Scalable Modeling and Simulation

An Overview of Extreme Scale Application Development at PNNL

August 17, 2011

Tjerk P. Straatsma, Laboratory Fellow

Darren J. Kerbyson, Laboratory Fellow

Pacific Northwest
NATIONAL LABORATORY

Toward Exascale Scientific Applications

Exascale computing will harness the power of next generation architectures with massive concurrency
Solving national and global challenges in energy, security, and environment
Enabling analysis and prediction of behavior of complex phenomena

- Our Science Missions
  - Depend on scientific computing
  - Need applications to be ready for next generation exascale computing resources
- Applications depend on advances in high-performance CS to achieve
  - Scalability
  - Robustness
- Exascale computers are expected to be based on
  - Massive numbers of processes
  - Hierarchical communication networks
  - Many-core and heterogeneous processors

Pacific Northwest
NATIONAL LABORATORY

PNNL's eXtreme Scale Computing Initiative (XSCI)

Purpose

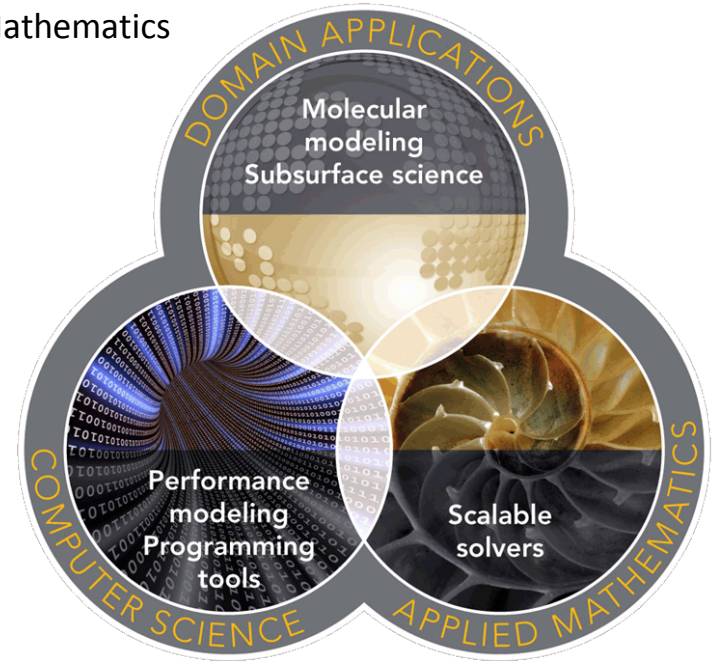
- Develop methods, algorithms and software enabling PNNL to use next generation massively parallel computer systems for addressing computational grand challenges in the areas of science, energy, environment and national security.

Approach

- Design and implement scalable methods for specific scientific domains
- Strengthen our High Performance Computer Science and Applied Mathematics

Impacts

- Leadership in solving client problems by the use of extreme scale computers
- New software tools for data management, load balancing, parallel I/O, scalable communication
- New math capabilities for analytics of complex, heterogeneous, and multi-physics problems
- Long-term collaborations with leading institutions and academia
- National leadership in Computer Science for extreme scale scientific computing and Computational Science

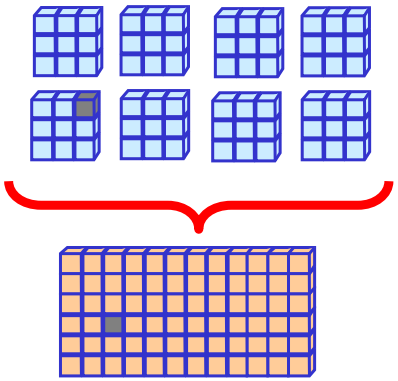


Pacific Northwest
NATIONAL LABORATORY

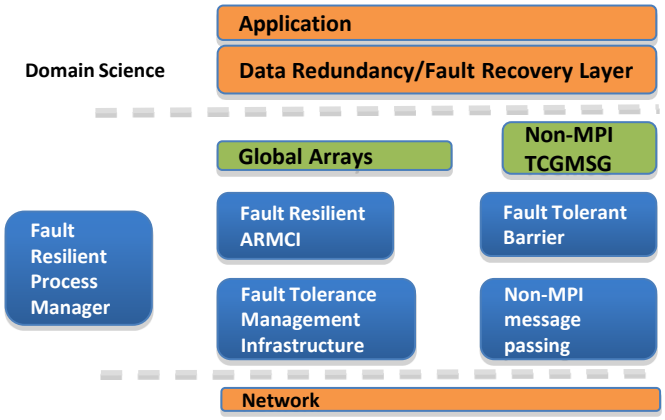
The GA/ARMCI Programming Model

Global Arrays: Distributed dense arrays that can be accessed through a shared memory-like style
ARMCI: One-sided asynchronous communication protocol

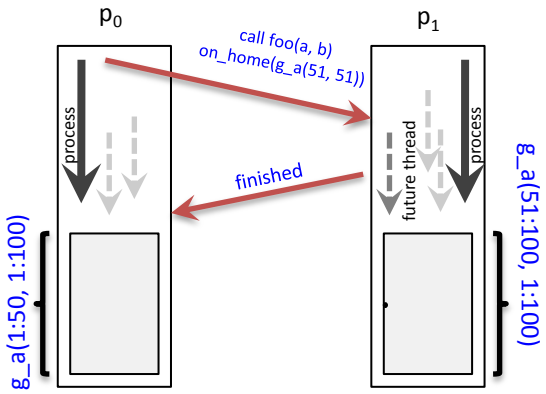
Physically distributed data



Fault Tolerance Infrastructure



Global Futures



- Globally shared view of multi-dimensional arrays
- Inter-operates with MPI
- Data-locality and granularity control is explicit with GA's get-compute-put model, unlike the non-transparent communication overheads with other PGAS models
- Library-based approach: does not rely upon smart compiler optimizations to achieve high performance
- Data consistency must be explicitly managed

- Computational Chemistry
  - **NWChem, GAMESS UK, MOLPRO, MOLCASS**
- Subsurface Transport Simulation
  - **eSTOMP**
- Bioinformatics
  - **ScalaBLAST**
- Computational Fluid Dynamics
  - **Thethys**

Pacific Northwest NATIONAL LABORATORY

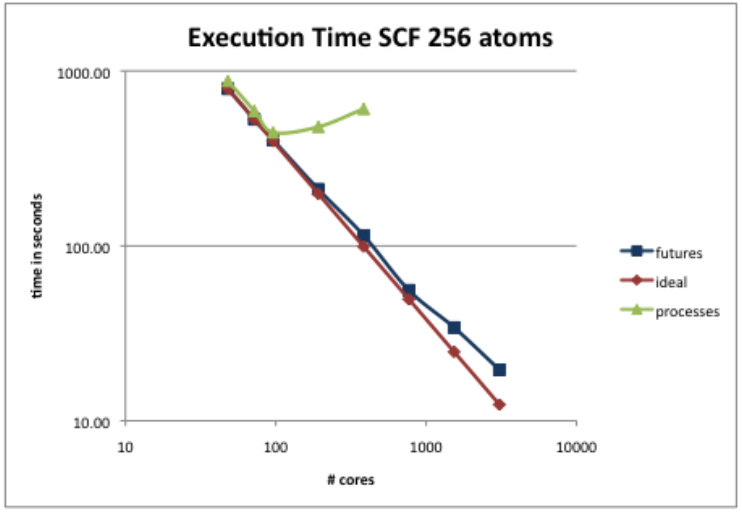
XSCI Technical Challenges

Scalability	Global Arrays: Reduced meta-data requirements Global futures: Moving computation to the data
Architectures	GPU accelerators: Novel CUDA implementations of application kernels System architecture: Taking advantage of future networks
Fault Tolerance	Data Replication techniques to handle node failures Increasingly looking towards soft-errors
Productivity	Diagnostics tools: Profiling Interface, Trace collection and visualization Community tools: Leveraging best in class tools for analysis
Power Management	Modeling/Optimization: DVFS & Interrupt driven techniques Run time and application specific optimizations

