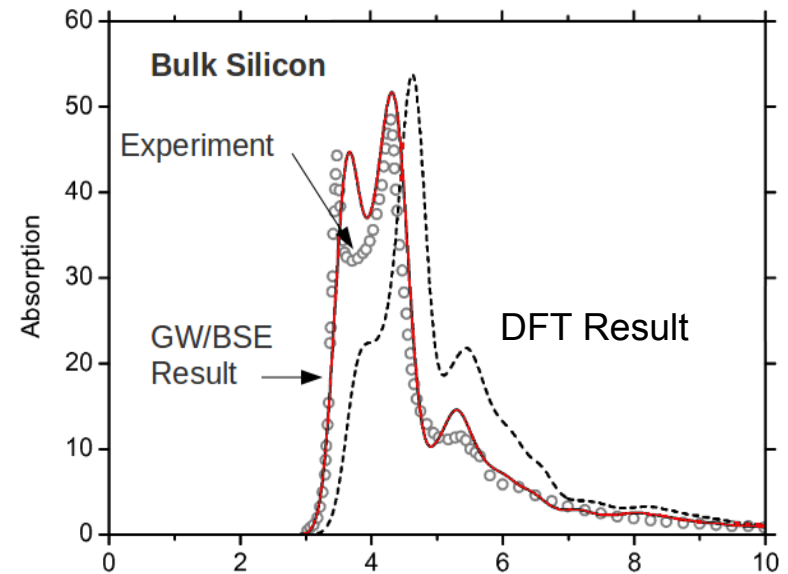
What is GW+BSE

Many-body effects extremely important in **Excited-State properties** of Complex Materials.

Includes screened-interaction for many-body effects

Accurately describes properties important for:

- Photovoltaics
- LEDs
- Junctions / Interfaces
- Defect Energy Levels
-





The Good:

Quantitatively accurate for quasiparticle properties in a wide variety of systems.

Accurately describes dielectric screening important in excited state properties.

The Bad:

Prohibitively slow for large systems. Usually thought to cost orders of magnitude more time than DFT.

Memory intensive and scales badly. Exhausted by storage of the dielectric matrix and wavefunctions. Limited ~50 atoms.

The Good:

Quantitatively accurate for quasiparticle properties in a wide variety of systems.

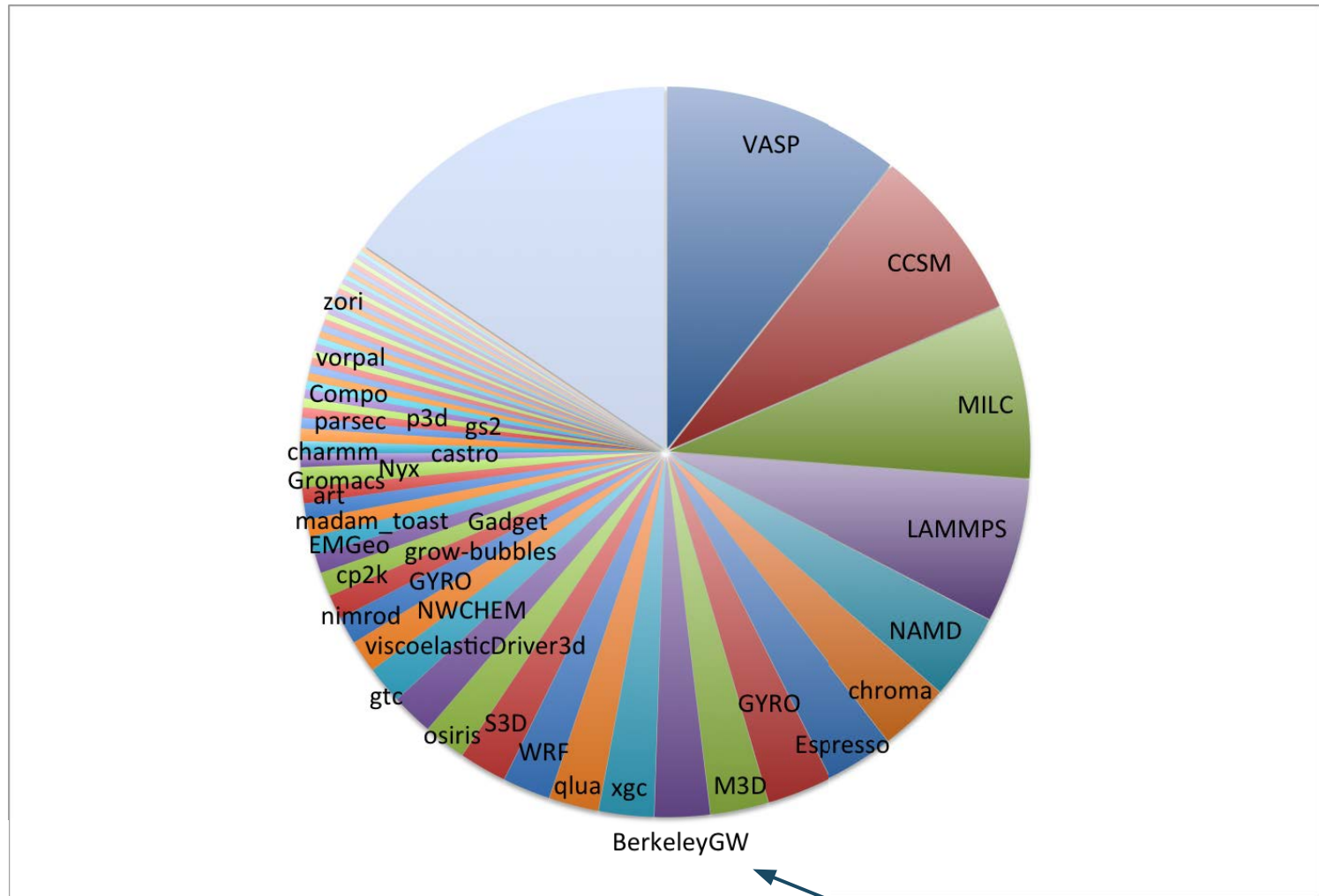
Accurately describes dielectric screening important in excited state properties.

The Bad:

Prohibitively slow for large systems. Usually thought to cost orders of magnitude more time than DFT.

Memory intensive and scales badly. Exhausted by storage of the dielectric matrix and wavefunctions. Limited ~50 atoms.

BerkeleyGW Usage at NERSC



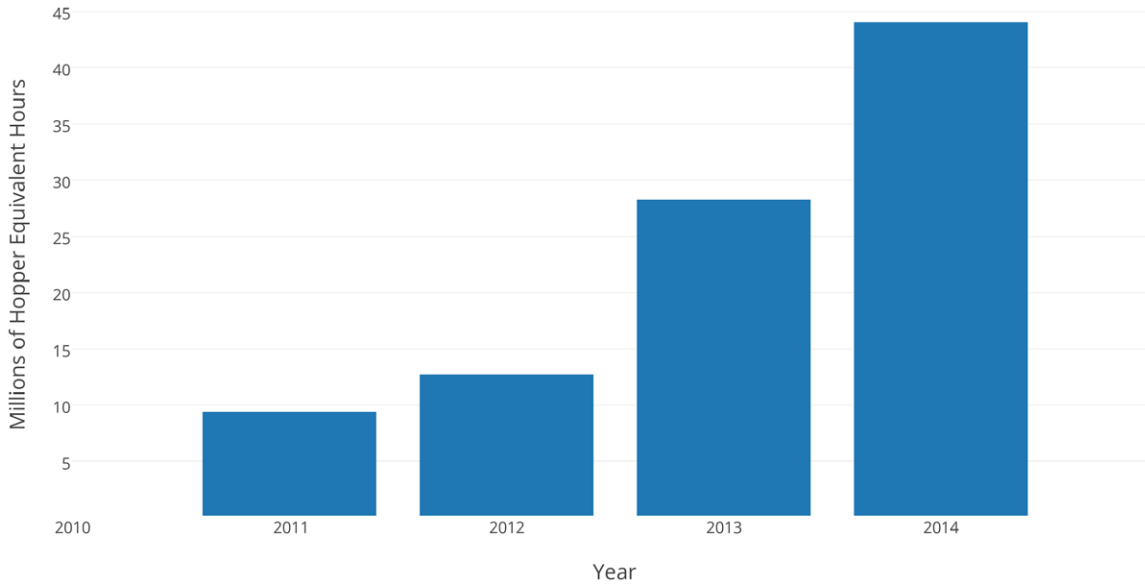
2.4%

NERSC Code Breakdown 2013

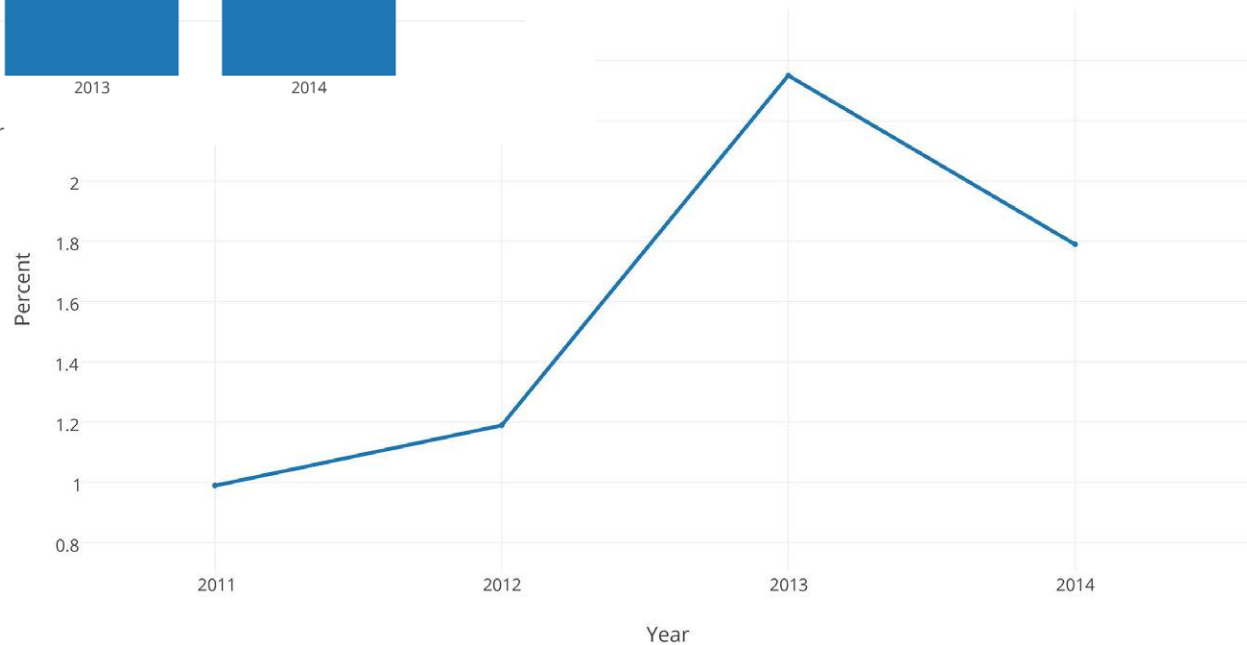
BerkeleyGW Usage at NERSC

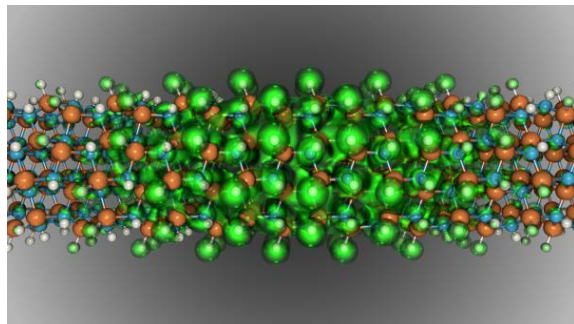


BerkeleyGW Millions of Hours used at NERSC



Percentage of NERSC Cycles





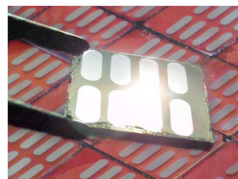
PI: Louie

First Ab Initio Method for Characterizing Hot Carriers Could Hold the Key to Future Solar Cell Efficiencies

Science Shorts Lynn Yarris • JULY 17, 2014

Tweet 112 Like 58 Facebook Like 78 Share 0 Email 7

One of the major road blocks to the design and development of new, more efficient solar cells may have been cleared. Researchers with the Lawrence Berkeley National Laboratory (Berkeley Lab) have developed the first *ab initio* method – meaning a theoretical model free of adjustable or empirical parameters – for characterizing the properties of “hot carriers” in semiconductors. Hot carriers are electrical charge carriers – electrons and holes – with significantly higher energy than charge carriers at thermal equilibrium.



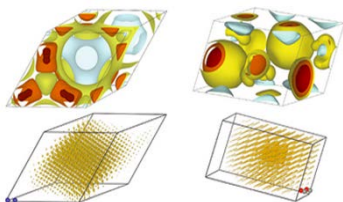
A new and better way to study “hot” carriers in semiconductors, a major source of efficiency loss in solar cells, has been developed by scientists at Berkeley Lab. (Photo by Roy Kaltschmidt)

PI: Kioupakis

TO BRIDGE LED'S GREEN GAP, SCIENTISTS THINK SMALL

PI: Wu

Cover Image: Phys. Rev. Lett. Vol. 110, Iss. 1



From the article:

Origin of the Variation of Exciton Binding Energy in Semiconductors

Marc Dvorak, Su-Huai Wei, and Zhigang Wu

Phys. Rev. Lett. **110**, 016402 (2013)

PI: Van De Walle / Cohen

IMPORTANT NEW METHOD FOR STUDYING SOLAR MATERIALS

The goal of research in this group in general has been to develop and use first-principles computational methods to understand, predict, and design novel electronic, optoelectronic, photovoltaic, and thermoelectric materials.

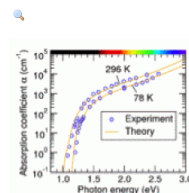
Using a NERSC NISE award, researchers were able to compute the phonon-assisted interband optical absorption spectrum of silicon entirely from first principles.

Nearly all commercially-available photovoltaic cells currently depend on this absorption process.

The new method is general enough to study fundamental physics of other optoelectronic and photovoltaic materials and can address questions that are not accessible by experiment.

Used the BerkeleyGW software written by new NERSC USG consultant Jack Deslippe.

Work done by Jesse Noffsinger, Emmanouil Kioupakis, Chris G. Van de Walle, Steven G. Louie, and Marvin L. Cohen.



Calculated (solid lines) and experimental (circles) absorption coefficient of silicon in the energy range between the indirect and direct gaps, for two temperatures.

PI: Prendergast

LASER, SUPERCOMPUTER MEASURE SPEEDY ELECTRONS IN SILICON

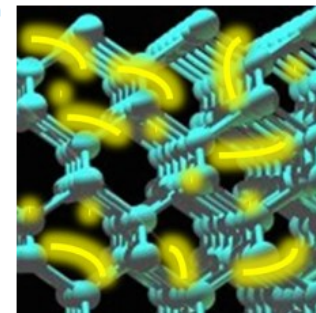
Simulations at NERSC Help Illuminate Attosecond Laser Experiment Findings

DECEMBER 19, 2014 | Tags: [Basic Energy Sciences \(BES\)](#), [Materials Science](#)
Contact: Robert Sanders, rsanders@berkeley.edu, (510) 643-6998

The entire semiconductor industry, not to mention Silicon Valley, is built on the propensity of electrons in silicon to get kicked out of their atomic shells and start to move through the material. These mobile electrons are routed and switched through transistors, carrying the digital information that characterizes our age.

An international team of physicists and chemists at the University of California, Berkeley, has for the first time taken snapshots of this ephemeral event using attosecond pulses of soft x-ray light lasting only a few billionths of a billionth of a second. The researchers then used supercomputing resources at Lawrence Berkeley National Laboratory's (Berkeley Lab) National Energy Research Scientific Computing Center (NERSC) to help them better understand their findings.

While earlier femtosecond lasers were unable to resolve the jump from the valence shell of the



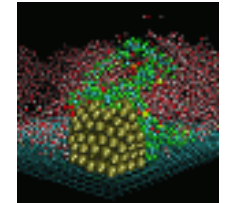
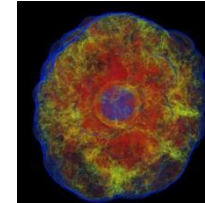
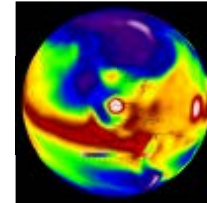
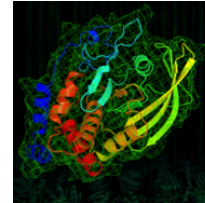
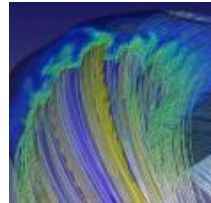
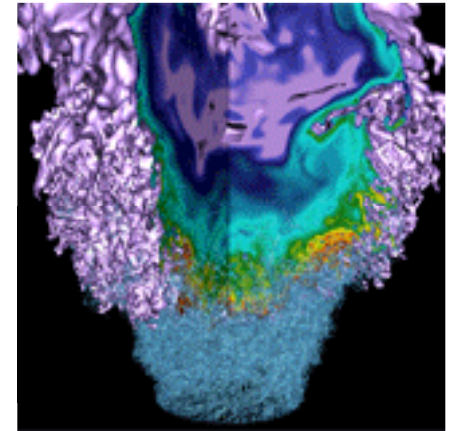
In silicon, electrons attached to atoms in the crystal lattice can be mobilized into the conduction band by light or voltage. Berkeley scientists have taken snapshots of this very brief band-gap jump and timed it at 450 attoseconds. Image: Stephen Leone



BerkeleyGW Workshops

- Emphasized integration with **PARSEC** code
- 3 times the number of applicants than space available! 45 Attendees.
- Survey results show that 100% of attendees found sessions useful of very useful.

