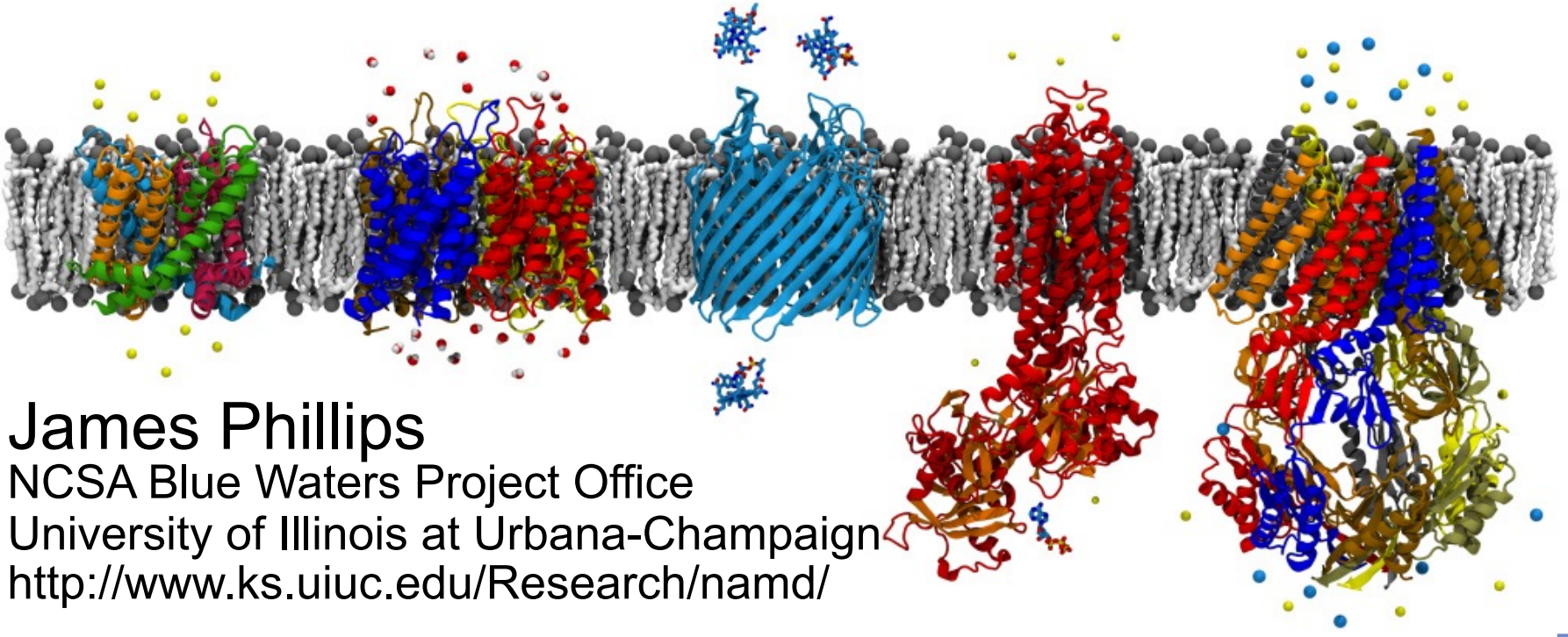
Experiences with Charm++ and NAMD on the Summit POWER9/Volta Supercomputer



James Phillips
NCSA Blue Waters Project Office
University of Illinois at Urbana-Champaign
http://www.ks.uiuc.edu/Research/namd/

The Blue Waters Project

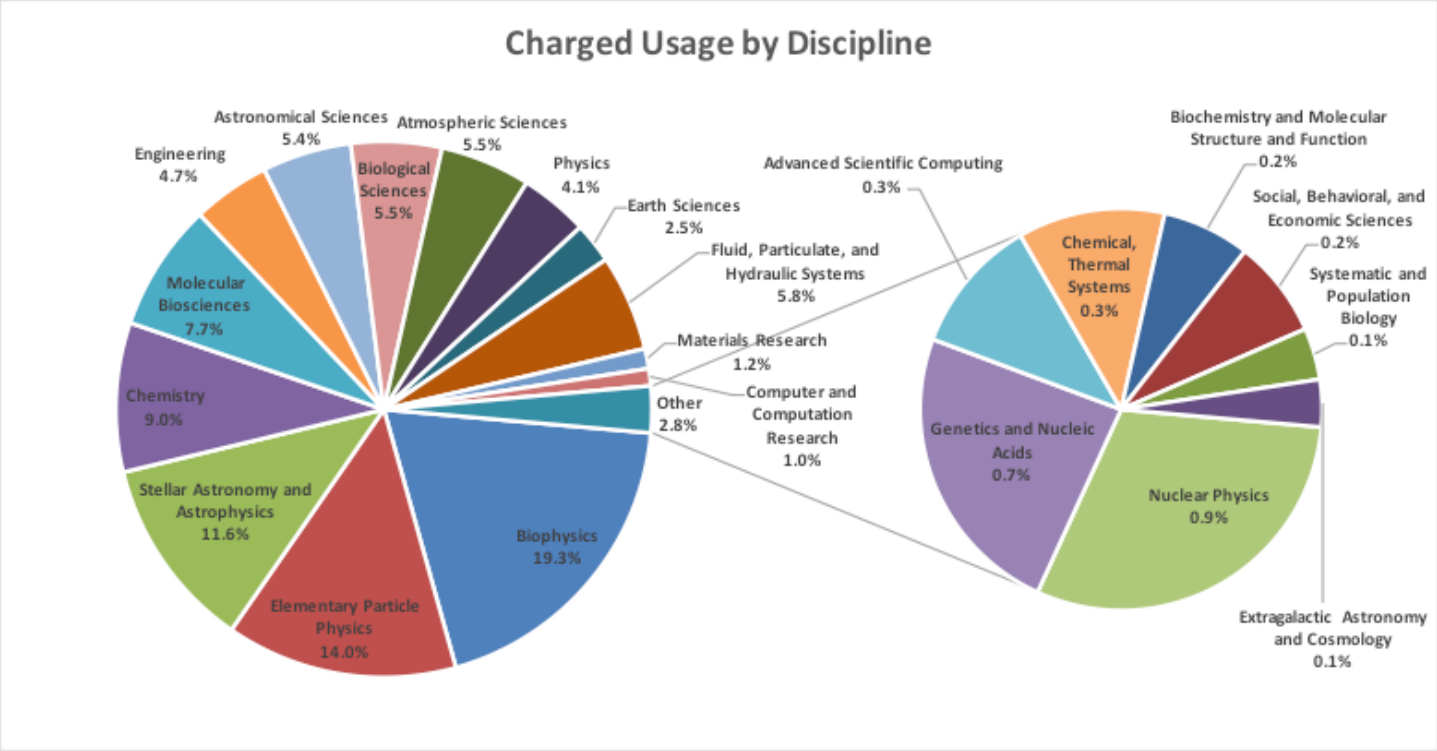
- Comprehensive development, deployment and service phases with co-design etc.
- The Blue Waters system is a top ranked system in all aspects of its capabilities.
- Diverse Science teams are able to make excellent use of those capabilities due to the system's flexibility and emphasis on sustained performance.
  - 45% larger than any system Cray has ever built
  - 22,640 CPU-only nodes, 4,224 GPU-accelerated nodes
  - Peak performance and delivered cycles are approximately the same as the aggregate of all the NSF XSEDE resources.
  - Ranks in the top systems in the world in peak performance – despite being over five years old
  - Largest memory capacity (1.66 PetaBytes) of any HPC system in the world! One of the fastest file systems (>1 TB/s) in the world!
  - Largest certified nearline tape system (>250 PB) in the world
  - Fastest external network capability (>420 Gb/s) of any open science site.

Blue Waters 2.0 ?

Towards a Leadership-Class Computing Facility - Phase 1

- “robust, well-balanced, and forward-looking computational asset for a broad range of research topics for which advances in fundamental understanding require the most extreme computational and **data analysis** capabilities”
- “at least two- to three-fold **time-to-solution** performance improvement over the current state of the art, the University of Illinois at Urbana-Champaign's (UIUC) Blue Waters system, for a broad range of existing and emerging computational and **data intensive** applications;”
- “scientific and technical evaluation of the Phase 1 system that will lead to an upgrade design of a leadership-class system, called the Phase 2 system,”
- “the Phase 2 system is expected to have a ten-fold or more time-to-solution performance improvement over the Phase 1 system;”
- See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503148

All comments are personal views only and do not represent NCSA.



Data From Blue Waters 2015-2016 Annual Report

NIH Biomedical Technology Research Center for Macromolecular Modeling and Bioinformatics

Developers of the widely used computational biology software **VMD** and **NAMD**

250,000 registered **VMD** users
80,000 registered **NAMD** users

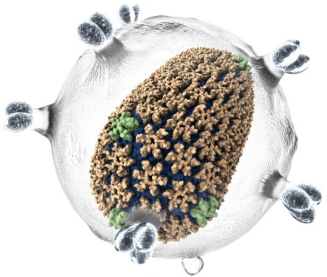
600 publications (since 1972)
over **54,000** citations

Perfect score (10.0) on 2017-2022 NIH renewal

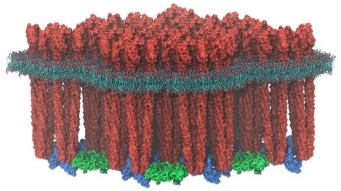
research projects include: virus capsids, bacteria, molecular motors, neurons and synapses, membrane transporters, bioenergetic membranes

4 faculty members
8 developers
1 systems administrator
17 postdocs
46 graduate students
2 administrative staff