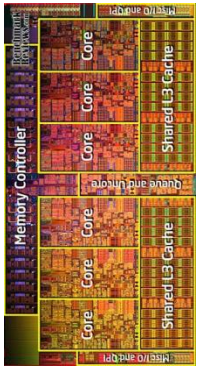
Pacific Northwest
NATIONAL LABORATORY

Scalability

- Inherent to most aspects of XSCI
  - Applications are capable of using current largest-scale systems (Molecular science, and sub-surface)
  - Designing for, and demonstrating, scalability in all aspects
- Not just about scale-out but also scale-in (node)
  - Expect O(100) increase in node capabilities vs. O(10) increase in #nodes
- GA metadata
  - Current requirements:
    -> 8bytes per process on each process (=1.6MB per GA instance @ 200K cores)
  - Optimization to share metadata on same shared memory domain (node)
    -> reduce by O(10) today and greater impact later
- Global Futures
  - Moving computation to the data
    - optimize (reduce) traffic flow and cost
  - API designed and tested
    - Demonstrated on SCF with much improved scalability
  - Much promise for wider use

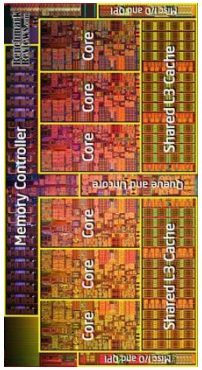


Architectures: which path leads to the future?

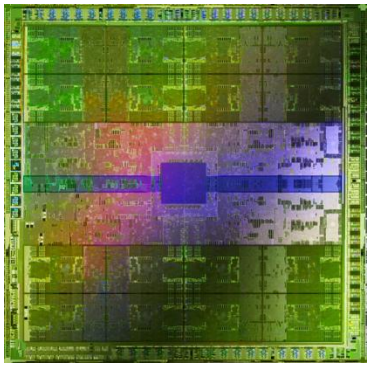


Multi-core
6-8 cores

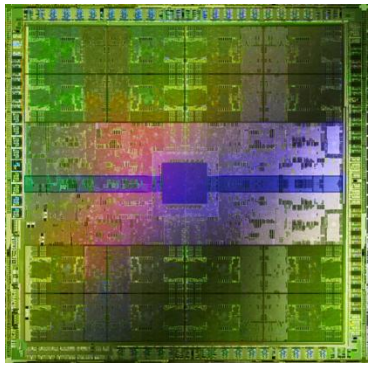
vs.



+

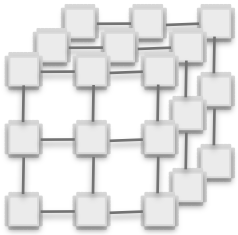


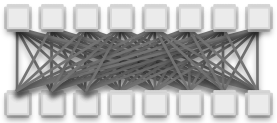
vs.

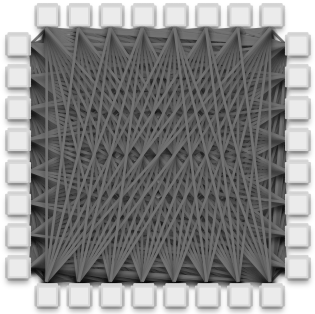


GPU: 16 cores
(internally 32 SIMD)

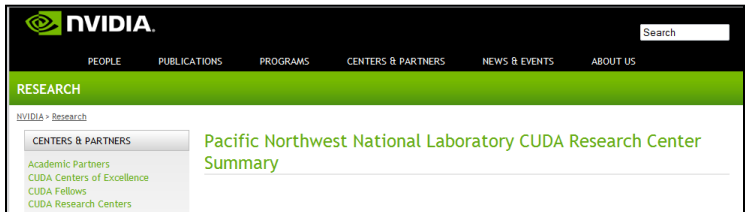
- Demonstrated performance improvements at system level
- GPU Optimizations for:
  - NWChem: TCE, CCSD(T), SCF, MD
  - eSTOMP: Chemistry reactions
- Significant contributions:
  - Optimizing data movement
  - Maximizing concurrency
  - On a micro & macro level







- Communication networks
  - 3D, 5D, 6D torus
  - Fat and skinny trees
  - Hierarchical networks: Dragonfly (Aries), IBM P7-IH
- Impact on applications
  - Data & Task Mapping
  - Data movement
  - Instruction movement



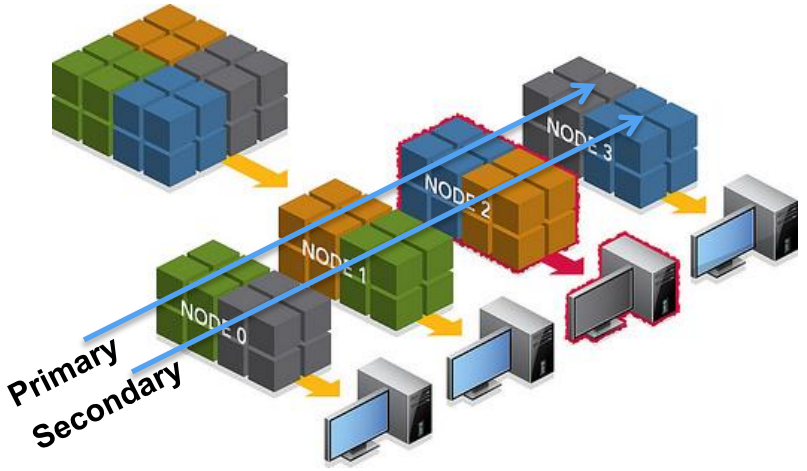
Pacific Northwest
NATIONAL LABORATORY

Fault Tolerance

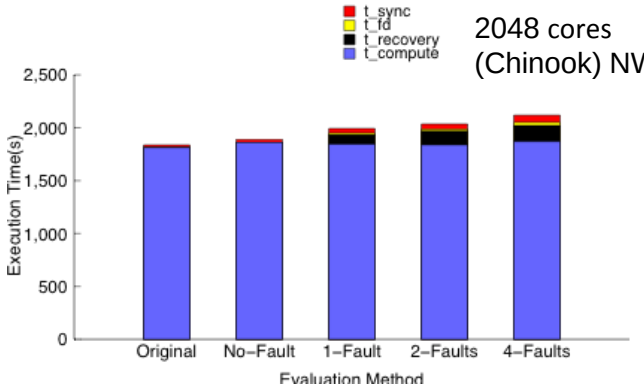
- Increased system size & inherent increases in technology faults will lead to a greater reliability issue

- Hard faults: node failures
  - 85% of all faults involve a single node
  - Combined application / run-time approach
  - Selective Data Replication
    - Re-computation of "lost" tasks
  - Resilient Process Manager
    - identify node failures &
    - ensure continued execution
  - Issues: fault detection (time), topology aware replica mapping
- Soft errors
  - Exploring techniques: e.g. Numerical assertions
  - Infrastructure development for asynchronous soft-error detection

