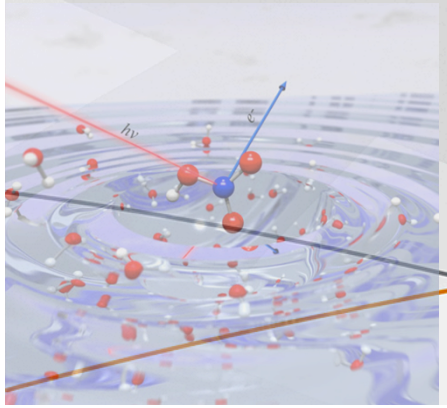
Chemical reactivity, solvation and multi-component heterogeneous processes in aqueous environments

Sotiris S. Xantheas

Physical Sciences Division
Pacific Northwest National Laboratory

US Department of *Energy*
Office of Science, Division of Chemical Sciences, Geosciences & Biosciences)

Mark R Pederson & Gregory J Fiechtner



NERSC BES Requirements for 2017
October 8-9, 2013
Gaithersburg, MD

Pacific Northwest
NATIONAL LABORATORY

Project Description

The program conducts fundamental research on

- use of model systems and unique methods to understand chemical processes on surfaces, in condensed complex media, and at interfaces
- first-principle calculations and advanced methods for modeling and simulation closely coupled with experiments to extend our understanding of chemical reactivity from the molecular scale to collective phenomena in complex systems

Facilitated via

- reliable descriptions of molecular interaction
- consistent chemical physics theories that link molecular interaction and physical observables of interest
- robust simulation techniques for efficient computations within the theoretical framework

Aims at obtaining a fundamental understanding of complex and condensed phase systems based on first-principles approaches

Pacific Northwest
NATIONAL LABORATORY

NERSC Computing Resources

- **DOE Production:**

  Chemical reactivity, solvation and multi-component heterogeneous processes in aqueous environments

  - **mp329: 1,825,000 hrs. (PI: S Kathmann)**

- **DOE Production:**

  Novel density functional theory approaches to the statistical mechanics of aqueous systems

  - **m452: 5,000,000 hrs. (PI: CJ Mundy)**

- **NISE award:**

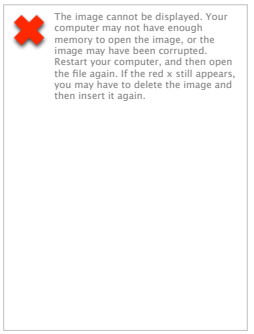
  Guest-host interactions in hydrate lattices: Implications for H_2 storage and CO_2 sequestration

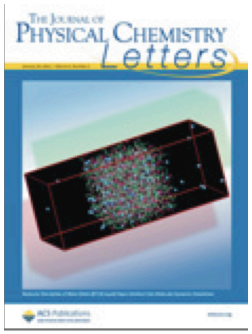
  - **m1513: 5,000,000 hrs. (PI: SS Xantheas)**

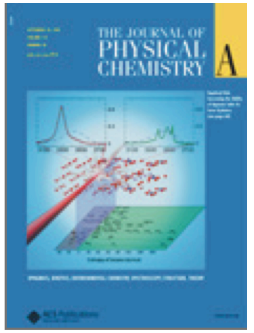
Pacific Northwest
NATIONAL LABORATORY

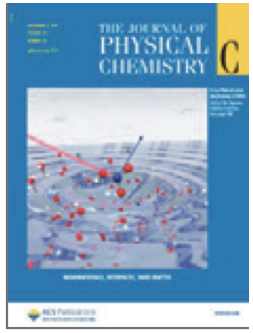
People / Synergism

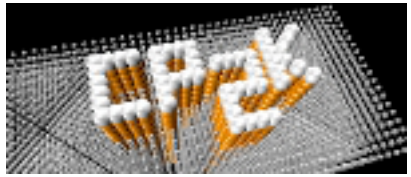
- **Molecular Theory group at PNNL**
  - Liem Dang
  - Shawn Kathmann
  - Christopher Mundy
  - Gregory Schenter (PI)
  - Marat Valiev
  - Sotiris Xantheas
- **NWChem Developers (EMSL, PNNL)**
  - Edoardo Apra
  - Niri Govind
  - Karol Kowalski (Capability Lead)
  - Marat Valiev
- **CP2K Developers (U of Zurich, ETH)**
  - Joost VandeVondele
  - Juerg Hutter







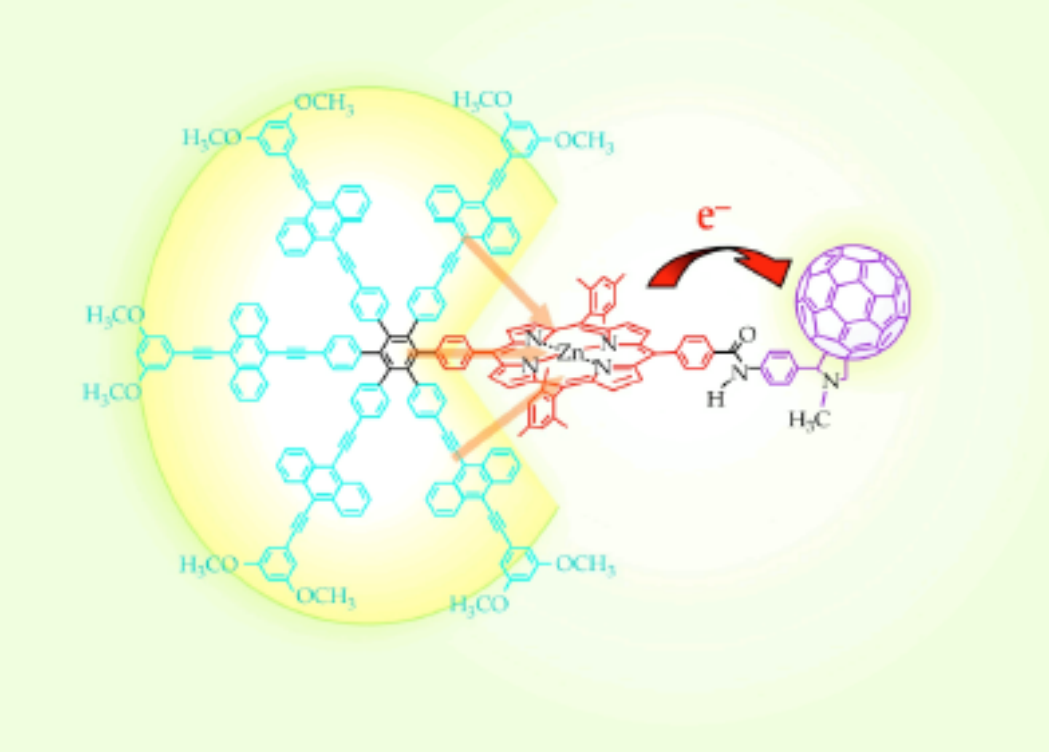




Pacific Northwest
NATIONAL LABORATORY

Some Future Scientific Challenges

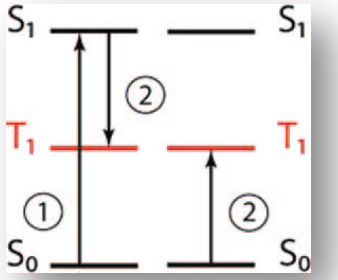
Light-harvesting antennas in dye-sensitized photocells

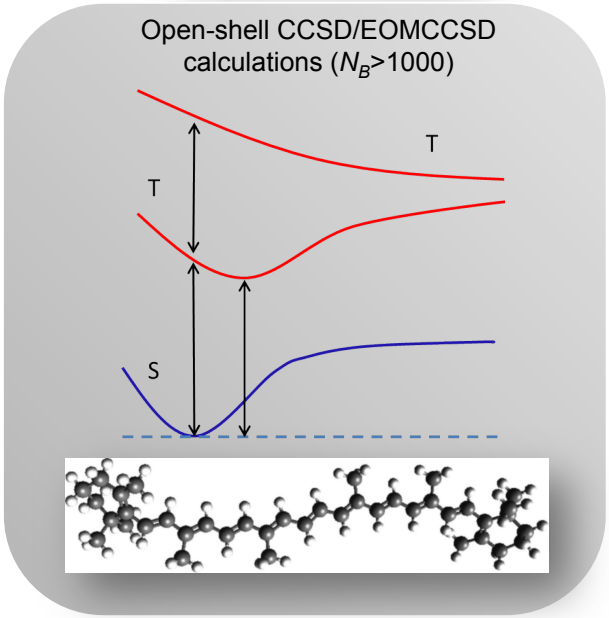


- Energy storage processes: conversion of solar energy
- Interaction of biological systems with radiation
- Localized excited states in solid state physics/chemistry

Singlet fission

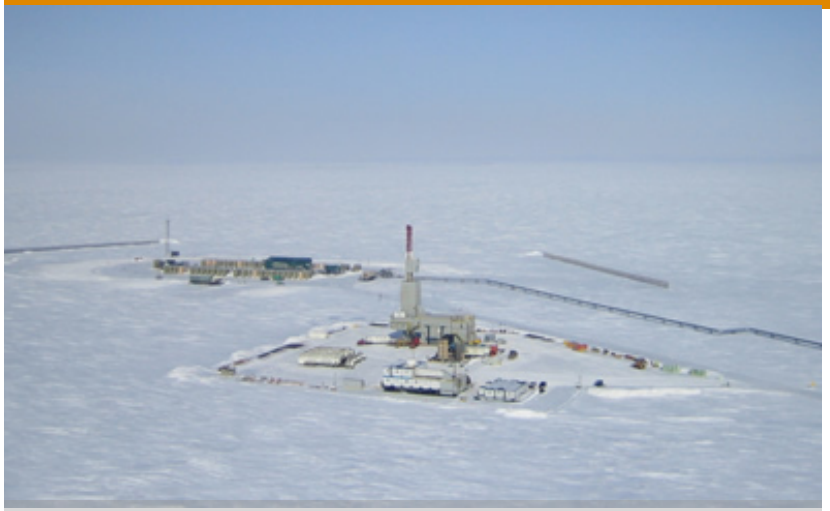
MB Smith, J Michl, *Chem. Rev.* **110**, 6891 (2010)





