OpenAtom:
First Principles GW method for electronic excitation

Minjung Kim, Subhasish Mandal, and Sohrab Ismail-Beigi
Yale University
Eric Mikida, Kavitha Chandrasekar, Eric Bohm, Nikhil Jain, and Laxmikant Kale
University of Illinois at Urbana-Champaign
Qi Li and Glenn Martyna
IBM T.J. Watson Research Center

Density Functional Theory (DFT)

Energy functional $E[n]$ of electron density $n(r)$

$$E[n] = KE + E_{ion} + E_H + E_{xc}$$

Minimizing over $n(r)$ gives exact

- Ground-state energy E_0
- Ground-state density $n(r)$

Minimum condition $\frac{\delta E}{\delta n(r)} = 0$ equivalent to Kohn-Sham equations

$$\left[-\frac{\nabla^2}{2} + V_{ion}(r) + V_H(r) + V_{xc}(r)\right]\psi_j(r) = \epsilon_j\psi_j(r) \qquad V_{xc}(r) = \frac{\delta E_{xc}}{\delta n(r)}$$

- LDA/GGA for E_{xc} : good geometries and total energies
- Bad band gaps and excitations

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

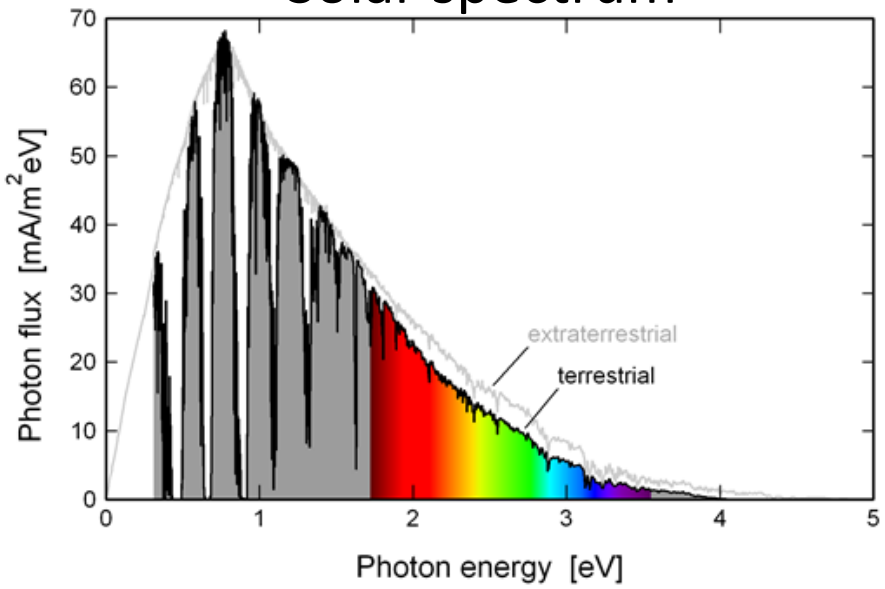
DFT: problems with excitations

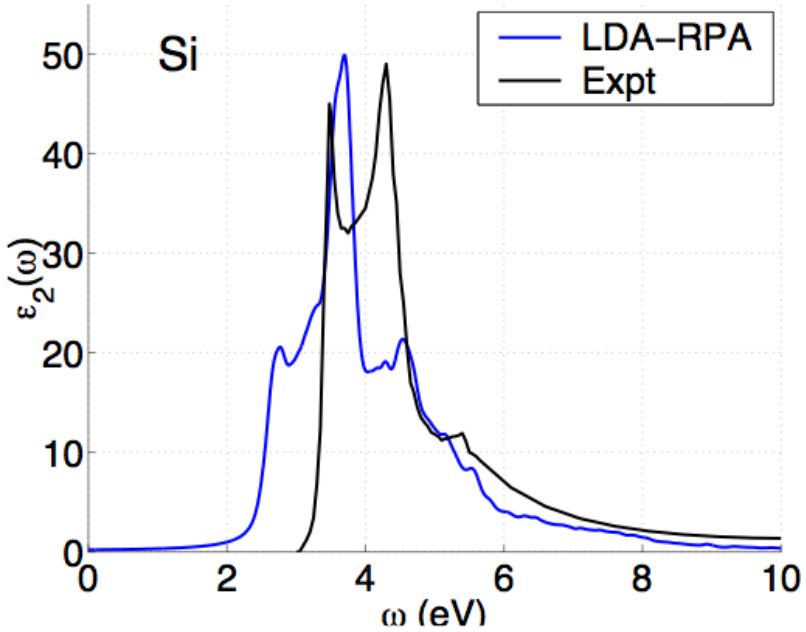
Energy gaps (eV)

Material	LDA	Expt. [1]
Diamond	3.9	5.48
Si	0.5	1.17
LiCl	6.0	9.4
$SrTiO_3$	2.0	3.25

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

Solar spectrum

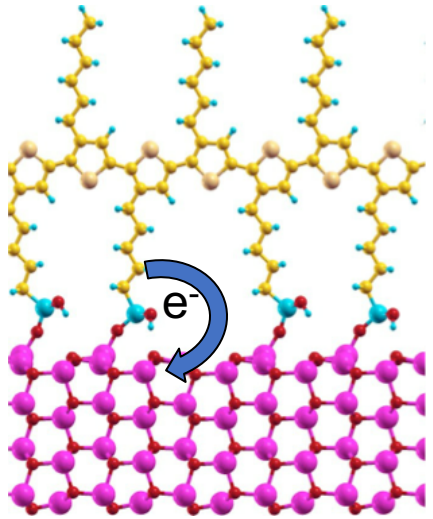


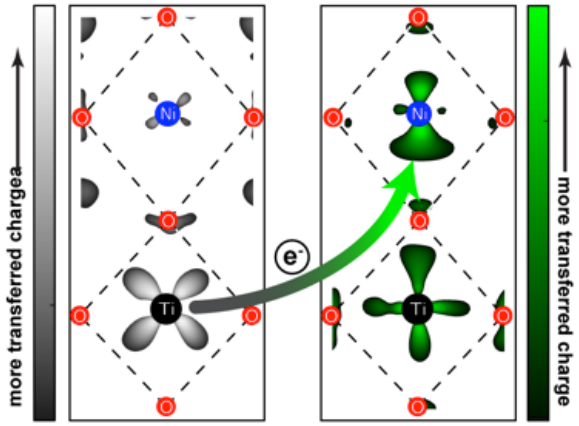


DFT: problems with energy alignment

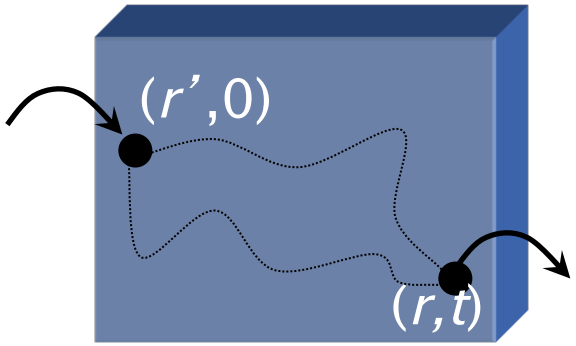
Interfacial systems:

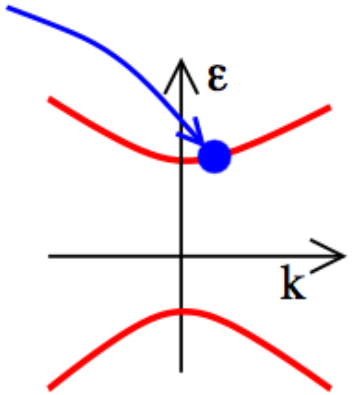
- Electrons can transfer across
- Depends on energy level alignment across interface
- DFT has errors in band energies
- Is any of it real?