- Production ready software
  - Close Collaboration with Cray: Jaguar (OLCF)
  - Infiniband (Chinook @ EMSL)
  - Working towards Cray Gemini, Power7-IH



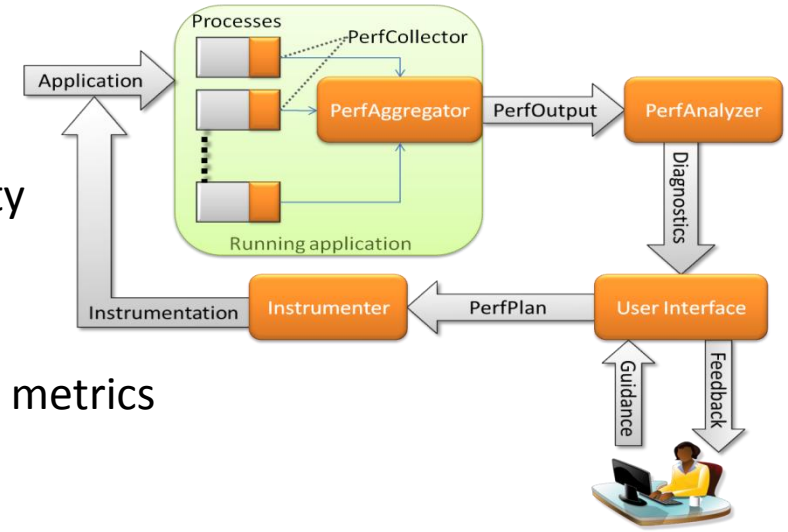
Example data mapped across 4 nodes: Node 2 dies, data recovery from secondary copy (node 3)



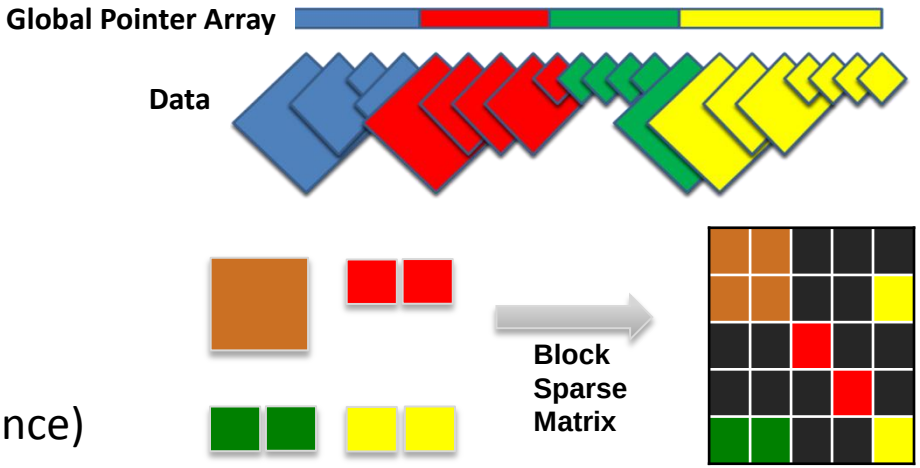
- Overheads (application dependent):
  - 2-3% time (no faults)
  - 5% space
- Improvement: $P^{-1/2}$ on MTBF

Productivity

- Performance Diagnostics
  - Weak bindings for GA API -> profiling
  - Leverage state-of-the-art tools in the community
    - TAU (U Oregon), Scalasca (Juelich)
  - Looking towards performance patterns
    - Identification of patterns using data centric metrics
  - Use within application teams



- Advanced Data decomposition
  - Global Pointers (GP)
  - Global view of distributed data structures/objects
  - Support for data re-distribution
  - Scalable data decomposition for:
    - Block-sparse Tensors (Molecular Science)
    - Linked Data Structures, Sparse arrays …



Pacific Northwest
NATIONAL LABORATORY

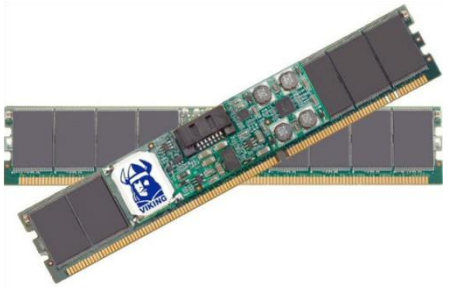
1MWyr = $1M

Power – the critical resource for the future

Current systems take up to 10MW (~ 1MW per Pflop)

Advances in hardware and software need to consider power requirements in future systems

- Developing capability to analyze, optimize and model power consumption
- Need to explore at all levels:
  - Application, System-software, Architecture
- Current activities
  - Application level power library
  - Microbenchmarks for power
  - Energy efficiency of GPUs
  - NVRAMs – SATADimms
  - Energy Templates
  - Power-aware runtime
- Modeling: bridge from small to large-scale



Pacific Northwest
NATIONAL LABORATORY

Scalable Applications

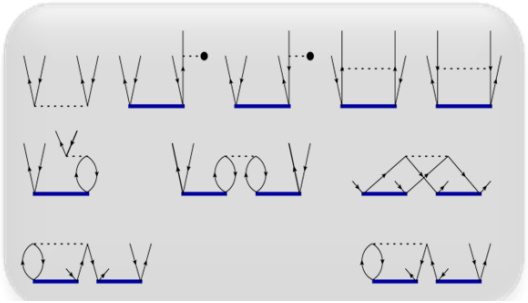
Developing scalable domain science applications in support of PNNL's major R&D activities

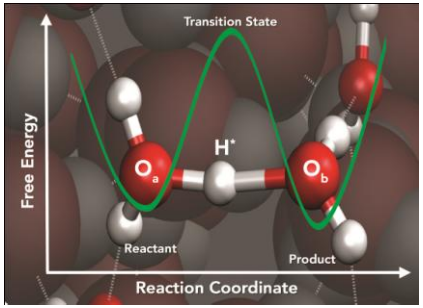
Distinguishing Modeling and Simulation Capabilities Developed at PNNL

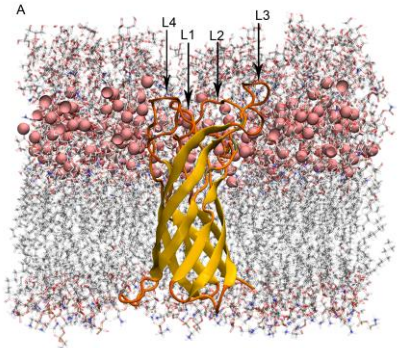
Molecular Science Applications – Energy, Environment		
EOM-CCSD(T) MR-CCSD(T)	Electronic structure of excited states and Multi-Reference states at high accuracy	Irregular data Compute-intensive Load balancing
PW-DFT	Electronic structure and dynamics	
MD, QM/MM	Classical and hybrid modeling of large systems	Irregular data Communication-intensive Synchronization Load balancing
Subsurface Science Applications – Environment		
eSTOMP	Modeling of subsurface contaminant fate and transport, carbon sequestration	Regular data Solver scalability
Engineering Applications – Energy, Environment		
ParaFlow	Lattice Boltzmann CFD modeling of tank mixing	

Pacific Northwest
NATIONAL LABORATORY

Computational Molecular Science







Simulating the Electronic Structure of Molecules

What do electrons do?

