BLUE WATERS

SUSTAINED PETASCALE COMPUTING

GPU Applications on Blue Waters

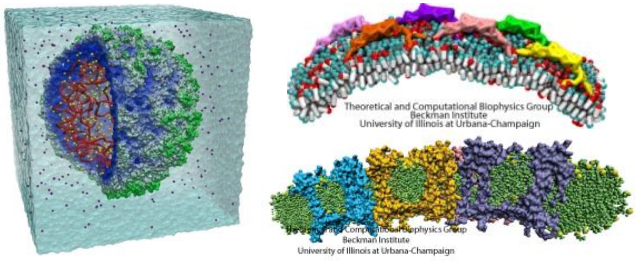
Cristina Beldica, PhD MBA
Blue Waters Executive Director
National Center for Supercomputing Applications, University of Illinois

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

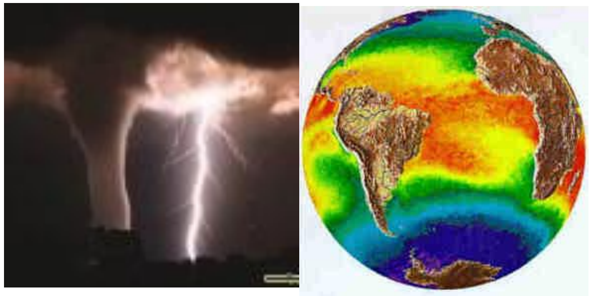
Sustained Petascale computing enables advances in a broad range of science and engineering disciplines

Examples include:

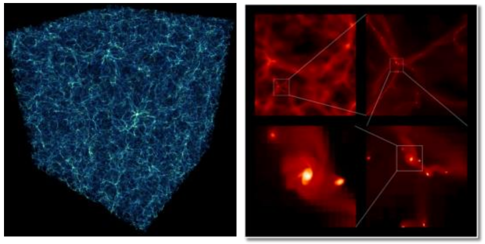
Molecular Science


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign
Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

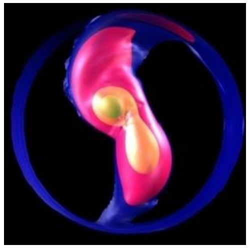
Weather & Climate Forecasting



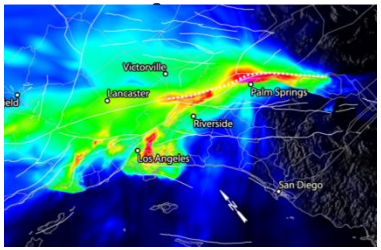
Astrophysics



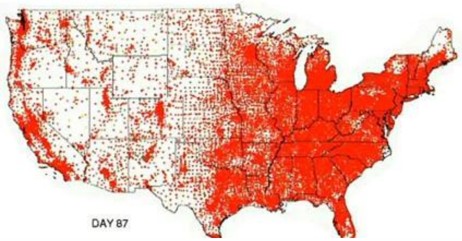
Astronomy



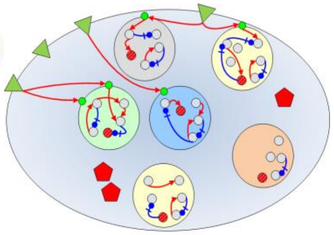
Earth



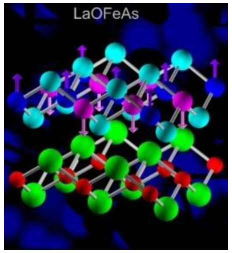
Health



Life Science

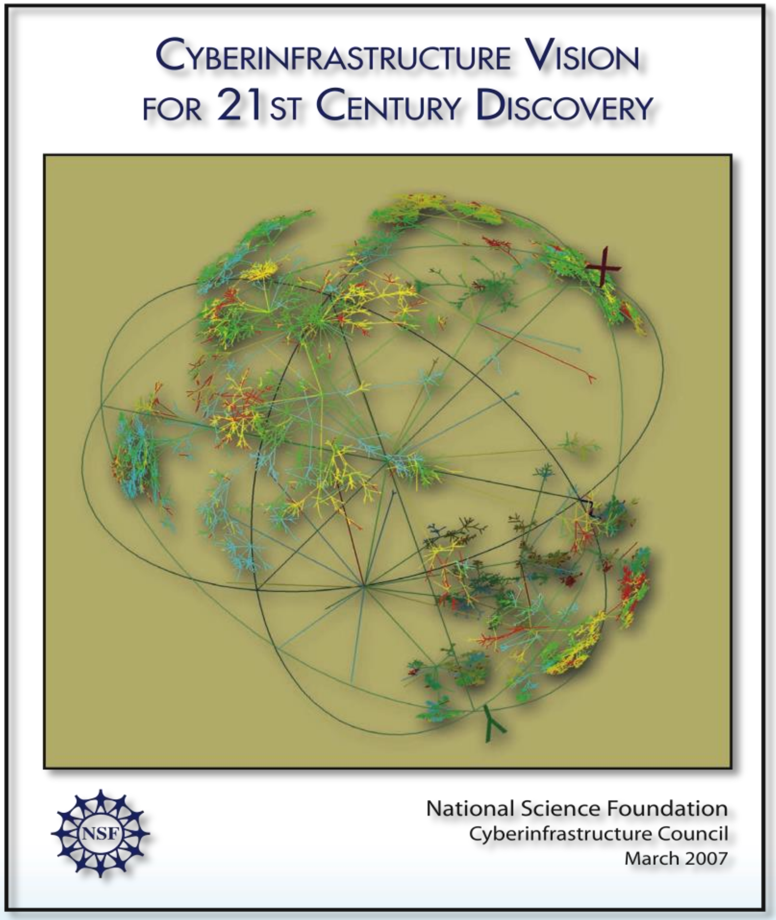


Materials



Background
NSF's Strategy for High-end Computing

- Three Resource Levels
  - Track 3: University owned and operated
  - Track 2: Several NSF-funded supercomputer & specialized computing centers (TeraGrid/XSEDE)
  - Track 1: NSF-funded leading-class computer center
- Computing Resources
  - Track 3: 10s–100 TF
  - Track 2: 500–1,000 TF
  - Track 1: see following slide



Cyberinfrastructure Vision for 21st Century Discovery

National Science Foundation
Cyberinfrastructure Council
March 2007

Background

NSF Track 1 Solicitation

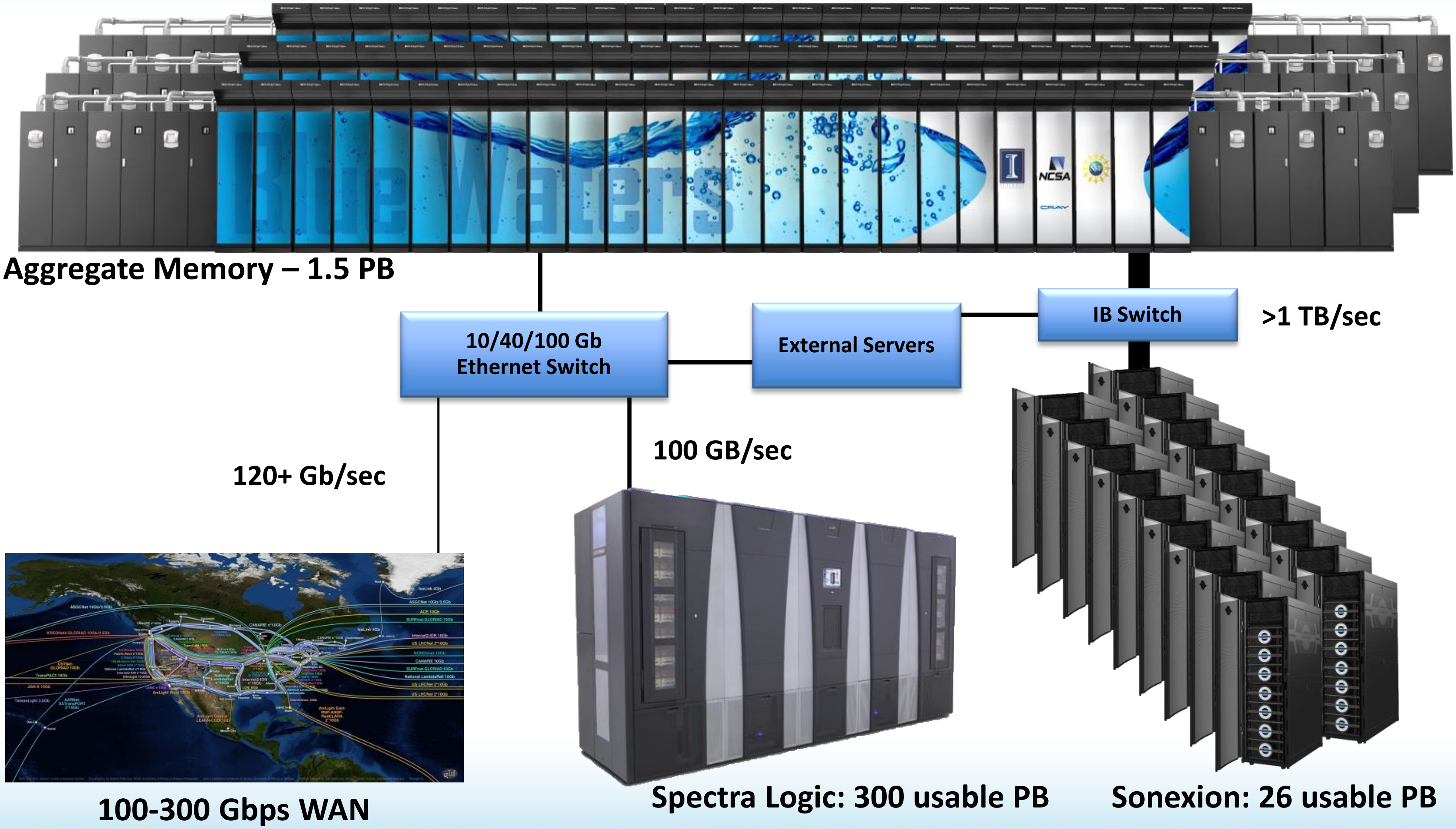
“The petascale HPC environment will enable investigations of computationally challenging problems that require computing systems capable of delivering **sustained performance approaching 10^{15} floating point operations per second** (petaflops) on real applications, that consume **large amounts of memory**, and/or that work with **very large data sets**.”