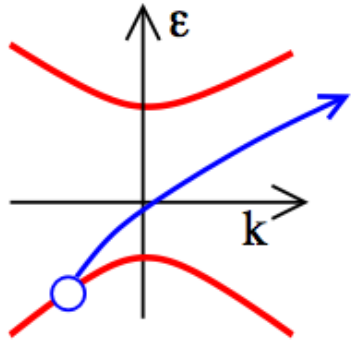




One particle Green's function







$$G_1(r,r',\omega) = \sum_j \frac{\psi_j(r)\psi_j(r')^*}{\omega - \epsilon_j}$$

Dyson Equation:

$$\left[\frac{-\hbar^2\nabla^2}{2m} + V_{ion}(r) + V_H(r)\right]\psi_j(r) + \int dr' \underline{\Sigma_{xc}(r,r',\epsilon_j)}\psi_j(r') = \epsilon_j\psi_j(r)$$

$$\Sigma \approx iG_1W \quad , \quad W = \varepsilon^{-1}(\omega) * v_c \quad (RPA)$$

DFT:

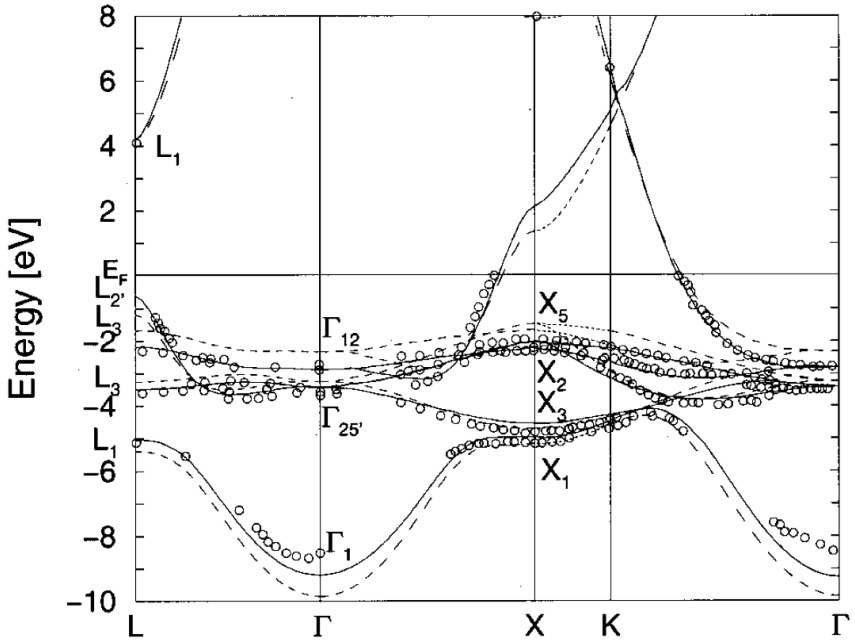
$$\left[-\frac{\nabla^2}{2} + V_{ion}(r) + V_H(r) + \underline{V_{xc}(r)}\right]\psi_j(r) = \epsilon_j\psi_j(r)$$

Green's function successes

Quasiparticle gaps (eV)

Material	LDA	GW	Expt.
Diamond	3.9	5.6*	5.48
Si	0.5	1.3*	1.17
LiCl	6.0	9.1*	9.4
$SrTiO_3$	2.0	3.4-3.8	3.25

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

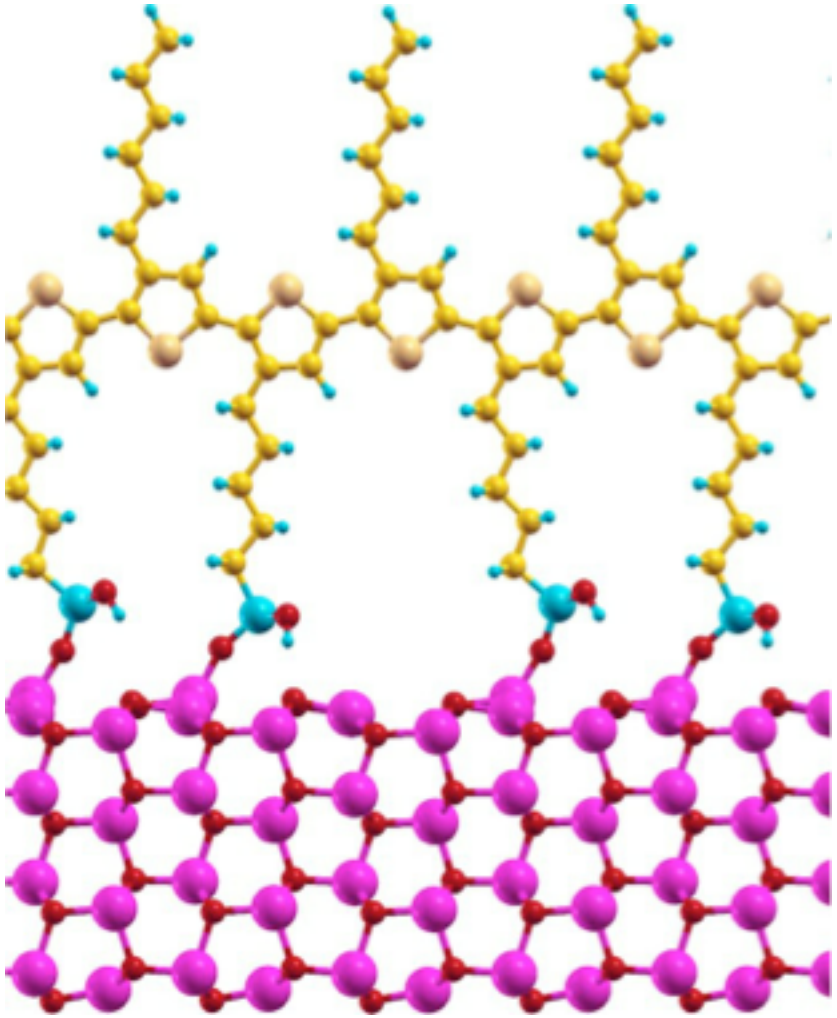


Band structure of Cu

Strokov *et al.*, PRL/PRB (1998/2001)

What is a big system for GW?

P3HT polymer



Zinc oxide nanowire

- Band alignment for this potential photovoltaic system?
- 100s of atoms/unit cell
- Not possible *routinely* (with current software)

GW is expensive

Scaling with number of atoms N

DFT:	N^3	
GW:	N^4	(gives better bands)
BSE:	N^6	(gives optical excitations)

But in practice the **GW** is the **killer**

a nanoscale system with 50-75 atoms (GaN)

DFT:	1	cpu x hours
GW:	91	cpu x hours
BSE:	2	cpu x hours

∴ **Focus on GW**

Steps for typical G_0W_0 calculation

Stage 1 : Run DFT calc. on structure → output : ε_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

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

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

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

What is so expensive in GW?

One key element : response of electrons to perturbation

$$P(r,r') = \frac{\partial n(r)}{\partial V(r')}$$

$P(r,r')$ = Response of electron density $n(r)$ at position r
to change of potential $V(r')$ at position r'

What is so expensive in GW?

