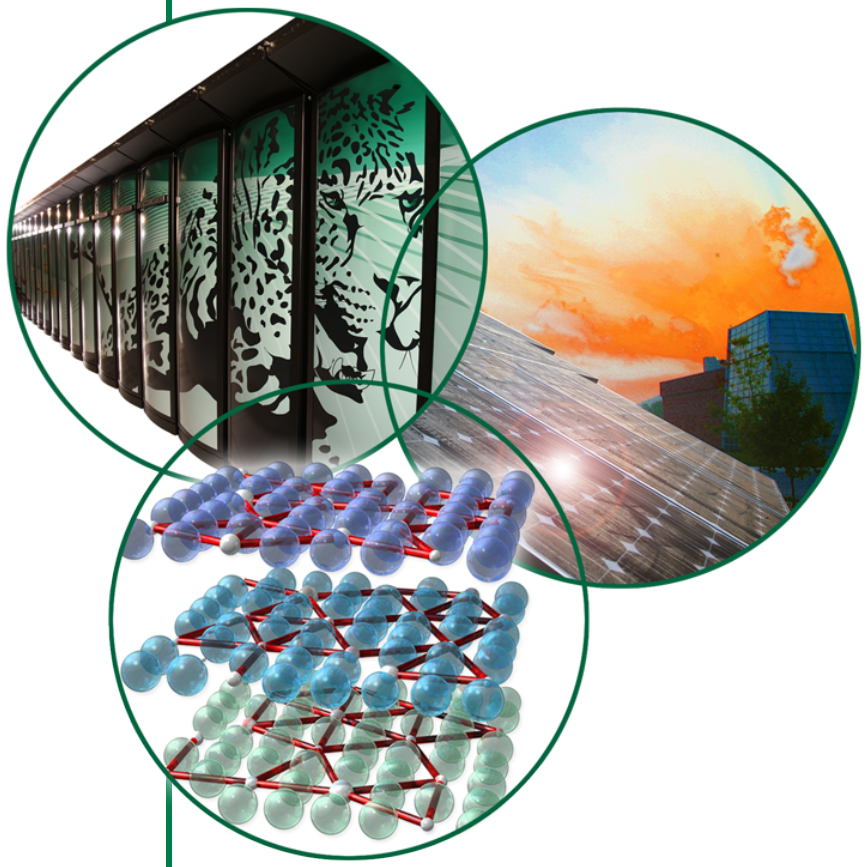
Molecular Dynamics Simulation of Protein Dynamics and Lignocellulosic Biomass (m906)

Loukas Petridis

Oak Ridge National Laboratory



U.S. DEPARTMENT OF ENERGY

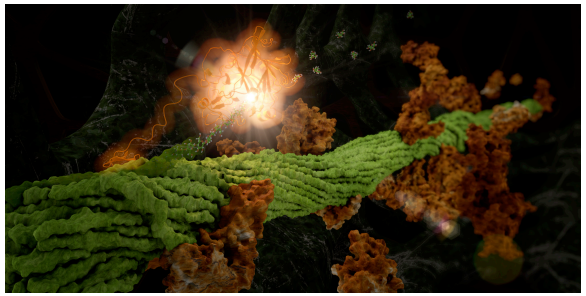
OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

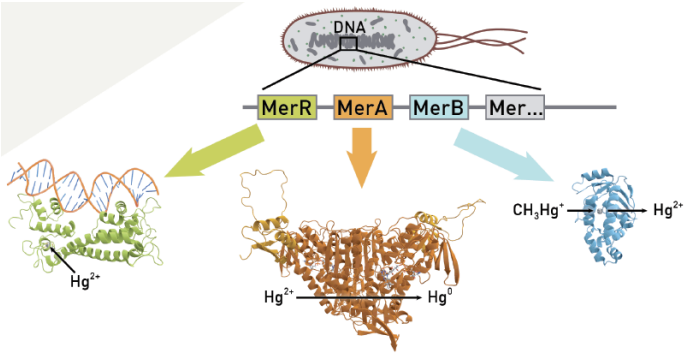
1. Project Description

PI: Jeremy C. Smith (ORNL)

