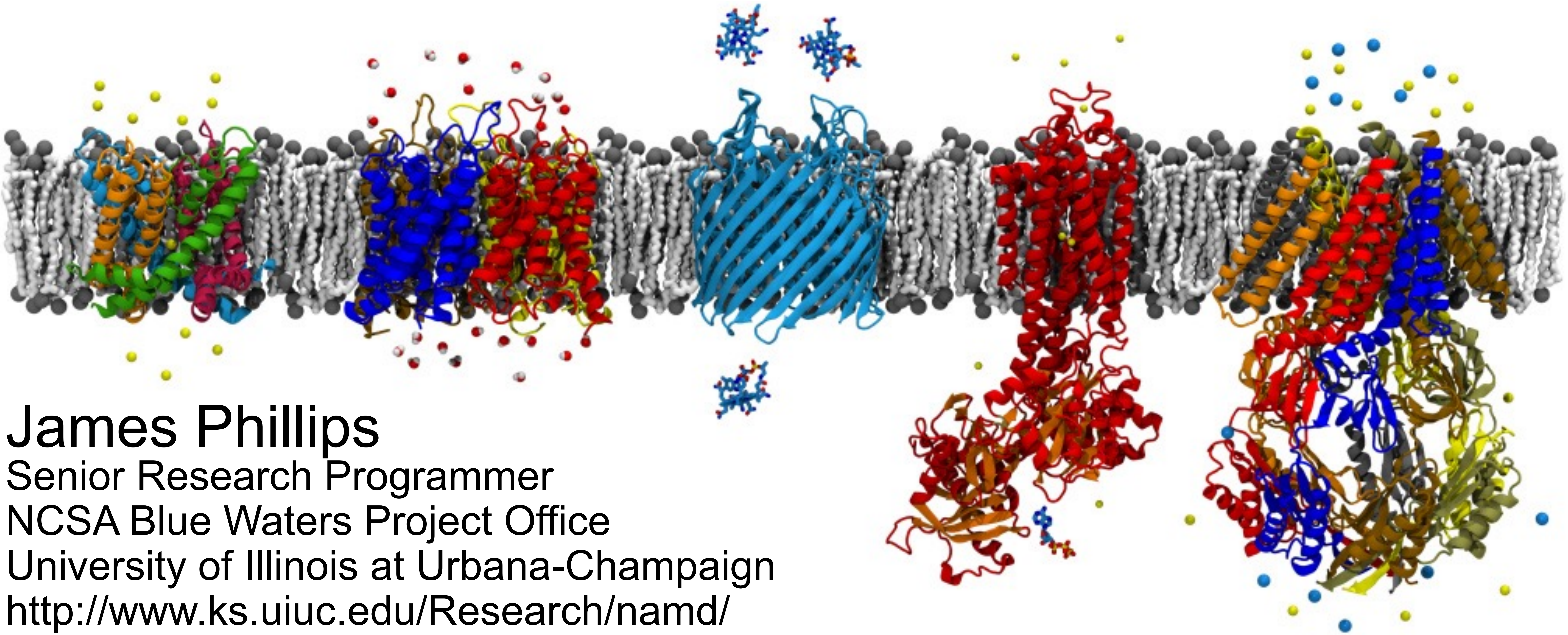
S9302: Petascale Molecular Dynamics Simulations on the Summit POWER9/Volta Supercomputer



James Phillips

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

GTC 2019

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
- Ranks in the top systems in the world – despite being over six years old
- Very large memory capacity (1.66 PetaBytes)
- Very fast file systems (>1 TB/s)
- Very large nearline tape system (>250 PB)
- Very high external network capability (>420 Gb/s)

Seven years of science: November 2012 through December 2019

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

THE HIV-1 CAPSID

Atomic structure of the AIDS pathogen's protein coat

PAGE 643

2013 *HPCwire* Editors' Choice Award for Best Use of HPC in Life Sciences

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**).

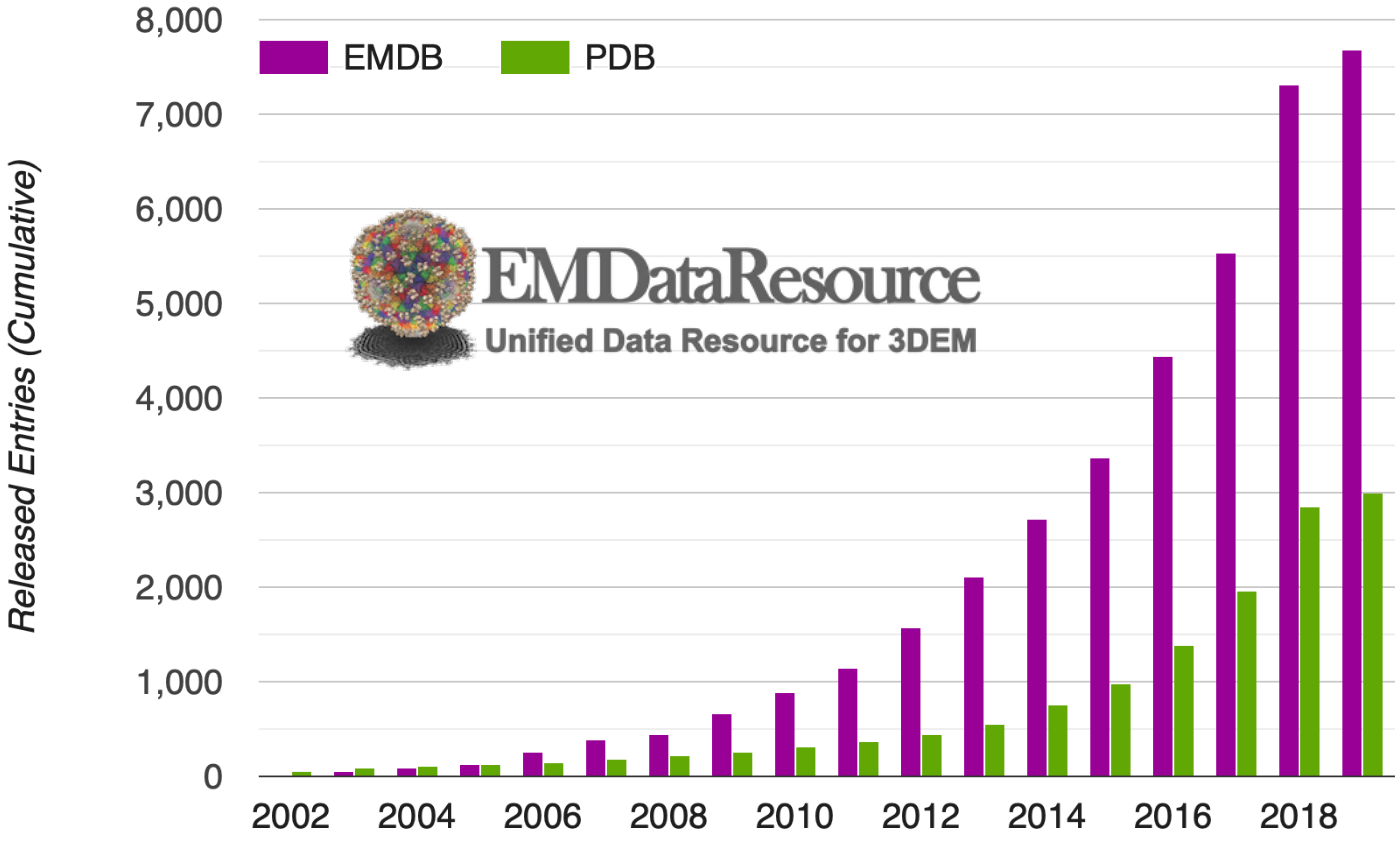
Application readiness/early science projects on
 - Argonne Theta (10 PF Cray KNL, completed)
 - Oak Ridge Summit (200 PF Power9/Volta, 2018)
~~- Argonne Aurora (200 PF Cray KNH, 2019)~~
 - Argonne Aurora (1 EF Intel Xeon + X^e, 2021)

KLAUS SCHULTEN
Center for the Physics of Living Cells
University of Illinois
2012 IEEE COMPUTER SOCIETY
SIDNEY FERNBACH
MEMORIAL AWARD
LAXMIKANT V. KALE
National Center for Supercomputing
Applications · University of Illinois

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

Meeting Emerging Needs of Experimental Structural Biology

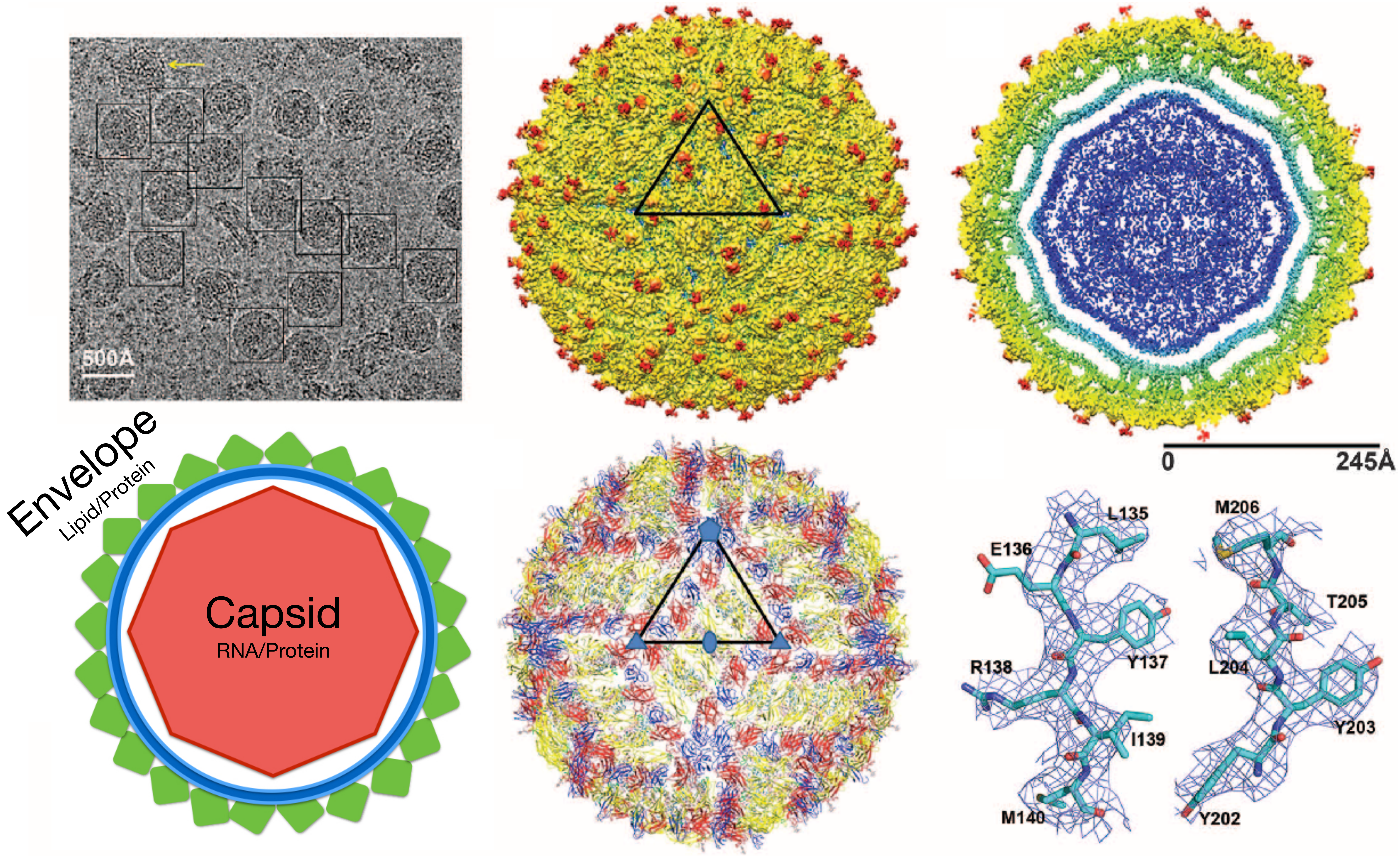
- Computational modeling is indispensable to ANY structural biological method to obtain high-resolution structures
  - X-ray, NMR
  - Cryo-EM, Cryo-ET, SAXS
  - EPR, FRET, MS, Cross-link data
  - Integrative Modeling
- Fast progression of experimental structural biology and other molecular biophysical techniques towards cellular processes
- Explosion of the data made available by techniques such as cryo-EM and cryo-ET



https://www.emdataresource.org/statistics.html

Ultimate Goal of Structural Biology

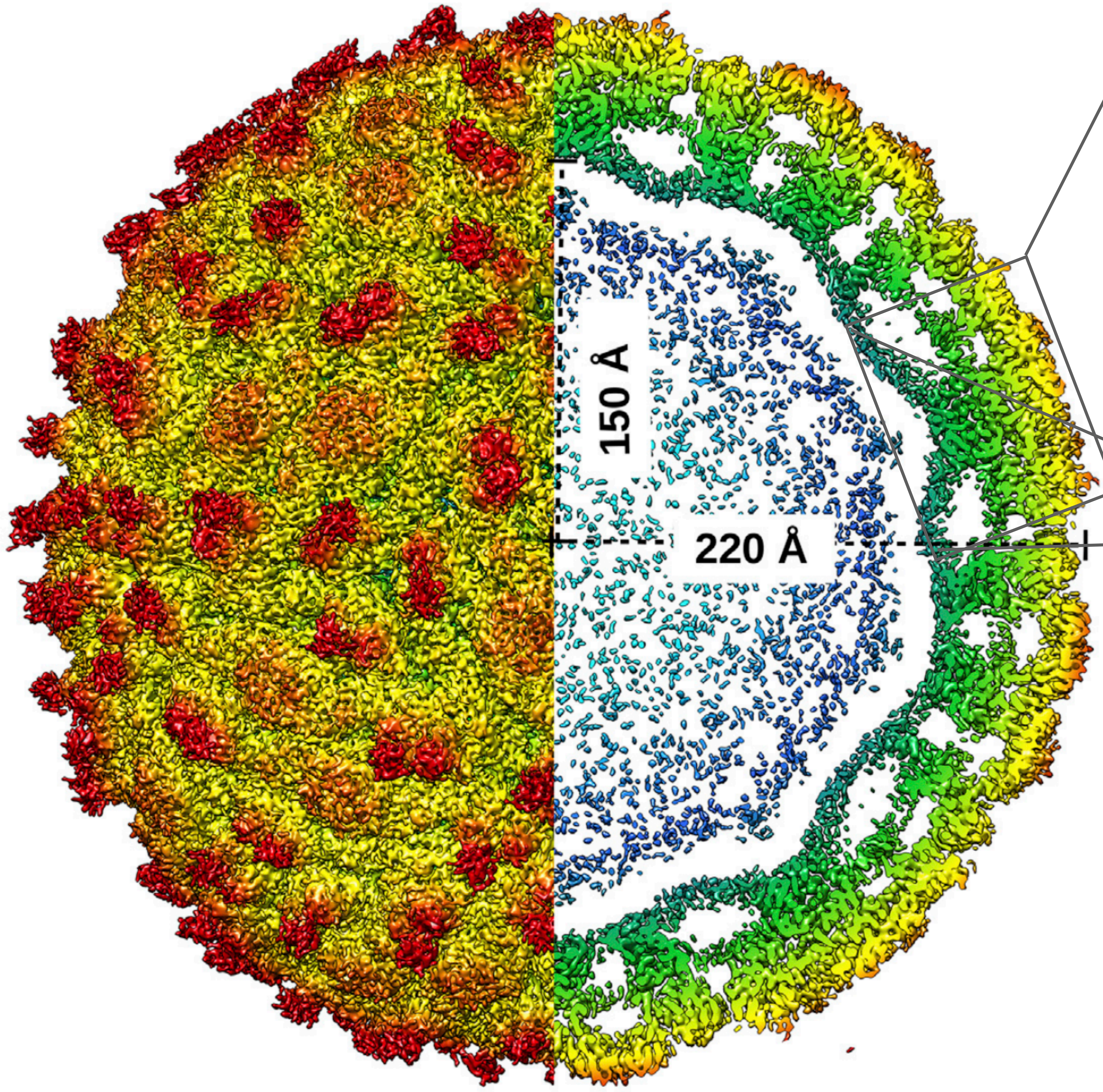
Construction of High-Resolution Structural Models