One key element : response of electrons to perturbation

$$P(r,r') = \frac{\partial n(r)}{\partial V(r')} = -2 \sum_{v}^{\text{filled}} \sum_{c}^{\text{empty}} \frac{\psi_v(r)\psi_c(r)\psi_v(r')\psi_c(r')}{\varepsilon_v - \varepsilon_c}$$

Standard perturbation theory expression

Problems:

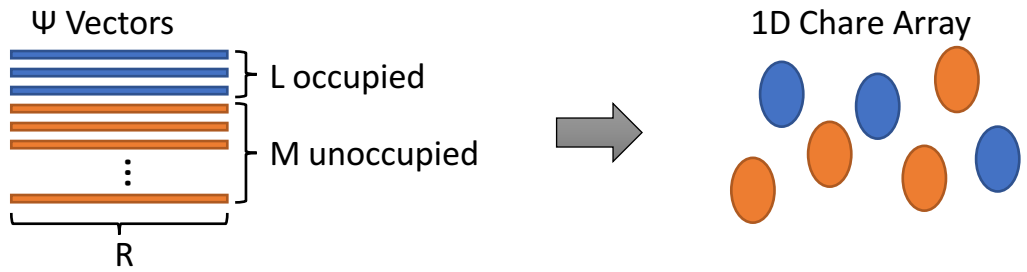
1. Must generate “all” empty states (sum over c)
2. Lots of FFTs to get functions $\psi_i(r)$ functions
3. Enormous outer produce to form P
4. Dense r grid : P huge in memory

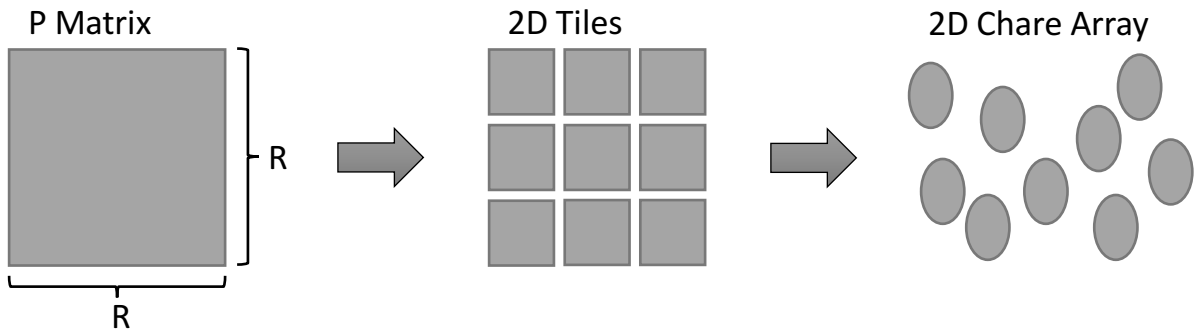
Computing P in Charm++

Basic Computation: $f_{lm} = \psi_l \times \psi_m^*$ for all l, m

$P \mathrel{+}= f_{lm}\ f_{lm}^{\dagger}$ for all f

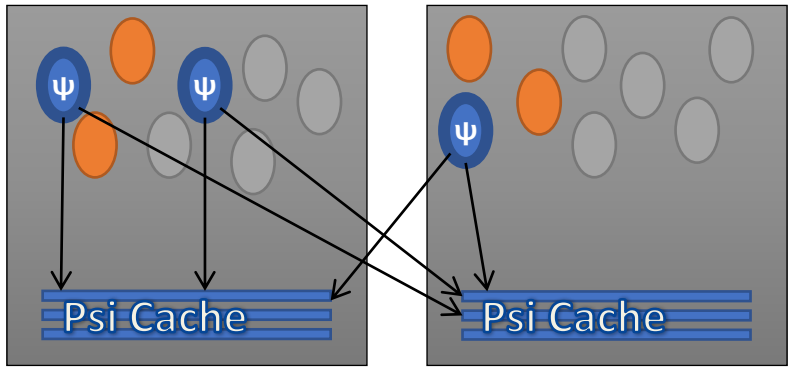
Parallel decomposition:





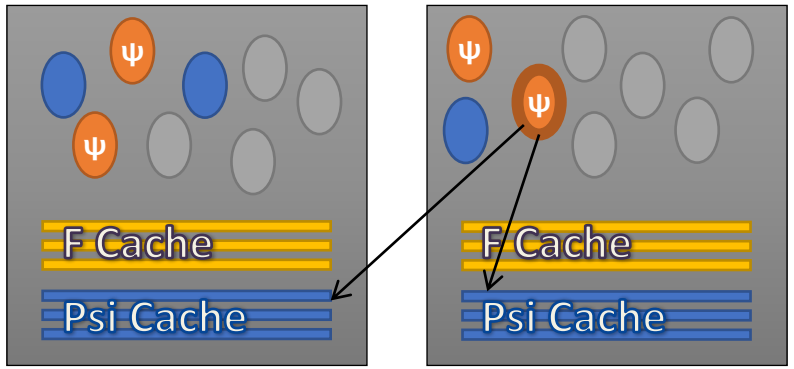
Computing P in Charm++

1. **Duplicate occupied states on each node**



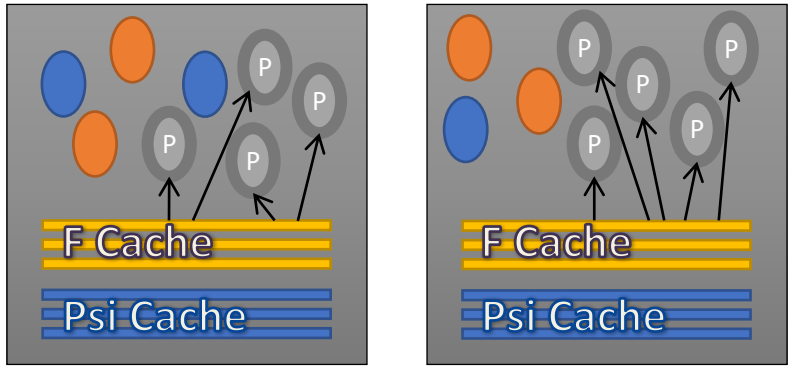
Computing P in Charm++

1. Duplicate occupied states on each node
2. **Broadcast an unoccupied state to compute f vectors**



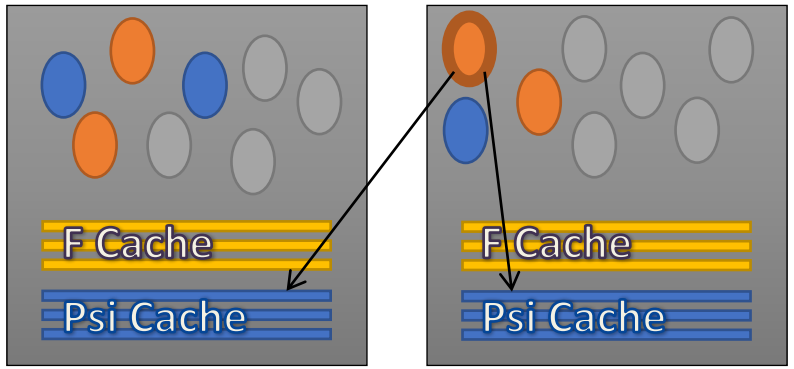
Computing P in Charm++

1. Duplicate occupied states on each node
2. Broadcast an unoccupied state to compute f vectors
3. **Locally update each matrix tile**



Computing P in Charm++

1. Duplicate occupied states on each node
2. Broadcast an unoccupied state to compute f vectors
3. Locally update each matrix tile
4. **Repeat step 2 for next unoccupied state**