Quantum Mechanics
CCSD(T)

Simulating Molecular Interactions

What do molecules do?

***ab initio* MD**
PWDFT

Simulating Molecular System Properties

What do systems do?

Statistical Mechanics
MD

Technical Challenges:

Irregular data
Load balancing
Compute intensive

Irregular data
Load balancing
Communication intensive
Synchronization

Pacific Northwest
NATIONAL LABORATORY

Numerical challenges for EOM-CC

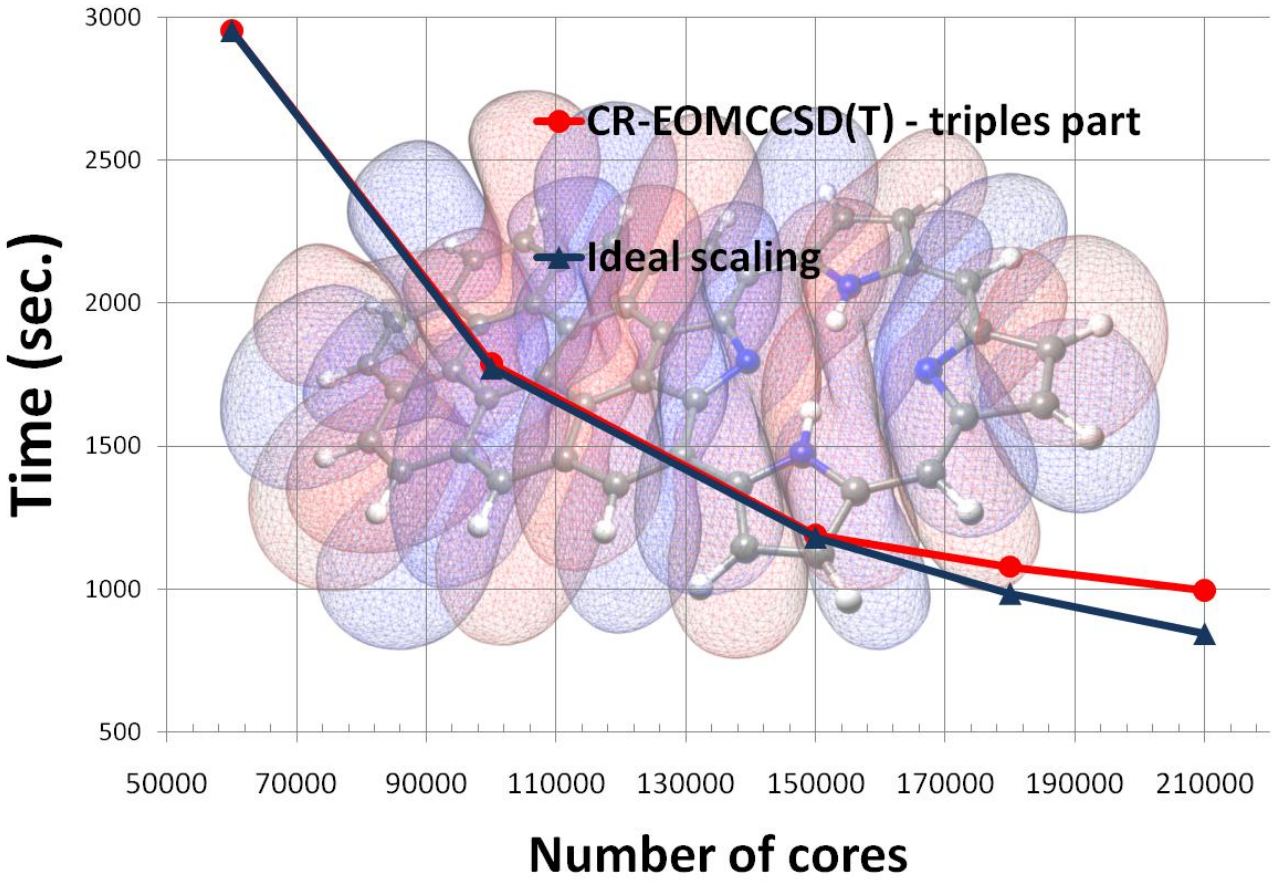
Computational requirements increase significantly with increased molecular systems size

Method	Memory demands	Numerical complexity	Applications
EOMCCSD	N^4	N^6	Singly excited states
CR-EOMCCSD(T)	N^4	N^7	Singly/double excited states; Potential energy surfaces (PES)
Active-space CR-EOMCCSD(T)	*small* N^4 $(n_o n_u^3)$	$N_{act}^5 N^2$	Singly/double excited states; PES
Löwdin-partitioning based CR-EOMCCSD(T)	N^4	N^7	As CR-EOMCCSD(T); larger number of roots can be scanned

Pacific Northwest
NATIONAL LABORATORY

Scalability of the non-iterative EOM-CC

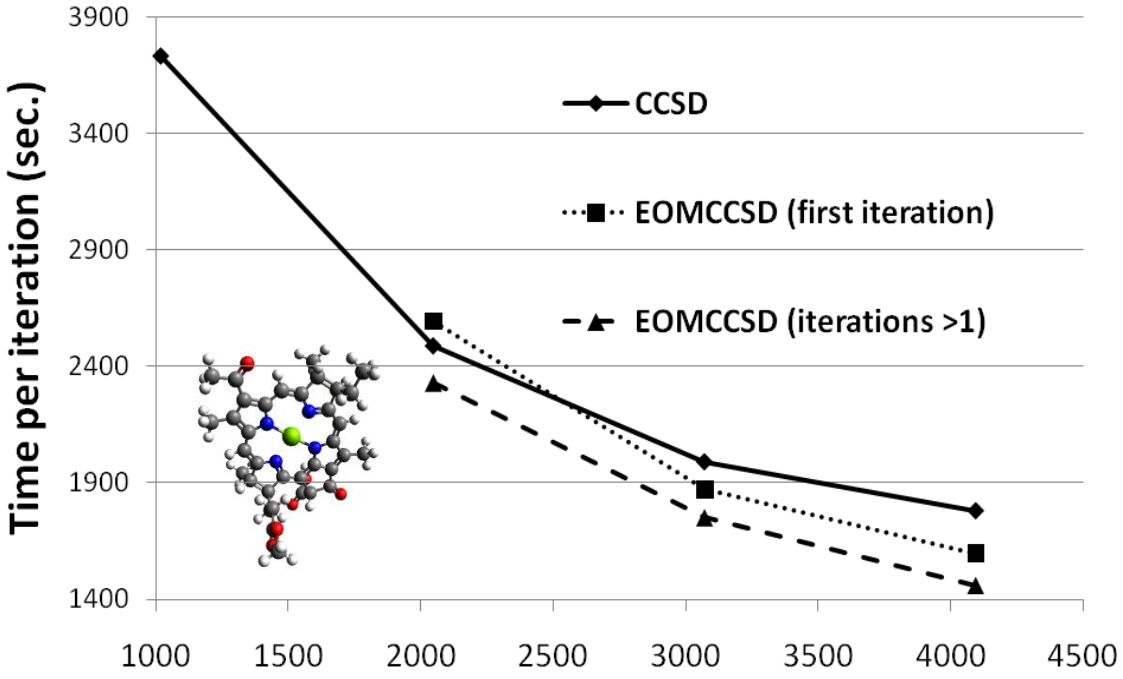
- Scalability of the triples part of the CR-EOMCCSD(T) approach for the FBP-f-coronene system in the AVTZ basis set.
- Timings were determined from calculations on the Jaguar Cray XT5 computer system at NCCS