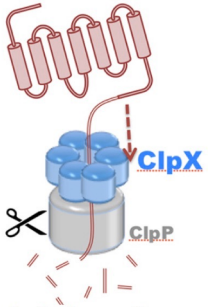
NIH Center Driving Projects 2017-2022



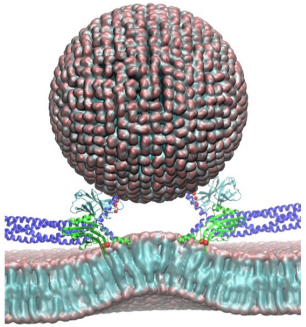
Viral Infection



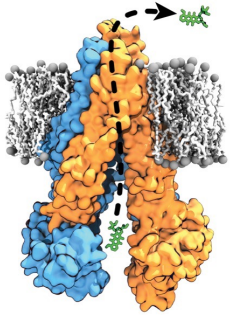
Symbiont Bacteria



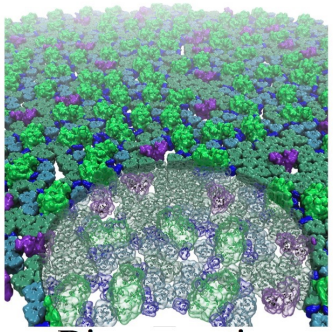
Molecular Motors



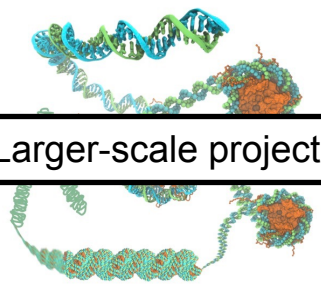
Neurons and Synapses



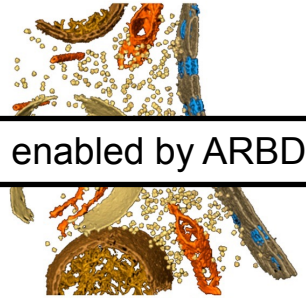
Membrane Transporters



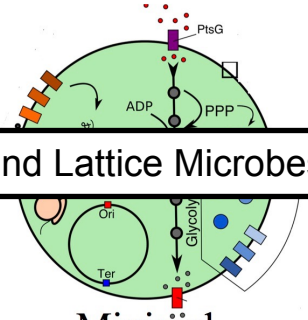
Bioenergetic Membranes



Chromatin



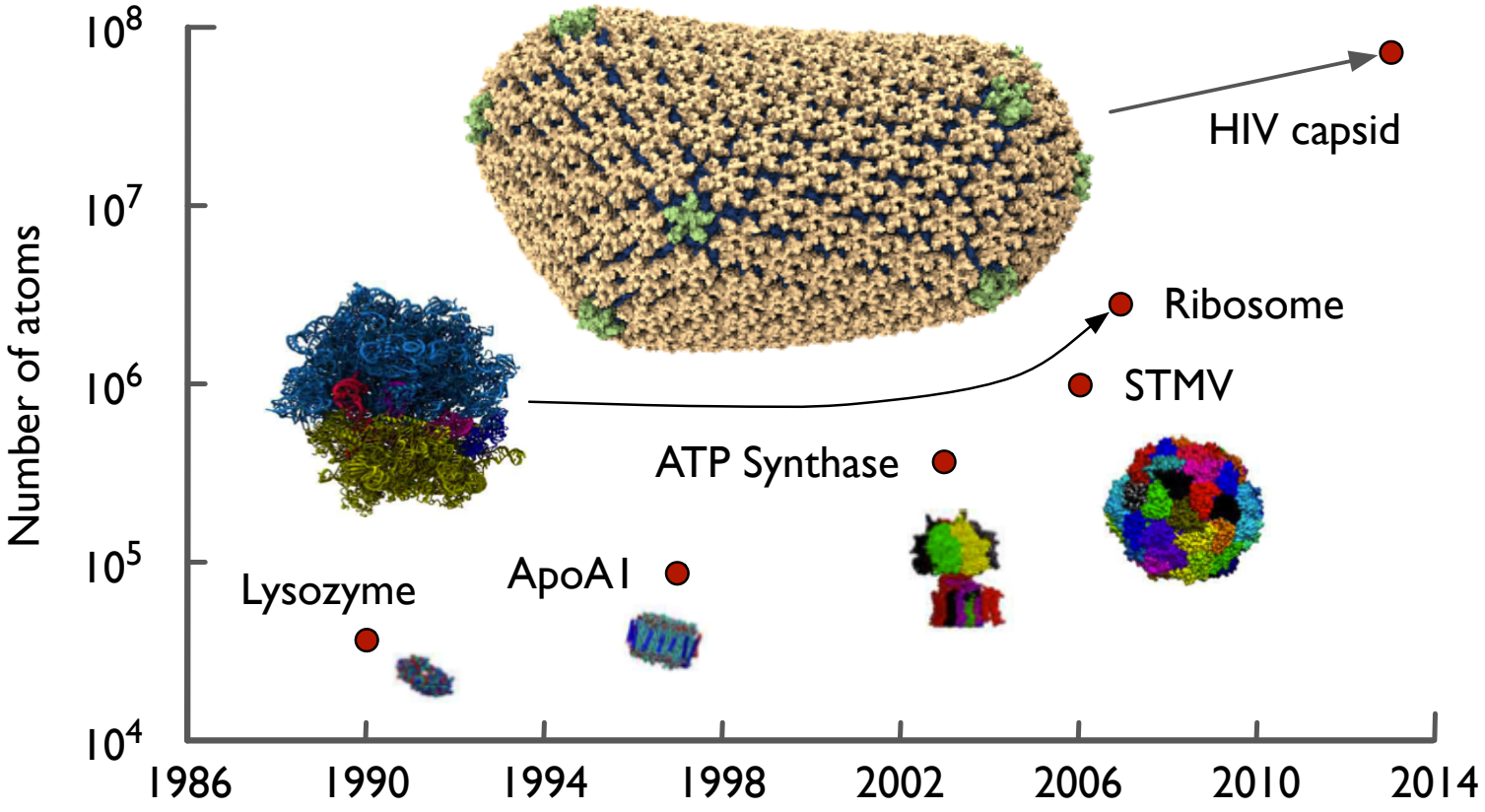
Bacterial & Eukaryal Systems



Minimal Cell

Larger-scale projects enabled by ARBD and Lattice Microbes

Need for petascale: Simulation follows structural discovery



NAMD: Practical Supercomputing for Biomedical Research

"**widest-used application**" on NCSA Blue Waters, NSF-specified benchmark for successor machine

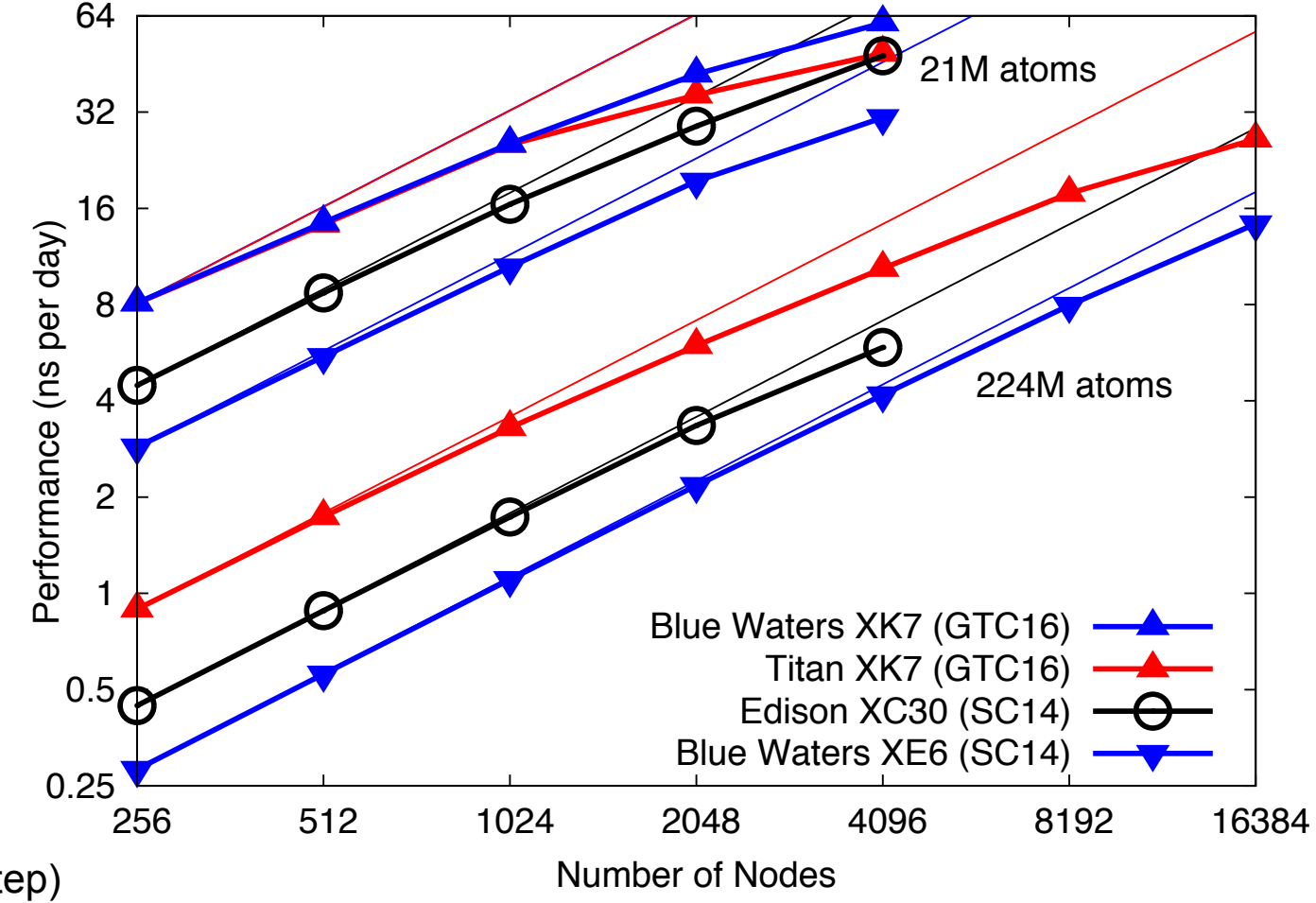
"**by a very large margin the most used code**" at Texas Advanced Computing Center (2nd largest)

Early adopters of workstation clusters (1993), Linux clusters (1998), and CUDA (**2007**).

Application readiness/early science projects on
- Argonne Theta (10 PF Cray KNL, completed)
- Oak Ridge Summit (200 PF Power9/Volta, 2018)
- ~~Argonne Aurora (200 PF Cray KNH, 2019)~~
- Argonne Aurora (1 EF Intel ???, 2021)

"For outstanding contributions to the development of widely used parallel software for large biomolecular systems simulation"

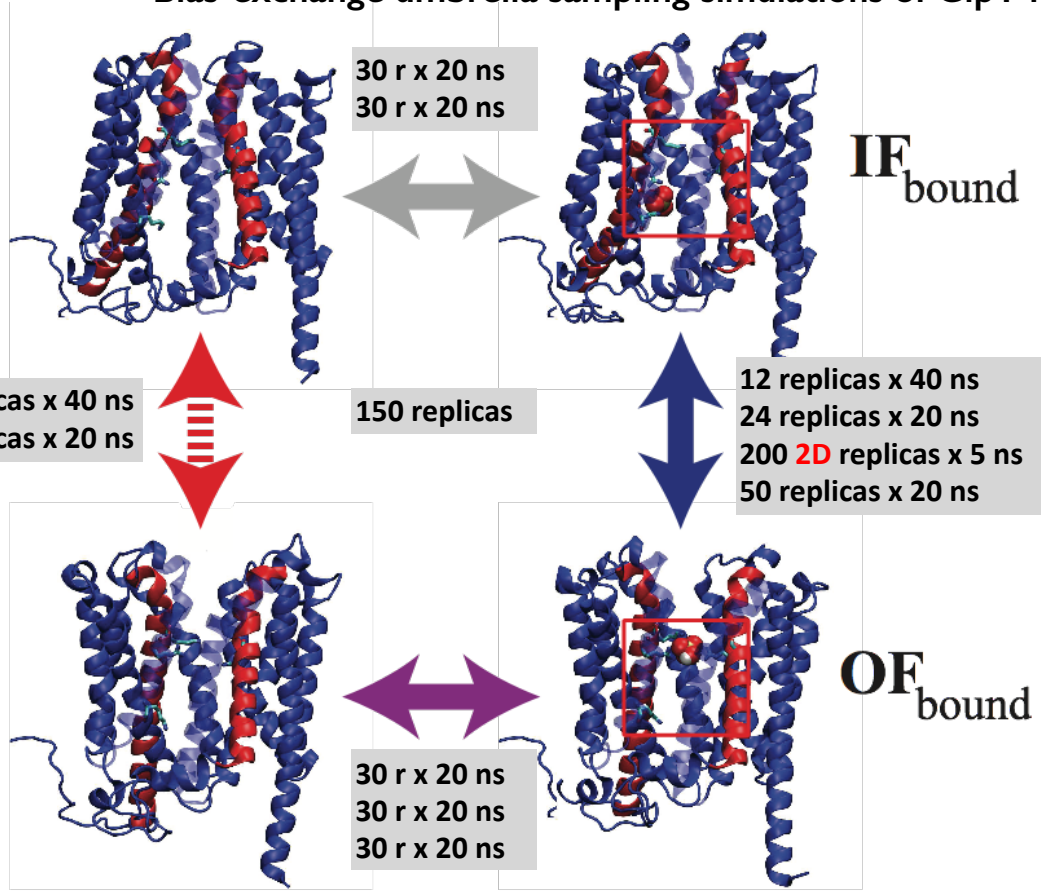
NAMD Runs Large Simulations Well

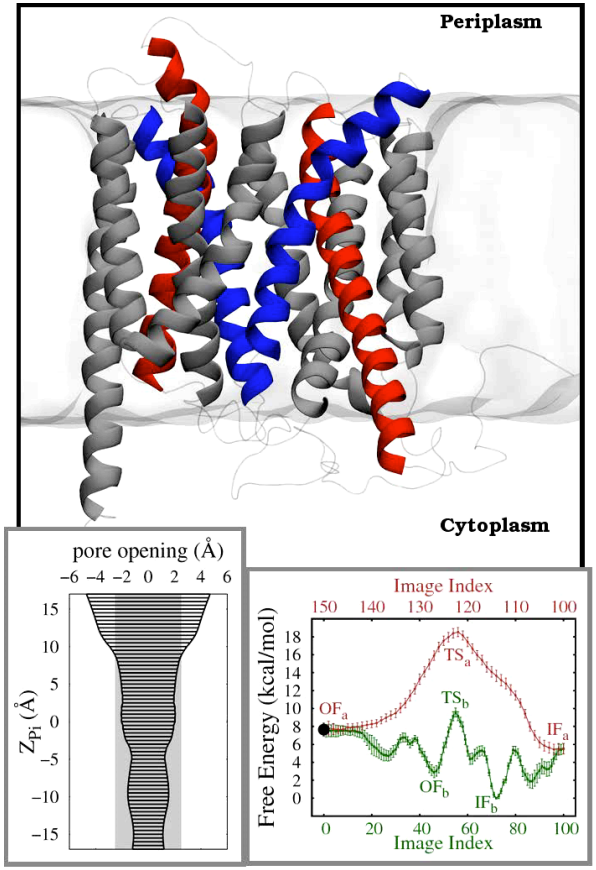


(2fs timestep)

