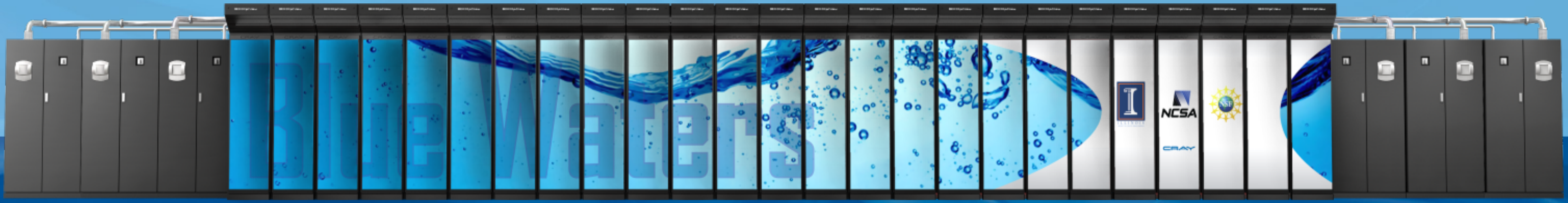
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Extreme Scale Computational Atmospheric Science – Insights for the Petascale Vantage Point

Dr. William Kramer

National Center for Supercomputing Applications, University of Illinois



GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

CRAY

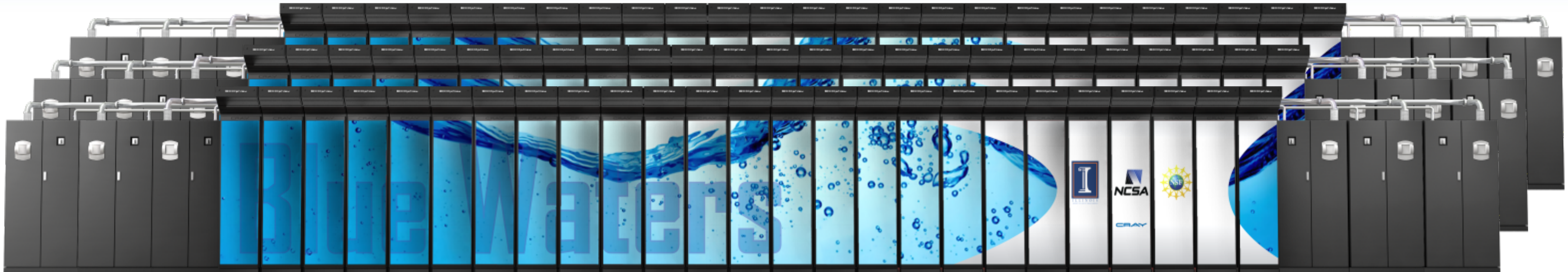
Outline

- A little about Blue Waters
- A little about what Sustained Performance means to us
- A little about usage
- Some Geoscience projects
- A few observations from the BW *experiment* that are important to future Extreme scale systems

Blue Waters Goals

- **Deploy a computing system capable of <u>sustaining</u> more than one petaflops or more for a <u>broad</u> 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 – 1.6 PB