Leadership-Class System Acquisition - Creating a Petascale Computing Environment for Science and Engineering
NSF 06-573

Blue Waters Goals

- **Deploy a computing system capable of sustaining more than one petaflops or more for a broad range of applications**
  - Cray system achieves this goal using a well defined metrics
- **Enable the Science Teams to take full advantage of the sustained petascale computing system**
  - Blue Waters Team has established strong partnership with Science Teams, helping them to improve the performance and scalability of their applications
- **Enhance the operation and use of the sustained petascale system**
  - Blue Waters Team is developing tools, libraries and other system software to aid in operation of the system and to help scientists and engineers make effective use of the system
- **Provide a world-class computing environment for the petascale computing system**
  - The NPCF is a modern, energy-efficient data center with a rich WAN environment (100-400 Gbps) and data archive (>300 PB)
- **Exploit advances in innovative computing technology**
  - Proposal anticipated the rise of heterogeneous computing and planned to help the computational community transition to new modes for computational and data-driven science and engineering

Blue Waters Computing System



Aggregate Memory – 1.5 PB

10/40/100 Gb Ethernet Switch

External Servers

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec

100-300 Gbps WAN

Spectra Logic: 300 usable PB

Sonexion: 26 usable PB

Description of Blue Waters' Scale

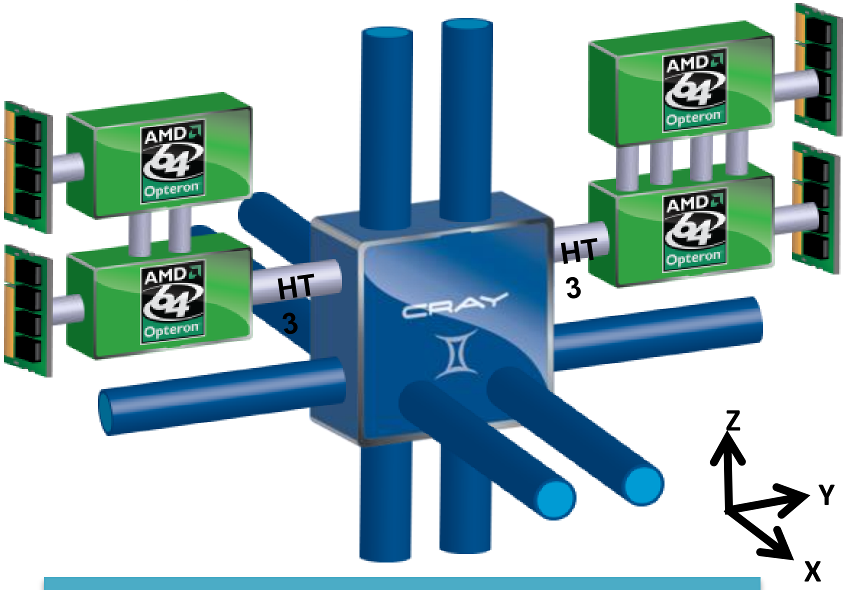
- What is a "core" these days?
- Rather than be imprecise, we express things in "node",
  - Two types of nodes – XE and XK
  - Smallest "schedulable" unit.
  - When we do use the term "core", we try to be precise, but if not, the default core definition is the Bulldozer full FMA cores in the AMD processor.

Racks	288
Nodes	27,648
Core-modules	397,824
x86 integer cores	795,648
K20x GPUs	4,224
CUDA cores/GPU	2,688
Total "cores"	12,547,584
Torus dimension	24x24x24
Gemini Routers	13,824
Physical Memory	1.6 PB

- Approx. 36 PB of raw on-line storage, 17,280 2TB disk drives, 730 LNET router nodes, 4,000 Intel SB cores.
- 72 Dedicated Cooling Rack Units
- 5 independent cooling control loops
- Inlet temperatures – Water 42-67˚F (5.6-19.4˚C), Air 78˚F (25.6˚C)

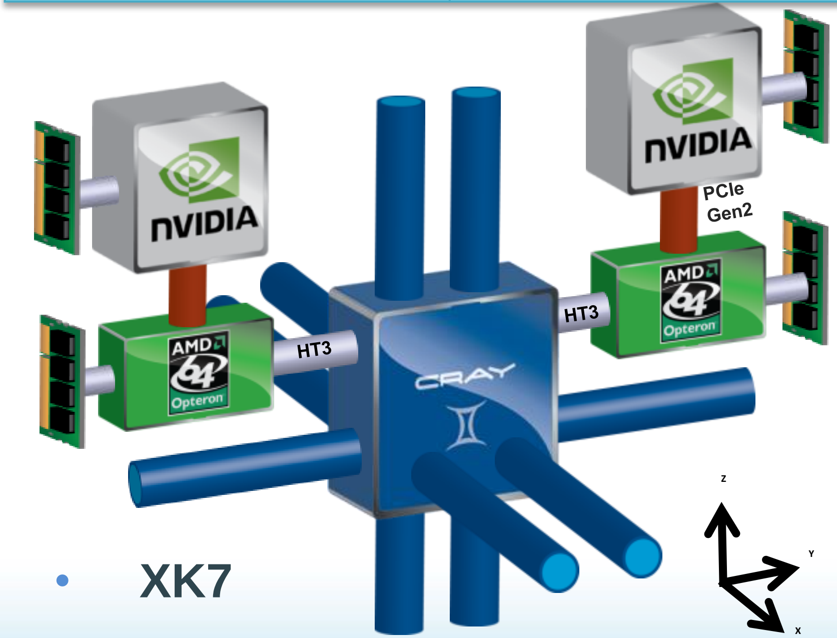
XE and XK Blue Waters Nodes

- XE6



Node Characteristics	
Number of Cores Modules	8
Peak Performance	156.8 Gflops/sec
Memory Size	32 GB per node
Memory Bandwidth (Peak)	51 GB/sec
Interconnect Injection Bandwidth (Peak)	9.6 GB/sec per direction