Multi-copy methodologies enable study of millisecond processes

Bias-exchange umbrella sampling simulations of GlpT membrane transporters



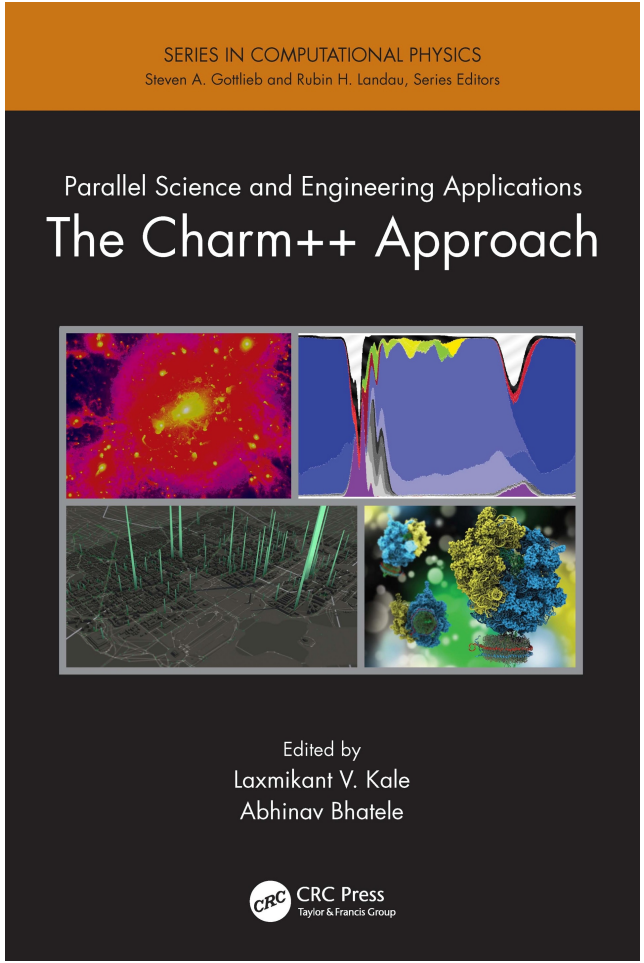


M. Moradi, G. Enkavi, and E. Tajkhorshid, *Nature Communications* **6**, 8393 (2015)

NAMD is based on Charm++

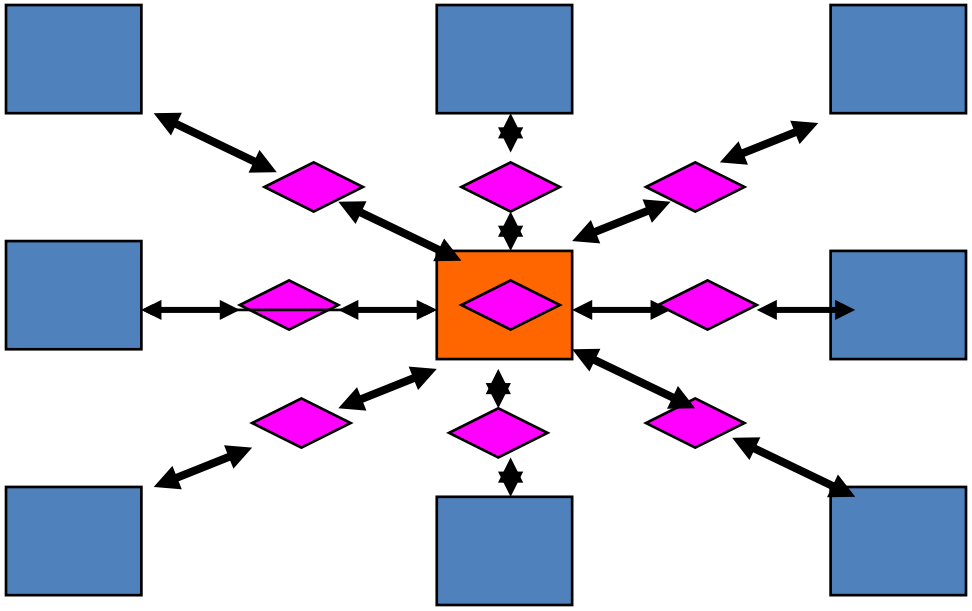
- Parallel C++ with *data driven* objects.
- Asynchronous method invocation.
- Prioritized scheduling of messages/execution.
- Measurement-based load balancing.
- Portable messaging layer.

Complete info at charmplusplus.org and charm.cs.illinois.edu



NAMD Hybrid Decomposition

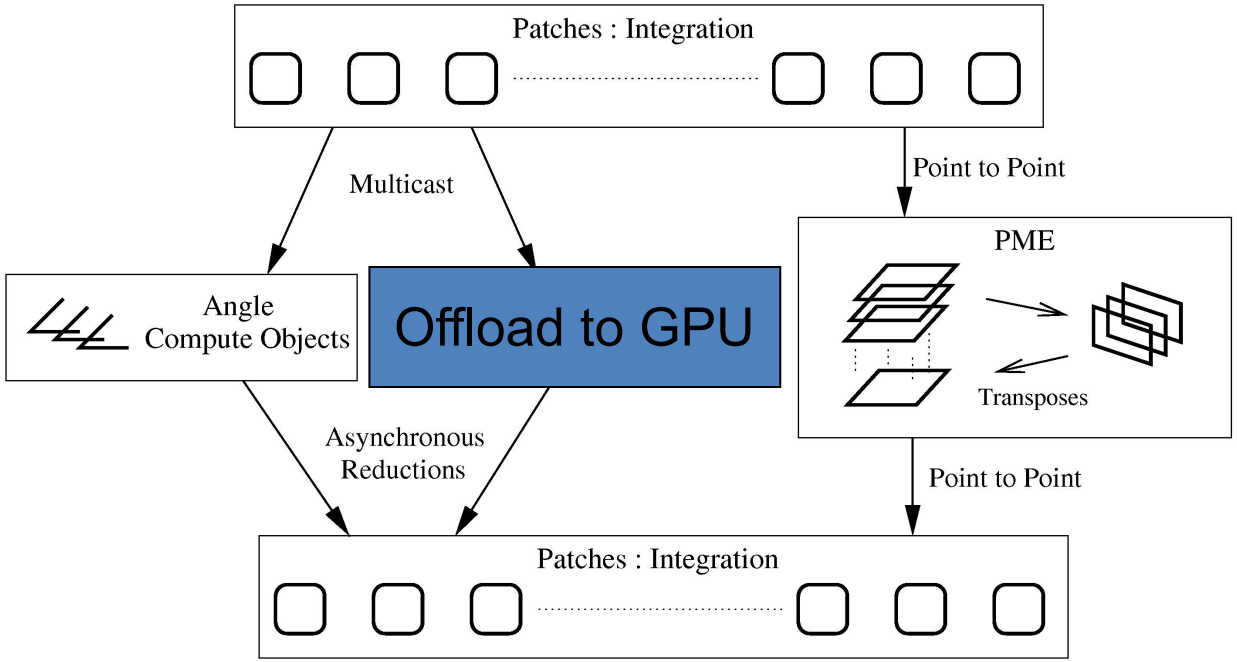
Kale *et al., J. Comp. Phys.* 151:283-312, 1999.



- Spatially decompose data and communication.
- Separate but related work decomposition.
- “Compute objects” facilitate iterative, measurement-based load balancing system.

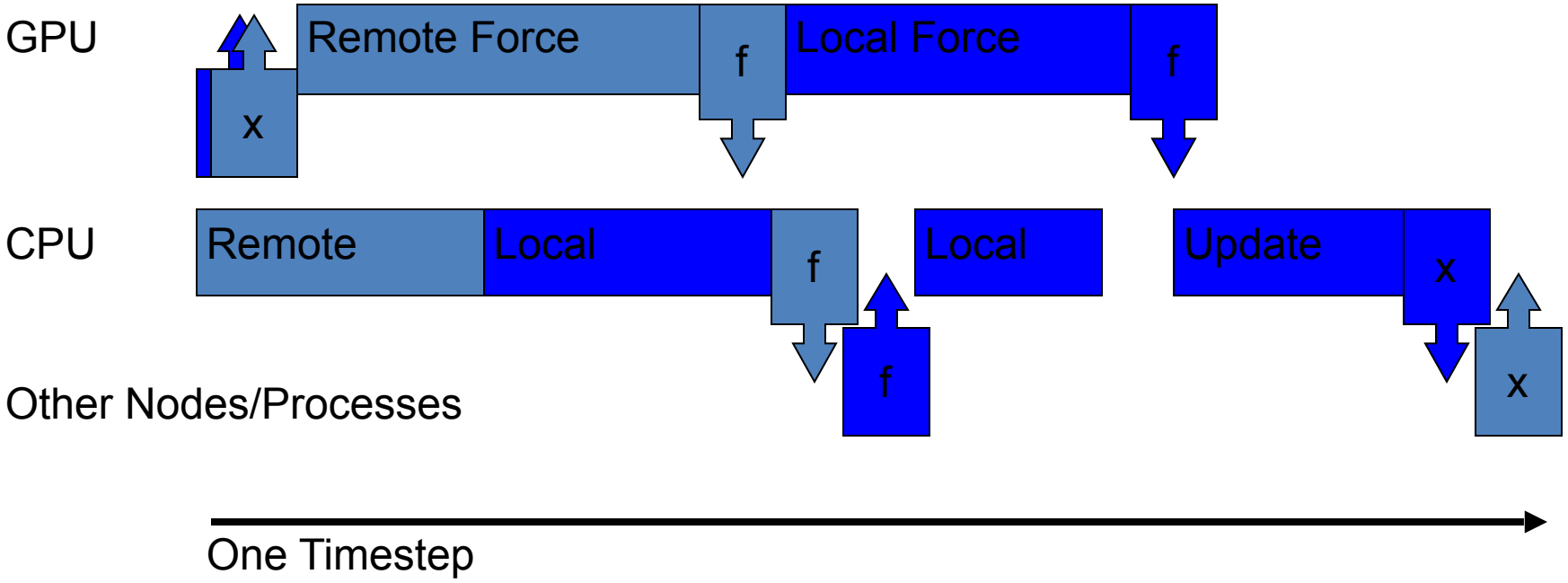
NAMD Overlapping Execution

Phillips *et al., SC2002*.



Objects are assigned to processors and queued as data arrives.