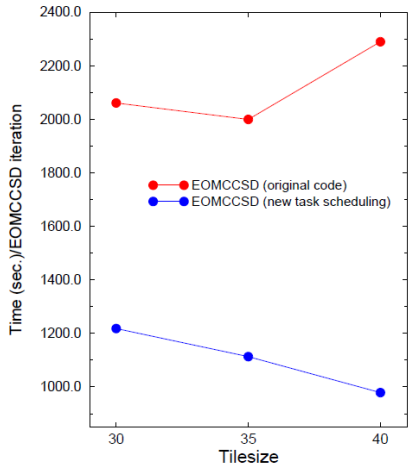


Pacific Northwest
NATIONAL LABORATORY

Scalability of the iterative EOMCC methods

- Storage and reuse of T-dependent recursive intermediates
- Alternative task schedulers – towards better load balancing
  - Current structure of the CCSD/EOMCCSD codes is semi-serial, which can aversely affect the parallel performance of these codes.
  - Independent procedures are grouped into several classes/layers
  - Instead of having N_i independent tasks characterizing *i*-th procedure/diagram, the layered model provides much larger task pool equal to $\sum_{i \in I} N_i$ for the *I*-th layer.



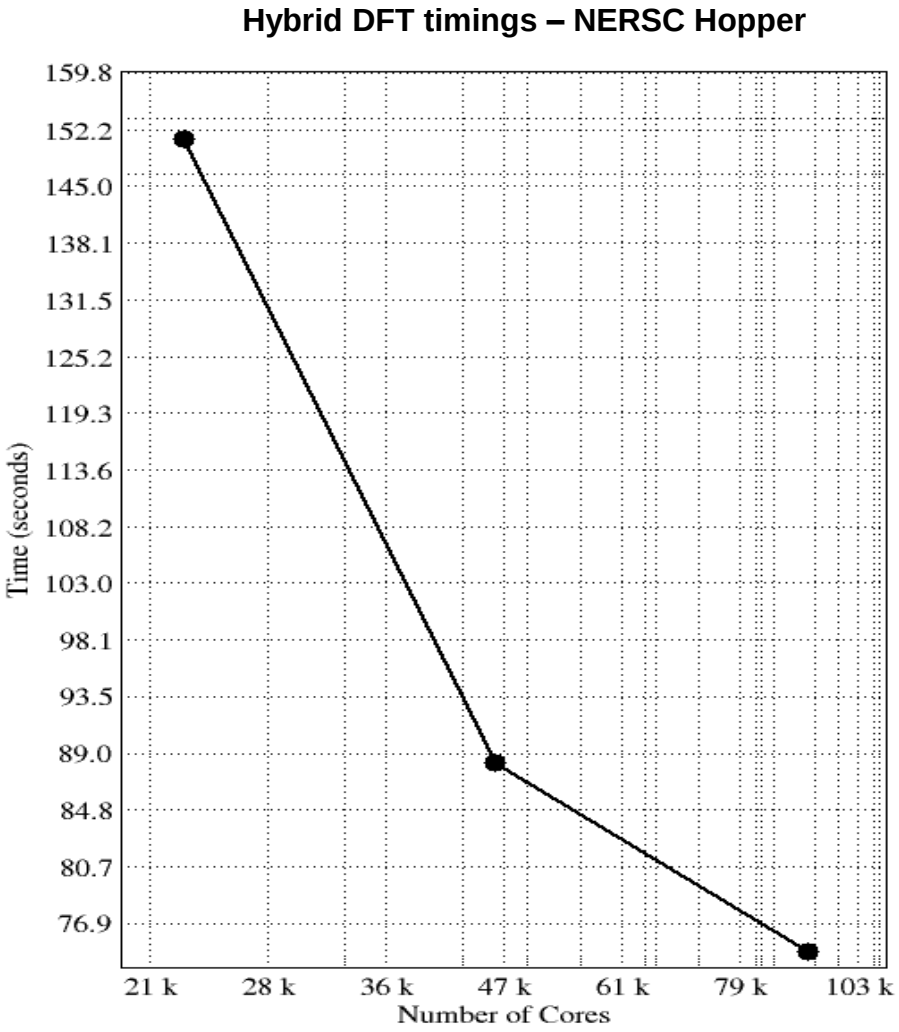


Pacific Northwest NATIONAL LABORATORY

Hybrid DFT AIMD

AIMD Performance Improvements

- New parallel algorithm for hybrid DFT reducing communication by ½
  - 10^5-10^6 (MP2 easily 10^9) 3d ffts are computed in seconds.
- Performance model for exact exchange algorithms
  - Algorithms will scale well over 100,000 cpus in the future
  - Bottleneck in the flop rate of the grouped parallel 3d FFTs (FLOP rates of parallel ffts is at most 15% even for small cpu numbers!)