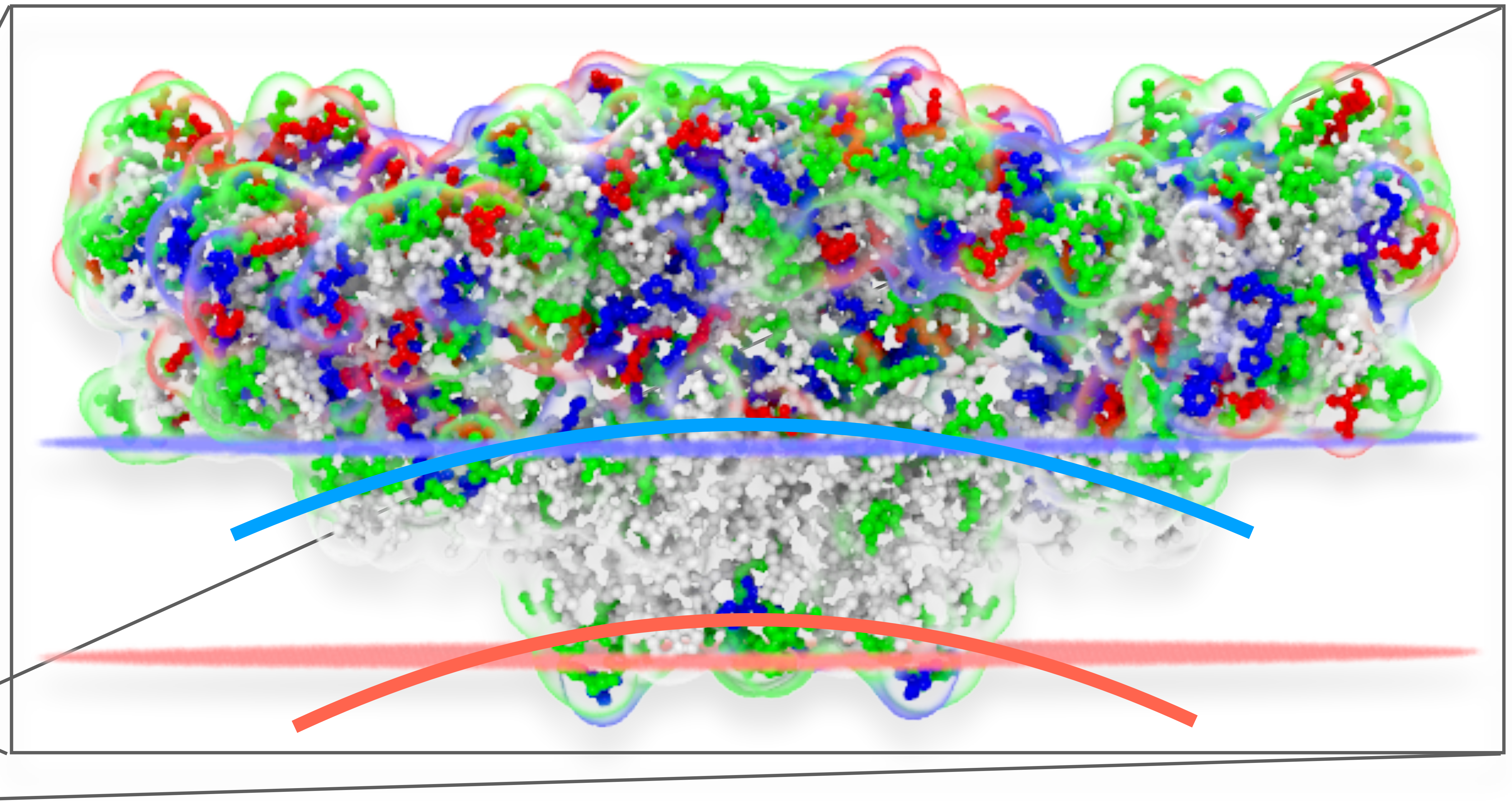


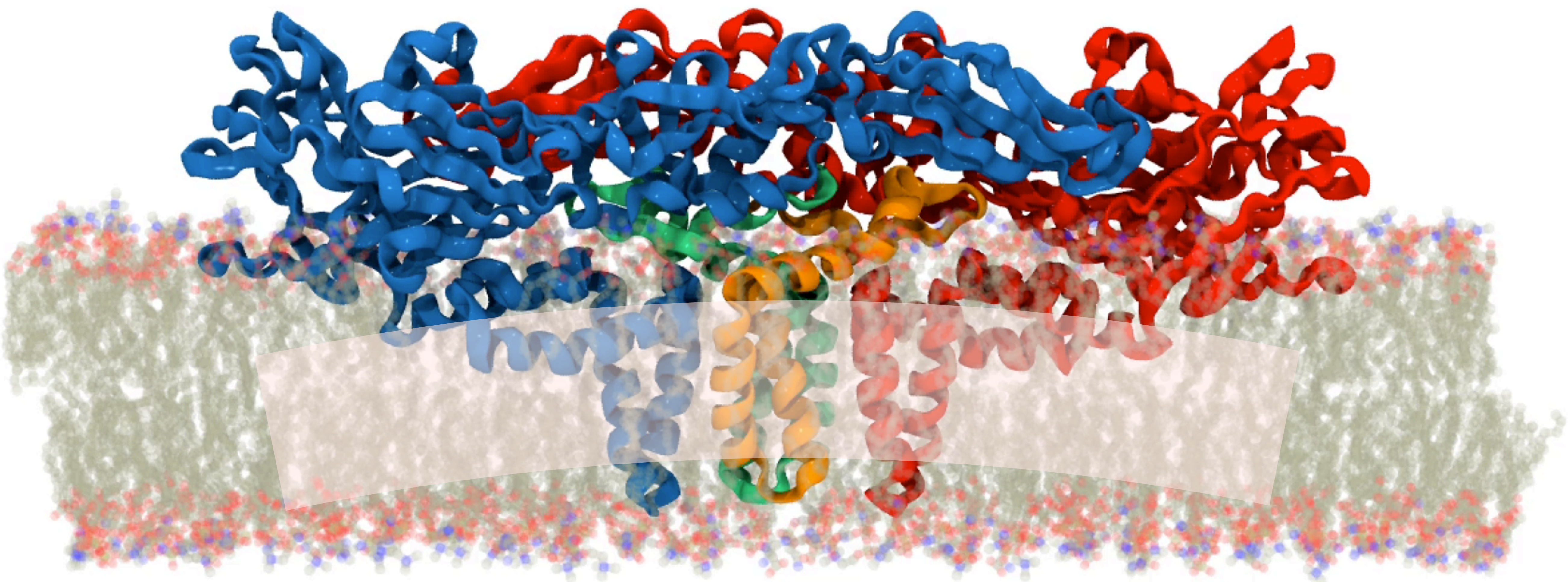
Zika Virus

The 3.8 Å resolution cryo-EM structure of Zika virus.
Sirohi, et al., *Science* 352: 467, 2016

Highly Localized Membrane Curvature Induced by Deeply Inserted Envelope Proteins

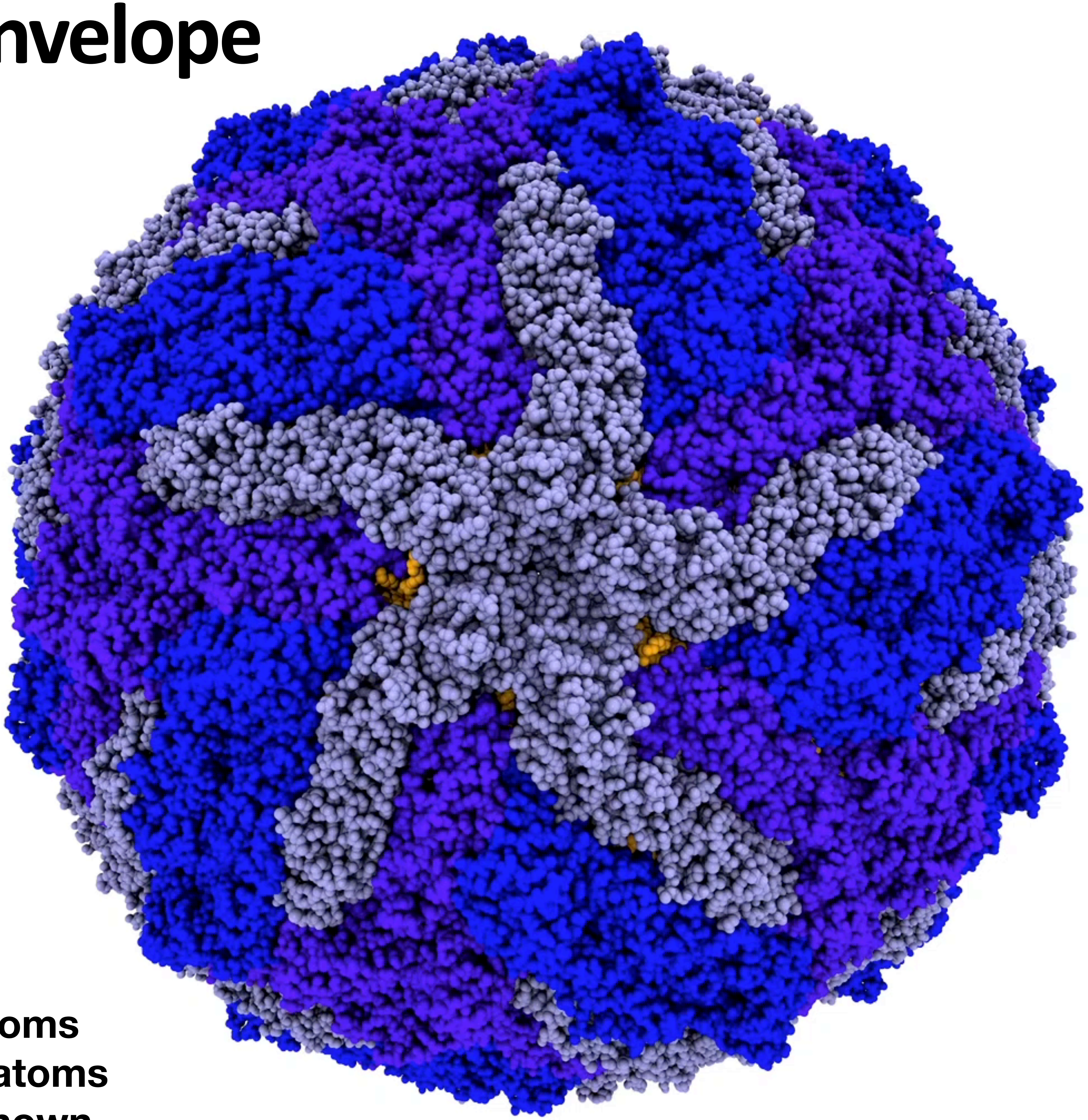






M. Sevvana et al., Refinement and Analysis of the Mature Zika Virus Cryo-EM Structure at 3.1 Å Resolution, *Structure*, Vol. 26, Issue 9, (2018).

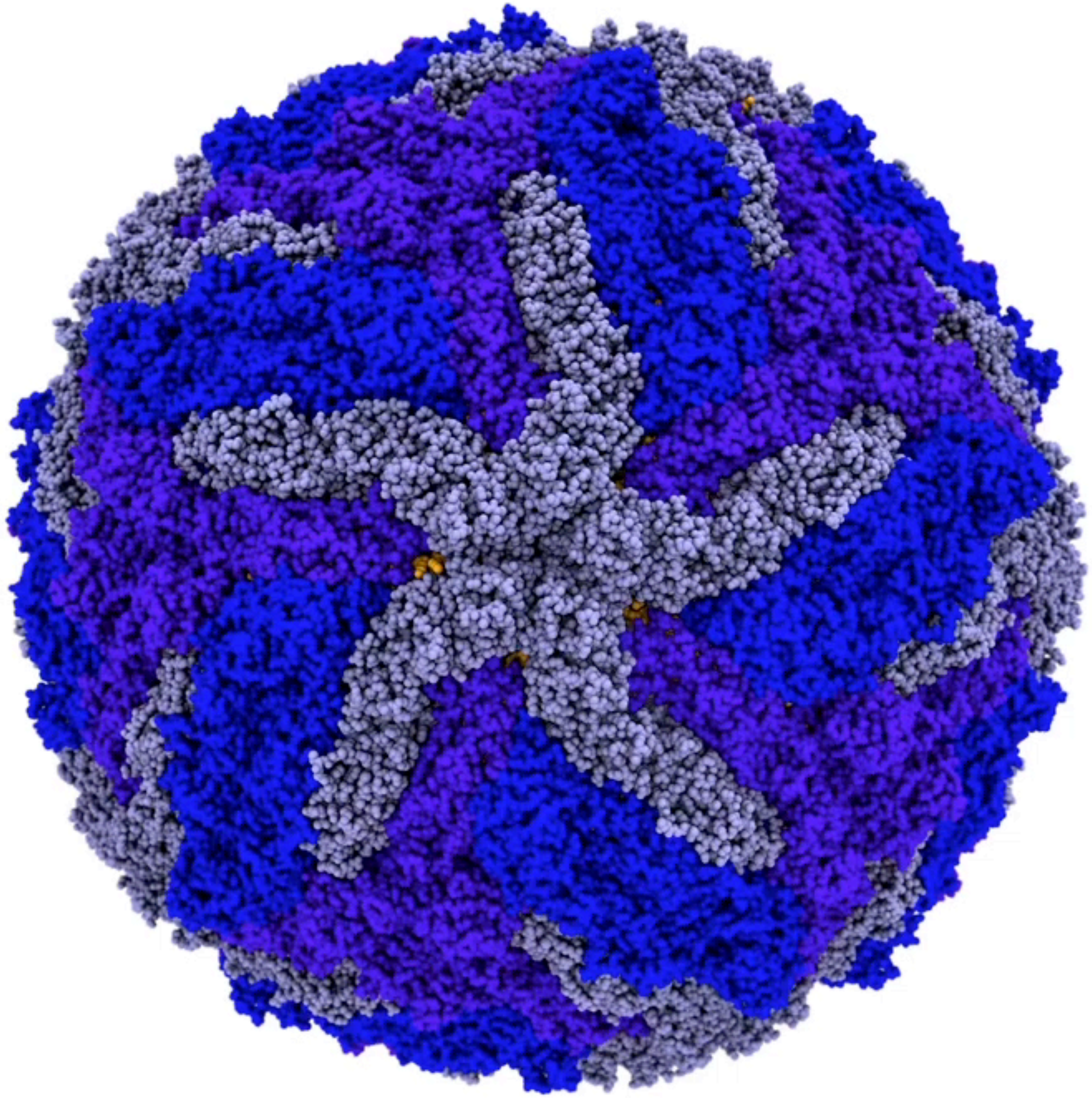
Full Zika Envelope



Envelope: 2.5M atoms
Full System ~ 20M atoms
Solvent/ions not shown

Emad
Tajkhorshid
Illinois

Full Zika Envelope



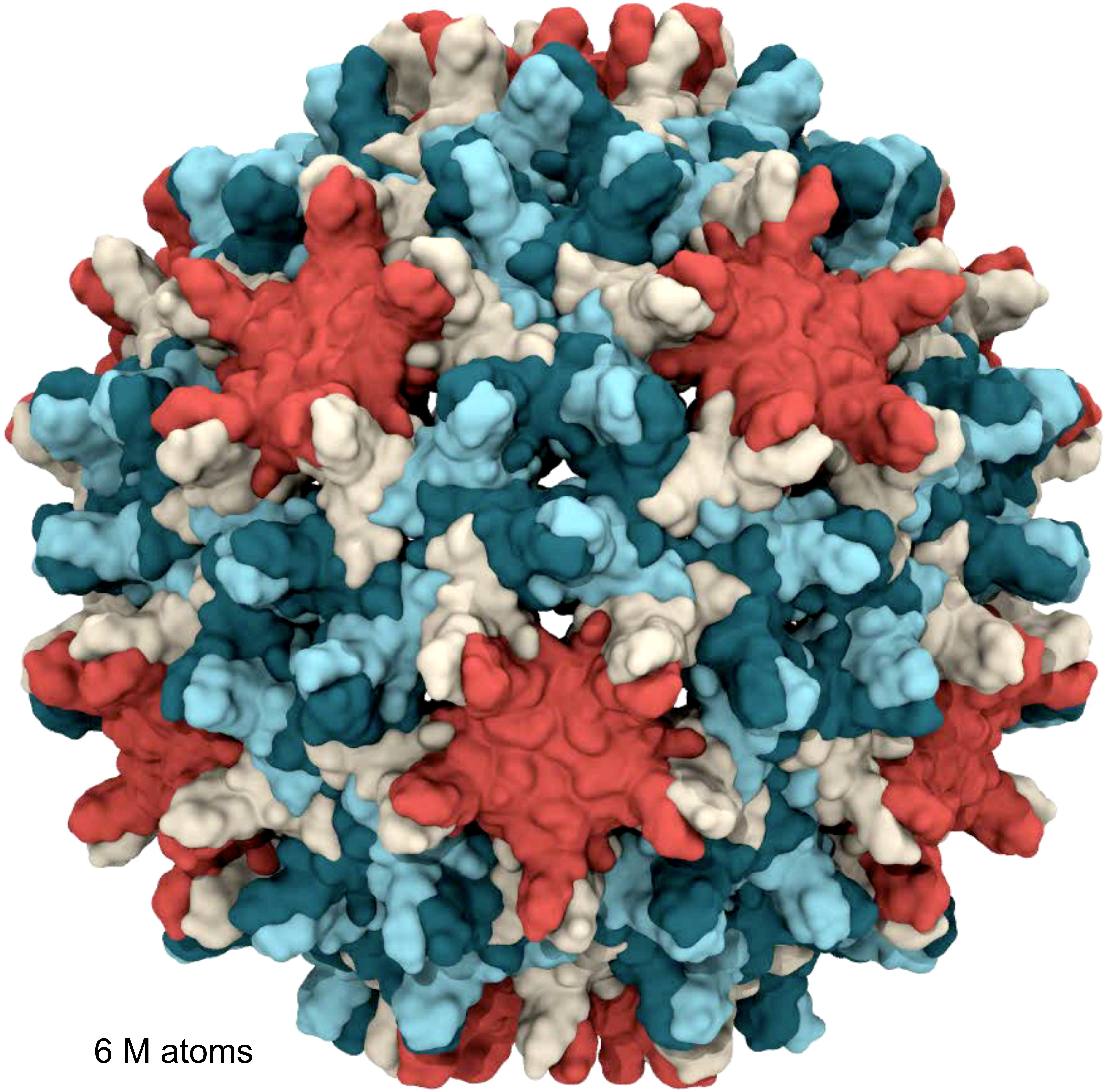
Caution with the setup!

Emad
Tajkhorshid
Illinois

Microsecond simulations of hepatitis B capsid

- Causes severe liver disease
- Chronic infection in 250 million people
- Vaccine available, but no cure
- Capsid is promising drug target
  - Drives genome delivery to cell nucleus

Jodi Hadden, University of Delaware

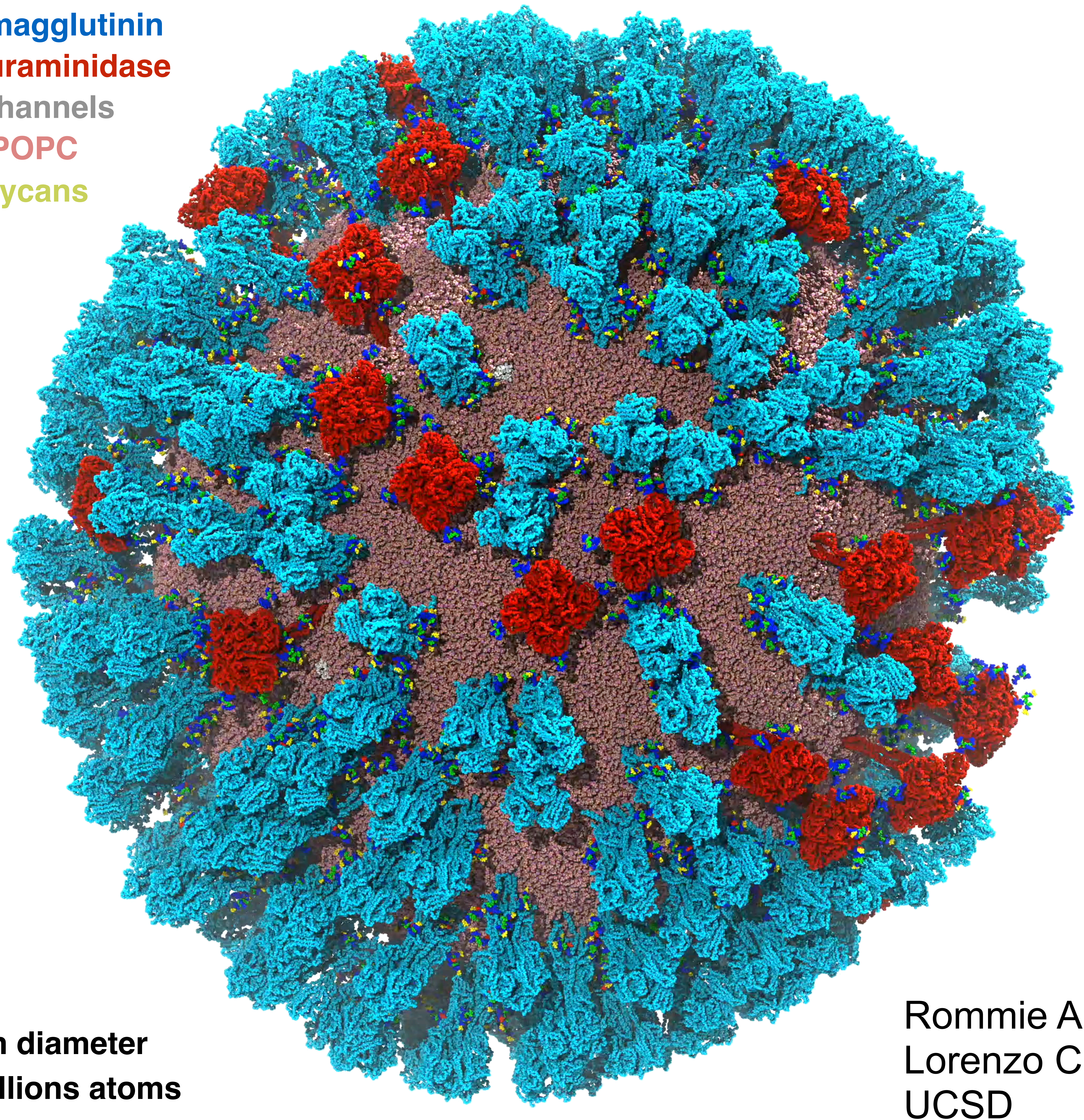


6 M atoms

Hadden, et al. *eLife* **2018**.