Overlapping GPU and CPU with Communication

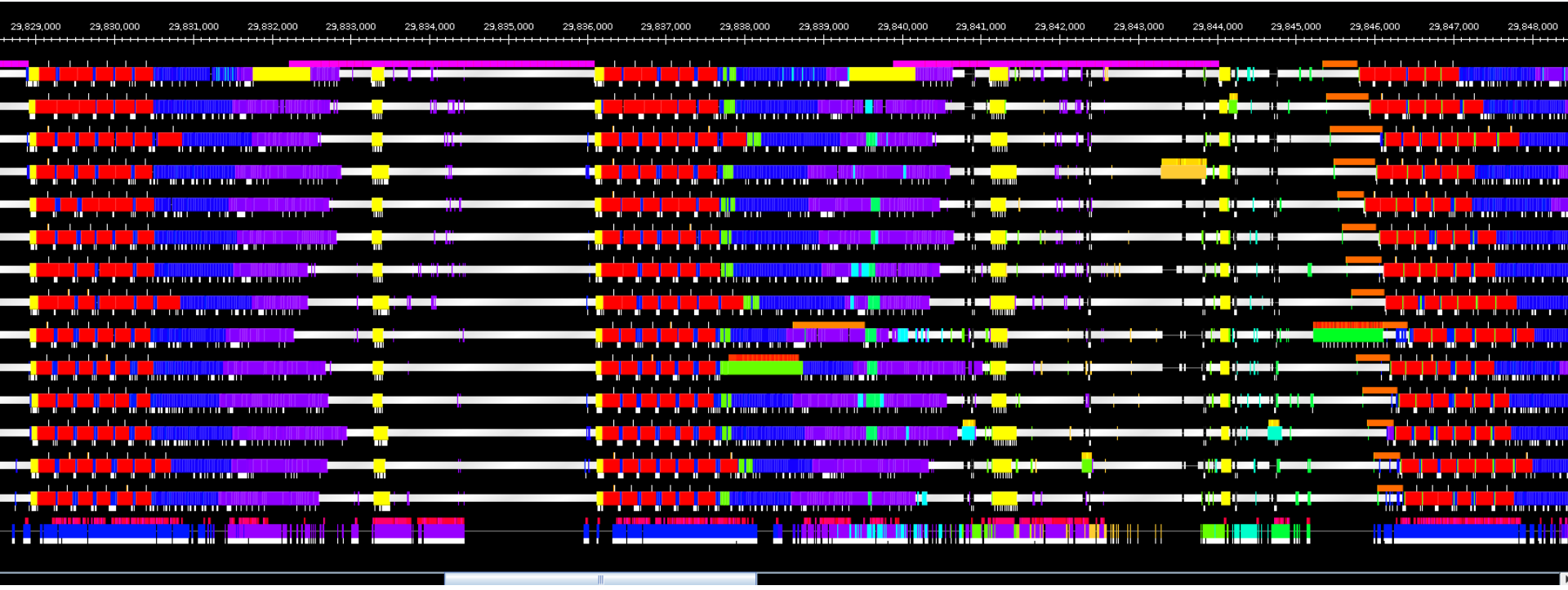


Phillips *et al.*, SC2008

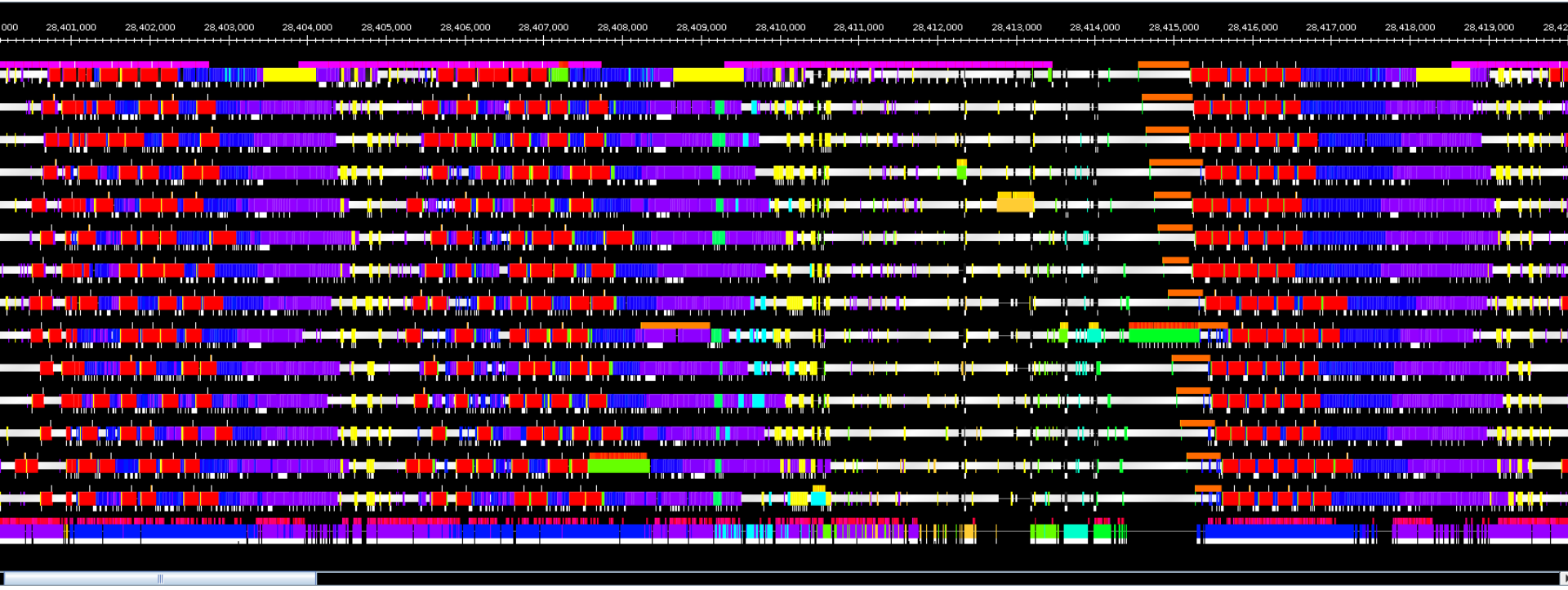
Streaming CPU Results to CPU

- Allows incremental results from a single grid to be processed on CPU before grid finishes on GPU
- Allows merging and prioritizing of remote and local work
- GPU side:
  - Write results to host-mapped memory (also without streaming)
  - __threadfence_system() and __syncthreads()
  - Atomic increment for next output queue location
  - Write result index to output queue
- CPU side:
  - Poll end of output queue (int array) in host memory

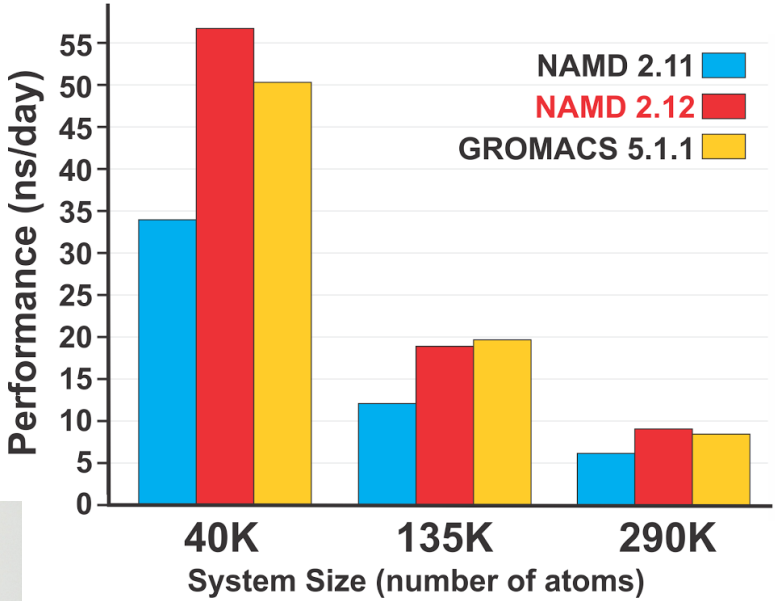
Non-Streaming Kernel

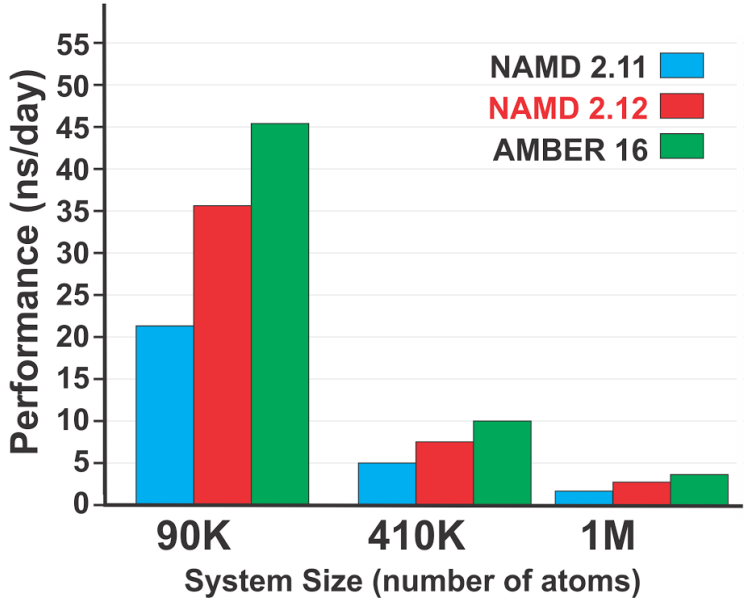


Streaming Kernel



Single-Node GPU Performance Optimization





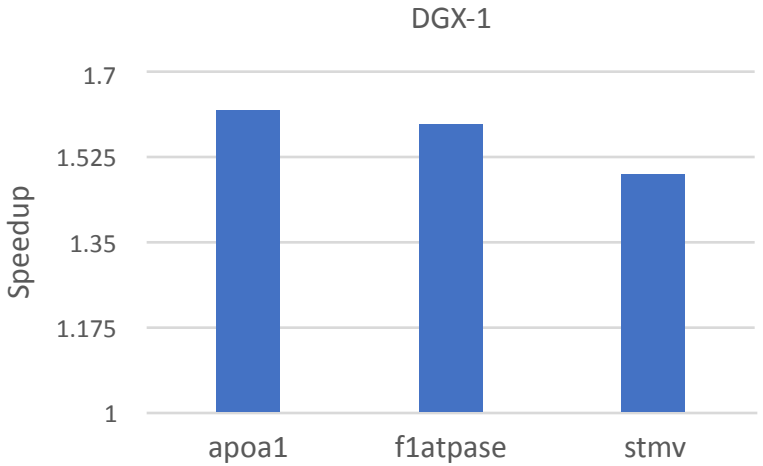
New kernels by **Antti-Pekka Hynninen**, formerly Oak Ridge, **NVIDIA**.

Stone, Hynninen, et al., *International Workshop on OpenPOWER for HPC (IWOPH'16)*, 2016.