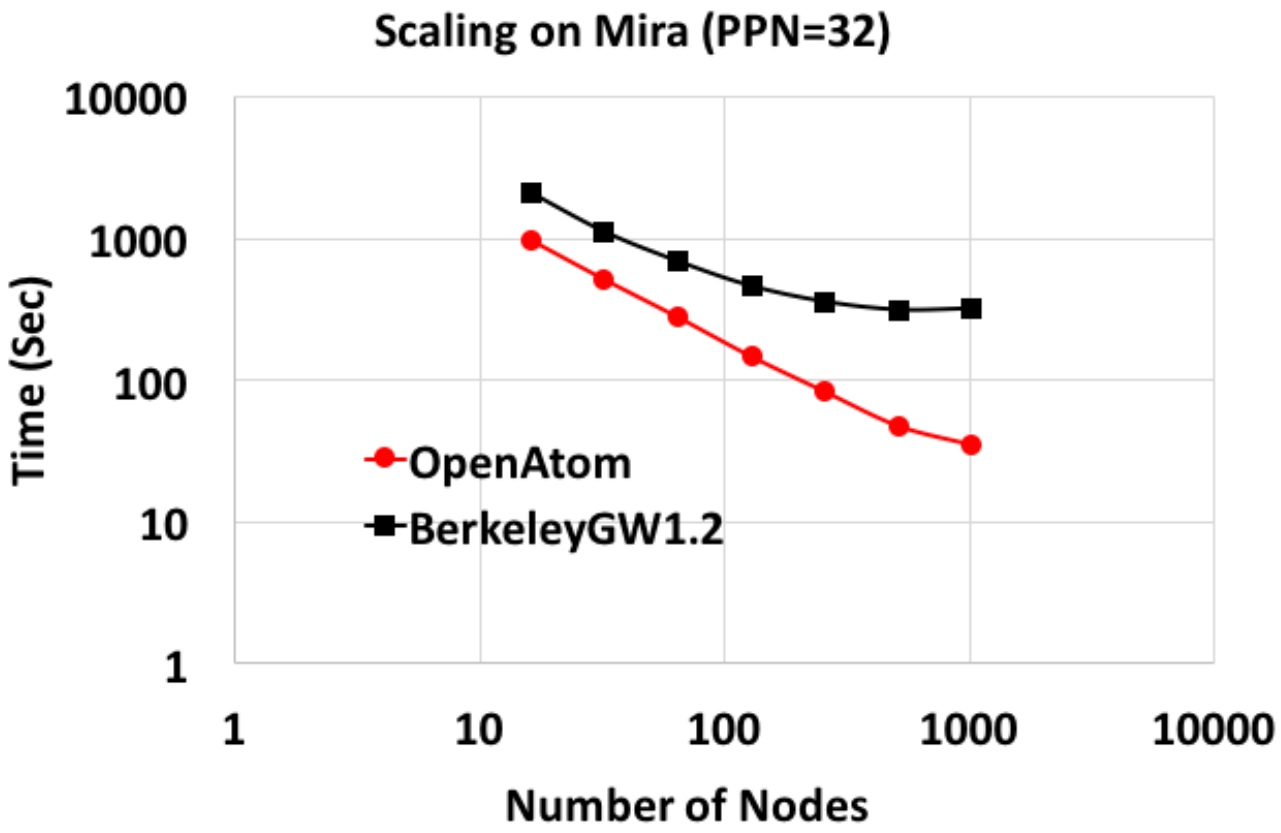


Parallel performance: P calculation

- 108 atom bulk Si
- 216 occupied
- 1832 unoccupied
- 1 k point
- **32** processors per node
- FFT grids: same accuracy

  OA 42x42x22

  BGW 111x55x55



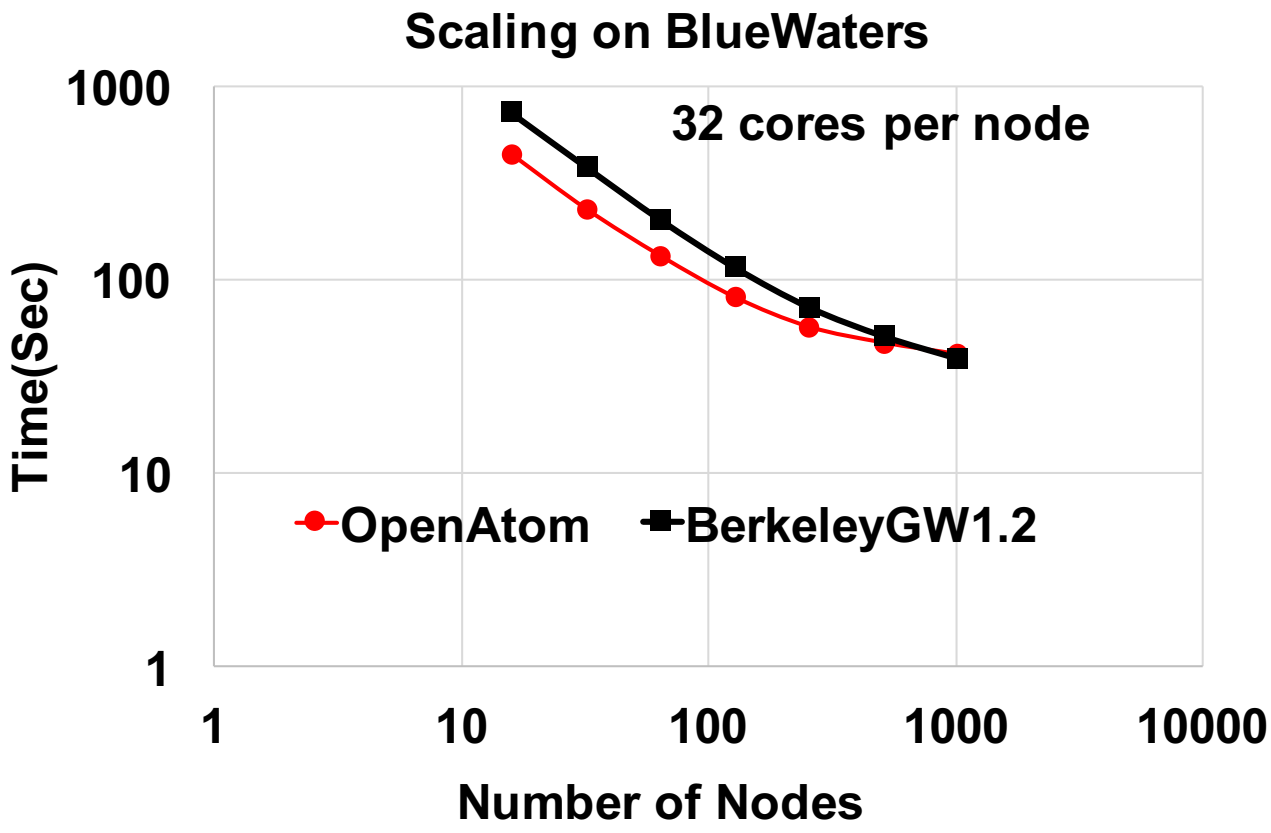
Supercomputer : Mira (ANL) : BQ BlueGene/Q

Parallel performance: P calculation

- 108 atom bulk Si
- 216 occupied
- 1832 unoccupied
- 1 k point
- **32** processors per node
- FFT grids: same accuracy

  OA 42x42x22

  BGW 111x55x55



Supercomputer : Blue Waters (NCSA) : Cray XE6

Reducing the scaling: quartic to cubic

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

- O(N^4) = $N_r^2 \times N_v \times N_c$
- Sum-over-state (i.e., sum over unoccupied c band) not to blame: removal of unocc. states still O(N^4) but lower prefactor*

- Working in r-space can reduce to O(N^3) [see also †]

* Bruneval and Gonze, PRB **78** (2008); Berger, Reining, Sottile, *PRB* **82** (2010)
* Umari, Stenuit, Baroni, *PRB* **81**, (2010)
* Giustino, Cohen, Louie, *PRB* **81**, (2010)
* Wilson, Gygi, Galli, *PRB* **78**, (2008); Govoni, Galli, *J. Chem. Th. Comp.*, **11** (2015)
* Gao, Xia, Gao, Zhang, *Sci. Rep.* **6** (2016)

† Foerster, Koval, Sanchez-Portal, *JCP* **135** (2011)
† Liu, Kaltak, Klimes and Kresse, *PRB* **94**, (2016)

What's special about r-space?