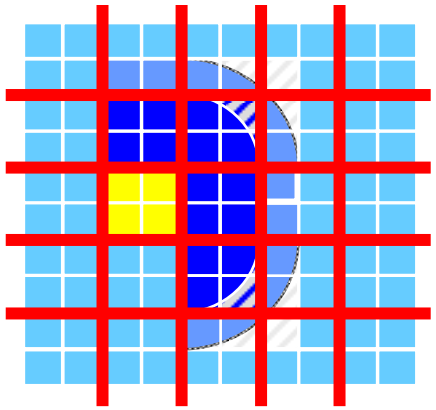


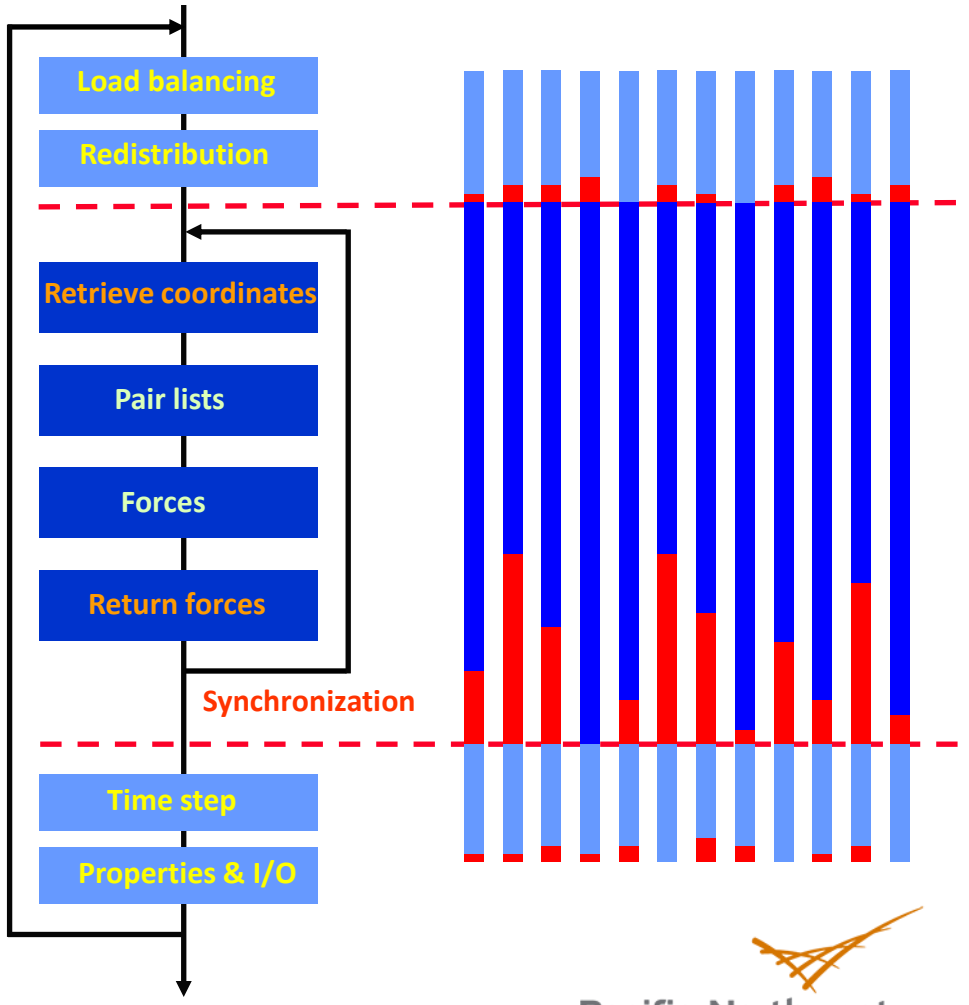
Pacific Northwest NATIONAL LABORATORY

Technical Challenges for MD

Time scale is bigger challenge than problem size: Strong scaling increases communication

- Synchronization between force evaluation and coordinate advance
- Effective assignment to processes of cell-cell interactions in domain decomposition
- Efficient load balancing in heterogeneous molecular systems
- Identification and appropriate single and multiple node mapping of concurrency
- Long range electrostatics corrections





Pacific Northwest
NATIONAL LABORATORY

Technical Approach

- **Identifying levels of parallelism in the MD kernel**
  - Task parallelism: MCTI, Parallel Tempering, Replica Exchange, Hamiltonian Exchange, Simulated Annealing
  - Functional parallelism: Direct Force Evaluation, Reciprocal Space (SPME), Trajectory Collection
  - Communication parallelism: Communication hiding through ordering of computation based on available data
  - Loop level parallelism: Accelerator and/or threading implementations
- **Approach**
  - Domain decomposition (provides linked cell advantage)
  - Topology aware distribution of cell pairs to processes
  - Single communication process per multi-core node
  - Parallelization within the node through threads (OpenMP, pthreads) and/or accelerators (GPU)
  - Dedicated core for GA server thread

Pacific Northwest
NATIONAL LABORATORY

Technical Approach: Cell Pair Distribution

- **For a molecular system with 1,000,000 atoms, domain decomposed:**
  - I: Cells distributed over all processes
    - Ia: 10x10x10= 1,000 cells
    - Ib: 20x20x20= 8,000 cells
    - Ic: 30x30x30=27,000 cells
  - II: Cell-cell pairs distributed over processes
    - IIa: 10x10x10= 1,000 cells: 14,000 cell-cell pairs
    - IIb: 20x20x20= 8,000 cells: 504,000 cell-cell pairs
  - III: Cell-cell pairs distributed over processes
    - III: 10x10x10= 1,000 cells: with duplicate pairs for load balancing: 22,000 cell-cell pairs

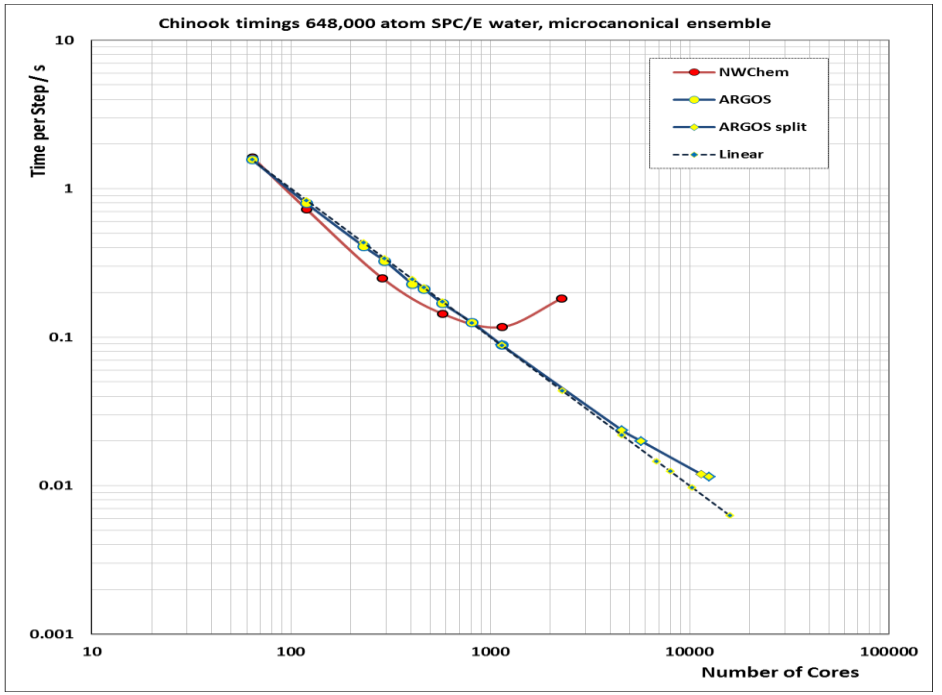
	Scenario Ia	Scenario Ib	Scenario Ic	Scenario IIa	Scenario IIb	Scenario III
Number of cells	1000	8000	27,000	1000	8000	1000
Number of cell pairs	14,000	504,000	9,261,000	14,000	504,000	22,000
Data moved per process / kB	610	364	297	94	12	94
Data moves per process	26	124	342	4	4	4
Total data moved / GB	0.6	2.8	7.6	1.2	5.5	1.9
Total data moves	$26*10^3$	$992*10^3$	$9,234*10^3$	$52*10^3$	$1,984*10^3$	$84*10^3$
Maximum number processes	1000	8000	27,000	14,000	504,000	22,000
Interactions per process	14,000,000	984,375	235,468	1,000,000	15,625	$\bar{N}$=636,636
Total interactions	$14,000*10^6$	$7,875*10^6$	$6,358*10^6$	$14,000*10^6$	$7,875*10^6$	$14,000*10^6$

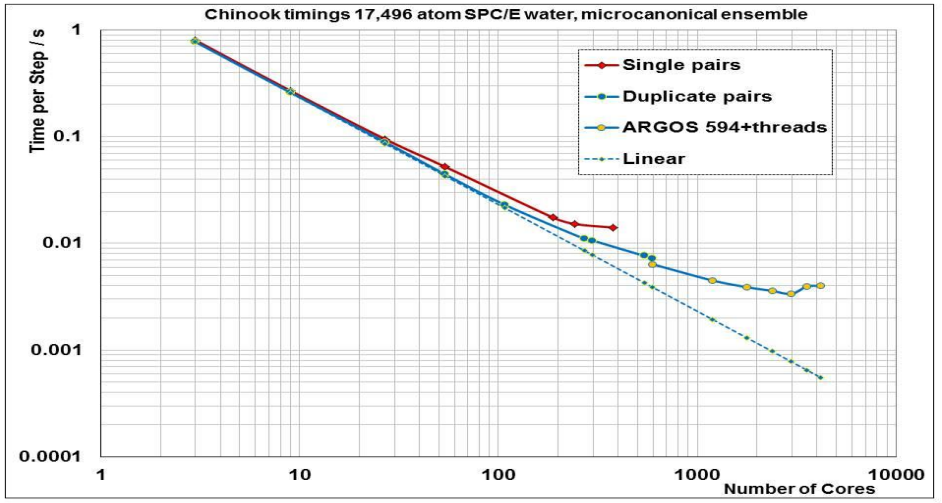
- Scenario Ia minimizes data moves and data moved, but is limited to <=1000 processes
- Scenario IIa,III minimizes data moves and data moved on >1000 processes