BerkeleyGW in the Many-Core Era



- **Cori will begin to transition the workload to more energy efficient architectures**
- **Cray XC system with over 9300 Intel Knights Landing (Xeon-Phi) nodes**
 - Self-hosted, (not an accelerator) manycore processor with over 60 cores per node + high-bandwidth memory
- **Data Intensive Science Support**
 - NVRAM Burst Buffer to accelerate applications



System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.

What is different about Cori?



Edison (Ivy-Bridge):

- 12 Cores Per CPU
- 24 Virtual Cores Per CPU
- 2.4-3.2 GHz
- Can do 4 Double Precision Operations per Cycle
- 2.5 GB of Memory Per Core
- ~100 GB/s Memory Bandwidth

Cori (Knights-Landing):

- 60+ Physical Cores Per CPU
- 240+ Virtual Cores Per CPU
- Much slower GHz
- Can do 8 Double Precision Operations per Cycle
- < 0.3 GB of Fast Memory Per Core
< 2 GB of Slow Memory Per Core
- Fast memory has ~ 5x DDR4 bandwidth

Optimization Strategy For Cori



Both **PARSEC** and **BerkeleyGW** are included in top tier of the NERSC Exascale Application Program (NESAP).

- Work with Cray, NERSC, Intel and SUPER staff
- Early access to simulators and hardware.

Strategy:

- A. Add OpenMP or other on-node parallelism
- B. Effectively use vector Instructions
- C. Identify/optimize memory bandwidth hotspots.

Optimization Strategy For Cori



Both **PARSEC** and **BerkeleyGW** are included in top tier of the NERSC Exascale Application Program (NESAP).

- Work with Cray, NERSC, Intel and SUPER staff
- Early access to simulators and hardware.

Strategy:

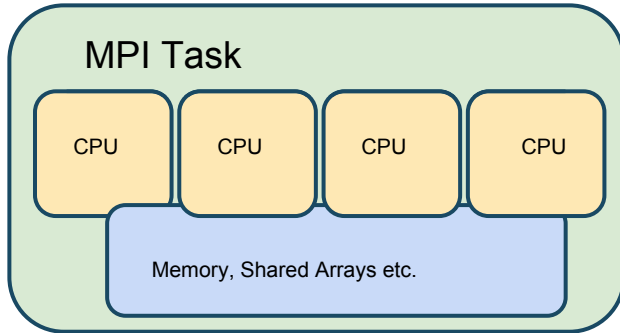
- Add OpenMP or other on-node parallelism
- Effectively use vector Instructions
- Identify/optimize memory bandwidth hotspots.

```
do i = 1, n
  a(i) = b(i) + c(i)
enddo
```



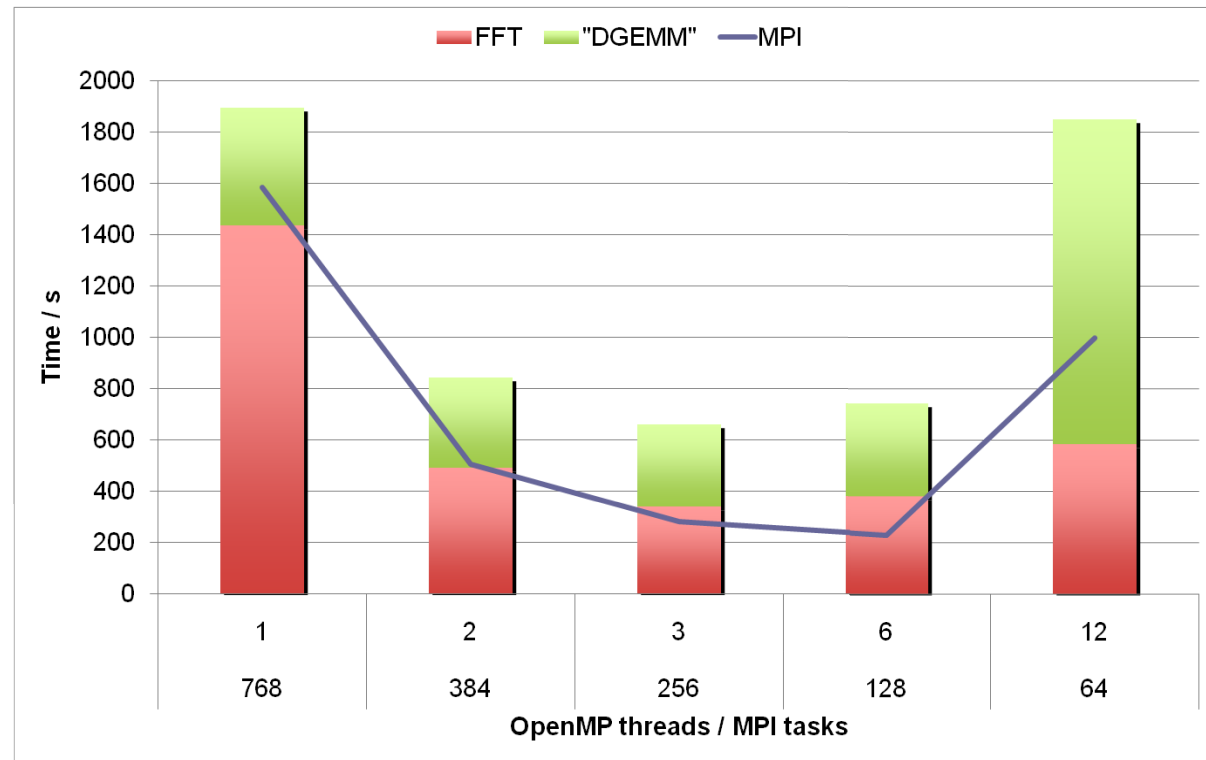
$$\begin{pmatrix} a_1 \\ \dots \\ a_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \dots \\ b_n \end{pmatrix} + \begin{pmatrix} c_1 \\ \dots \\ c_n \end{pmatrix}$$

Example Use Case For OpenMP in BerkeleyGW and PARSEC



Parallel FFTs involve MPI all-to-all communication (small messages, latency bound).

Reducing the number of MPI tasks in favor of OpenMP threads makes large improvement in overall runtime.



Work by Andrew Canning

Significant Bottleneck is large matrix reduction like operations. Turning arrays into numbers.

$$\langle n\mathbf{k} | \Sigma_{\text{CH}}(E) | n'\mathbf{k} \rangle = \frac{1}{2} \sum_{n''} \sum_{\mathbf{q}\mathbf{G}\mathbf{G}'} M_{n''n}^*(\mathbf{k}, -\mathbf{q}, -\mathbf{G}) M_{n''n'}(\mathbf{k}, -\mathbf{q}, -\mathbf{G}') \\ \times \frac{\Omega_{\mathbf{G}\mathbf{G}'}^2(\mathbf{q}) (1 - i \tan \phi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}))}{\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}) (E - E_{n''\mathbf{k}-\mathbf{q}} - \tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}))} v(\mathbf{q} + \mathbf{G}')$$

Targeting Intel Xeon Phi Many Core Architecture

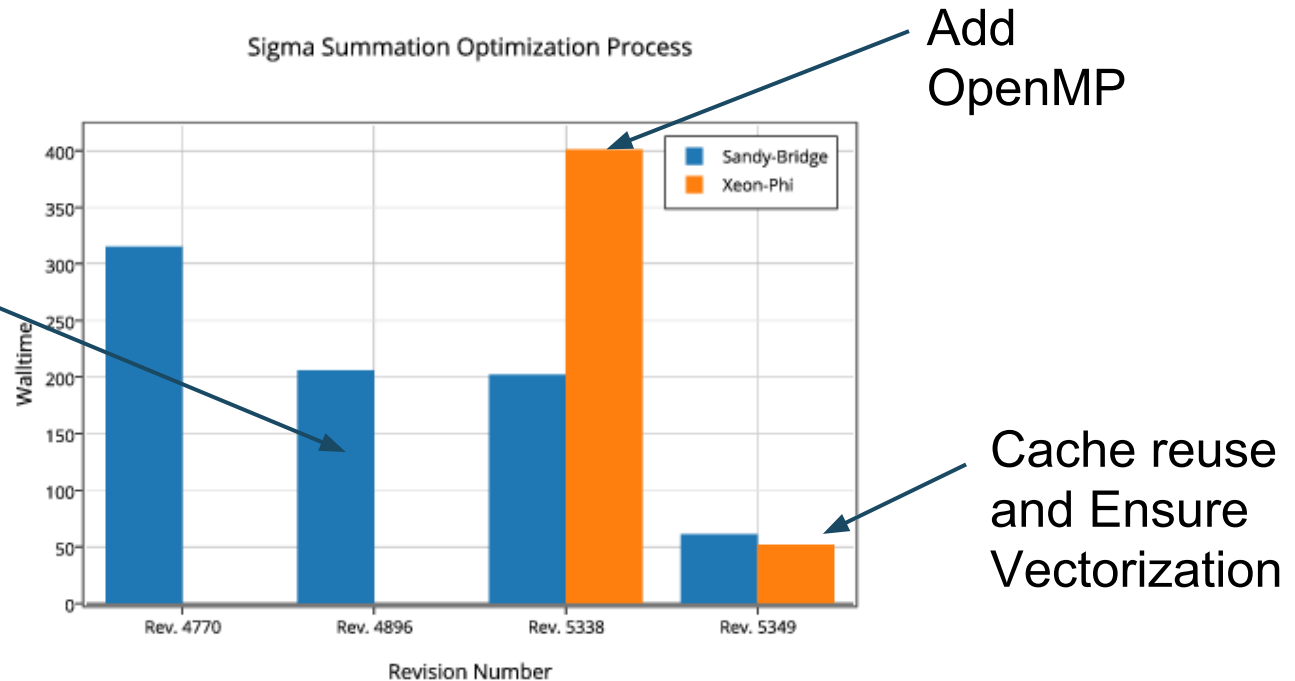


1. Target more on-node parallelism. (MPI model already failing users)
2. Ensure key loops/kernels can be vectorized.

Example: Optimization steps for Xeon Phi Coprocessor

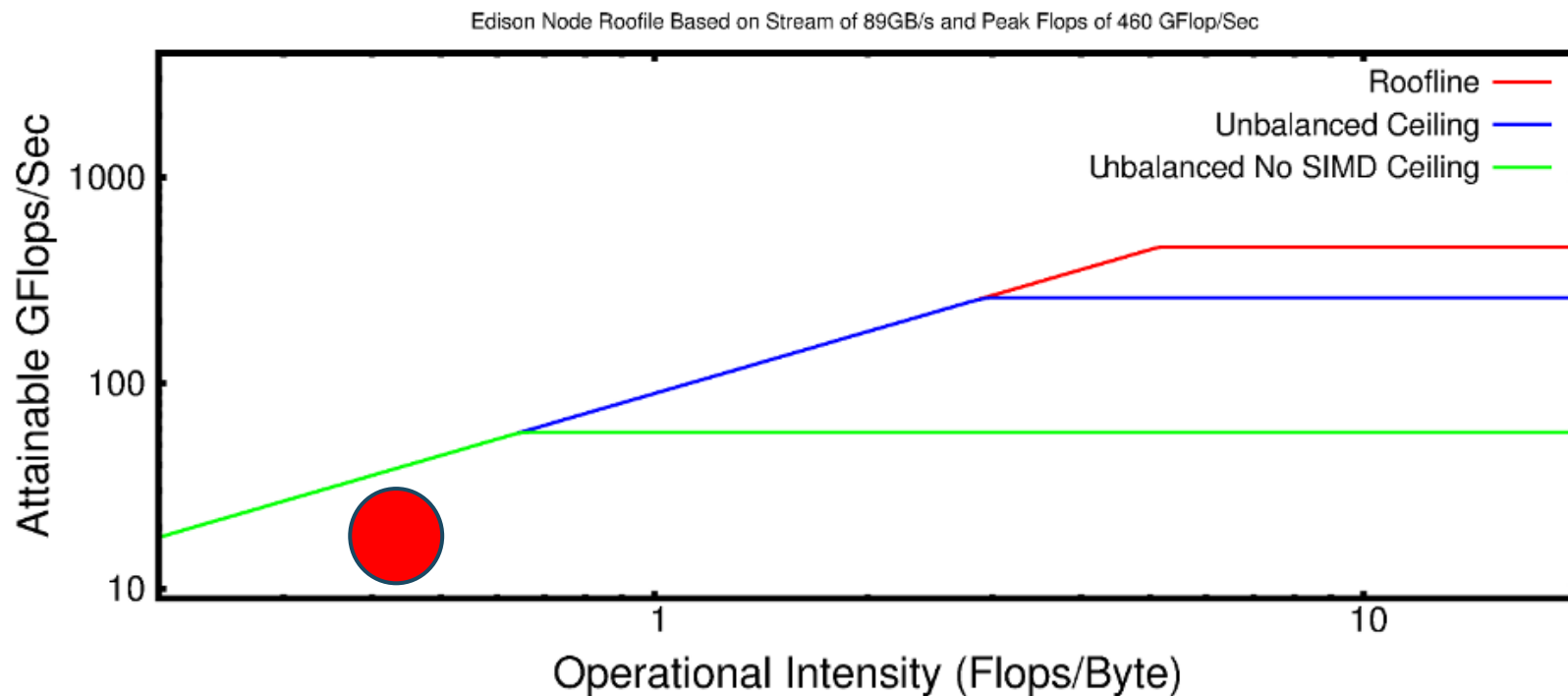
Refactor to Have 3
Loop Structure:

Outer: MPI
Middle: OpenMP
Inner: Vectorization



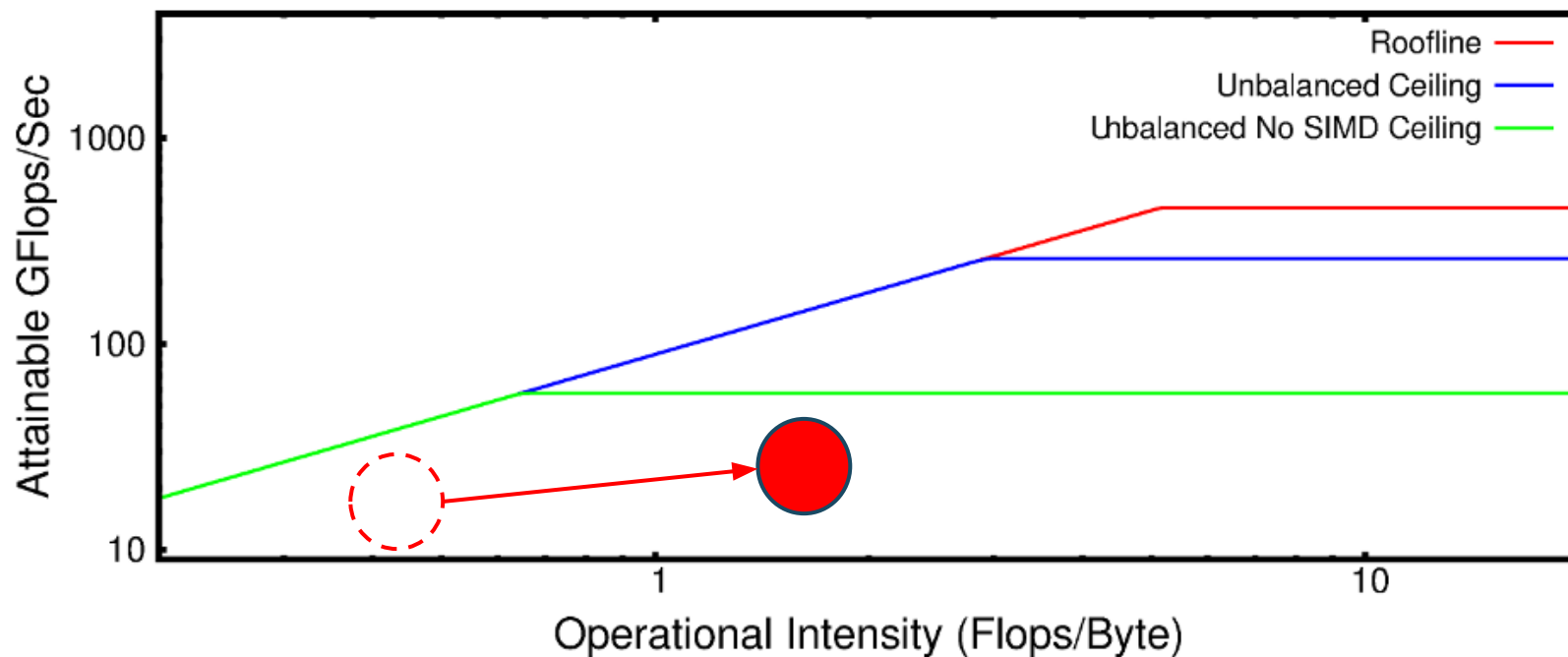
Added additional layers of cache-blocking to improve locality. Important on Xeon-Phi which lacks L3.