Quasi-philosophical: all basis good in quantum mechanics, why is r-space special?

Observable is diagonal in the best basis $P(r,r') = \frac{\partial n(r)}{\partial V(r')} \quad n(r) = \sum_v |\psi_v(r)|^2$

Practical: P is separable in r-space

$$P(r,r') = -\sum_{v,c} \psi_c(r)^* \psi_v(r) \psi_c(r') \psi_v(r)^* \frac{2}{\epsilon_v - \epsilon_c} \qquad N_r^2 N_c N_v \quad \propto \quad N^4$$

$$\frac{1}{\epsilon_c - \epsilon_v} = \int_0^\infty dx\, e^{-(\epsilon_c - \epsilon_v)x}$$

$$P(r,r') = -2 \int_0^\infty dx \sum_c \psi_c^*(r) \psi_c(r') e^{-\epsilon_c x} \sum_v \psi_v(r) \psi_v^*(r') e^{\epsilon_v x}$$

separable

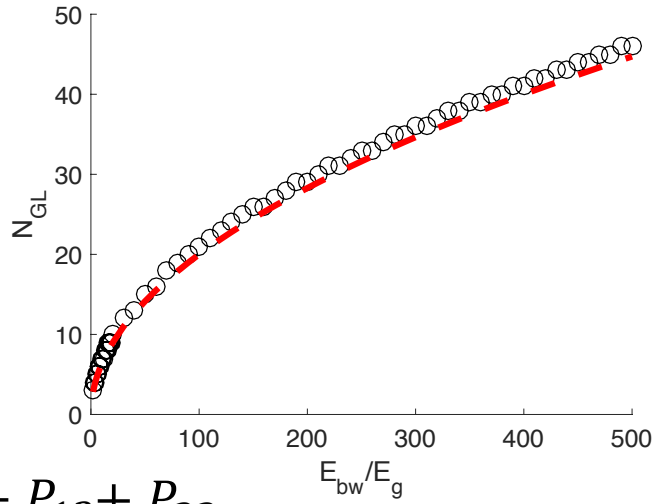
Gauss-Laguerre quadrature: $\int_0^\infty f(z) e^{-z}\, dx \approx \sum_k^{N_L} \omega_k f(z_k)$

$$P(r,r') = -2 \sum_k^{N_L} \omega_k e^{x_k} \sum_c \psi_c^*(r) \psi_c(r') e^{-\epsilon_c x_k} \sum_v \psi_v(r) \psi_v^*(r') e^{\epsilon_v x_k}$$

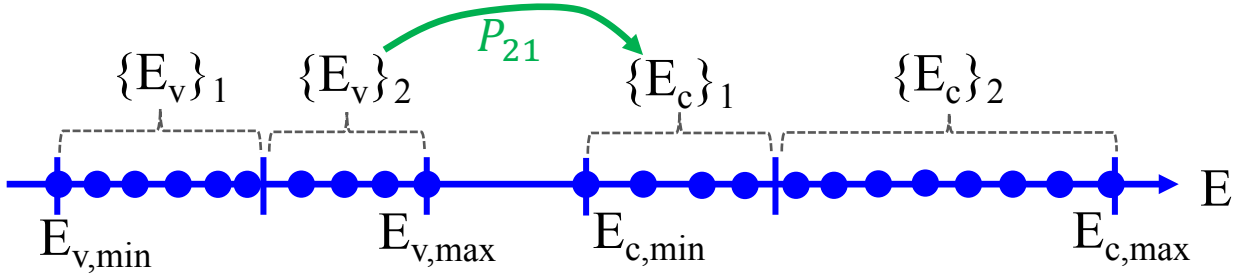
N_L is intensive $\quad N_r^2 N_L (N_c + N_v) \quad \propto \quad N^3$

Windowed cubic Laplace method

- N_{GL} depends on $\frac{E_{bw}}{E_{gap}}$ $E_{bw} = E_{cmax} - E_{vmin}$
- Largest error: $E_c - E_v = E_g$ or E_{bw}



- Example: 2 by 2 windows $P = P_{11} + P_{21} + P_{12} + P_{22}$



$$P(r,r') = \sum_{l}^{N_{wv}} \sum_{m}^{N_{wc}} P_{lm}(r,r')$$

N_{wv} : # windows for E_v

N_{wc} : # of windows for E_c

- Save computation: small N_{GL} for each window pair
- Especially for materials with small band gaps

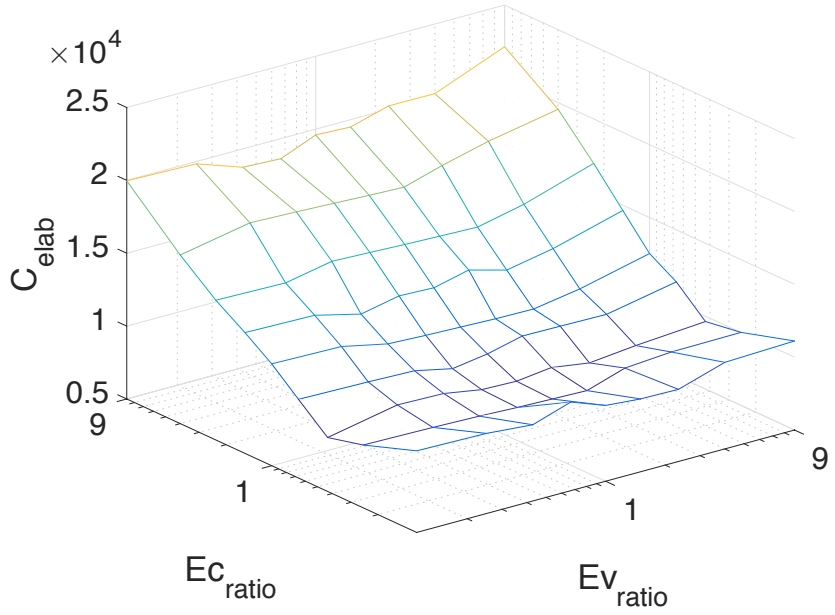
Estimate the computational costs

Computation cost can be estimated with E_{bw} and E_g:

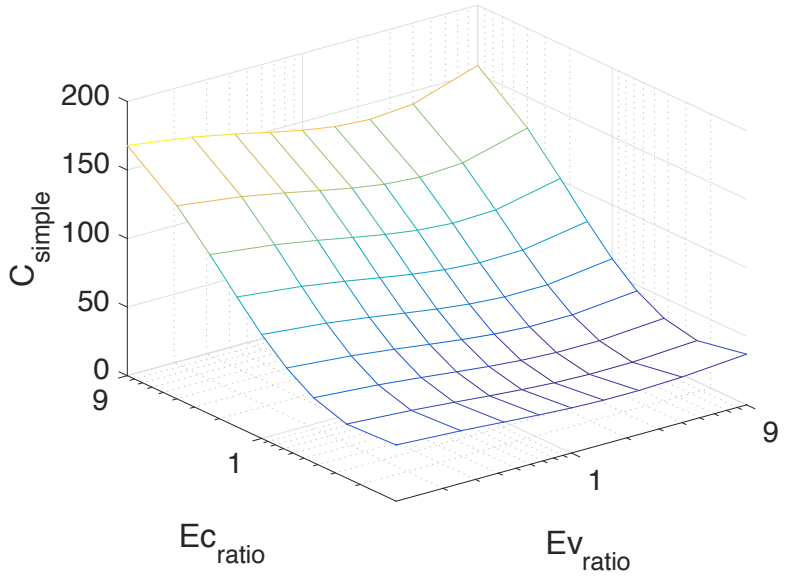
$$C \propto \sum_{l}^{N_{vw}} \sum_{m}^{N_{cw}} \sqrt{\frac{E_{bw}^{lm}}{E_g^{lm}}} \left(\frac{E_{vl}^{max} - E_{vl}^{min}}{E_v^{max} - E_v^{min}} N_v - \frac{E_{cm}^{max} - E_{cm}^{min}}{E_c^{max} - E_c^{min}} N_c \right)$$

Example: 2x2 window

Real computational costs



Estimated computational costs

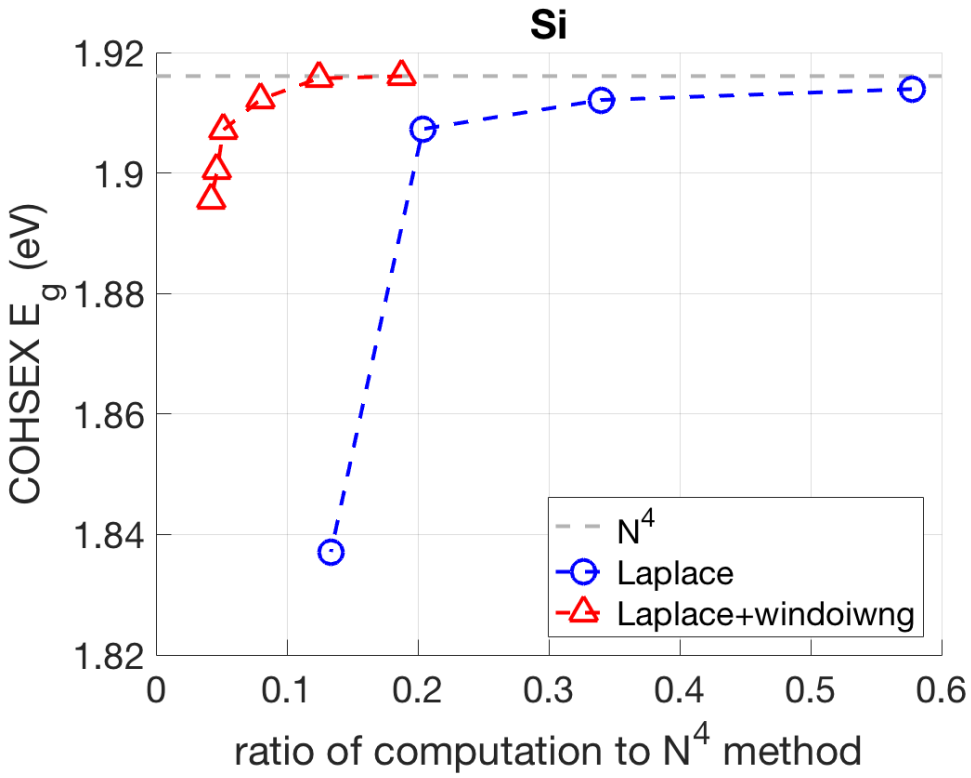


$$E_{v,ratio} = \frac{E_v^* - E_{v,min}}{E_{v,max} - E_v^*}$$

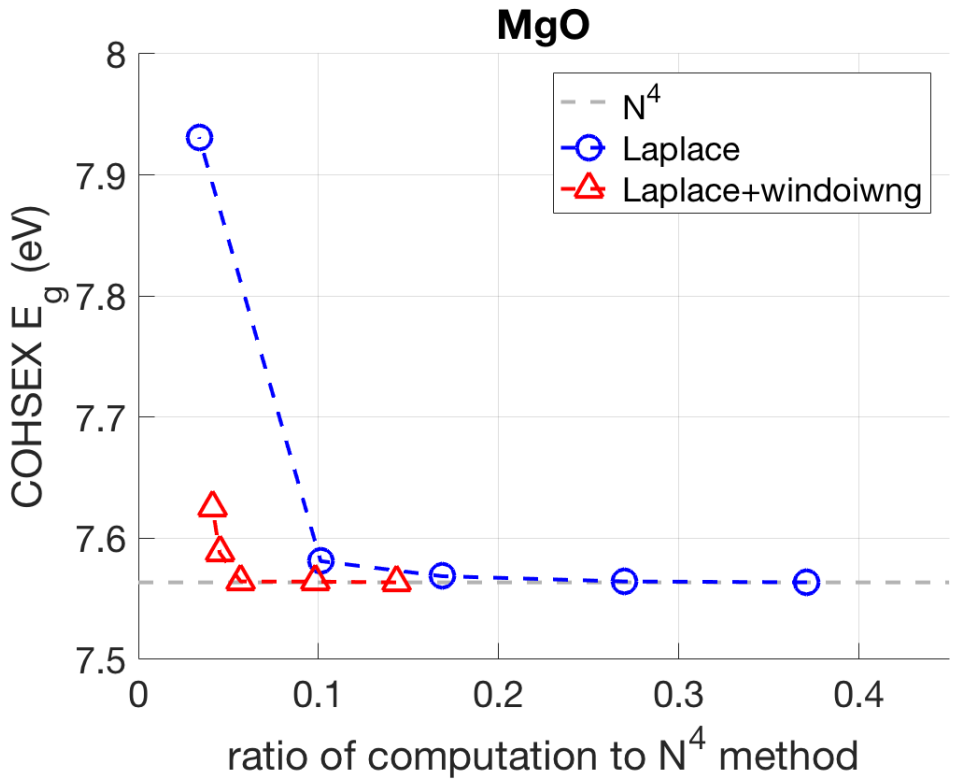
$$E_{c,ratio} = \frac{E_c^* - E_{c,min}}{E_{c,max} - E_c^*}$$