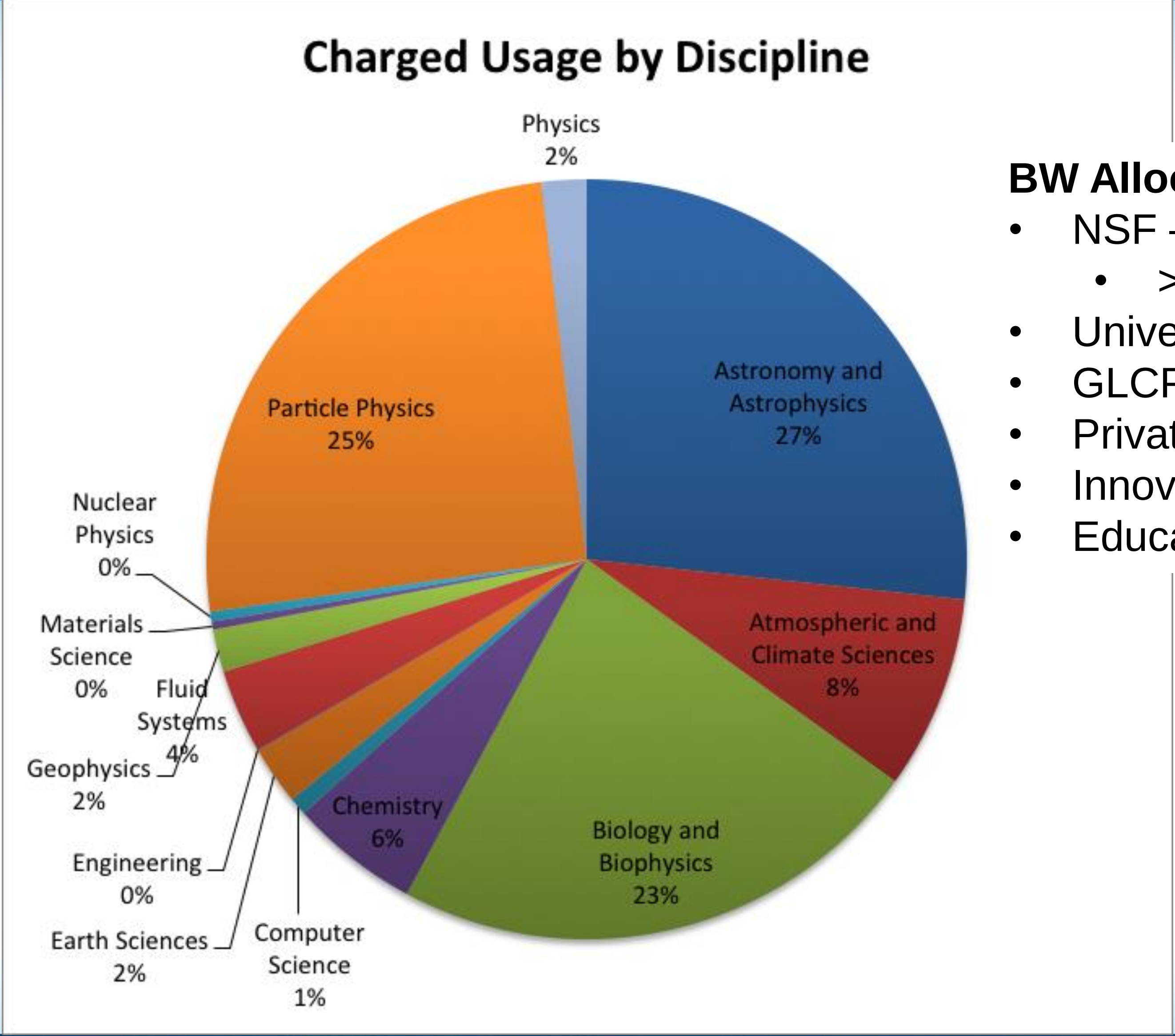
XK7 Compute Node Characteristics	
Host Processor	AMD Series 6200 (Interlagos)
Host Processor Performance	156.8 Gflops
Kepler Peak (DP floating point)	1.4 Tflops
Host Memory	32GB 51 GB/sec
Kepler Memory	6GB GDDR5 capacity 180 GB/sec



- XK7

BW Focus on Sustained Performance

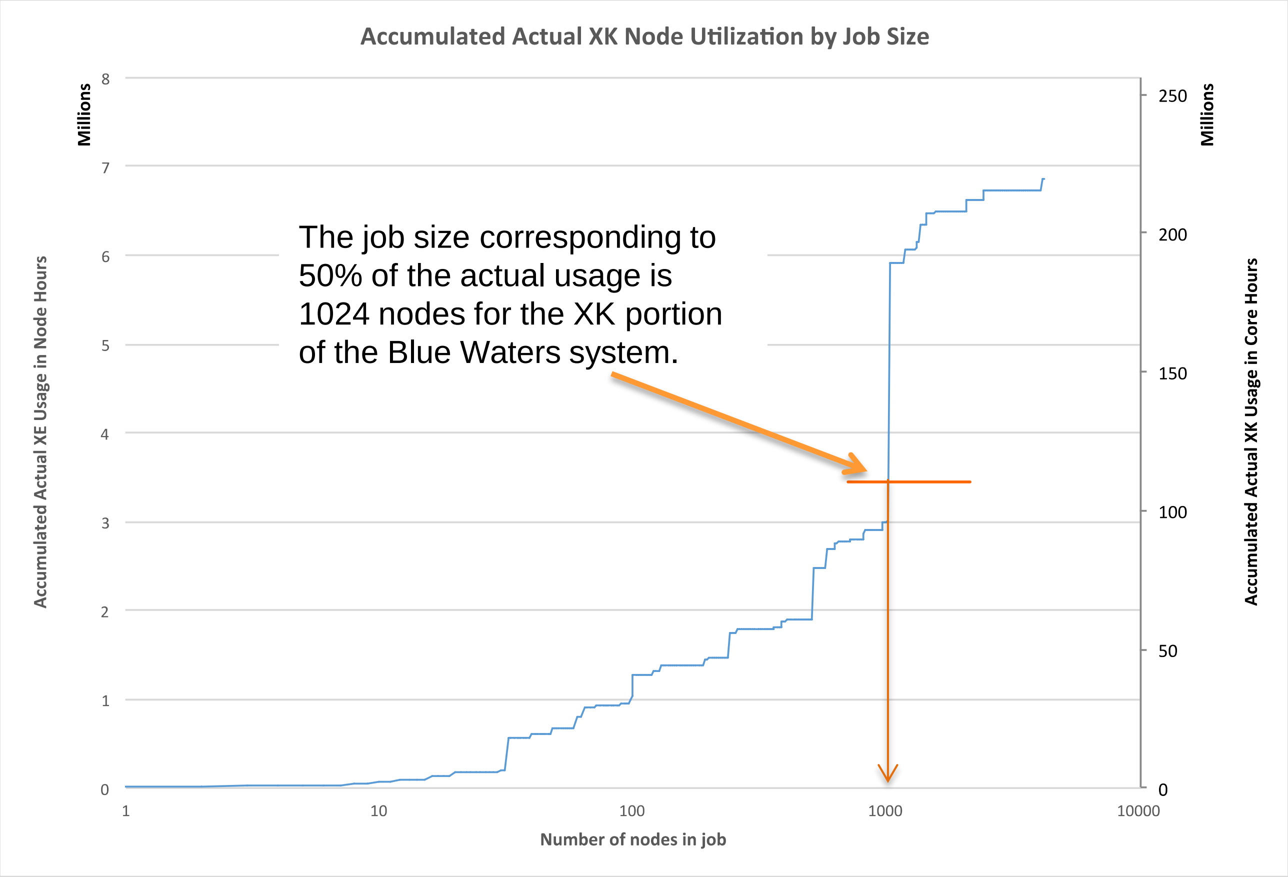
- **Blue Water's and NSF are focusing on *sustained* performance in a way few have been before.**
- *Sustained* is the computer's useful, consistent performance on a broad range of applications that scientists and engineers use every day.
  - Time to solution for a given amount of work is the important metric – not hardware Ops/s
  - Sustained performance (and therefore tests) include time to read data and write the results
- Full Scale SPP XE Codes
  - In addition to the NSF Petascale tests, four SPP tests ran above 1 PF using the full XE node section of the system; two of the four ran above 1.2 PF
  - Scale ranges from 21,417 to 22,528 nodes
- SPP XK codes x86 to Kepler Speed ups
  - Four XK SPP codes (NAMD, Chroma, QMCPACK, and GAMESS) all show a runtime improvement between 3.1-4.9x over x86 version running at same scale.
    - Scale ranges from 700 to 1,536 nodes
  - Three codes were CUDA implementation, one code (GAMESS) was an OpenACC implementation



