Heterogenous Computing with **Titan**

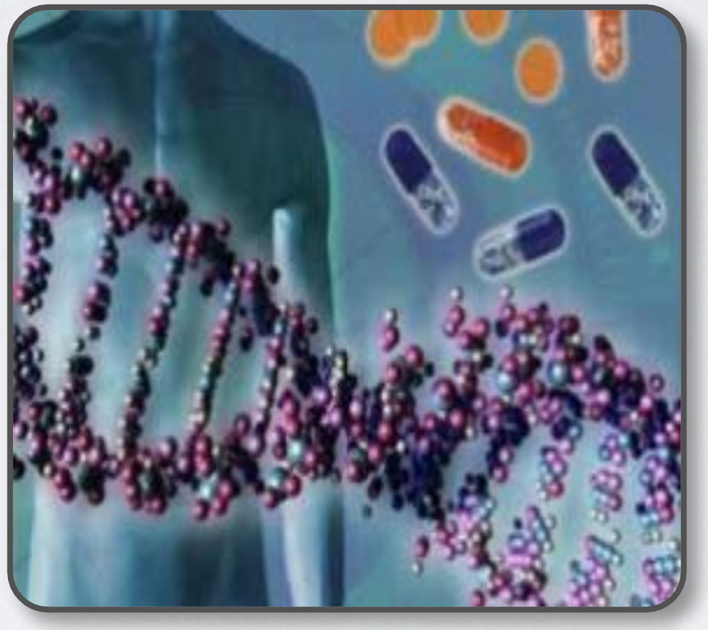
Fernanda Foertter
Oak Ridge Leadership Computing Facility (OLCF)

20 Years of Excellence in Computational Science

OLCF
OAK RIDGE LEADERSHIP COMPUTING FACILITY
1992–2012

BIG PROBLEMS REQUIRE BIG SOLUTIONS

Energy

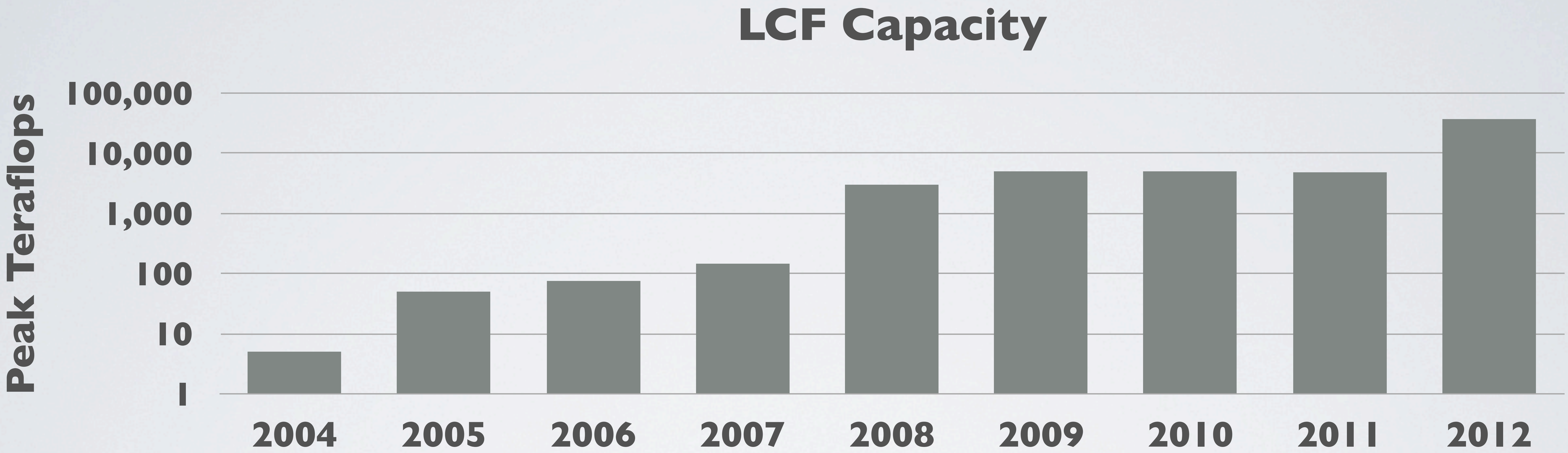


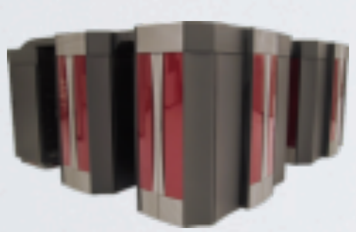
Healthcare

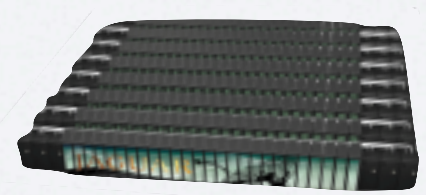
Competitiveness

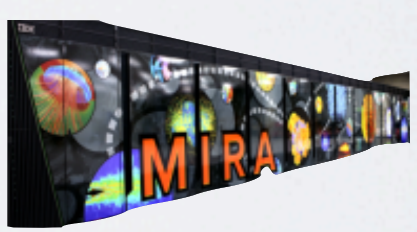
INCREASED OUR SYSTEM CAPABILITY BY 10,000X

since 2004







SCIENCE BREAKTHROUGHS AT THE LCF

2x allocated vs requested (2007–2010)

3x allocated vs requested (2011–2013)

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Hours allocated	4.9M	6.5M	18.2M	95M	268M	889M	1.6B	1.7B	1.7B	5B
Projects	3	3	15	45	55	66	69	57	60	61

2005: Researchers solved the 2D Hubbard model and presented evidence that it predicts HTSC behavior, ***Phys. Rev. Lett*** *(2005)* .

2008: Modeling of molecular basis of Parkinson's disease named #1 computational accomplishment, ***Breakthroughs*** *(2008).*

2009:

Largest simulation of a galaxy's worth of dark matter, showed for the first time the fractal-like appearance of dark matter substructures, ***Nature*** (2008), ***Science*** *(2009).*

World's first continuous simulation of 21,000 years of Earth's climate history, ***Science*** *(2009).*

Largest-ever LES of a full-sized commercial combustion chamber used in an existing helicopter turbine, **Compte Rendus de Mecanique** (2009).

2010:

OMEN breaks the petascale barrier using more than 220,000 cores, ***Proceedings SC10.***

Unprecedented simulation of magnitude-8 earthquake over 125-square miles, ***Proceedings SC10.***

2011: New method to rapidly determine protein structure, with limited experimental data, ***Science*** *(2010)*, ***Nature*** *(2011).*

2012:

NIST proposes new standard reference materials from LCF concrete simulations ***Eur Phys J E Soft Matter*** *(2012).*

Calculation of the number of bound nuclei in nature, ***Nature*** *(2012).*

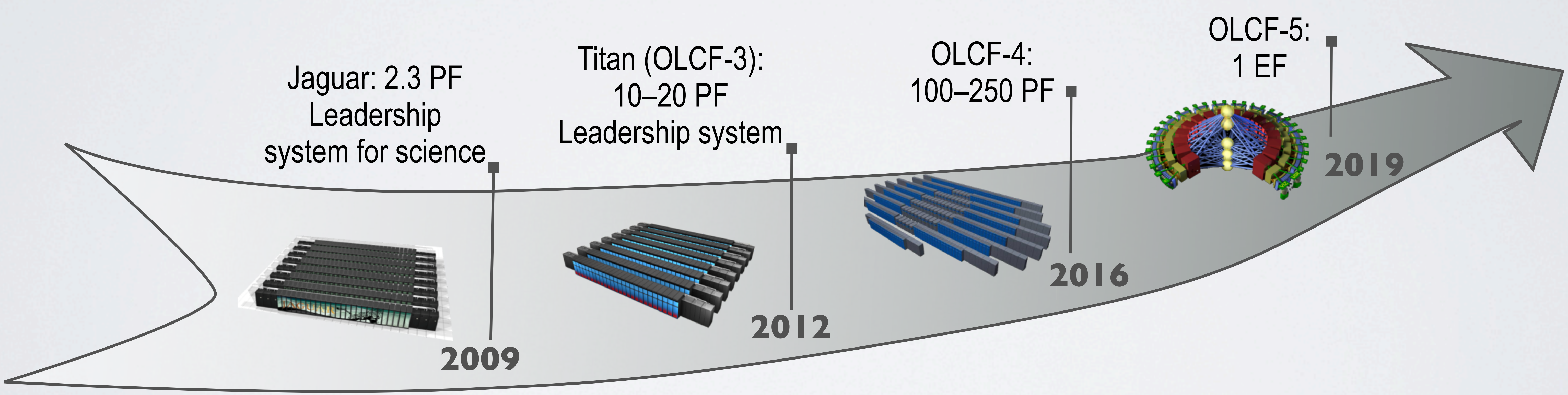
SCIENCE REQUIRES EXASCALE CAPABILITY THIS DECADE

Mission: Deploy and operate the computational resources required to tackle global challenges

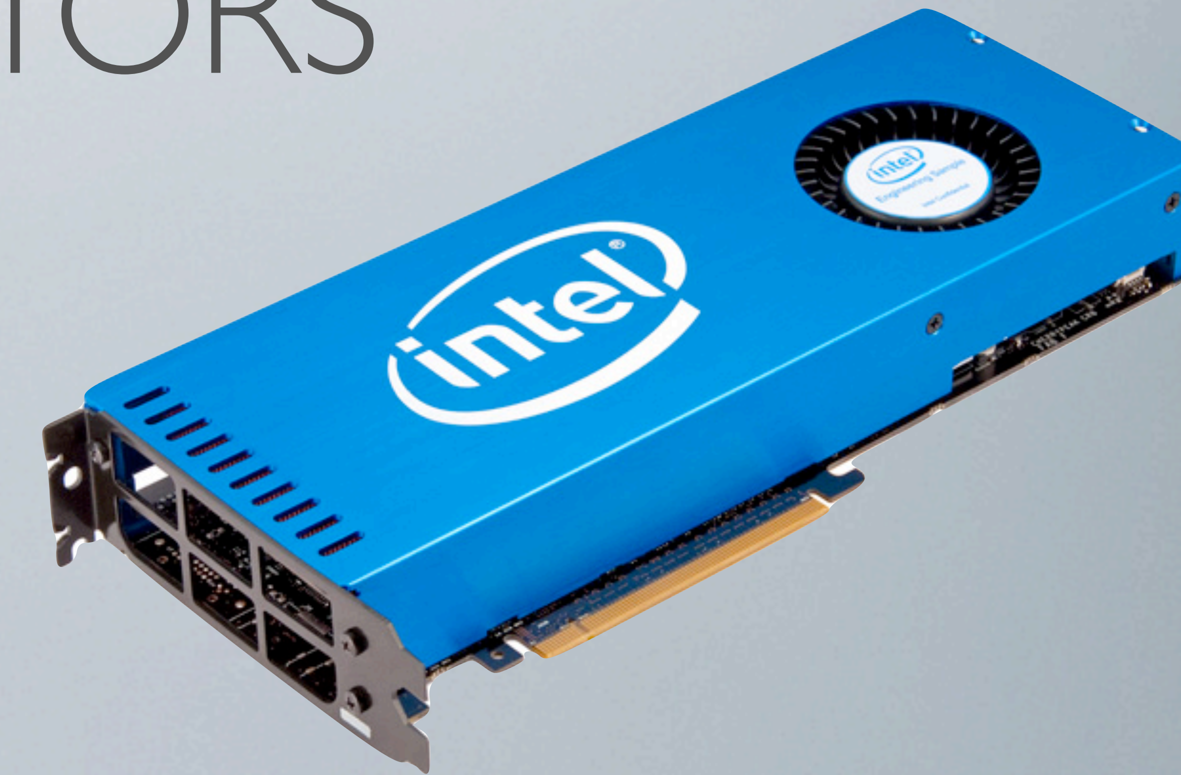
- Deliver transforming discoveries in climate, materials, biology, energy technologies, etc.
- Enabling investigation of otherwise inaccessible systems, from regional climate impacts to energy grid dynamics

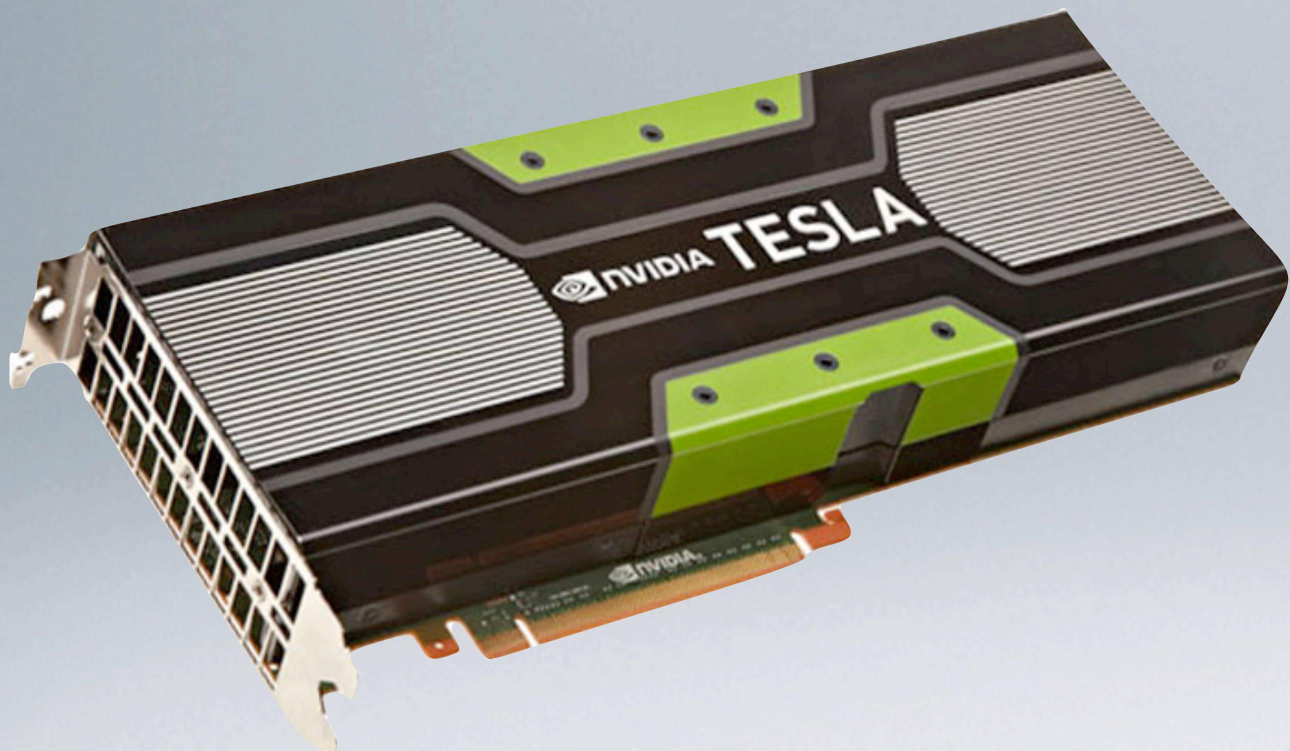
Vision: Maximize scientific productivity and progress on largest scale computational problems

- World-class computational resources and specialized services for the most computationally intensive problems
- Stable hardware/software path of increasing scale to maximize productive applications development