2013 - Low Operational Intensity



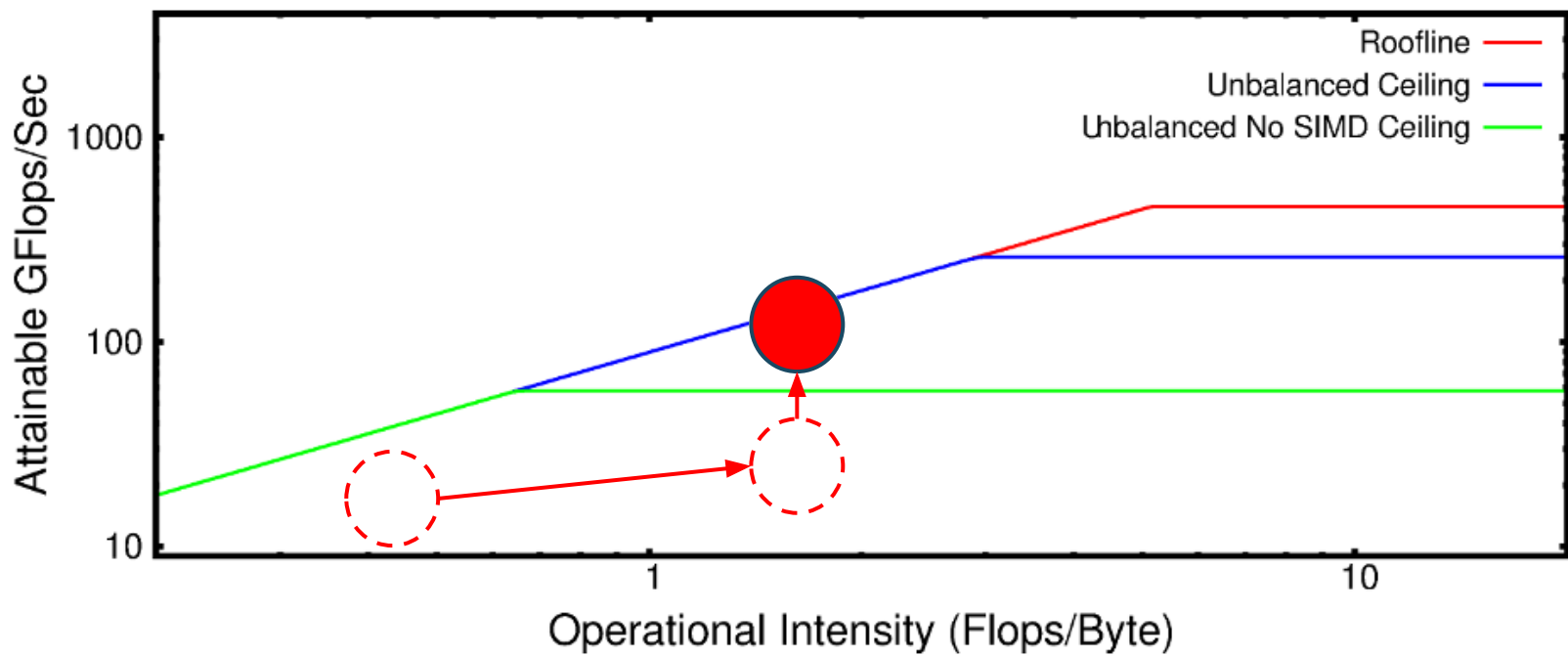
2014 - Refactored loops, improved locality

Edison Node Roofline Based on Stream of 89GB/s and Peak Flops of 460 GFlop/Sec

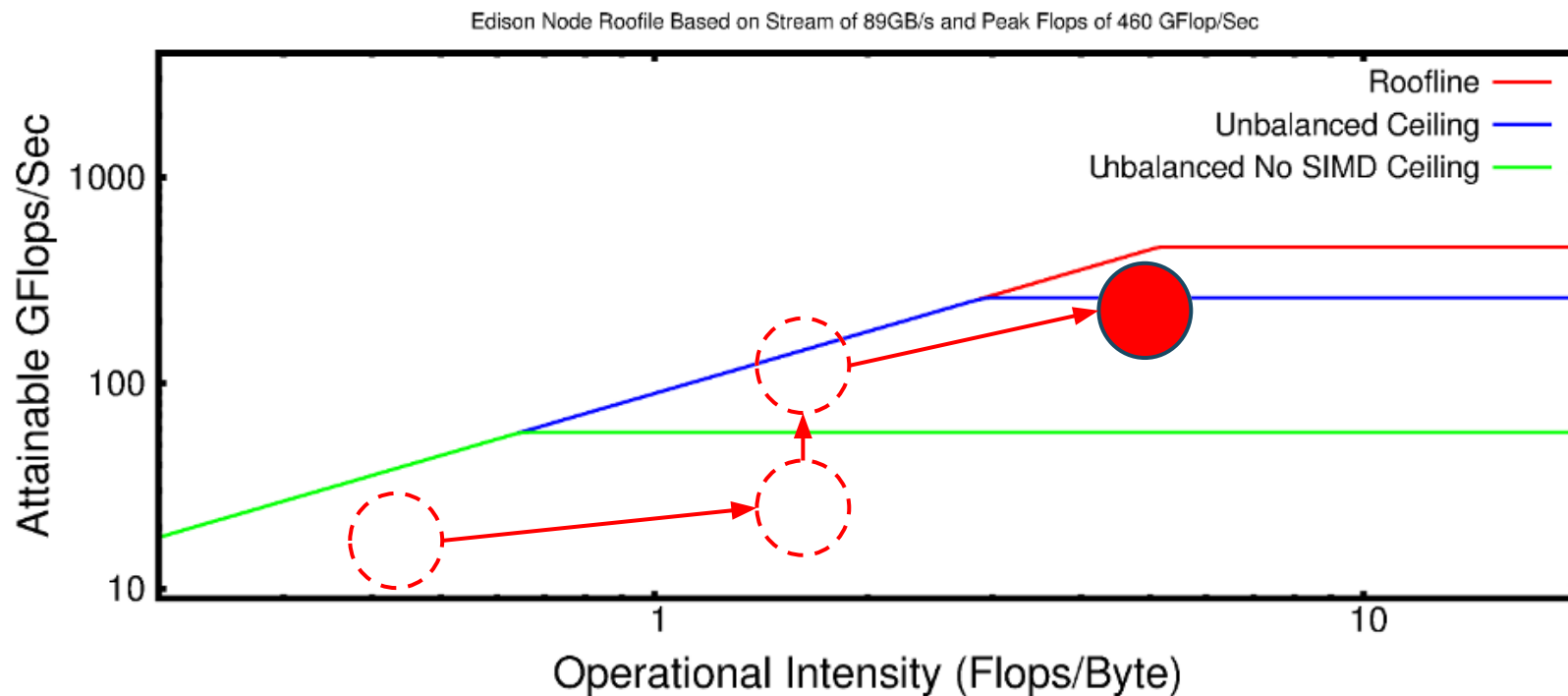


2014 - Vectorized Code

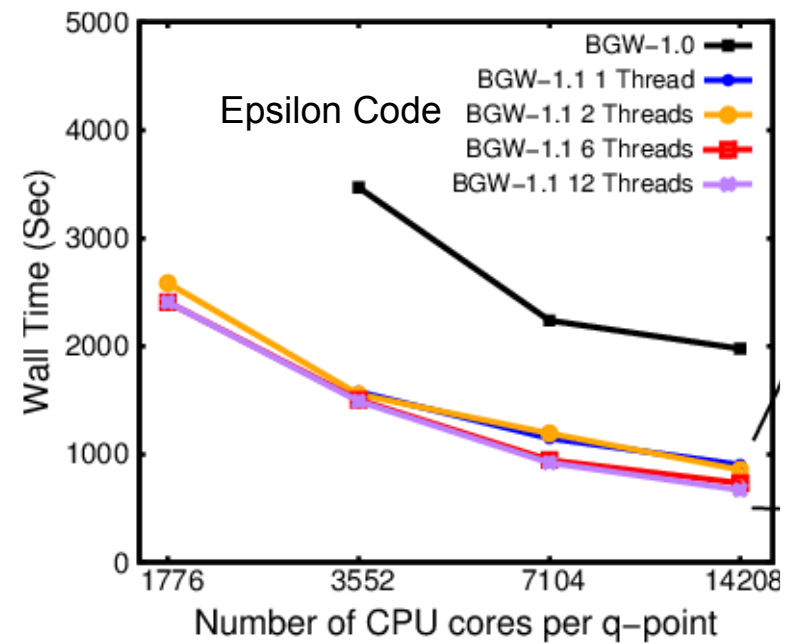
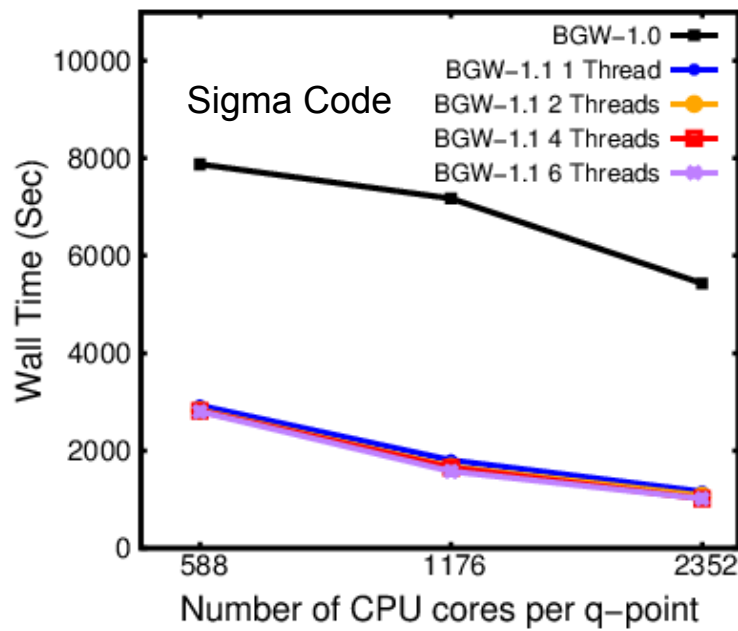
Edison Node Roofline Based on Stream of 89GB/s and Peak Flops of 460 GFlop/Sec



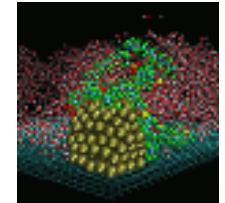
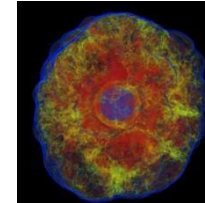
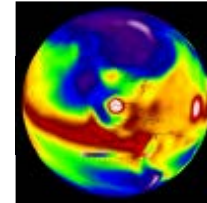
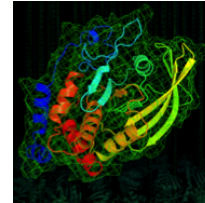
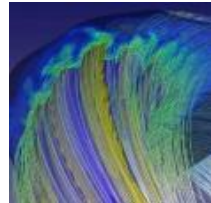
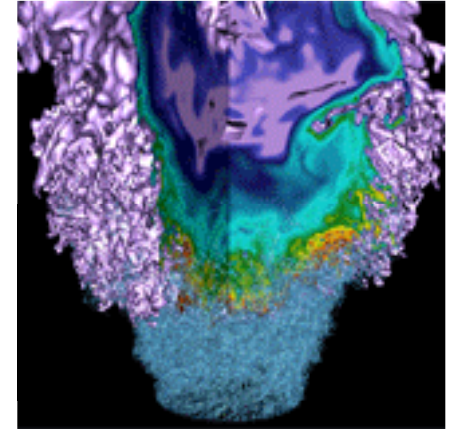
2015 - Cache Blocking



Hybrid MPI-OpenMP Scaling Improvements.



Science/Method Advances



New Features



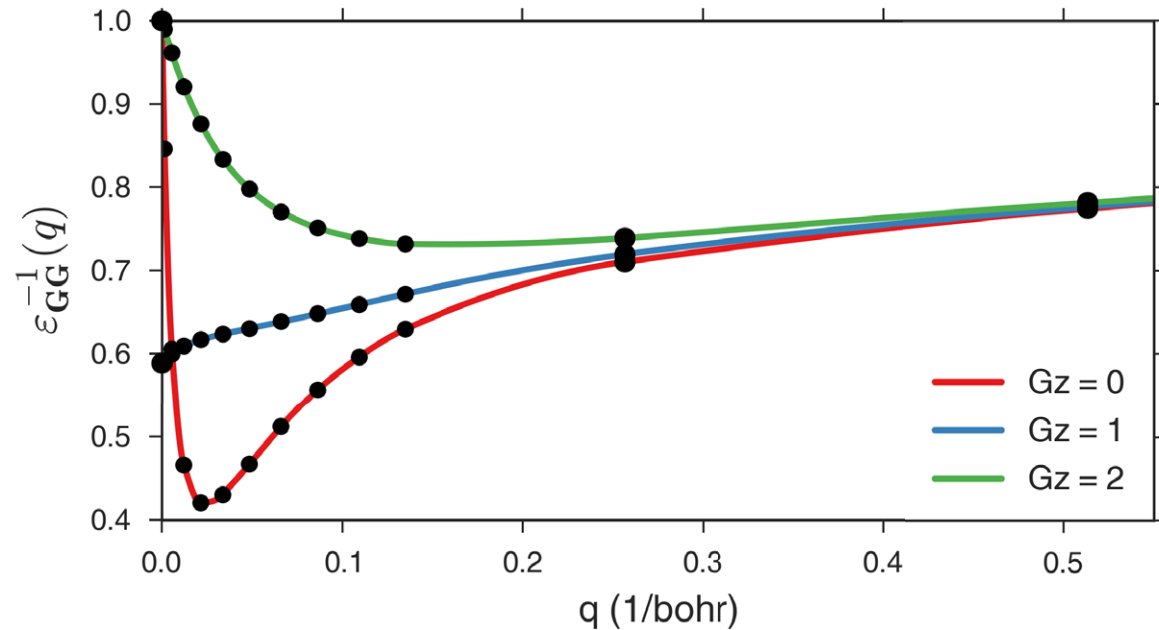
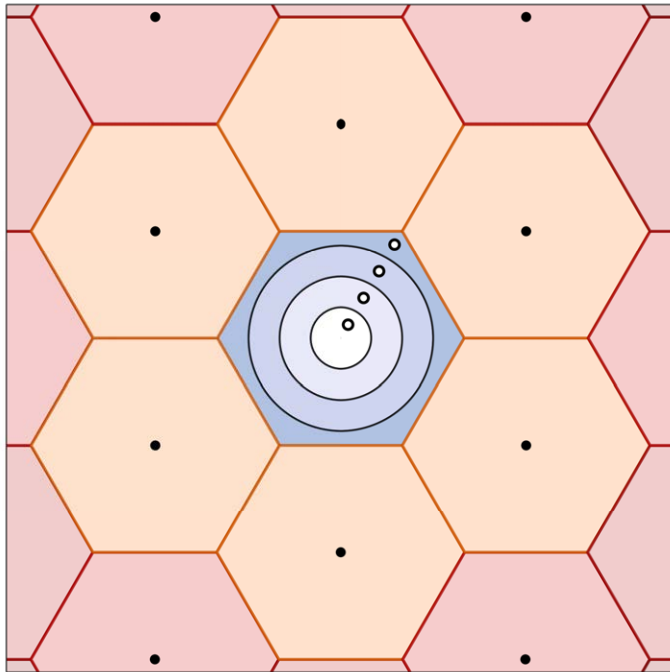
- Exciton Bandstructures (Finite Center of Mass Momentum Excitons)
- Full-Frequency Calculations With $\sim 2X$ GPP Cost (Hilbert Transform and Contour Deformation Approach)
- Parallelization Over Frequencies
- Parallel IO and Transferable File Format
- Parallel reduced size FFTs
- **Vastly Improved K-Point Convergence**
- Support for PARSEC Input, Abinit Input, RMG Input
- Support for Static COHSEX Starting Point
- Empty State Requirement Reduction
- **Full BSE Calculations in Parallel**
- **Accurate/efficient GPP models for Informatics**

- Over 2500 Commits since BerkeleyGW 1.0. Many performance improvements, bug fixes and new features

Improving K-Point Integration



How to efficiently capture long wavelength features in the dielectric matrix?