Pacific Northwest
NATIONAL LABORATORY

Some Future Scientific Challenges (*cont.*)



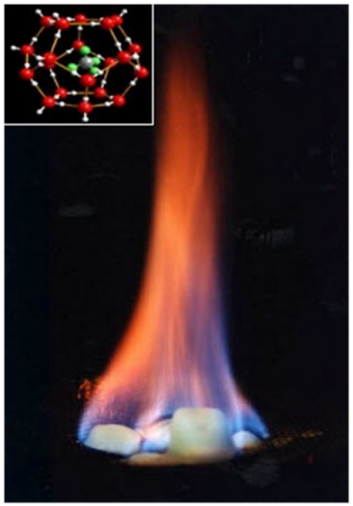
Ignik Sikumi #1 field trial well site, Prudhoe Bay, Alaska

Conoco Phillips for US DoE (NETL)

Field trial for methane (CH_4) hydrate production methodology, whereby carbon dioxide (CO_2) molecules are exchanged in situ for the methane molecules within a hydrate structure, releasing the methane for production.

The objective was to evaluate the viability of this hydrate production technique and to understand the implications of the process at a field scale

Guest/host interactions in complex systems with emphasis in energy applications



Example: Hydrates as natural resources for

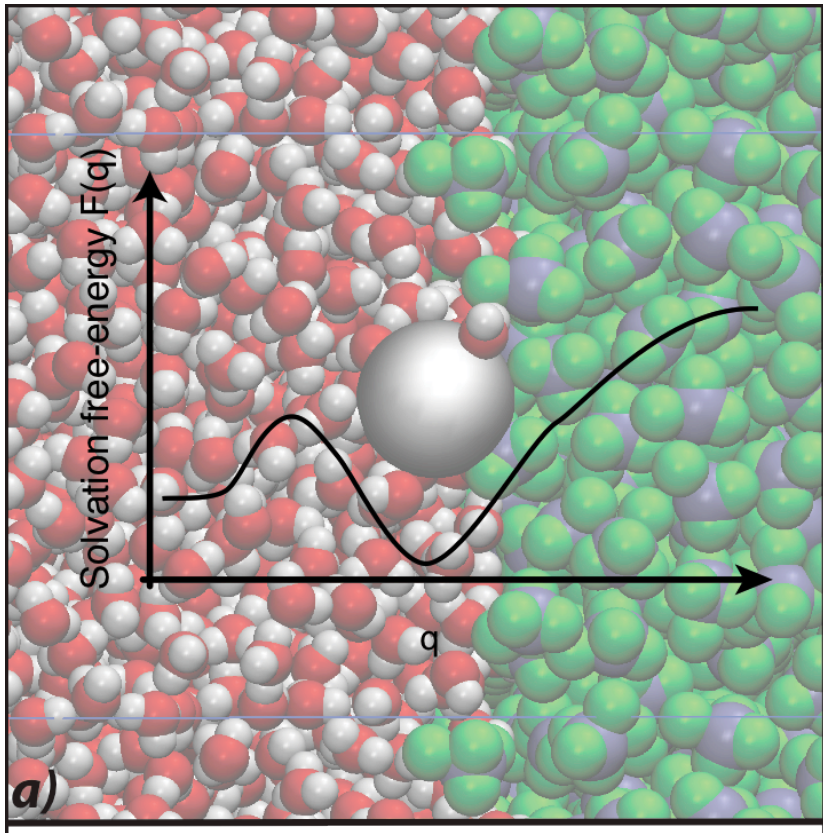
- H_2 storage
- CH_4 production
- CO_2 sequestration

Scientific issues

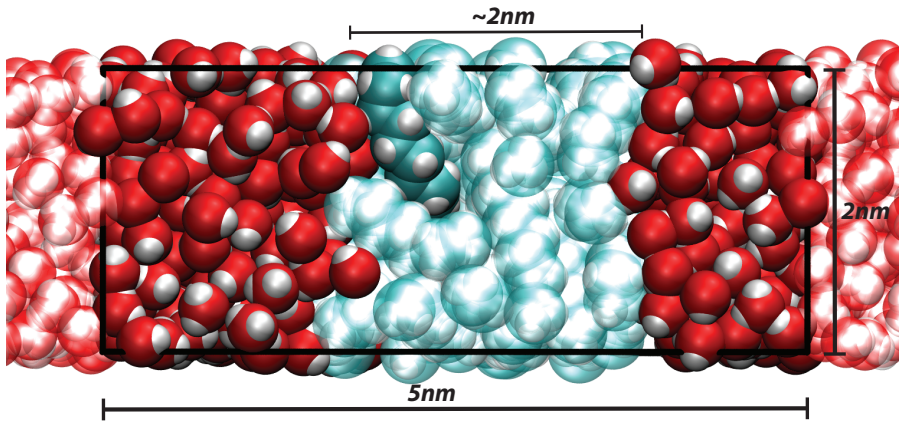
- nature and magnitude of interaction between hollow cages of 3-D lattices and guest molecules
- mechanism for diffusion / exchange of guest molecules between cages
- enhancement of storage capacity

Pacific Northwest
NATIONAL LABORATORY

Some Future Scientific Challenges (*cont.*)



Seeking principles of self-assembly through ion adsorption to water-hydrophobic interfaces



n-dodecane / water systems

- New equilibrium theories of colloid stability can predict experimentally determined flocculation as a function of electrolyte concentration
- Salient microscopic principle: Ions adsorbed on the hydrophobic interface modify the electrostatics between surfaces

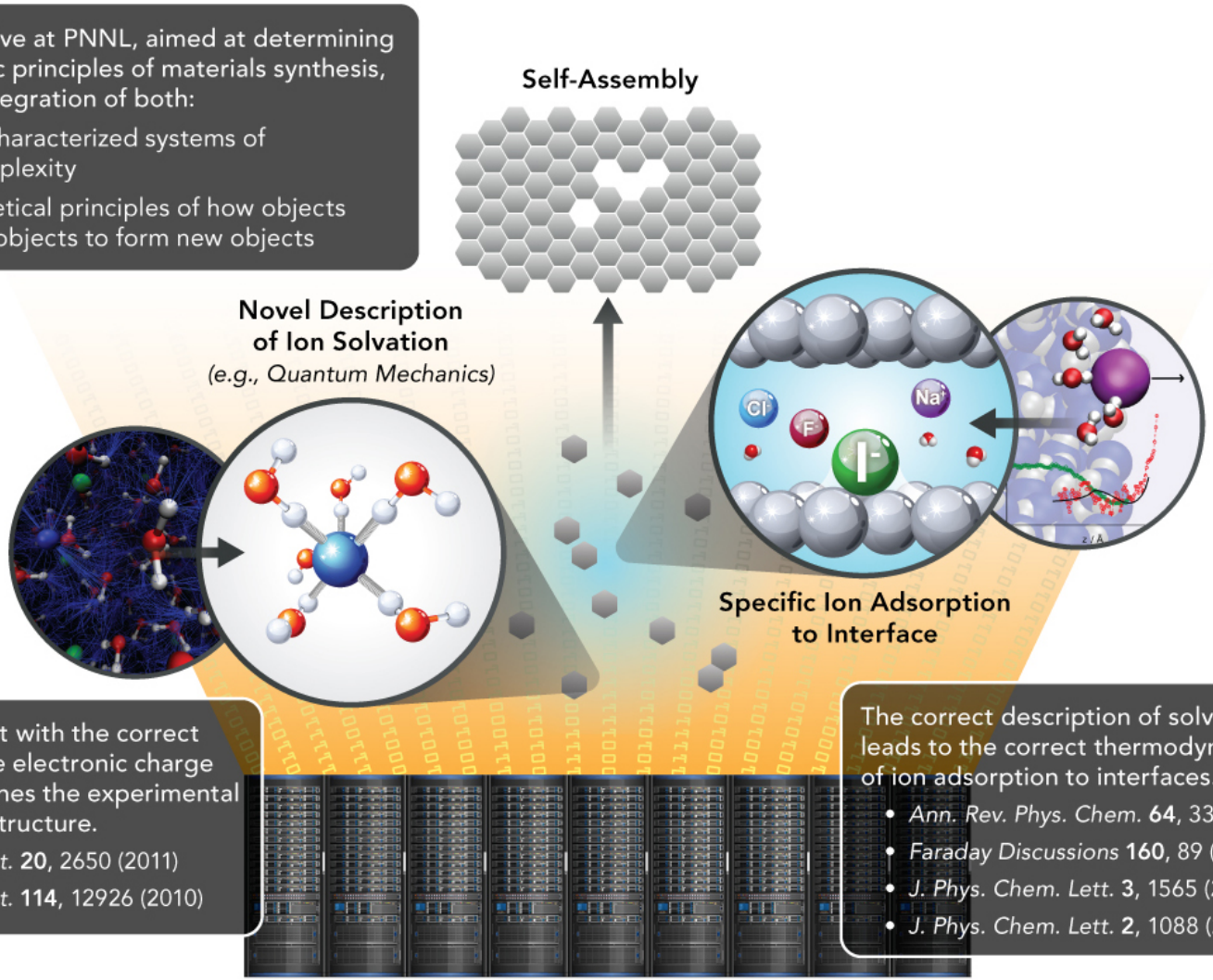
Pacific Northwest
NATIONAL LABORATORY

Investments already in place at PNNL

The foundations of our insights are built on access to petascale computing

The MS^3 Initiative at PNNL, aimed at determining the microscopic principles of materials synthesis, requires the integration of both:

- New well-characterized systems of higher complexity
- New theoretical principles of how objects encounter objects to form new objects



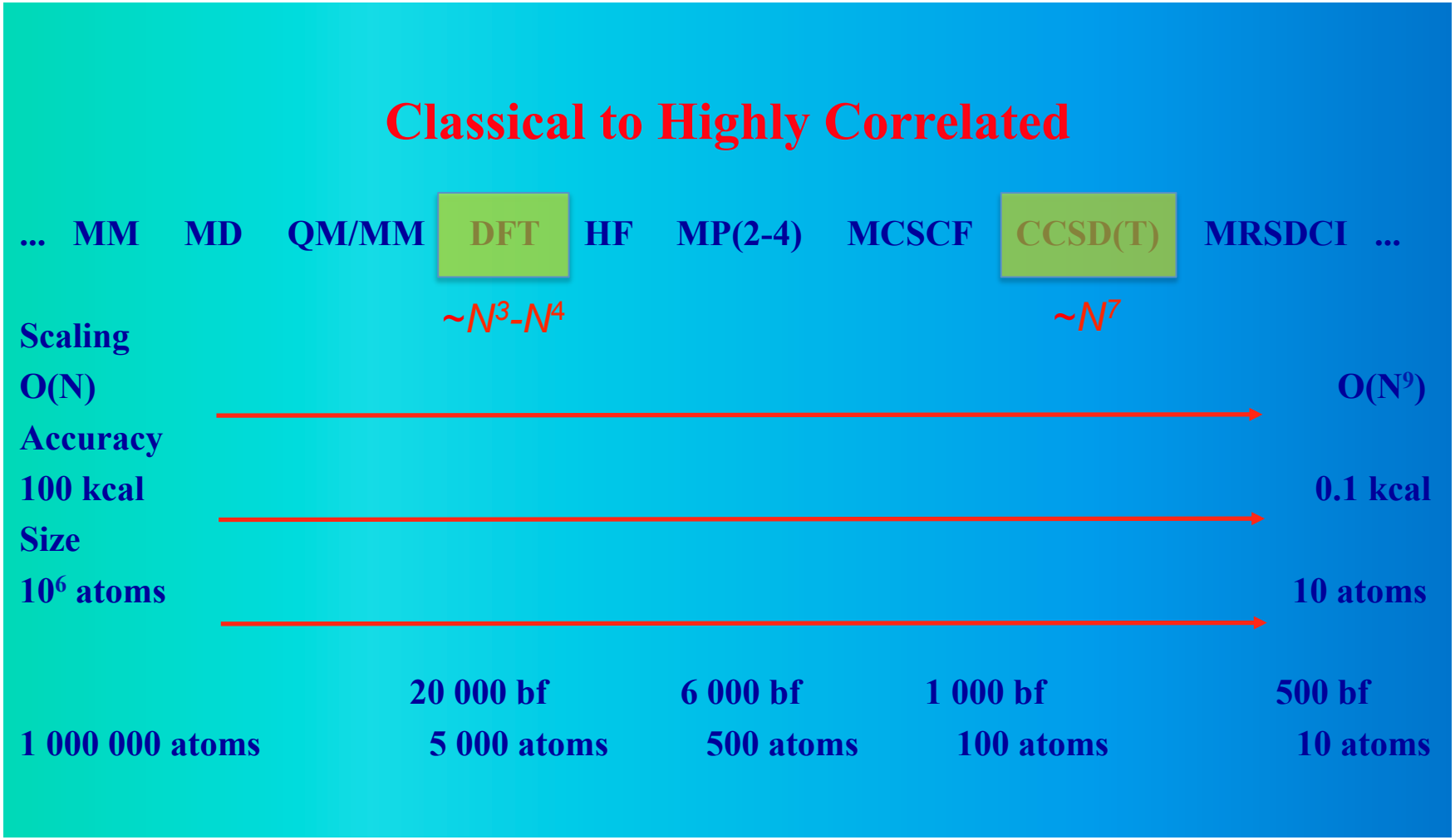
These principles start with the correct representation of the electronic charge density that determines the experimental observed solvation structure.

- *J. Phys. Chem. Lett.* **20**, 2650 (2011)
- *J. Phys. Chem. Lett.* **114**, 12926 (2010)

The correct description of solvation leads to the correct thermodynamics of ion adsorption to interfaces.

- *Ann. Rev. Phys. Chem.* **64**, 339 (2013)
- *Faraday Discussions* **160**, 89 (2013)
- *J. Phys. Chem. Lett.* **3**, 1565 (2012)
- *J. Phys. Chem. Lett.* **2**, 1088 (2011)

Methods for describing intermolecular interactions



Pacific Northwest
NATIONAL LABORATORY

A Path Forward

Electronic structure

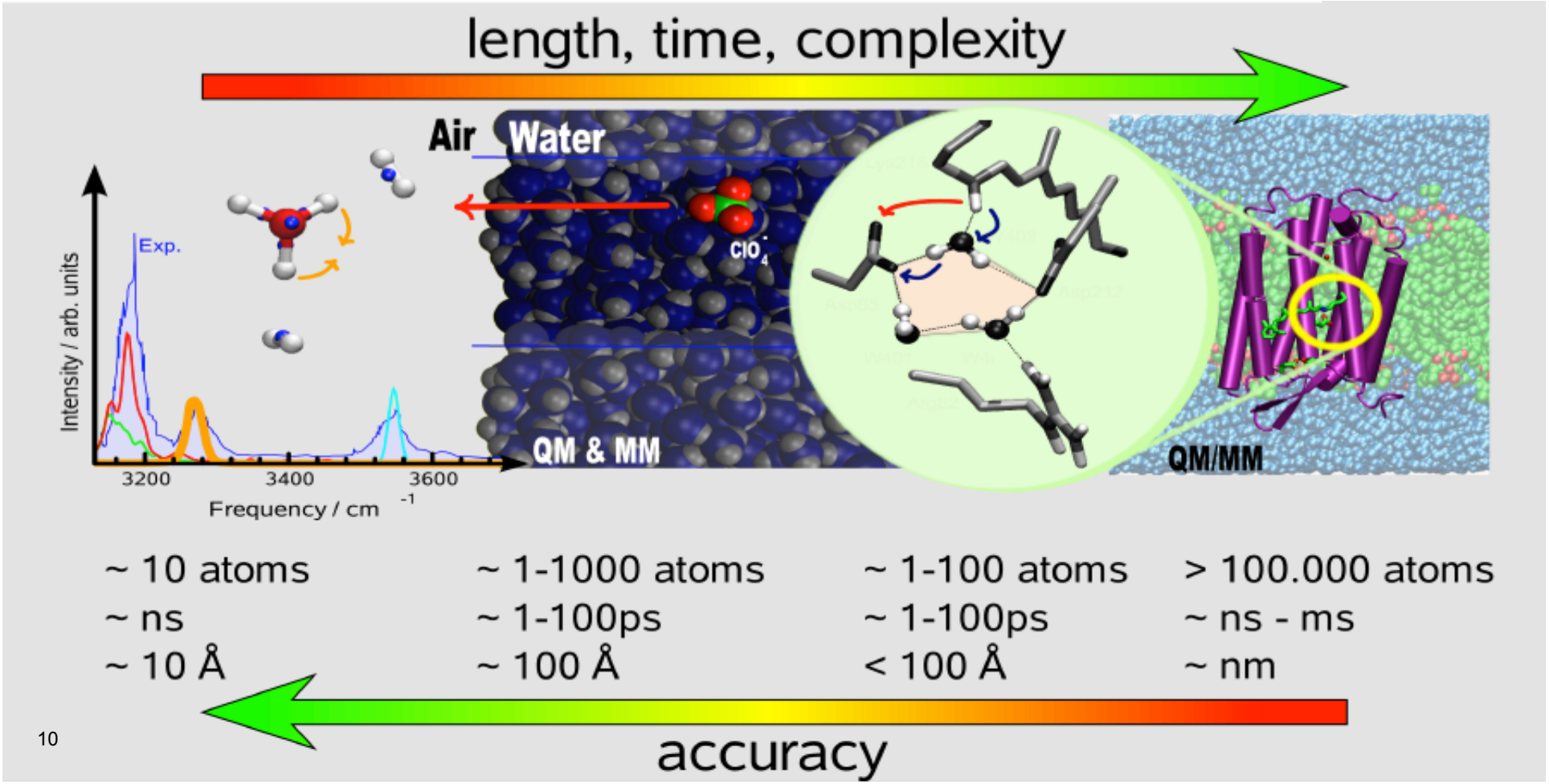
QM/MM

NWChem

Statistical Mechanics

Coarse grained models

CP2K



Scalable gas phase *ab-initio* calculations (NWChem)

I. Ground States

Edo Apra (PNNL)