10/40/100 Gb Ethernet Switch

External Servers

IB Switch

>1 TB/sec

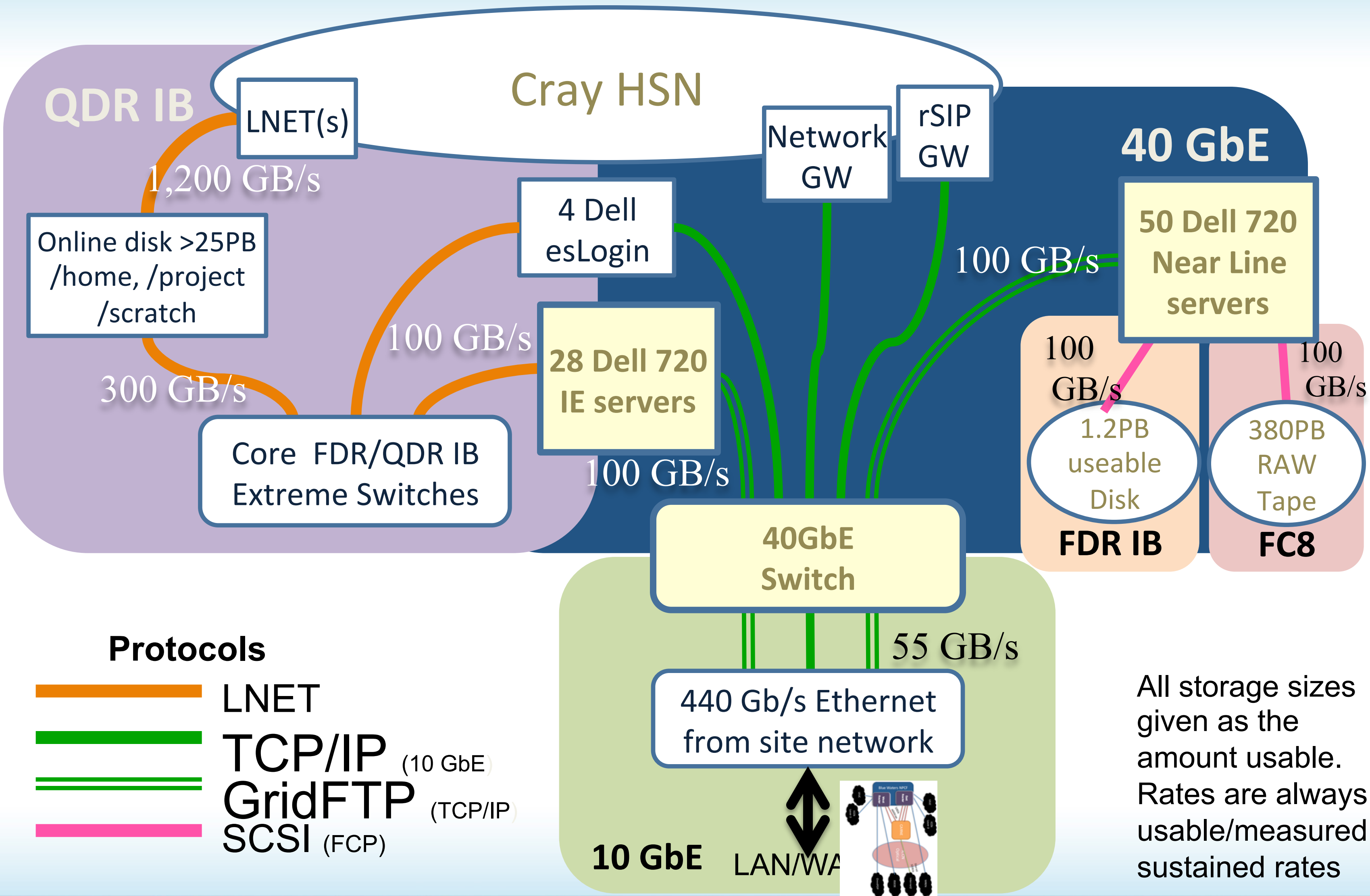
120+ Gb/sec

100 GB/sec

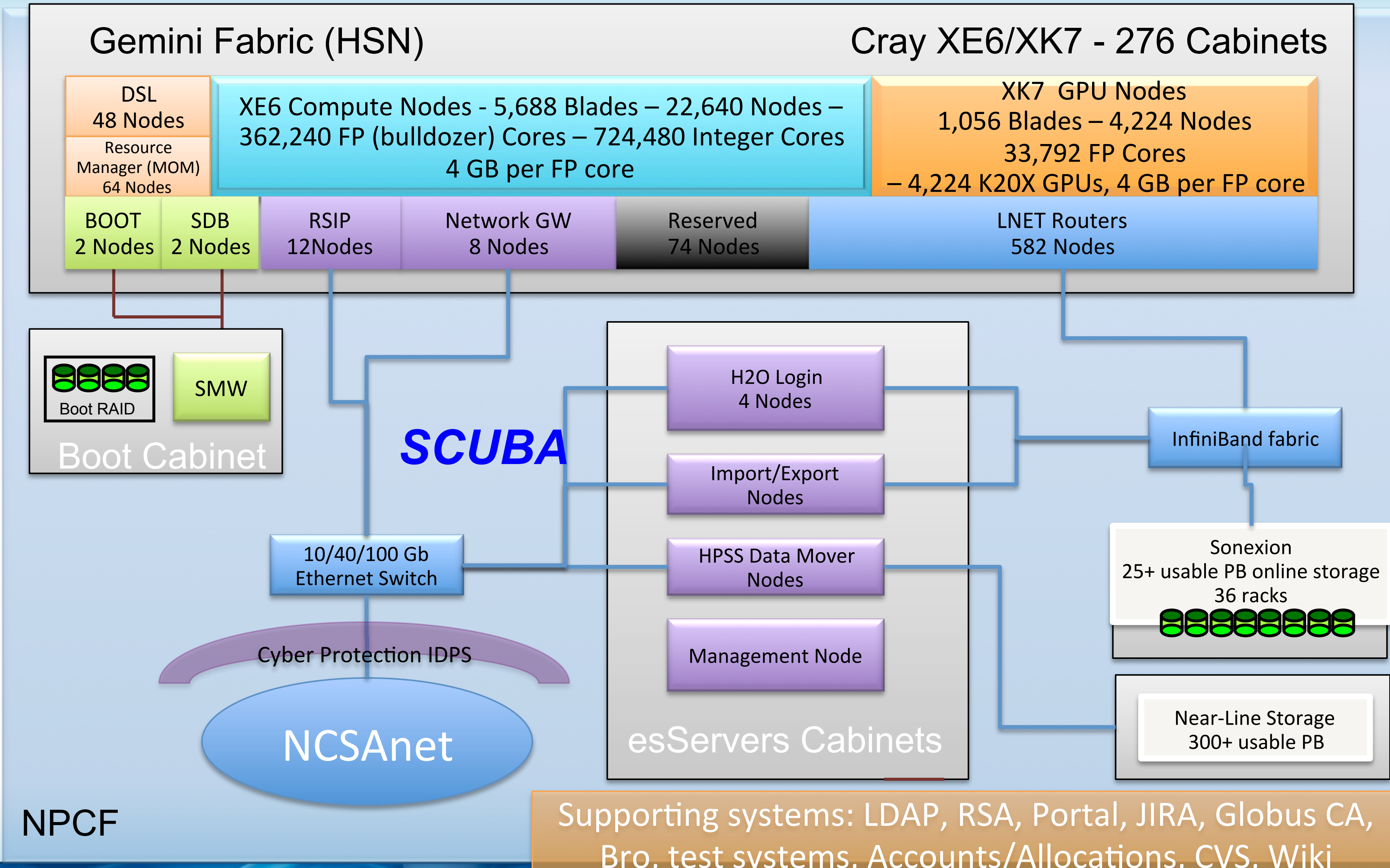
100-300 Gbps WAN