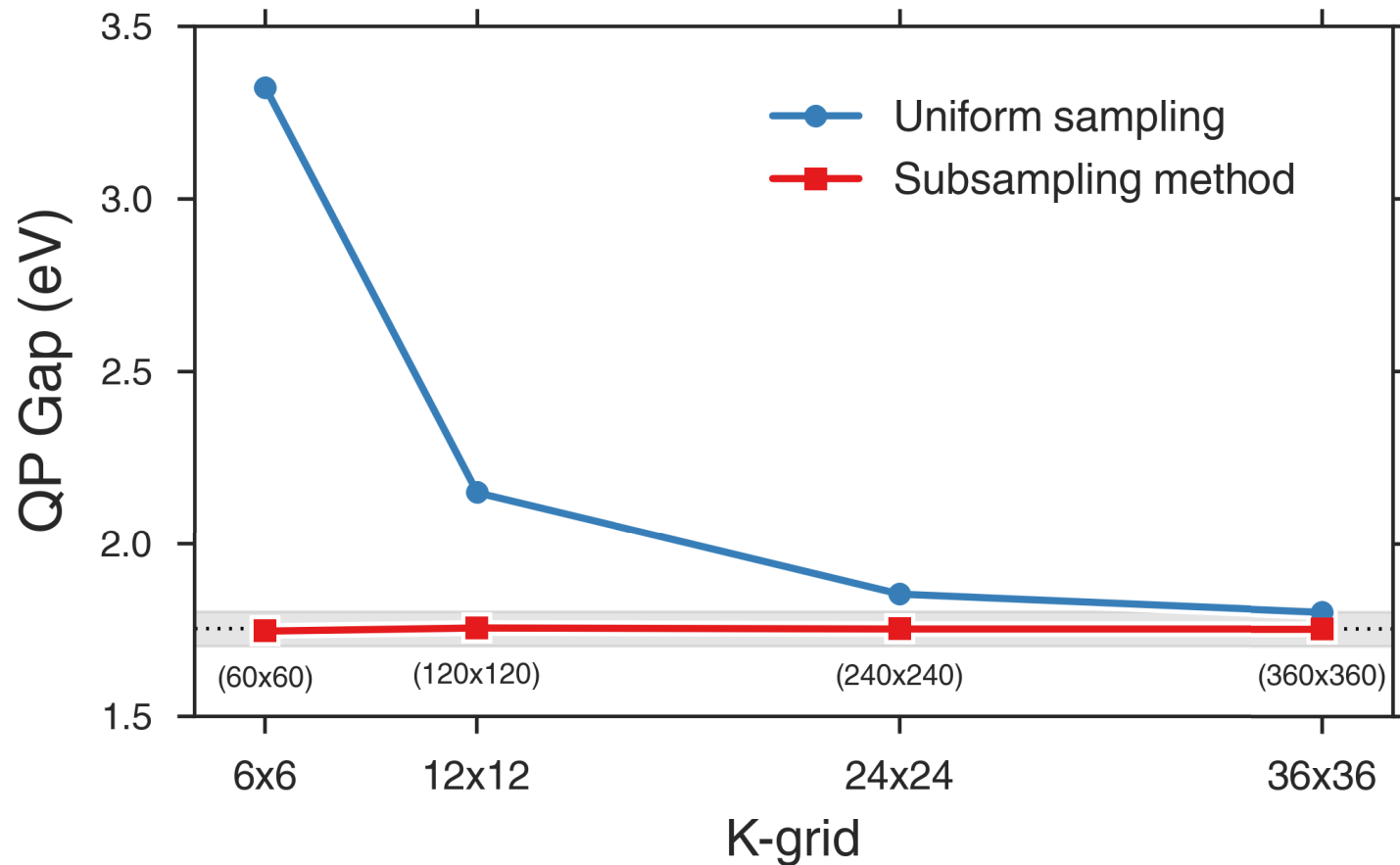
Our solution: subsampling method - a hybrid sampling of the Brillouin Zone.

Improving K-Point Integration

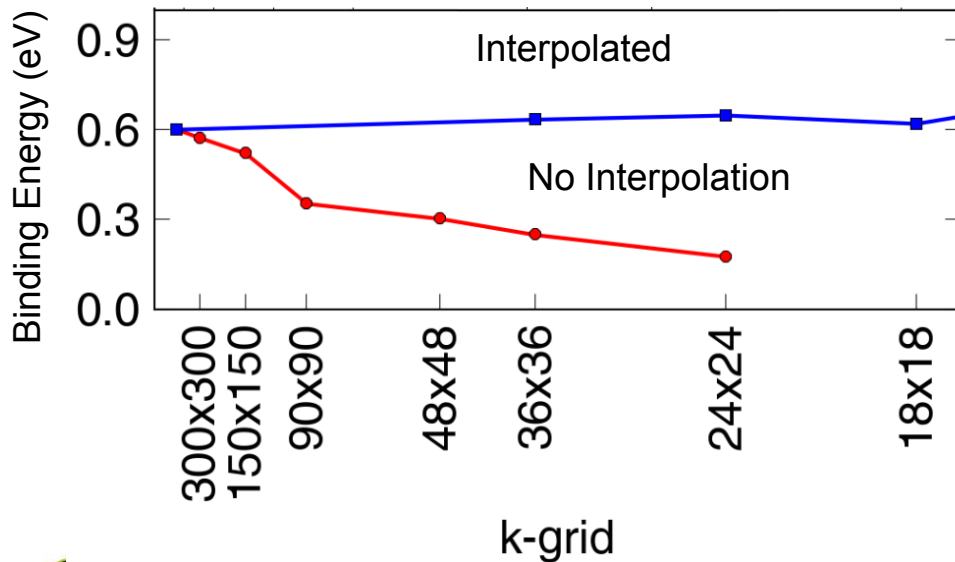
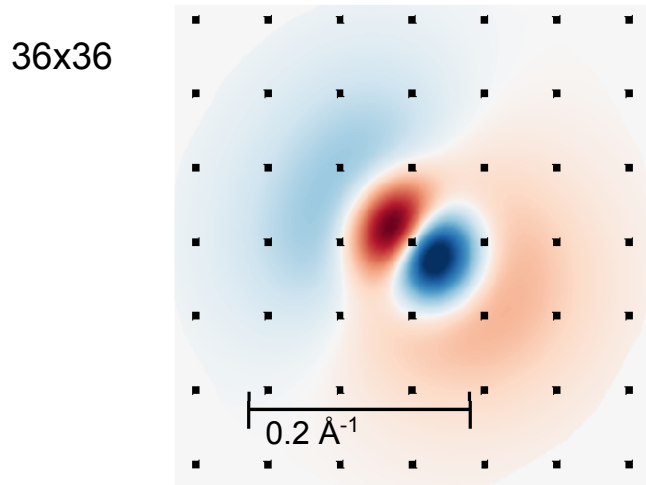


Subsampling

MoS₂

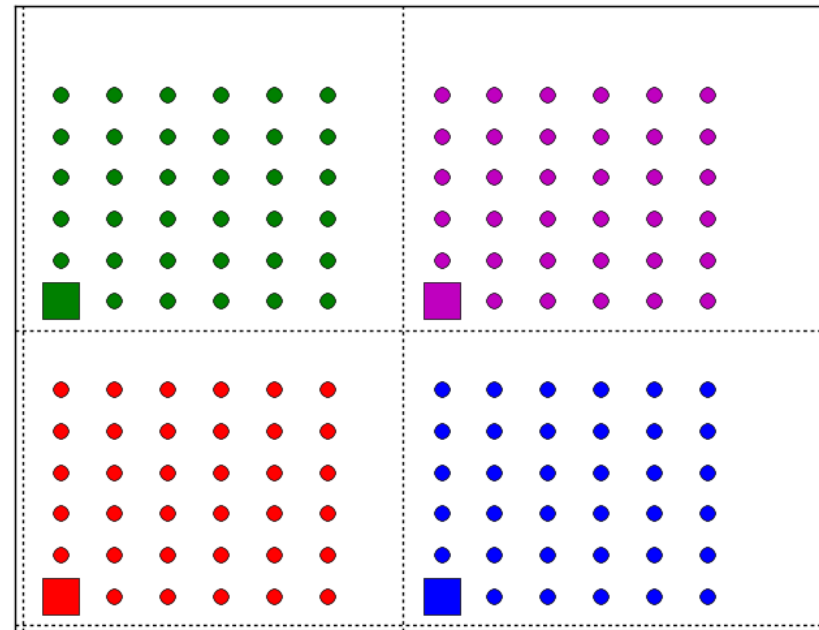


Improving K-Point Integration



Improved cluster interpolation method for BSE

■ Coarse point ● Cluster point



Efficient Full BSE Calculations



The Full BSE Hamiltonian is complex, non Hermitian

$$H_{\text{BSE}} = \begin{bmatrix} R & C \\ -C^* & -R^* \end{bmatrix}$$

We need a parallel diagonalization routine for matrix of this form.

Want cost to be similar to the
Tamm-Dancoff approximation:

$$H_{\text{BSE}}^{\text{TDA}} = [R]$$

Efficient Full BSE Calculations



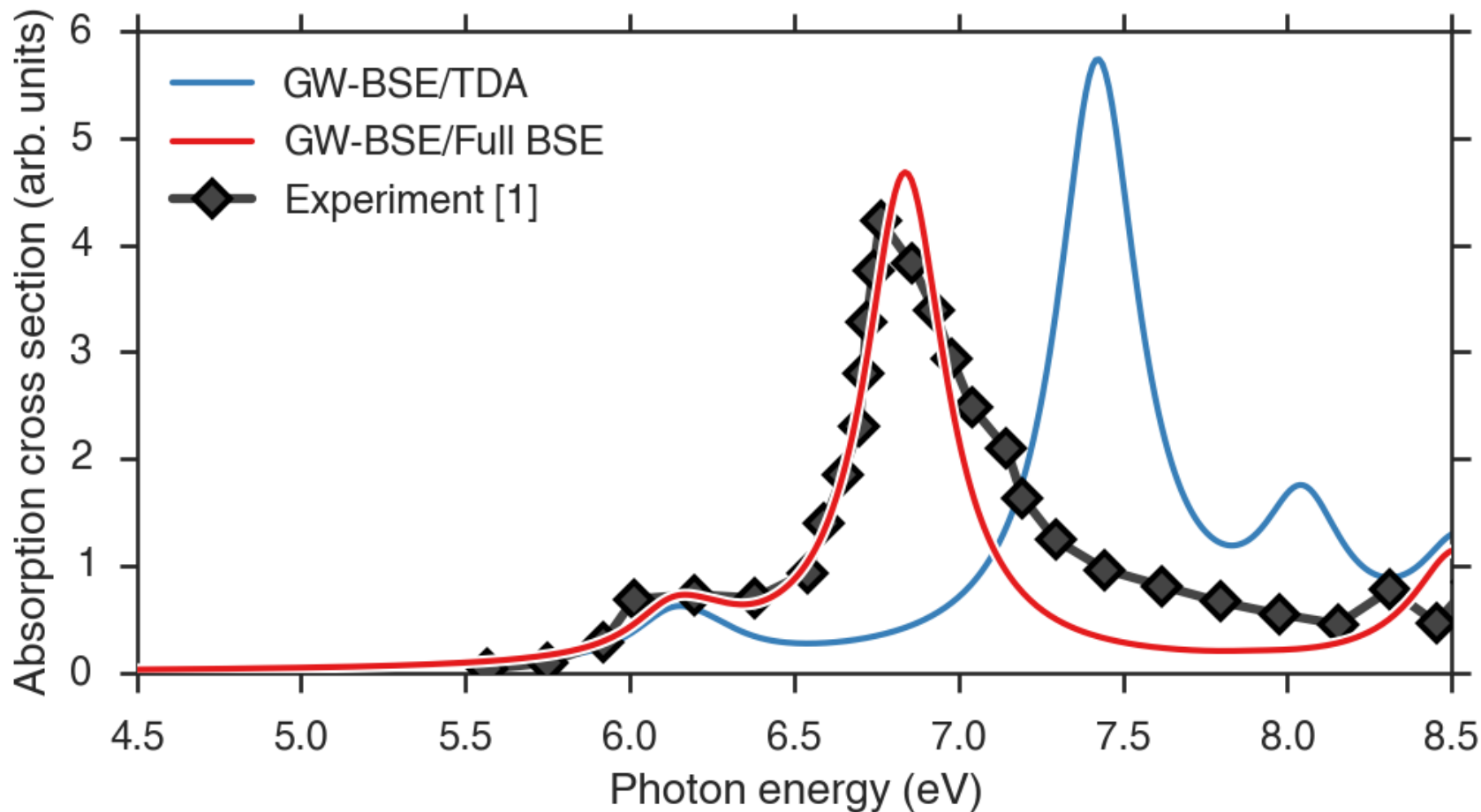
- Challenge: diagonalize non-Hermitian matrix (2x the rank):
 - Efficiently
 - In parallel
 - Preserving structure of the solutions
- Solution: **new solver** written by Meiyue Shao and Chao Yang.
- Si (real matrix), nmat=24 000, #PEs=128

	TDA	Our solver	Generic solver
Time (s)	78.163	243.481 (3.1x)	1198.535 (15.3 x)

- Naphthalene (complex matrix), nmat=8 000, #PEs=72

	TDA	Our solver	Generic solver
Time (s)	41.088	259.309 (6.3x)	593.593 (14.4x)