Elucidating the impact of glycans on the A/Shandong/2009 (H1N1) influenza virus

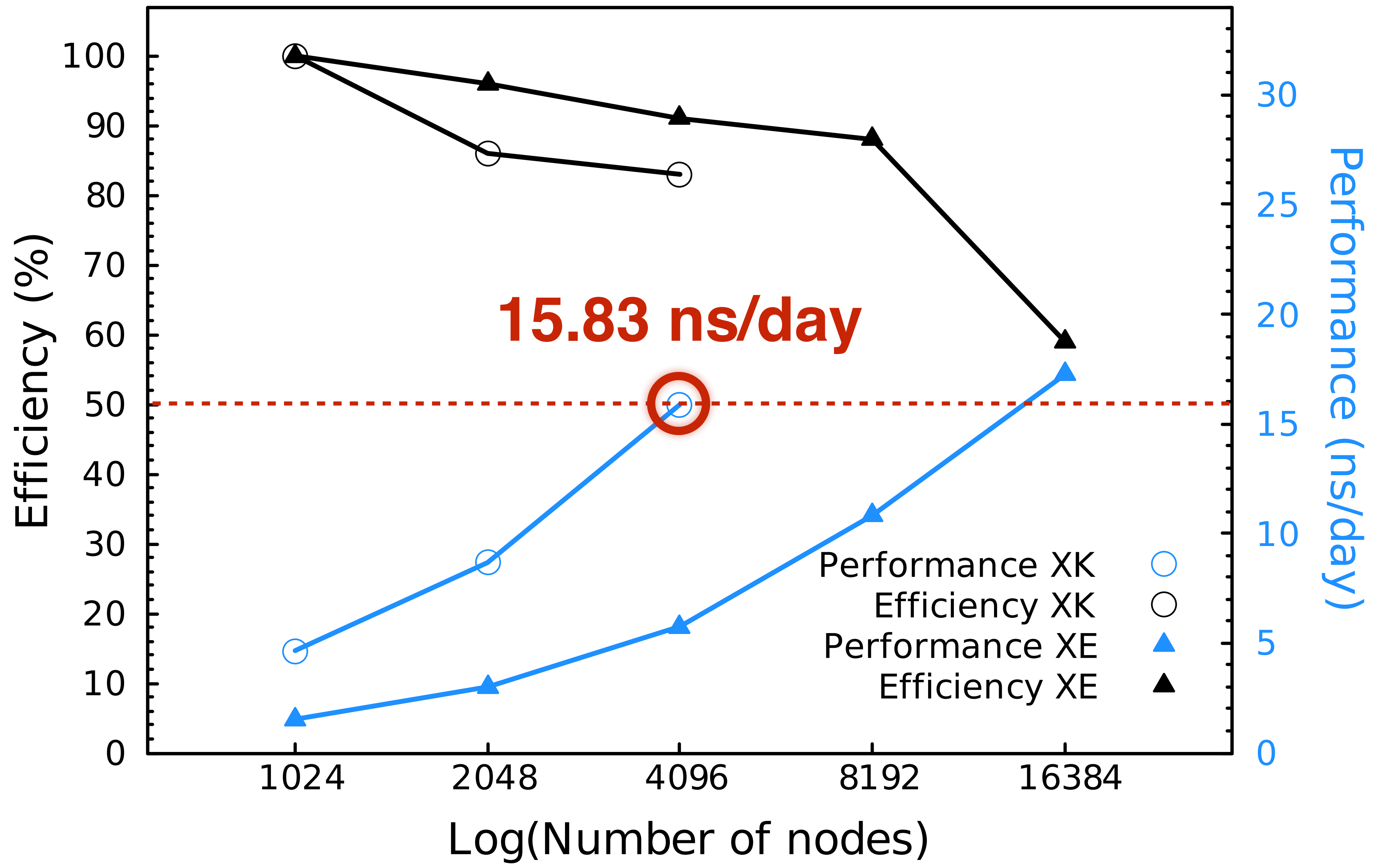
708 Hemagglutinin
120 Neuraminidase
11 M2 channels
48,043 POPC
1,509 glycans



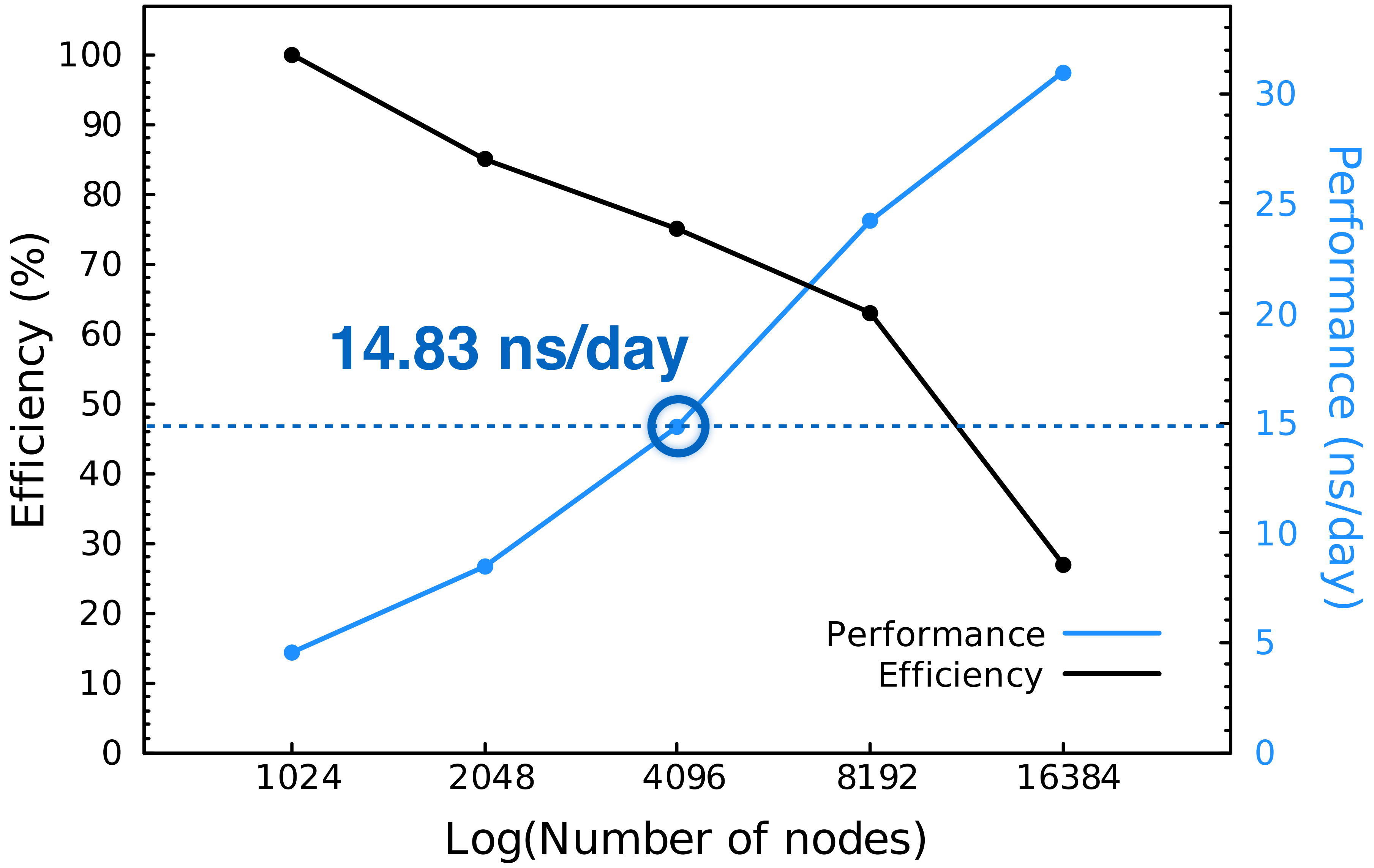
~110 nm diameter
~160 millions atoms
Explicit water (115 nm x 120 nm x 116 nm)

Rommie Amaro
Lorenzo Casalino
UCSD

BLUE WATERS



TITAN (OAK RIDGE)



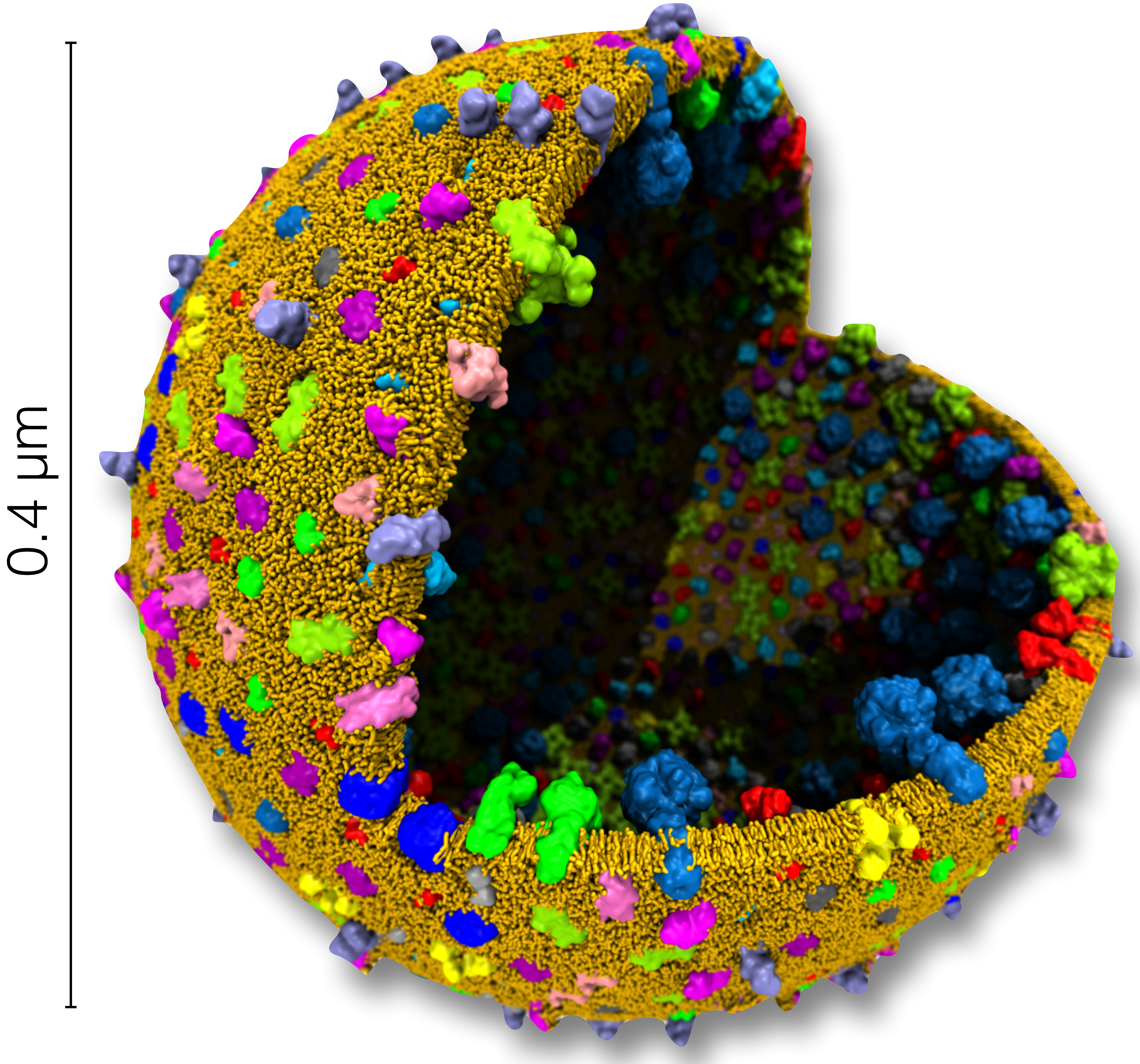
Summit will replace Titan as the OLCF’s leadership supercomputer

summit

- 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

Summit Early Science: Modeling of a Minimal Cell Envelope



	Protein Components	Copy #
●	Aquaporin Z	97
●	Copper Transporter (CopA)	166
●	F1 ATPase	63
●	Lipid Flipase (MsbA)	29
●	Molybdenum transporter (ModBC)	130
●	Translocon (SecY)	103
●	Methionine transporter (MetNI)	136
●	Membrane chaperon (YidC)	126
●	Energy coupling factor (ECF)	117
●	Potassium transporter (KtrAB)	148
●	Glutamate transporter (Glt$_{Tk}$)	41
●	Cytidine-Diphosphate diacylglycerol (Cds)	50
●	Membrane-bound protease (PCAT)	57
●	Folate transporter (FolT)	134
●		
		1,397

3.7 M lipids, 1,400 proteins,
416 M water molecules, 2.4 M ions

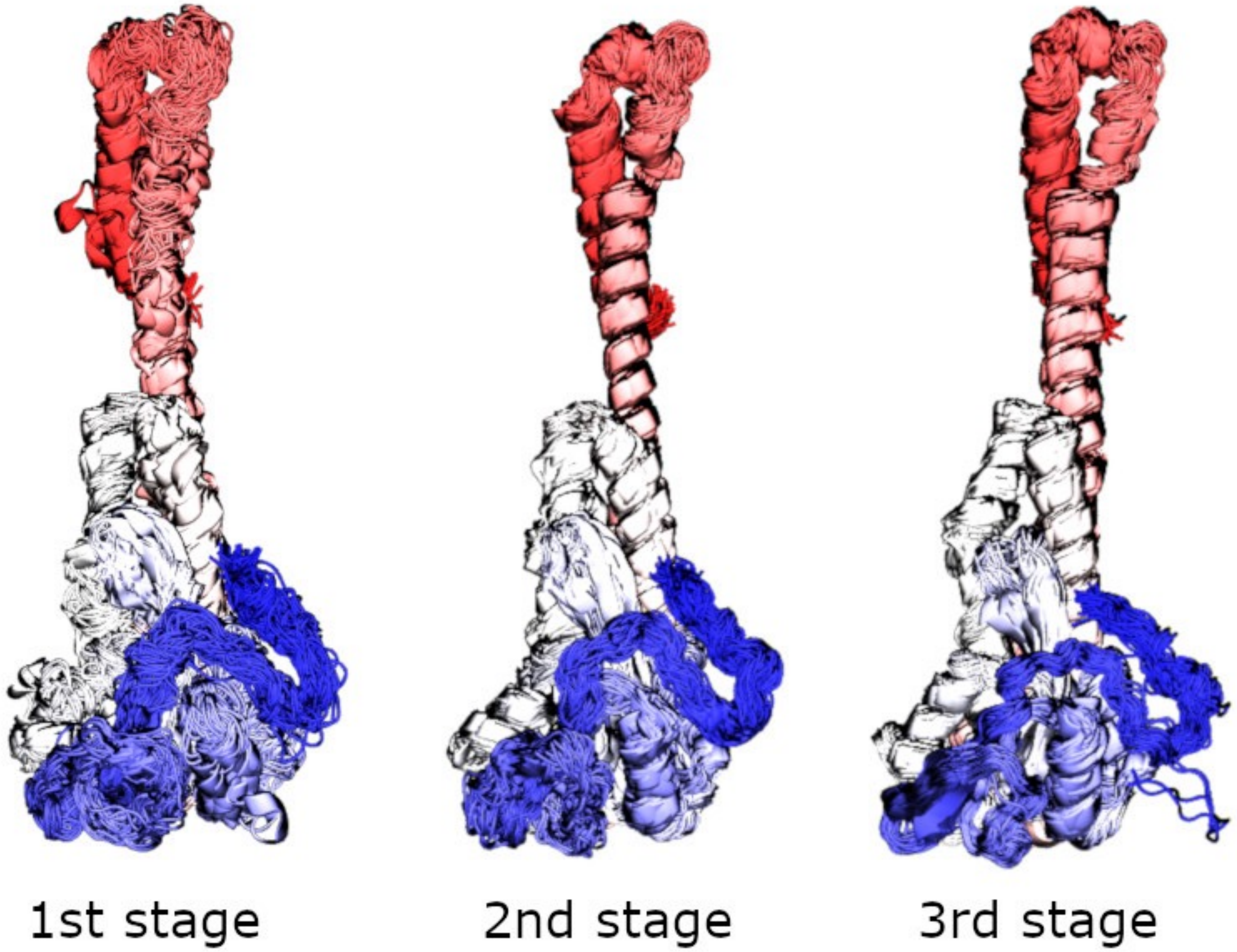
Multi-Copy NAMD Application 1: Protein Folding