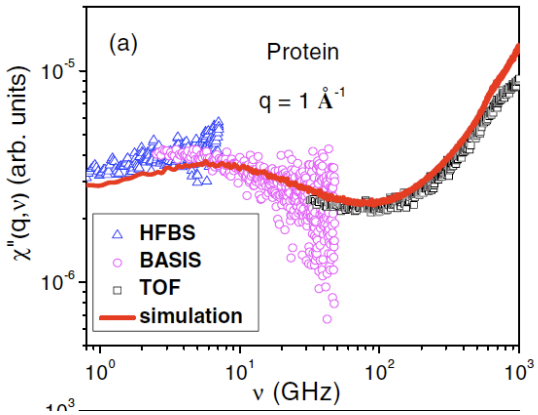
Insight into physical processes leading to biological function in critical research missions:



Bioenergy

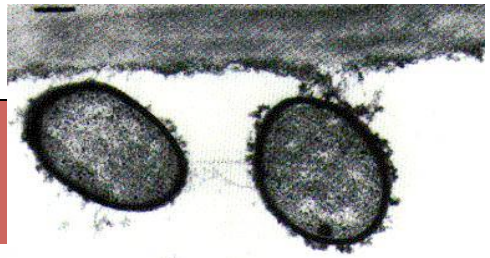


Bioremediation



Neutron scattering

Relevant time and lengthscales



1μm

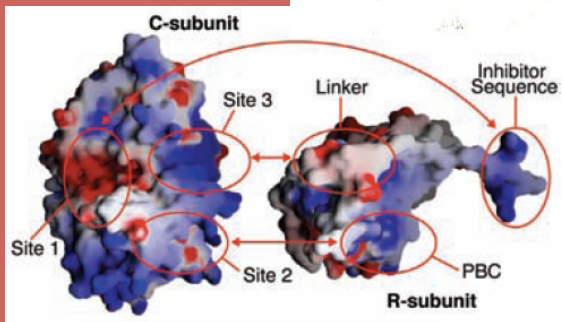
100 *nm*

1 *nm*

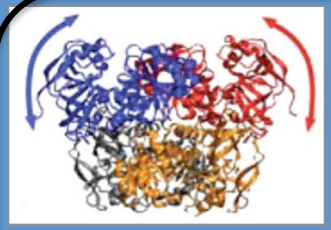
0.1 *nm*

NERSC 8

multi-component

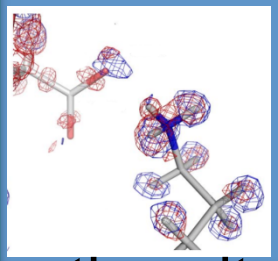


Hopper



domain motions

structural transitions



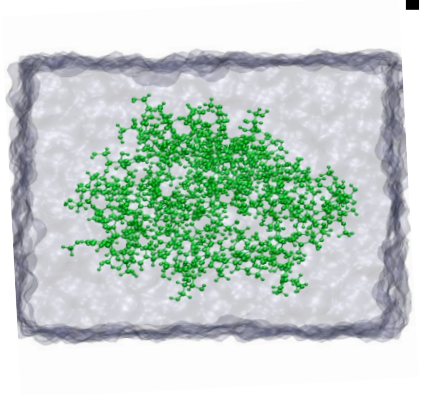
active site

local dynamics

1 *fs* 1 *ps* 1 *ns* **1 µs**

2. Computational Strategies

Molecular Dynamics Simulation



time t: coord & vel

Potential → **Energy**

force

Integrate → **Eq. motion**

time t+Δt: coord & vel

Codes: GROMACS and NAMD

Algorithms:

Integration of equations of motions: verlocity Verlet.

N-body algorithms with neighborlist.

Domain and force decomposition for multi-level parallelization.

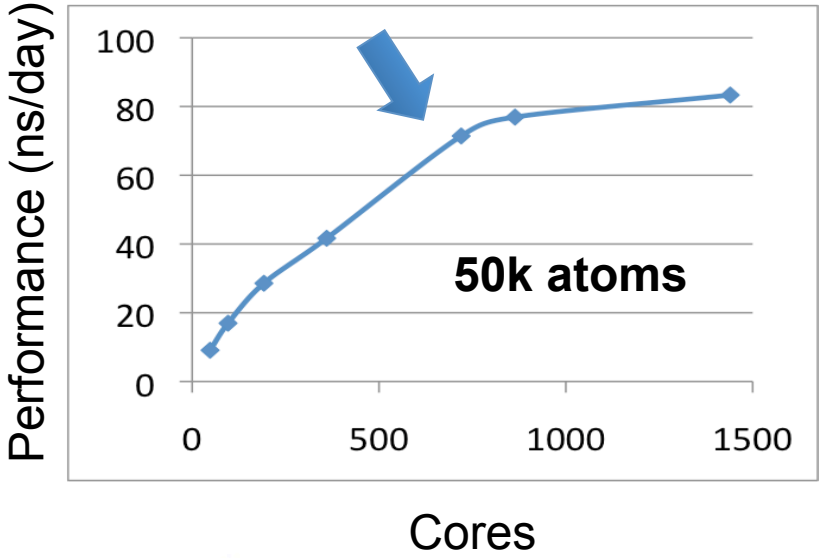
Grid-based electrostatics: Particle Mesh Ewald (FFT).