Described at GTC 2016 **S6623** - Advances in NAMD GPU Performance

Coming in NAMD 2.13: Bonded force offloading

- GPU offloading for bonds, angles, dihedrals, impropers, exclusions, and crossterms
- Computation in single precision
- Forces are accumulated in 24.40 fixed point
- Virials are accumulated in 34.30 fixed point
- Code path exists for double precision accumulation on Pascal and newer GPUs
- **Reduces CPU workload and hence improves performance on GPU-heavy systems**



New kernels by **Antti-Pekka Hynninen, NVIDIA.**

Summit will replace Titan as the OLCF's leadership supercomputer

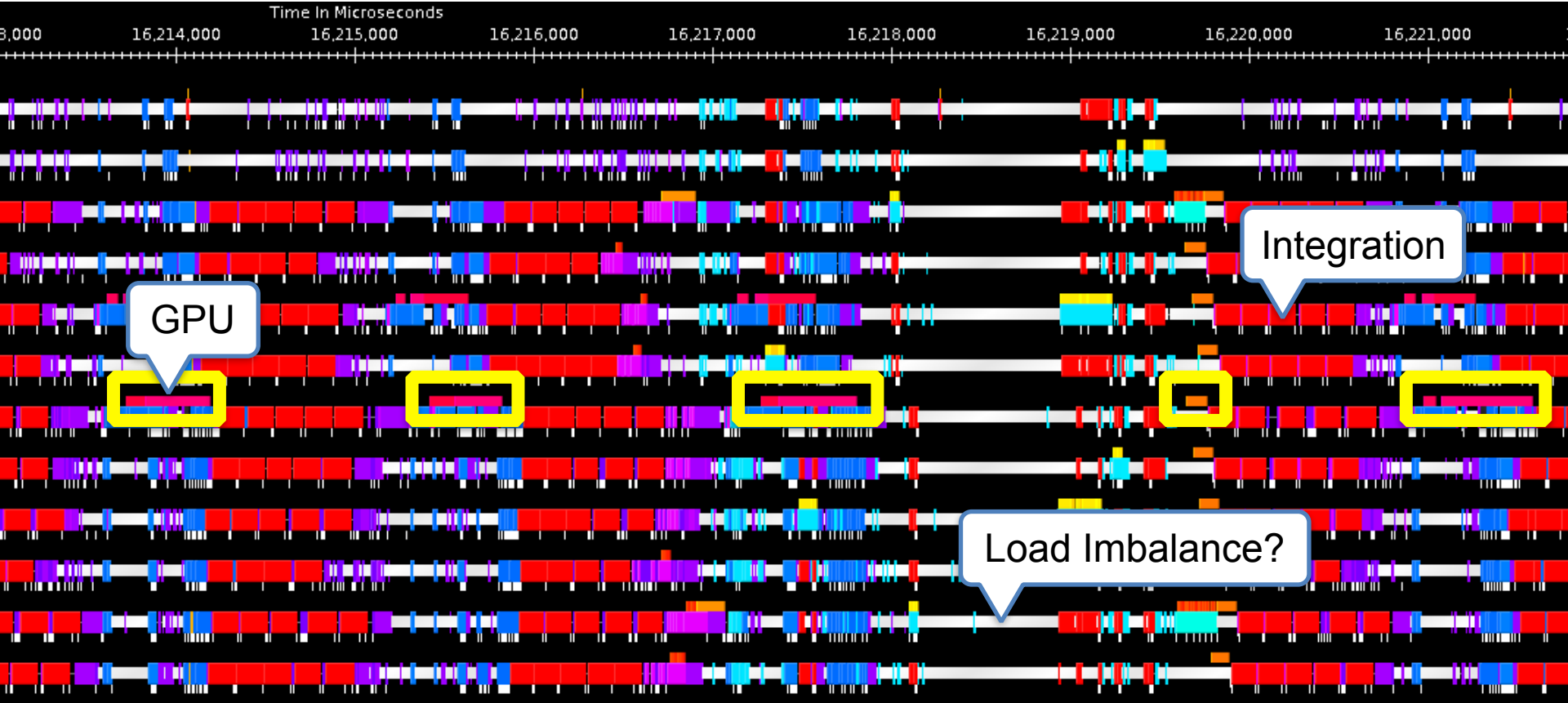
- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and total system memory
- Faster interconnect
- Much higher bandwidth between CPUs and GPUs
- Much larger and faster file system

Feature	Titan	Summit
Application Performance	Baseline	5-10x Titan
Number of Nodes	18,688	~4,600
Node performance	1.4 TF	> 40 TF
Memory per Node	32 GB DDR3 + 6 GB GDDR5	512 GB DDR4 + HBM
NV memory per Node	0	1600 GB
Total System Memory	710 TB	>10 PB DDR4 + HBM + Non-volatile
System Interconnect (node injection bandwidth)	Gemini (6.4 GB/s)	Dual Rail EDR-IB (23 GB/s)
Interconnect Topology	3D Torus	Non-blocking Fat Tree
Processors	1 AMD Opteron™ 1 NVIDIA Kepler™	2 IBM POWER9™ 6 NVIDIA Volta™
File System	32 PB, 1 TB/s, Lustre®	250 PB, 2.5 TB/s, GPFS™
Peak power consumption	9 MW	15 MW

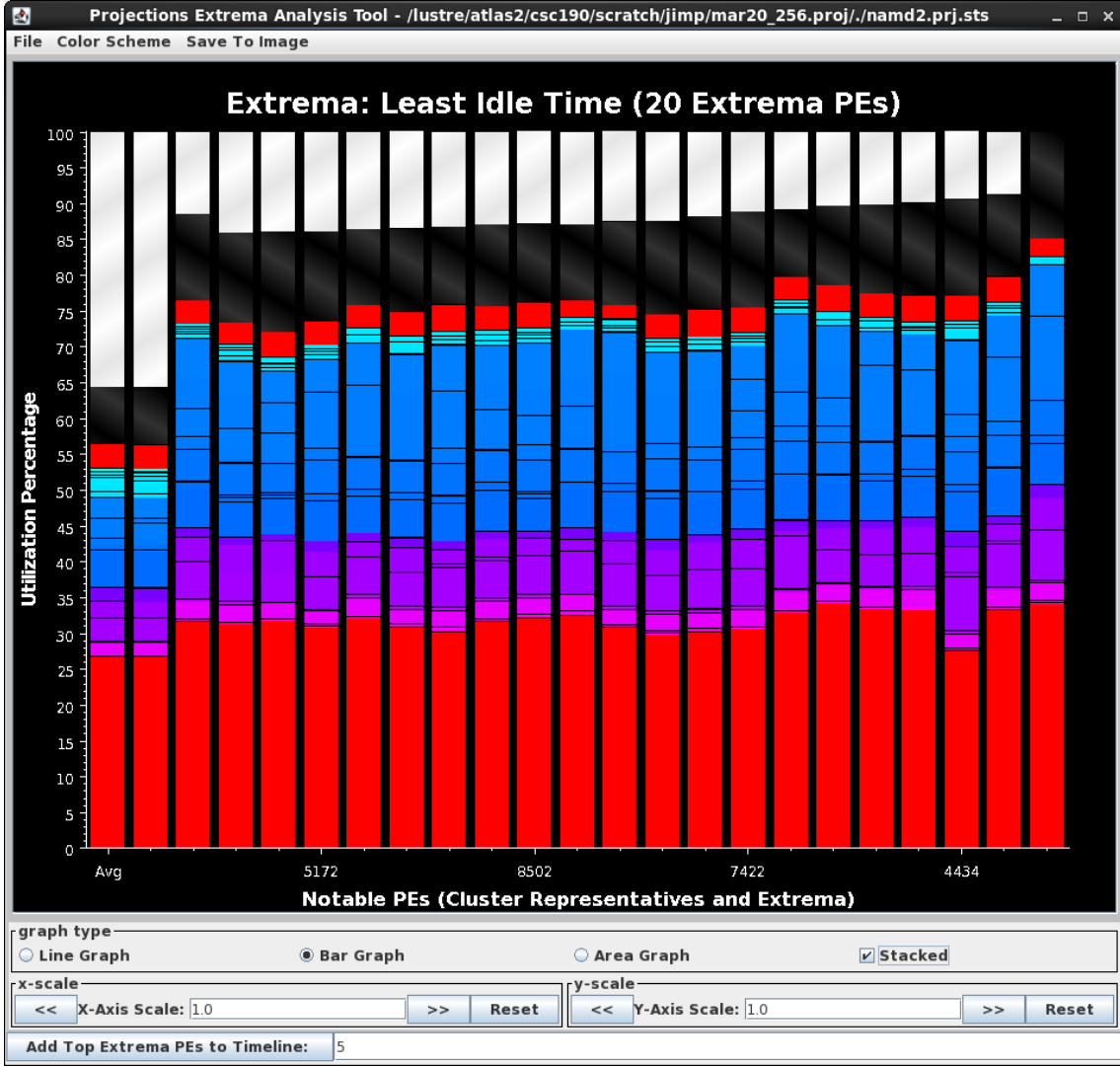
Charm++/NAMD configuration

- IBM PAMI SMP machine layer
  - Initially developed for Blue Gene series
  - No dedicated communication thread
- Single GPU per process (6 processes per node, 6 threads per process)
  - Leaving one core free per resource set seems to reduce noise
  - One core per socket is reserved by jsrun, so 8 unused cores per node
- With thread to core affinity:
  - jsrun -r6 -g1 -c7 namd2 +ignoresharing +ppn 6 +pemap 4-27:4,32-55:4,60-83:4,92-115:4,120-143:4,148-171:4
- Or without (expected to run slower, but sometimes faster):
  - jsrun --bind rs -r6 -g1 -c7 namd2 +ignoresharing +ppn 6