soluble systems

Titan + Summit

transmembrane-systems

flippase

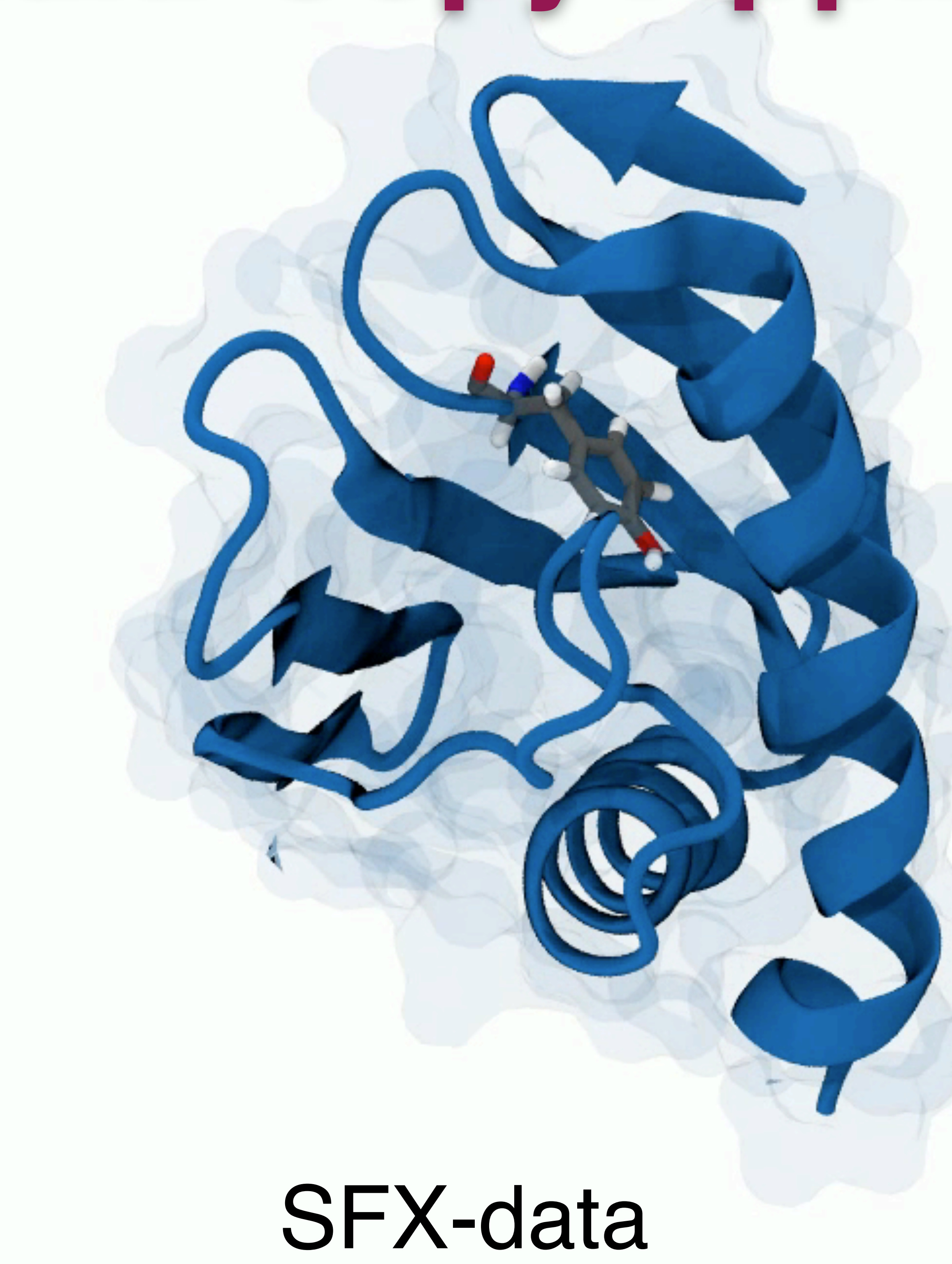


CorA- Mg Channel

MaxEnt methods: Dill, Tajkhorshid, Perez, Kihara

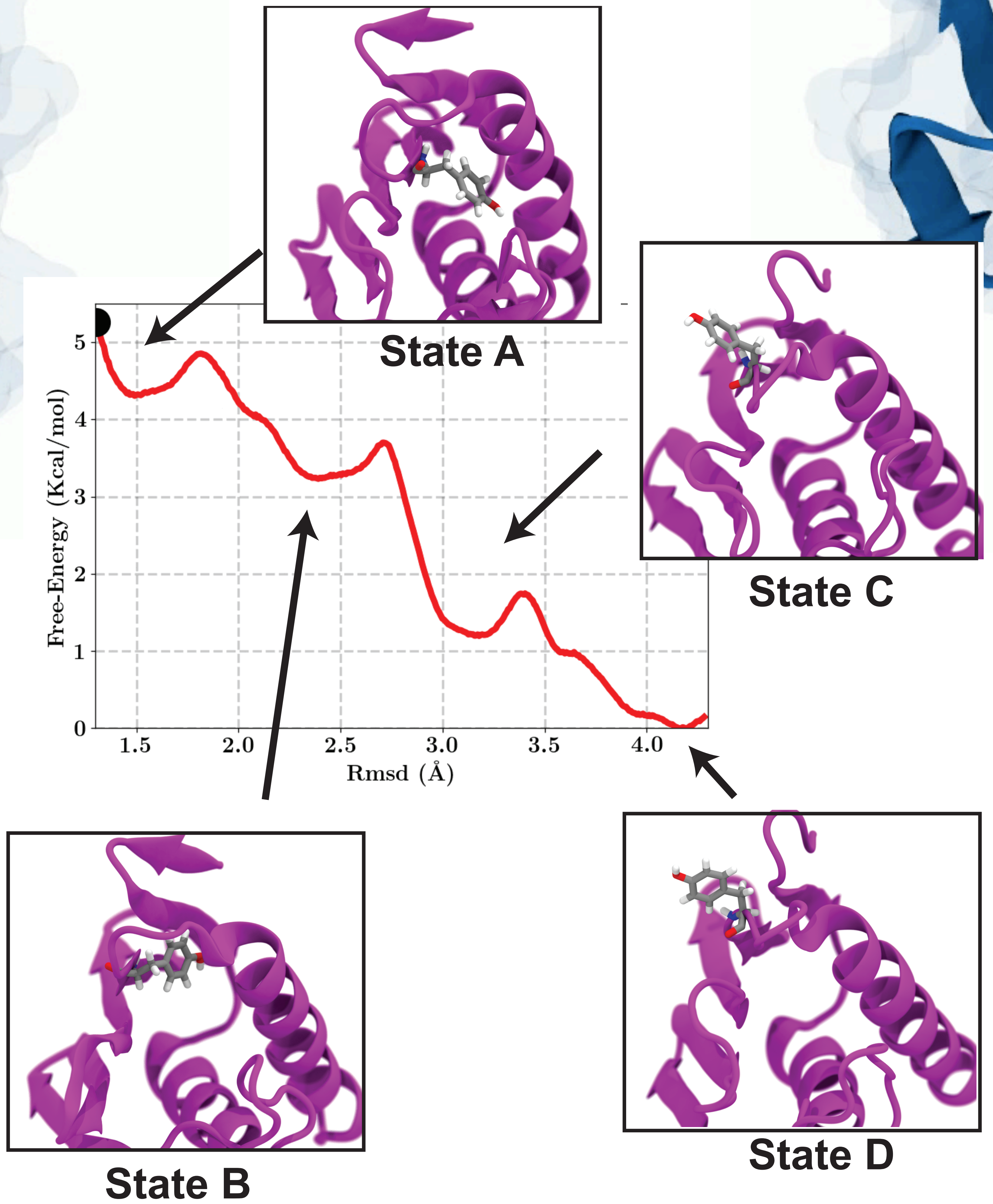
Abhishek Singharoy ASU

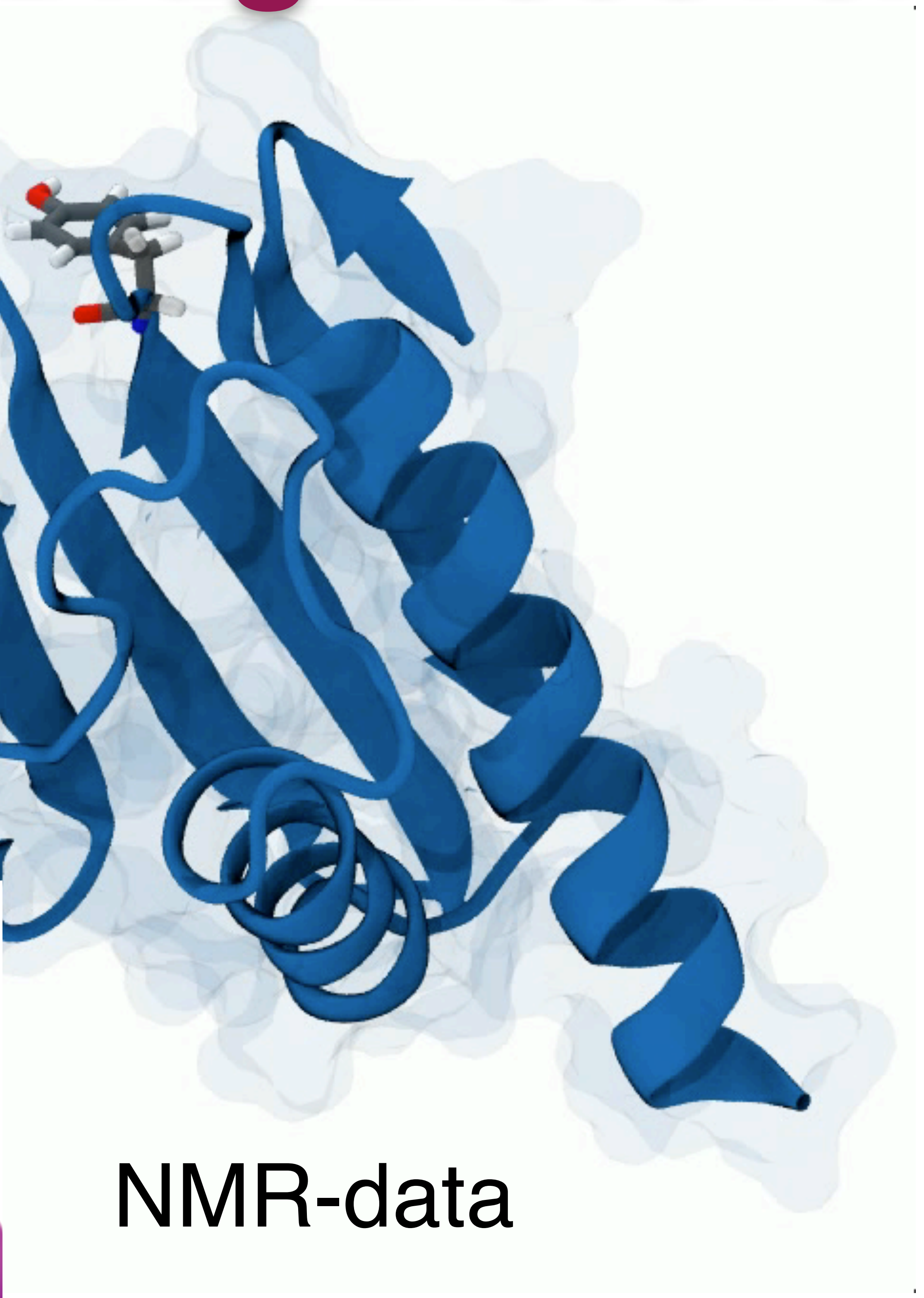
Multi-Copy Application 2: Tularemia Drug Discovery



SFX-data

<u>Titan + Summit</u>





NMR-data

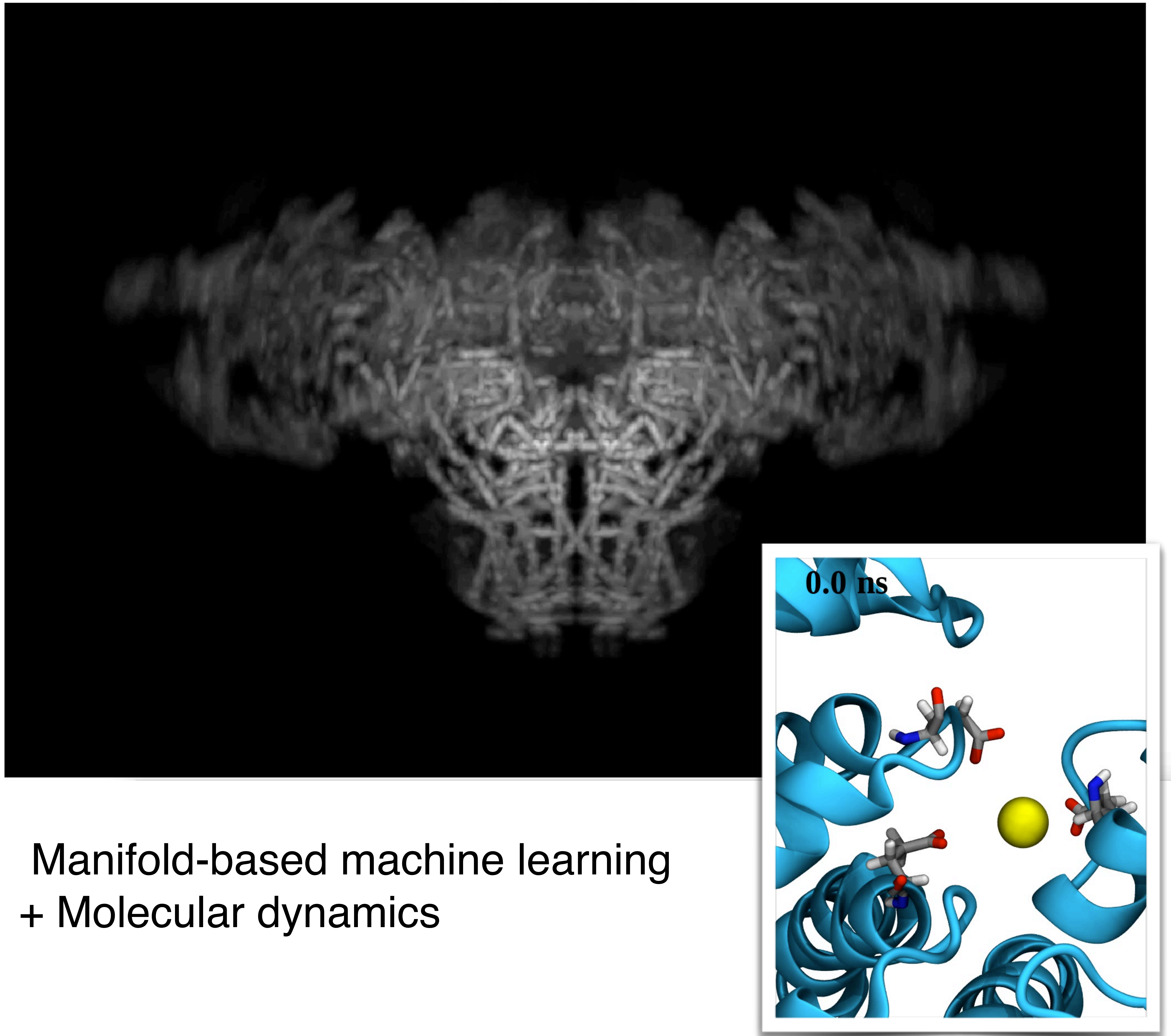
Hamiltonian exchange umbrella sampling

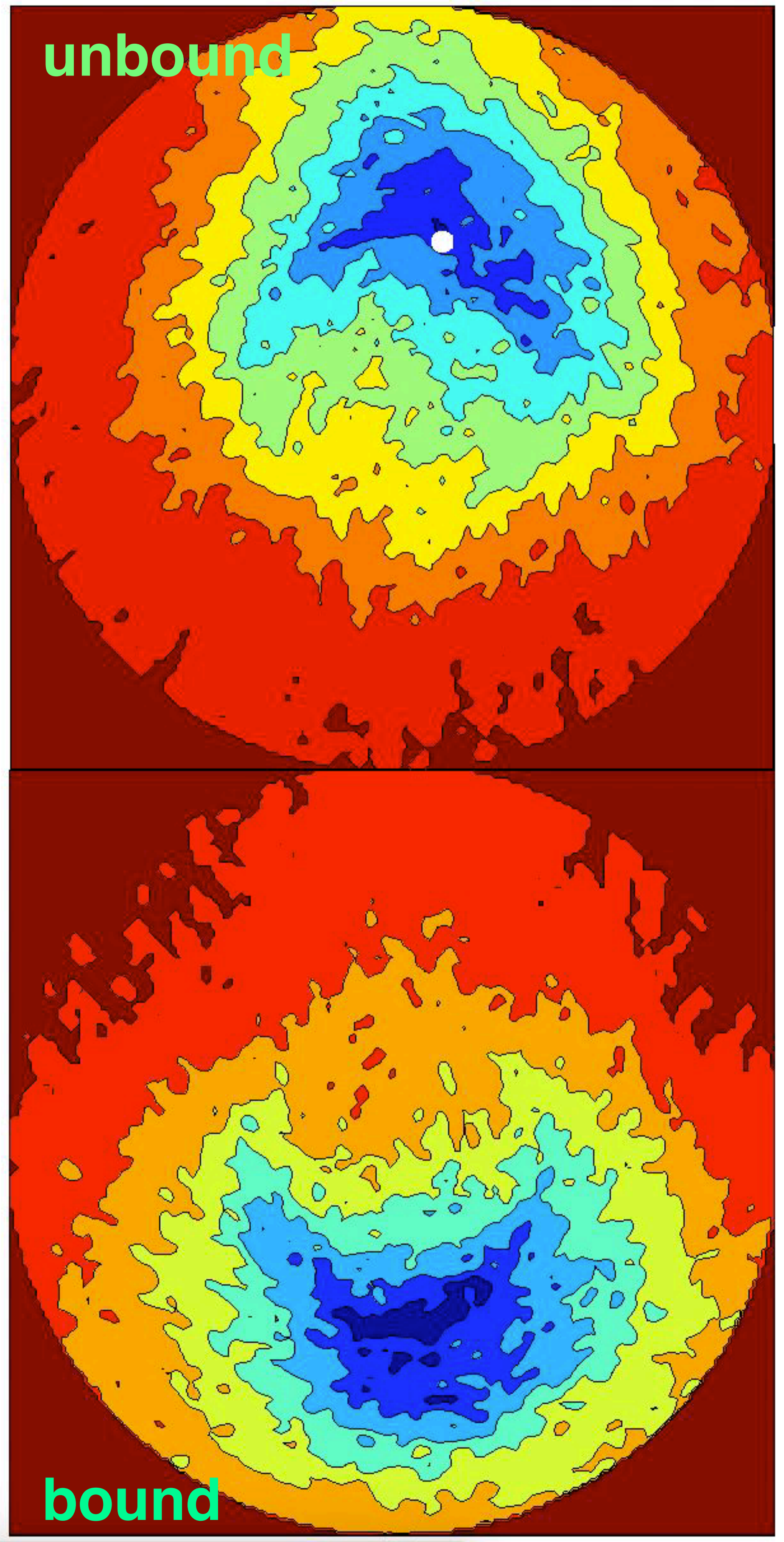
Fromme

Zook

Abhishek Singharoy
ASU

Multi-Copy Application 3: Ion Channels





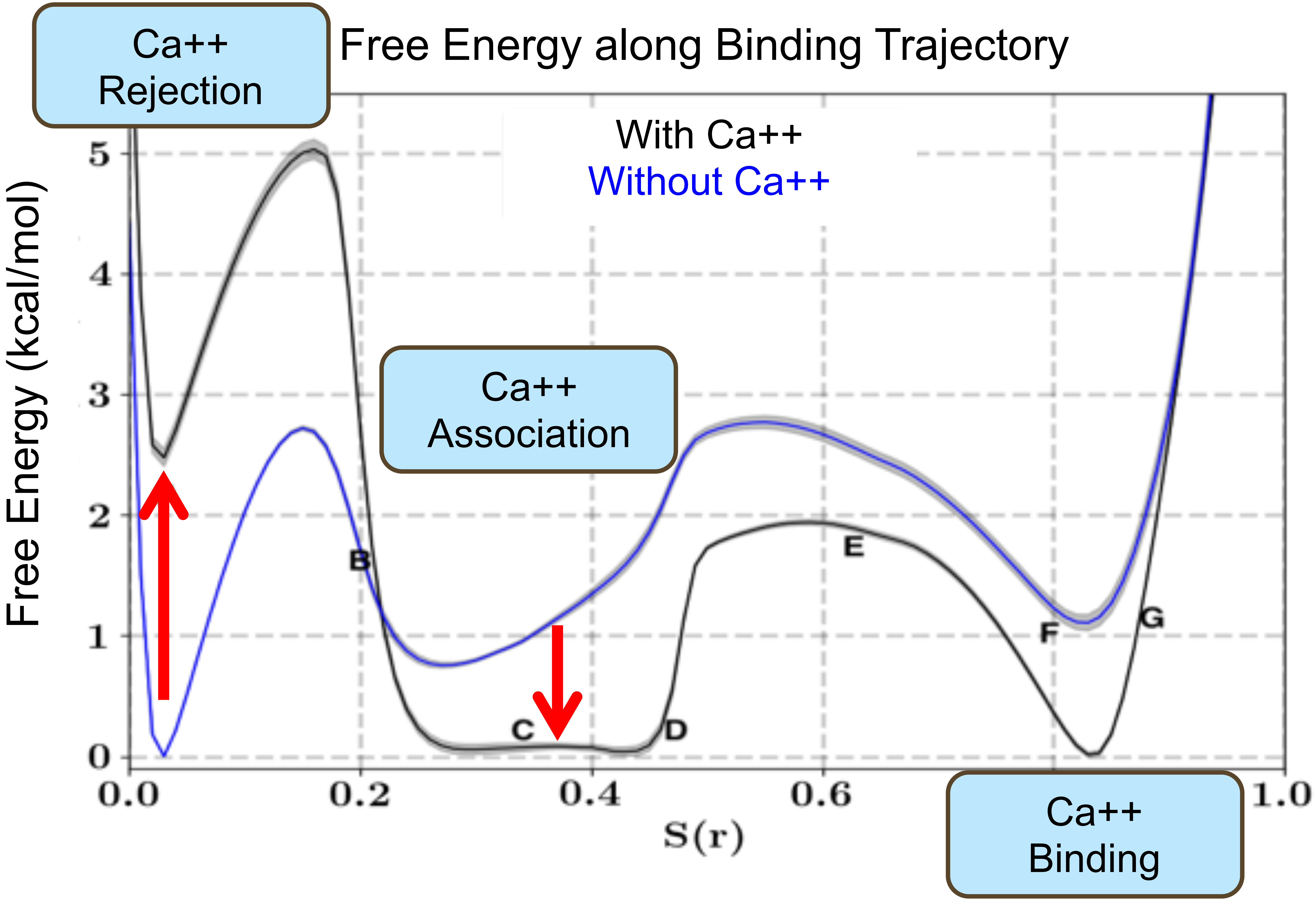
Manifold-based machine learning
+ Molecular dynamics