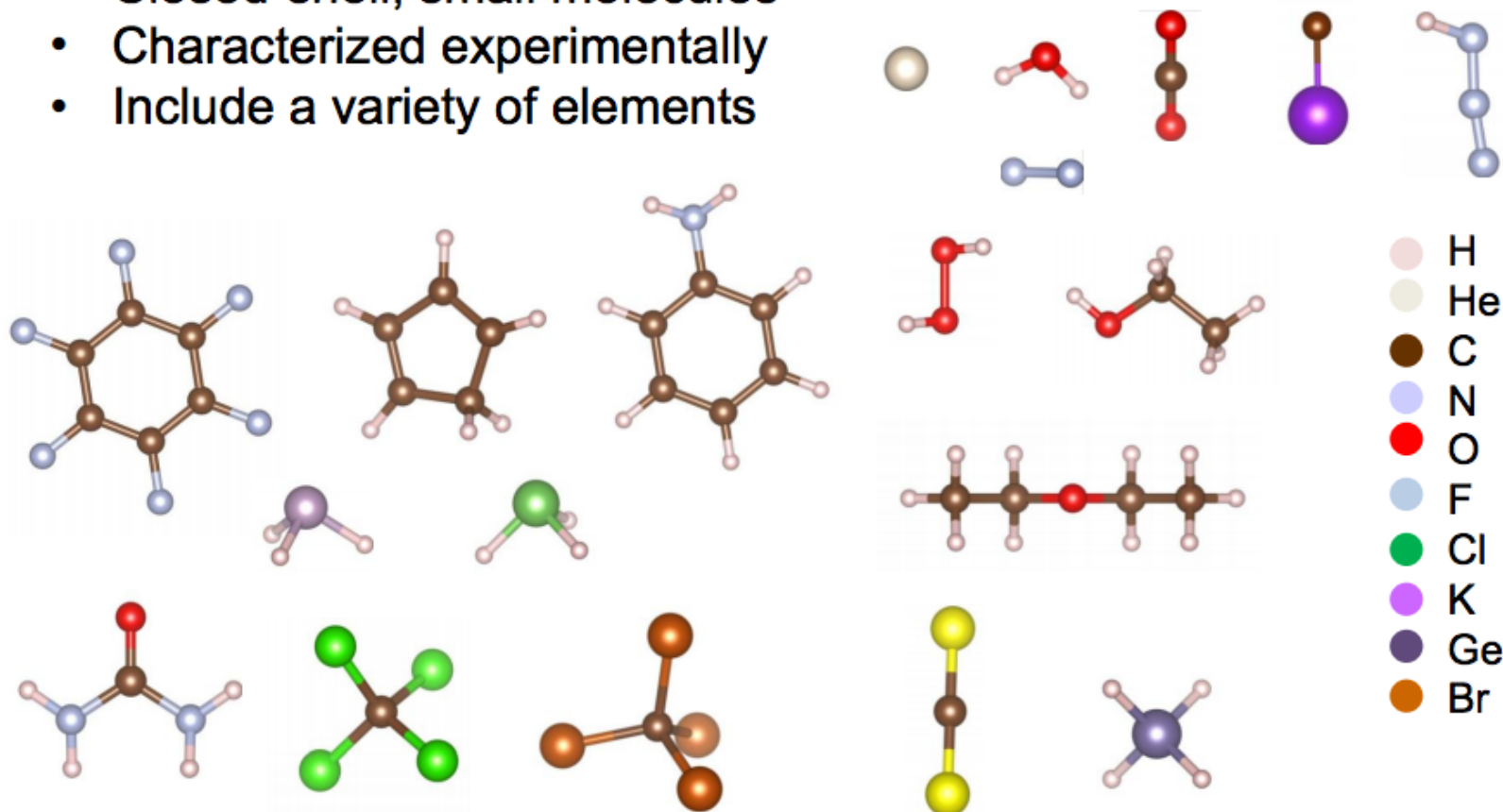
Example: Benzene



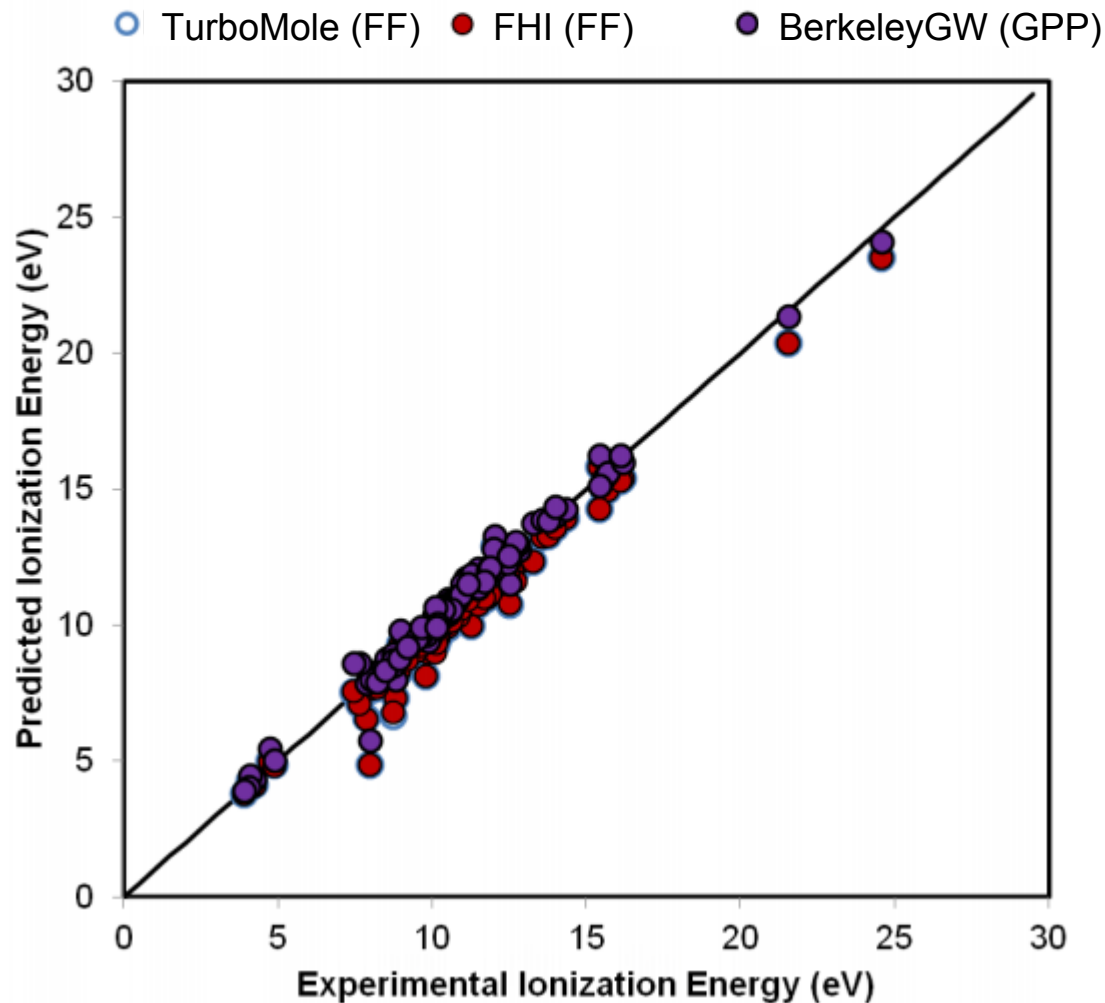
GW - Towards Informatics



- Closed-shell, small molecules
- Characterized experimentally
- Include a variety of elements



GW - Towards Informatics



GW - Towards Informatics



Full-Frequency Calculations within G0W0
Systematically underestimate band gaps.

GPP gives generally very good energies for
informatics based approaches.

Poles in FF G0W0 ~

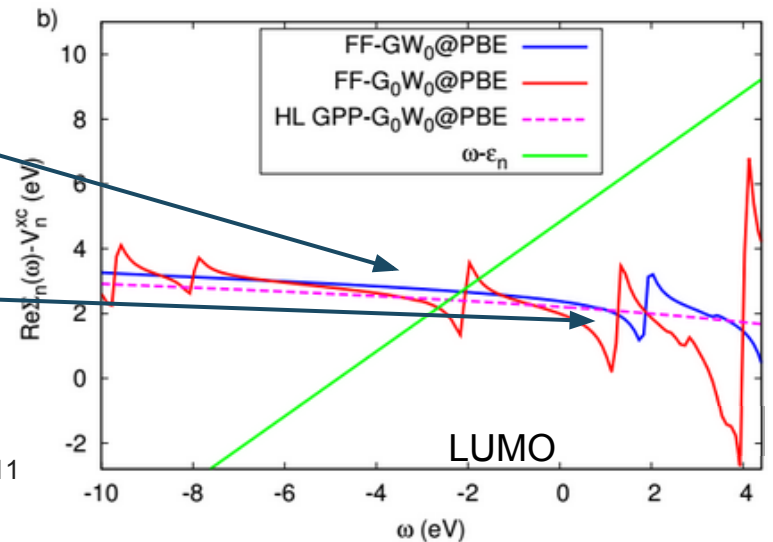
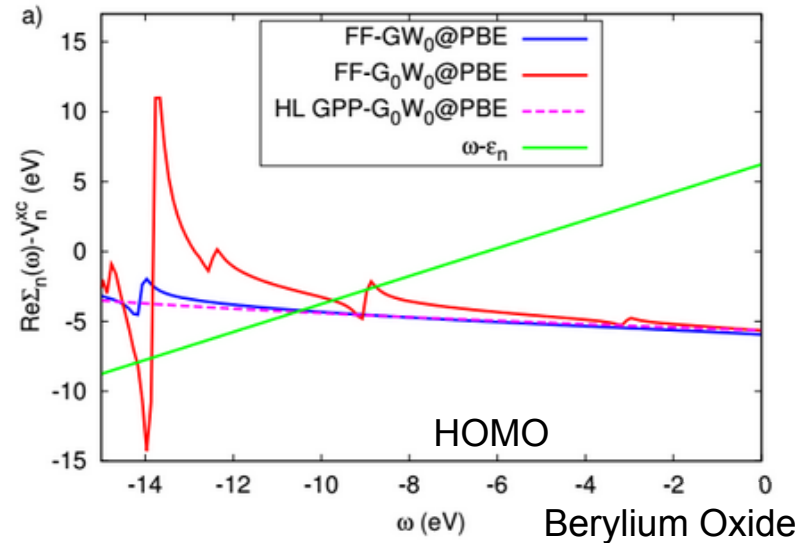
$$E_{v_lda} - E_{gap}$$

$$E_{c_lda} + E_{gap}$$

Poles in FF GW0 ~

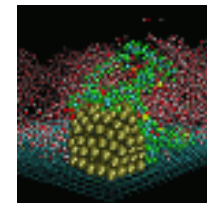
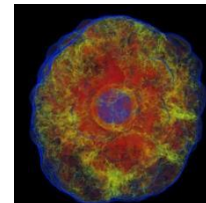
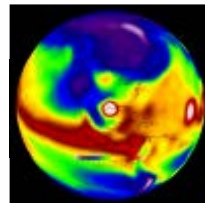
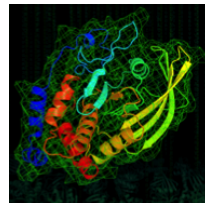
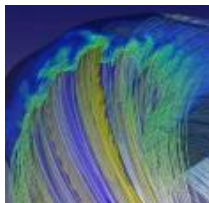
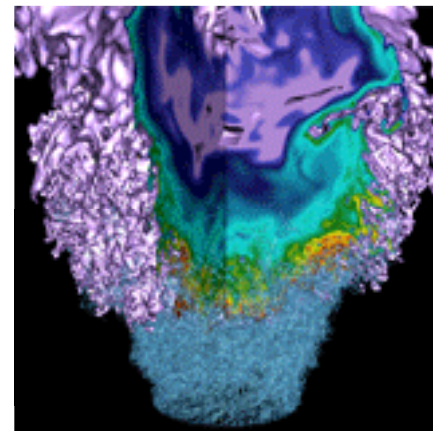
$$E_{v_QP} - E_{gap}$$

$$E_{c_QP} - E_{gap}$$



Lischner, Johannes ... Jack Deslippe, J.B.
Neaton, S.G. Louie. *Physical Review B* 90.11
(2014): 115130.

Conclusions



Conclusions:

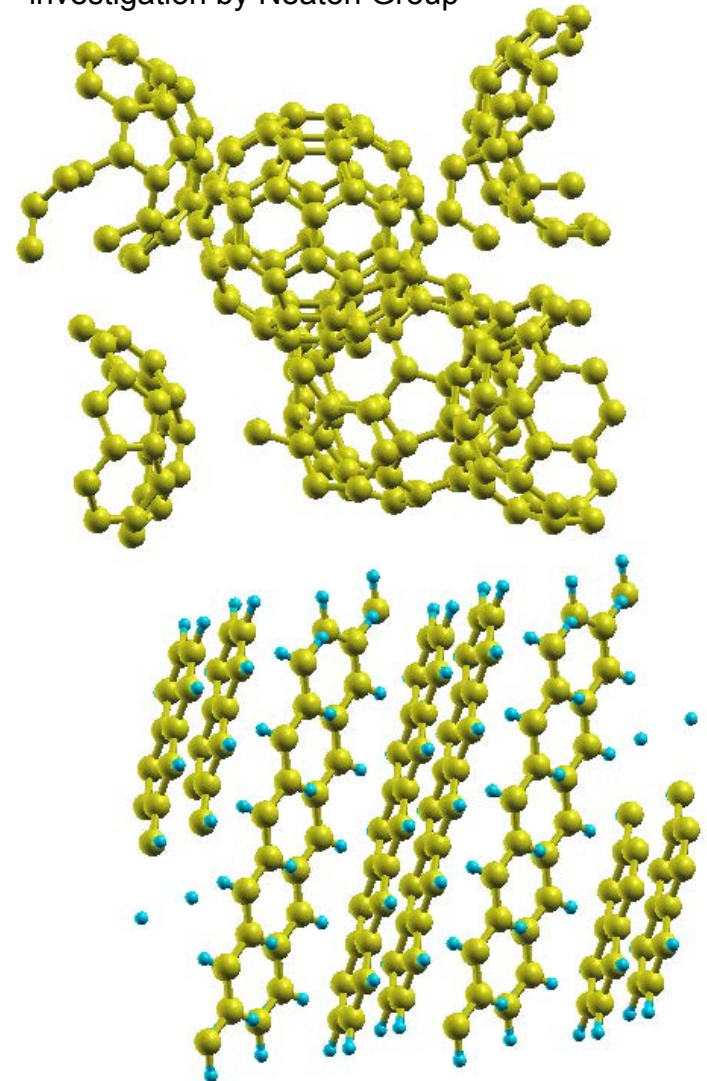
BerkeleyGW routinely run on systems with hundreds of atoms.

Order 1000 atoms possible with DOE HPC resources like Edison and Mira.

BerkeleyGW is ahead of the pack of GW codes.

- Foundry users find Full-Frequency calculation of MgO takes hours with BerkeleyGW, weeks with abinit
- Yambo celebrates passing the 1000 CPU core in 2014, BerkeleyGW commonly run on 10-100x that scale.
- VASP GW limited in size by memory requirements of (G, G') matrices.

C60 Pentacene Interface under investigation by Neaton Group



Our Posters



SciDAC Program: Scalable Computational Tools for Discovery and Design – Excited State Phenomena in Energy Materials

Ultrafast Dynamics of Excited Electrons in Materials for Energy Application

Marco Bernardi, Derek Vigil-Fowler, Jamal Mustafa, Chin Shen Ong, Jeffrey B. Neaton and Steven G. Louie
Department of Physics, University of California at Berkeley, and Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley CA 94720

Ultrafast Dynamics of Excited Carriers

Two ultrafast (<1ps) mechanisms for hot carriers to lose energy

Electron – phonon scattering

Electron – electron scattering
"Impact ionization" & Auger processes

Use perturbation theory

$$g_{\mathbf{k}\mathbf{q}} = -\frac{1}{\Omega} \sum_{\mathbf{r}} e^{i\mathbf{q}\cdot\mathbf{r}} \nabla V_{\mathbf{r}} \cdot \mathbf{e}_{\mathbf{k}} \langle \mathbf{k} | \mathbf{r} | \mathbf{k} + \mathbf{q} \rangle$$

all sp matrix elements
interpolate on fine grids

$$(\tau_{\text{e-ph}})^{-1} = \text{Im}(\sum_{\mathbf{q}} \sum_{\mathbf{k}'} g_{\mathbf{k}\mathbf{q}}^2 \rho_{\mathbf{k}\mathbf{k}'}^{\text{ph}}) = \text{Im}(g^2 G D)$$

DFT + Wannier electron-phonon approach

$$(\tau_{\text{e-e}})^{-1} = \text{Im}(\sum_{\mathbf{q}} \sum_{\mathbf{k}'} V_{\mathbf{q}} \rho_{\mathbf{k}\mathbf{k}'}^{\text{ee}}) = \text{Im}(G^{\text{ee}}(E_{\text{hot}}) F)$$

Ab initio GW method

Relaxation time of carrier in state \mathbf{k} : $(\tau_{\text{rel}})^{-1} = \text{Im}(\sum_{\mathbf{q}} \sum_{\mathbf{k}'} V_{\mathbf{q}} \rho_{\mathbf{k}\mathbf{k}'}^{\text{ee}}) + \text{Im}(\sum_{\mathbf{q}} \sum_{\mathbf{k}'} g_{\mathbf{k}\mathbf{q}}^2 \rho_{\mathbf{k}\mathbf{k}'}^{\text{ph}})$

Understand excited ("hot") carriers in materials:

- Energy distribution vs. time
- Timescale (10–100 fs) for energy loss
- Transport and mean free paths

Codes: Quantum ESPRESSO, BerkeleyGW, EPW. We used in-house modified versions

Hot Carriers in Silicon Solar Cells

Hot carrier relaxation time

Hot carrier mean free path

Fast (10 fs) away from band edge
Slower (100 fs) near the band edge

Mean free paths ~5–10 nm in Si
Anisotropy: (100) most favorable

Calculate agreement with experiment:
thermalization in 250–300 fs near CBM

Highlight of the DOE-BES, press articles by LBNL and several journals and websites

First Ab Initio Method for Characterizing Hot Carriers Could Hold the Key to Future Solar Cell Efficiencies

Reference: M. Bernardi, D. Vigil-Fowler, J. Li, J. Neaton, J.B. Neaton, S.G. Louie, *Phys. Rev. Lett.* **112**, 257402 (2014)

Hot Electrons in GaAs

Compute ~50 billion of sp matrix elements
Where did the "old" (semispherical) calculations go wrong?

Semispherical: Multiple parameters for sp-ph coupling

Exact agreement with experiment:
thermalization in 250–300 fs near CBM

Excellent agreement with pump-probe experiments
Help resolve a long-standing controversy on HCs in GaAs

Reference: M. Bernardi, D. Vigil-Fowler, C.S. Ong, J.B. Neaton, S.G. Louie, *PNAS* **112**, 5291 (2015)

Generation and relaxation of hot carriers from surface plasmons in noble metals

Energy and momentum conserving transitions induced by surface plasmon polaritons (SPPs) generate hot carriers with a distribution of energies (in a probabilistic sense)

Results for Au, Ag and Cu show similar trends

Application to photocatalysis with hot carriers

Interband ($\sigma \rightarrow d$)
Intraband ($d \rightarrow d$)

Step size may be better than for hot carriers from SPPs

Reference: M. Bernardi, J. Mustafa, J.B. Neaton, S.G. Louie, *Nature Commun.* **6**, 7044 (2015)

Carrier Transport from First Principles

Electron-phonon (e-ph) scattering controls the resistivity of metals at room temperature
Using importance sampling, we map the e-ph relaxation times (RT) on the Fermi surface

Conductivity $\sigma_{xx} = e^2 \sum_{\mathbf{k}} v_{\mathbf{k}}^2 \tau_{\text{rel}}(\mathbf{k}) \rho_{\mathbf{k}\mathbf{k}}$

Typical ab initio calculations combine DFT bands with the RT as a parameter
Using GW and ab initio e-ph RTs, we can predict the resistivity of noble metals within 10%.

Accuracy

Material	Calculated	Experimental
Cu	1.27	1.20
Ag	1.90	1.69
Au	1.95	2.36

SC = single crystal; FC = polycrystalline

Faster convergence

This research was supported by SciDAC Program on Excited State Phenomena in Energy Materials funded by the U.S. DoE, Office of Basic Energy Sciences, and of Advanced Scientific Computing Research, under Contract No. DE-AC02-05CH11231 at LBNL and partly by the NSF under grant DMR10-1006184. The research used resources at DoE's National Energy Research Scientific Computing Center.

