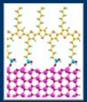
Scalable, Extensible, and Open Framework for Ground and Excited State Properties of Complex Systems



L. Kale[1], S. Ismail-Beigi[2], and G.J. Martyna[3]

(1) Department of Computer Science, University of Illinois (2) Department of Applied Physics, Yale University (3) Physical Sciences, IBM TJ Watson Research Center

The need for fast excited state methods

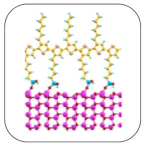
Introduction:

Electronic excitations can play a significant role in designing novel materials. Although density functional theory (DFT) describes ground state properties accurately, it can fail to provide the right description of some basic properties such as the band gap of a material, the relative alignment of energy bands between two materials, or the optical spectrum of a material. In terms of *ab initio* methods, the GW-BSE approach in many-body perturbation theory has proved to improve the description of excitations.

To date, GW-BSE has been applied mainly to bulk materials. The main reason is the computational expense which limits most simulations to systems with tens of atoms in the simulation cell.

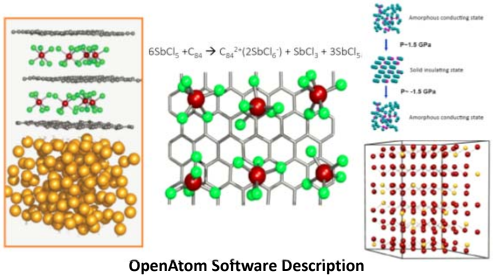
However, there is a great need to apply these methods to describe complex materials including hybrid organic/inorganic photovoltaic systems, metal-organic framework storage materials, surface-molecule systems for electron transport, etc.

The figure shows a potential photovoltaic system where the energy band alignment between two material is critical, but present GW approaches are too costly to be used even on modern computers without new methods and software.



1.

Fast sampling, treatment of nuclear quantum effects on the BO surface


6SbCl$_5$ +C$_{84}$ → C$_{84}$$^{2+}$(2SbCl$_6$$^-$) + SbCl$_3$ + 3SbCl$_5$
Amorphous conducting state
P~1.5 GPa
Solid insulating state
P~ -1.5 GPa
Amorphous conducting state

OpenAtom Software Description

- Suitable for light elements – adding PAWS as part of proposal to treat transition metals.
- Reaches long time scales through new methods and fine grained parallelism.
- E-structure – plane-wave based Density Functional theory (GGA/LDA – hybrids/vdW to do).
- Capabilities to be used to impact Science and Technology (see apps above and slide 8).
  - Path Integrals for nuclear quantum effects.
  - k-point sampling.
  - CPAIMD / BOMD (improvements part of this work).
  - LDA / LSDA / GGA.
  - Parallel tempering with BOMD (part of this work).

2.

Parallelism of OpenAtom + GW under charm++ : General Concerns

Object-based parallel programming

- Over decomposition of work.
- Asynchrony.
- Message-driven.

Empowers introspective runtime system

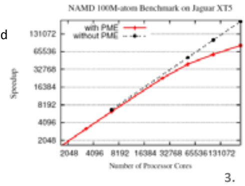
- Load balancing.
- Fault tolerance.
- Topology awareness.

Programmer productivity

- Work and data natural to algorithm.
- Parallelism and Science well encapsulated experts can work in their own space.

Many scalable science applications

- NAMD
- EpiSimdemics
- ChanGA

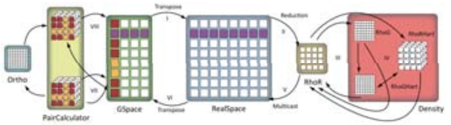


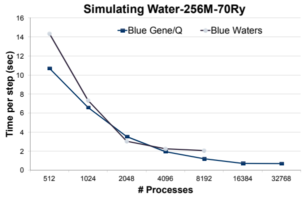
3.

OpenAtom - Fine Grained Parallel Software for E-Structure

Close coupling of science and programming yields outstanding performance:

- Eight phases drive by data and work
- Objects representing physical entities and tasks: electronic states, density, energy computation, ortho-normalization
- Numerous concurrent FFTs, multicasts, and reductions





4.