Charm++ *Projections* tool shows bottleneck

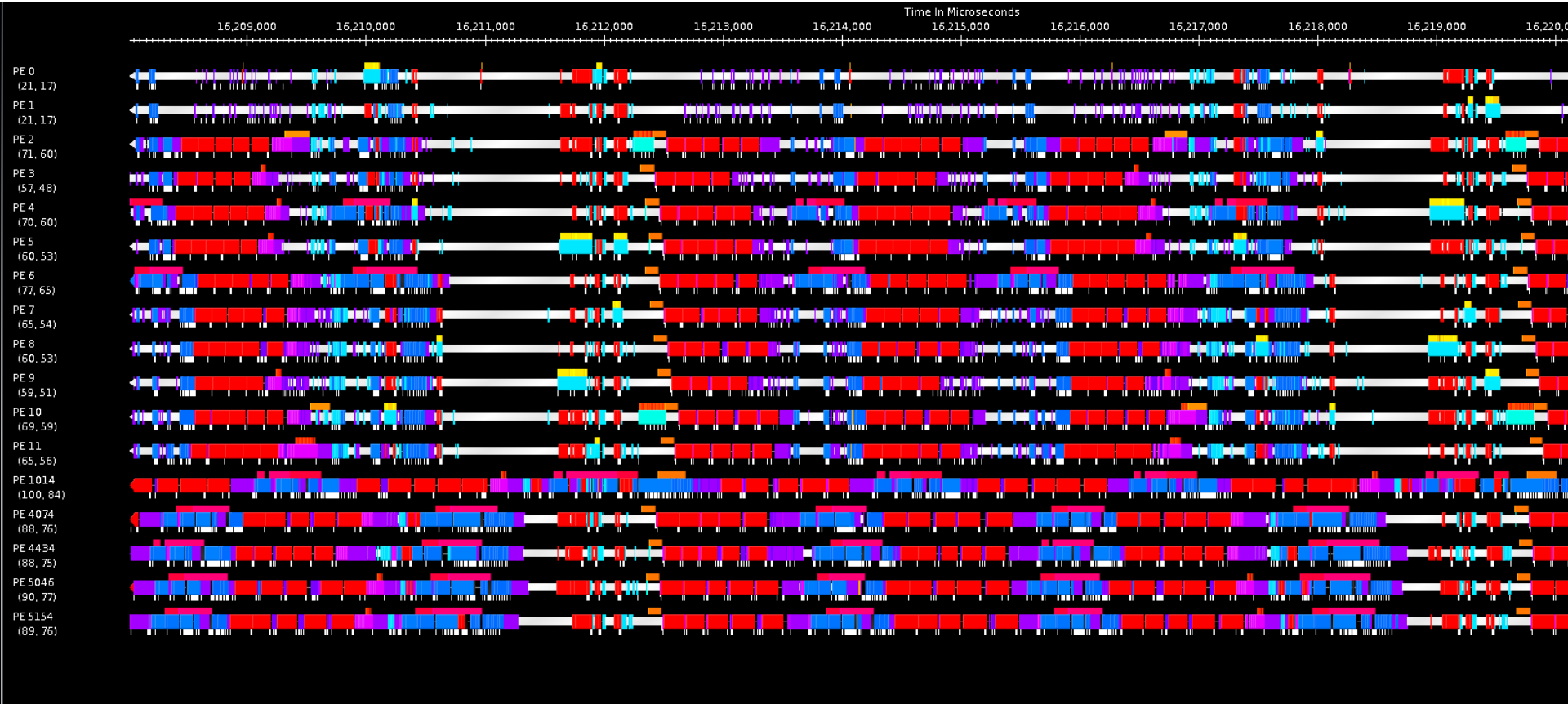


Charm++ *Projections* Extrema Tool Finds Problem PEs



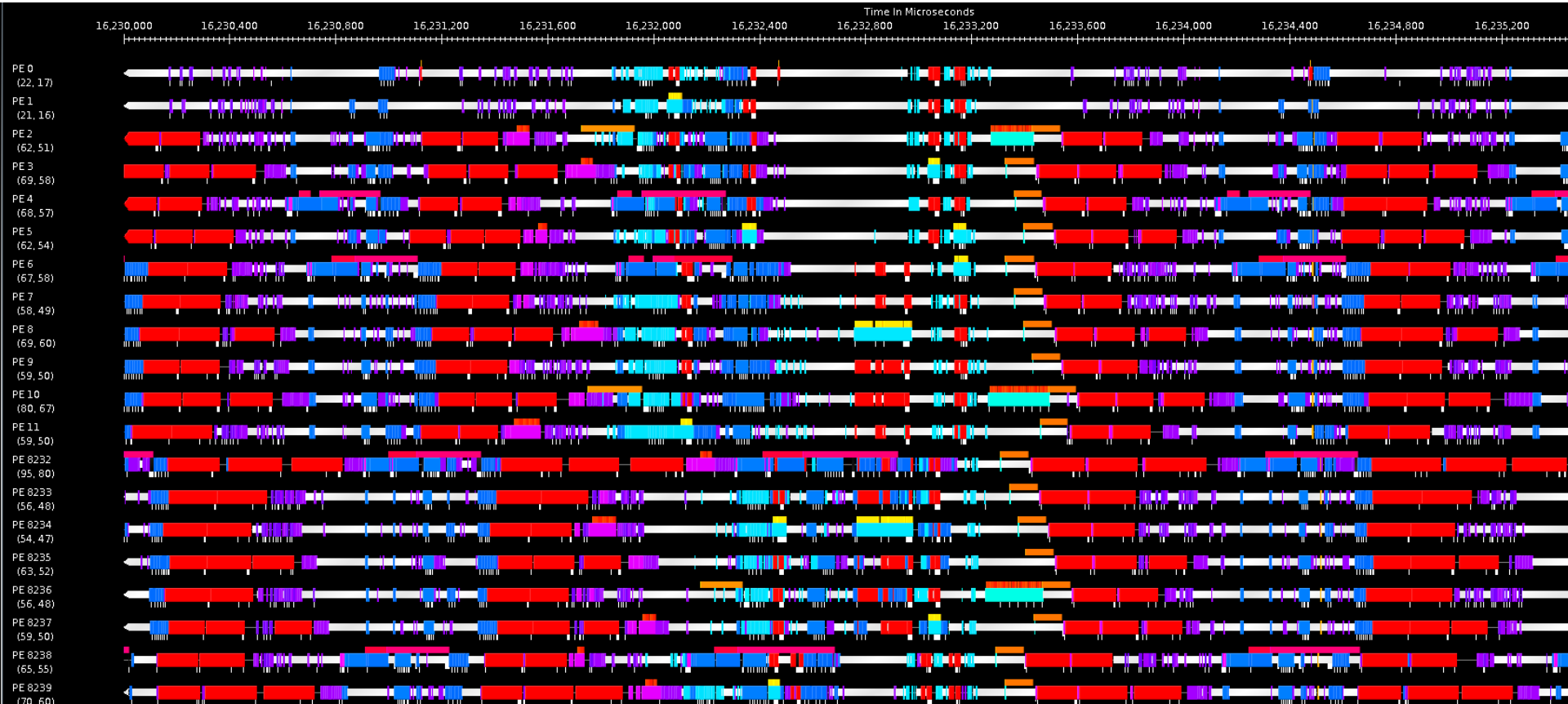
One PE has no idle time!