Summit - friendly

Frank, Singharoy, Ourmazd

Abhishek
Singharoy
ASU

Ensemble-refinement and pathway information



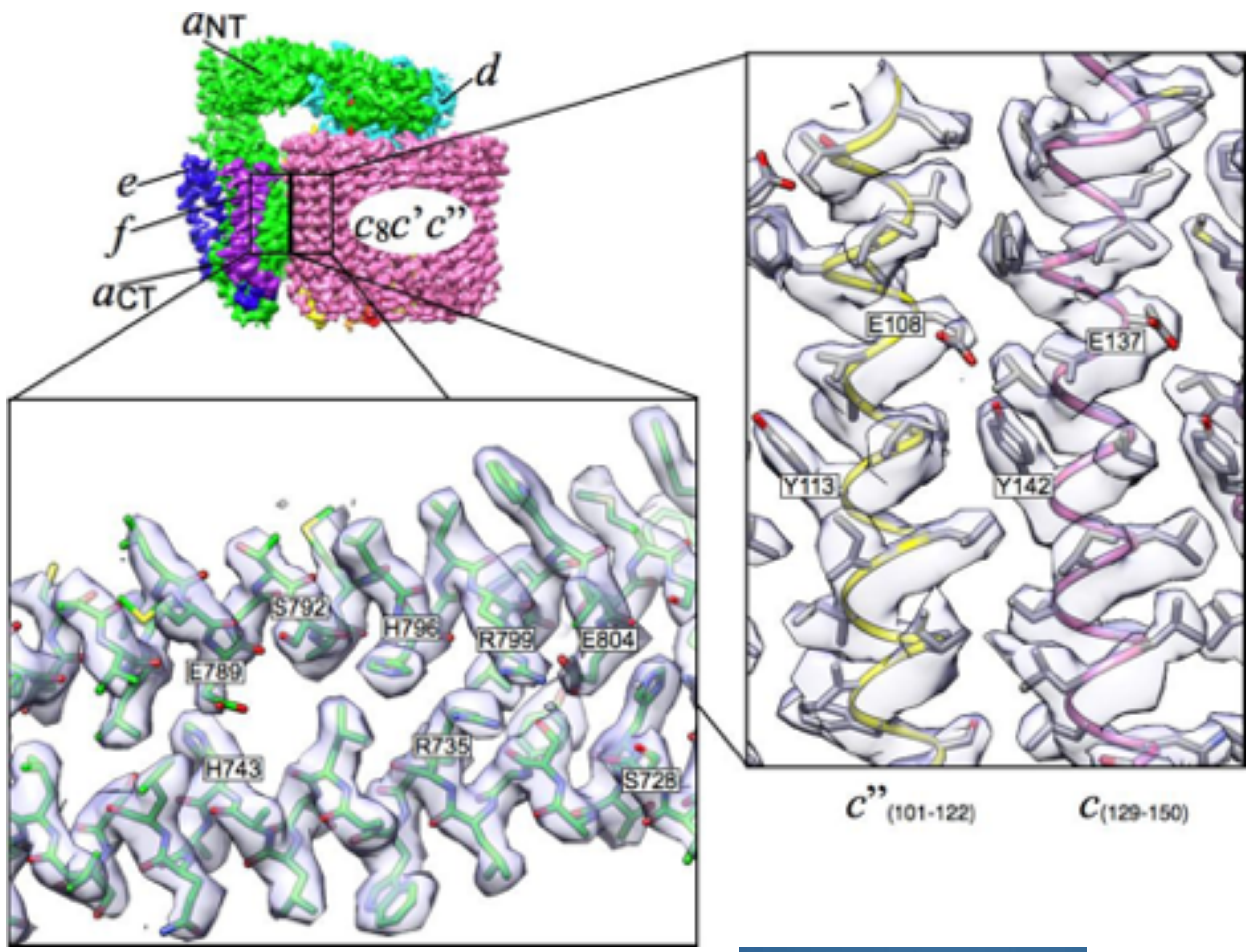
String + Adaptive biasing force

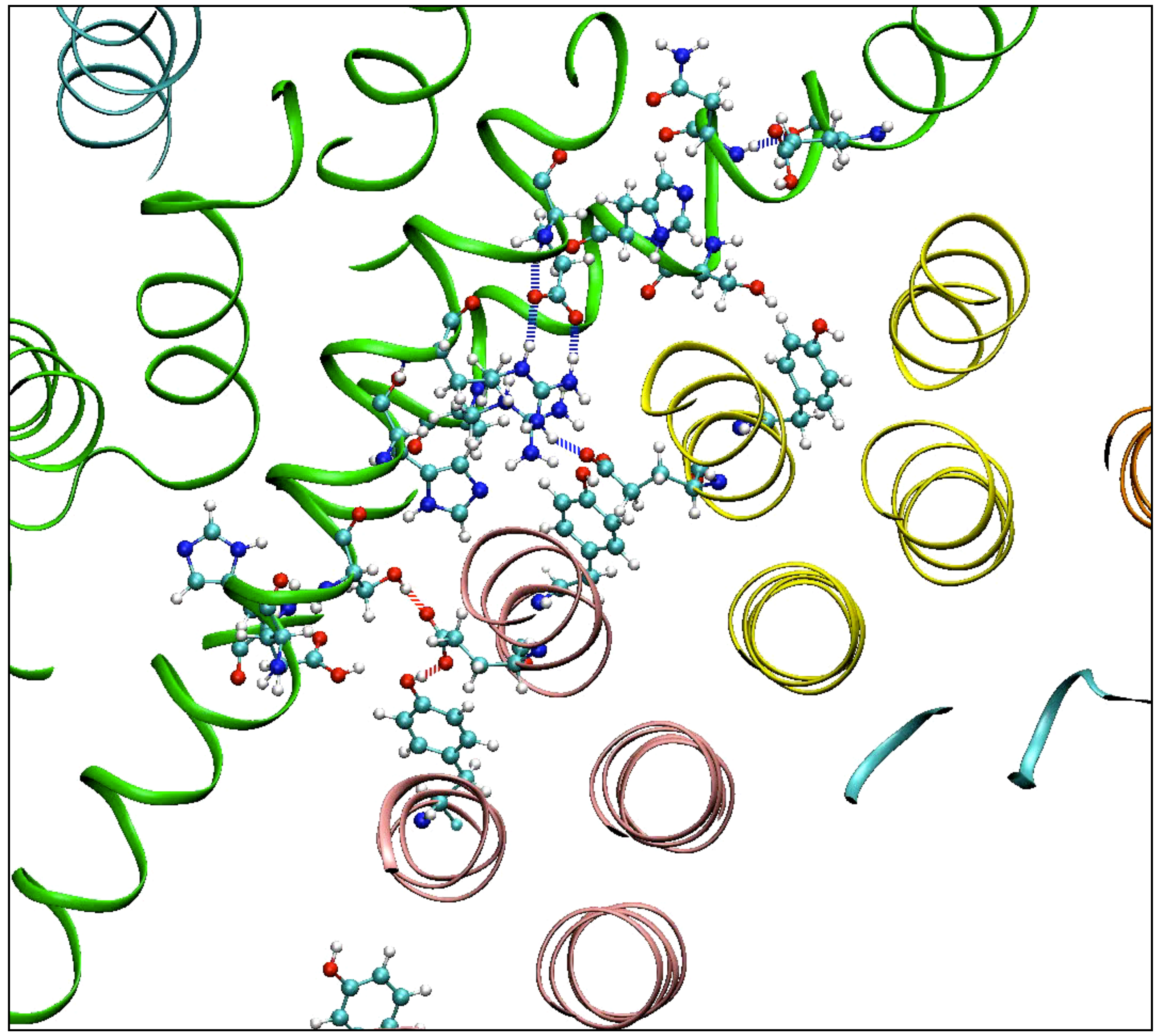
Summit - friendly + ESP

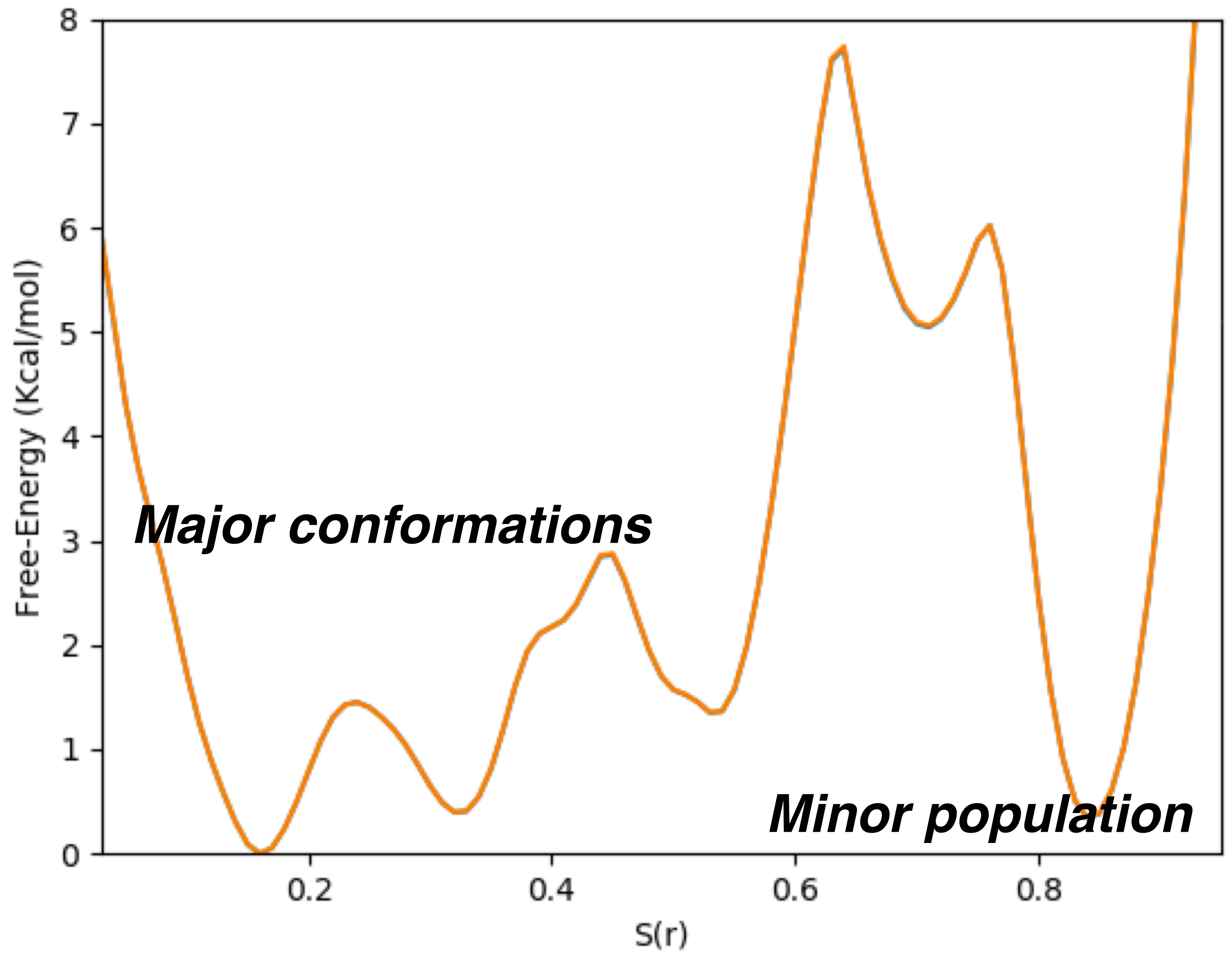
Abhishek Singharoy ASU

Multi-Copy Application 4: ATP Synthase of Tuberculosis-Causing Bacteria

Summit - ESP + INCITE







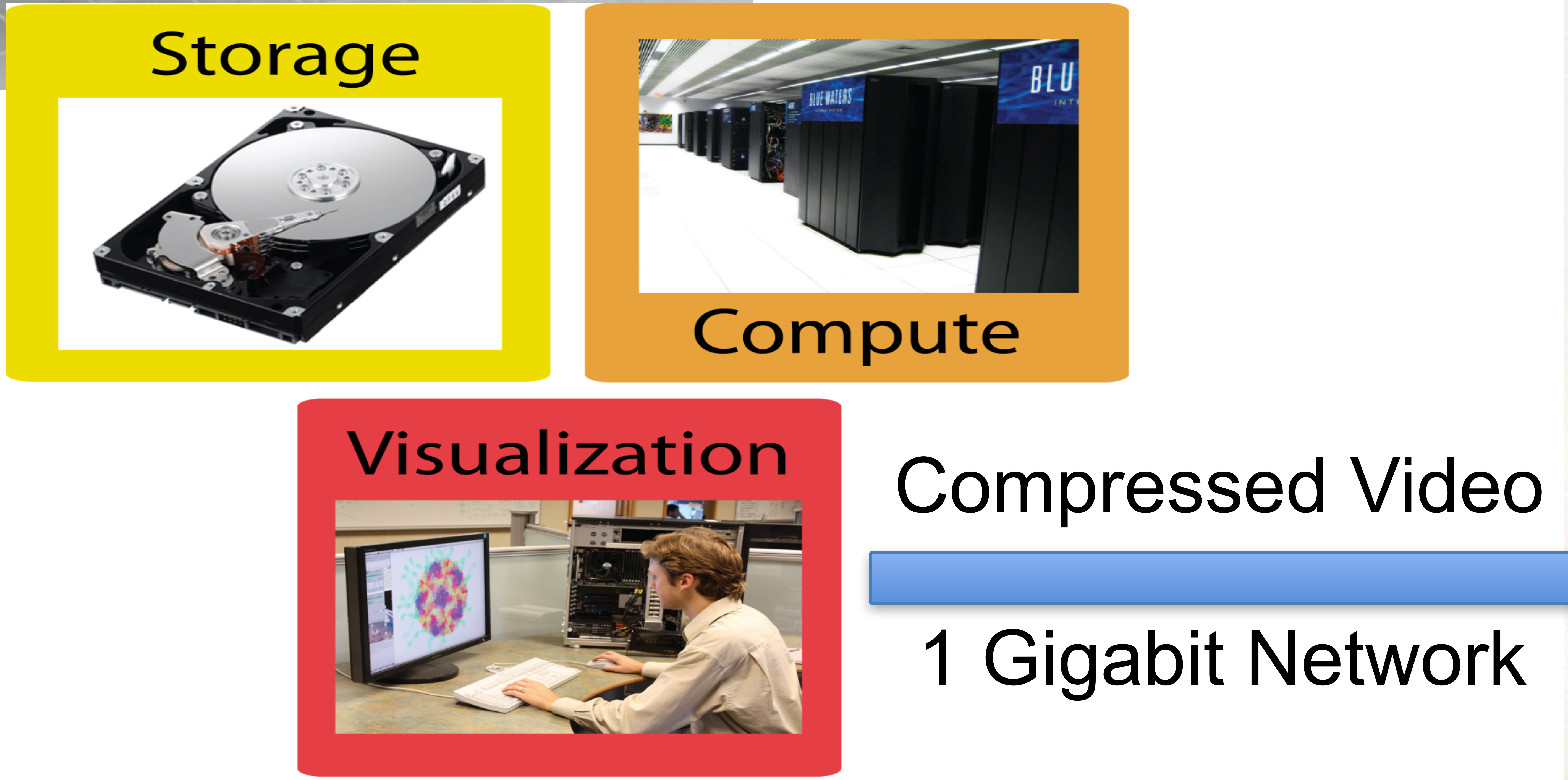
String +
Adaptive biasing force

Chiu, Wilkens

Abhishek
Singharoy
ASU

GPUs are critical for visualization and analysis

Large memory GPU-accelerated remote visualization must be ***embedded at supercomputer centers***. Available now! See bluewaters.ncsa.illinois.edu/dcv and OLCF Rhea docs.