Algorithms for Electronic Structure Calculations in Real Space

Charles Lena, N. Scott Bobbitt, Grady Schofield, James R. Chelikowsky
University of Texas at Austin

Large Scale Ground State Calculations

$$\left[-\frac{\hbar^2 \nabla^2}{2m} + V_{\text{ion}}(\mathbf{r}) + V_{\text{ext}}(\mathbf{r}) + V_{\text{ex}}(\mathbf{r}) \right] \psi_{\mathbf{k}}(\mathbf{r}) = E_{\mathbf{k}} \psi_{\mathbf{k}}(\mathbf{r})$$

Goals: Provide fast and accurate ab initio computational methods for designing and predicting new energy related materials. We want to reliably apply a massively parallel treatment of pseudopotential-density functional theory to complex systems in the 100,000+ atom range.

Utilizing massively parallel computers at NERSC, we can now handle systems with over 10,000 atoms employing algorithms based on subspace filtering [1].

Example shown here: Large silicon nanocrystal passivated with hydrogen atoms. The nanocrystal density of state has converged to that expected for a bulk crystal.

High Order Force Calculations

Atomic forces require a finer grid spacing than total energy calculations. The cost of the computation increases roughly as the cube of the grid spacing.

We apply an improved integration technique to the force calculation and demonstrate that we can calculate accurate, converged forces using a coarser grid spacing with this new technique [2].

$$F = \sum_{\mathbf{r}} \frac{Z_{\mathbf{r}} Z_{\mathbf{r}'} e^2}{R^3} + \int \rho(\mathbf{r}) \frac{dV_{\text{ex}}(\mathbf{r}-\mathbf{R})}{d\mathbf{R}}$$

Integration was originally done with a Riemann sum of cubes at each grid point.

The wave function is approximated on intermediate points by several high order Taylor series.

Combining with a finite difference scheme provides a smooth approximation of the wave function.

A 20 point Gaussian quadrature rule is applied for the integration.

Improving Integration: Impact

Left: The improved integration scheme reduces variations in energy as a benzene molecule is translated through space.

Right: Vibrational frequencies for CO₂. The high order result attains comparable accuracy as the low order result, but uses a much coarser grid.

Low Order	Low Order	High Order	Exp. [Ref. 3]
h=0.15	h=0.30	h=0.30	
2415	2301	2400	2349
1346	1517	1335	1333
642	605	628	667
641	603	628	667

Computational Savings

The force between two Si atoms in a cluster as a function of grid spacing. With the high order integration scheme, the force is converged up to grid spacing with $h=0.50$, while low order results diverge at $h=0.35$.

Using the high order scheme, we can compute the vibrational spectrum of Si₂₀H₄₀ using a grid spacing of $h=0.40$ instead of $h=0.20$ with the low order scheme.

References

1. Y. Zhou, Y. Savd, M.L. Tiago, J.R. Chelikowsky, *Phys. Rev. E*, **74**, 2096; G. Schofield, J.R. Chelikowsky, Y. Savd, *Comp. Phys. Comm.* **183**, 497 (2012).
2. N.S. Bobbitt, G. Schofield, C. Lena, J.R. Chelikowsky, *Physical Chemistry Chemical Physics*, **2015**, DOI: 10.1039/C5CP02651A
3. T. Shimouchi, Tables of molecular vibrational frequencies consolidated, volume 1. Technical report, DTIC Document, 1972

Acknowledgments

We wish to acknowledge support provided by the Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy, Office of Science Advanced Scientific Computing Research and Basic Energy Sciences under award number DESC0008677. This research used resources of the National Energy Research Scientific Computing Center.

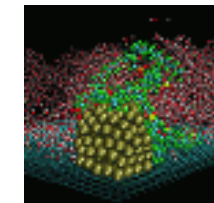
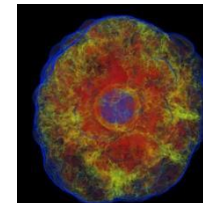
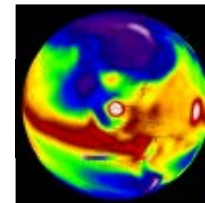
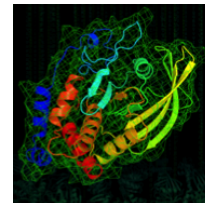
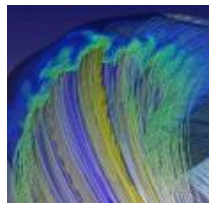
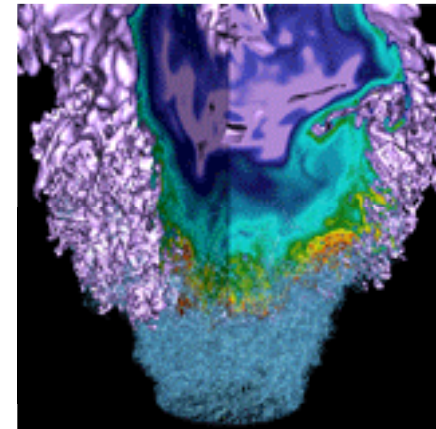
Acknowledgements



This work could not have been done without SCIDAC!

SCIDAC Program on Excited State Phenomena in Energy Materials funded by the U. S. Department of Energy, Office of Basic Energy Sciences and of Advanced Scientific Computing Research, under Contract No. DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory and under Award No. DESC0008877 at University of Texas, Austin

Extra Slides

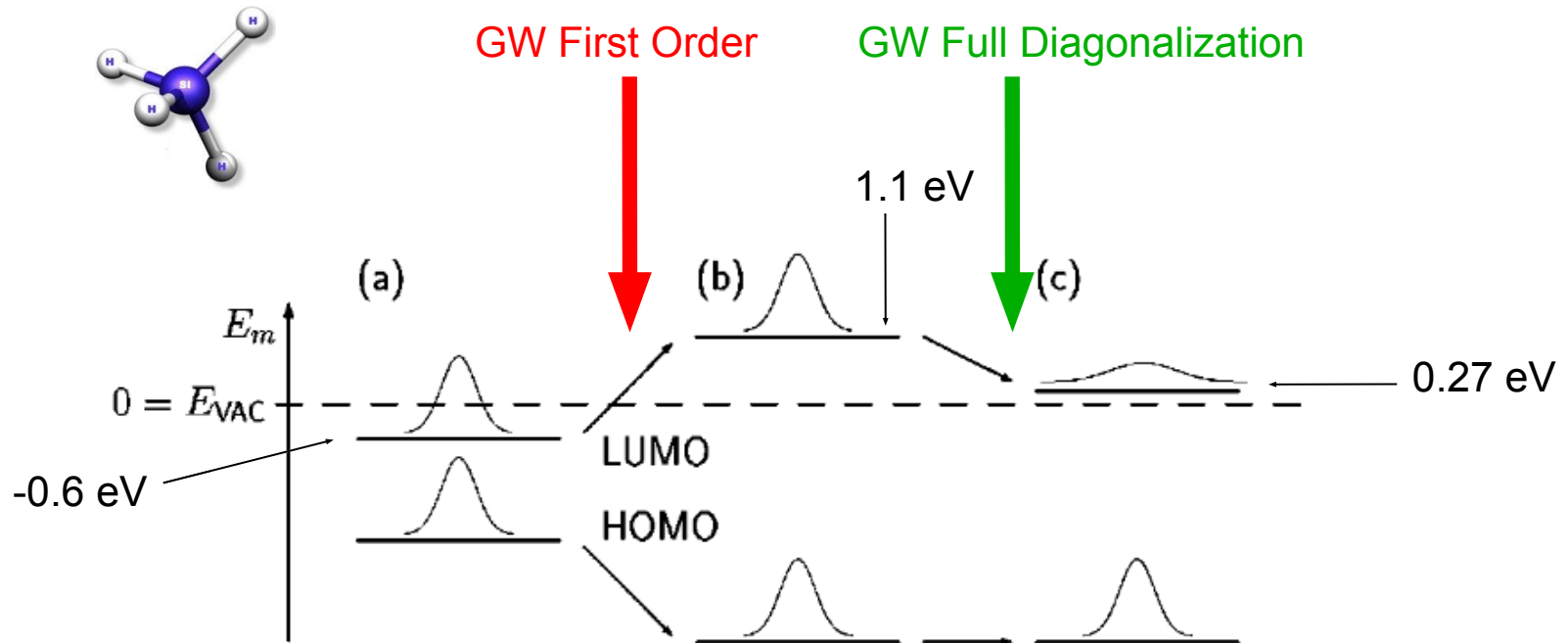


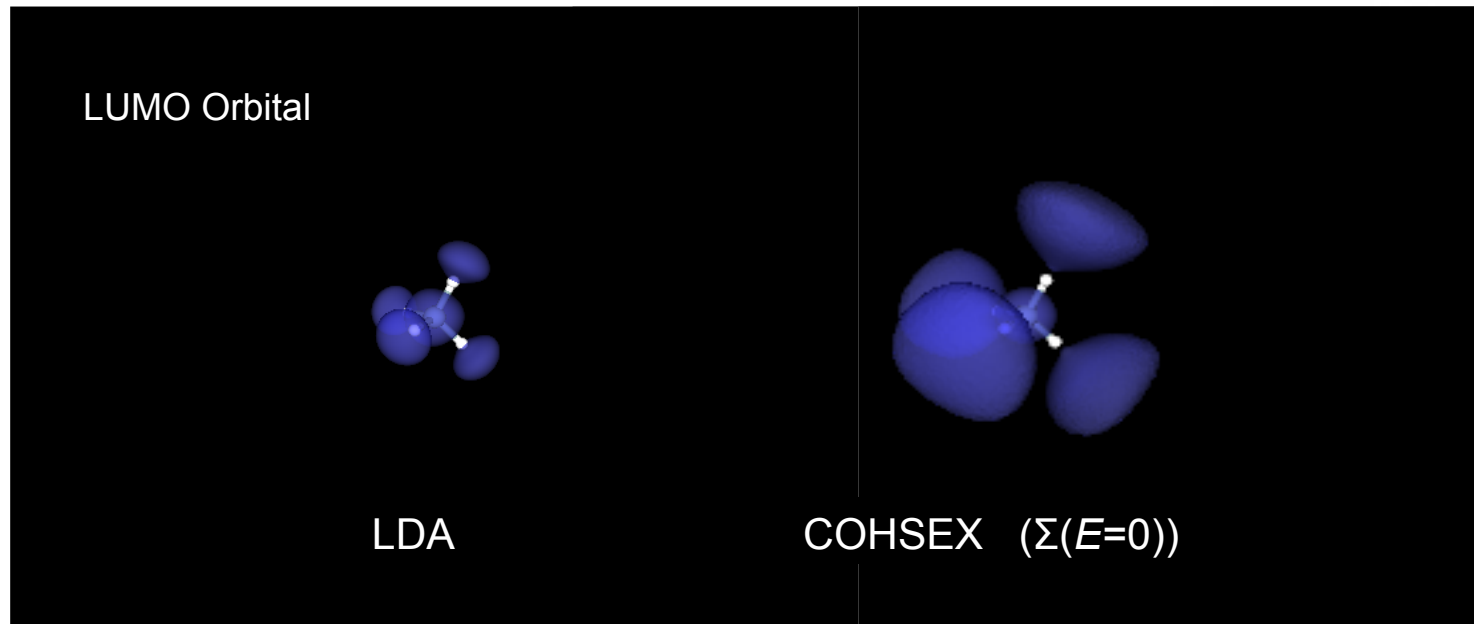
Using the Static Limit To Improve GW

For a typical GW calculation, the LDA starting point is sufficient:

$$E_n^{QP} \approx \langle \Psi_n^{MF} | H_{Hartree} | \Psi_n^{MF} \rangle + \langle \Psi_n^{MF} | \Sigma | \Psi_n^{MF} \rangle$$

Notable exceptions - Silane:

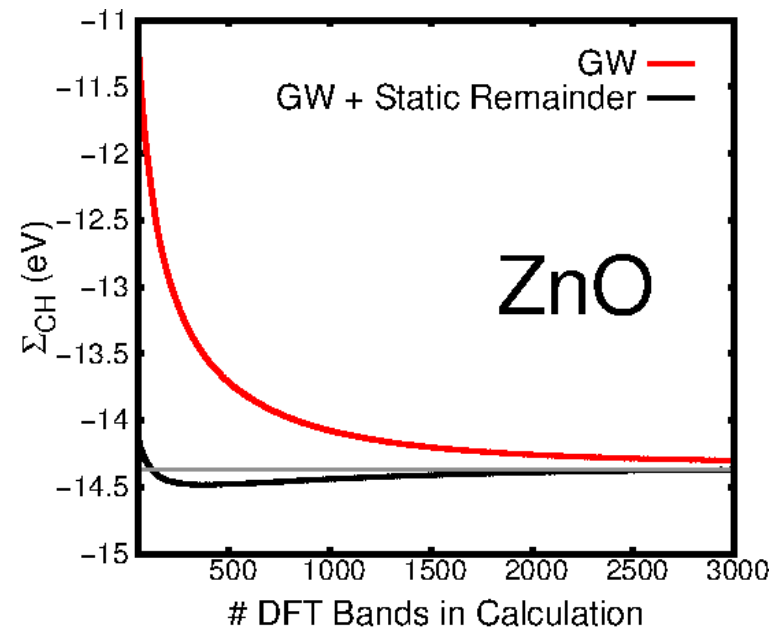
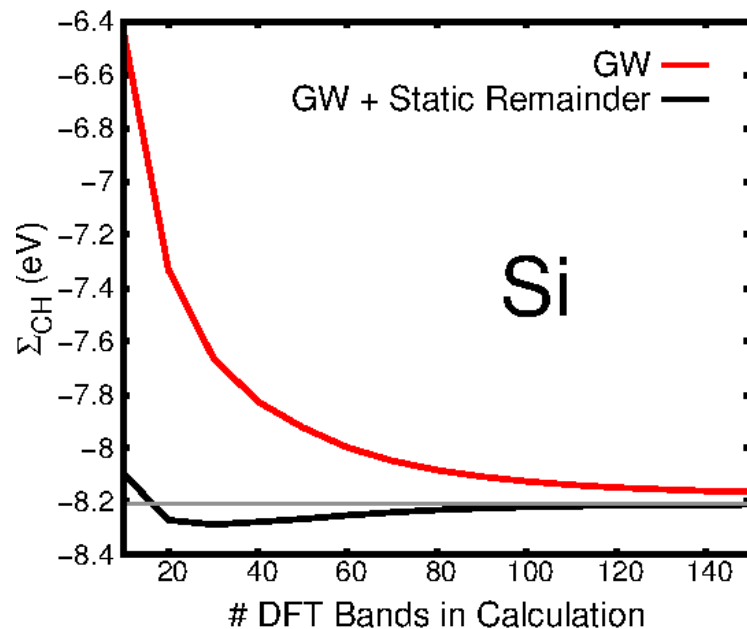




	LDA	LDA+GW	CSX	CSX+GW
HOMO	-8.52	-12.80	-13.2	-12.80
LUMO	-0.465	1.02	0.1	0.29
QP gap	8.06	13.82	13.3	13.10

Using the Static Limit to Improve GW

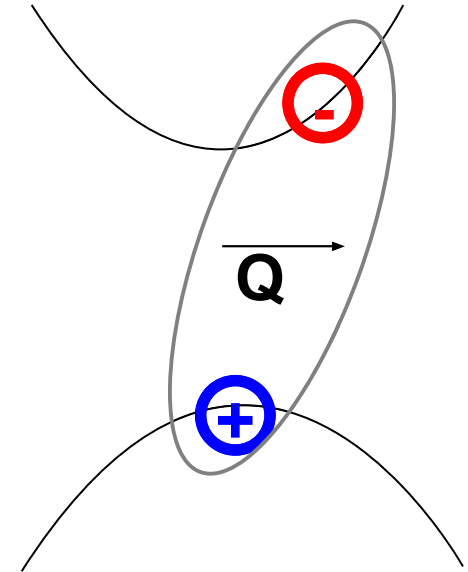
$$\langle n\mathbf{k} | \Sigma_{\text{CH}}^{\infty}(\mathbf{r}, \mathbf{r}'; E) | n'\mathbf{k} \rangle = \langle n\mathbf{k} | \Sigma_{\text{CH}}^N(\mathbf{r}, \mathbf{r}'; E) | n'\mathbf{k} \rangle + \frac{1}{2} \left(\langle n\mathbf{k} | \Sigma_{\text{CH}}^{\text{Coh}/\infty}(\mathbf{r}, \mathbf{r}') | n'\mathbf{k} \rangle - \langle n\mathbf{k} | \Sigma_{\text{CH}}^{\text{Coh}/N}(\mathbf{r}, \mathbf{r}') | n'\mathbf{k} \rangle \right).$$



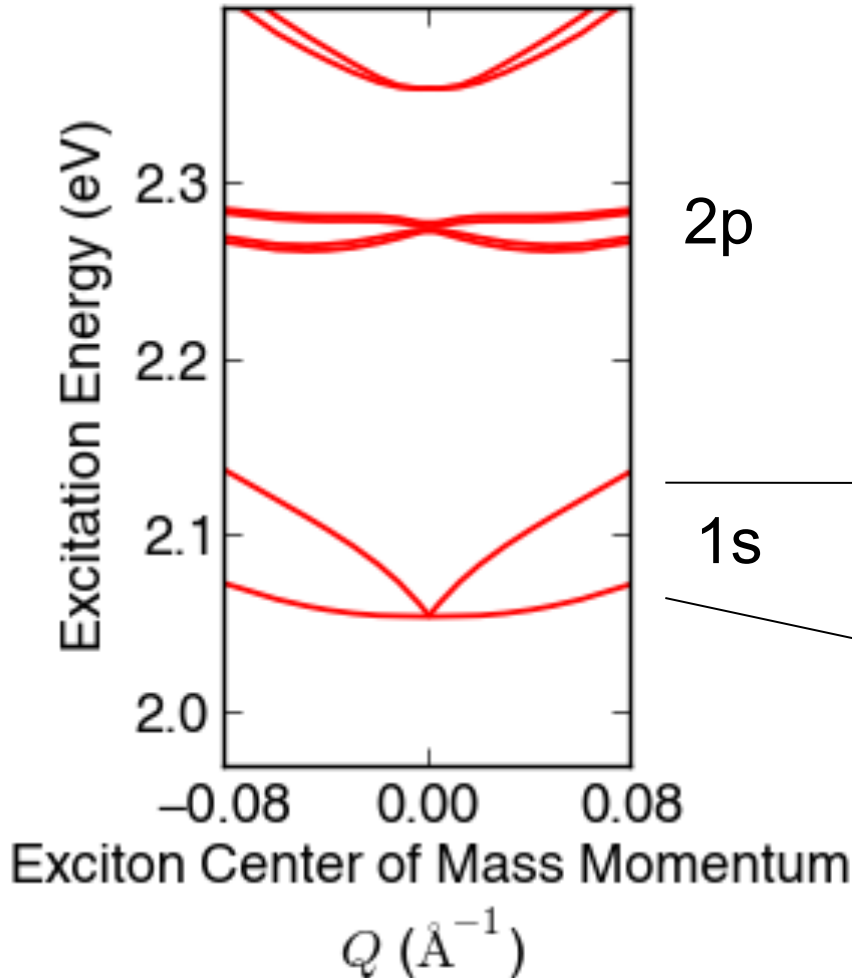
Finite Momentum Excitons



- $Q \rightarrow 0$
 - Optical absorption $\sim \text{Im } \epsilon_M$
- $Q \neq 0$
 - Optically inactive
 - Energy loss $\sim -\text{Im } \epsilon^{-1}$
 - Important for exciton dynamics, relaxation, ...