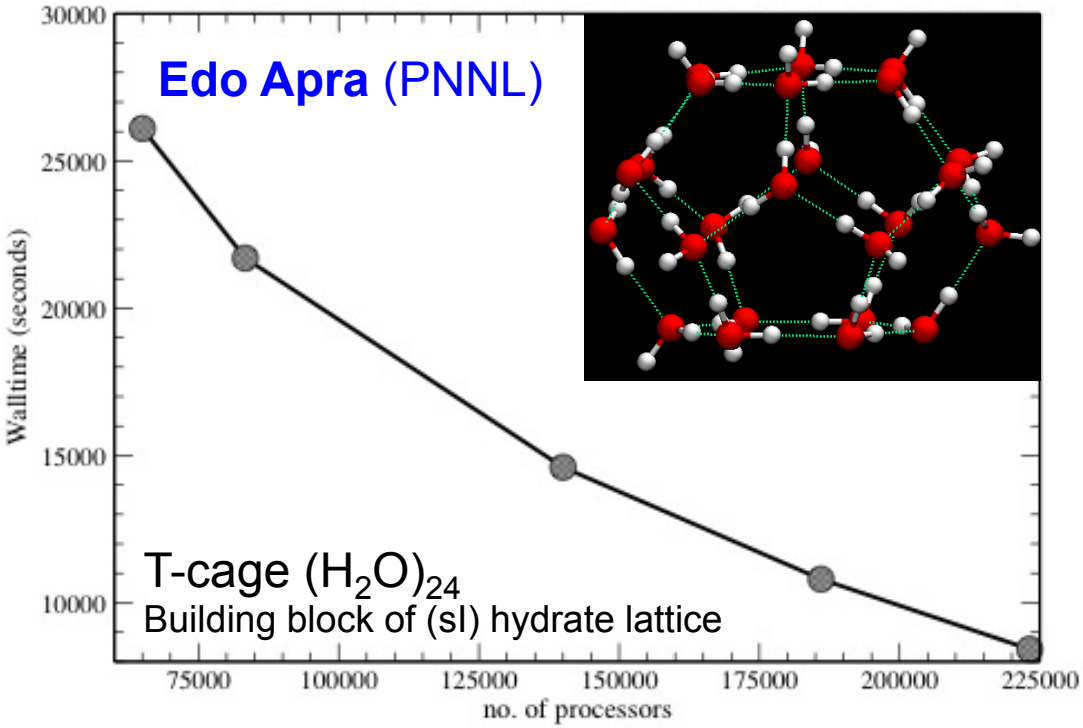
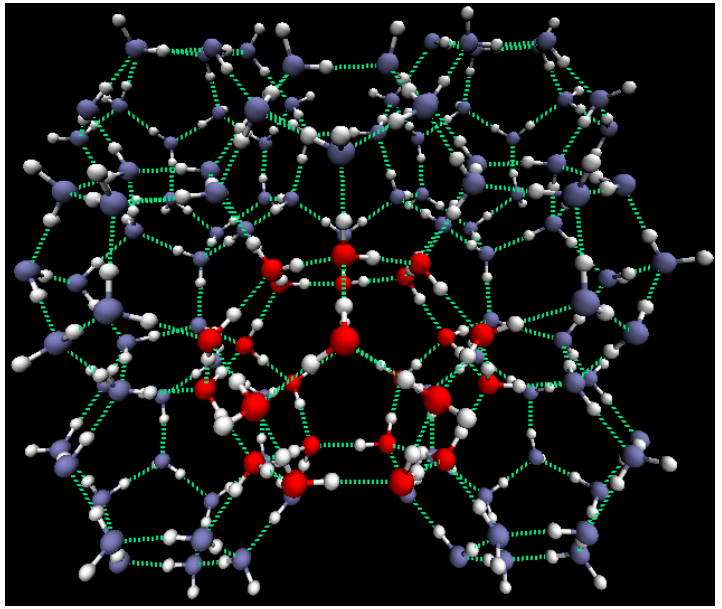


Figure 2: Walltime for the CCSD(T) calculation of total energy of $(H_2O)_{24}$ as a function of the number of processors

- Run on ORNL's "jaguar"
- 150 TB of memory
- 223,200 processors
- approx. 3 hrs. for (T)
- Double precision
- **1.39 PetaFLOP/s** (sustained)



S Yoo, MV Kirov, SS Xantheas, *JACS* **131**, 7564 (2009)

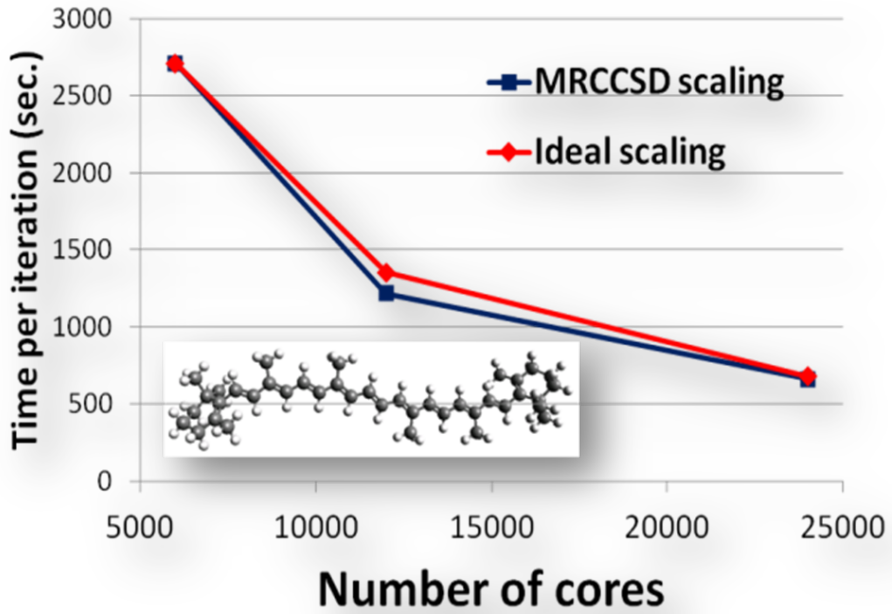
E Aprà, RJ Harrison, WA deJong, AP Rendell, V Tipparaju, SS Xantheas, *SC'09: Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis,* SESSION: Gordon Bell finalists, article No. 66, Published by ACM, New York, NY USA (2009). ISBN:978-1-60558-744-8.
DOI: http://doi.acm.org/10.1145/1654059.1654127

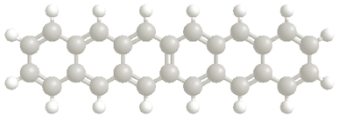
Pacific Northwest NATIONAL LABORATORY

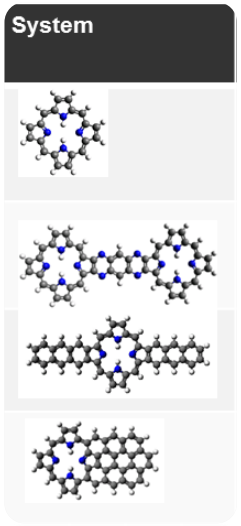
Scalable gas phase *ab-initio* calculations (NWChem)