} network communication

Computational Challenges for Capacity-Class Simulations

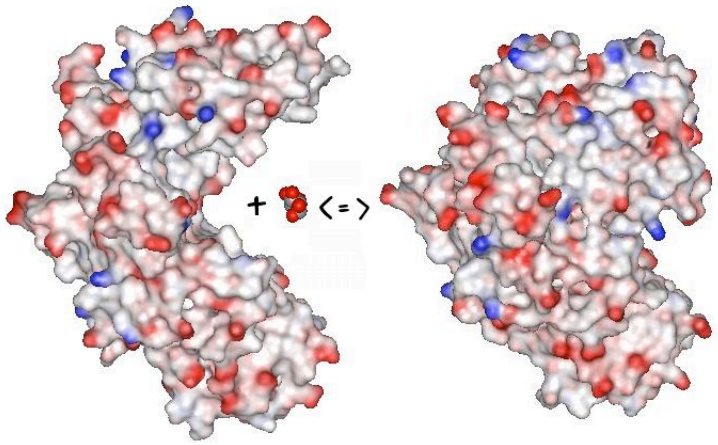
1. Biggest computational challenge: strong scaling is limited by network latency.

➔ limits accessible timescales

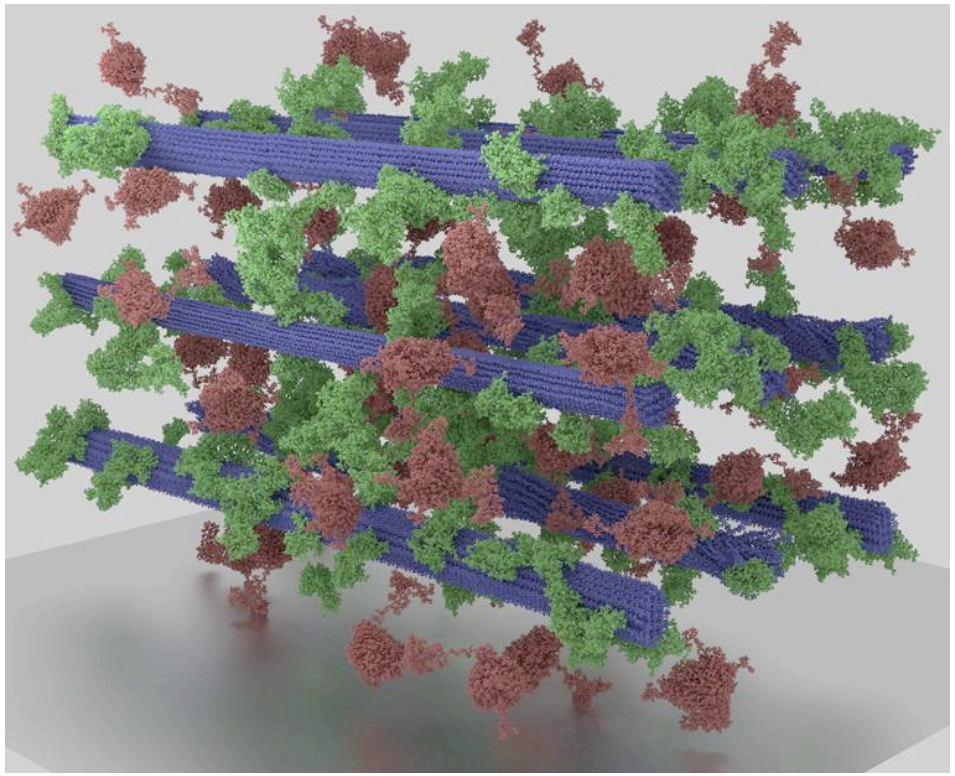


2. Ensemble-based methods to enhance sampling of structural transitions substantial increase in computational cost

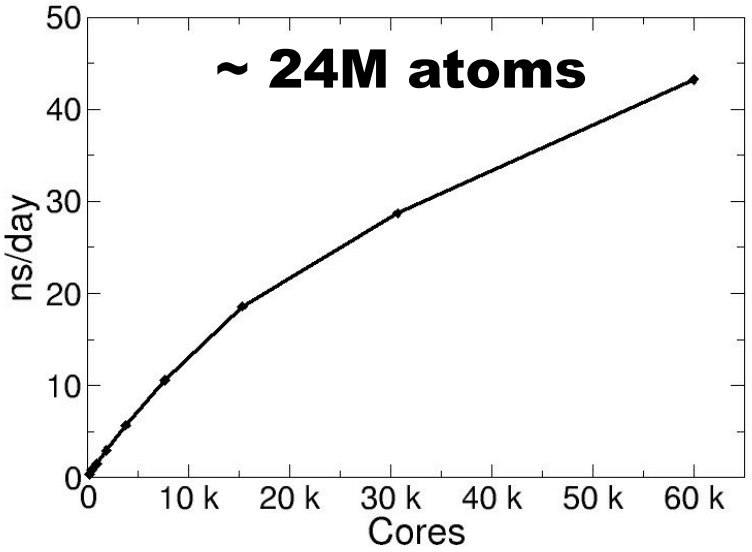
➔ currently not applicable in multi-component systems

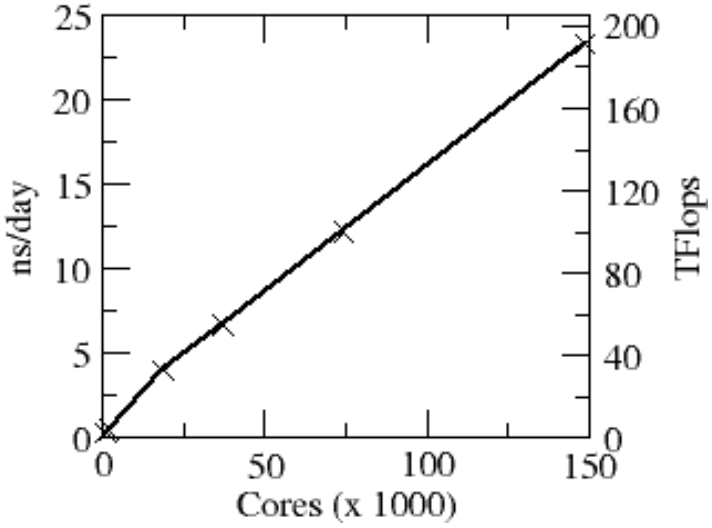


(Capability-Class Simulations)



24M-atom simulation of enzyme binding to pretreated lignocellulose. 2012: 50 M INCITE hours on OLCF Jaguar XT5





~ 100M atoms

3. Current HPC Usage