Spectra Logic: 300 usable PB

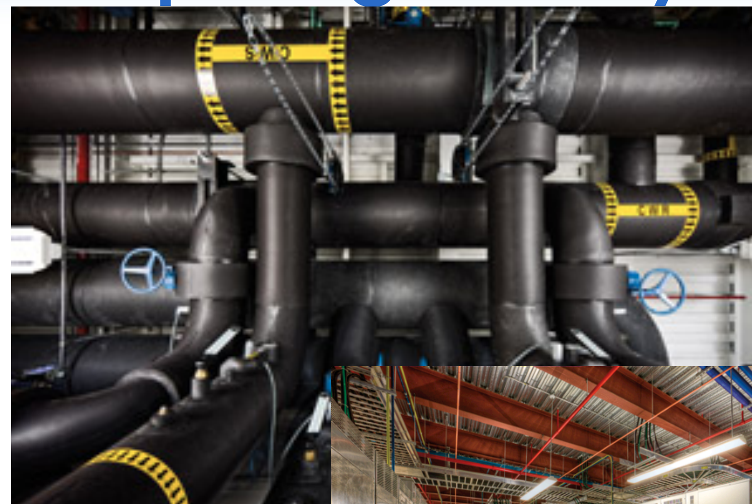
Sonexion: 26 usable PB



All storage sizes given as the amount usable. Rates are always usable/measured sustained rates