BW Allocation Mechanisms

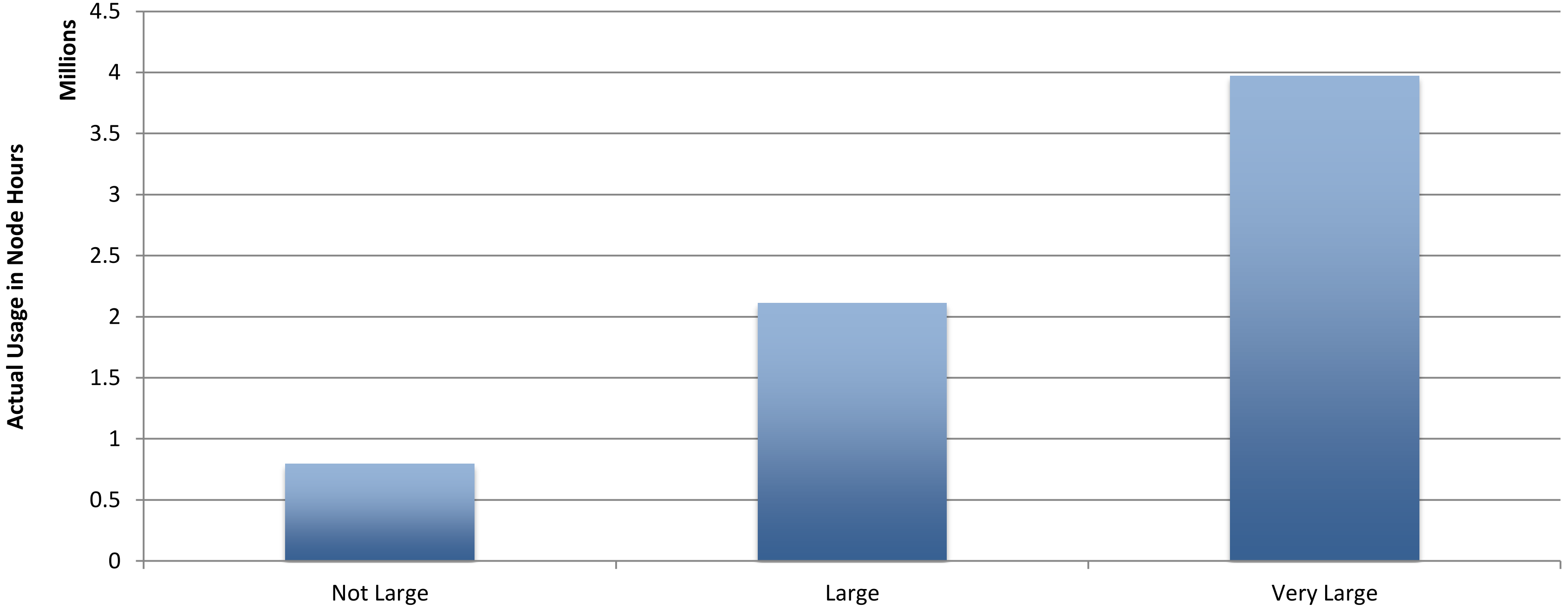
- NSF – PRAC process
  - > 80% of available time
- University of Illinois (7%)
- GLCPC (2%)
- Private sector (5%)
- Innovation and Exploration (5%)
- Education (1%)

Accumulated usage for the XK nodes per job size



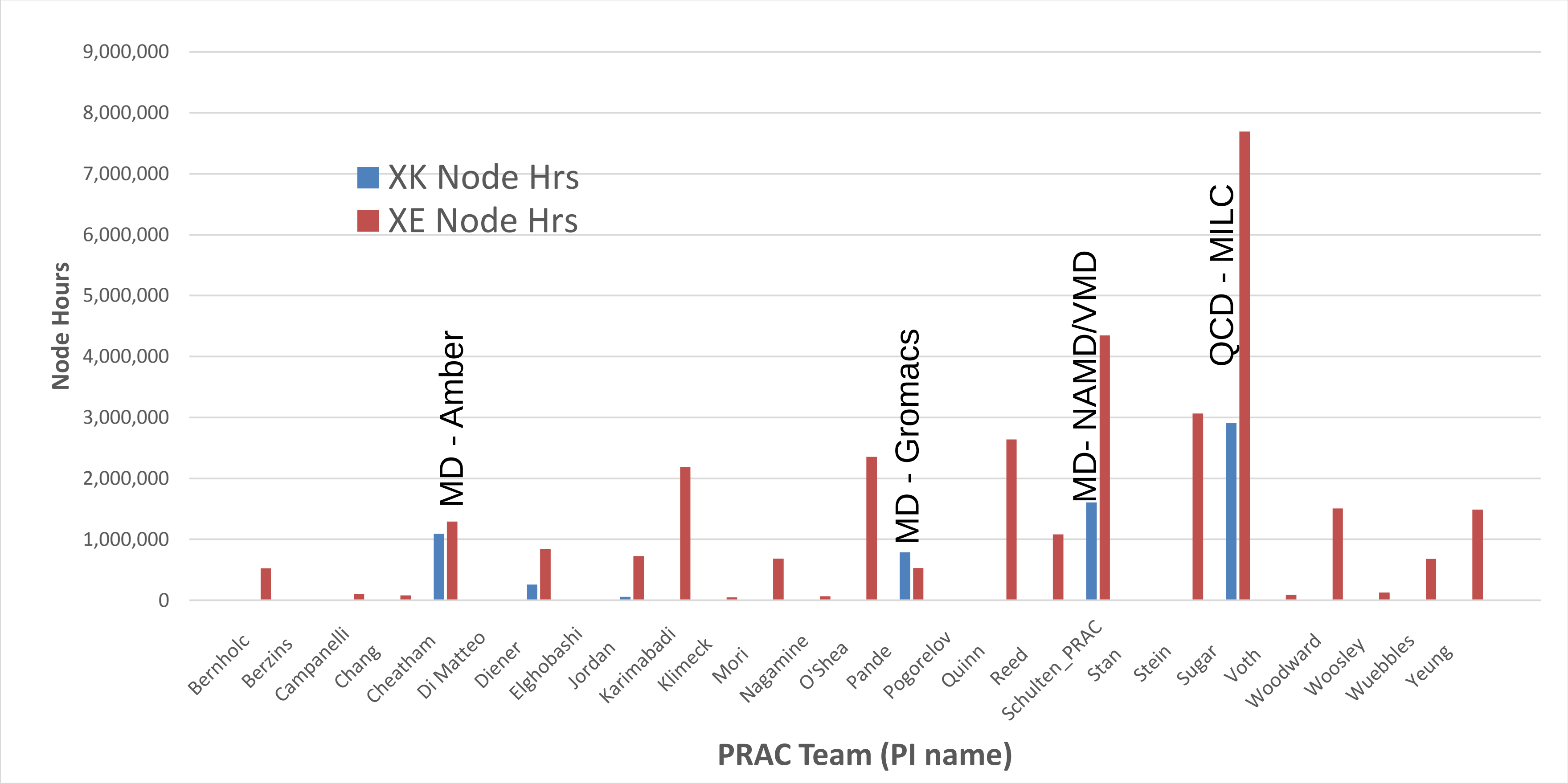
The job size corresponding to 50% of the actual usage is 1024 nodes for the XK portion of the Blue Waters system.