Overcoming Bottlenecks in OpenAtom Scaling

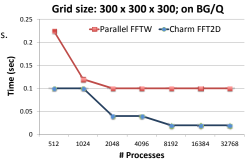
I. Fine-grained FFT using 2-D thick pencil decomposition – implementing in OpenAtom

I.A) 1-D decomposition:

1. 1 transpose, O(N) parallelism
2. Divide 3D-grid (X,Y,Z) into XY-planes along Z.
3. Perform X and Y FFT.
4. Transpose the XY-planes to construct YZ-planes.
5. Perform the Z FFT.

I.B) 2-D decomposition:

1. 2 transposes, O(N^2) parallelism.
2. Divide 3D-grid (X,Y,Z) into X-lines in XY planes.
3. Perform X FFT.
4. Transpose to construct Y-lines in XY planes.
5. Perform the Y FFT.
6. Transpose to construct Z-lines in YZ-planes.
7. Perform the Z FFT.



II. Use of GPUs to speed up matrix related computation:

1. GEMM operations in orthonormalization - bottleneck for large systems.
2. GPUs/MICs provides more compute capability but requires a driver CPU.
3. Charm++ RTS on driver CPU distributes work to the CPUs and the GPUs/MICs.
4. OpenAtom will requests the RTS to offload GEMMs to GPUs/MICs.
5. Reduction in the execution time ≈ 50% for orthonormalization including GPU load.
6. Future GPUs and MICs that share memory to avoid load phase are promising.

5.

GW – Why is it computationally expensive

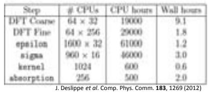
Why is present GW-BSE so costly?

- Huge number of FFTs to get wave functions
- Dense FFT grids require have huge memory footprints
- Large and dense matrix multiplications

Theoretical scaling

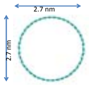
DFT	N^3
GW	N^4
BSE	N^6

Practical example for (20,20) SWCNT:



Step	# CPUs	CPU hours	Wall hours
DFT Coarse	64 × 32	19000	9.1
DFT Fine	64 × 256	29000	1.8
epsilon	1600 × 32	61000	1.2
sigma	960 × 16	46000	3.0
kernel	1024	600	0.6
absorption	256	500	2.0

J. Deslippe *et al*. Comp. Phys. Comm. **183**, 1269 (2012)



Observations:

- BSE has the worst theoretical scaling
- For actual nanosystems, GW is the main bottleneck
- Accelerating GW is thus the primary goal at present

Tasks:

- Developing new algorithmic advances to reduce FFTs
- Deploying efficient parallel FFTs and linear algebra
- Effective memory parallelization

6.

GW in OpenAtom

Strategy for RPA static polarizability calculations: G-space vs. R-space

The most time consuming part of GW calculations is obtaining the polarizability *P* which encodes the response of the electron distribution to potential changes. This is a dense matrix in G space $P_{GG'}$.

Standard G-space approach:

Directly compute *P* in G space. This requires a huge number of FFTs.

$$P_{GG'} = \sum_{v,c} \underbrace{\langle c|e^{-iG\bullet r}|v\rangle}_{FFT^{-1}[\psi_c^*(r)\psi_v(r)]}\langle v|e^{iG'\bullet r}|c\rangle \frac{2}{\varepsilon_v-\varepsilon_c}$$

Number of FFTs ≈ 2 x nv × nc

v = an occupied electron state
c = an empty electron state
ε_j = energy of electron state
$\psi_j(r)$ = electron state wavefunction

nv = number of occupied electron states
nc = number of unoccupied electron states

New R-space approach:

Compute *P* in real-space $P_{rr'}$ as an intermediate then FFT to G-space $P_{GG'}$.

$$P_{rr'} = \sum_{v,c}\psi_c^*(r)\psi_v(r')\psi_v^*(r)\psi_c(r')\frac{2}{\varepsilon_v-\varepsilon_c} \xrightarrow[FFT^{-1}]{} P_{GG'}$$

Number of FFTs ≈ nv + nc + 8 × nc

For large systems where nv is at least many hundreds, R-space much more efficient

Example:
nv = 500, nc = 1500
Number of FFTs: **14,000** (R-space) vs. **1,500,000** (G-space)

Progress and Future plans

- R-space polarizability approach and dielectric matrix computation developed and tested.
- Near complete implementation of Coulomb truncation for confined nanosystems.
- Developing algorithms to reduce large number of empty states (nc) in GW.
- Creating an interface between Charm++ and GW data structure.
- Once GW is under control, implement the BSE for optical excitation calculations.