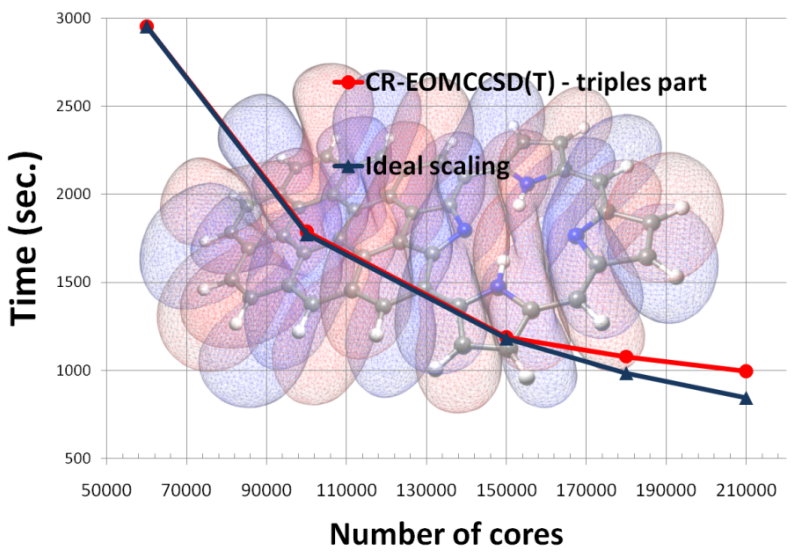
II. Excited States

Karol Kowalski (PNNL) - Extreme Scale Initiative





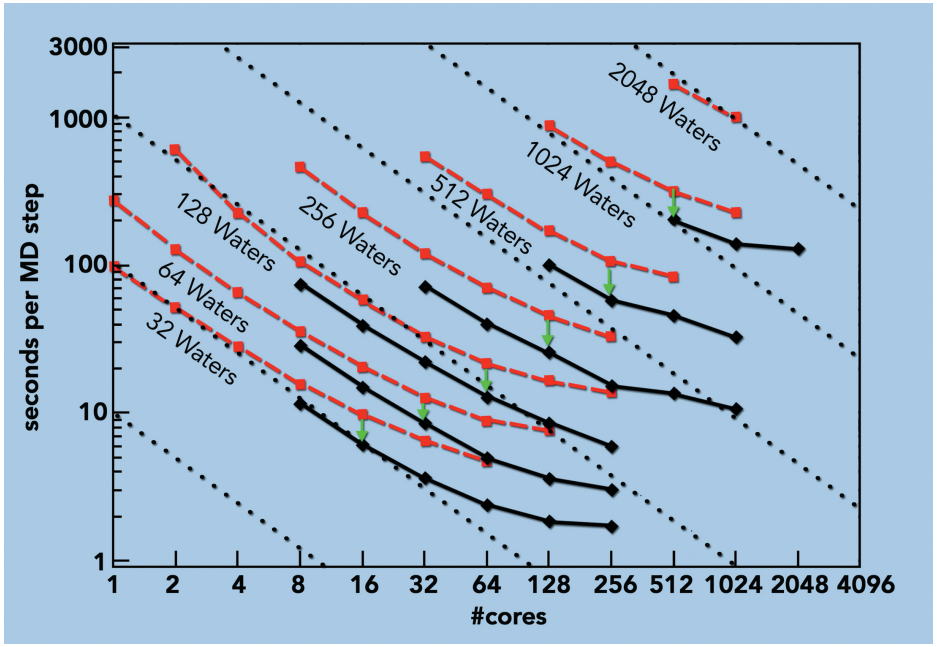


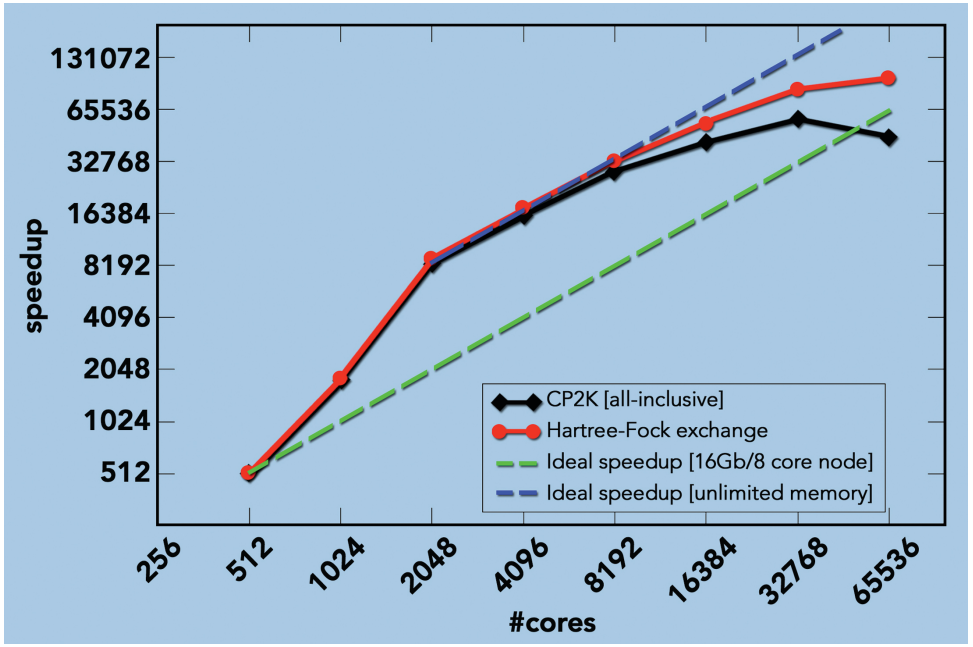


Pacific Northwest NATIONAL LABORATORY

Scalable condensed phase *ab-initio* calculations (CP2K)

J VandeVondele (ETH) **J Hutter** (Zurich)





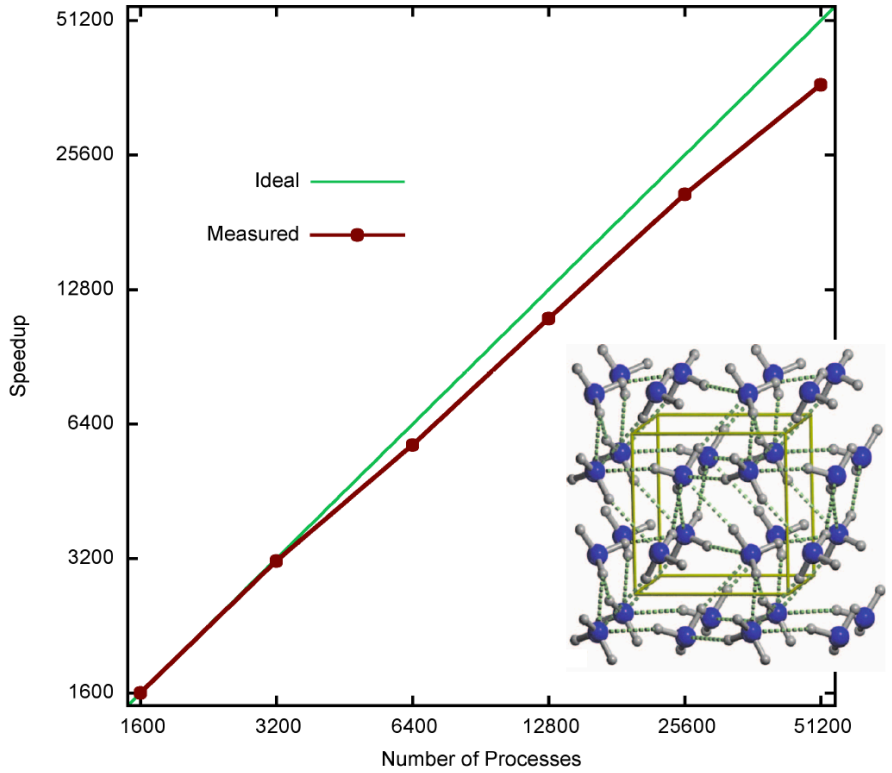
Black lines are improvement from XT3 to XT5
Standard DFT using GPW (e.g. QUICKSTEP)

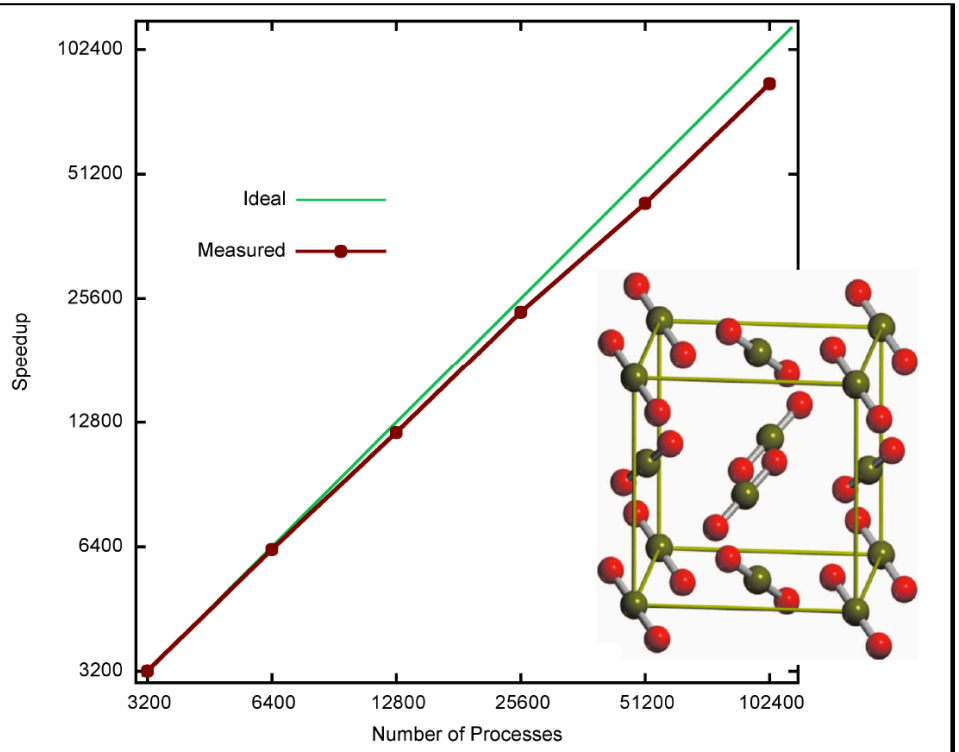
Pacific Northwest
NATIONAL LABORATORY

Scalable condensed phase *ab-initio* calculations (CP2K)

J VandeVondele (ETH) **J Hutter** (Zurich)

MP2- GPW Scalability





NH3: 74s on 51200 cores, ~cc-pVTZ : 2272BF, 128 occupied orbitals
CO2: 518s on 102400 cores, ~cc-pVQZ : 5184BF, 256 occupied orbitals

Pacific Northwest NATIONAL LABORATORY

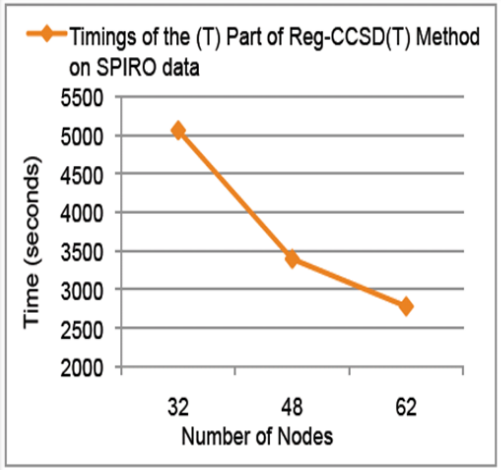
Recent GPU implementation of CCSD(T)

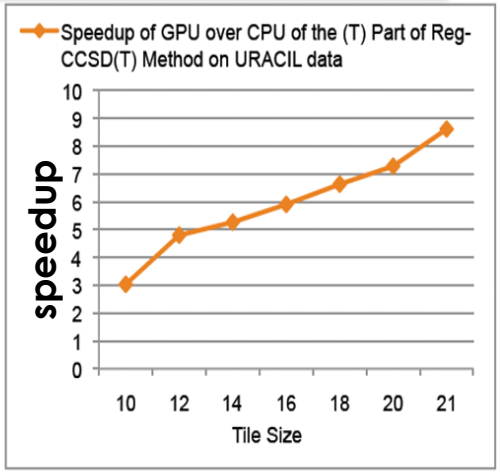
- The Reg-CCSD(T) code has been rewritten in order to take advantage of GPGPU accelerators
- Recent tests on Titan show a good speedup of the most expensive (N^7) part of the CCSD(T) approach (**5-6 ×**)