Actual XK Usage by Job Size

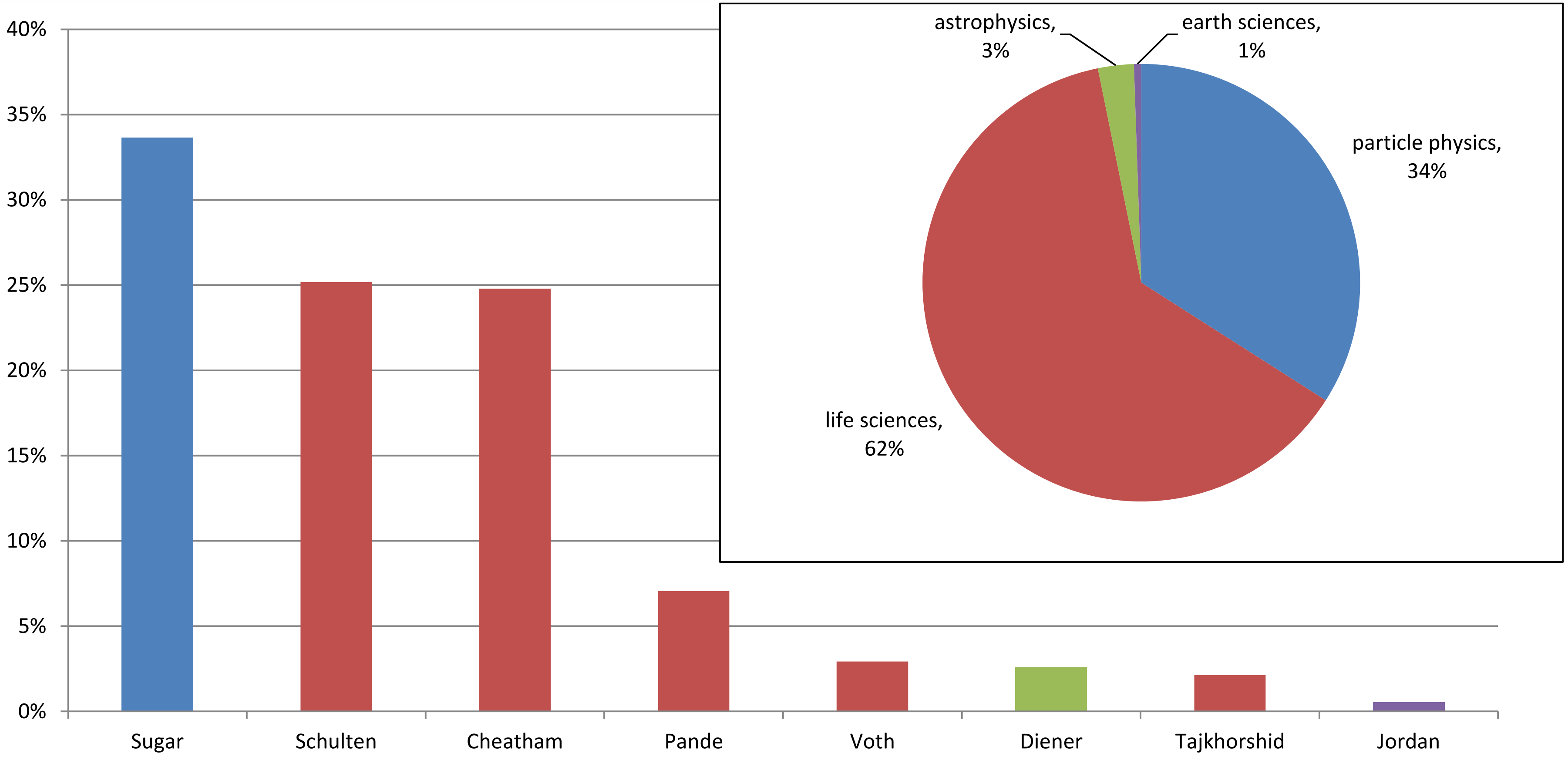


	Not Large	Large	Very Large
XE nodes	1- 511 nodes	512 - 4,528 nodes	4,529 - 26,864 nodes
XK nodes	1 - 63 nodes	64 - 845 nodes	846 – 4,224 nodes

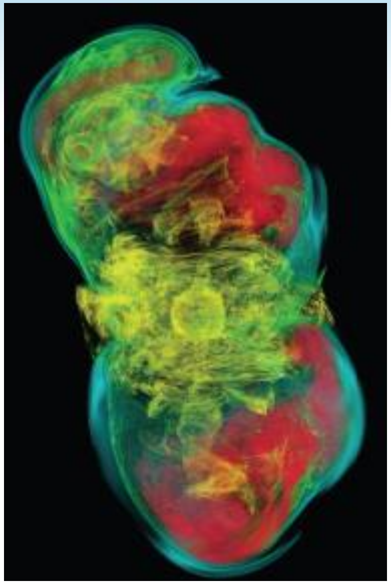
XE and XK usage by PRAC teams



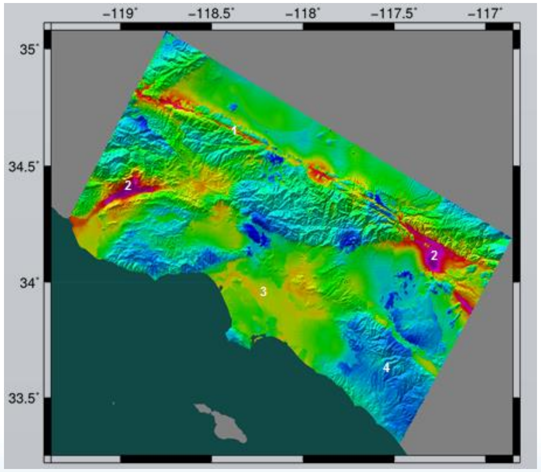
GPU Power Users on Blue Waters



Projects Using GPUs on BW

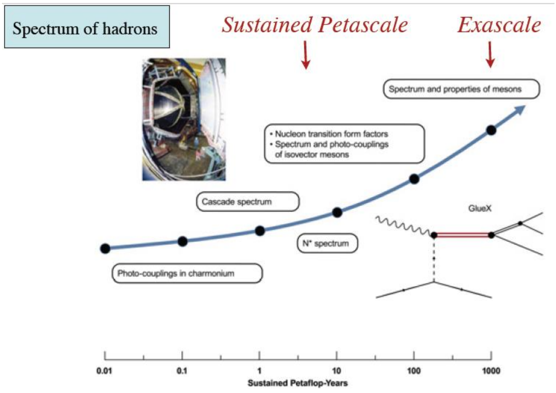


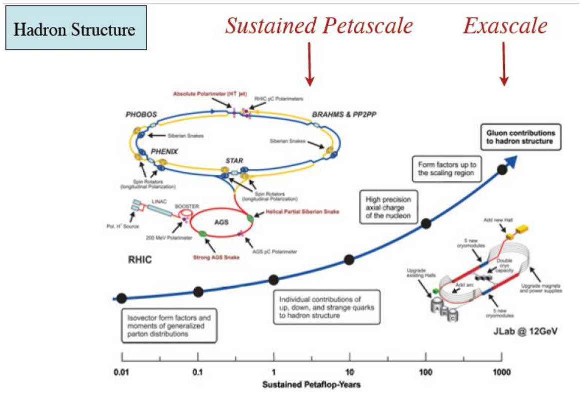
- *Lattice QCD on Blue Waters* - Robert Sugar, University of California, Santa Barbara
- *The Computational Microscope* - Klaus Schulten, University of Illinois at Urbana-Champaign
- *Hierarchical molecular dynamics sampling for assessing pathways and free energies of RNA catalysis, ligand binding, and conformational change,* Thomas Cheatham, Univ. of Utah
- *Simulating vesicle fusion on Blue Waters*, Vijay Pande, Stanford
- *Petascale Multiscale Simulations of Biomolecular Systems* , Gregory Voth, University of Chicago
- *From Binary Systems and Stellar Core Collapse To Gamma-Ray Bursts* - Peter Diener, Louisiana State University
- *Characterizing Structural Transitions of Membrane Transport Proteins at Atomic Details* - Emad Tajkhorshid, University of Illinois at Urbana-Champaign
- *Petascale Research in Earthquake System Science on Blue Waters (PressOn)* - Thomas Jordan, University of Southern California

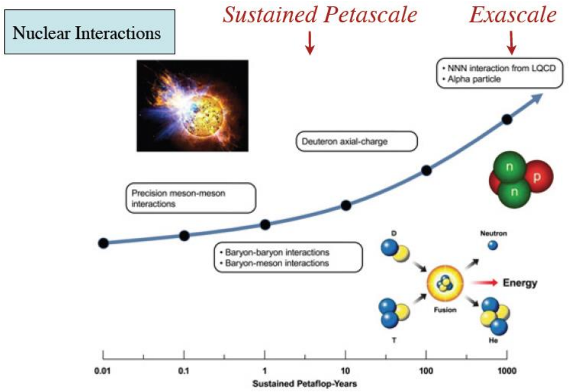


LQCD’s need for speed (FLOPs)

- LQCD – Lattice Quantum Chromo-Dynamics

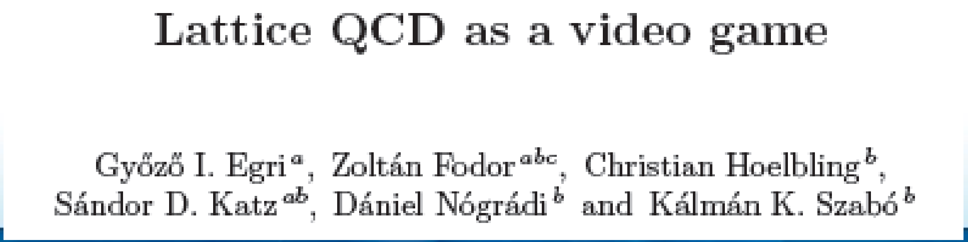






Source: Joo “GPU Computing and LQCD”, 2011

- Recognized early on the potential for doing LQCD on GPU: Egri et al. Comput. Phys. Commun. 177:631-639, 2007