Storage

Compute

Visualization

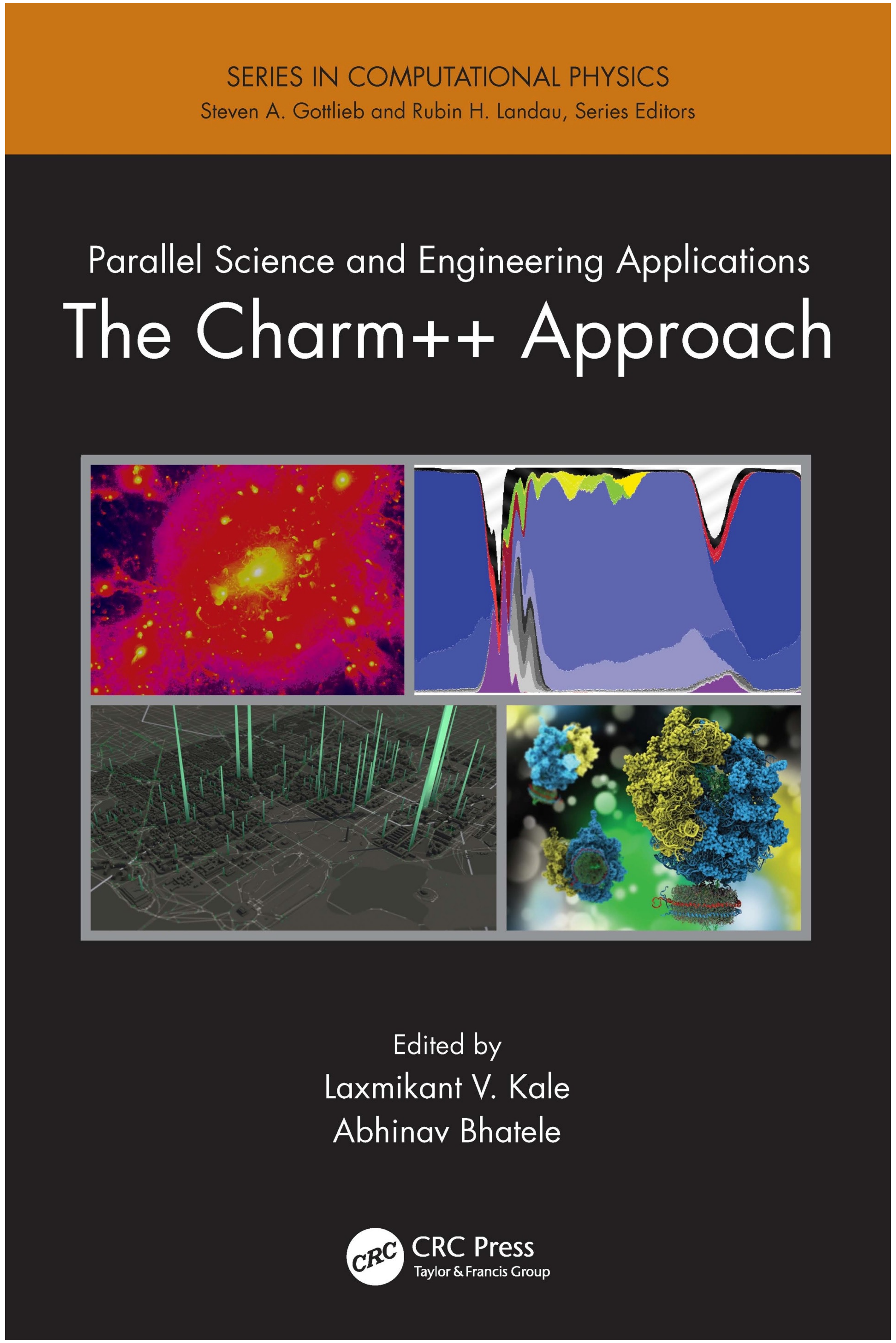
Compressed Video

1 Gigabit Network

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



SERIES IN COMPUTATIONAL PHYSICS
Steven A. Gottlieb and Rubin H. Landau, Series Editors

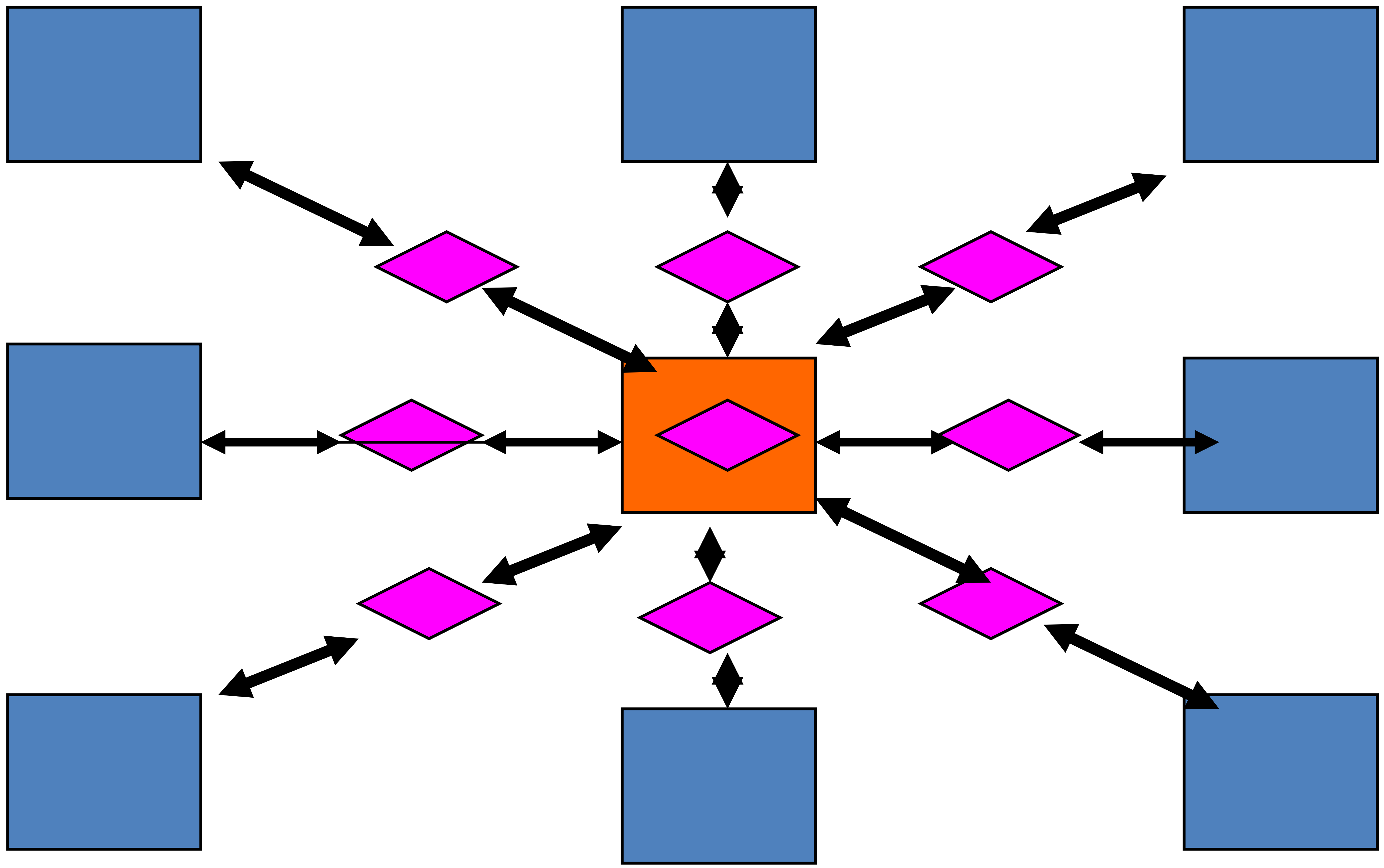
Parallel Science and Engineering Applications
The Charm++ Approach

Edited by
Laxmikant V. Kale
Abhinav Bhatele

CRC Press
Taylor & Francis Group

NAMD Hybrid Decomposition

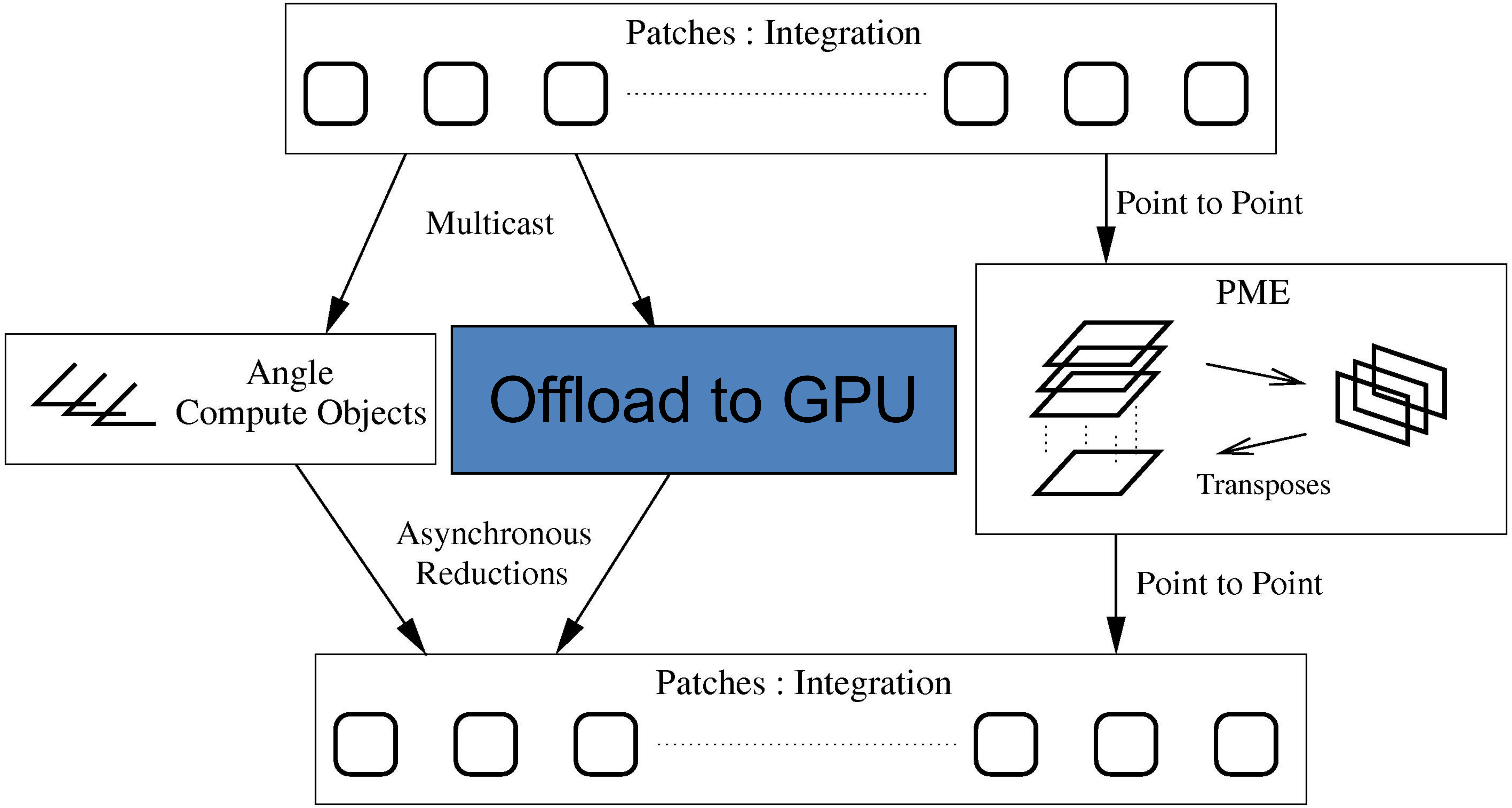
Kale *et al., J. Comp. Phys.* 151:283-312, 1999.



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

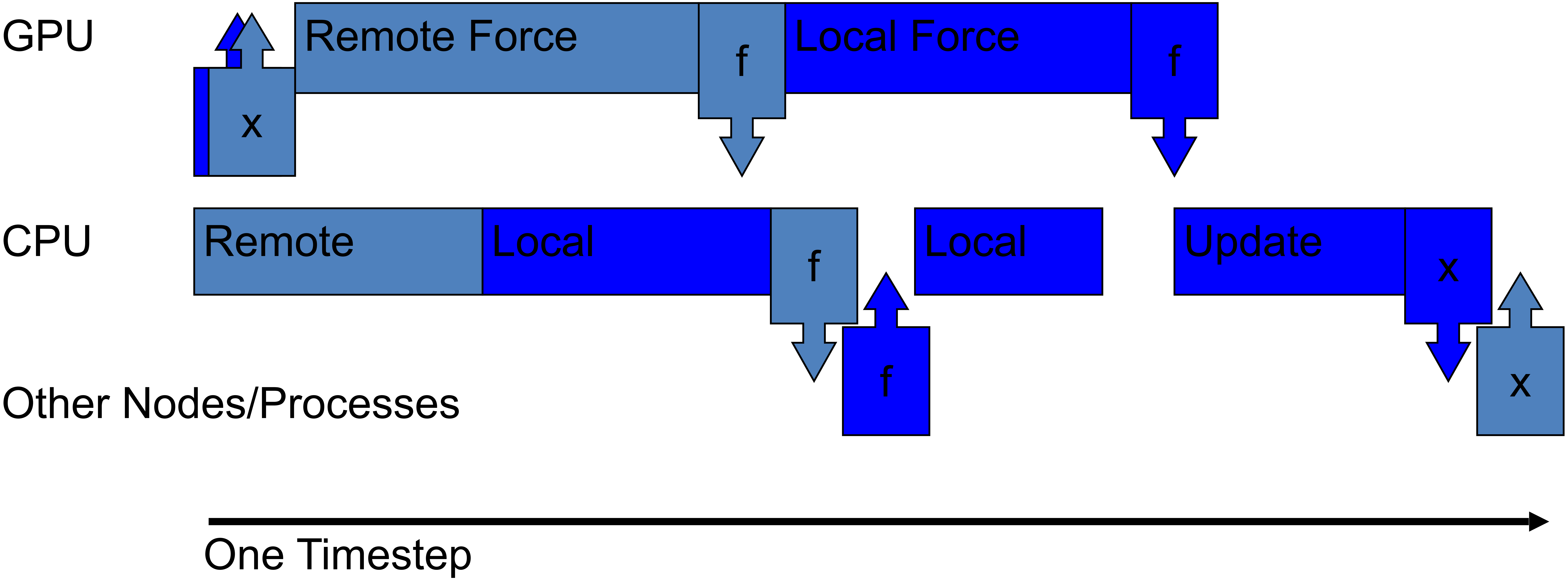
NAMD Overlapping Execution

Phillips *et al., SC2002*.



Objects are assigned to processors and queued as data arrives.

Overlapping GPU and CPU with Communication



GTC 2019

Phillips *et al.*, SC2008

Streaming GPU 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

29,829,000
29,830,000
29,831,000
29,832,000
29,833,000
29,834,000
29,835,000
29,836,000
29,837,000
29,838,000
29,839,000
29,840,000
29,841,000
29,842,000
29,843,000
29,844,000
29,845,000
29,846,000
29,847,000
29,848,000

Charm++ *Projections* performance-analysis tool

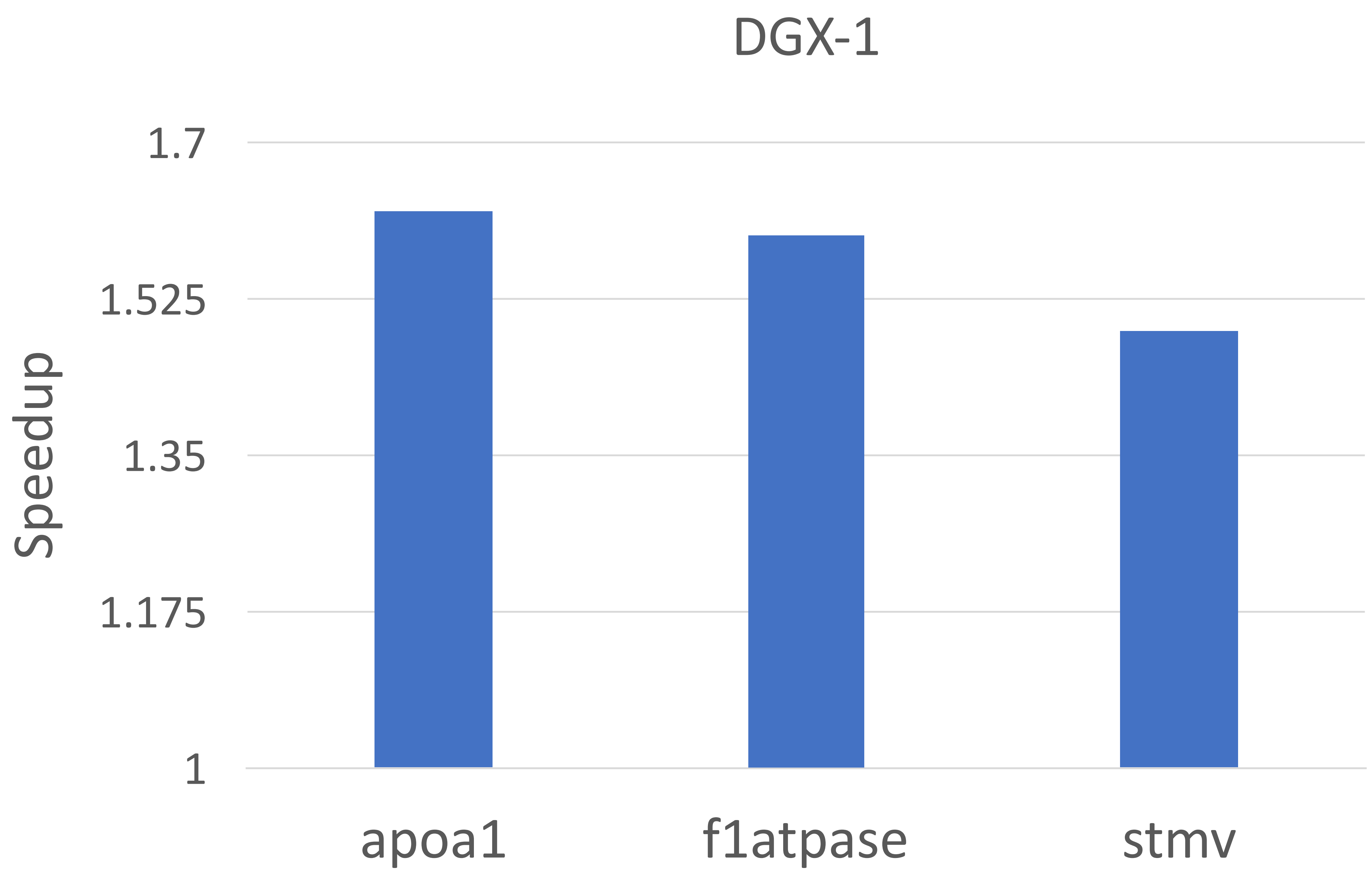
Streaming Kernel

000
28,401,000
28,402,000
28,403,000
28,404,000
28,405,000
28,406,000
28,407,000
28,408,000
28,409,000
28,410,000
28,411,000
28,412,000
28,413,000
28,414,000
28,415,000
28,416,000
28,417,000
28,418,000
28,419,000
28,42

Charm++ *Projections* performance-analysis tool

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.**

NAMD 2.13 released Nov 9

- First release since December 2016, many improvements
- All force calculation now done on GPU
- CUDA 9 and Volta compatibility
- IBM PAMI SMP machine layer
- Support for two-billion-atom simulations
- New constant pH, improved QM-MM
- Improved core binding of CUDA CPU threads
- Improved CUDA error reporting, **print hostname on Cray**

GTC18: Summit has a noise problem - now fixed!