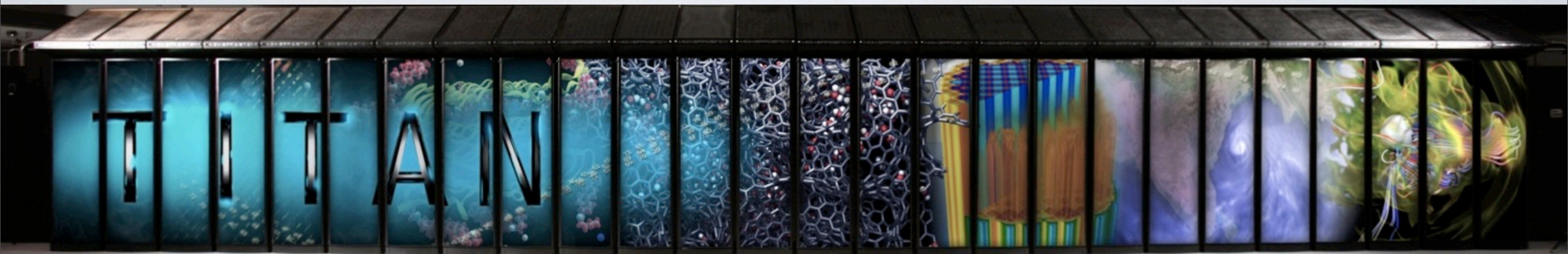


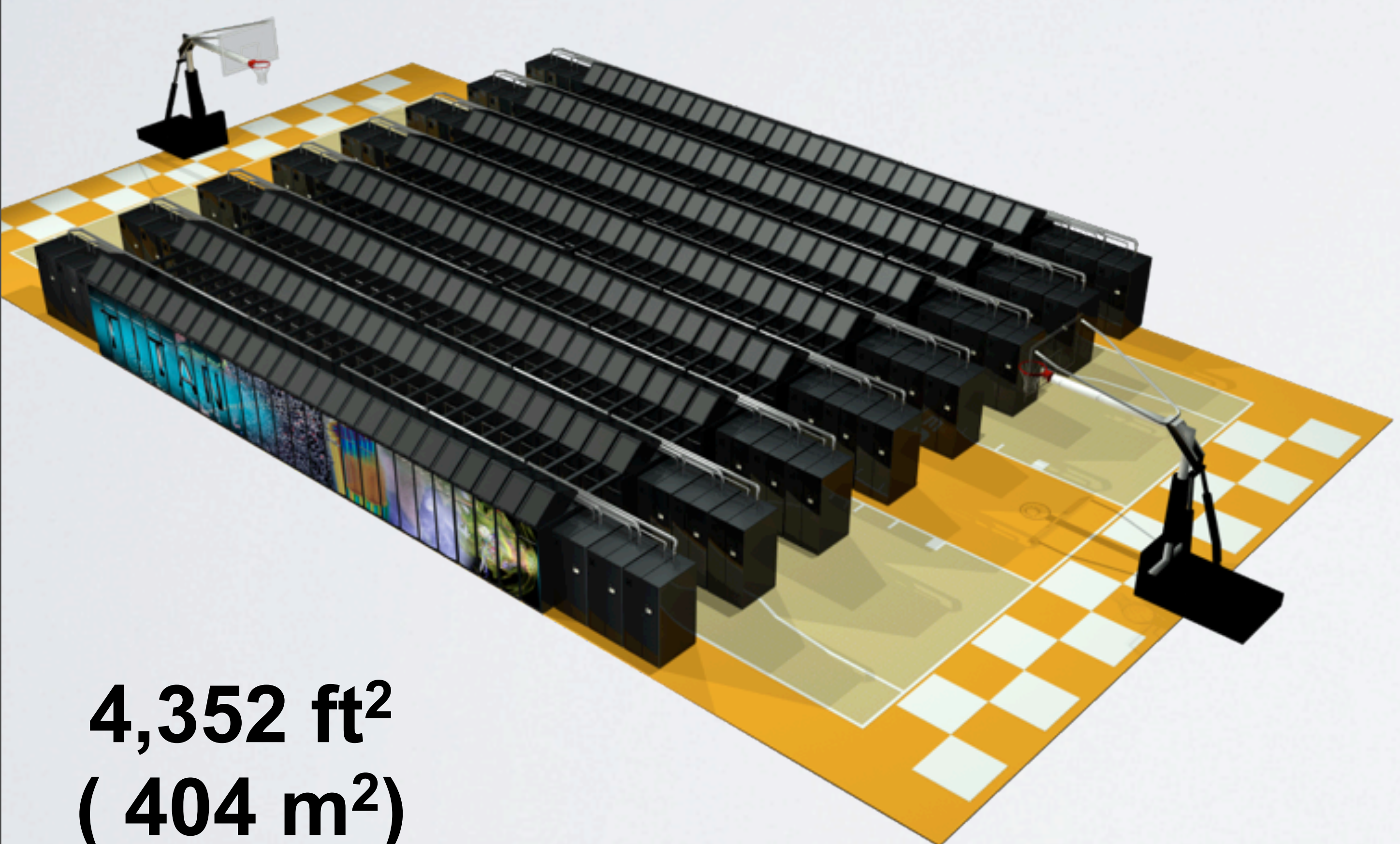
ACCELERATORS





ORNL'S "TITAN" HYBRID SYSTEM



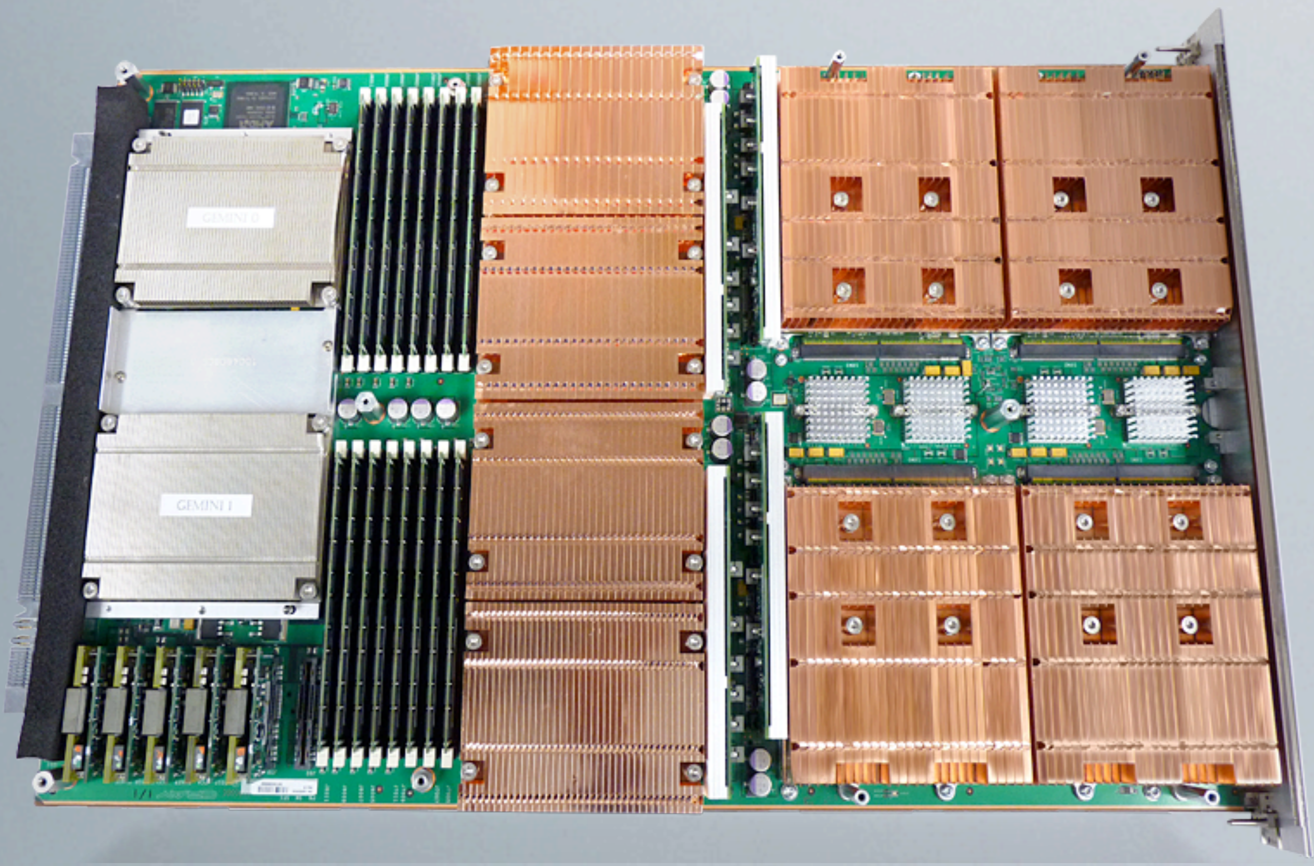


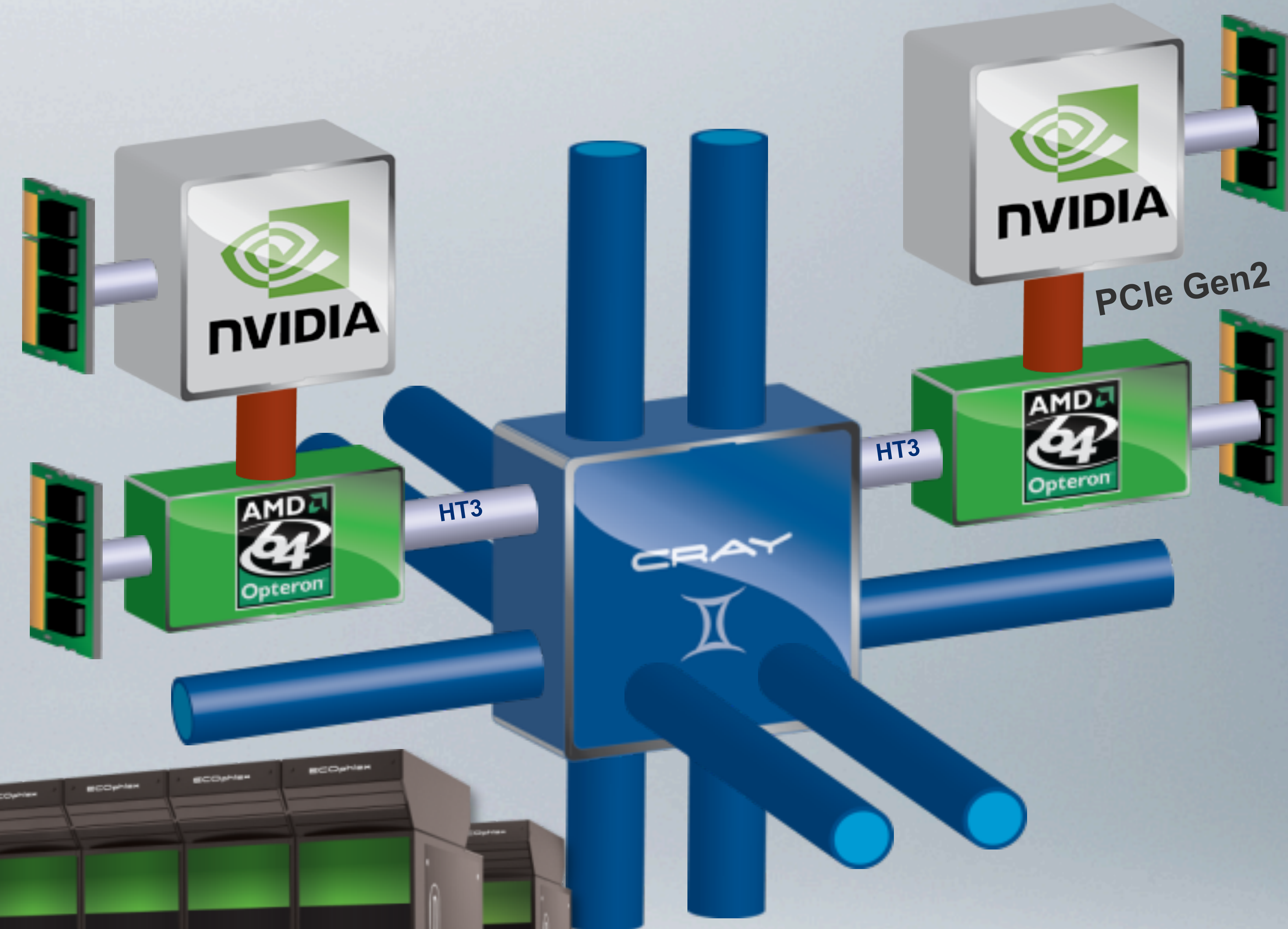
4,352 ft^2
(404 m^2)

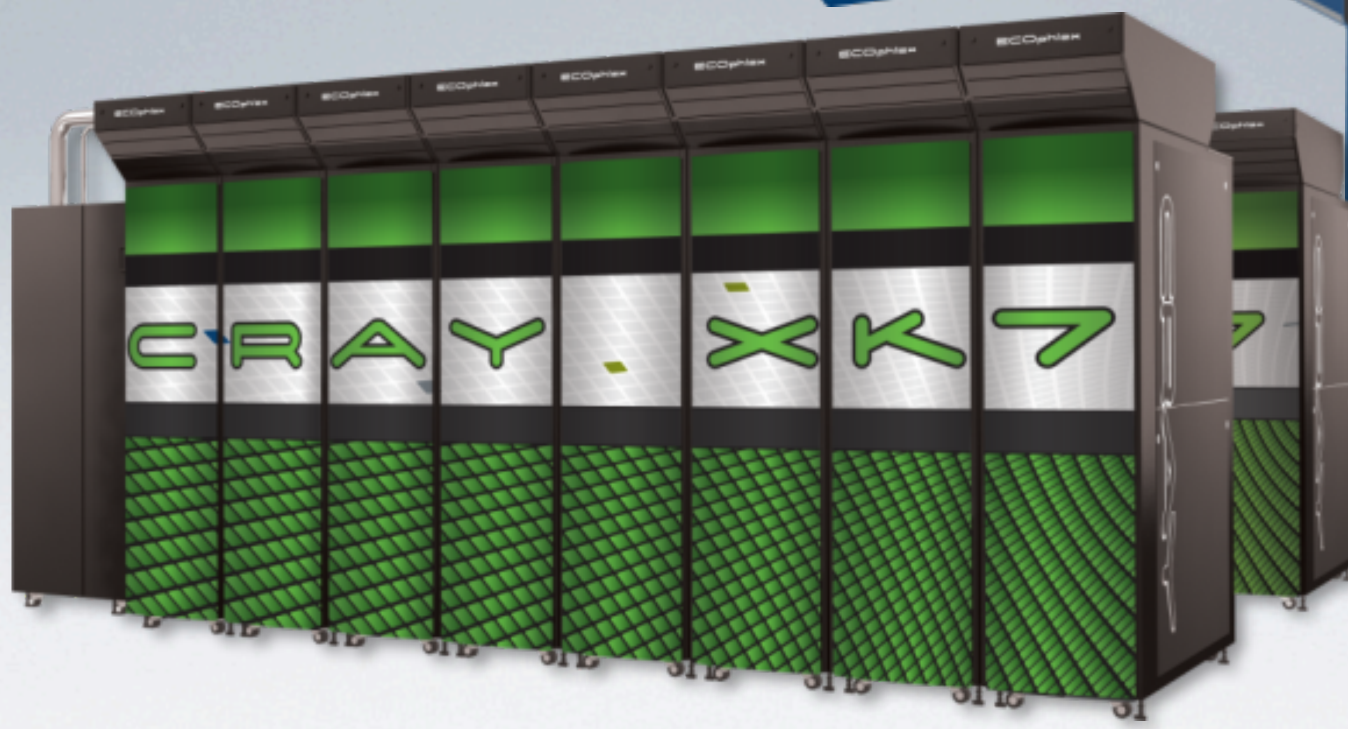
SYSTEM SPECIFICATIONS:

- Peak performance of 27.1 PF
  - 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla "K20x" GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak power

	Titan Nodes		
Node	AMD Opteron 6200 Interlagos (16 cores)	2.2 GHz	32 GB (DDR3)
Accelerator	Tesla K20x (2688 CUDA cores)	732 MHz	6 GB (DDR5)
Network	Gemini High Speed Interconnect	3D Torus	
Storage	Luster Filesystem	5 PB	
Archive	High-Performance Storage System (HPSS)	29 PB	