Windowed Laplace: example

- Si crystal (16 atoms)
- Number of bands: 399
- N_{wv}=1, N_{wc}=4



- MgO crystal (16 atoms)
- Number of bands: 433
- N_{wv}=1, N_{wc}=4

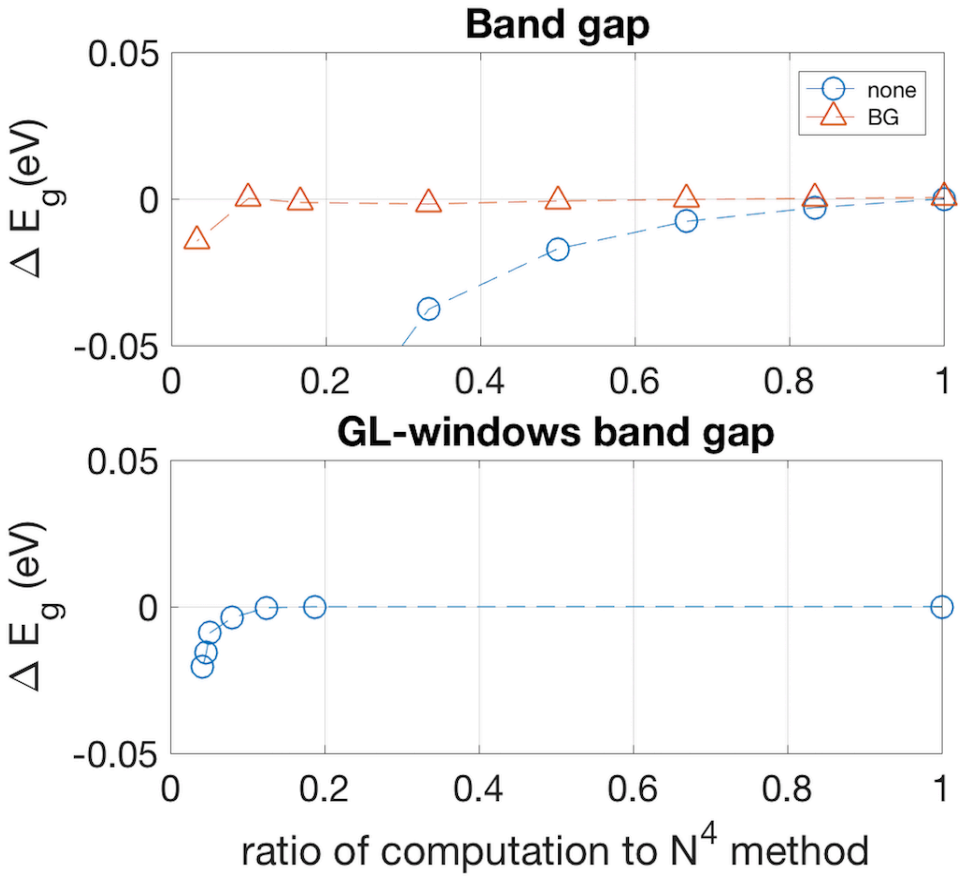


Compared to O(N^4) method, for bigger system ratio is $\frac{Above\ ratio}{N_{at}/16}$

Do I care in practice?

Correct practical comparison:

- Our N^3 method vs. available N^4 method with acceleration
- Crossover is at very few atoms: N^3 method already competitive for small systems



- 2 atoms Si , 8 k-points
- Yambo N^4 GW software
- BG* acceleration

* Bruneval & Gonze, *PRB* **78** (2008)

Windowed Laplace method for self-energy

Dynamic GW self-energy:

$$\Sigma(\omega)^{dyn}_{r,r'} = \sum_{p,n} \frac{B^p_{r,r'}\psi_{rn}\psi^*_{r'n}}{\omega - \epsilon_n + sgn(\mu - \epsilon_n)\omega_p}$$

$B^p_{r,r'}$: residues

ω_p: energies of the poles of $W(r)_{r,r'}$

$$= \sum_{p,n} B^p_{r,r'}\psi_{rn}\psi^*_{r'n}F(\omega - \epsilon_n \pm \omega_p)$$

$$F(x) = \frac{1}{x}$$

$$\frac{1}{\omega - \epsilon_n \pm \omega_p} > 0 \quad \text{OR} \quad \frac{1}{\omega - \epsilon_n \pm \omega_p} < 0 \rightarrow$$

Gauss-Laguerre quadrature not appropriate

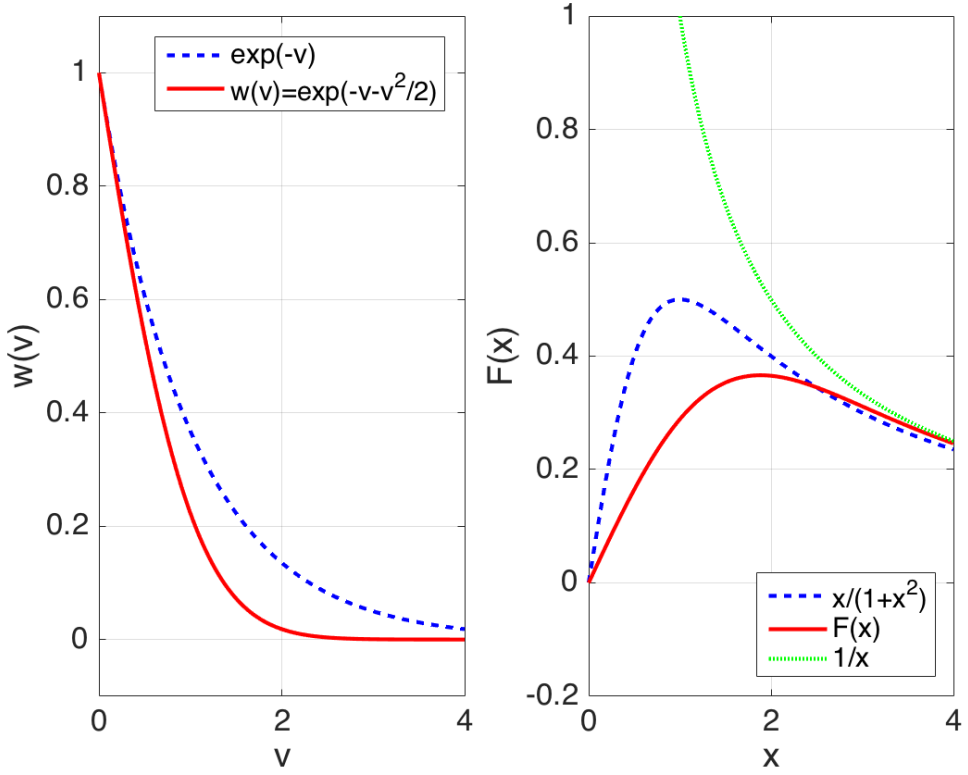
$$\Sigma(\omega) = \sum_{l}^{N_{pw}} \sum_{m}^{N_{nw}} \Sigma(\omega)^{lm}$$