Time In Microseconds
14,200,000
14,240,000
14,280,000
14,320,000
14,360,000
14,400,000
14,440,000
14,480,000
14,520,000
14,560,000
14,600,000
14,640,000
14,680,000
PE 410 (41, 38)
PE 412 (41, 38)
PE 618 (33, 28)
PE 1014 (38, 32)
PE 1446 (34, 30)
PE 2889 (37, 35)
PE 3588 (34, 30)
PE 4074 (35, 30)
PE 4362 (33, 28)
PE 4434 (35, 29)
PE 5046 (34, 29)
PE 5082 (34, 29)
PE 5154 (34, 29)
PE 5172 (33, 28)
PE 5592 (43, 39)
PE 6918 (33, 28)
PE 7224 (33, 29)

80 ms

GTC18 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

GTC19 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~~ **7 threads per process**)
  - ~~Leaving one core free per resource set seems to reduce noise~~
  - One core per socket is reserved by jsrun, so ~~8~~ **2 unused cores per node**
- With thread to core affinity (plus resource-set binding for CUDA thread):
  - jsrun **--bind rs -a1** -r6 -g1 -c7 namd2 +ignoresharing **+ppn 7 +pemap 0-83:4,88-171:4** ~~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~~

“Words of wisdom and comfort on the loss of 90% of your supercomputer performance”

or

“When bad OS updates happen to good scientific applications”

Helpful Activities

- DON’T PANIC
- Recompile
- Try MPI instead of PAMI communication layer
- Report issue to user support
- Periodically ask for updates
- Escalate at every opportunity
- Allow unaffected multi-copy early science to run

Neutral Activities

- Blame <vendor>
- Curse <vendor>
- Wonder if this is related to your contact leaving
- Hope she wasn’t the only one who knows the code
- “Not my circus, not my monkeys.”
- “No, I will not fix your supercomputer.”
- Update Charm++ to bleeding edge…

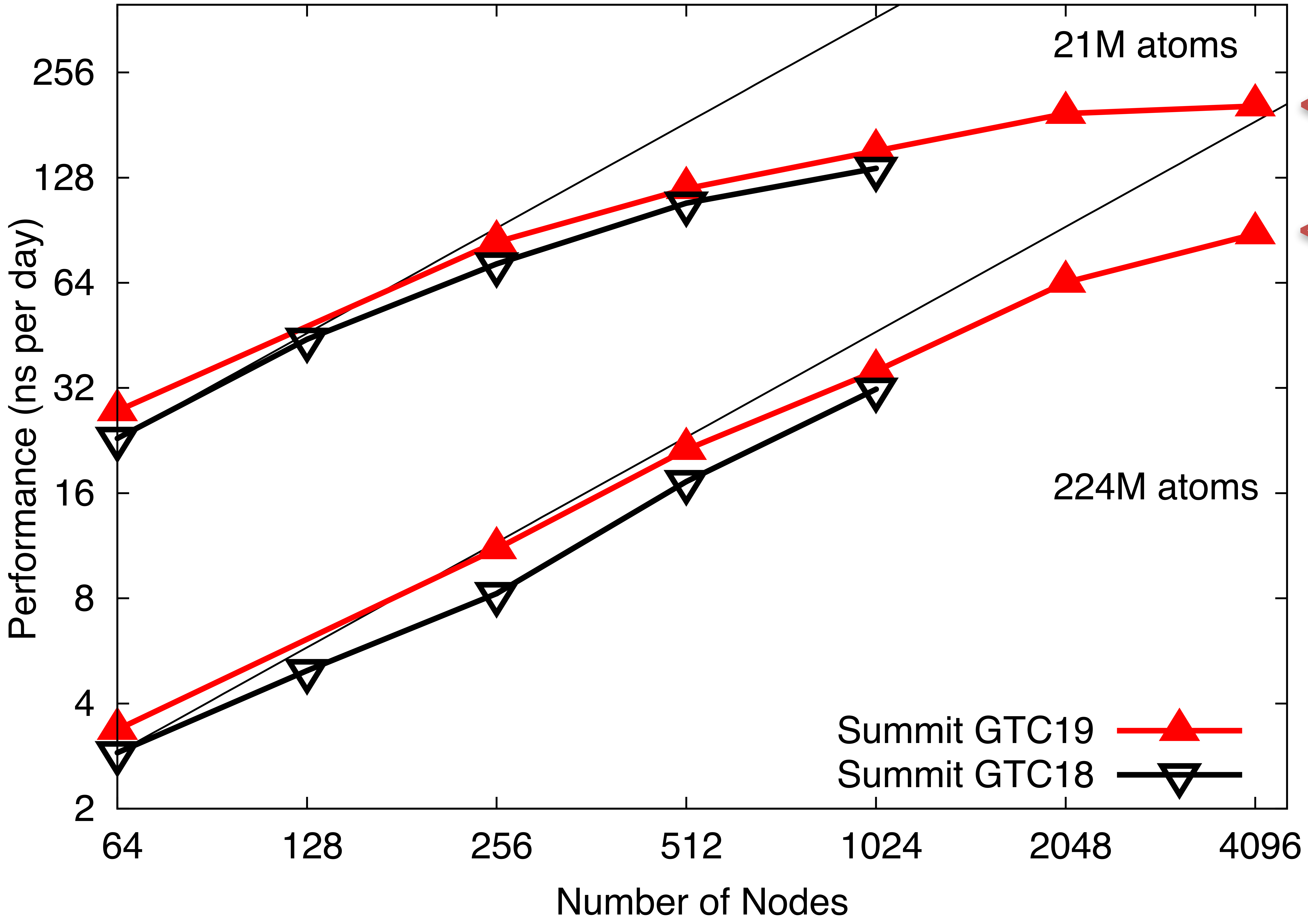
Unhelpful Activities

- Forget you updated Charm++
- Blame instability with new Charm++ on compiler
- Change integrator build flag to -O0 as workaround
- Forget you changed build flag to -O0
- When <vendor> fixes PAMI library, don’t check performance until Friday before GTC
- Fantasize about throwing <vendor> under bus

Helpful Activities (2)

- Remember -O0 change to integrator
- Realize binary from November works fine now
- Notice compiler from November is still available
- Notice compiler from November doesn’t work now
- Realize that Charm++ from November works
- “git log src/archpami-linux-ppc64le”
- “git revert …”