NVIDIA
NVIDIA
PCIe Gen2
AMD 64 Opteron
AMD 64 Opteron
HT3
HT3
CRAY


CRAY XK7

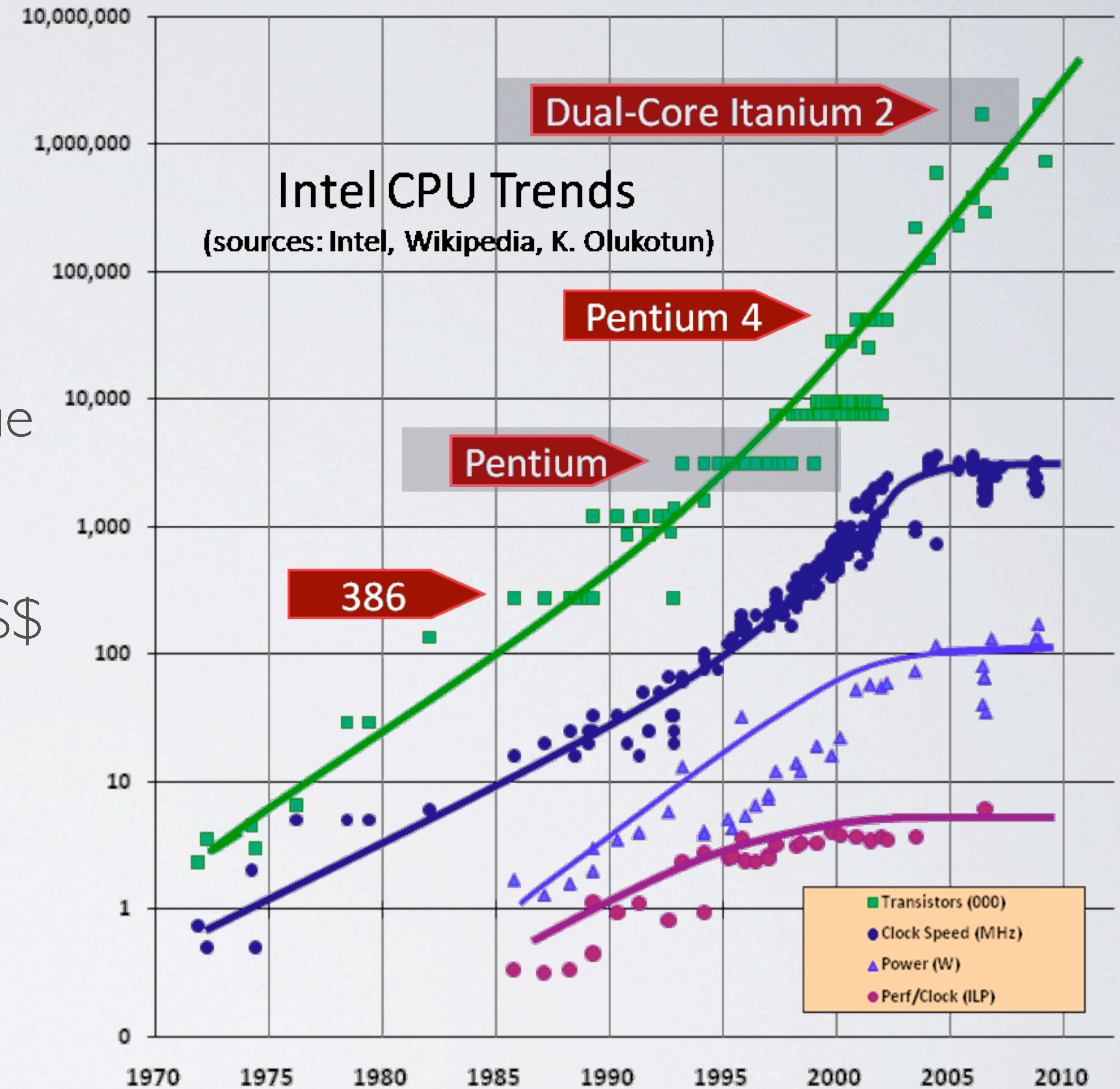
TITAN UPDATE

- Jaguar to Titan upgrade was in place
- Titan is still going through acceptance

Date	Nodes
Feb 2nd	9,716 (CPU Only)
March 11	8,972 (GPUs available)
Early April	0 (Acceptance)
May	18,688 (ALL)

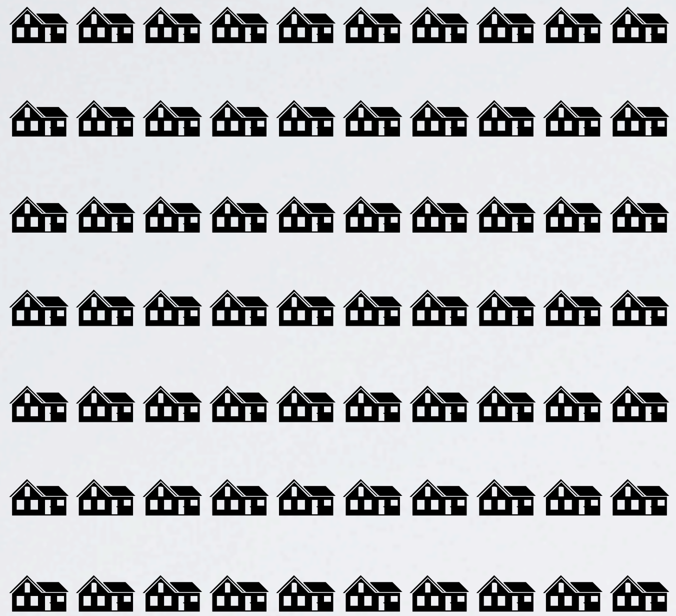
THE POWER WALL

- **Moore's Law** continues, while **CPU clock rates** stopped increasing in 2003 due to **power constraints**.
- **Power** is capped by heat dissipation and $$$
- Performance increases have been coming through increased parallelism



Herb Sutter: Dr. Dobb's Journal: http://www.gotw.ca/publications/concurrency-ddj.htm

POWER IS THE PROBLEM

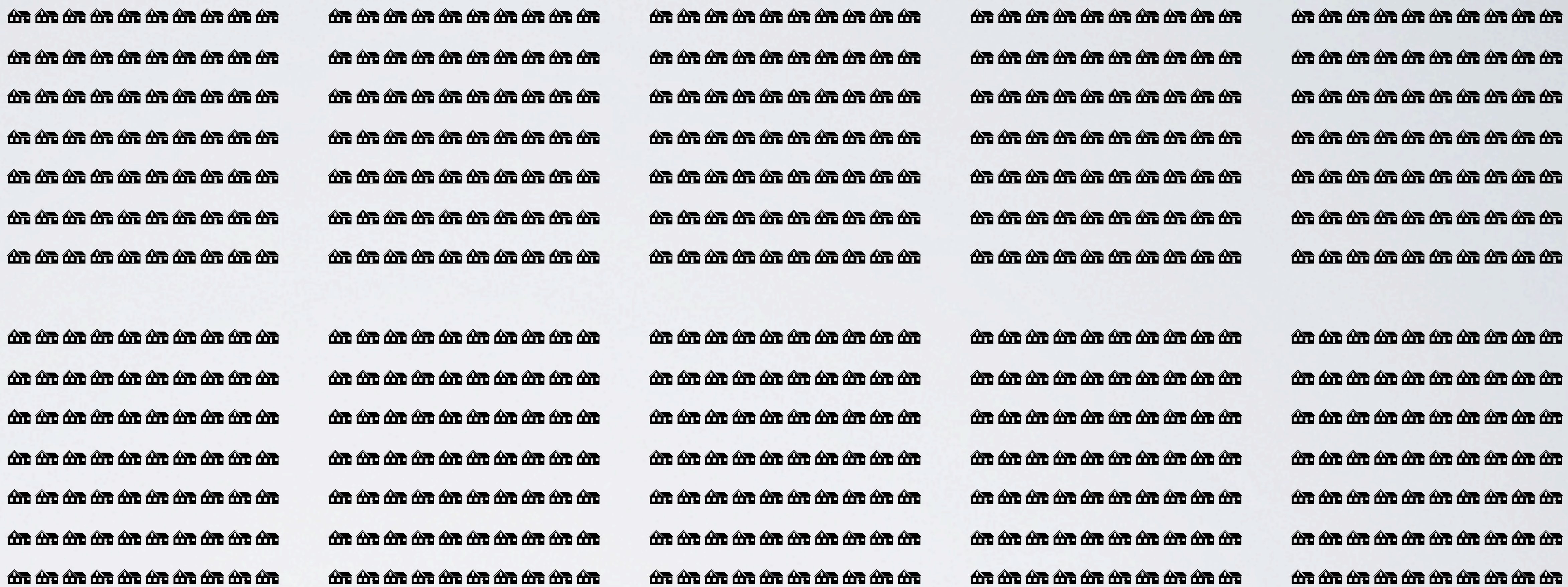


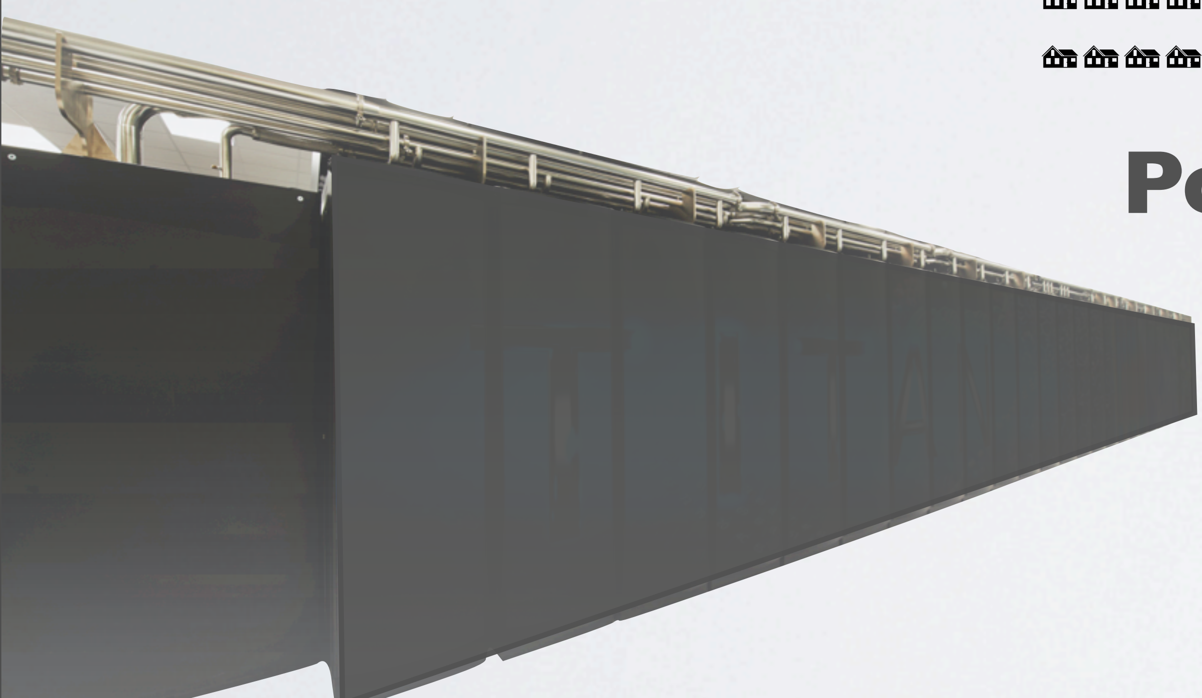
Power consumption of 2.3 PF Jaguar

7 megawatts

equivalent to a small city (~7,000 homes)

POWER IS THE PROBLEM



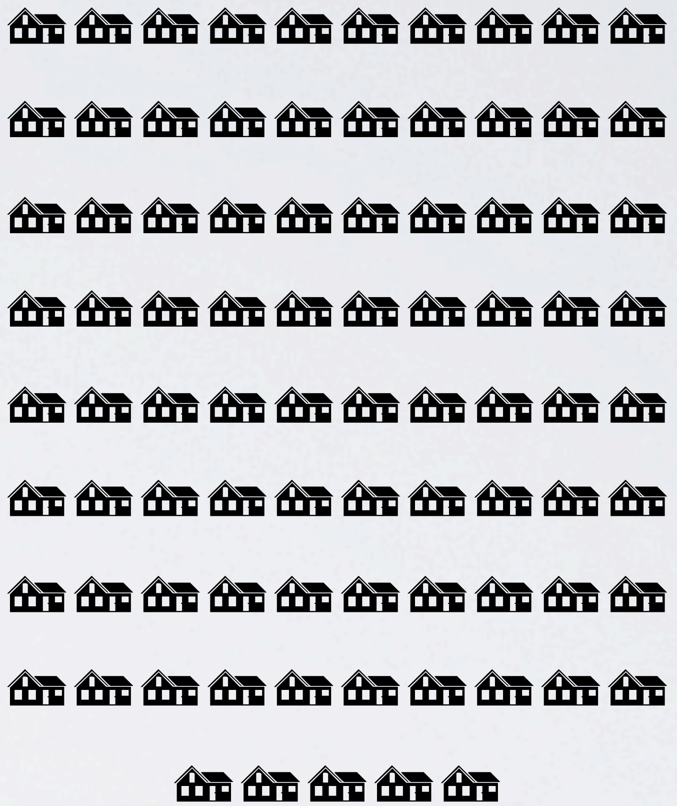


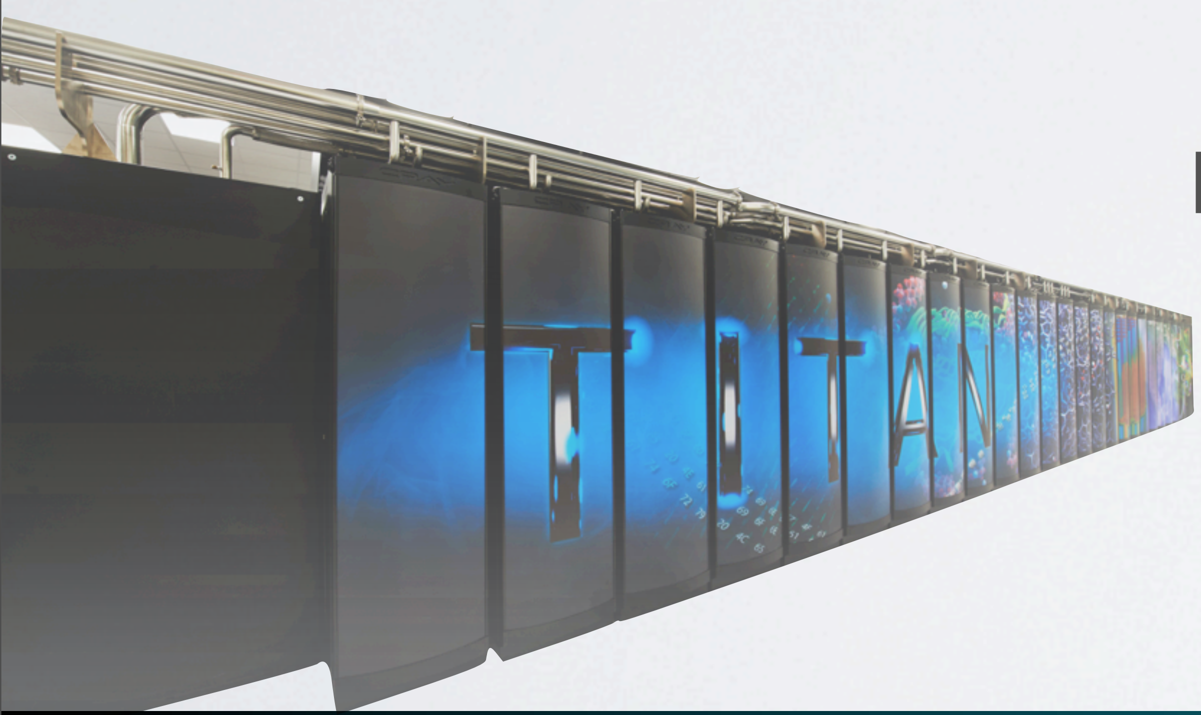
Power consumption of a 27 PF CPU-only system