Gemini Fabric (HSN)
Cray XE6/XK7 - 276 Cabinets
DSL
48 Nodes
Resource Manager (MOM)
64 Nodes
XE6 Compute Nodes - 5,688 Blades – 22,640 Nodes – 362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core
XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores
– 4,224 K20X GPUs, 4 GB per FP core
BOOT
2 Nodes
SDB
2 Nodes
RSIP
12Nodes
Network GW
8 Nodes
Reserved
74 Nodes
LNET Routers
582 Nodes
Boot RAID
SMW
Boot Cabinet
SCUBA
H2O Login
4 Nodes
Import/Export Nodes
HPSS Data Mover Nodes
Management Node
esServers Cabinets
InfiniBand fabric
Sonexion
25+ usable PB online storage
36 racks
Near-Line Storage
300+ usable PB
10/40/100 Gb Ethernet Switch
Cyber Protection IDPS
NCSAnet
NPCF
Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA, Bro, test systems, Accounts/Allocations, CVS, Wiki

National Petascale Computing Facility



- Only Facility in the world of this scale on an Academic Campus
- Capable of sustained 24 MW today
- Expandable in space, power and cooling [50,000 ft^2 (4,645+ m^2) machine room gallery and sustained 100 MW]

- Modern Data Center
  - 90,000+ ft^2 (8,360+ m^2) total
  - 30,000 ft^2 (2,790+ m^2) raised floor
    20,000 ft^2 (1,860+ m^2) machine room gallery

- Energy Efficiency
  - LEED certified Gold
  - Power Utilization Efficiency, PUE = 1.1–1.2
  - Staff participating in Energy Efficient HPC working group and

WHAT IS SUSTAINED PERFORMANCE

SPP Is a Quantitative Method for "Sustained"

- Sustained Performance is accomplishing an amount of work in a elapsed time.
  - It is not a hardware rate
  - It is not the work needed to scale
  - It is reflection of the work needed completing meaningful problems
- SPP **Performance is proportional to runtime only**
  - Better SPP = short run time for same work
- A benchmark == **application + problem** (input set)
  - Just the application is never a test or benchmark
  - Different problems may determine a completely code path and algorithmic method

Sustained Petascale Performance (SPP)

- SPP is an instance of the Sustained System Performance (SSP) **Method** of Evaluating systems
  - Method is a process or recipe
    - A process to evaluate performance for a range of applications
  - SSP evolved over time at NERSC over multiple procurements benchmark test implements
    - Tim Pugh talked about it earlier
  - The method was formally defined, generalized and expanded at Berkeley to cover any scale, any workload, heterogeneity and any architecture
    - http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-143.pdf
  - Specifics are determined by the implementation of the method based on specific workload, systems, etc.
- SPP is the Blue Waters/NSF implementation of the SSP Method

The Sustained Petascale Performance (SPP) Metric

- Establish a set of application codes that reflect the intended work the system will do
  - Can be any number of tests as long as they have a common measure of the amount of work
- A test consists of a complete code and a problem set reflecting the science teams' intentions
- Establish the reference amount work (ops, atoms, years simulated, etc.) the problem needs to do for a fixed concurrency
- Time each test takes to execute
  - Concurrency and/or optimization can be fixed and/or varied as desired
- Determine the amount of work done for a given "schedulable unit" (node, socket, core, task, thread, interface, etc.)
  - Work = Total work (operations) /total time/number of scalable units
  - Work per unit= Total work/number of scalable units used for the test
- Composite the work per schedulable unit for all tests
  - Composite functions based on circumstances and test selection criteria
  - Can be weighed or not as desired
  - BW is using the Geometric mean – lowest of all means and reduces impact of outliers
- Determine the SPP of a system by multiplying the composite work per schedulable unit by the number of schedulable units in the system
- Determine the *Sustained Petascale Performance*

General SSP/SPP Measures Time to Solution

Per processor performance for code *i*, with test *j* on system *s*

Work Operations for code *i*, with test *j*

$$p_{s,i,j} = \frac{f_{i,j}}{m_{i,j} * t_{s,i,j}} = \frac{f_{i,j}}{m_{i,j}} * \frac{1}{t_{s,i,j}}$$

Concurrency for code *i*, with test *j*

Wall clock execution time for code *i*, with test *j* on system *s*

$$\frac{SSP_s}{SSP_{s'}} = \frac{N_s * \sum \frac{f_{i,j}}{m_{i,j}} * \frac{1}{t_{s,i,j}} \Big/ (I*J)}{N_{s'} * \sum \frac{f_{i,j}}{m_{i,j}} * \frac{1}{t_{s',i,j}} \Big/ (I*J)} = \frac{\sum \frac{f_{i,j}}{m_{i,j}} \Big/ (I*J)}{\sum \frac{f_{i,j}}{m_{i,j}} \Big/ (I*J)} * \frac{\left(\sum \frac{1}{t_{s,i,j}}\right) \Big/ (I*J)}{\left(\sum \frac{1}{t_{s',i,j}}\right) \Big/ (I*J)} = \boxed{\sum \frac{t_{s',i,j}}{t_{s,i,j}}}$$