- **Hopper: 7.5M CPU hours in 2012 on single-molecule simulations**

Cores	200-2,000
Checkpoint data	1.2 to 24 MB
Checkpoint bandwidth	2 MB/sec
Data I/O*	0.2 – 2 Gb
I/O bandwidth	500 kB/sec
Project directory space	1TB
Archival data	10 TB
Memory per node	~MB

system size: ~50k-500k atoms

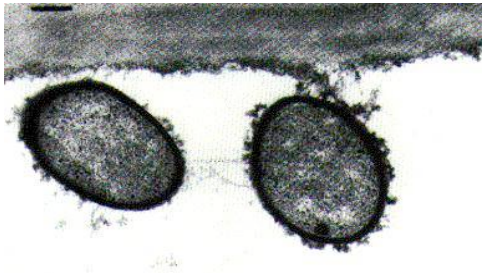
~10k one-hour runs per year

*based on 1 hour run

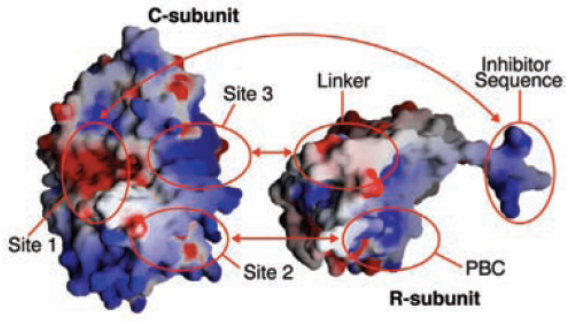
4. HPC Requirements for 2017

- **60 M hours.**

- **Parallel concurrency increase x10:**



extend lengthscales: biomass-microbe interactome



ensemble methods in complexes systems: signaling pathways

- **Modest memory requirements (~10MB) and I/O bandwidth ~5MB/sec**
- **~10k files, size ~ 10 MB – 10 GB**