82 megawatts

equivalent to ~80,000 homes

POWER IS THE PROBLEM





Power consumption of a 27 PF Hybrid system

8.2 megawatts

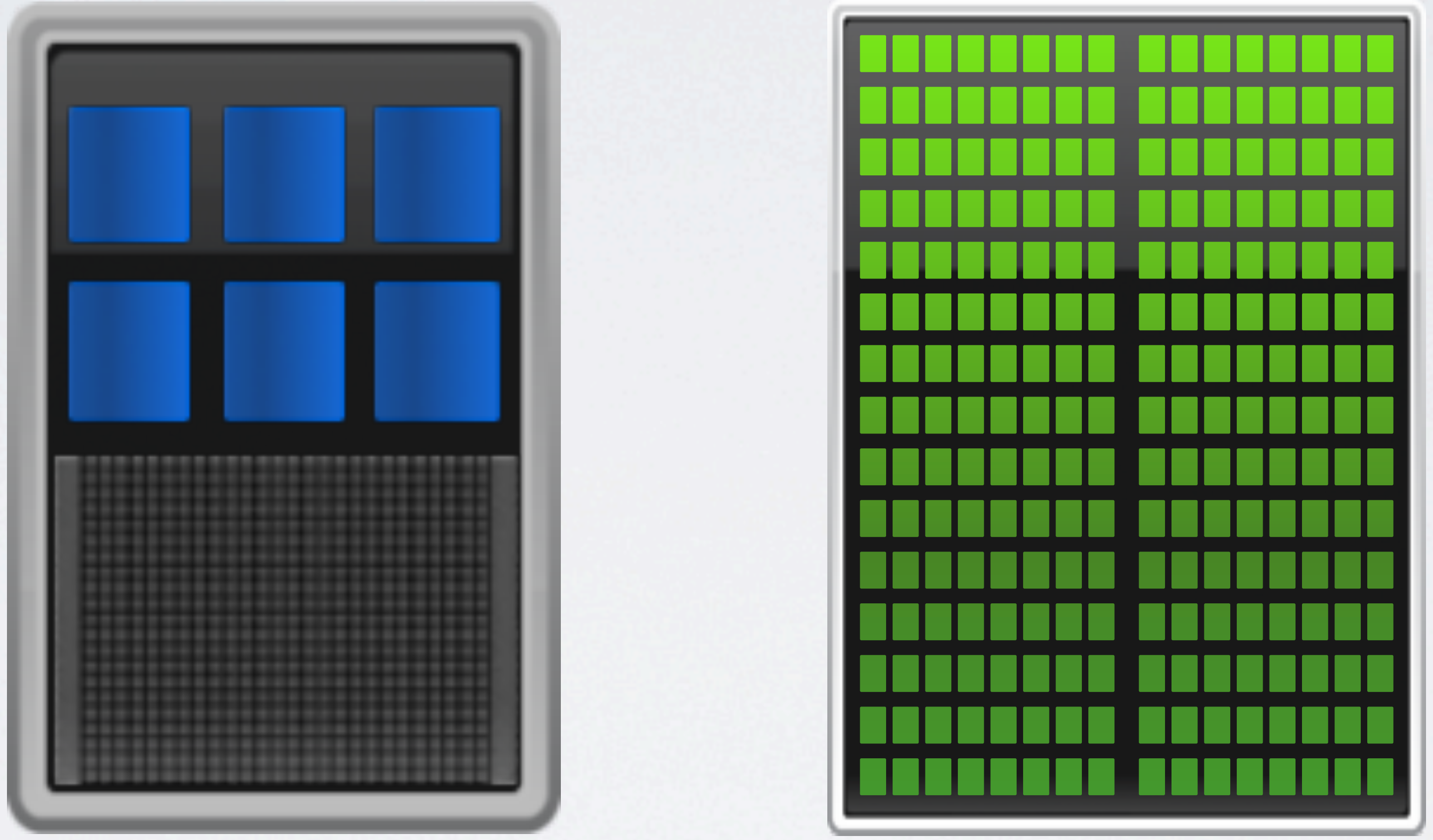
equivalent to ~8,000 homes

WHY GPUs ?

High performance and power efficiency on path to exascale

CPU

GPU



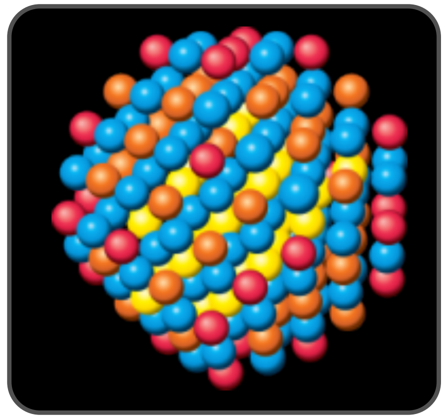
10x performance per socket

10x the energy-efficiency

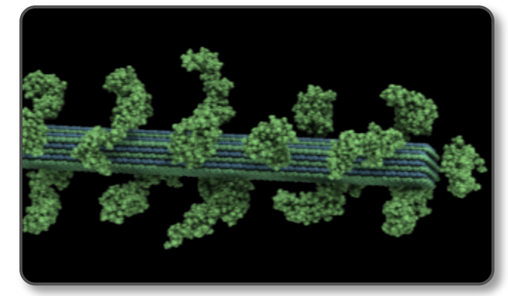
Optimized for multitasking

Optimized for throughput

CENTER FOR ACCELERATED APPLICATION READINESS (CAAR)

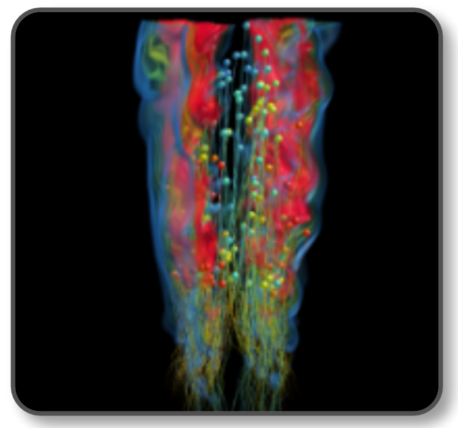


Material Science (WL-LSMS)
Illuminating the role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

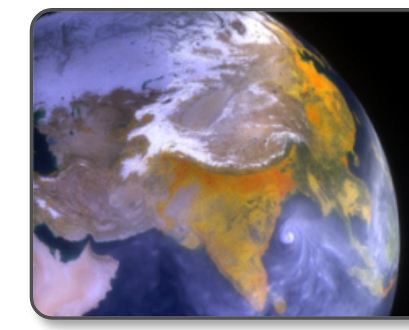


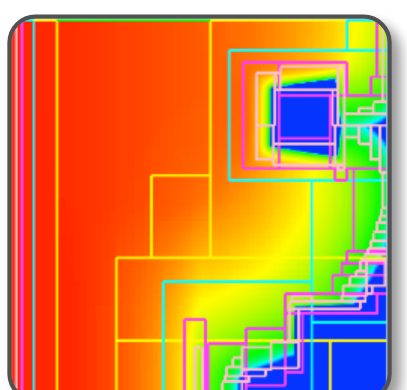
Molecular (LAMMPS)
A molecular description of soft materials, with applications in biotechnology, medicine and energy.

Combustion (S3D)
Understanding turbulent combustion through direct numerical simulation with complex chemistry.

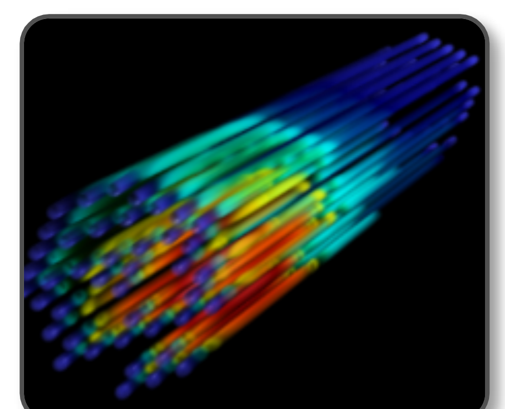


Climate Change (CAM-SE)
Answering questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns / statistics and tropical storms.





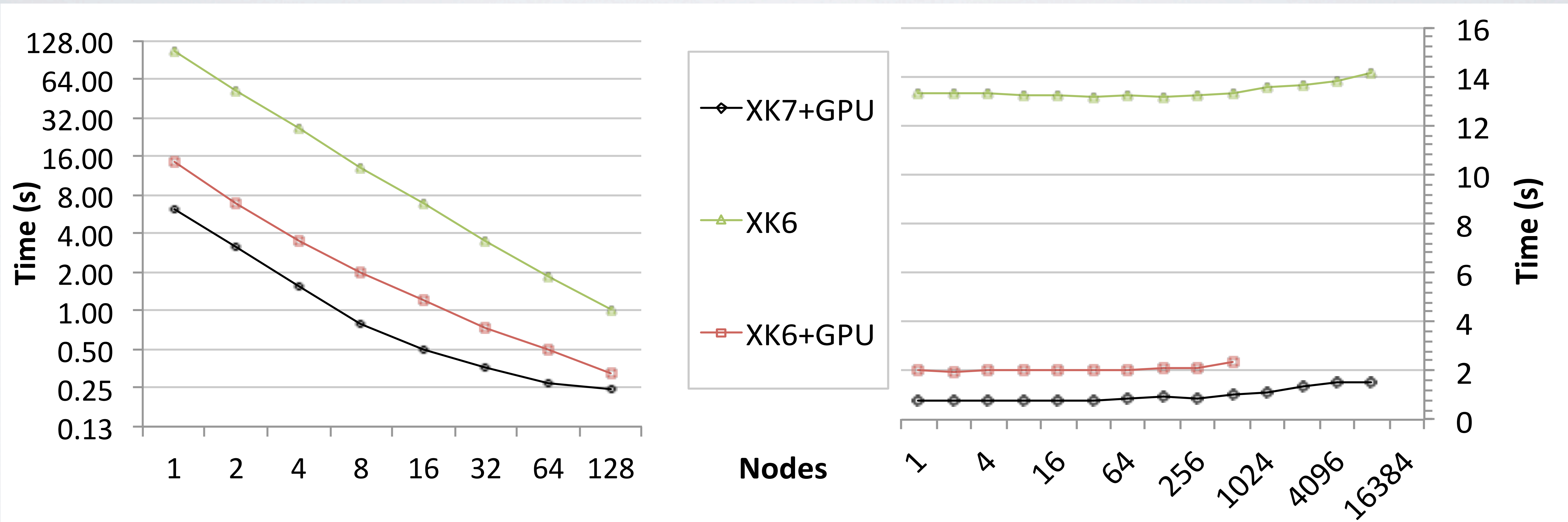
Astrophysics (NRDF)
Radiation transport – important in astrophysics, laser fusion, combustion, atmospheric dynamics, and medical imaging – computed on AMR grids.



Nuclear Energy (Denovo)
Discrete ordinates radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

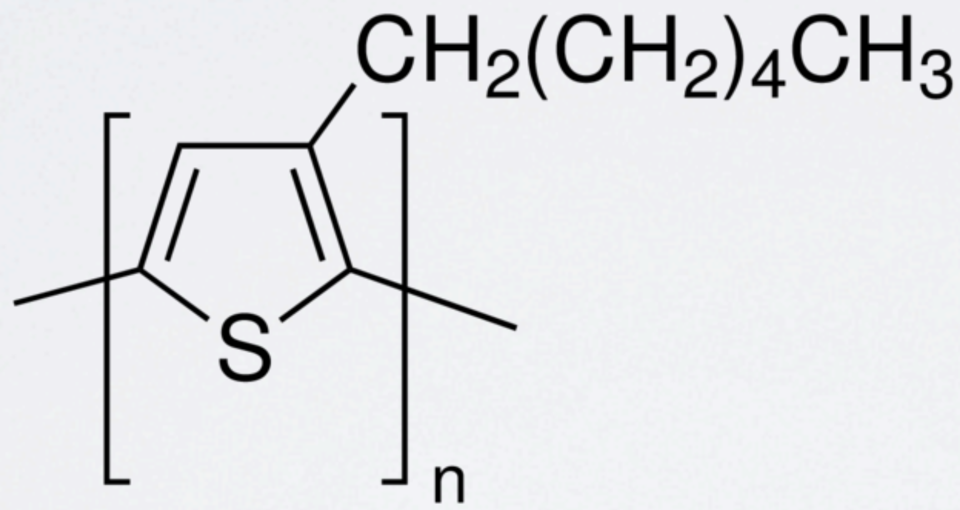
LAMMPS EARLY RESULTS

- **Liquid crystal** mesogens are represented with biaxial ellipsoid particles, Gay-Berne potential, isotropic phase, isothermal-isobaric ensemble, 4σ cutoff with a 0.8σ neighbor skin (High arithmetic intensity)

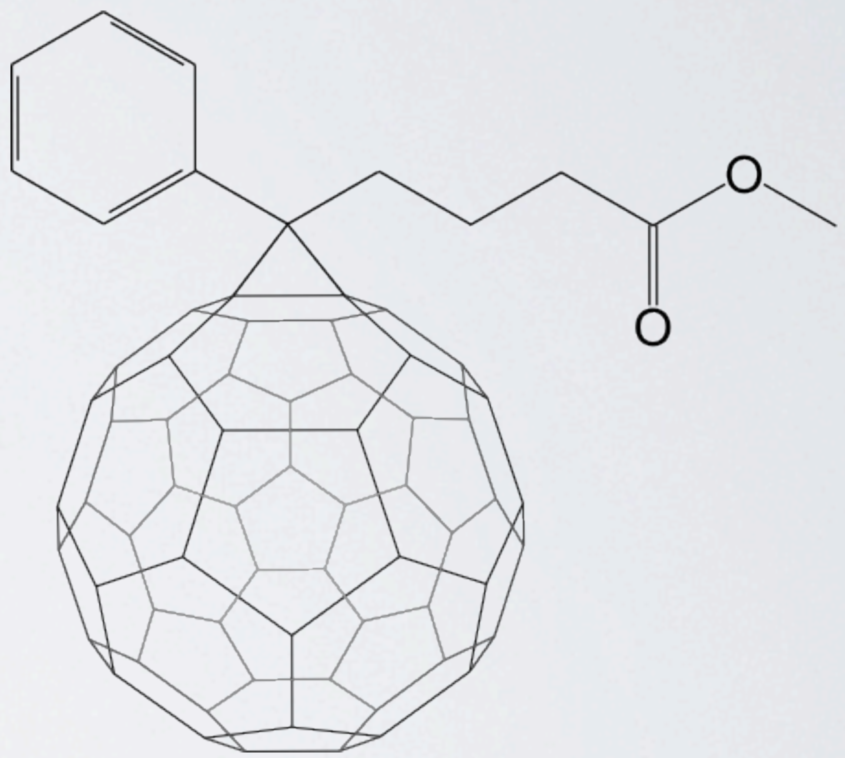


EFFICIENT ORGANIC PHOTOVOLTAIC MATERIALS

- Organic photovoltaic (OPV) solar cells are promising renewable energy sources:
- Low costs, high-flexibility, and light weight
- Bulk-heterojunction (BHJ) active layer is critical for device performance
- High ratios of donor/acceptor interfaces per volume
- Detailed structure of BHJ is unknown
- Use Titan to converge early pioneering MD simulations of BHJ interfaces

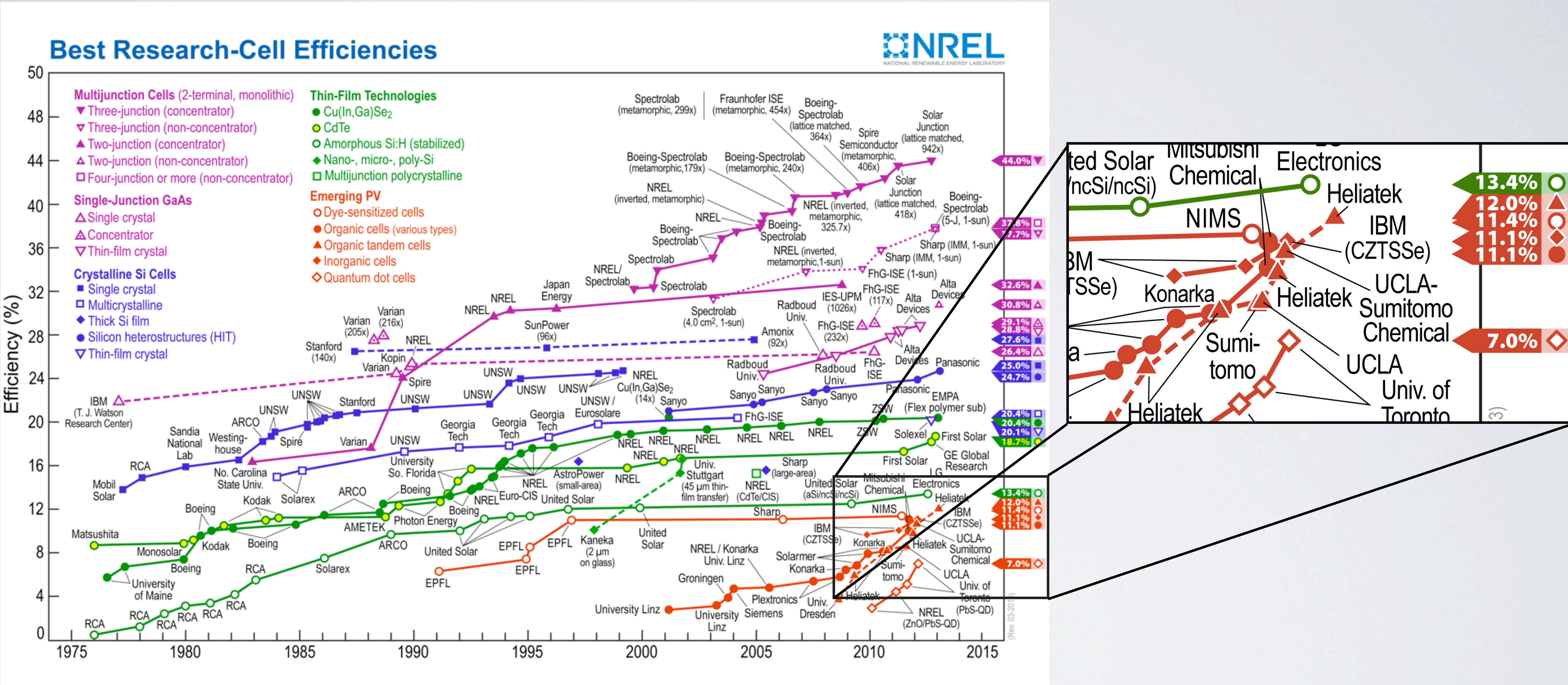


P3HT (electron donor)