Assume Number of Schedulable Units in Systems are equal

Number of tests

Challenges for SPP Implementation

- Representative workload
- Heterogeneous system work units
  - XE and XK nodes
- Unprecedented scale – drives unprecedented problem definition
- Determining the reference operation count
- Added Criteria to the Method
  - Runs at full scale of the SPP codes
  - Comparing same application/problem performance of XE and XK performance on a node basis

SPP Metric Definition for BW

- SPP metric is a geometric mean of per node performance rates for a suite of applications, each running in dedicated mode on a 1/5 to a 1/2 of the full number of compute nodes on the Blue Waters system, multiplied by the total number of compute nodes in the system.
- Each set of nodes of a given type is has the SPP contribution calculated independently and those sustained measures are summed to obtain the full system SPP value.
  - More precisely, for a given set of benchmark codes, the performance rate of the i-th code expressed in units of GFLOPS per node of type a, $P_{\alpha,i,}$, is calculated by dividing the reference FLOP count for that benchmark by the number of nodes of that type used to run the problem and by the total wall clock time for that run.
  - For a given number of nodes of a given type α, N_α, the contribution to the SSP from nodes of type a is the geometric mean of $P_{\alpha,i}$ over all applications, multiplied by N_α.
  - The total SSP is the sum of the contributions for each node type. For Blue Waters, α is two for the XE and XK node types. N_{XE} = 22640 and N_{xk}= 4224.
  - The number of GFLOPS per node was computed for the i-th benchmark running on the XE nodes, $P_{XE6,i}$ and the jth benchmark running on the XK nodes, $P_{XK7,j}$.
  - The contribution to the SSP for a given node type is the geometric mean of the $P_{\{XE6,XK7\},i \text{ or } j}$ values times the corresponding numbers of nodes of each type in the full system.
  - Thus, the total SSP of the XE/XK system is:
  - *SSP = Geometric Mean for all i* $(P_{XE6,i}) \times N_{XE6}$ + *Geometric Mean for all j* $(P_{XK7,j}) \times N_{XK7}$

From Method To Implementation

- Sustained Petascale Performance Metric is the Blue Waters/NSF implementation of the SSP Method
- To move from the Method to Metric
  1. Select number and instances of applications and problem sets
  2. Select Input sets that determine the code paths
  3. Establish Reference Counts
  4. Optimize (or not)
  5. Run Tests
  6. Composite
  7. Evaluate
  8. Repeat 4 thru 7 or 2 thru 7 or until complete

Determining Reference Operation Counts

- Determining the total number of reference work operations (e.g. FLOPs) required for each SPP science problem requires specifying the code version and the input problem data set.
- The GigaFLOP value used to calculate $P_{\alpha,i}$ is based on reference FLOP counts obtained using *best practices*. In order of preference, these best practices are:
  - hand-counting the floating-point operations within the code (where feasible),
  - using developer-implemented measures of the number of FLOPs executed, or
  - collecting hardware counter data collected by running the problem on Interlagos processors. When hardware performance counters are collected, the hardware counter data was compared to hand counts or developer-implemented measures (where available) for validation.
  - In order to avoid including extra FLOPs that may result from the extra operations used for scaling such as redundant computations, etc., scaling assessments were collected and compared hardware counter data obtained from multiple runs at different node counts for the same total problem size.
  - Enabled determination of whether the FLOP count for a fixed total problem size increases with the number of nodes, as well as how to eliminate any superfluous FLOPs from FLOP counts obtained at the desired scale.