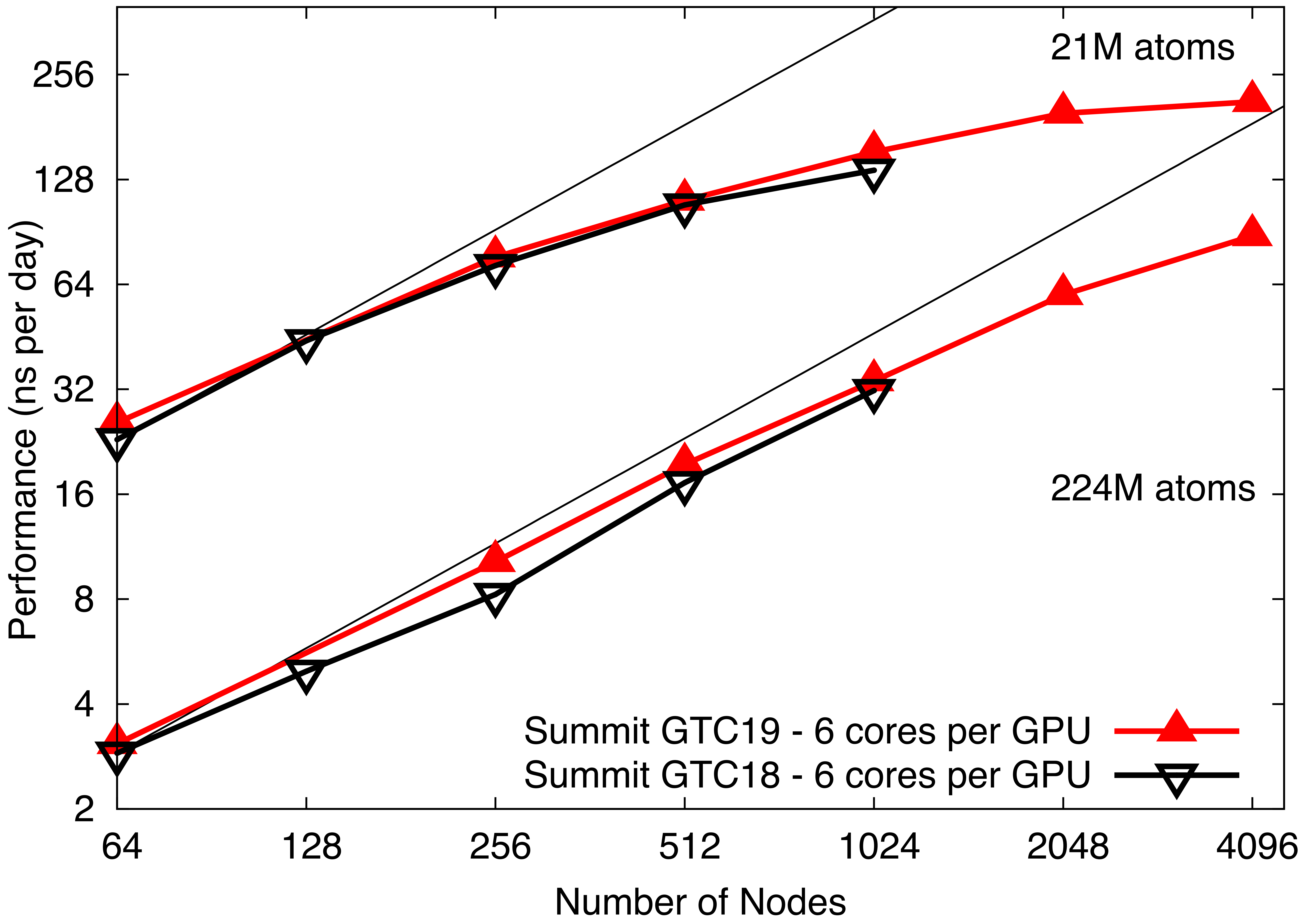
Comparison vs GTC 2018



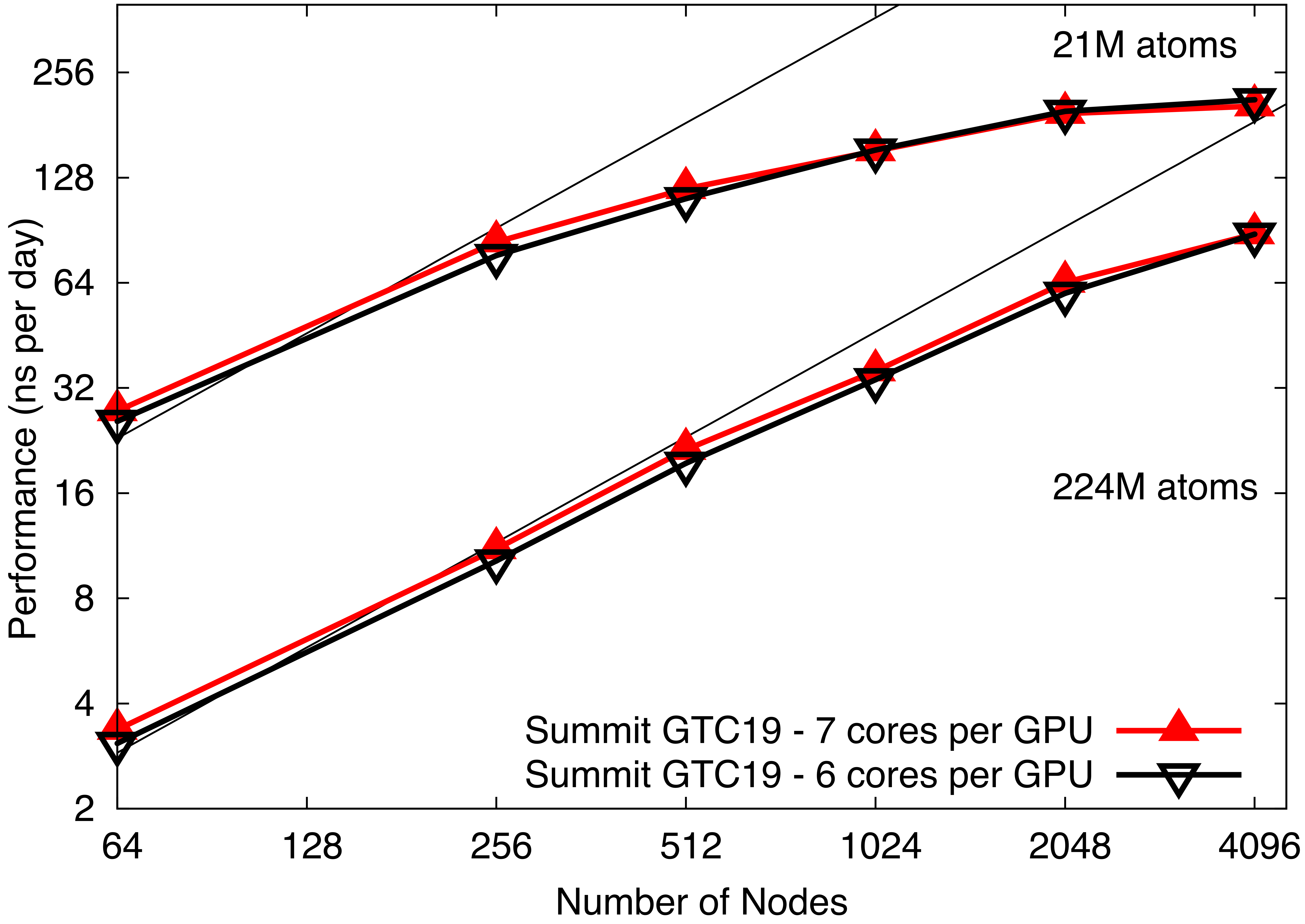
0.8 ms/step
210 ns/day

2.0 ms/step
90 ns/day

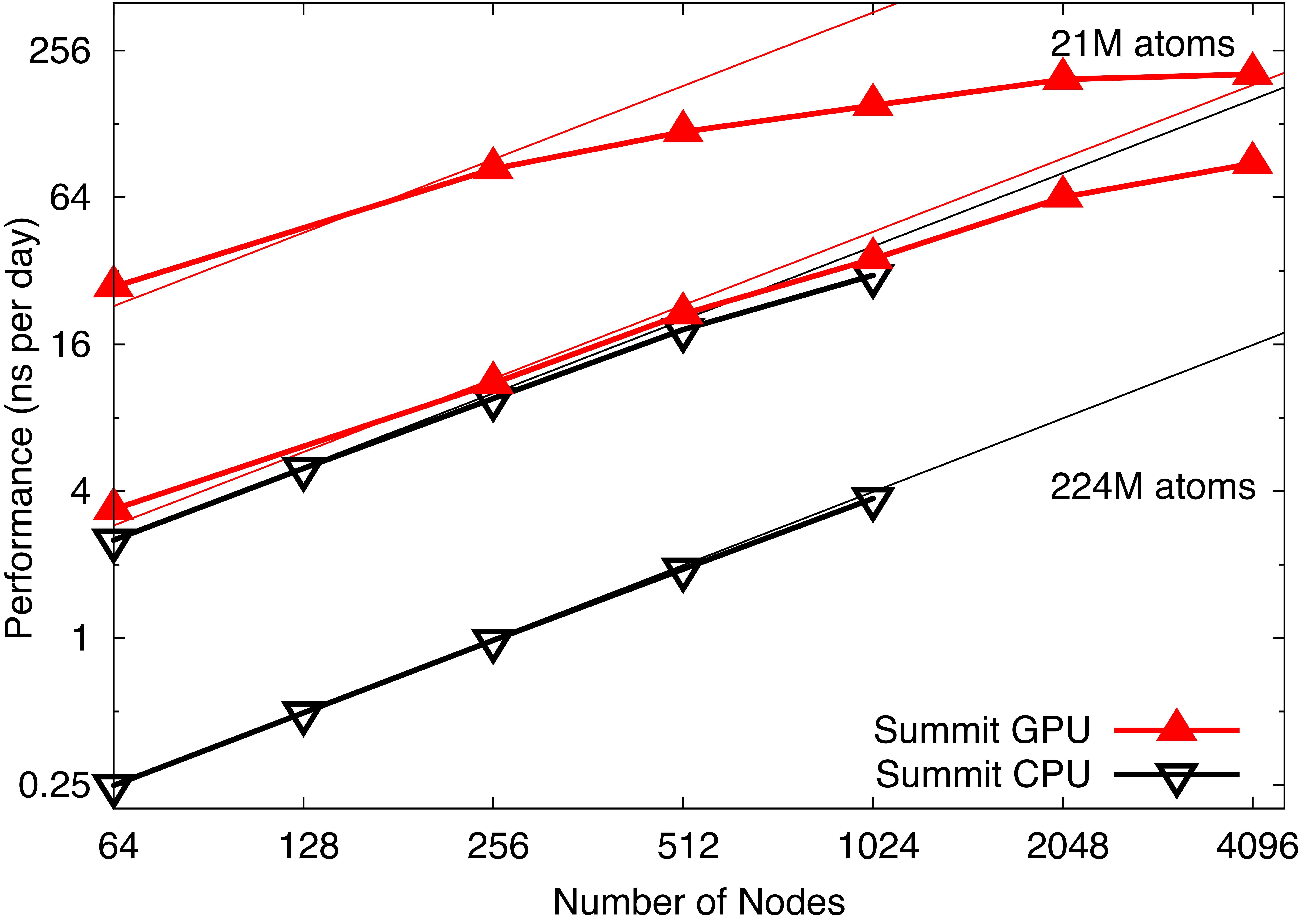
Fairer Comparison vs GTC 2018



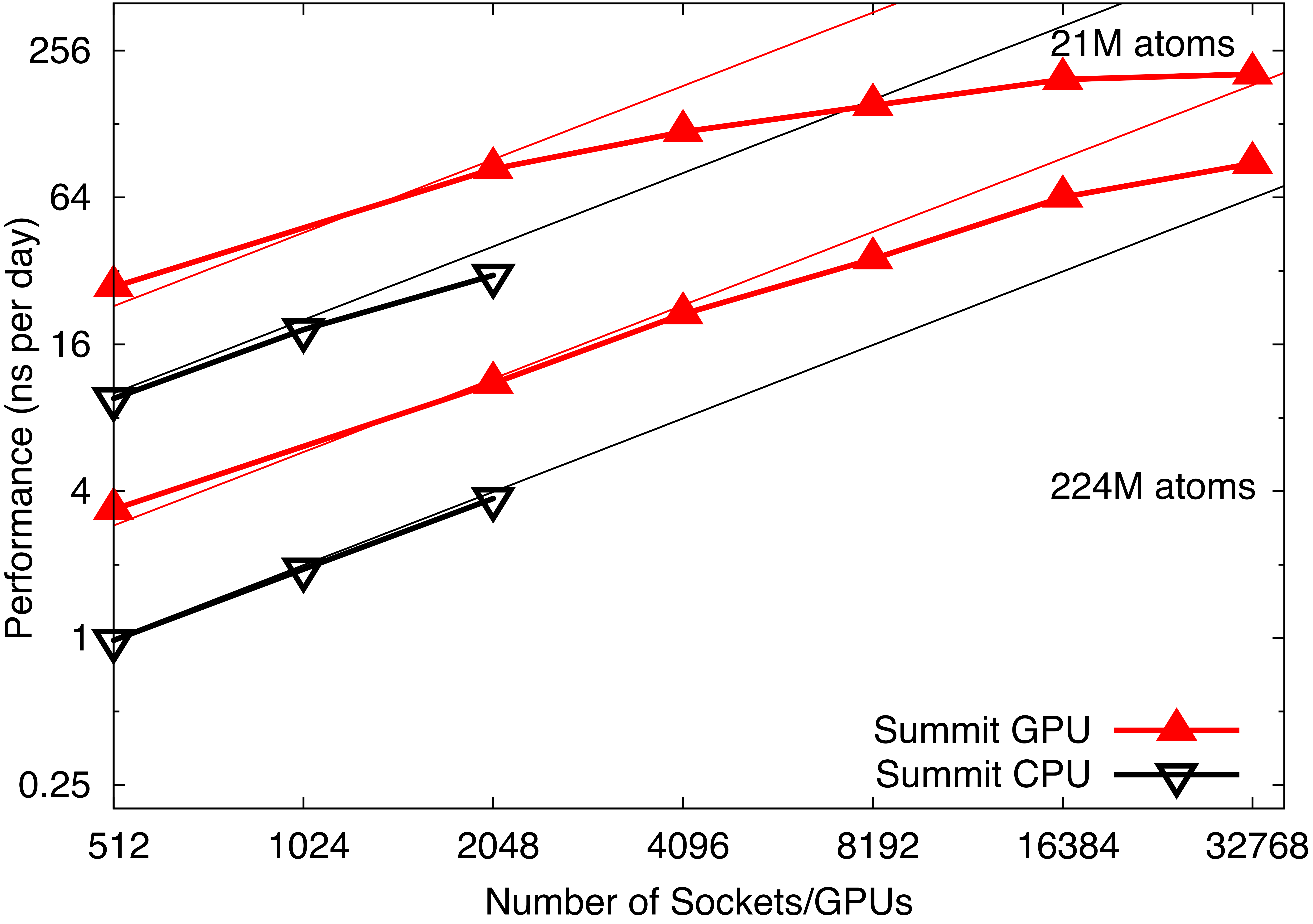
Comparison 7 vs 6 Cores per GPU



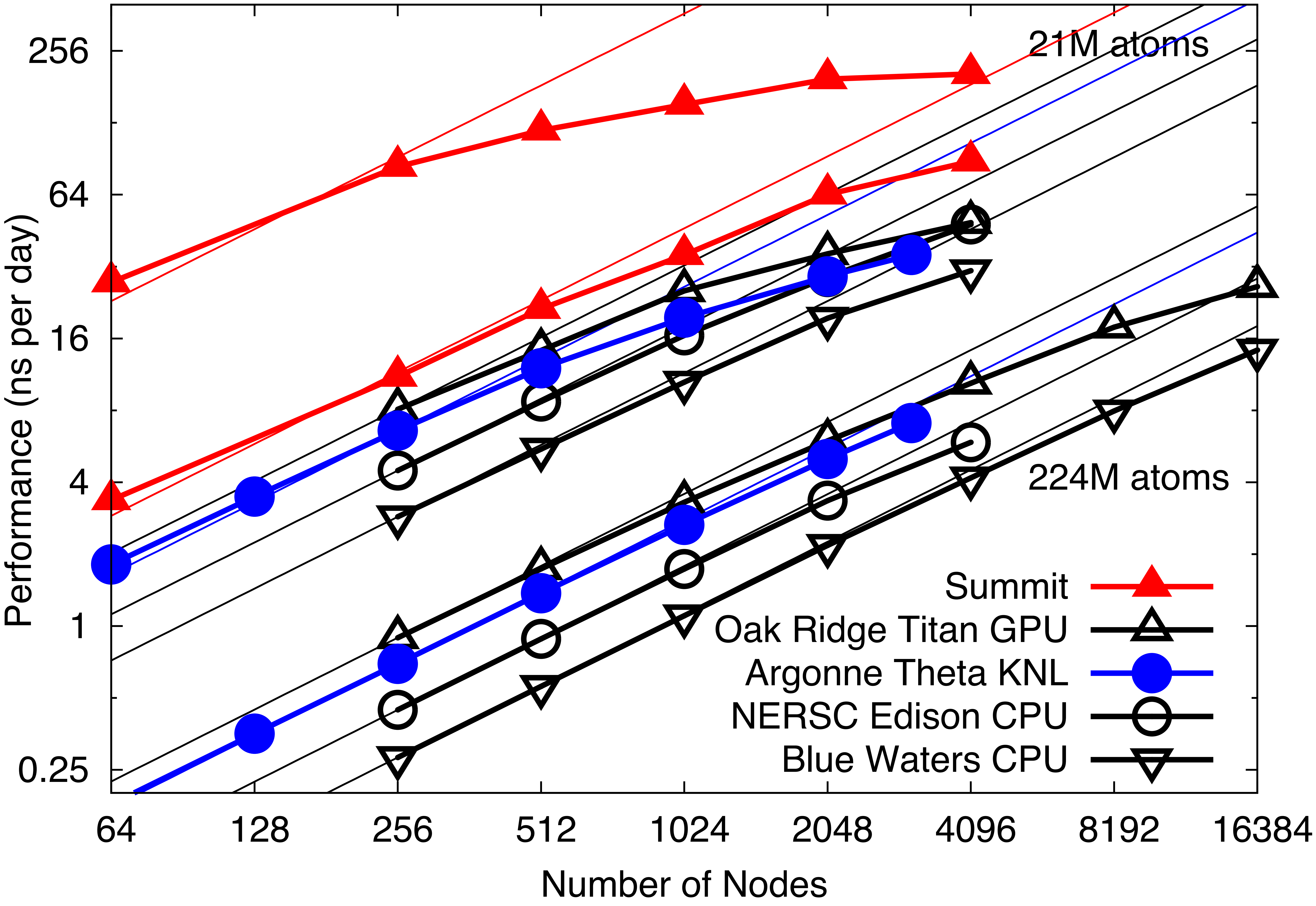
Comparison for large benchmarks



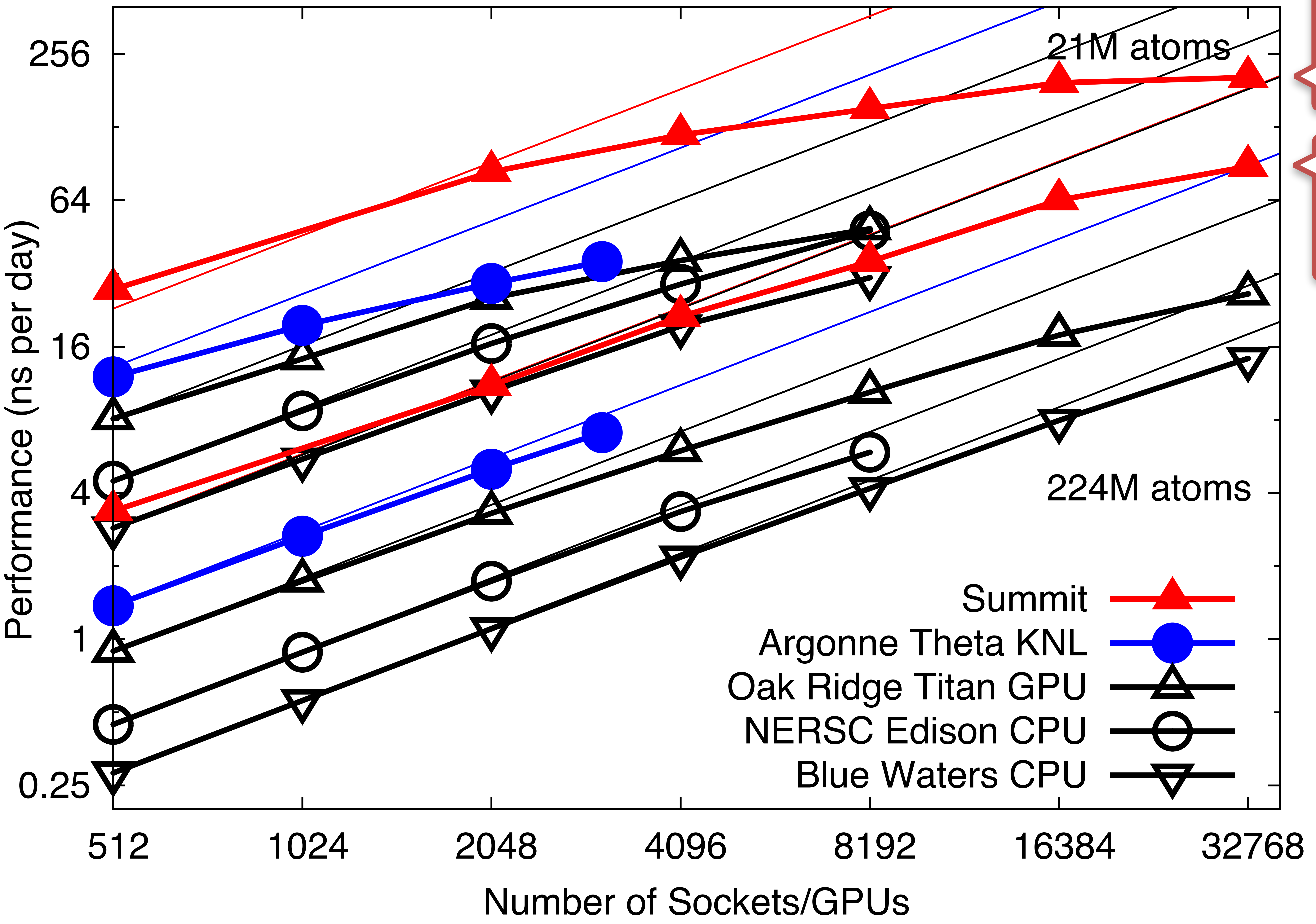
“Fair” comparison for large benchmarks



Comparison for large benchmarks



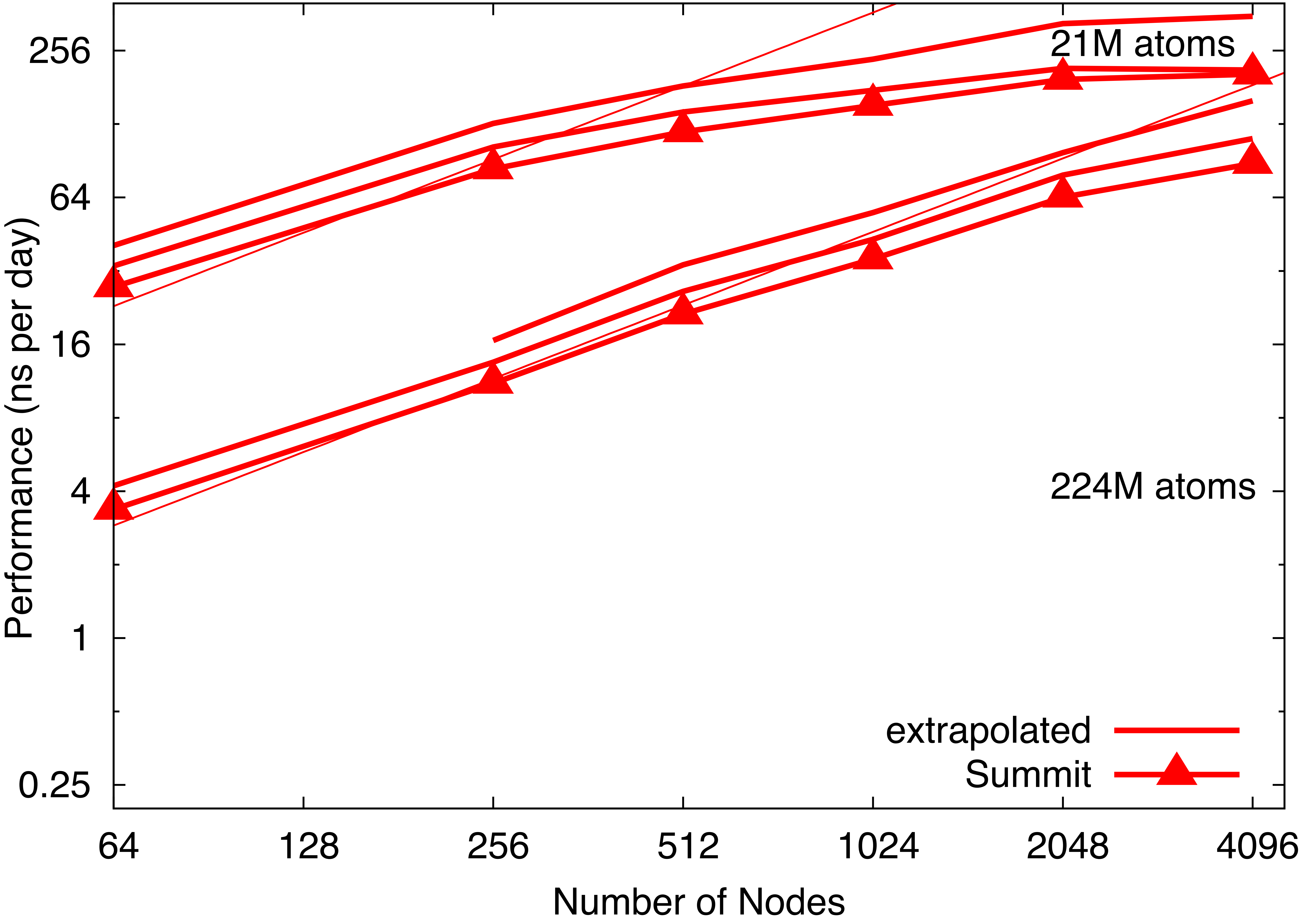
“Fair” comparison for large benchmarks



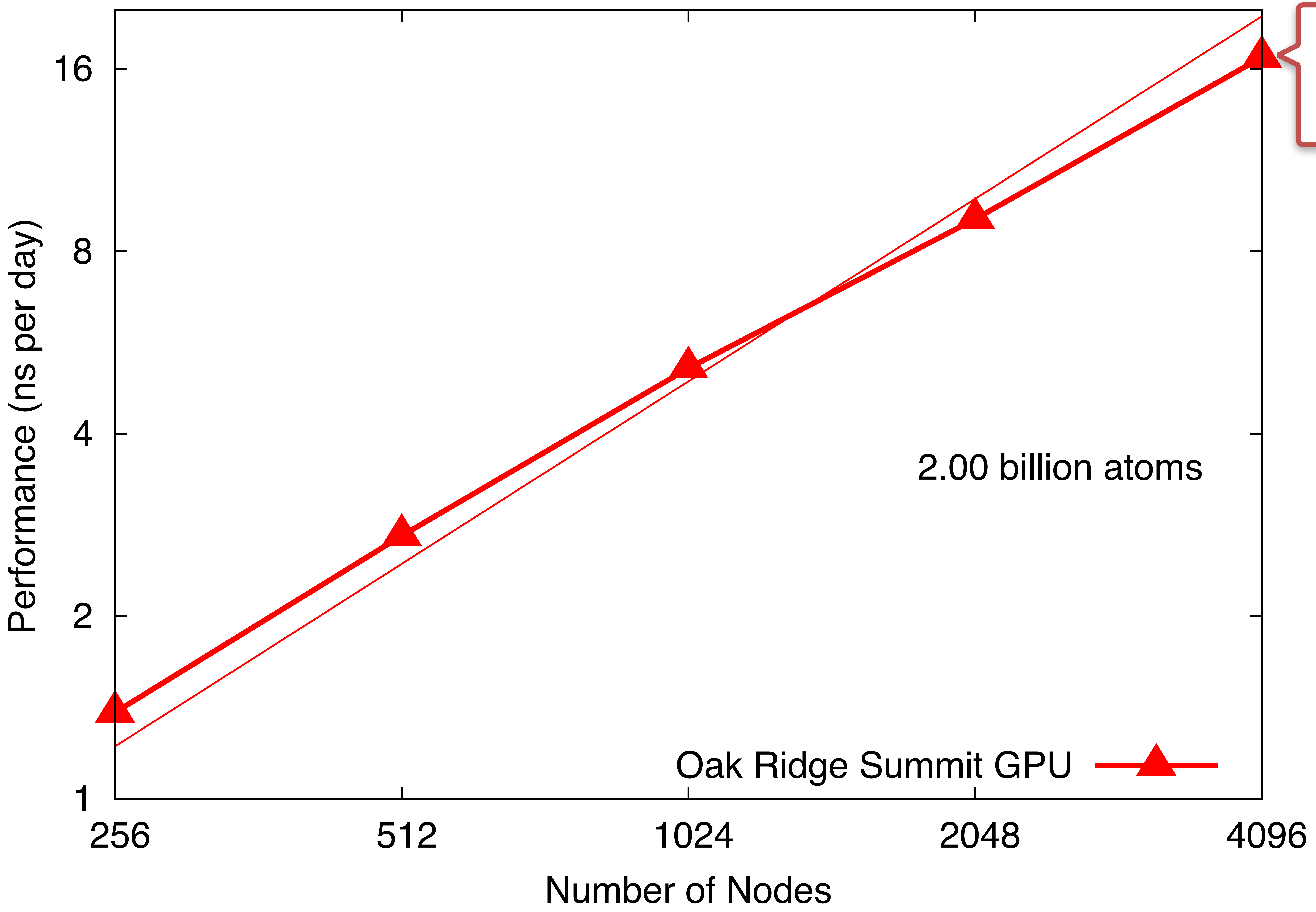
0.8 ms/step
210 ns/day

2.0 ms/step
90 ns/day

“Fix” problems with simpler integrator



Two billion atoms



10 ms/step
17 ns/day

Charm++ *Projections* tool shows bottlenecks

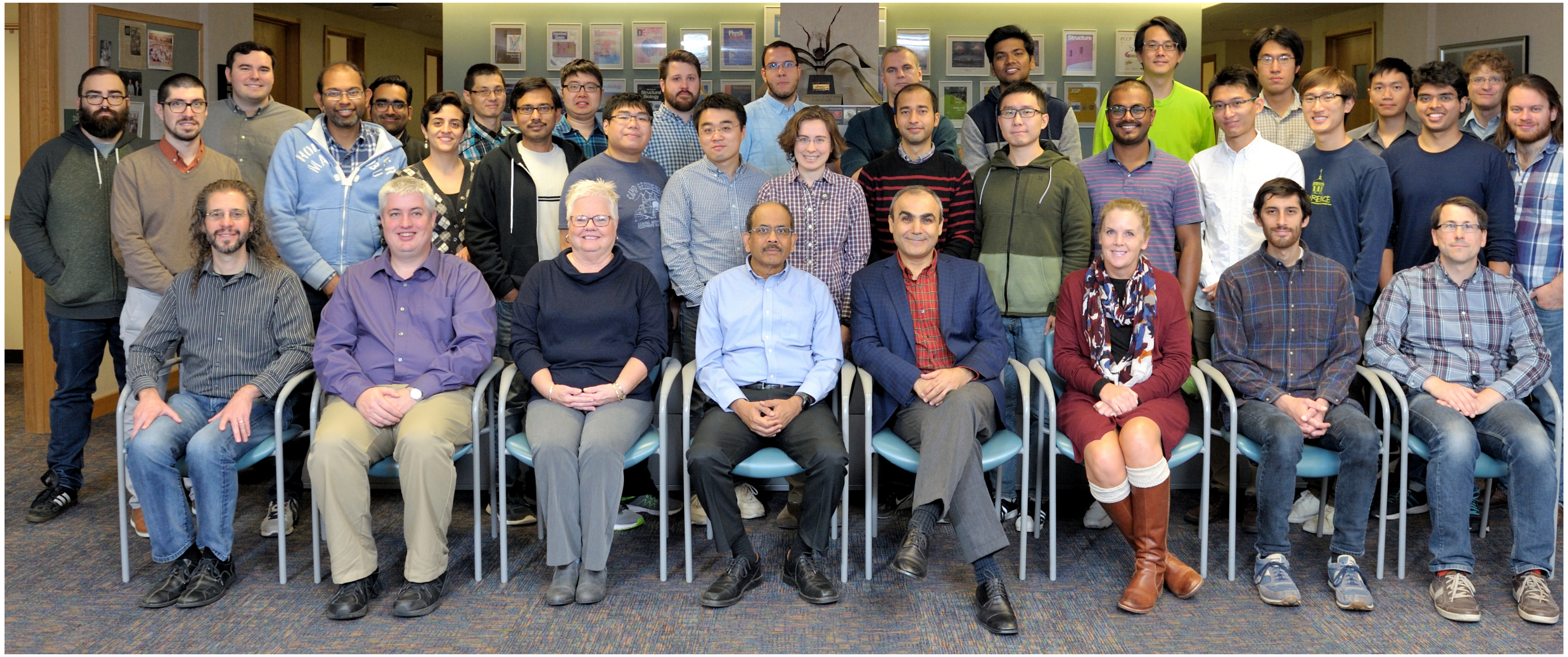
Time In Microseconds
3,000
16,214,000
16,215,000
16,216,000
16,217,000
16,218,000
16,219,000
16,220,000
16,221,000
GPU
Integration
Load Imbalance?

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 now scales well
  - Change is bad, performance regression tests are good

Acknowledgments

Antti-Pekka Hynninen,
Ke Li, & Peng Wang, NVIDIA
Sameer Kumar &
Bilge Acun, IBM
Tjerk Straatsma, OLCF
William Kramer, NCSA
Jodi Hadden, Delaware
Rommie Amaro, UCSD
Lorenzo Casalino, UCSD
Abhi Singharoy, ASU



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

NATIONAL INSTITUTES OF HEALTH

nsf

NCSA

OAK RIDGE
National Laboratory

Argonne
NATIONAL LABORATORY

Related talks

- All earlier today but streaming soon:
  - S9503 - Using Nsight Tools to Optimize the NAMD Molecular Dynamics Simulation Program
  - S9589 - Interactive High-Fidelity Biomolecular and Cellular Visualization with RTX Ray Tracing APIs
  - S9594 - Bringing State-of-the-Art GPU-Accelerated Molecular Modeling Tools to the Research Community