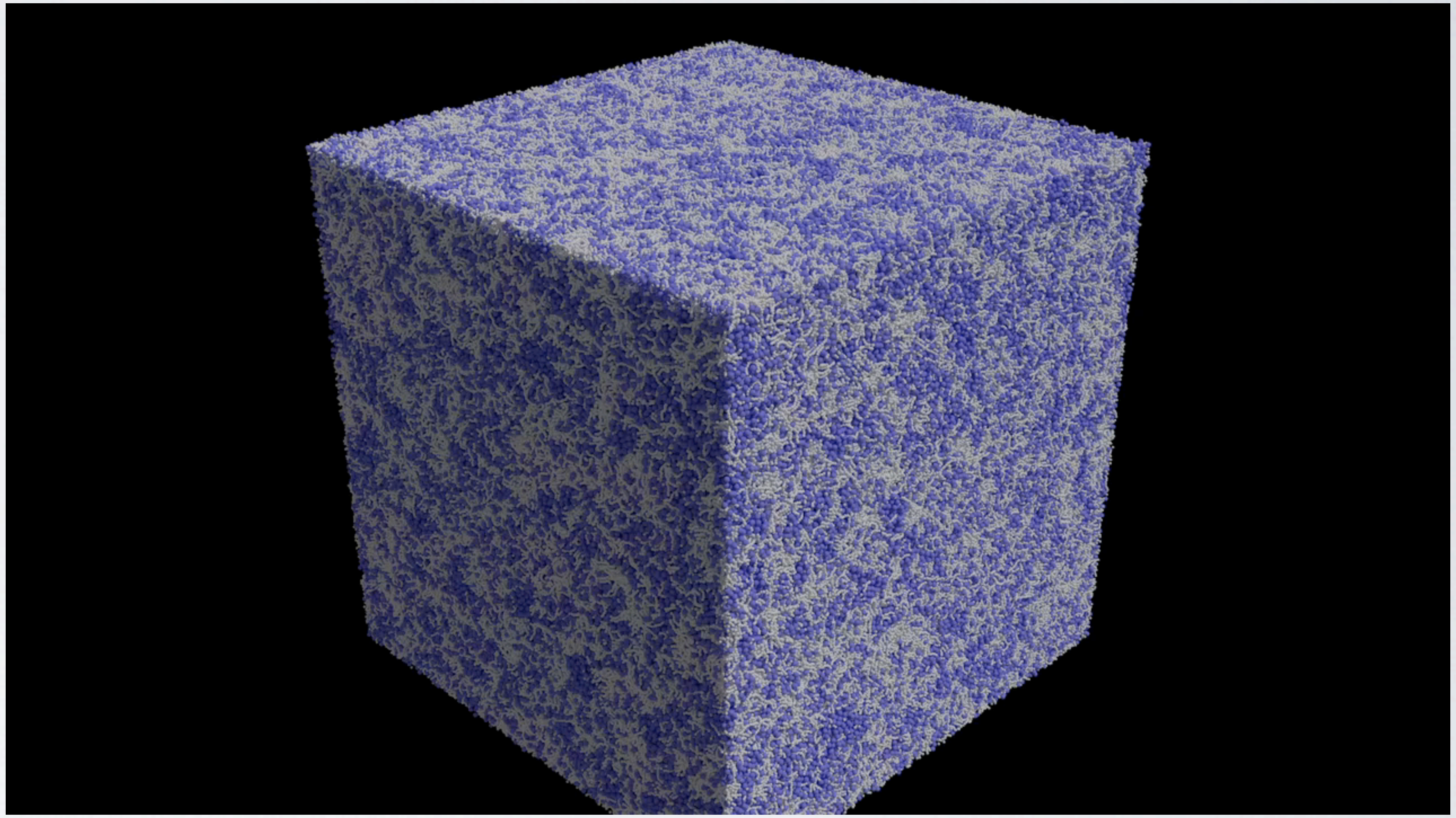
PCBM (electron acceptor)

EFFICIENT ORGANIC PHOTOVOLTAIC MATERIALS



COARSED-GRAIN MD SIMULATION OF P3HT:PCBM HETEROJUNCTION

217 Cray XK7 nodes per simulation during March 2013



Speedup of 2.5-3x for OPV simulation used here

- Acceleration for neighbor-list, short-range forces, and long-range electrostatics
- Portability: Builds with CUDA or OpenCL
- Speedups on Titan (GPU+CPU vs. CPU: 2X to 15x (mixed precision) depending upon model and simulation
- Titan simulations are 27x larger and 10x longer
- Converged P3HT:PCBM separation in 400ns CGMD time
- Increasing polymer chain length will decrease the size of the electron donor domains
- PCBM (fullerene) loading parameter results in an increasing, then decreasing impact on P3HT domain size

SPINODAL DEWETTING ON TITAN

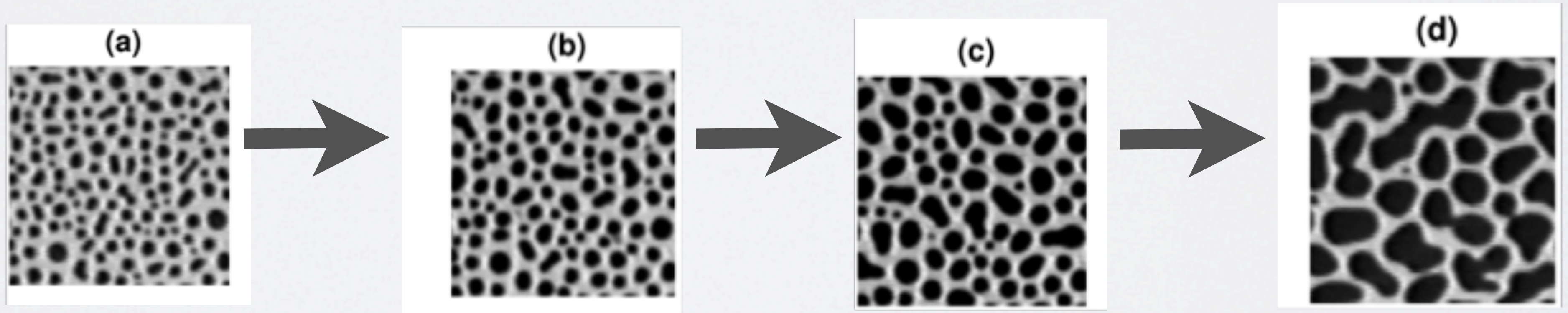
- Model the liquid crystal as a Gay-Berne mesogen (liquid crystal unit) interacting with a Lennard-Jones subtrate
- Allows us to study the mechanism of dewetting at the molecular level at large size scales
- We can study the impact of the size and aspect ratio of the characteristic mesogen on the dewetting process as well as the impact of changes in the relative mesogen interaction strengths along the optical axis
- We can study local phase transitions that occur with dewetting and the formation of complex patterns
- We can study the effect of substrate properties, polymer grafting, non-LC solute, etc. on the dewetting process

SPINODAL DEWETTING ON TITAN

Titan Simulation of LC Dewetting using (3:1) Characteristic Mesogen on 4900 Nodes

Simulation Trajectory (Left) **Simulation Layer Height (Right)**

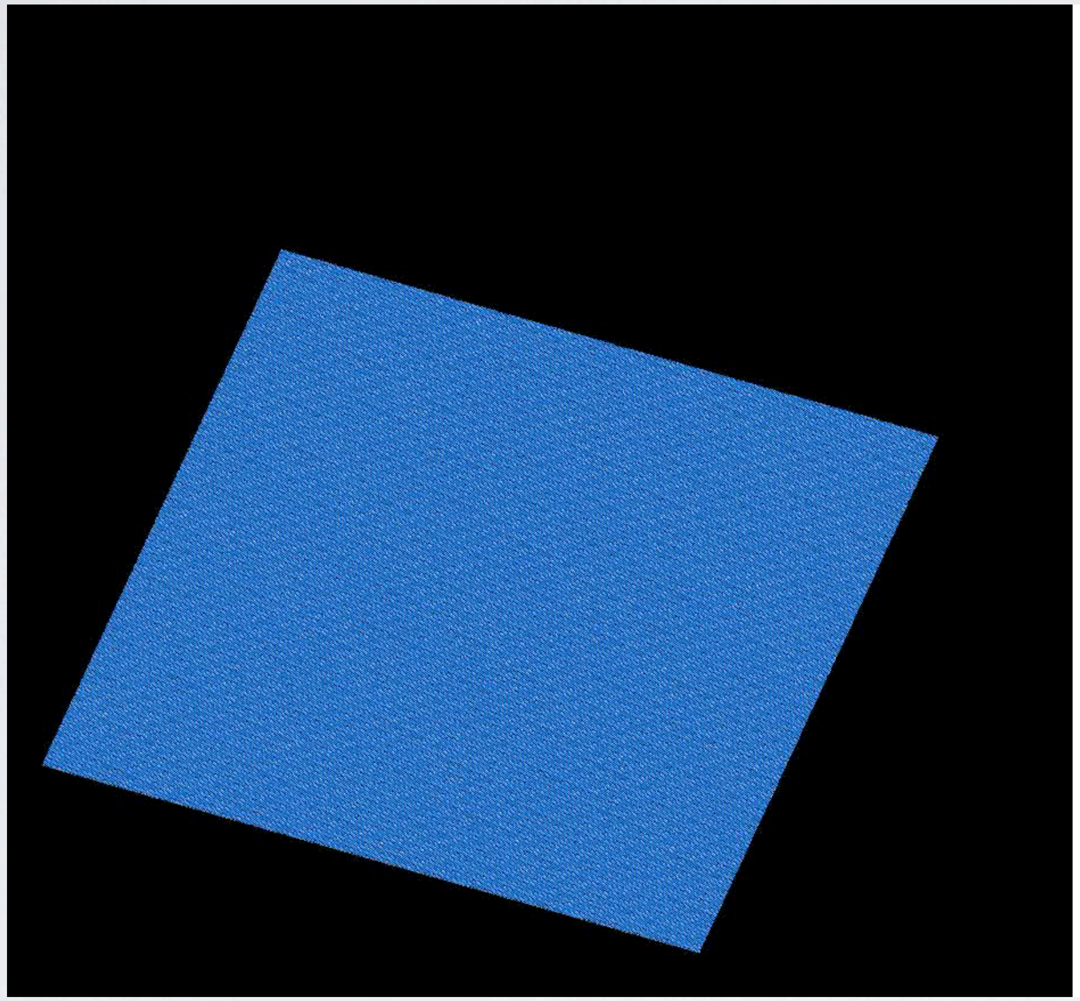
Time Progression of 5CB Dewetting on Silicon Wafer from Experiment

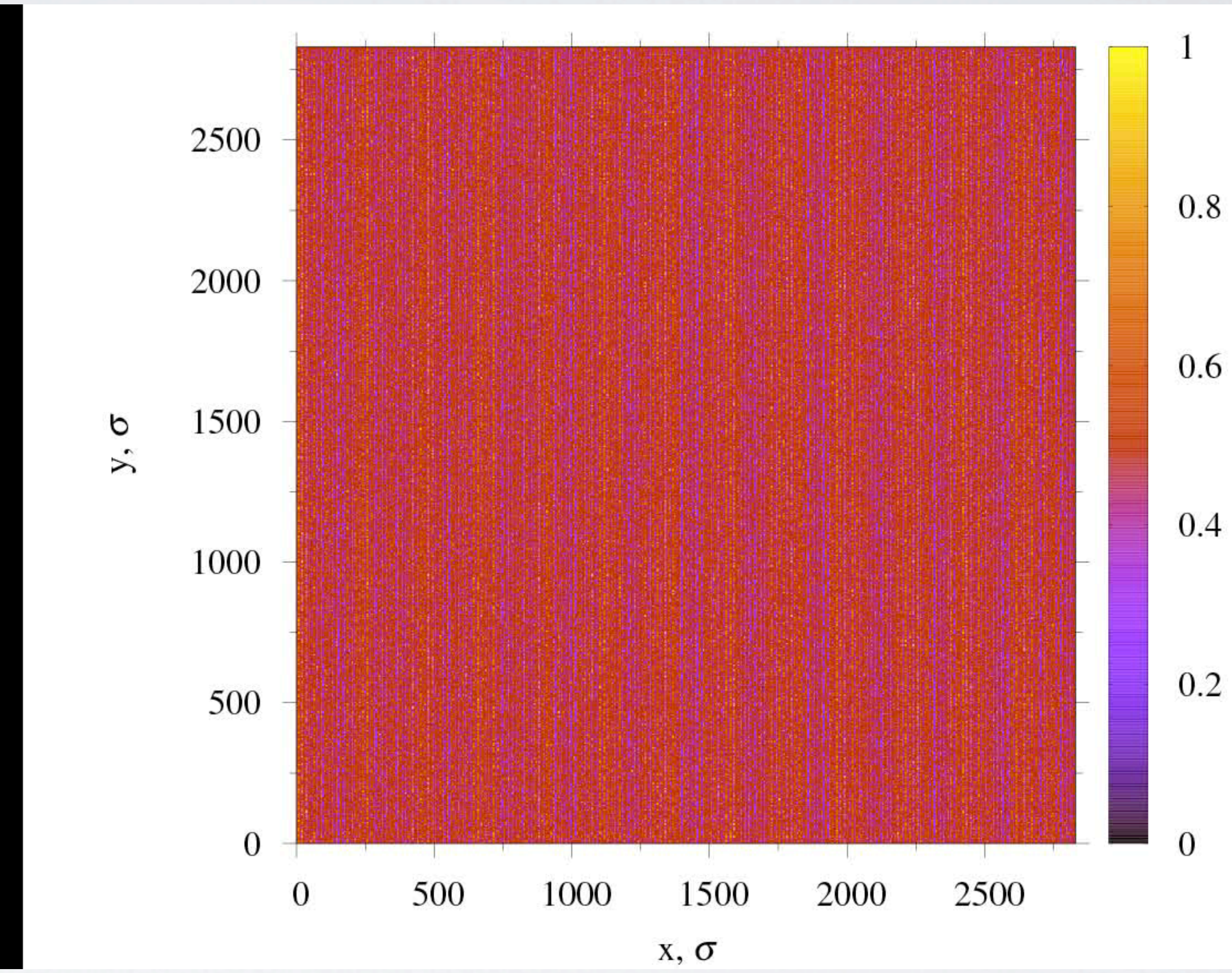


SPINODAL DEWETTING ON TITAN

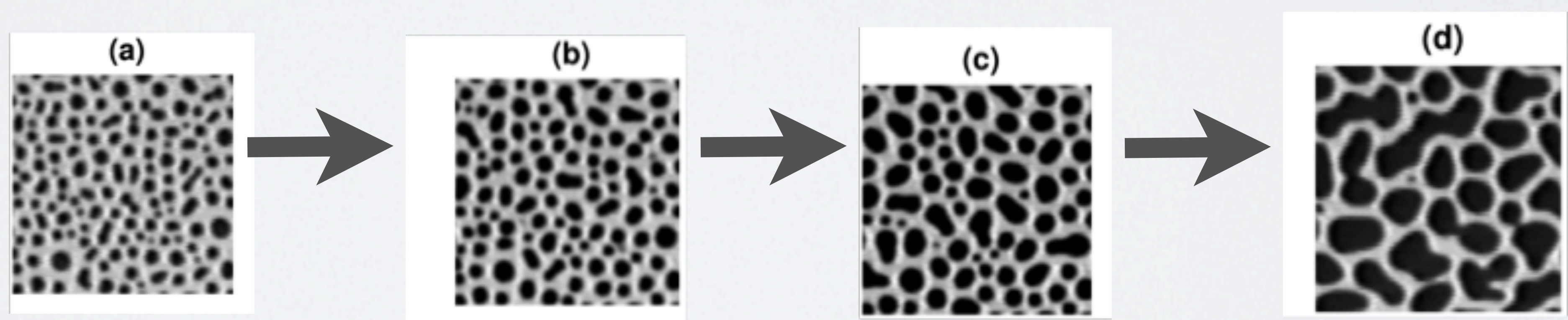
Titan Simulation of LC Dewetting using (3:1) Characteristic Mesogen on 4900 Nodes