Also, overloaded PEs are all GPU hosts



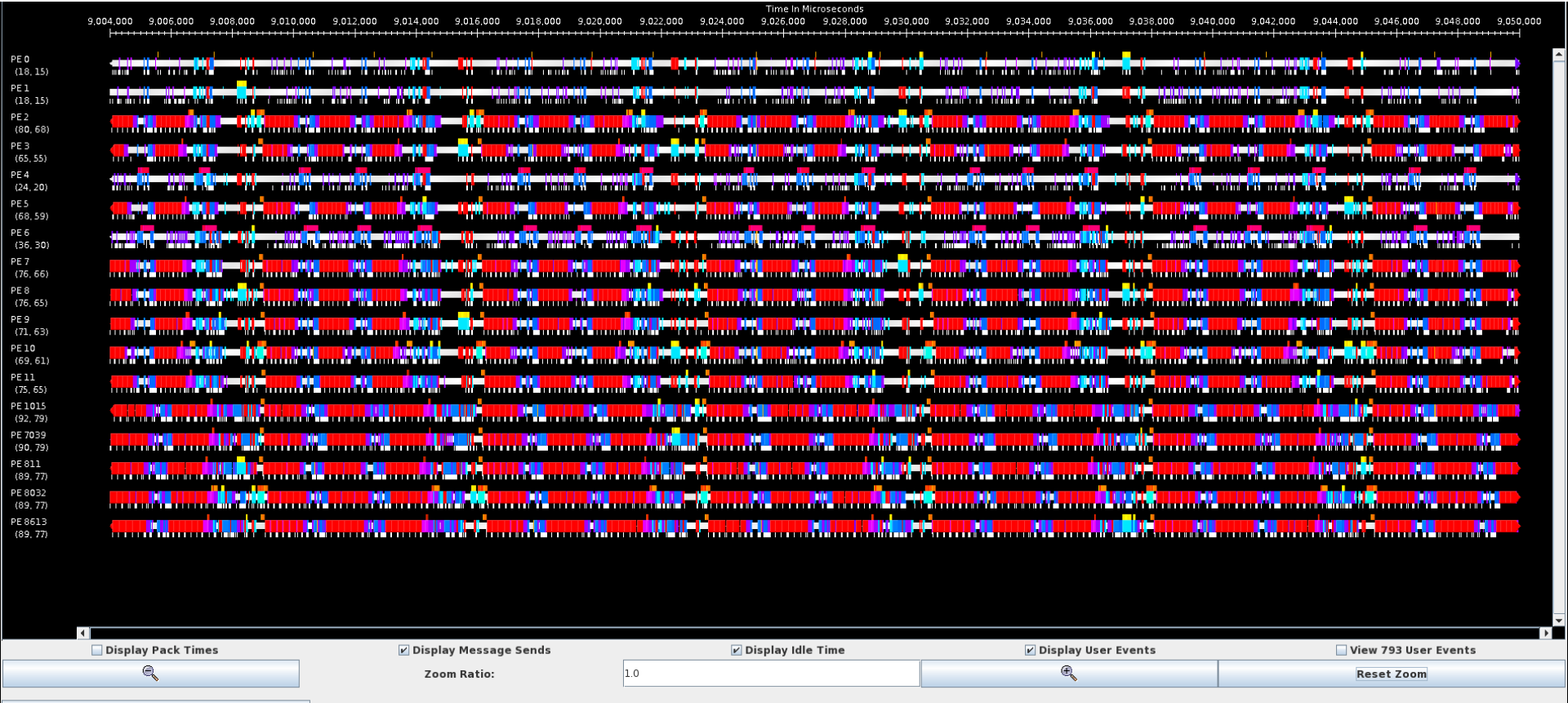
Same issue on 512 nodes

Now showing all PEs on process



Try removing patches from GPU host PEs

Overloaded PEs (256 nodes) are no longer GPU hosts



Overloaded PEs still have idle time

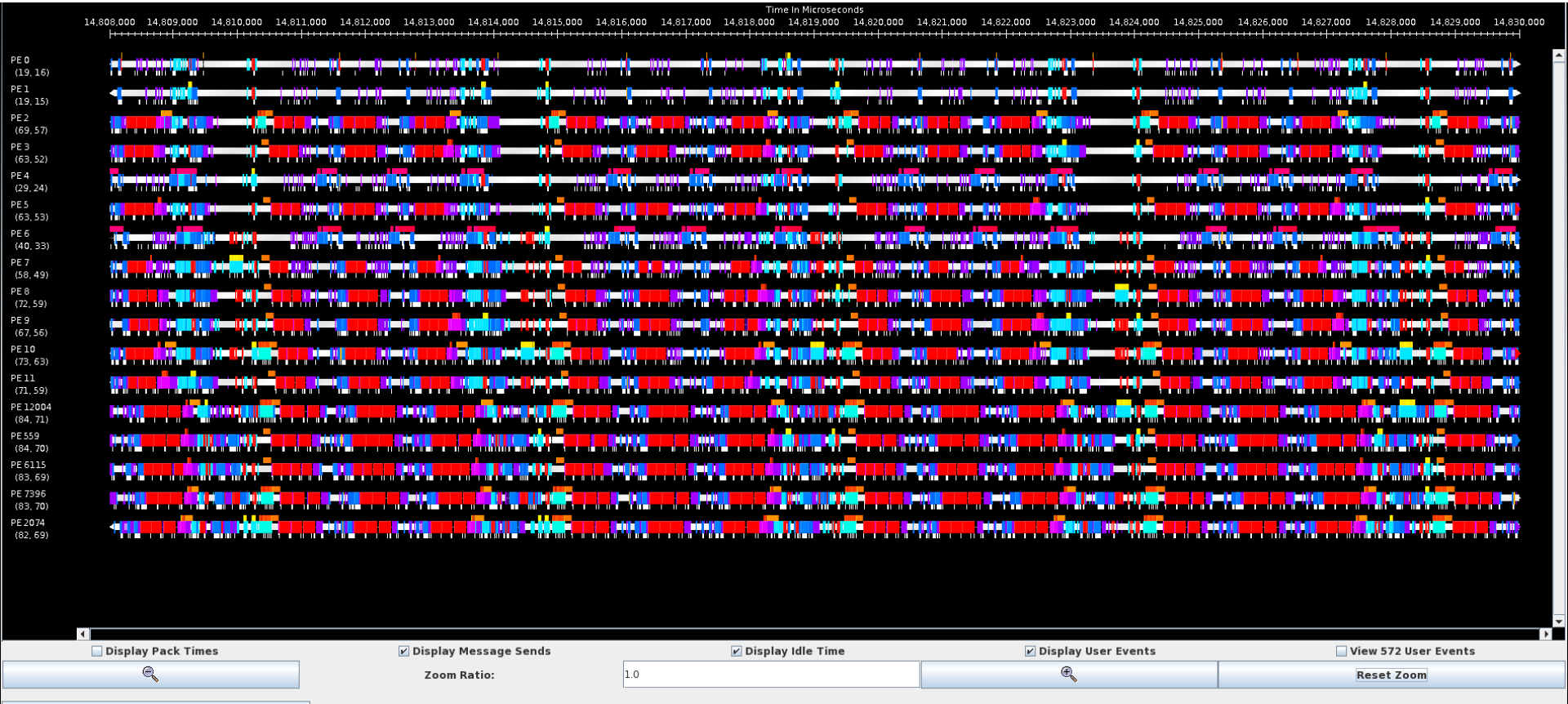
Now showing all PEs on process

Load imbalance delayed until PME steps

256 nodes zoomed in

512 nodes similarly improved

Showing overloaded PEs



512 nodes similarly improved

Now showing all PEs on process

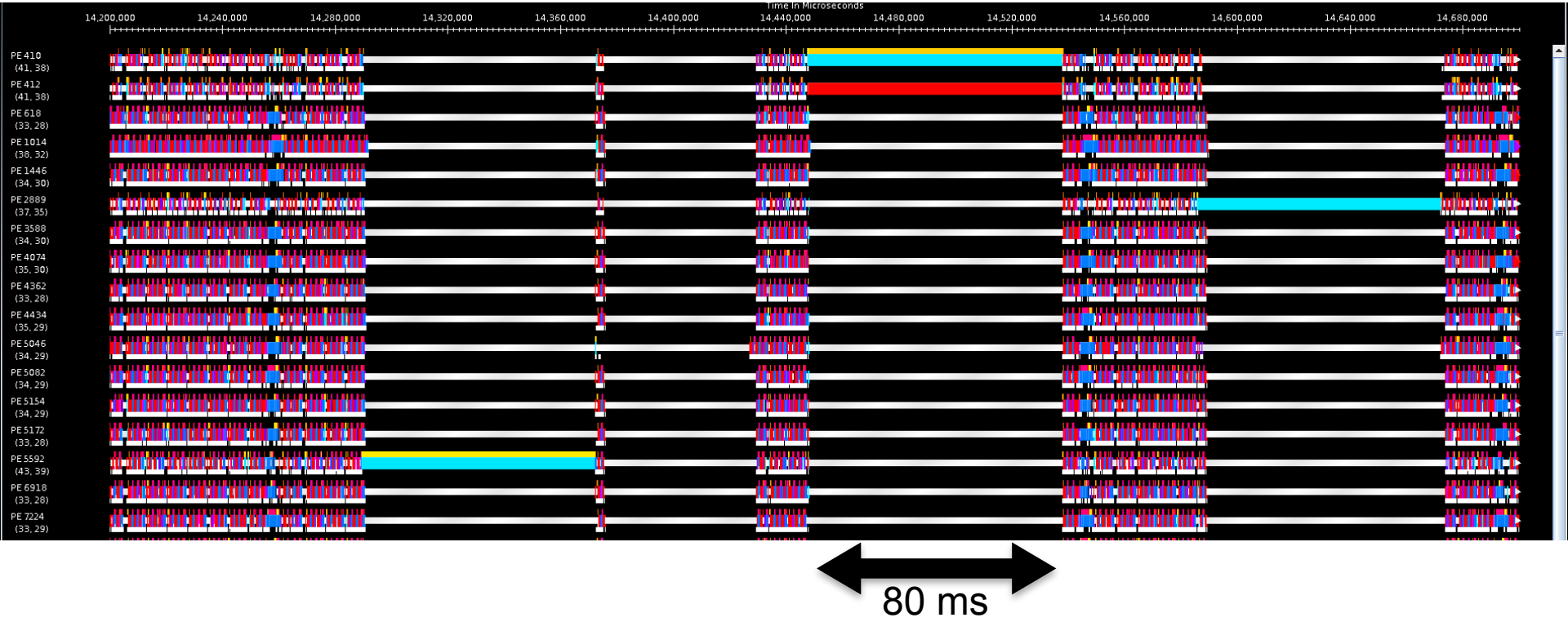
GPU overlaps with communication only

512 nodes zoomed in

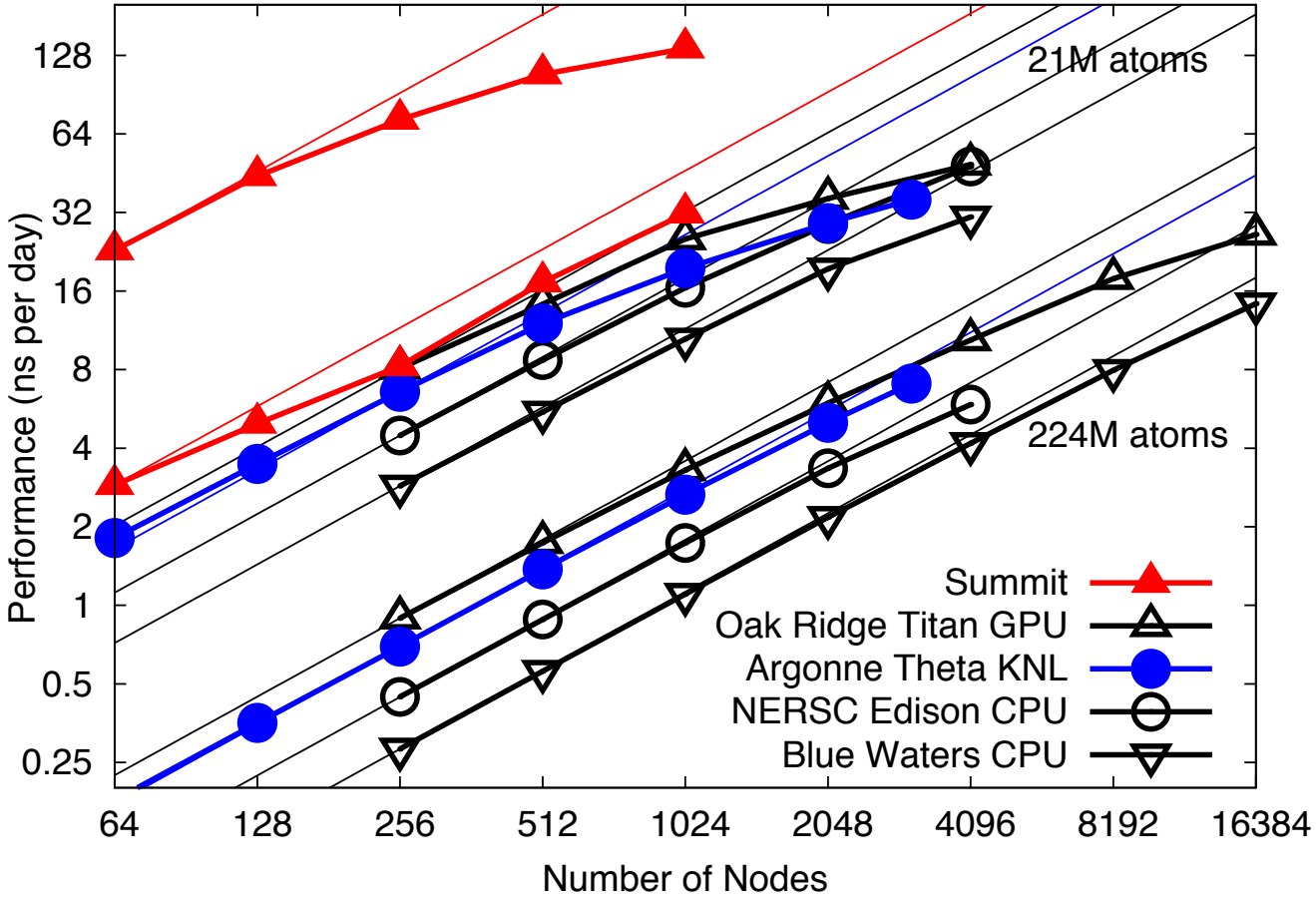
Notes on Benchmarks

- All results are early and preliminary.
- We've had access for less than two months.
- Acceptance is not until this summer.
- Only 1/4 of the nodes are available.
- Installation and software testing continue.
- New platforms always have issues.
- The Volta GPUs are really fast!

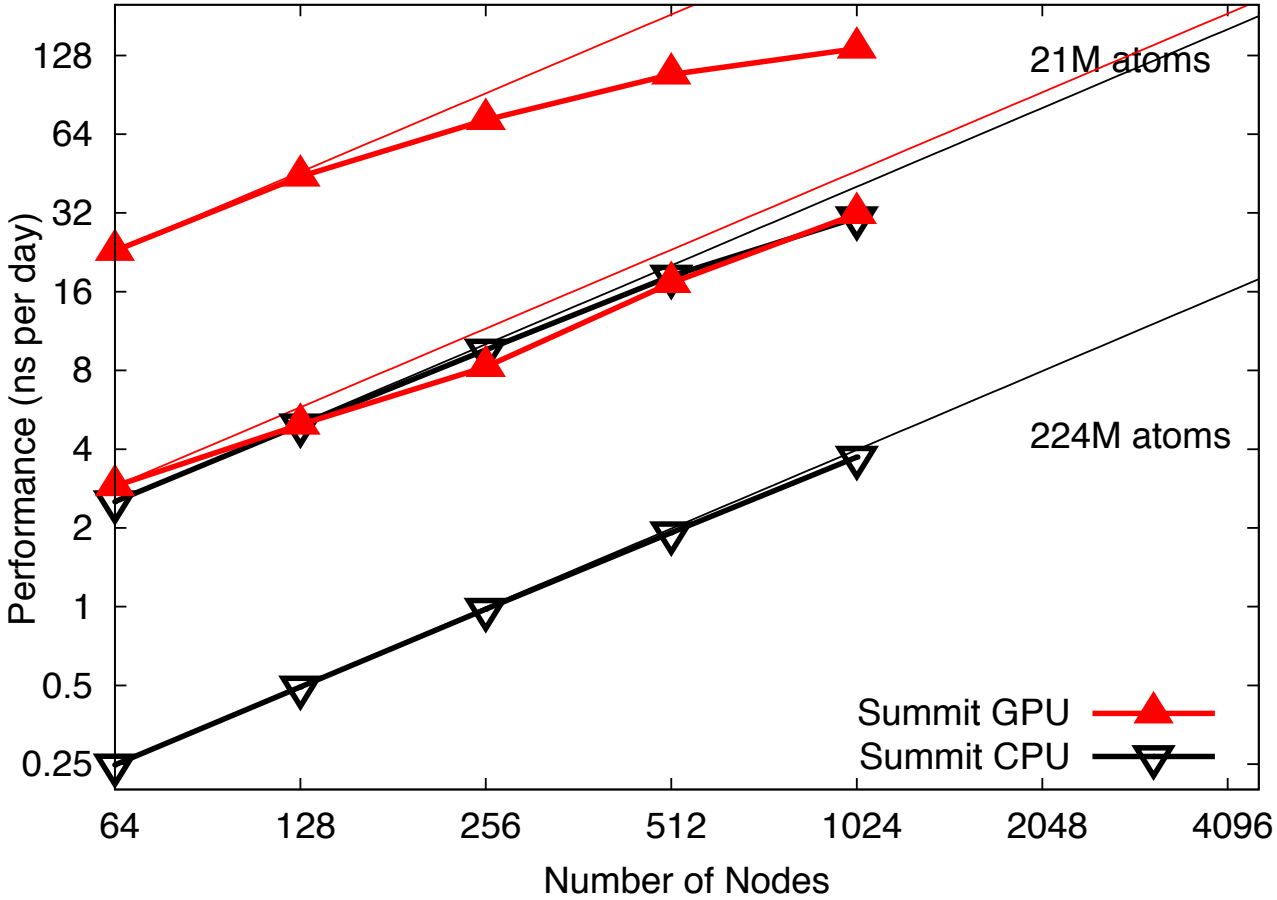
We’re ignoring the noise problem for now