$$e_m^{min} \leq \omega - \epsilon_n < e_m^{max}$$

$$\Omega_l^{min} \leq \pm\omega_p < \Omega_l^{max}$$

New quadrature for overlapping windows

New quadrature



$$F(x) = Im \int_0^\infty w(v) e^{ivx} dv$$

$w(v) = e^{-v}$

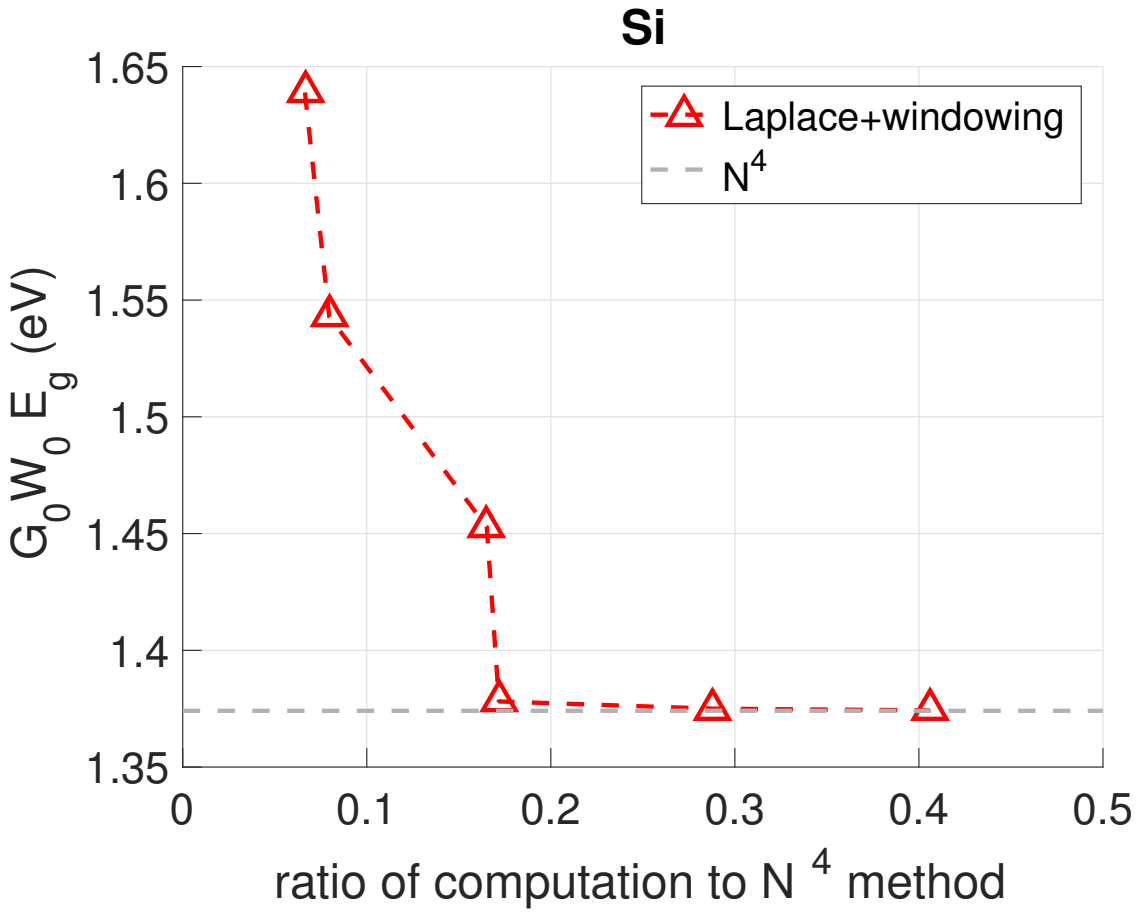
$w(v) = e^{-v-v^2/2}$

Size of quadrature grid

% error	n_q (e^{-v})	n_q $(e^{-v-v^2/2})$
5	6	1
1	24	1
0.1	124	5
0.01	547	15
0.001	2216	36

Results - G_0W_0 gap

- Si crystal (16 atoms)
- Number of bands: 399
- N_{pw}=15, N_{nw}=30



Where we are with OpenAtom GW

Phase		Serial	Parallel
1	Compute P in RSpace	Complete	Complete
2	FFT P to GSpace	Complete	Complete
3	Invert epsilon	Complete	Complete
4	Plasmon pole	Complete	In Progress
5	COHSEX self-energy	Complete	Complete
6	Dynamic self-energy	Complete	In Progress
7	Coulomb Truncation	Future	Future

Aim to release parallel COHSEX version late spring 2018

Summary

- OpenAtom framework
- r-space has many advantages for GW
- Charm++ run time library
  - Reduces parallelization/porting/refactoring headaches
  - Good performance, very good scaling
- r-space separability leads to N^3 scaling GW
  - Straightforward change to sum-over-states methods
  - Crossover with N^4 for N_{atoms} ~ 5-10

Back up slides

G vs. R space P calculation

G-space:

$$P(G,G') = -\sum_{v,c} \langle c|e^{-iG\cdot r}|v\rangle \langle c|e^{-iG'\cdot r}|v\rangle^* \frac{2}{\epsilon_v - \epsilon_v}$$

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

- Directly compute *P* in *G* space
- Many FFTs : $N_v N_c$
- Big multiply: $N_v N_c N_G^2$ = O(N^4)

  N_v : # occupied states
  N_c : # unoccupied states
  N_G : # of G vectors

- $N_v N_c$ FFTs needed
- Big O(N^4) matrix multiply

R-space:

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

Big multiply: $N_v N_c N_r^2$ = O(N^4)

$P(r,r')$

↓ FFT N_r rows

$P(G,r')$

↓ FFT N_r columns

$P(G,G')$

N_r: # r grid
$N_r \approx 4N_c$

- $N_v+N_c+8N_c$ FFTs needed
- Big O(N^4) matrix multiply

"Physicist" programming

$$P(r,r') = \frac{\partial n(r)}{\partial V(r')} = -2 \sum_{v}^{\text{filled}} \sum_{c}^{\text{empty}} \frac{\psi_v(r)\psi_c(r)\psi_v(r')\psi_c(r')}{\varepsilon_v - \varepsilon_c}$$

Consider two key steps

a. Many FFTs → $\psi_i(r)$
b. Outer product → P

Typical MPI / OpenMP: working <u>explicitly</u> with # of processors

1. Divide $\psi_i(r)$ among procs
2. Do pile of FFTs on each proc
3. Divide (r,r') among procs (e.g. ScaLAPACK)
4. Do outer product

Problems

- $N_i > N_{proc}$ and $N_i < N_{proc}$ need different parallelizations:

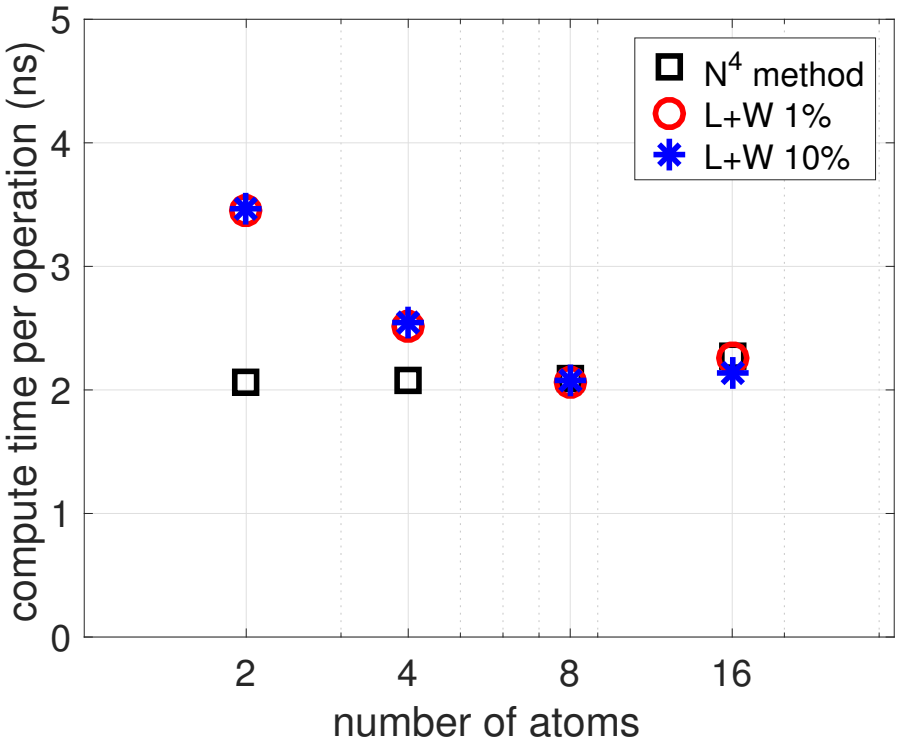
  explicitly different coding

- Typical programmer does 1. & 2. *then* 3. & 4. ; hard to interleave
- Machines/fashion change: need to recode parallelization…

  (GPUs, SMPs, few cores, multicores, etc.)

Where is crossover in scaling?

Si 16 atom calculation



- Number of computations
  N^4: $N_v N_c N_r^2$
  L+W: $\sum_{lm} N_{GL}^{lm} (N_c^m + N_v^l) N_r^2$
- Comparable prefactor
- Speedup for small $N_{atoms} \gtrsim 10$