Finite Momentum Excitons



- Ab-initio exciton bandstructure of MoS2
- Access novel physics:
 - Nonanalytic dispersion due to 2D Coulomb interaction
 - Valley quantum phase with valley pseudospin winding number = 2
- Couples to longitudinal electric field
- Couples to transverse electric field

GW Parallelism



GW is an ideal case for Many-Core / Exascale. Many levels of parallelism can be exploited. Ideal for many-core.

$$\begin{aligned}
 \epsilon_{\mathbf{G}\mathbf{G}'}^{r/a}(\mathbf{q}, E) &= \delta_{\mathbf{G}\mathbf{G}'} - v(\mathbf{q} + \mathbf{G}) \\
 &\times \sum_n^{\text{occ}} \sum_{n'}^{\text{emp}} \sum_{\mathbf{k}} M_{nn'}^*(\mathbf{k}, \mathbf{q}, \mathbf{G}) M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}') \\
 &\times \frac{1}{2} \left[\frac{1}{E_{n\mathbf{k}+\mathbf{q}} - E_{n'\mathbf{k}} - E - i\delta} + \frac{1}{E_{n\mathbf{k}+\mathbf{q}} - E_{n'\mathbf{k}} + E + i\delta} \right]
 \end{aligned}$$

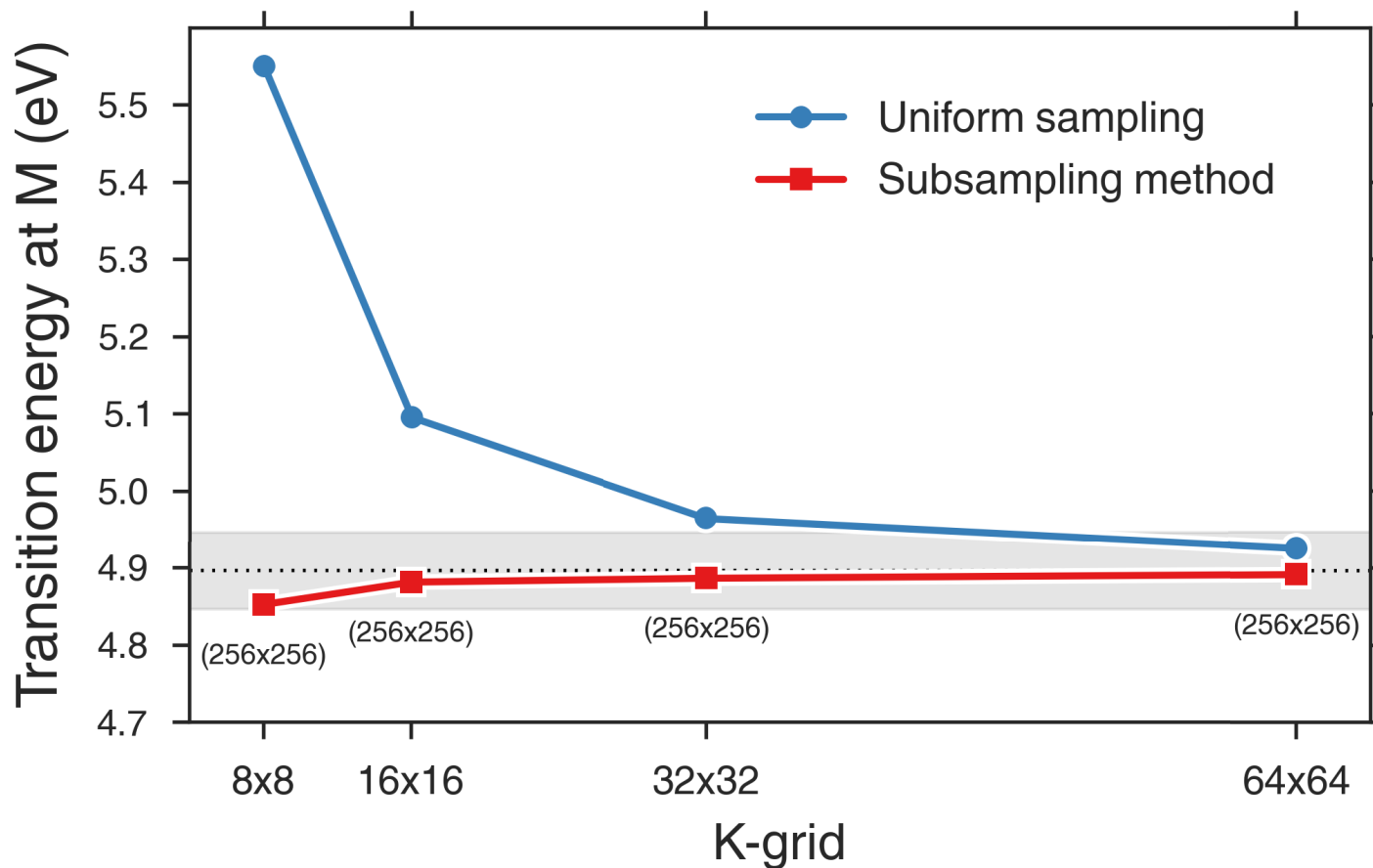
Inner Dimension of Hybrid MPI-OpenMP ZGEMM (OpenMP new to 1.2)

Outer Dimension of Hybrid MPI-OpenMP ZGEMM (OpenMP new to 1.2)

MPI Group Level Parallelization over Frequencies (New to 1.2)

Trivially parallelization over \mathbf{q} points

Graphene



Simplified Final Loop Structure



```
!$OMP DO reduction(+:achtemp)
do my_igp = 1, ngpown
...
do iw=1,3
  scht=0D0
  wxt = wx_array(iw)
  do ig = 1, ncouls
    !if (abs(wtilde_array(ig,my_igp) * eps(ig,my_igp)) .lt. TOL) cycle
    wdiff = wxt - wtilde_array(ig,my_igp)
    delw = wtilde_array(ig,my_igp) / wdiff
    ...
    scha(ig) = mygpvar1 * aqsntemp(ig) * delw * eps(ig,my_igp)
    scht = scht + scha(ig)
  enddo ! loop over g
  sch_array(iw) = sch_array(iw) + 0.5D0*scht
enddo
achtemp(:) = achtemp(:) + sch_array(:) * vcoul(my_igp)
enddo
```

ngpown typically in 100's to 1000s. Good for many threads.

Original inner loop. Too small to vectorize!

ncouls typically in 1000s - 10,000s. Good for vectorization. Don't have to worry much about memory alignment.

Attempt to save work breaks vectorization and makes code slower.

1. Compute via $n \times n'$ FFTs (N^3 Step. Big Prefactor.):

$$M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) = \langle n\mathbf{k} + \mathbf{q} | e^{i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r}} | n'\mathbf{k} \rangle$$

$$M_{nn'}(\mathbf{k}, \mathbf{q}, \{\mathbf{G}\}) = FFT^{-1} (\phi_{n, \mathbf{k} + \mathbf{q}}(\mathbf{r}) * \phi_{n', \mathbf{k}}^*(\mathbf{r}))$$

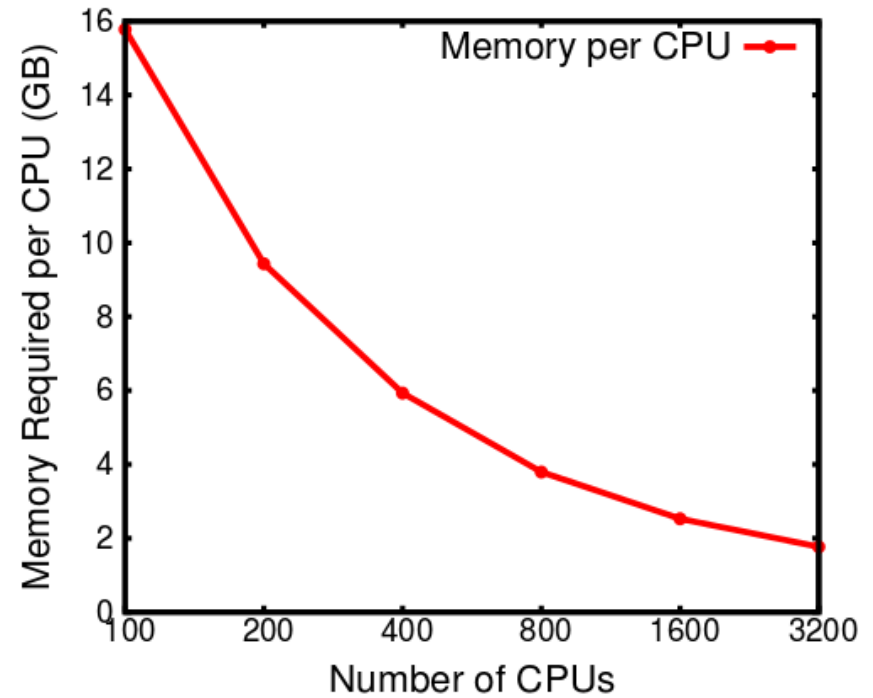
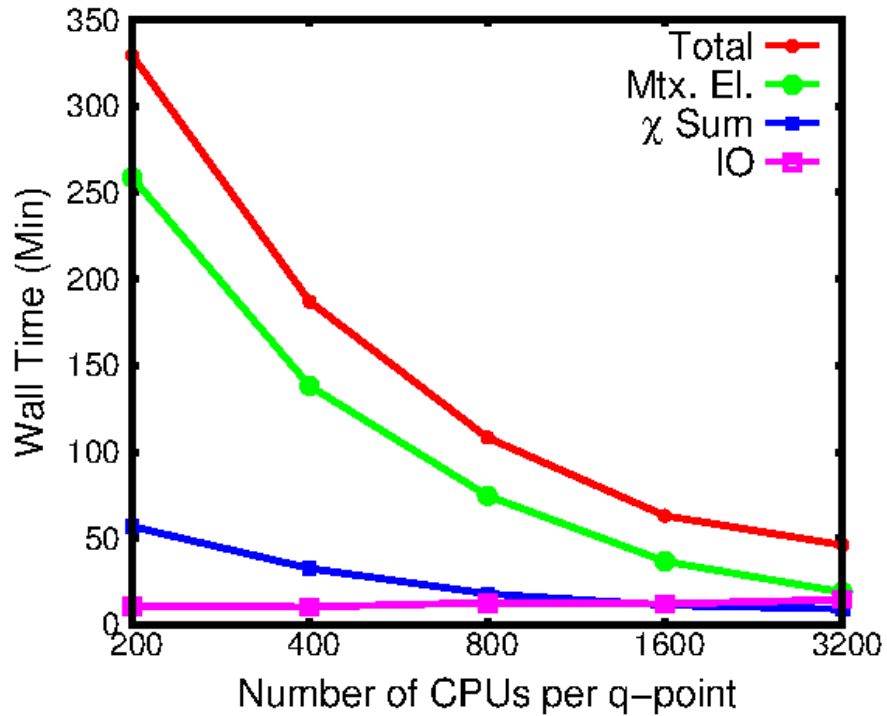
2. Compute sum via large ZGEMM (N^4 Step. Small Prefactor. All to All Communication Done):

$$\chi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}; 0) = \mathbf{M}(\mathbf{G}, \mathbf{q}, (n, n', \mathbf{k})) \cdot \mathbf{M}^T(\mathbf{G}', \mathbf{q}(n, n', \mathbf{k}))$$

$$\text{Where, } \mathbf{M}(\mathbf{G}, \mathbf{q}, (n, n', \mathbf{k})) = M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) \cdot \frac{1}{\sqrt{E_{n\mathbf{k} + \mathbf{q}} - E_{n'\mathbf{k}}}}$$

3. Matrix Inversion. ScaLAPACK

MPI Scaling of Epsilon Code:



(Sigma GPP Option)

4. Manual loop reductions to compute sum for self-energy.

N^3 x <number of bands of interest>

$$\langle n\mathbf{k} | \Sigma_{\text{SX}}(E) | n'\mathbf{k} \rangle = - \sum_{n''}^{\text{occ}} \sum_{\mathbf{q}\mathbf{G}\mathbf{G}' } M_{n''n}^*(\mathbf{k}, -\mathbf{q}, -\mathbf{G}) M_{n''n'}(\mathbf{k}, -\mathbf{q}, -\mathbf{G}') \times \left[\delta_{\mathbf{G}\mathbf{G}'} + \frac{\Omega_{\mathbf{G}\mathbf{G}'}^2(\mathbf{q}) (1 - i \tan \phi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}))}{(E - E_{n''\mathbf{k}-\mathbf{q}})^2 - \tilde{\omega}_{\mathbf{G}\mathbf{G}'}^2(\mathbf{q})} \right] v(\mathbf{q} + \mathbf{G}')$$

$$\langle n\mathbf{k} | \Sigma_{\text{CH}}(E) | n'\mathbf{k} \rangle = \frac{1}{2} \sum_{n''} \sum_{\mathbf{q}\mathbf{G}\mathbf{G}' } M_{n''n}^*(\mathbf{k}, -\mathbf{q}, -\mathbf{G}) M_{n''n'}(\mathbf{k}, -\mathbf{q}, -\mathbf{G}') \times \frac{\Omega_{\mathbf{G}\mathbf{G}'}^2(\mathbf{q}) (1 - i \tan \phi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}))}{\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}) (E - E_{n''\mathbf{k}-\mathbf{q}} - \tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}))} v(\mathbf{q} + \mathbf{G}')$$

Compilers want to “vectorize” your loops whenever possible. But sometimes they get stumped. Here are a few things that prevent your code from vectorizing:

Loop dependency:

```
do i = 1, n
    a(i) = a(i-1) + b(i)
enddo
```

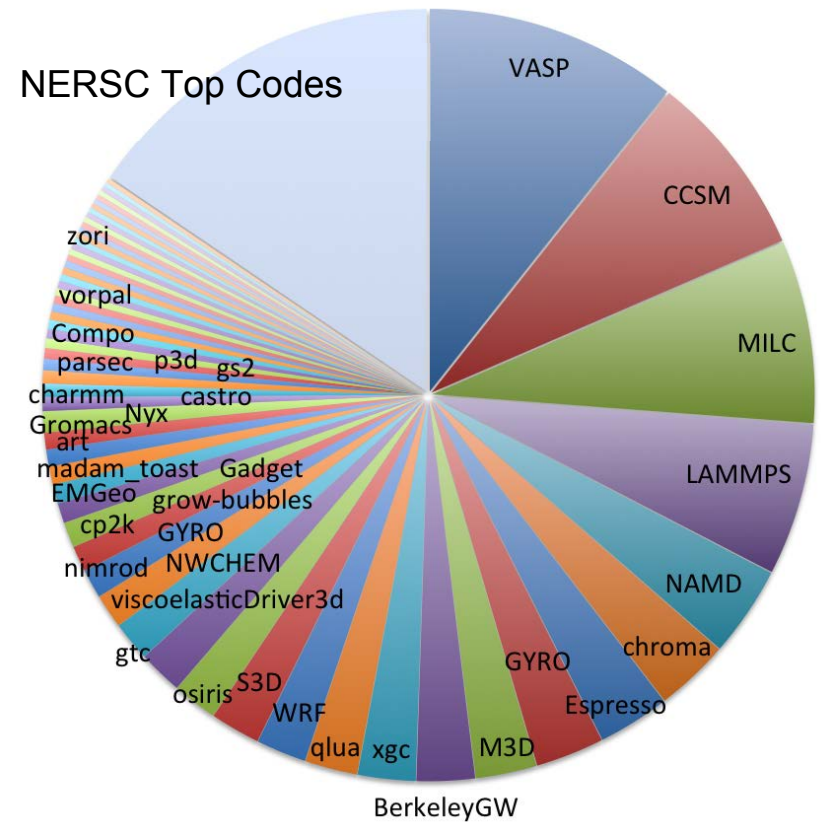
Task forking:

```
do i = 1, n
    if (a(i) < x) cycle
    if (a(i) > x) ...
enddo
```