Simulation Trajectory (Left) **Simulation Layer Height (Right)**





Time Progression of 5CB Dewetting on Silicon Wafer from Experiment

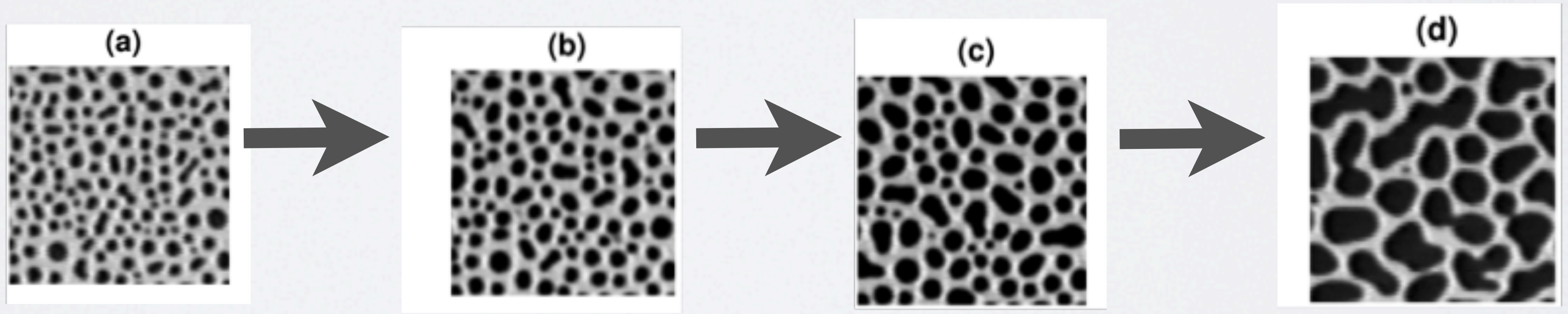


SPINODAL DEWETTING ON TITAN

Titan Simulation of LC Dewetting using (3:1) Characteristic Mesogen on 4900 Nodes

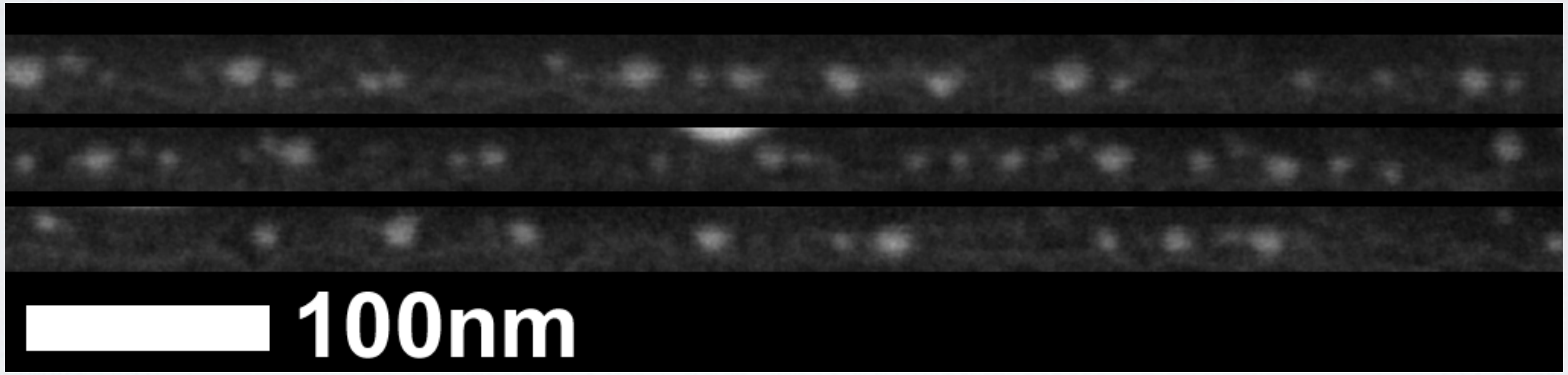
Simulation Trajectory (Left) **Simulation Layer Height (Right)**

Time Progression of 5CB Dewetting on Silicon Wafer from Experiment



RAYLEIGH-PLATEAU LIQUID INSTABILITY FOR COPPER LINES ON GRAPHITE

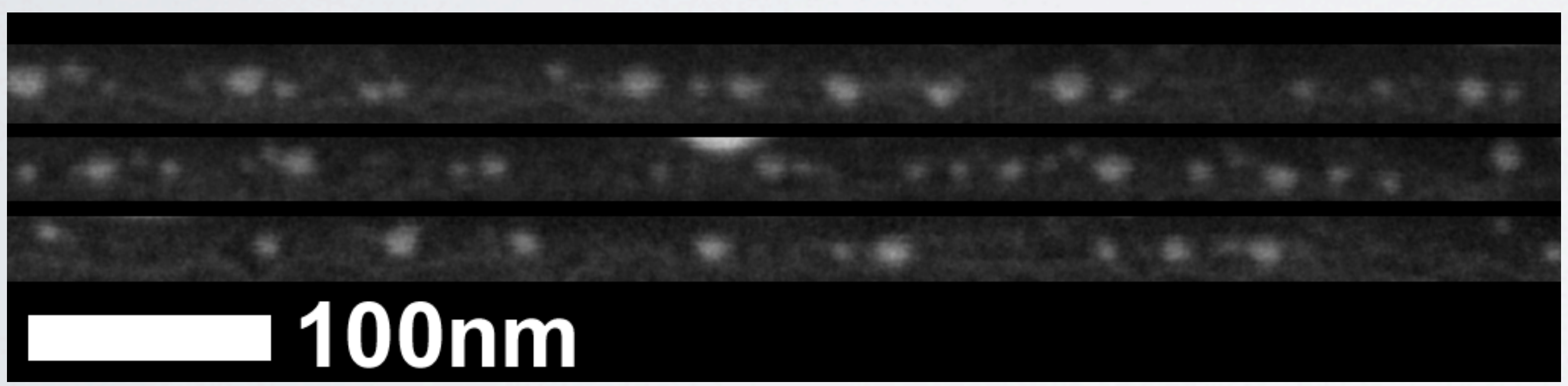
- Pulsed laser melting offers a unique opportunity to dictate materials assembly where rapid heating and cooling rates and ns melt lifetimes are achievable
- Using both experiment and theory we have investigated ways of controlling how the breakage occurs so as to control the assembly of metallic nanoparticles

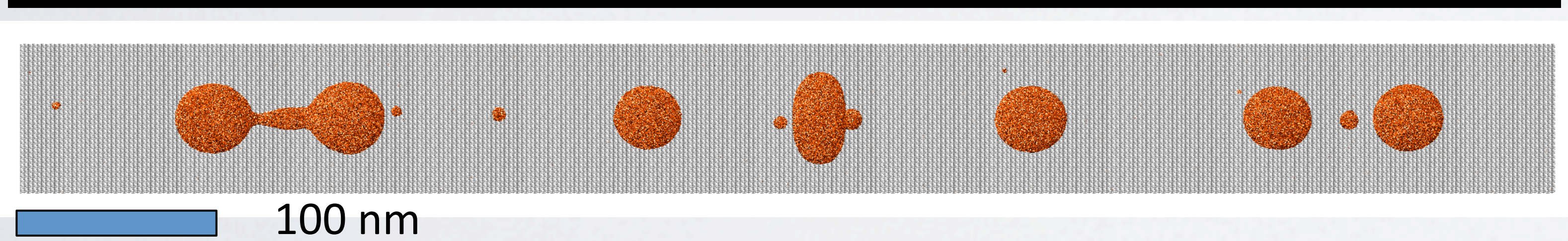


RAYLEIGH-PLATEAU LIQUID INSTABILITY FOR COPPER LINES ON GRAPHITE

- 11.4M Cu Atom Simulations on Graphitic Substrate
- 2.7X Faster than 512 XK6 w/out Accelerators

Simulations were performed with GPU acceleration on Jaguar at the same scales as experiment

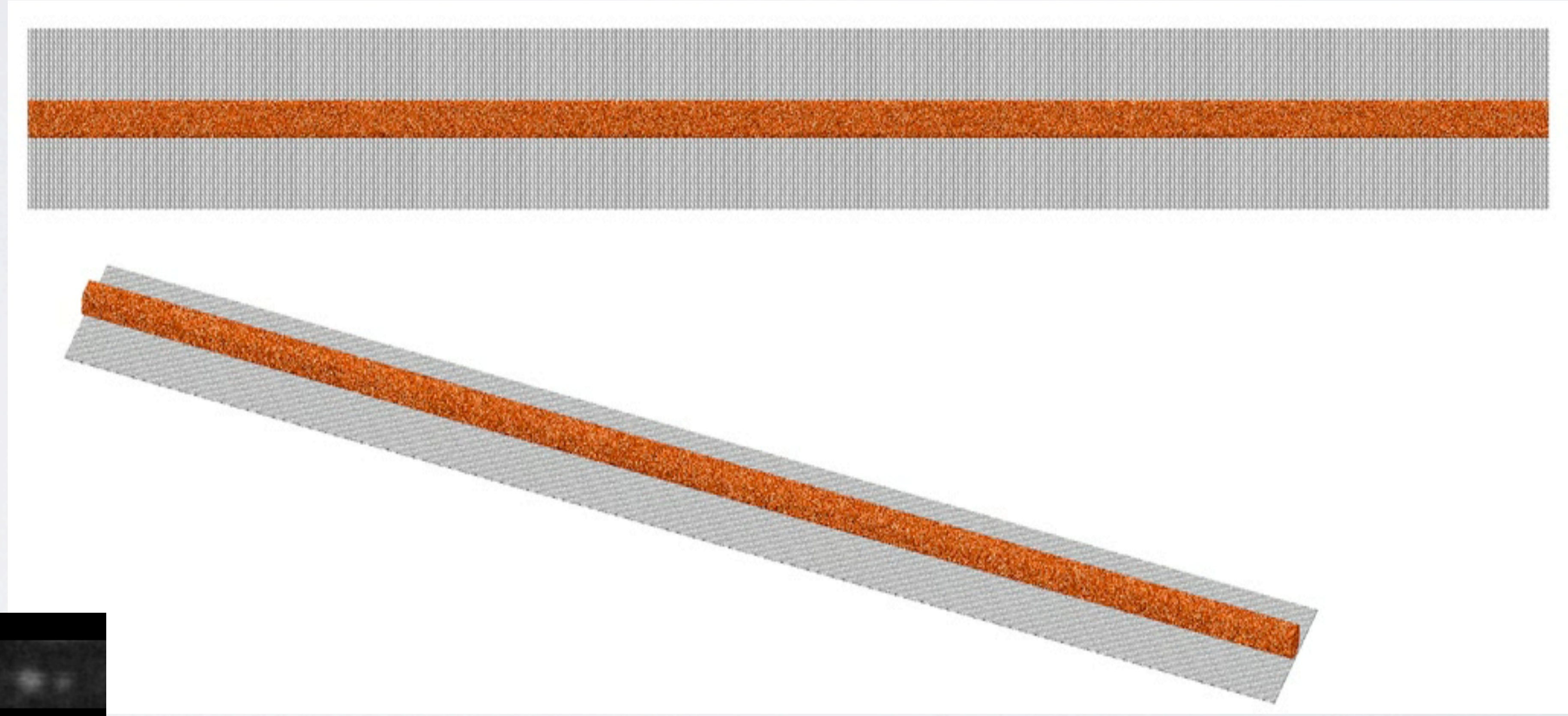


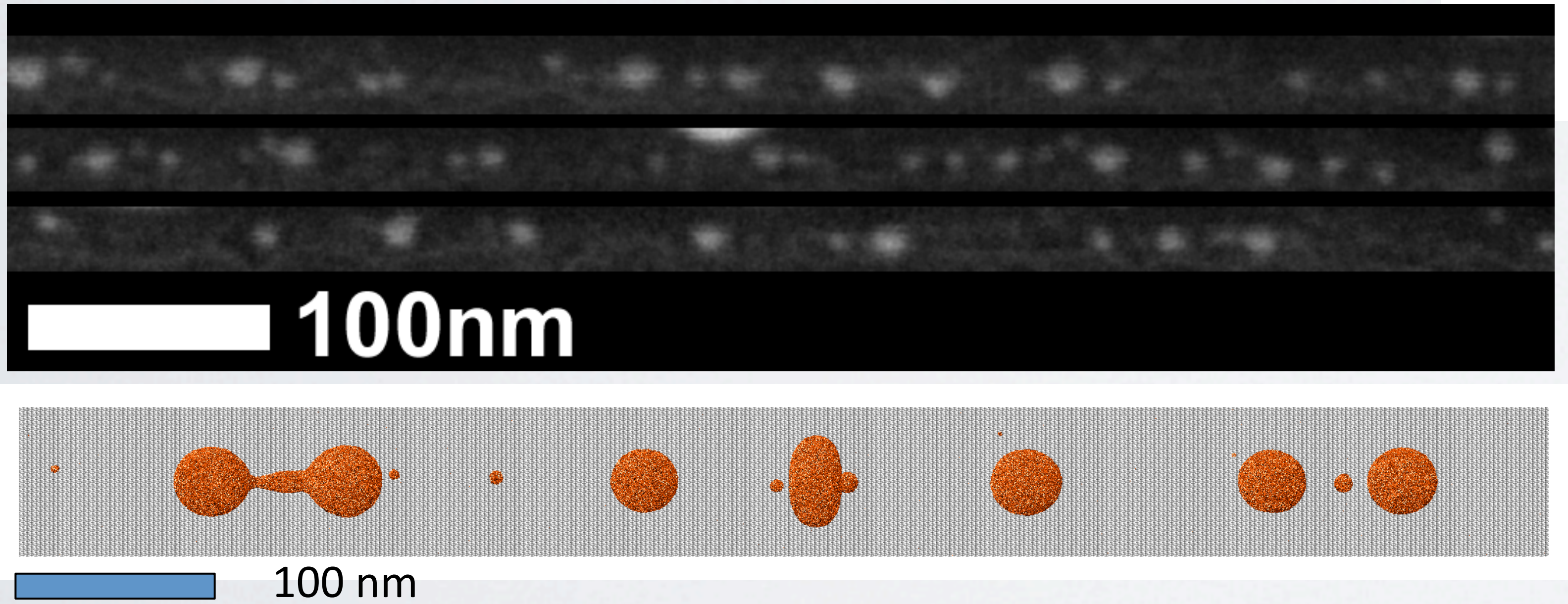


RAYLEIGH-PLATEAU LIQUID INSTABILITY FOR COPPER LINES ON GRAPHITE

- 11.4M Cu Atom Simulations on Graphitic Substrate
- 2.7X Faster than 512 XK6 w/out Accelerators

Simulations were performed with GPU acceleration on Jaguar at the same scales as experiment