$C_{22}H_{14}$ (378 basis fncts., C_1 symmetry)
(96 nodes: 8 cores per node + 1GPU)

```
Using CUDA CCSD(T) code
Using 0 device(s) per node
CCSD[T] correction energy / hartree = -0.150973754992986
CCSD[T] correlation energy / hartree = -3.067917061062492
CCSD[T] total energy / hartree = -844.403376796441080
CCSD(T) correction energy / hartree = -0.147996460406684
CCSD(T) correlation energy / hartree = -3.064939766476190
CCSD(T) total energy / hartree = -844.400399501854849
Cpu & wall time / sec 9229.9 9240.3
```

```
Using CUDA CCSD(T) code
Using 1 device(s) per node
CCSD[T] correction energy / hartree = -0.150973754993019
CCSD[T] correlation energy / hartree = -3.067917061062597
CCSD[T] total energy / hartree = -844.403376796441307
CCSD(T) correction energy / hartree = -0.147996460406693
CCSD(T) correlation energy / hartree = -3.064939766476270
CCSD(T) total energy / hartree = -844.400399501854963
Cpu & wall time / sec 1468.0 1630.7
```





- Preliminary tests on ORNL's Titan (2013) show ~6x speedup
- The MR-CCSD(T) implementation based on three levels of parallelism (reference level for MR part, task level within each reference and GPU for (T) contribution) is currently under development

Karol Kowalski (PNNL)

Pacific Northwest
NATIONAL LABORATORY

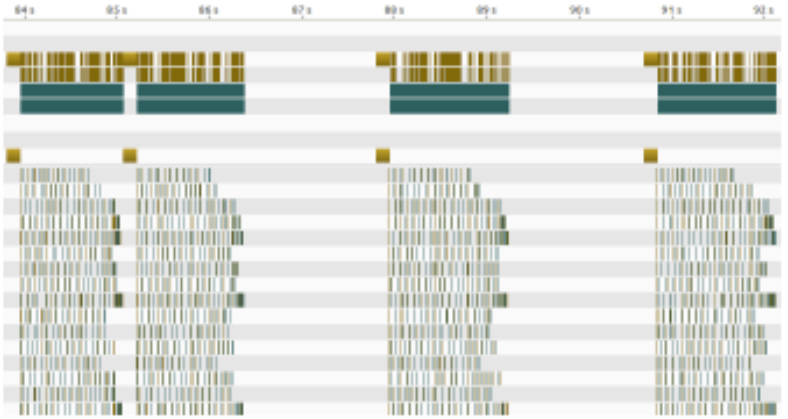
GPU Application Benchmark (CP2K)

- >400 multiplications for 1 run.
- Additional thresholding in multiplications (less flops for same data)
- This week's results.... subject to change

20736 atoms (6912 water molecules), matrix dim 159000, on 576 nodes XK6,
~60 matrix multiplications / iter.

XK6 without GPU : 1965s per iteration
XK6 with GPU : 924s per iteration

Speedup 2.12x



MPI performance (bandwidth) appears to be the bottleneck (e.g. 50% slowdown without custom rank reordering) :

- Still need to figure out MPI performance (incl. effectiveness of overlap).
- Is the dynamic linking still an issue ?
- Any interference between GPU+CPU ?
- One Communication thread per node enough ?

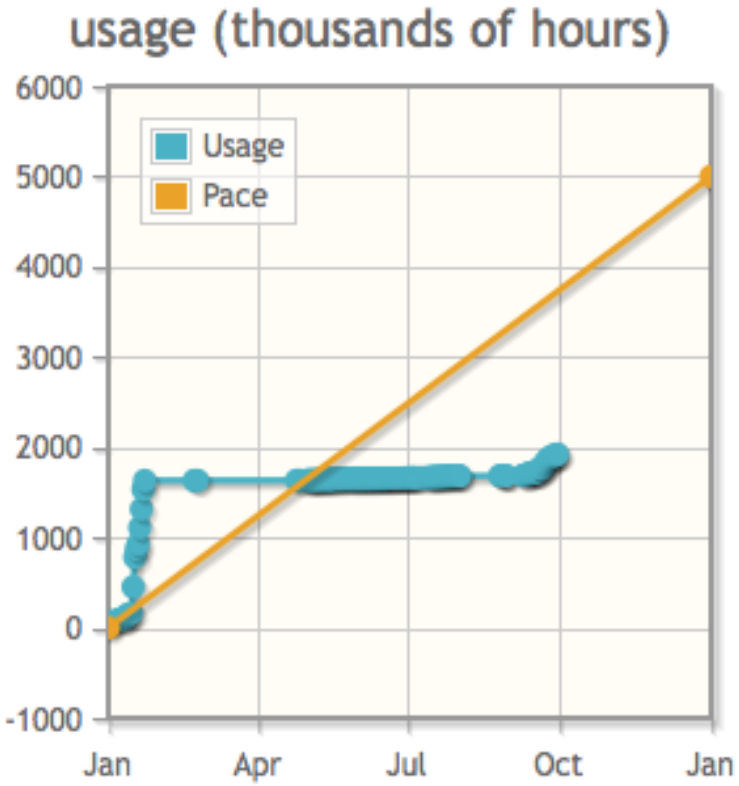
J VandeVondele (ETH) **J Hutter** (Zurich)

Present and Future

	NWChem		CP2K	
	Present	**Future**	**Present**	**Future**
Molecular systems	Model	Real	homogeneous	heterogeneous
CPU hours	6.8 M	500 M	5 M	20 M
Compute cores	500 – 20,000	50,000 – 250,000	1,000 – 2,000	10,000
Max. # of cores	200,000+	1,000,000+	64,000	100,000
High throughput benefit?	No	Yes	Yes	Yes
S/W scaling	Both (Strong)	Both (Strong)	Both	Both
Scratch I/O	< 10TB	< 200 TB		
I/O bandwidth		1 TB / sec		
Shared data	No	< 100 GB	Yes	5 TB
% of I/O		< 1%		
Scratch data		1 PB		
Archival Storage	< 500 GB	< 200 TB		10 TB / yr.
Memory		5x		16 GB / node

Usage of repo m1513

PI: Sotiris Xantheas
Category: Chemistry
Initial Allocation: 5,000,000
Current Allocation: 5,000,000
Charged Usage: 1,923,932
Total Usage: 2,031,006
Charge Factor: 0.899355
Percent Used: 38



Usage of repo m1513
(NISE award)
as of 10/08/2013

Pacific Northwest
NATIONAL LABORATORY

Synopsis

- Detailed plan to move from model ("toy") to more realistic systems
- Exploit algorithmic dependence of the "*golden standard of quantum chemistry*" [CCSD(T) / MR-CCSD(T)] to target efficient scaling over 1M+ cores
- Provide the benchmarks by which lower scaling methods (such as DFT and/or semi-empirical methods) will be chosen for the statistical sampling needed to obtain macroscopic properties
- NERSC is a major player in this process
  - early access
  - test parallelization software (Global Arrays) early on
  - address fault tolerance issues
  - provide RDMA extensions in MPI
  - efficient parallel linear algebra for heterogeneous architectures
  - consulting & account support crucial
  - job scheduling policies that "favor" jobs requesting more resources
- We are confident that the synergy between the Molecular Theory group, the NWChem / CP2K developers will
  - result in new paradigms in several science domains
  - serve, support and further enhance the DoE's mission

Backup Slides

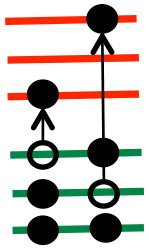
Coupled Cluster formalism

- **CC method**

(Coester, Kümmel, Cizek,Paldus,Bartlett,…)

$$|\Psi\rangle = e^{T}|\Phi\rangle \qquad T = T_1 + T_2 + T_3 + \ldots + T_N$$

$$T_2|\Phi\rangle = \sum_{\substack{i<j \\ a<b}} t_{ab}^{ij}$$



- CC is size-extensive (scales properly with the number of electrons)
- Approximate higher excitations through the products of cluster operators
- Exact limit exists
- Applicable to: nuclear matter, molecular systems, polymers, solid state

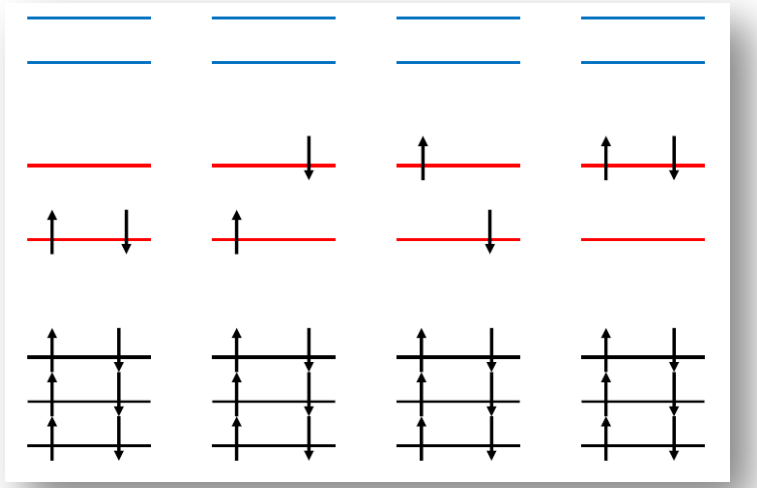
Pacific Northwest
NATIONAL LABORATORY

MRCC theory in a nutshell

$$\left|\Phi\right\rangle \longrightarrow M_0 = ls\left\{\left|\Phi_\mu\right\rangle\right\}_{\mu=1}^{M}$$

Reference function

Model space



Schematic representation of the complete model space corresponding to two active electrons distributed over two active orbitals (red lines). Only determinants with M_S=0are included in the model space.