Application Readiness



- NERSC partnering with selected projects (~20) to help prepare application codes for Cori
- The program will provide:
 - early access to NERSC-8 hardware and testbed systems
 - special vendor (Cray + Intel) training and optimization sessions
 - NERSC Staff support and training



Things that prevent vectorization in your code



Example From NERSC User Group Hackathon - (Astrophysics Transport Code)

```
for (many iterations) {  
  ... many flops ...  
  et = exp(outcome1)  
  tt = pow(outcome2,3)  
  IN = IN * et + tt  
}
```



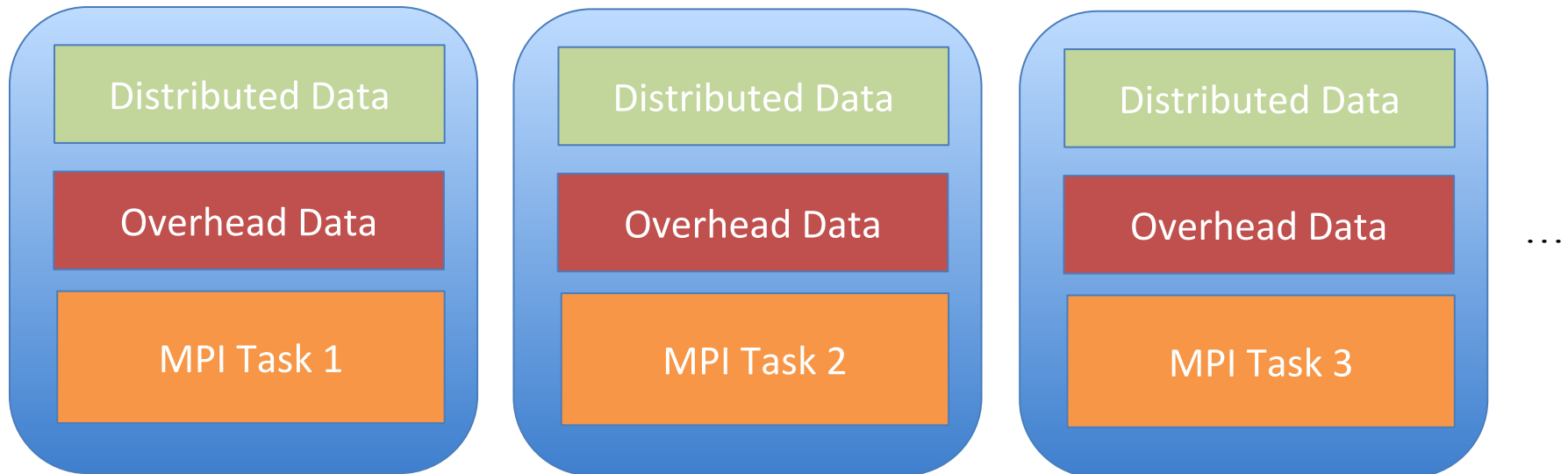
```
for (many iterations) {  
  ... many flops ...  
  et(i) = exp(outcome1)  
  tt(i) = pow(outcome2,  
3)  
}  
for (many iterations) {  
  IN = IN * et(i) + tt(i)  
}
```

**30% speed up for entire
application!**

Failure of the MPI-Only Programming Model in BerkeleyGW

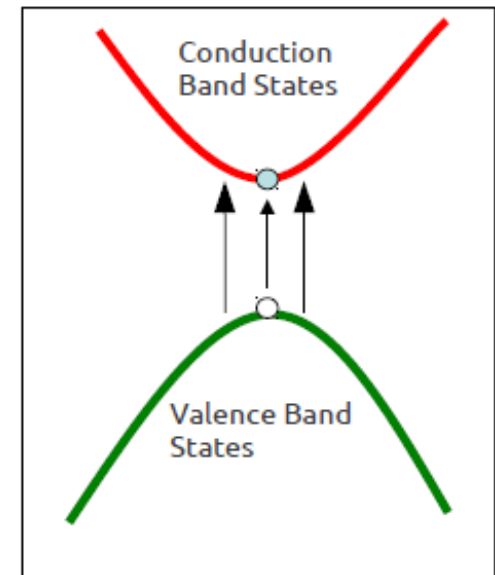
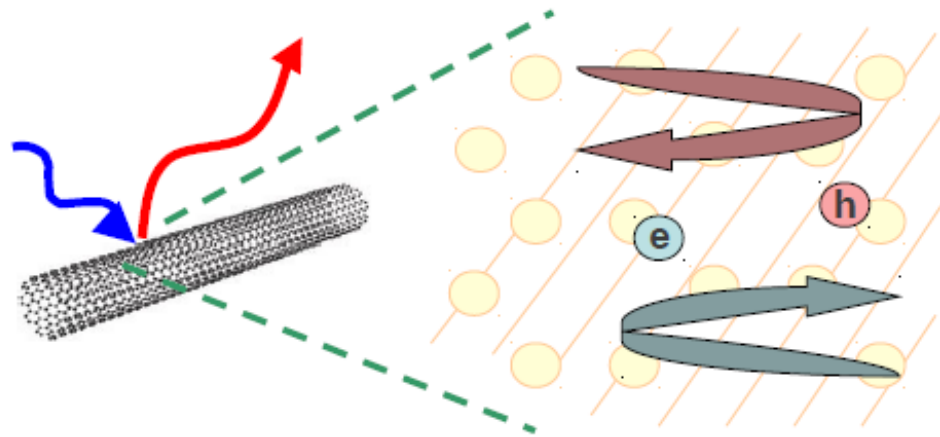


- ★ Big systems require more memory. Cost scales as N_{atm}^2 to store the data.
- ★ In an MPI GW implementation, in practice, to avoid communication, data is duplicated and **each MPI task has a memory overhead.**
- ★ On Edison, users sometimes forced to use 1 of 24 available cores, in order to provide MPI tasks with enough memory. **90% of the computing capability is lost.**



What is GW+BSE

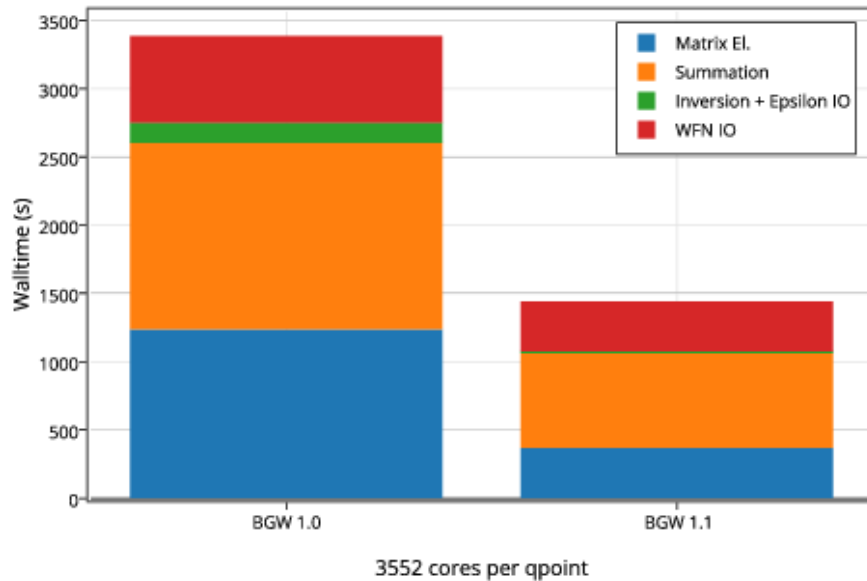
$$|N, S\rangle = \sum_v \sum_c^{\text{hole elec}} A_{vc}^S a_v^+ b_c^+ |N, 0\rangle + \dots$$



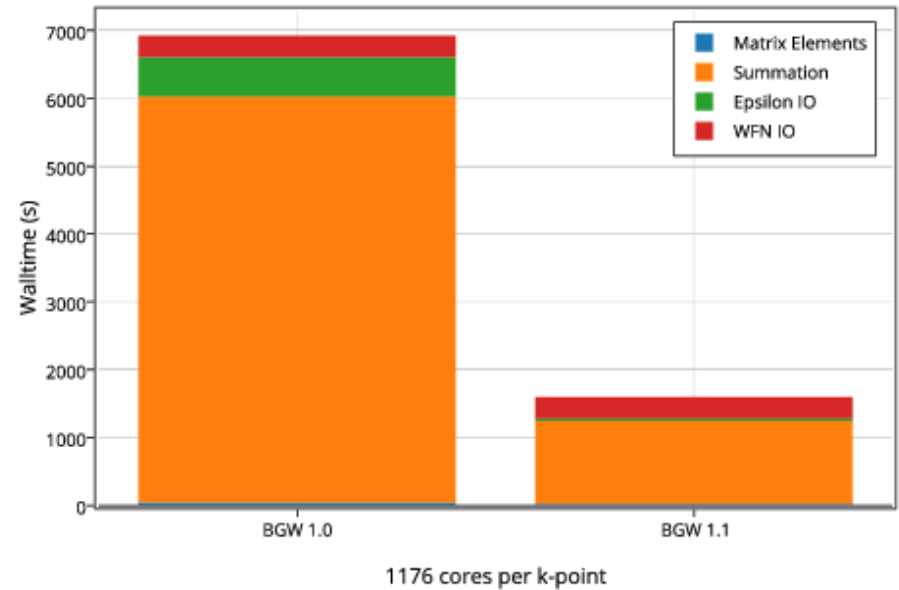
$$\left(E_{ck}^{QP} - E_{vk}^{QP} \right) A_{vck}^S + \sum_{k'v'c'} \langle vck | K^{eh} | v'c'k' \rangle A_{v'c'k'}^S = \Omega^S A_{vck}^S$$

Epsilon/Sigma Improvements

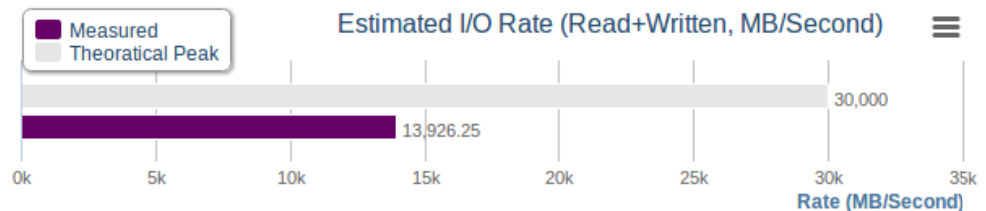
BGW 1.0 vs. 1.1 Epsilon Performance



BGW 1.0 vs. 1.1 Sigma Performance



- Performance improvements from:
 - Parallel IO (HDF5)
 - Vectorization
 - Memory-locality improvements
 - FFT Size/Performance Improvements





NESAP Participation

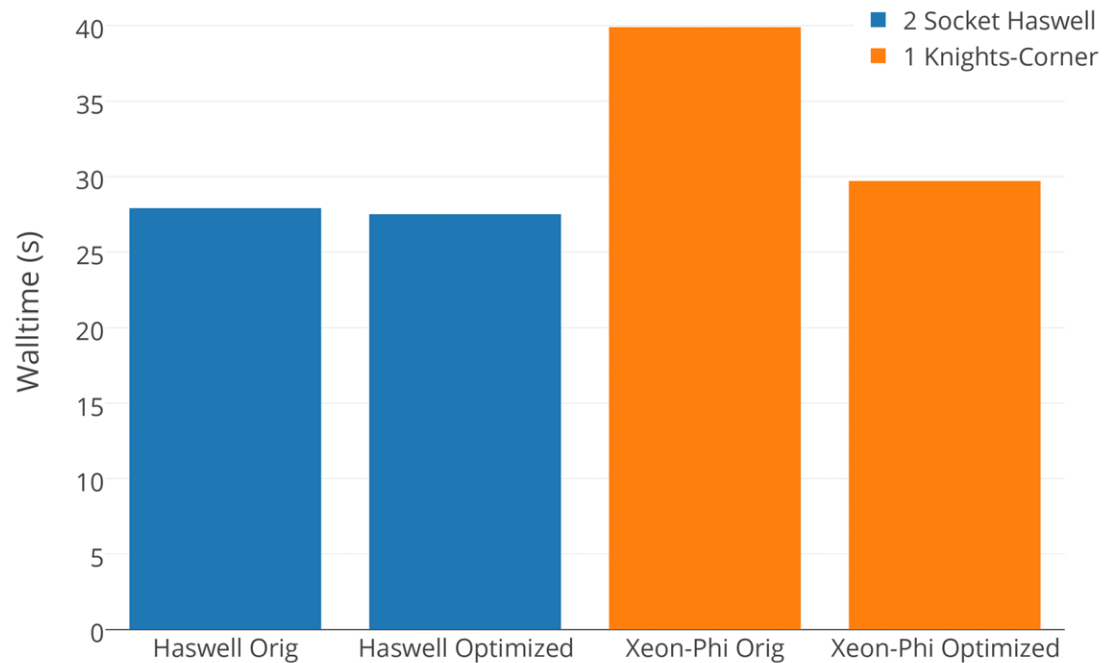
Both **PARSEC** and **BerkeleyGW** are included in the top tier of the NERSC Exascale Science Application Program

- **Early access to hardware**
 - Early access on the full Cori system
- **Technical deep dives**
 - Access to Cray and Intel staff on-site
 - Multi-day deep dive ('dungeon' session) with Intel staff at Oregon Campus
- **User Training Sessions**
 - From NERSC, Cray and Intel staff on OpenMP, vectorization, application profiling
 - Knights Landing architectural briefings from Intel

NESAP Advances with Cray, Intel and SUPER



- Worked with Cray, Intel and SUPER* to identify further bottlenecks in BerkeleyGW kernels.
- Added additional layers of cache-blocking to improve locality. Important on Xeon-Phi which lacks L3.

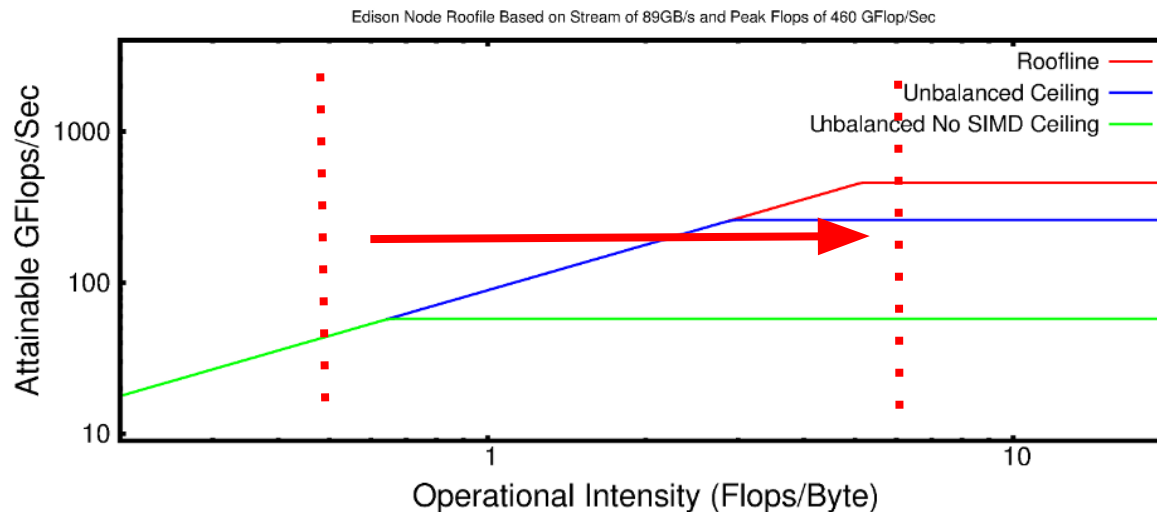


Memory Bandwidth Optimizations



What to do if your code is memory bandwidth bound?

1. Try to improve memory locality, cache reuse



2. Identify arrays leading to high bandwidth usage and make sure they are/will-be allocated in HBM on Cori.