MEMBRANE FUSION

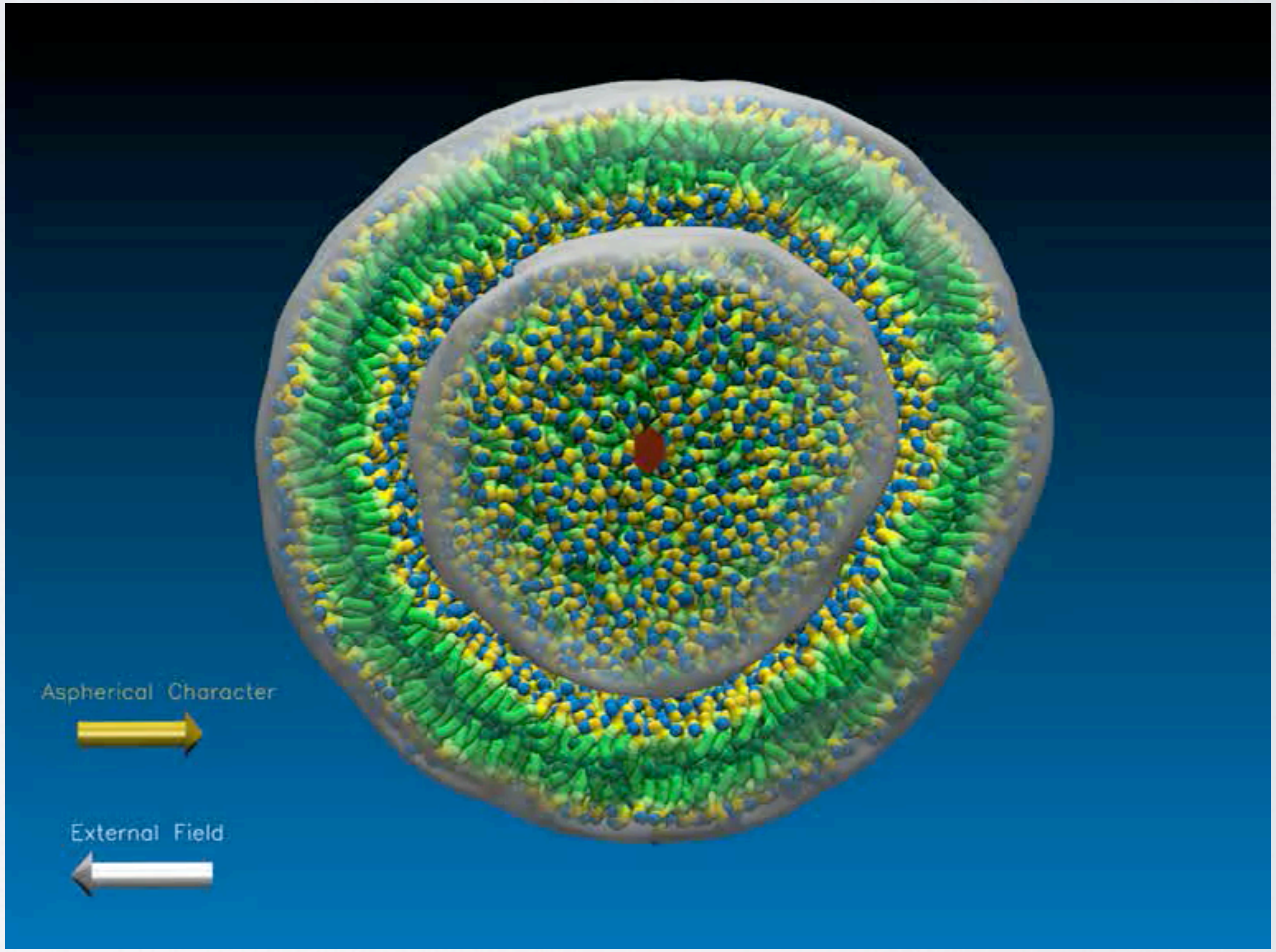
39M Particle Liposome System2.7X Faster than 900 XK6 w/out Accelerators

- Membrane fusion, which involves the merging of two biological membranes in a controlled manner, is an integral part of the normal life cycle of all living organisms.
- Viruses responsible for human disease employ membrane fusion as an essential part of their reproduction cycle.
- Membrane fusion is a critical step in the function of the nervous system
- Correct fusion dynamics requires realistic system sizes

MEMBRANE FUSION

39M Particle Liposome System2.7X Faster than 900 XK6 w/out Accelerators

- Membrane fusion, which involves the merging of two biological membranes in a controlled manner, is an integral part of the normal life cycle of all living organisms.
- Viruses responsible for human disease employ membrane fusion as an essential part of their reproduction cycle.
- Membrane fusion is a critical step in the function of the nervous system
- Correct fusion dynamics requires realistic system sizes



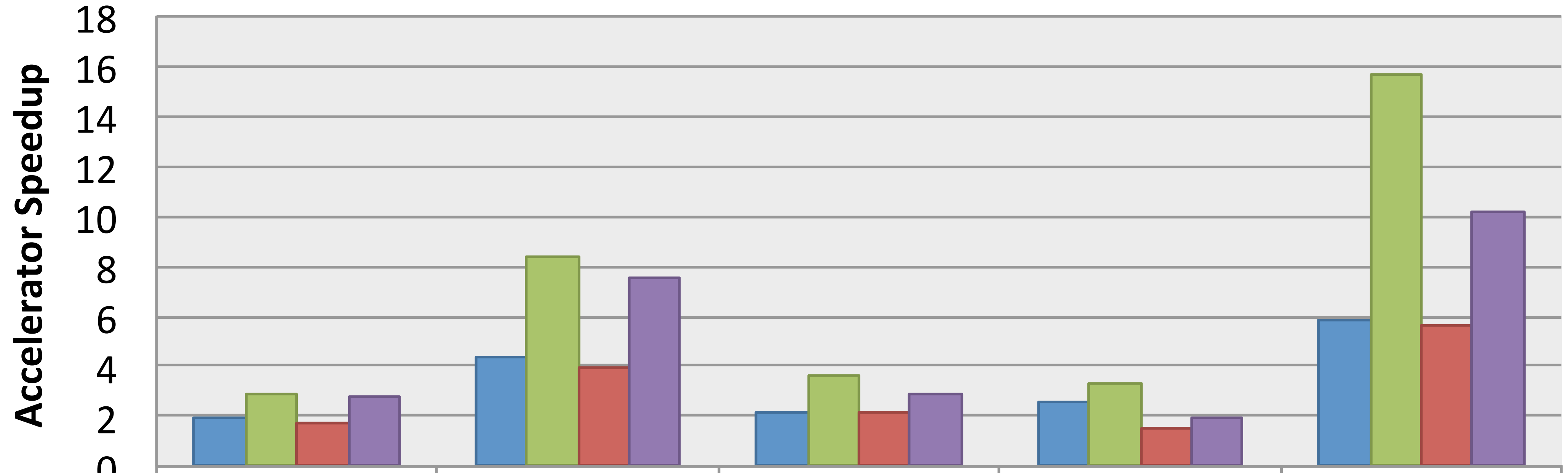
MEMBRANE FUSION

39M Particle Liposome System2.7X Faster than 900 XK6 w/out Accelerators

- Membrane fusion, which involves the merging of two biological membranes in a controlled manner, is an integral part of the normal life cycle of all living organisms.
- Viruses responsible for human disease employ membrane fusion as an essential part of their reproduction cycle.
- Membrane fusion is a critical step in the function of the nervous system
- Correct fusion dynamics requires realistic system sizes

LAMMPS ACCELERATOR SPEEDUP

Speedup with Acceleration on XK6/XK7 Nodes
1 Node = 32K Particles
900 Nodes = 29M Particles



	Atomic Fluid (cutoff = 2.5σ)	Atomic Fluid (cutoff = 5.0σ)	Bulk Copper	Protein	Liquid Crystal
XK6 (1 Node)	1.92	4.33	2.12	2.6	5.82
XK7 (1 Node)	2.90	8.38	3.66	3.36	15.70
XK6 (900 Nodes)	1.68	3.96	2.15	1.56	5.60
XK7 (900 Nodes)	2.75	7.48	2.86	1.95	10.14

HOW EFFECTIVE ARE GPUS ON SCALABLE APPLICATIONS?

Very early performance measurements

OLCF-3 early science codes compared to performance on Jaguar

Application	Description	Jaguar workload	Speedup
S3D	Turbulent combustion	6%	1.8
Denovo sweep	Sweep kernel of 3D neutron transport for nuclear reactors	2%	3.8
LAMMPS	High-performance molecular dynamics	1%	7.4*
WL-LSMS	Statistical mechanics of magnetic materials	2%	3.8**
CAM-SE	Community atmosphere model	1%	~1.8

*mixed precision **gordon bell winner

ACTION PLAN FOR CODE PORTING

We developed a plan for porting these applications, which involved the following steps:

1. Multidisciplinary code team for each code – OLCF application lead, Cray engineer, NVIDIA developer, also cross-cutting support from tool and library developers
2. Early testbed hardware –white box GPU cluster "yona" for code development
3. Code inventory for each code to understand characteristics – application code structure, code suitability for GPU port, algorithm structure, data structures and data movement patterns. Also code execution profile – are there performance "hot spots" or is the profile "flat"
4. Develop parallelization approach for each application – ascertain which algorithm and code components to port to GPU, how to map work to GPU threads, how to manage data motion CPU-GPU and between GPU main memory and GPU caches/shared memory
5. Decide GPU programming model for port to GPU, e.g., CUDA for more close-to-the-metal programming, OpenACC for a higher abstraction level and a more incremental porting approach, OpenCL for portability advantages, or libraries when appropriate
6. Address code development issues – rewrite vs. refactor, managing portability to other platforms, incorporating GPU code into build system, relationship to the code repository main trunk
7. Representative test cases, e.g., early science problems, formulated as basis for evaluating code performance and setting priorities for code optimization. Also formulate comparison metric to measure success, e.g., time to solution on dual Interlagos Cray XE6 vs. Titan Cray XK7 Interlagos+Kepler

APPLICATION CHARACTERISTICS INVENTORY

App	Science Area	Algorithm(s)	Grid type	Programming Language(s)	Compiler(s) supported	LOC	Comm Libraries	Math Libraries
CAM-SE	climate	spectral finite elements, dense & sparse linear algebra, particles	structured	F90	PGI, Lahey, IBM	500K	MPI	Trilinos
LAMMPS	Biology, materials	molecular dynamics, FFT, particles	N/A	C++	GNU, PGI, IBM, Intel	140K	MPI	FFTW
S3D	combustion	Navier-Stokes, finite diff, dense & sparse linear algebra, particles	structured	F77, F90	PGI	10K	MPI	None
Denovo	nuclear energy	wavefront sweep, GMRES	structured	C++, Fortran, Python	GNU, PGI, Cray, Intel	46K	MPI	Trilinos, LAPACK, SuperLU, Metis
WL-LSMS	nanoscience	density functional theory, Monte Carlo	N/A	F77, F90, C, C++	PGI, GNU	70K	MPI	LAPACK (ZGEMM, ZGTRF, ZGTRS)
NRDF	radiation transport	Non-equilibrium radiation diffusion equation	structured AMR	C++, C, F77	PGI, GNU, Intel	500K	MPI, SAMRAI	BLAS, PETSc, Hypre, SAMRSolvers

CAAR: LESSONS LEARNED

- Repeated themes in the code porting work:
  - finding more threadable work for the GPU
  - Improving memory access patterns
  - making GPU work (kernel calls) more coarse-grained if possible
  - making data on the GPU more persistent
  - overlapping data transfers with other work
- Helpful to use as much asynchronicity as possible, to extract performance (CPU, GPU, MPI, PCIe-2)
- Codes with unoptimized MPI communications may need prior work in order to improve performance before GPU speed improvements can be realized
- Some codes need to use multiple MPI tasks per node to access the GPU (e.g., via proxy)—others use 1 MPI task with OpenMP threads on the node
- Code changes that have global impact on the code are difficult to manage, e.g., data structure changes. An abstraction layer may help, e.g., C++ objects/ templates
- Two common code modifications are:
  - Permuting loops to improve locality of memory reference
  - Fusing loops for coarser granularity of GPU kernel calls
- Tools (compilers, debuggers, profilers) were lacking early on in the project but are becoming more available and are improving in quality
- Debugging and profiling tools were useful in some cases (Allinea DT, CrayPat, Vampir, CUDA profiler)

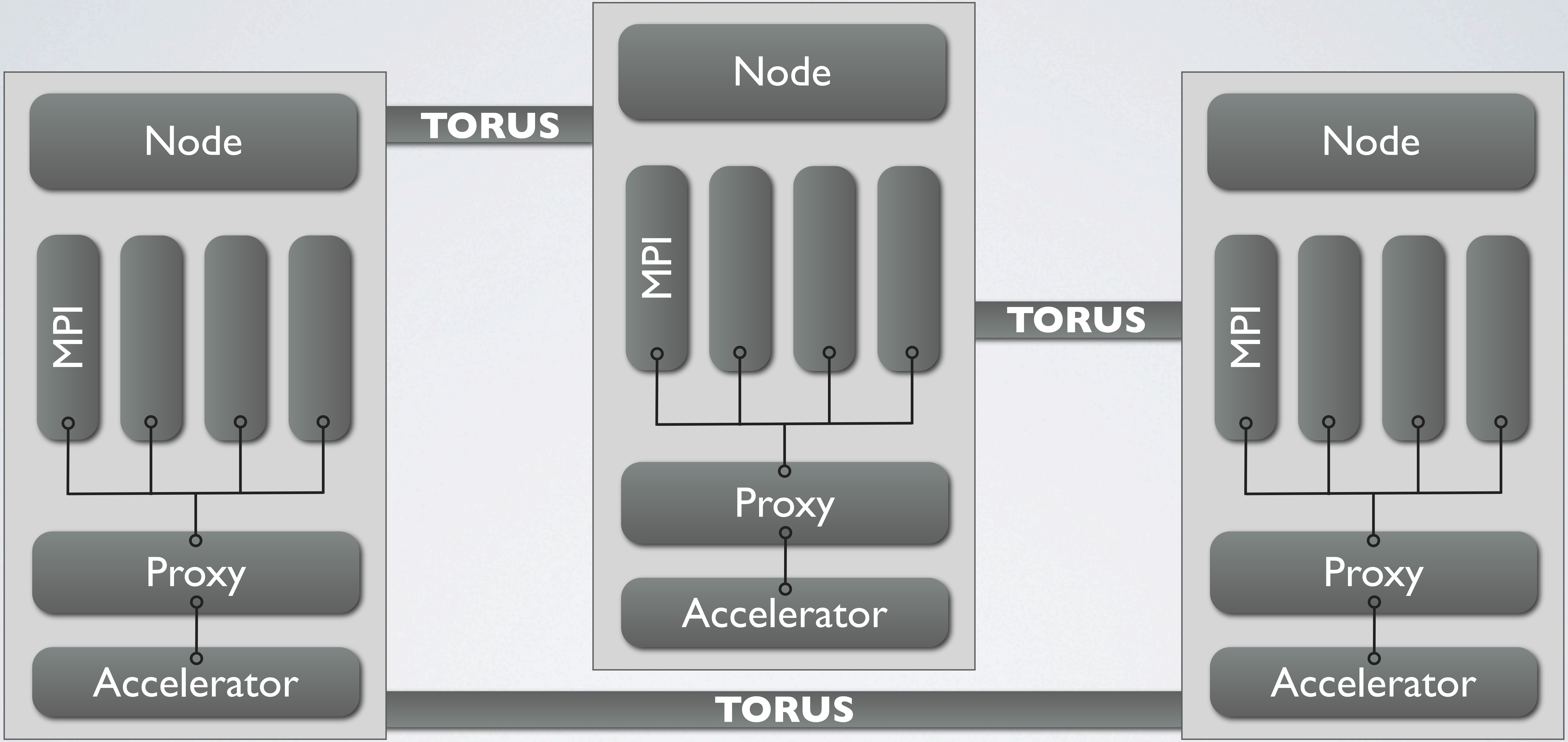
CAAR: SELECTED LESSONS LEARNED

- The difficulty level of the GPU port was in part determined by:
  - Structure of the algorithms—e.g., available parallelism, high computational intensity
  - Code execution profile—flat or hot spots
  - The code size (LOC)
- Since not all future code changes can be anticipated, it is difficult to avoid significant code revision for such an effort
- Up to 1-3 person-years required to port each code
  - Takes work, but an unavoidable step required for exascale
  - Also pays off for other systems—the ported codes often run significantly faster CPU-only (Denovo 2X, CAM-SE >1.7X)
- We estimate possibly 70-80% of developer time is spent in code restructuring, regardless of whether using CUDA / OpenCL / OpenACC / …
- Each code team must make its own choice of using CUDA vs. OpenCL vs. OpenACC, based on the specific case—may be different conclusion for each code
- Science codes are under active development—porting to GPU can be pursuing a "moving target," challenging to manage
- More available flops on the node should lead us to think of new science opportunities enabled—e.g., more DOF per grid cell
- We may need to look in unconventional places to get another ~30X thread parallelism that may be needed for exascale—e.g., parallelism in time