Pacific Northwest
NATIONAL LABORATORY

Current Progress in MD Parallel Performance

- Removal of explicit global synchronization from the basic MD time steps
- More effective hiding of communication latency and bandwidth through computation scheduling
- Improved scalability through cell-cell pair in stead of cell distribution
- Improved initial load balancing through duplication compute-intensive pairs in the cell-cell pair list
- Assignment of multiple cell pairs per process that minimizes communication
- Improved scalability by loop level parallelization using threading

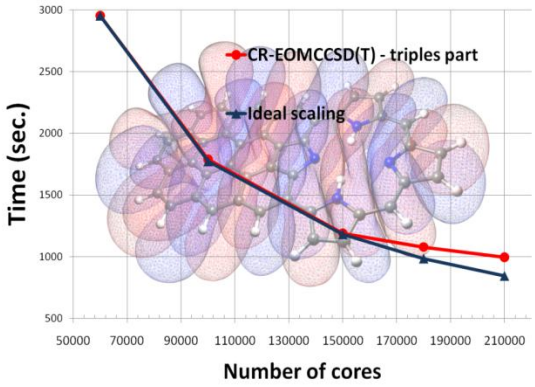




Summary: Computational Molecular Science

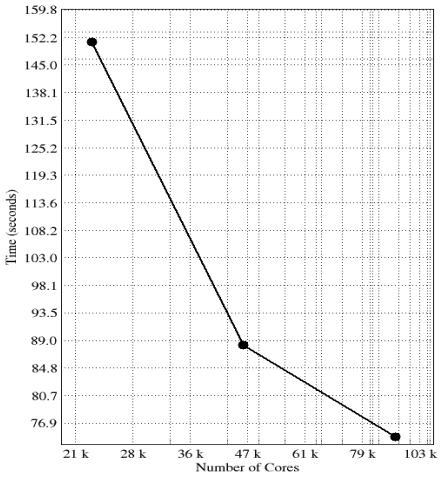
EOM-CCSD(T)

- Excellent scalability of the most computationally intensive (N^7) perturbative methods (triples)
- New parallel models for task scheduling of the iterative CC/ EOMCC methods
- Implementations of CC for GPU architectures
- Demonstrated fault tolerance of the CC implementation
- MR-CCSD(T) has potential to scale to exascale



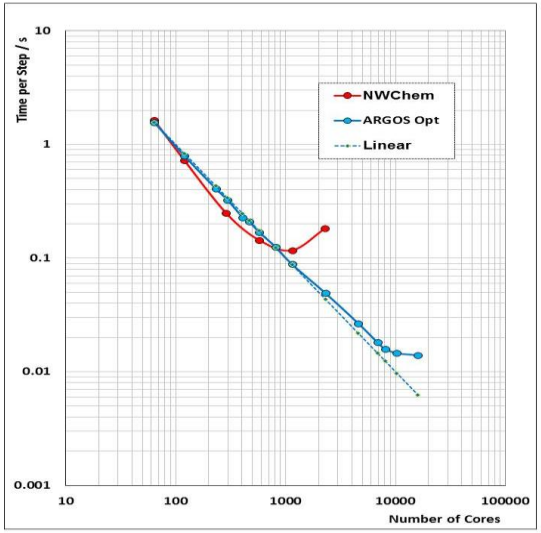
PWDFT

- Reduced communication algorithm
- Scalability to 100k cores
- Performance model for exact exchange algorithms
- Parallel in time algorithm for *ab initio* MD



MD

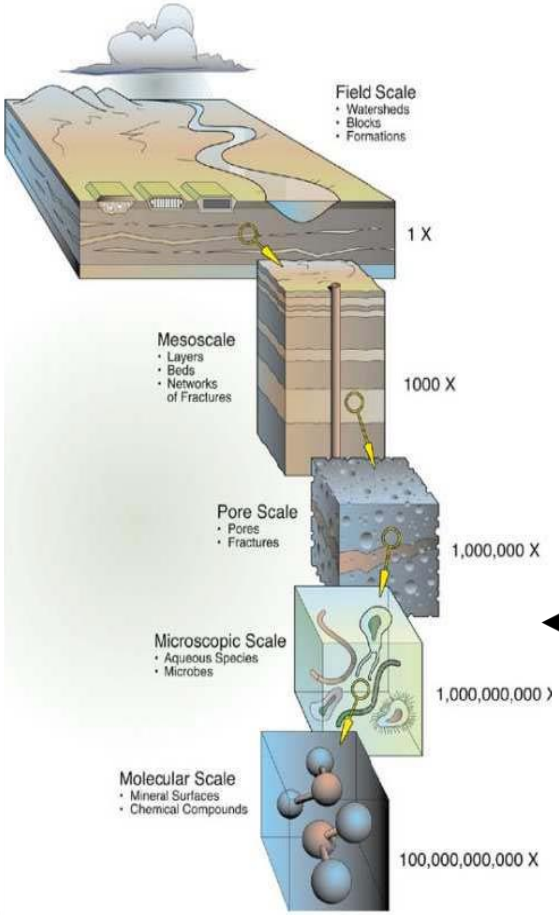
- Removal of explicit synchronization from the basic MD time steps
- New approach to more effectively hiding communication latency and bandwidth
- Increased scalability through data distribution by cell pairs



Pacific Northwest NATIONAL LABORATORY

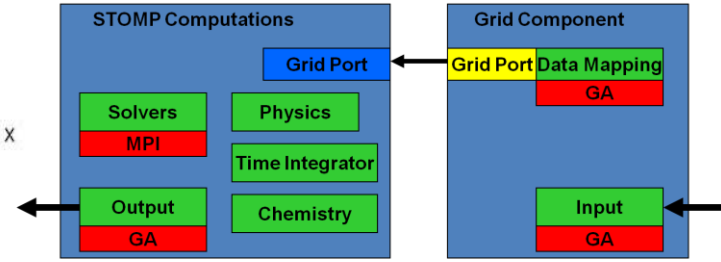
Computational Subsurface Modeling

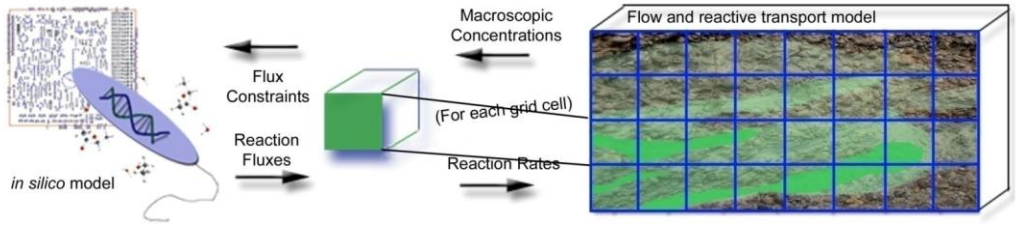
- **Predictive Modeling Capability**
  - Contaminant fate and transport
  - Carbon Sequestration
- **Current Technical Capabilities**
  - Hybrid multi-scale approach couples pore-scale and continuum (i.e. field scale)-scale models in a single simulation
  - Coupling of genome-scale models of microbial metabolism with reactive transport simulators
- **Computational complexities**
  - Subsurface science (ASCEM)
  - Parallel code development (XSCI-eSTOMP)
  - FY11 DOE ASCR Joule Metric (e-STOMP)