Pacific Northwest
NATIONAL LABORATORY

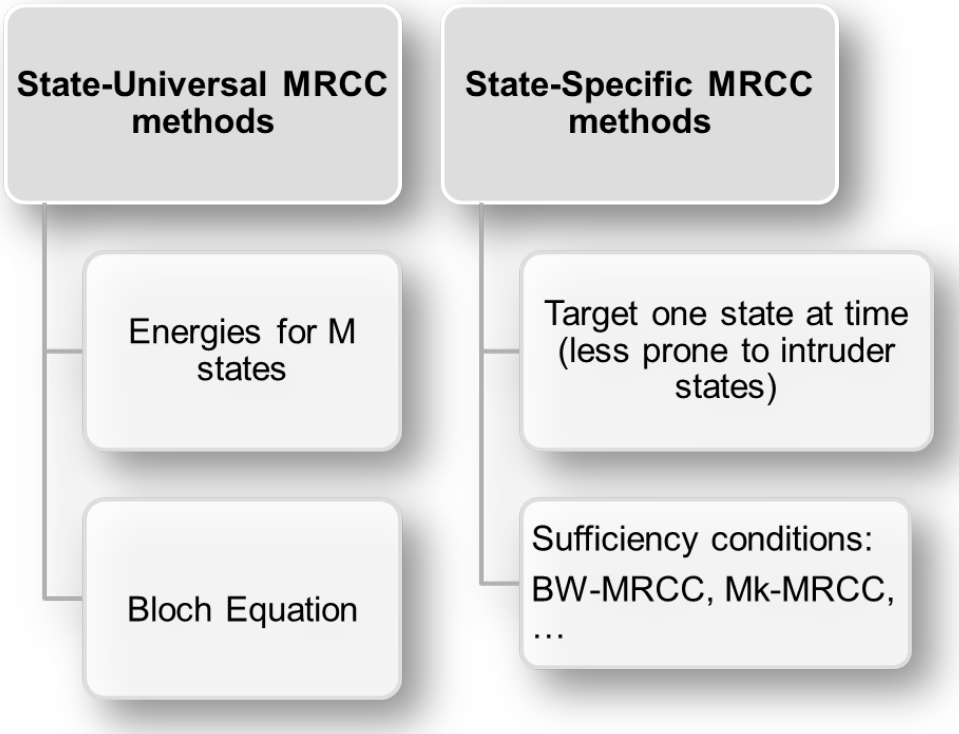
MRCC theory in a nutshell

$$|\Psi\rangle = \sum_{\mu=1}^{M} c_\mu e^{T(\mu)} |\Phi_\mu\rangle$$

MRCC wavefunction expansion (Jeziorski, Monkhorst)

$$H^{eff} c = Ec$$

$$H^{eff} = [\langle \Phi_\nu | (He^{T(\mu)})_C | \Phi_\mu \rangle]$$



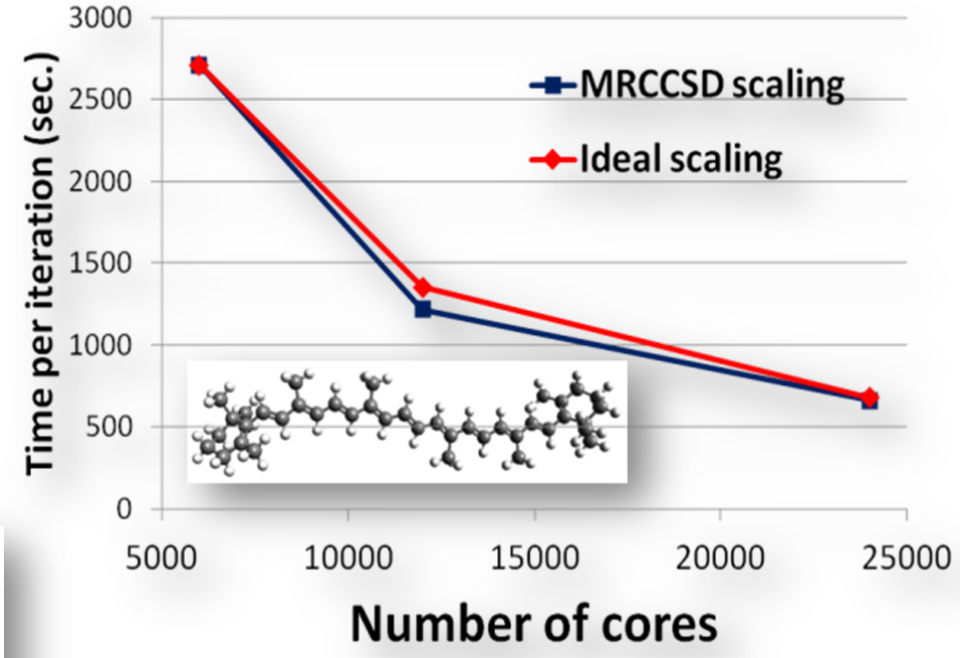
Main challenges

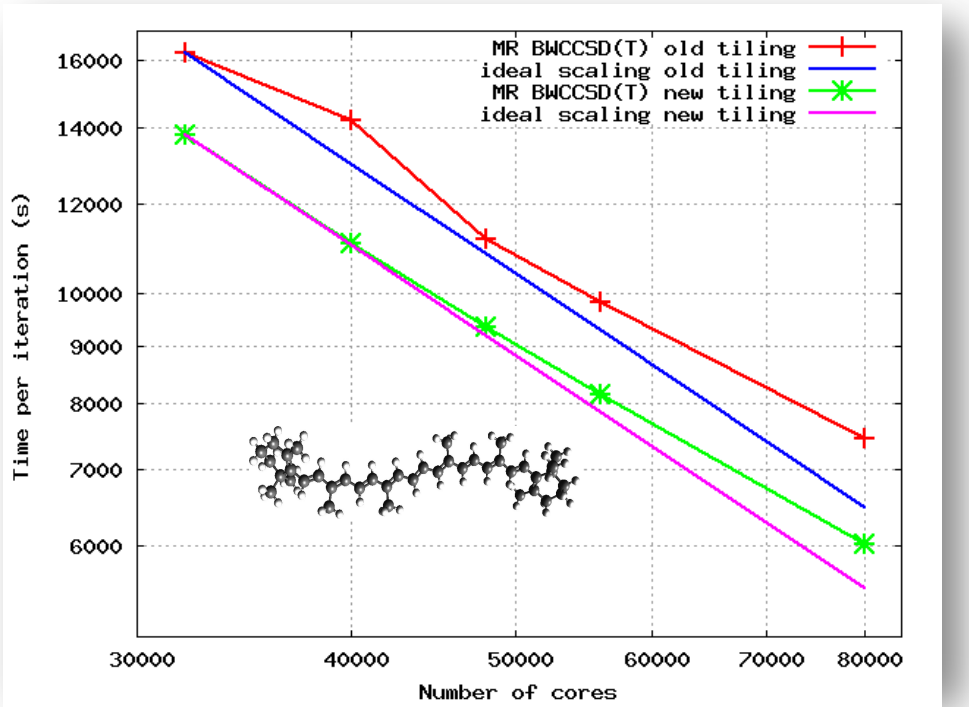
- Intruder-state/intruder-solution problems
  - Schucan & Weidenmuller
- Complete model space
  - Huge dimensionality
- Overall cost of the MRCC methods
  - $M \times N^6$ (iterative MRCCSD)
  - $M \times N^7$ (non-iterative MRCCSD(T))
- Algebraic complexity of the MRCC methods

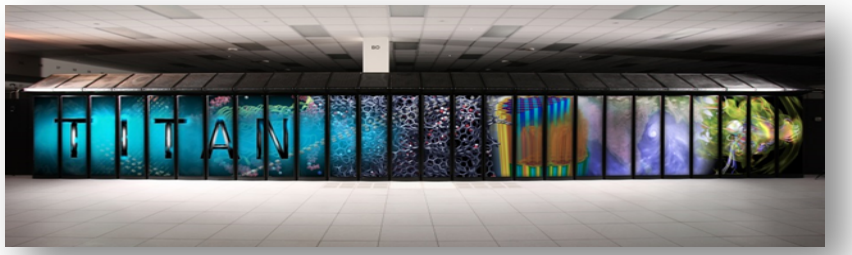
Pacific Northwest
NATIONAL LABORATORY

Processor group (PGs) and reference level parallelism

- Scalability of the BW-MRCCSD/ BW-MRCCSD(T) methods for β-carotene in the 6-31G basis set (~470 basis set functions); (4,4) complete model space model space (20 reference functions) was used







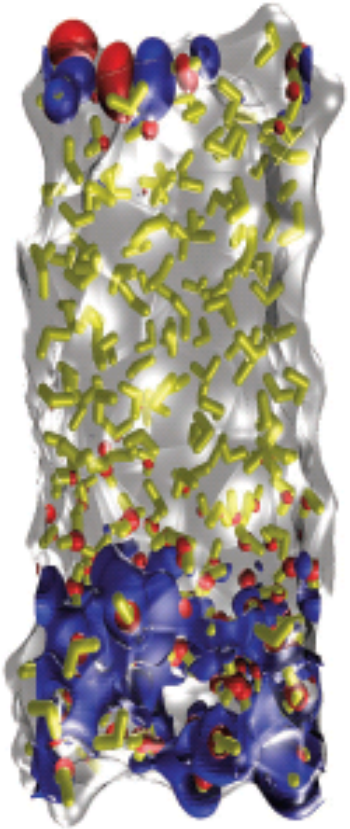
Pacific Northwest
NATIONAL LABORATORY

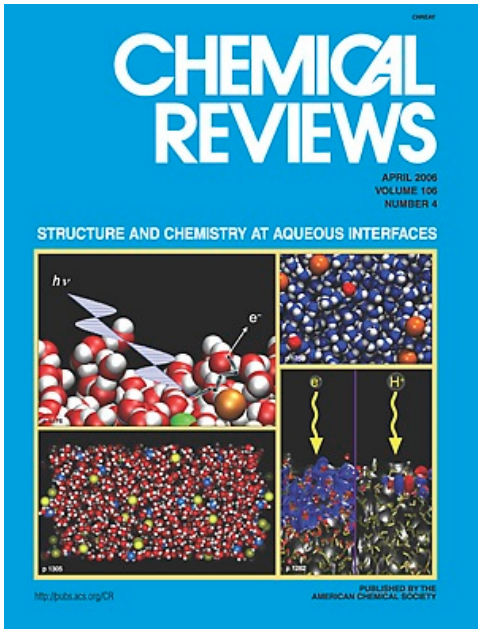
We continue to develop protocols for large-scale DFT simulation for open aqueous systems