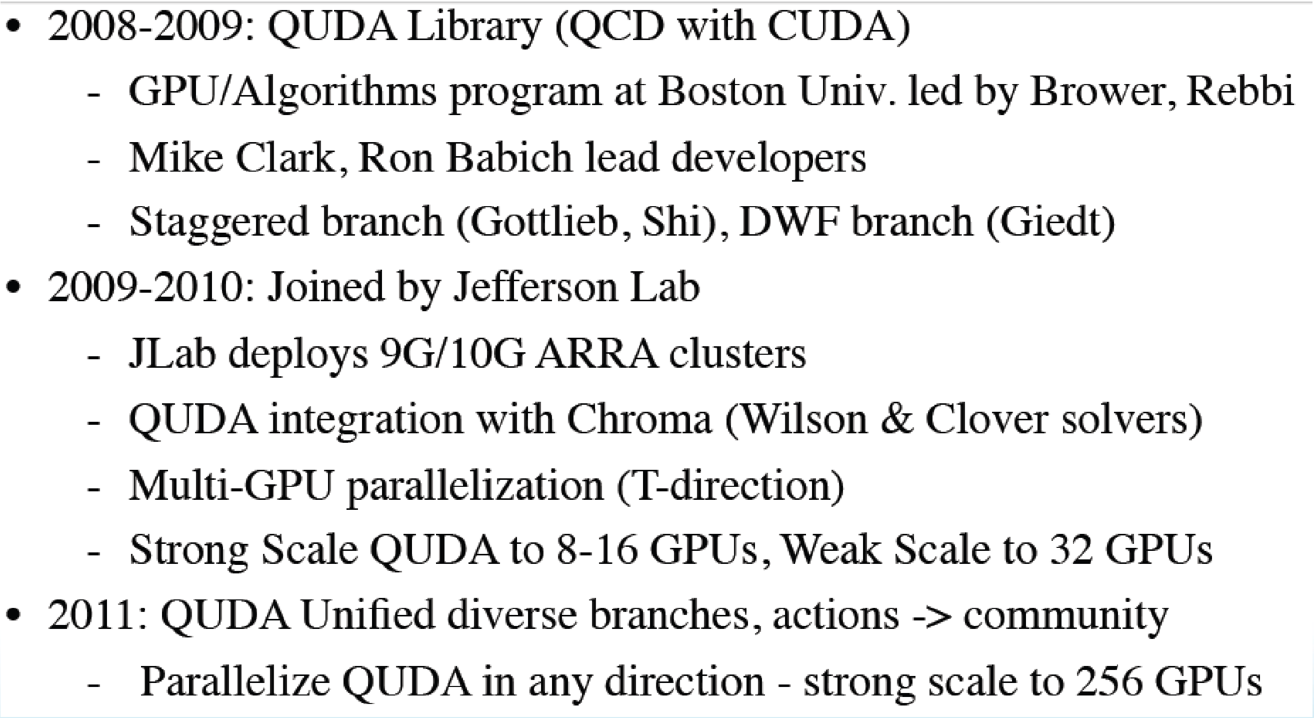
Lattice QCD as a video game

Győző I. Egri[a], Zoltán Fodor[abc], Christian Hoelbling[b],
Sándor D. Katz[ab], Dániel Nógrádi[b] and Kálmán K. Szabó[b]

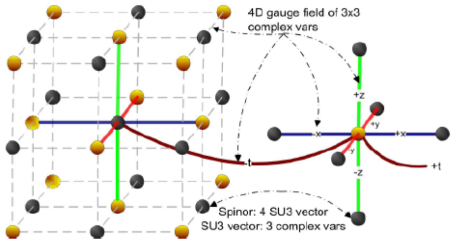
LQCD’s need for speed (FLOPs)

- Good community codes development: USQCD
- Common CUDA library for LQCD: QUDA

https://github.com/lattice/quda



- 2008-2009: QUDA Library (QCD with CUDA)
  - GPU/Algorithms program at Boston Univ. led by Brower, Rebbi
  - Mike Clark, Ron Babich lead developers
  - Staggered branch (Gottlieb, Shi), DWF branch (Giedt)
- 2009-2010: Joined by Jefferson Lab
  - JLab deploys 9G/10G ARRA clusters
  - QUDA integration with Chroma (Wilson & Clover solvers)
  - Multi-GPU parallelization (T-direction)
  - Strong Scale QUDA to 8-16 GPUs, Weak Scale to 32 GPUs
- 2011: QUDA Unified diverse branches, actions -> community
  - Parallelize QUDA in any direction - strong scale to 256 GPUs


4D gauge field of 3x3 complex vars
+z
+y
+x
-x
-y
-z
+t
-t
Spinor: 4 SU3 vector
SU3 vector: 3 complex vars

Four dimensional space-time Lattice QCD.

Source: Ibrahim IRISA/INRIA

Source: Joo “GPU Computing and LQCD”, 2011

LQCD’s need for speed (FLOPs)

- LQCD PRAC project on Blue Waters
  - MILC based applications
  - Chroma based application
  - Largest BW user: allocation ~ 30 M node-hours per year
  - More than 20% node-hours on XK nodes

  - Chroma benchmark
    - 3.9x speed-up XK CPU+GPU to XK CPU
    - 2.4x speed-up XK CPU+GPU to XE CPU+CPU

Source: Hwu “GPU Computing in Blue Waters” 2013

NAMD: GPUs accelerate molecular dynamics

History:

- 2007: GPU development started
- 10/2010: NAMD 2.7 released with CUDA support for key computations
- 2010-2014: More computation moved to GPUs

Challenges:

- Communication time (node to node, main memory to GPU memory) limit what calculations GPU can perform efficiently
- GPU support for new features lag CPU version

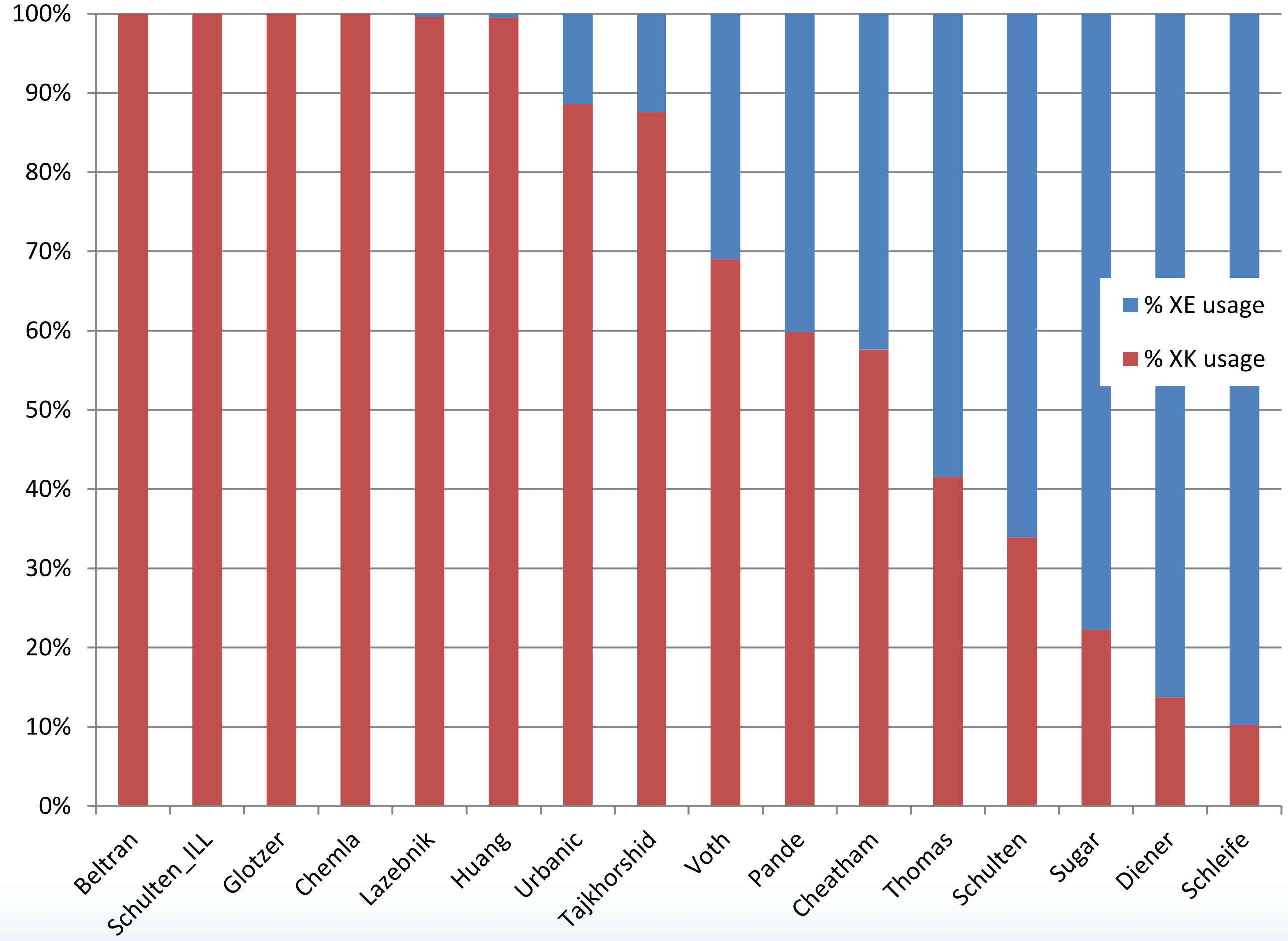
NAMD on Blue Waters: GPUs enable unprecedented simulation and analysis

HIV-1 Capsid simulation:

- 64 million atoms
- Scales to full BW XK7 GPU partition (4,224 nodes) and beyond
- XK7 3-4 times faster than XE6 CPU nodes at scale
- BW XK7s also critical for parallel analysis of simulation results using VMD (incl. OpenGL on BW*)

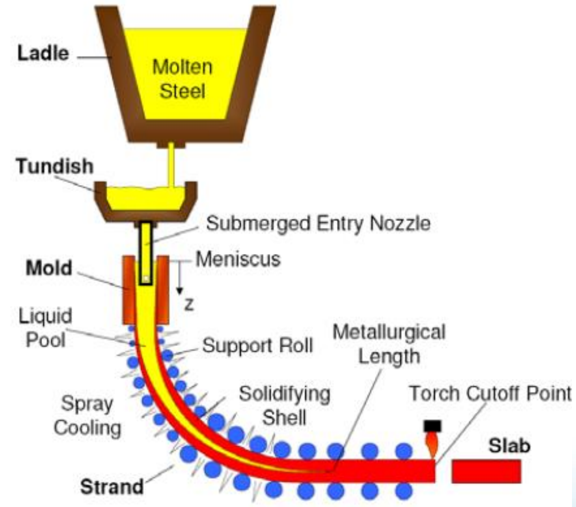
Zhao et al., Nature 497: 643-646 (2013)

GPU compared to CPU usage – top users



Other Applications Using GPUs on BW

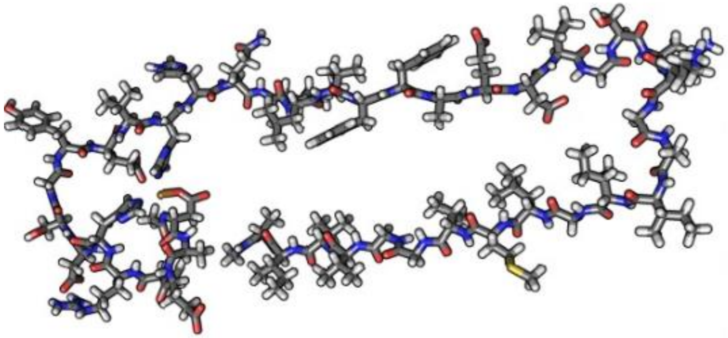
- *GPU-Accelerated and Monte Carlo Based Robust Optimization for Spot Scanning Proton Therapy* - Chris Beltran, Mayo Clinic
- *Petascale Quantum Simulations of Nano Systems and Biomolecules* - Jerzy Bernholc, North Carolina State University at Raleigh
- *C. Crescentus Cell Division Using Our In-House Lattice Microbe Simulation Program AND Interactions Between Ribosomal Signatures and 5' and Central Domain of the Ribosomal Small Subunit Using NAMD 2.9 Accelerated by GPUS* - Zaida Luthey-Schulten, University of Illinois at Urbana-Champaign
- *Fluid-Flow and Stress Analysis of Steel Continuous Casting* - Brian Thomas, University of Illinois at Urbana-Champaign
- *Large-Scale Incremental Visual Learning Using Rich Feature Hierarchies* - Svetlana Lazebnik, University of Illinois at Urbana-Champaign
- *Feature Learning by Large-Scale Heterogeneous Networks with Application to Face Verification* - Thomas S. Huang, University of Illinois at Urbana-Champaign
- *Implicitly-Parallel Functional Dataflow for Productive Hybrid Programming on Blue Waters*, Michael Wilde, University of Chicago

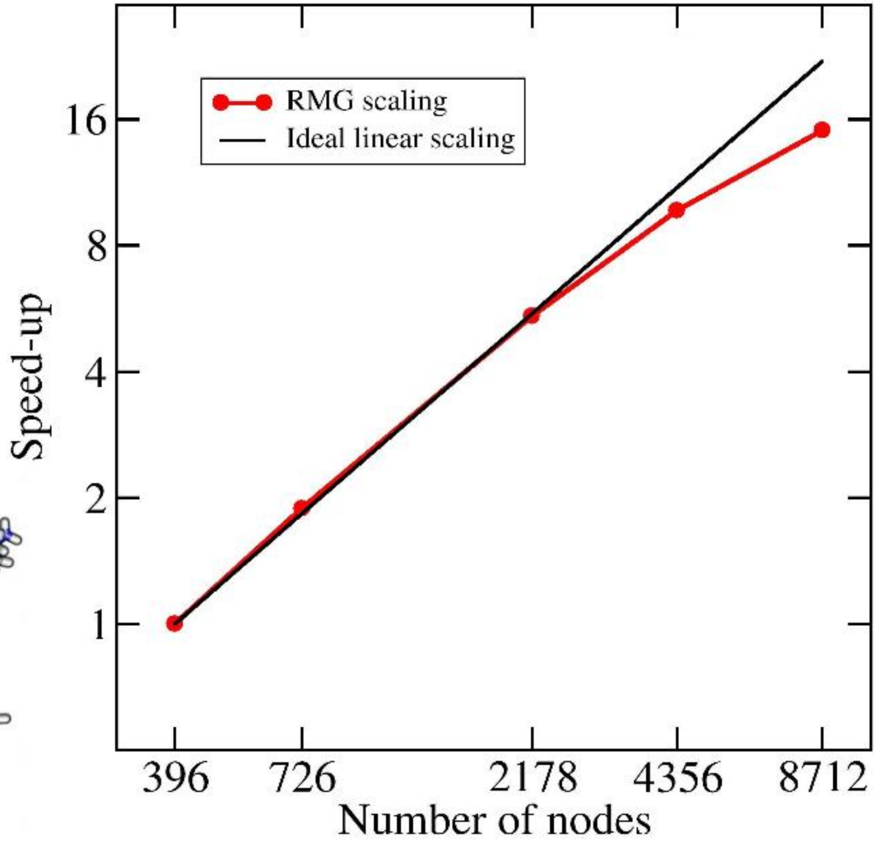


Scaling Results for RMG (Bernholc)

Test problem: Gas phase Amyloid Beta 1-42 protein
Test system: Cray XK7
1 node = 16 Opeteron cores + 1 Nvidia K20x GPU accelerator
Strong scaling

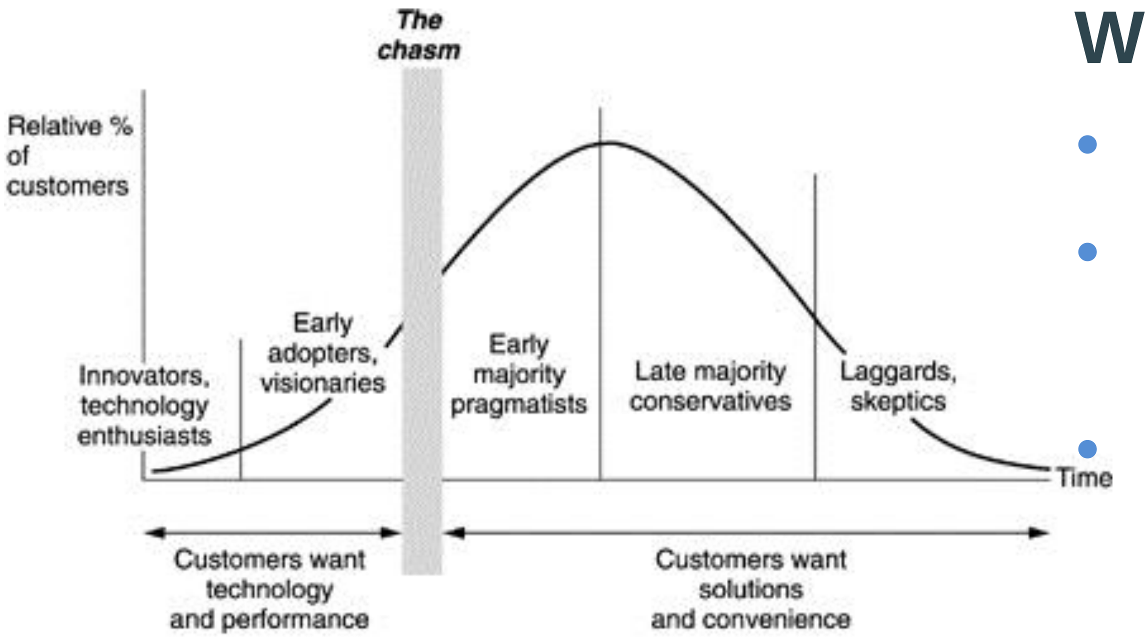
Largest run used **139,392** CPU cores and **8712** GPU's.