Grid component mapped onto GA

- Calculation modes not affected
- Support for irregular grids
- Multilevel parallelism for UQ
- GPU implementation of chemistry modes

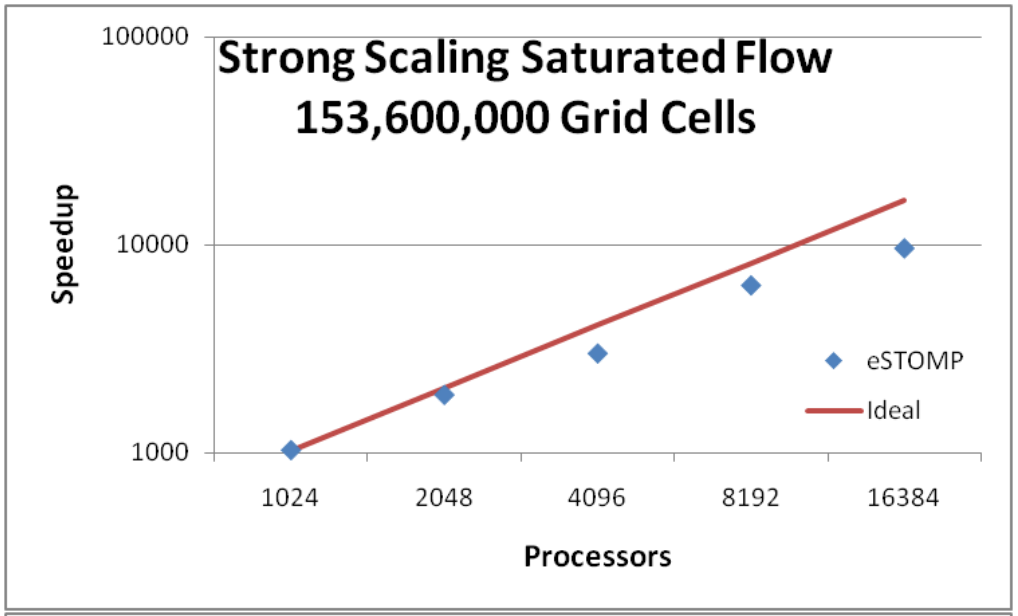


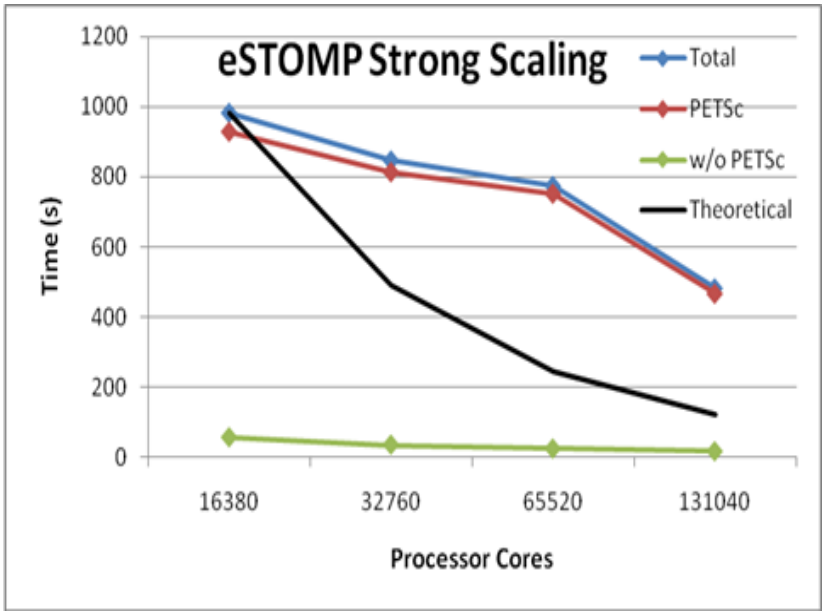


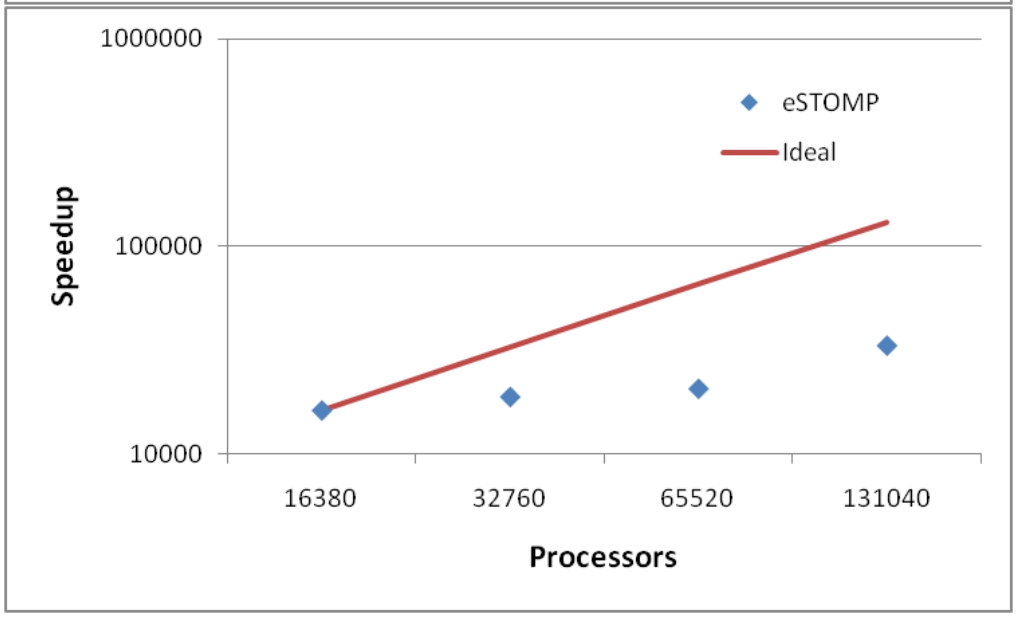
Pacific Northwest NATIONAL LABORATORY

eSTOMP: Single Phase Saturated Flow

- Efficient scaling of single phase flow and transport
  - Removal STOMP scaling bottlenecks
  - Remaining issues: solver scaling
- Demonstrated on
  - EMSL Chinook
  - NERSC Franklin
  - OLCF Jaguar







Acknowledgments

Principal and Key Investigators on the XSCI:

Kevin Barker
Eric Bylaska
Daniel Chavarría
Huub van Dam
Bert de Jong
Karol Kowalski
Sriram Krishnamoorthy
Manoj Krishnan
Bruce Palmer
Marat Valiev
Oreste Villa
Abhinav Vishnu
Steve Yabusaki

Other Development Teams:

NWChem
STOMP
CASS-MT
Performance Modeling
Global Arrays

Pacific Northwest
NATIONAL LABORATORY