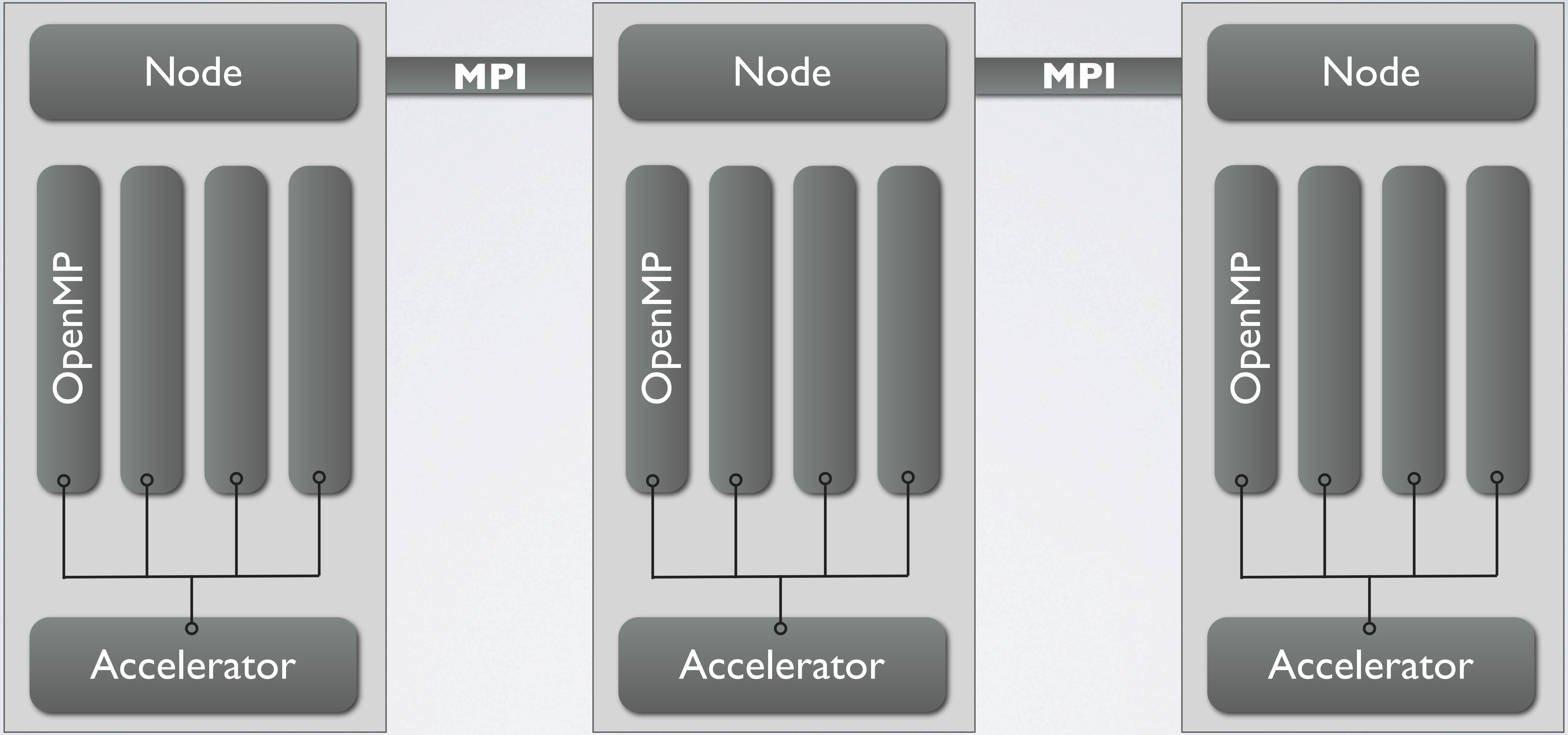
HYBRID PROGRAMMING MODEL

- On Jaguar, with 299,008 cores, we were seeing the limits of a single level of MPI scaling for most applications
- To take advantage of the vastly larger parallelism in Titan, users need to use hierarchical parallelism in their codes
    - Distributed memory: MPI, SHMEM, PGAS
    - Node Local: OpenMP, Pthreads, local MPI communicators
    - Within threads: Vector constructs on GPU, libraries, OpenACC
- These are the same types of constructs needed on all multi-PFLOPS computers to scale to the full size of the systems!

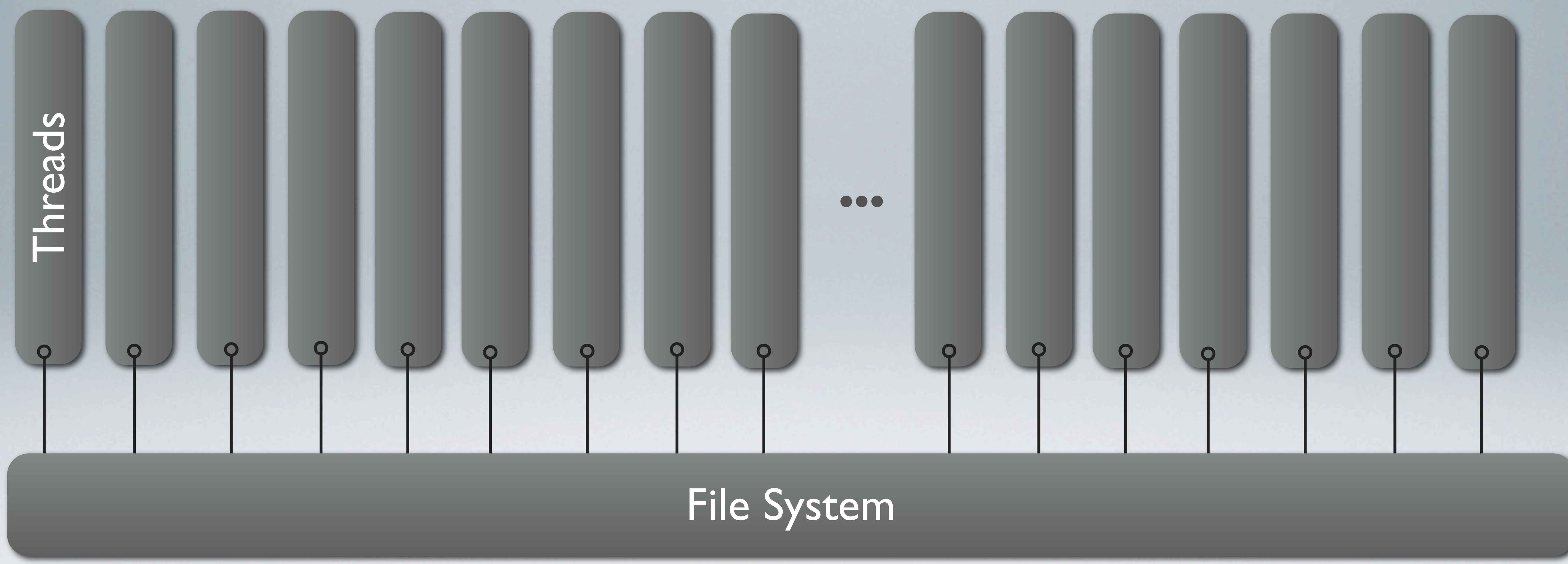
HYBRID PROGRAMMING MODEL



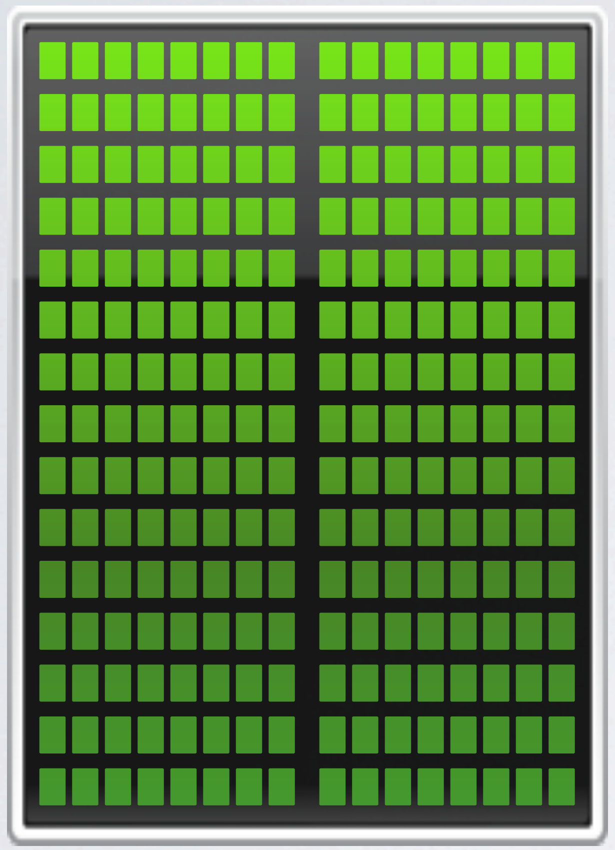
HYBRID PROGRAMMING MODEL

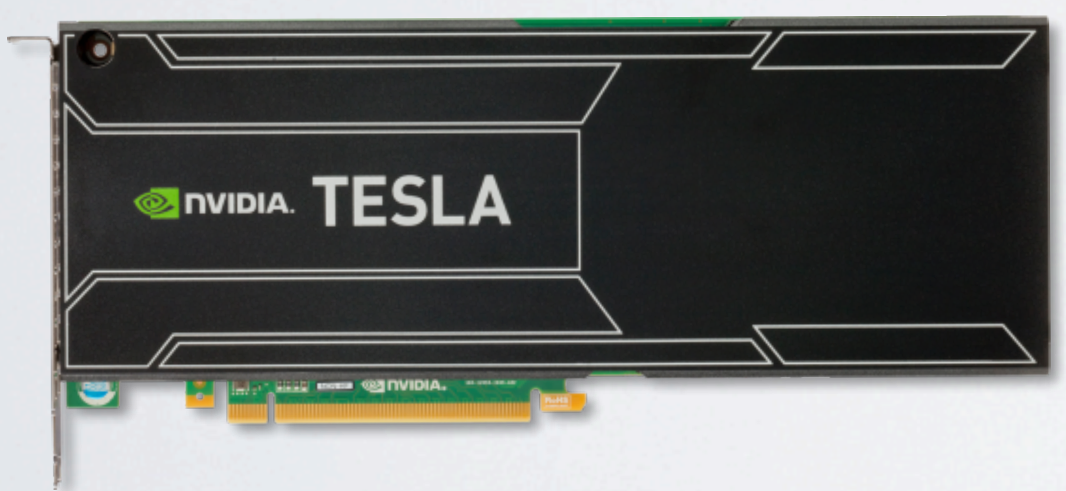


INPUT AND OUTPUT



GPUs: PATH TO EXASCALE





Hierarchical parallelism
Improve scalability of applications

Explicit data management
Between CPU and GPU memories

Data locality: Keep data near processing
GPU has high bandwidth to local memory and large internal cache

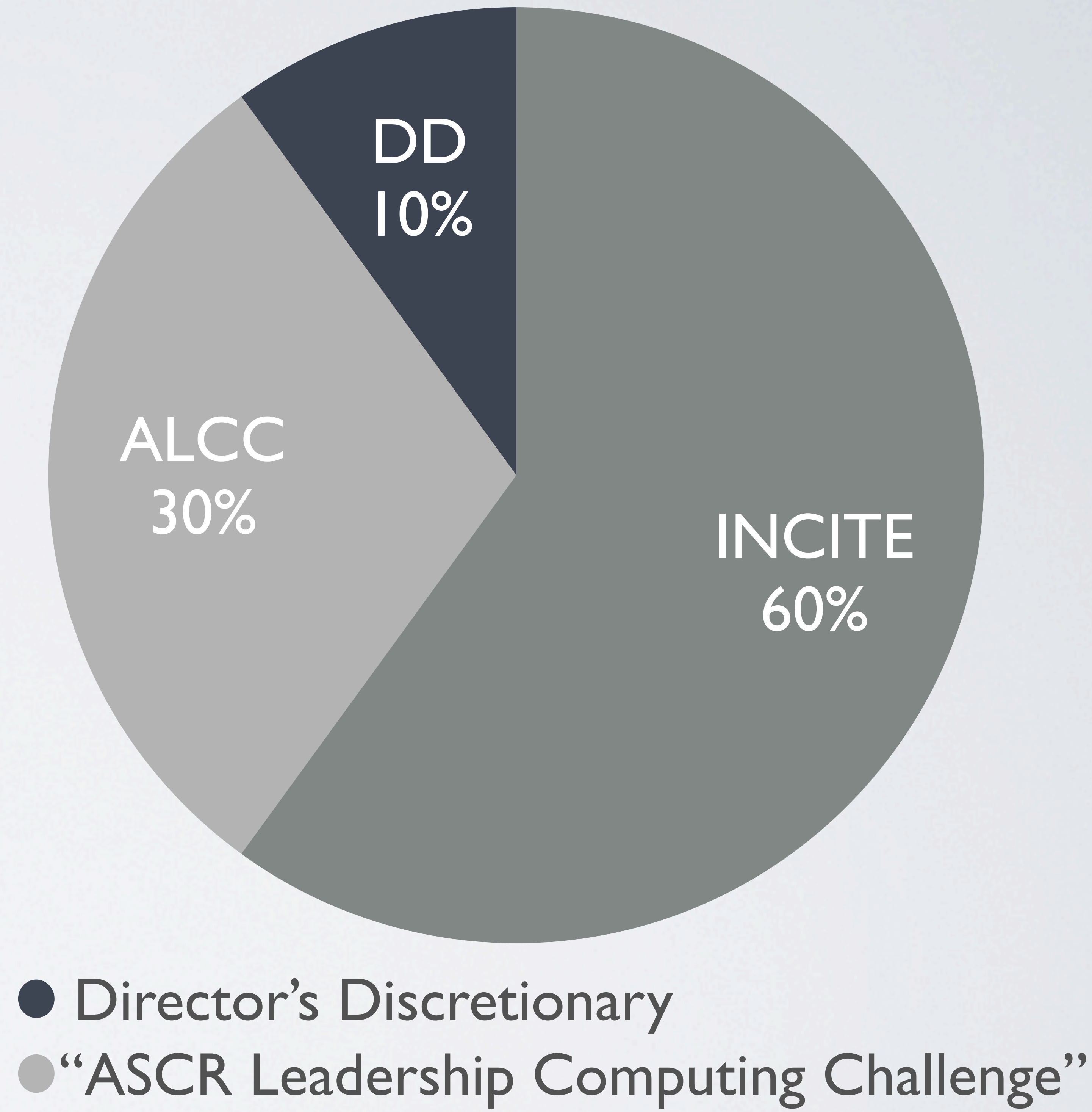
Expose more parallelism
Code refactoring and source code directives can double performance

Heterogeneous multicore processor architecture
Using right type of processor for each task

DOE ALLOCATION POLICY FOR LEADERSHIP FACILITIES

Primary Objective:

"Provide substantial allocations to the open science community through an peered process for a small number of high-impact scientific research projects"



OLCF ALLOCATION PROGRAMS

	INCITE		ALCC		Director's Discretionary	
Mission	High-risk, high-payoff science that requires LCF-scale resources		High-risk, high-payoff science aligned with DOE mission		Strategic LCF goals	
Call	1x/year (Closes June)		1x/year (Closes February)		Rolling	
Duration	1-3 years, yearly renewal		1 year		3m, 6m, 1 year	
Typical Size	30 - 40 projects	20M - 100M core-hours/yr.	5 - 10 projects	1M - 75M core-hours/yr.	100s of projects	10K -1M core-hours
Review Process	Scientific, Peer-Review	Computational Readiness	Scientific, Peer-Review	Computational Readiness	Strategic impact and feasibility	
Managed by	INCITE management committee (ALCF & OLCF)		DOE Office of Science		OLCF management	
Availability	Open to all scientific researchers and organizations including industry					

ABOUT OLCF SERVICES

	Software		
	Tools	Performance	Communications
Future	Training	Scaling	Internships
Everest	Advocacy	Collaboration	Tours
Viz Tech Ops	User Assistance	Liasons	Outreach

Oak Ridge Leadership Computing Facility

ACKNOWLEDGEMENTS

- OLCF-3 CAAR Team: Bronson Messer, Wayne Joubert, Mike Brown, Matt Norman, Markus Eisenbach, Ramanan Sankaran
- OLCF Users: Jackie Chen, Tom Evans, Markus Eisenbach,
- OLCF-3 Hardware Vendor Parters: Cray, AMD, and NVIDIA

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

QUESTIONS