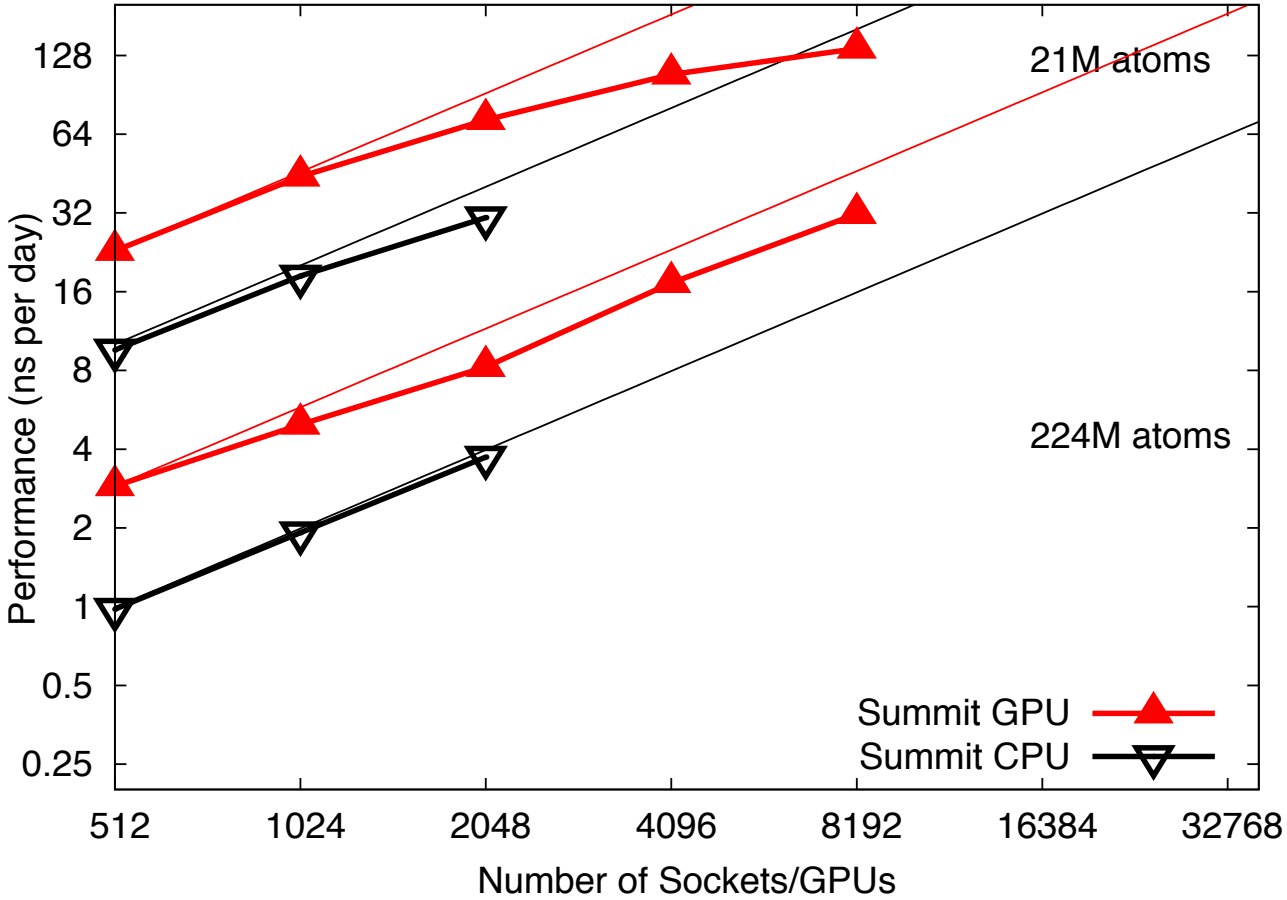
Comparison for large benchmarks



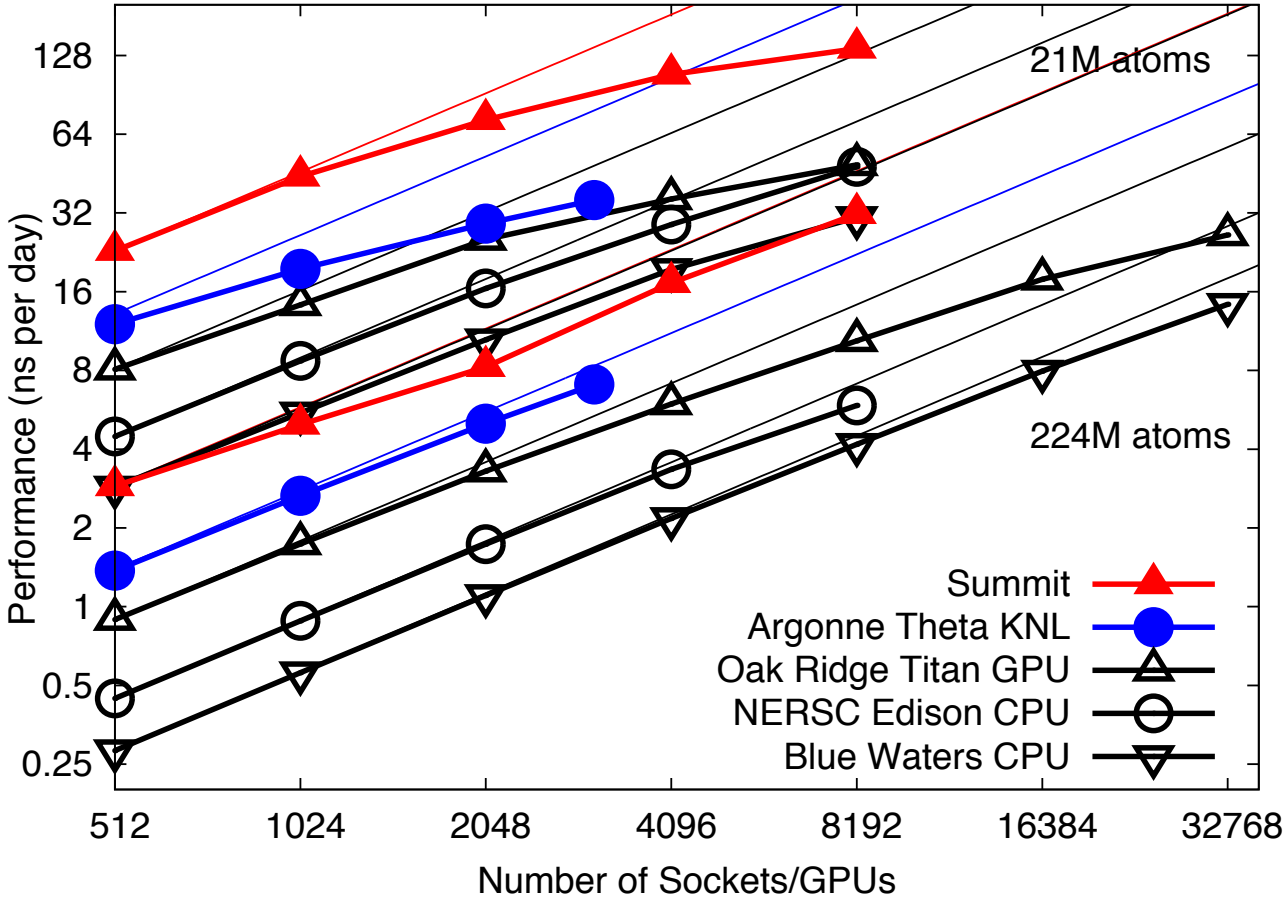
Comparison for large benchmarks



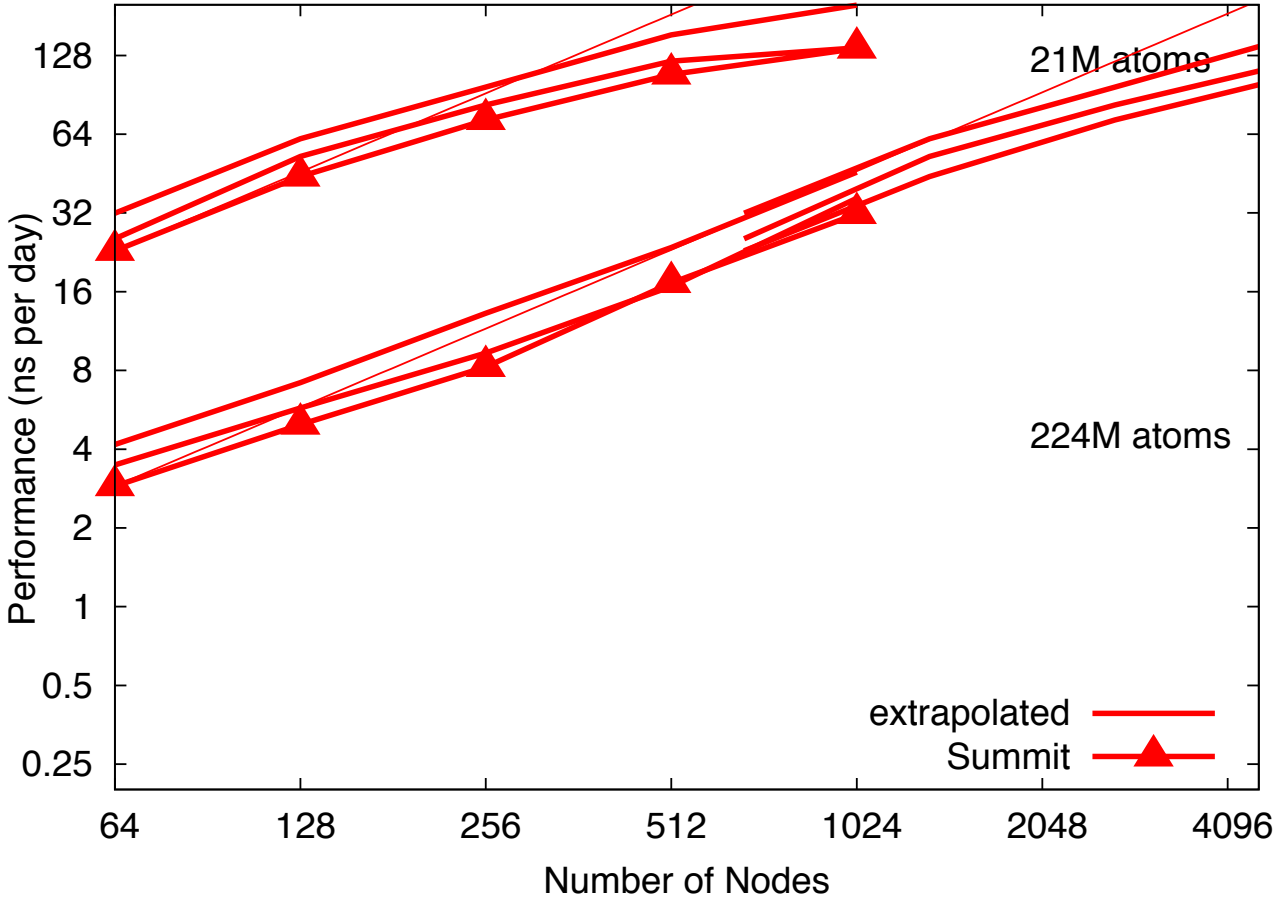
“Fair” comparison for large benchmarks



“Fair” comparison for large benchmarks



“Fix” problems, extrapolate from 21M atom results

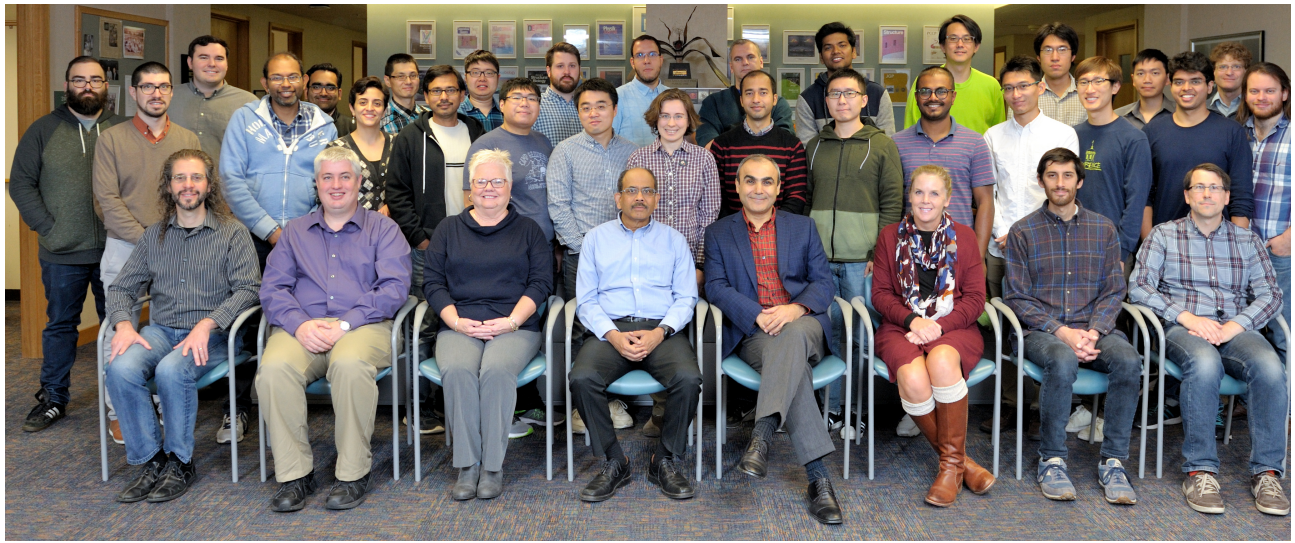


Conclusions and Future Work

- Summit represents a new era in GPU acceleration
  - The CPU will be the bottleneck for many codes
  - Optimizing/vectorizing/parallelizing on the CPU not enough
  - Offload everything practical to the GPUs
- Worry about optimizing the CUDA code last
  - Stage/stream data to reduce CPU/network bottlenecks
- A supercomputer is not just a large cluster
  - IBM knows this (Blue Gene series), Summit should scale well

Acknowledgments

Antti-Pekka Hynninen & Ke Li, NVIDIA
Sameer Kumar & Bilge Acun, IBM
Tjerk Straatsma, OLCF
William Kramer, NCSA
Alexander Bobyr & Michael Brown, Intel
Abhi Singharoy, ASU



NIH Center for Macromolecular Modeling and Bioinformatics
University of Illinois at Urbana-Champaign