SPP Method Coverage

Science Area	Struct Grids	Unstruct Grids	Dense Matrix	Sparse Matrix	N-Body/ Agent	Monte Carlo	FFT	PIC	Significant I/O
Climate and Weather	X	X		X		X			X
Plasmas/Magnetosphere	X				X		X		X
Stellar Atmospheres and Supernovae	X			X	X	X		X	X
Cosmology	X			X	X				
Combustion/Turbulence	X						X		
General Relativity	X			X					
Molecular Dynamics			X		X		X		
Quantum Chemistry			X	X	X	X			X
Material Science			X	X	X	X			
Earthquakes/Seismology	X	X			X				X
Quantum Chromo Dynamics	X		X	X	X		X		
Contagion (Social) Networks					X				
Evolution									
Engineering/System of Systems						X			
Computer Science		X	X	X			X		X

BW SPP Test Components

- SPP – is a time to solution metric that is using the planned applications on representative parts of the Science team problems
  - **Represents end to end problem run including I/O, pre and post phases, etc.**
  - Coverage for science areas, algorithmic methods, scale
- SPP Application Mix (details and method available)
  - NAMD – molecular dynamics
  - MILC, Chroma – Lattice Quantum Chromodynamics
  - VPIC, SPECFEM3D – Geophysical Science
  - WRF – Atmospheric Science
  - PPM – Astrophysics
  - NWCHEM, GAMESS – Computational Chemistry
  - QMCPACK – Materials Science
- Minimum SPP for x86 processors plus
- Kepler processors have to add at least 13% more above the x86 SPP
- At least three SPP benchmarks run at full scale

BW SPP Test Components

XE

Area	Code - version	Run Scale (XE Nodes) (Multiply by 16 or 32 to get cores)	Features
Molecular Dynamics	NAMD v2.0	5,000	C++, Charm++
Quantum Monte-Carlo	QMCPACK v52	4,800	C++/Fortran, MPI+OpenMP
Quantum Chromodynamics	MILC 7.6.3	4,116	C/C++, MPI/ pthreads
Quantum Chemistry	NWChem 6.1	5,000	C/Fortran, GA
Climate/ Weather	WRF 3.3.1	4,560	C/Fortran, MPI +OpenMP
Earthquakes/ Seismology	SpecFEM3D 5.13	5,419	F90/C++, MPI
Stellar Atmospheres and Supernovae	VPIC	4,608	Fortran/C, MPI +OpenMP
Plasmas/ Magnetosphere	PPM – 7/2/12	8,256	Fortran, MPI +OpenMP

XK

Area	Code	Run Scale	Method
Molecular Dynamics	NAMD	768	Cuda
Quantum Monte-Carlo	QMCPACK	700	Cuda
Quantum Chromodynamics	CHROMA	768	Cuda
Quantum Chemistry	GAMESS	1,536	OpenACC

- **Composite System SPP – 1.31 PF/s**
  - **x86 SPP Contribution – 1.10 PF/s**
  - **Kepler SPP Contribution – 0.21 PF/s**

Additional SPP Test Results

- Full Scale SPP XE Codes
  - In addition to the NSF Petascale tests, 4 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 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, 1 code was an OpenACC implementations

Some Other SPP Lessons

- Take all published performance projections with a large grain of salt
- Take all claims of code porting/optimizing to new architectures with a large box of salt
- Modeling applications and systems can significantly improve performance projections
- Balance run times with optimal performance
  - need to have ability to do tuning and improvement

Example for SPP
Blue Waters & Titan Computing Systems

System Attribute (2012)	**UIUC/NCSA** ***Blue Waters***	**DOE/ORNL** ***Titan***
Vendor(s)	Cray/AMD/NVIDIA	Cray/AMD/NVIDIA
Processors	Interlagos 2.3 GHz/Kepler K20X	Interlagos 2.1 GHz /Kepler K20X
Total Peak Performance (PF/s)	*13.1*	*27.11*
Total Peak Performance (CPU/GPU)	7.6/5.5	2.63/24.5
Number of Nodes	27,648	19,200
Number of CPU Modules (8 cores/Module)	49,504	18,688
Number of GPU Chips	4,224	18,688
SPP Sustained Performance (PF/s)	**1.31**	**0.64**
Amount of CPU Memory (TB)	1,660	710
Interconnect	Gemini 3-D Torus	Gemini 3-D Torus
Dimensions	24x24x24	25x16x24
Amount of Usable On-line Disk Storage (PB)	26	>10
2013 upgrade		~40 shared
Sustained Disk Transfer (TB/sec)	1.2	0.245
2013 upgrade		~1 shared
Amount of Near-line/Archival Storage (Usable/Maximum) (PB)	300/400	125/250
2013 upgrade		150/300
Protection from single point of tape failure	Yes	No
Sustained Tape Transfer (GB/sec)	88	18