7.

OpenAtom Demonstrator System – Petascale Challenge

CPAIMD investigation of hydrogen storage in Metal-Organic-Frameworks (MOF)

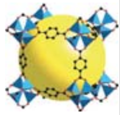
Introduction: A goal of science and technology research is to deliver a source(s) of clean, renewable energy. The hydrogen economy is ideal as hydrogen combustion produces water and hydrogen is the 3rd most abundant on earth. Challenges to realizing the hydrogen powered world include extracting hydrogen from its bound forms (.e.g. water, hydrocarbons) to generate the molecular moiety, delivery, ***storage,*** efficient fuel cells among others as depicted to the right.

Necessary elements of the hydrogen economy.

The hydrogen molecule

Effective H_2 storage requires attaining high mass and volumetric density, preferably at ambient pressure and temperature. The ideal hydrogen storage material is thus light weight, and can reversibly and rapidly store molecular hydrogen or novel typically by physisorption.

Materials with large surface areas and low densities, such as metal–organic frameworks (MOFs) pictured to the right, have been found to adsorb hydrogen up to 4.5% mass density at low temperature and at low pressure (20 bar). However, the DOE 2015 targets for a hydrogen storage system have not yet been reached such as a capacity of 40 g H_2 per L. This suggests developing computational methods to aid discovery.



(MOF-5) $Zn_4O(BDC)_3$
(BDC = 1,4 benzenedicarboxylate)

1. atoms: 424.
2. electrons: 1872.
3. g-vectors: > 9 million.
4. path Integral beads: 16-32.
5. Tempering : TBD

Current MOF simulations are performed by fitting an empirical force field to *ab initio* data on fragments (with methods that can handle Van der Waals interactions) and then using Grand Canonical Monte Carlo sampling to determine loading characteristics. Complementary long time sampling fully *ab initio* studies including nuclear quantum effects with OpenAtom would provide additional important scientific and technical insight into strong candidates – new binding modes, for instance, and allow more complex materials choices that might not be simple to model using an empirical potential. Our MOF demonstrator system is pictured to the right.

8.

Euler Exponential Spline based supersoft/PAW implementation

Comparison of pseudopotential forms:

Norm-conserving pseudopotentials (NCPP)

- Pseudo- & all-electron wavefnc. same norm within and match exactly outside.
- Match 1st energy derivative of phase shift at core.
- NCPP requires large plane-wave cutoff.
- KB form more efficient with plane-waves.

Hamman, Schluter, and Chiang (1979); Kleinman & Bylander (1982)

Ultrasoft pseudopotentials (USPP)

- Norm-conservation constraint relaxed, localized atom-centered augmentation charges make up the charge deficit.
- Transferability tricky for systems with strong charge transfer, polarization, …

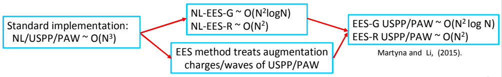
Vanderbilt (1990).

Projector Augmented Waves (PAW) method

- Linear transformation from pseudo- to all electron wavefnc.
- Work directly with all electron wavefnc using extra radial grids.
- Augmentation charges can be extended spatially – comp. cost.
- Easier transferability / construction than USPP.

P. Blochl (1994)

Fast, reduced order methods based on Euler Exponential Splines



Standard implementation:
NL/USPP/PAW ~ O(N^3)

NL-EES-G ~ O(N^{2}logN)
NL-EES-R ~ O(N^2)

EES method treats augmentation charges/waves of USPP/PAW

EES-G USPP/PAW ~ O(N^2 log N)
EES-R USPP/PAW ~ O(N^2)

Martyna and Li, (2015).

Martyna et al, *Chem. Phys. Chem.* (2005).

9.

Acknowledgements

NSF SI2 Award Number: 1339715 (Kale group).
NSF SI2 Award Number: 1339804 (Ismail-Beige group).
IBM Research (Martyna group).

10.