Source: http://es13.wm.edu/talks/Briggs.pdf

Increasing adoption rates


The chasm
Relative % of customers
Innovators, technology enthusiasts
Early adopters, visionaries
Early majority pragmatists
Late majority conservatives
Laggards, skeptics
Time
Customers want technology and performance
Customers want solutions and convenience

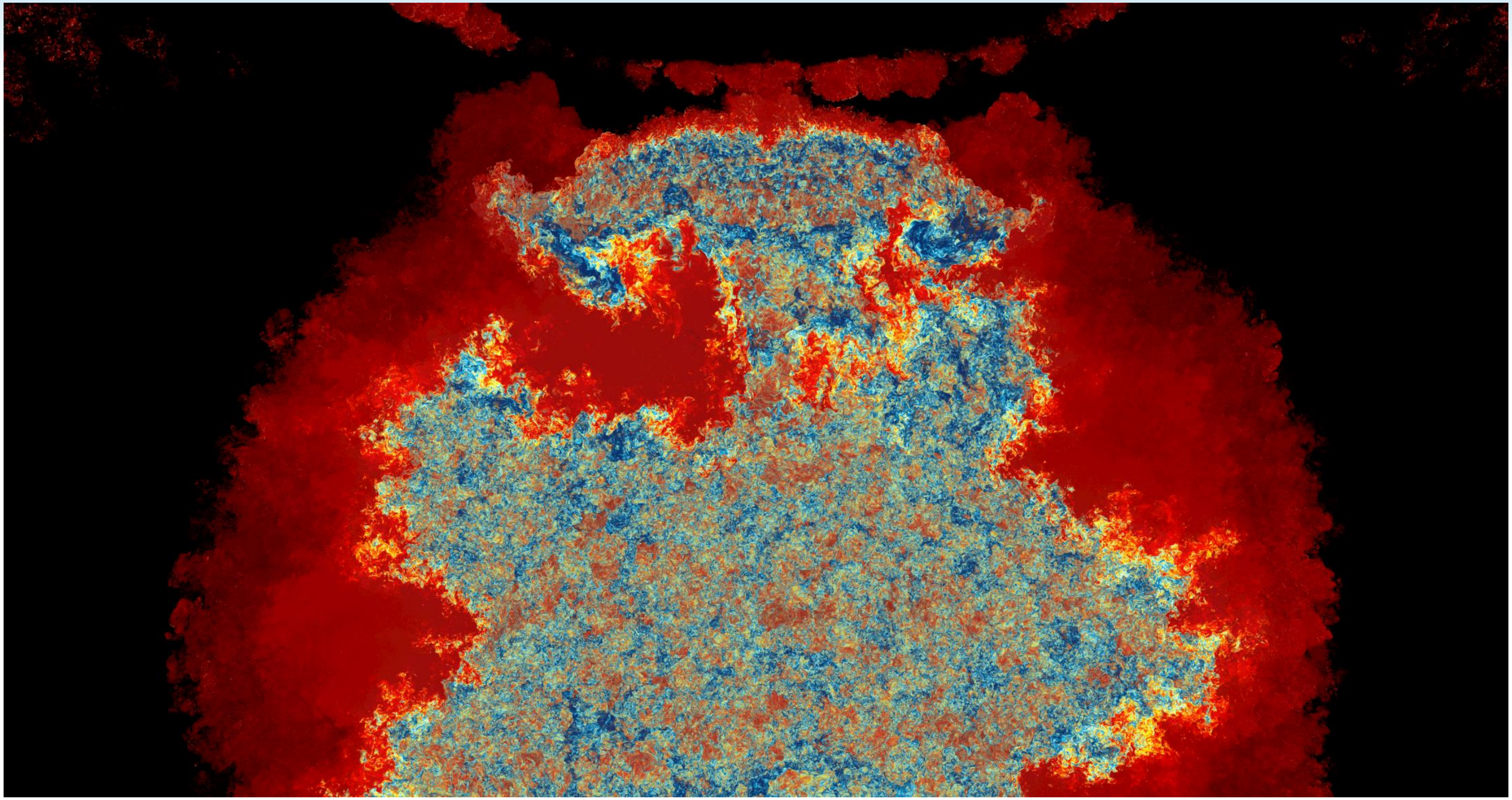
Wish List

- Training
- Single Source for multiple architectures
- Tools to automate code conversion and analysis

Impact of OpenGL on XK

- NCSA, following S&E team suggestions, convinced Cray and NVIDIA to run a Kepler driver that enables OpenGL applications
  - Allows visualization tools to run directly on XK nodes
  - First system to do this
- Two early impacts
  - Schulten – NAMD
    - 10X to 50X rendering speedup in VMD.
    - OpenGL render backup to ray tracing. Used to fill in failed ray traced frames.
    - Potential for interactive remote display.
  - Woodward
    - Eliminate need to move data.
    - Created movies of large data simulation in days rather than months.

John E. Stone, Barry Isralewitz, and Klaus Schulten. "Early experiences scaling VMD molecular visualization and analysis jobs on Blue Waters." In Proceedings of the 2013 Extreme Scaling Workshop.



PI: Woodward: Image 1000 of the movie of 1711 frames. This is one eye's view, and the image has 4096x2160 pixels. It shows the mixing fraction between ICF capsule material and the enclosed fuel in a simplified numerical experiment where both materials are represented by ideal, monatomic gases. Pure fluids of either type are transparent. As the concentration by volume of dense capsule gas increases, colors go from dark blue, through aqua, white, yellow, to red.

Real Improvement in Time to Solution

- 10,560^3 grid Inertial confinement fusion (ICF) calculation with multifluid PPM
- Rendered 13,688 frames at 2048x1080 pixelsv4 panels per view & 2 views per stereo image @ 4096x2160 pixels. Stereo movie is 1711 frames

	Local (Minnesota)	Remote (NCSA)
Raw data transfer time	26TB @ 20MB/s = 15 days	0 secs
Rendering time 13,688 frames	Estimated 33 days (6 nodes with 1 GPU/node)	24 hours (128 GPUs)
Visualized data transfer	0 secs	38GB @ 20MB/s = 32 minutes
Total time	Min 33, max 48 Days	24.5 hours **About 40x speedup + better analysis**

Here we compare the time required to generate the movie of our trillion-cell ICF simulation with the multifluid PPM code, performed in Dec., 2012. Rendering the movie using Blue Waters' GPU nodes produces a movie in one day that would otherwise take a month and a half of continuous data transfer and image rendering at Minnesota.

New Generation of Scientists and Engineers

- Undergraduate Education
  - Professional Development Workshops for Undergraduate Faculty
  - Research Experiences for Undergraduates
  - Undergraduate Materials Development by Undergraduate Faculty
- Virtual School of Computational Science and Engineering (VSCSE)
  - Proven Algorithmic Techniques for Many-core Processors (2008, 2009, 2010, 2011, 2012, 2014)
  - More than 2,000 students participated
- OpenAcc XSEDE workshops
  - John Urbanic (Nov 2013, Apr 2014, Aug 2014)
- Graduate Fellowships
  - 10 fellowships awarded for 2014-2015

Example: Ariana Minot, Harvard University, will study the implementation of a GPU-accelerated particle filter algorithm and the potential of more general sequential Monte Carlo methods to better understand the behavior of the electric power grid under a high penetration of renewable energy sources and distributed smart loads.

NCSA/UIUC Enhanced Intellectual Services for Petascale Performance – NEIS-P^2

- Larger computational S&E community needs assistance to take full advantage of the capability provided by today's computing systems
- Increasing performance requires dramatic increases in parallelism that then generates complexity challenges for science and engineering teams
  - Scaling applications to large core counts
  - Effectively using accelerators and "many-core" processors
  - Using general purpose and heterogeneous (accelerated) nodes in a single, coordinated simulation
  - Effectively using parallel IO systems for data-intensive applications
  - Enhancing application flexibility to increasing effective, efficient use of systems
- NEIS-P2 Component 1 - Direct PRAC Funding
  - Provide funding to NSF PRAC teams to re-engineer applications, develop new algorithmic methods, and expand use of heterogeneous computing - 21 PRAC teams
  - Several used this funding to design and implement GPU versions of their code modules (https://bluewaters.ncsa.illinois.edu/neis-p2-final-reports)
- NEIS-P2 Component 3

NEIS-P^2 Component 3

- Prof. Wen-mei Hwu – co-PI of the Blue Waters Project
  - PI of the first NVIDIA CUDA Center of Excellence (since 2008 at UIUC)
  - SW development efforts
    - GMAC library
    - DL compiler-based memory layout transformation tool
    - TC compiler-based tool for thread coarsening and data tiling
  - VSCSE Summer School courses, semester-long CIC course, Coursera
- Forum for Developing Collaborative Opportunities
- Support for Science Teams to Modify Codes to Use New Methods
- Technical Guidance, Collaborative Support, and Quality Assessment

Questions

Blue Waters Portal

http://bluewaters.ncsa.illinois.edu

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

The work described was achieved through the efforts of many other teams.