4. HPC Requirements for 2017

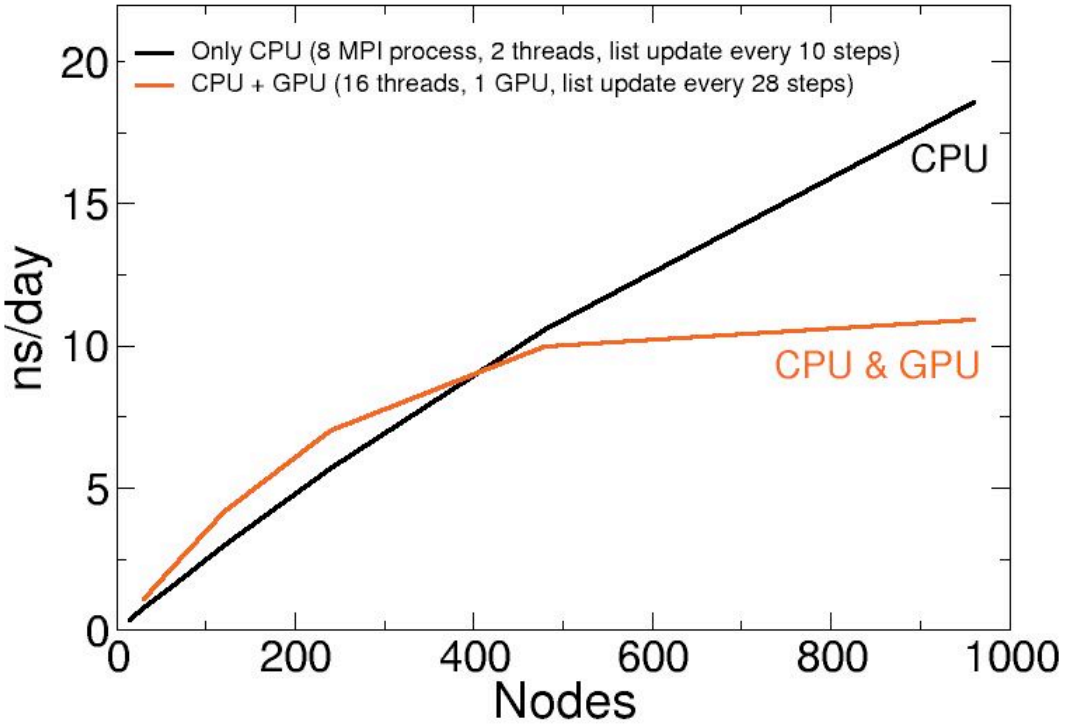
Cores	2,000-20,000
Checkpoint data	12 to 120 MB
Checkpoint bandwidth	20 MB/sec
Data I/O*	2 – 20 Gb
I/O bandwidth	5 MB/sec
Project directory space	2TB
Archival data	20 TB
Memory per node	~MB

write less frequently

Applications: Gromacs, NAMD, VMD
Development: C++, Boost, libxml, Cmake, Git, FFTW, Cuda (or equivalent), Eclipse/PTP

5. Strategies for New Architectures

- MD codes already compatible with hybrid CPU/GPU architectures



TitanDEV
(ORNL)

- Currently two CPUs as fast as CPU+GPU. By 2017 expect GPUs will provide 2x speedup.
- Implicit solvent calculations potentially benefit from GPU

5. Summary

- NERSC8 ➔ new biology
  - Increase in accessible lengthscales
  - Ensemble-based methods for multi-component complexes.
- Recommendations on NERSC architecture:
  - Vendors: tightly integrated (ideally shared cache) CPU + GPU enabling fine-grain split of workload between CPU and GPU.
  - Strong scaling of MD simulations is limited by network latency: small cluster with lowest possible network latency.
  - Simulations are globally synchronized, therefore the lowest network connection is slowing down the entire simulation. A task placement that ensures nodes can communicate with low network latency.