IFW Kuo and CJM, *Science* 303, 658 (2004)

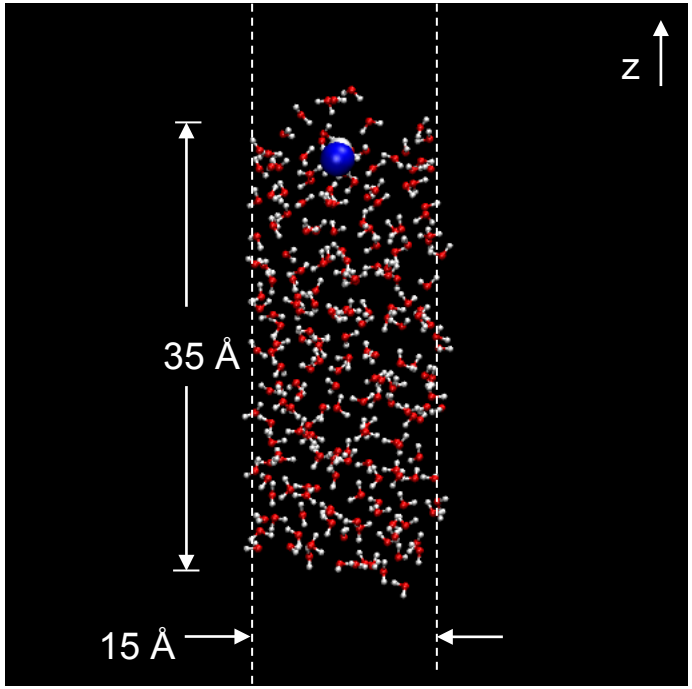
An ab Initio Molecular Dynamics Study of the Aqueous Liquid-Vapor Interface

I-Feng W. Kuo and Christopher J. Mundy*





CJM, IFW Kuo, *Chemical Reviews* 106, 1282 (2006)



- First large-scale heterogeneous interface with DFT performed at LLNL in 2004. 216 waters are needed to yield a true interfacial system (1000's of processors on MCR)
- Results have stood test of time. Larger systems are needed in order to simulate the effects of dilute salt solutions and their chemical reactions

CP2K Computational Kernels

- Cholesky
- Matrix-Matrix multiplication
- Diagonalization

•Sparse linear algebra:

- Matrix Matrix
- Matrix vector

•Regular grids:

- multigrids and FFT

•Time integration:

- MD simulation

•Chemical kernels:

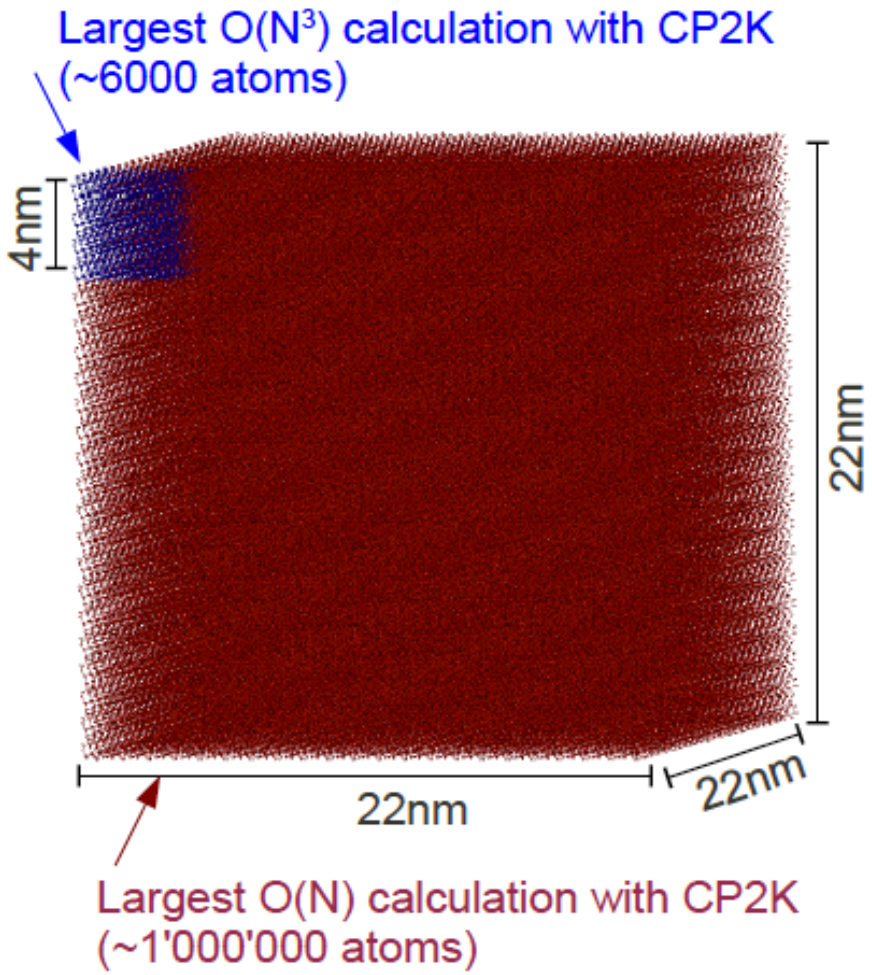
- Grid collocation & integration
- four center integrals
- MP2

A single kernel rarely dominates
Scaling is O(N)...O(N**5)

Linear Scaling SCF (CP2K)

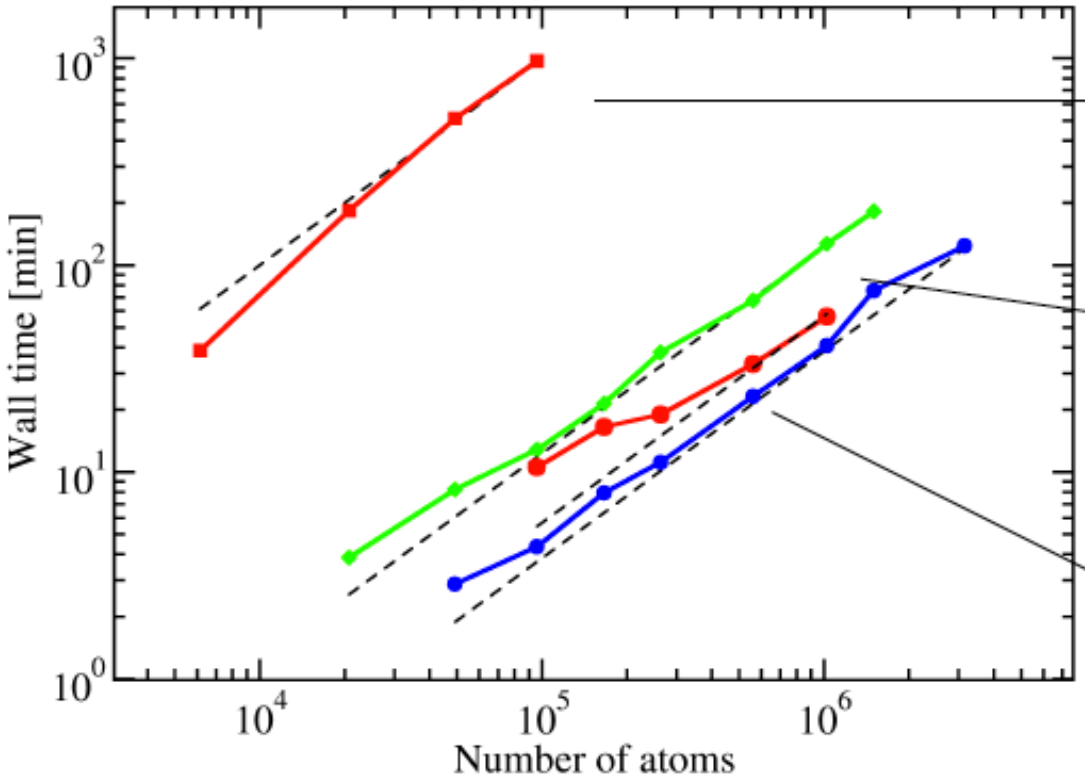
Traditional approaches to solve the self-consistent field (SCF) equations are $O(N^3)$ limiting system size significantly.

A newly implemented algorithm is O(N), allowing for far larger systems to be studied.



Pacific Northwest
NATIONAL LABORATORY

System size in the condensed phase (CP2K)



Accurate basis sets, DFT
46656 cores

The electronic structure
O(10^6) atoms in < 2 hours

Minimal basis sets:
DFT, NDDO, DFTB
9216 cores

Bulk liquid water. Dashed lines represent ideal linear scaling.

VandeVondele, Borstnik, Hutter, JCTC, http://pubs.acs.org/doi/abs/10.1021/ct200897x