Information compiled by: Thomas Dunning and Bill Kramer, UIUC/NCSA and Buddy Bland ORNL

MAJOR SCIENCE AND ENGINEERING TEAMS AND APPLICATIONS

NSF Major PRAC S&E Teams

- Super instruction architecture for petascale computing. - Rodney J. Bartlett, University of Florida
- Petascale Quantum Simulations of Nano Systems and Biomolecules - Jerzy Bernholc, North Carolina State University at Raleigh
- Collaborative Research: Simulation of Contagion on Very Large Social Networks with Blue Waters - Keith Bisset, Virginia Polytechnic Institute and State University
- Computational Relativity and Gravitation at Petascale: Simulating and Visualizing Astrophysically Realistic Compact Binaries - Manuela Campanelli, Rochester Institute of Technology
- Hierarchical molecular dynamics sampling for assessing pathways and free energies of RNA catalysis, ligand binding, and conformational change - Thomas Cheatham, University of Utah
- Petascale Cosmology with Gadget: Modeling the Formation of the First Quasars with Blue Waters - Tiziana Di Matteo, Carnegie Mellon University
- From Binary Systems and Stellar Core Collapse To Gamma-Ray Bursts - Peter Diener, Louisiana State University
- Direct Numerical Simulation of Fully Resolved Vaporizing Droplets in a Turbulent Flow - Said Elghobashi, University of California-Irvine
- System Software for Scalable Applications - Bill Gropp, University of Illinois at Urbana-Champaign
- Petascale Research in Earthquake System Science on Blue Waters (PressOn) - Thomas Jordan, University of Southern California
- Enabling Breakthrough Kinetic Simulations of the Magnetosphere via Petascale Computing - Homayoun Karimabadi, University of California-San Diego
- Accelerating Nano-scale Transistor Innovation - Gerhard Klimeck, Purdue University
- Computational Chemistry at the Petascale - Monica Lamm, Iowa State University
- Petascale plasma physics simulations using PIC codes - Warren Mori, University of California-Los Angeles
- Peta-Cosmology: galaxy formation and virtual astronomy - Kentaro Nagamine, University of Nevada-Las Vegas
- Formation of the First Galaxies: Predictions for the Next Generation of Observatories - Brian O'Shea, Michigan State University
- Simulating vesicle fusion on Blue Waters - Vijay Pande, Stanford University

Seehttps://bluewaters.ncsa.illinois.edu/science-teams for full summaries

NSF Major PRAC S&E Teams

- Modeling Heliophysics and Astrophysics Phenomena with a Multi-Scale Fluid-Kinetic Simulation Suite - Nikolai Pogorelov, University of Alabama, Huntsville
- Evolution of the Small Galaxy Population From High Redshift to the Present - Thomas Quinn, University of Washington
- Collaborative Research: Petascale Design and Management of Satellite Assets to Advance Space Based Earth Science - Patrick Reed, Pennsylvania State University; Eric Wood, Princeton University
- The Computational Microscope - Klaus Schulten, University of Illinois at Urbana-Champaign
- Collaborative Research: Testing Hypotheses about Climate Prediction at Unprecedented Resolutions Using the Blue Waters System - Cristiana Stan, Center for Ocean-Land-Atmosphere Studies; David Randall, Colorado State University
- Ab Initio Models of Solar Activity - Robert F. Stein, Michigan State University
- Lattice QCD on Blue Waters - Robert Sugar, University of California-Santa Barbara
- Evolution of intricate multi-scale biological systems - Ilias Tagkopoulos, University of California-Davis
- Petascale Multiscale Simulations of Biomolecular Systems - Gregory Voth, University of Chicago
- Enabling Large-Scale, High-Resolution, and Real-Time Earthquake Simulations on Petascale Parallel Computers - Liqiang Wang, University of Wyoming
- Understanding Tornadoes and Their Parent Supercells Through Ultra-High Resolution Simulation/Analysis - Robert Wilhelmson, University of Illinois at Urbana-Champaign
- Petascale Simulation of Turbulent Stellar Hydrodynamics - Paul R. Woodward, University of Minnesota
- Type Ia Supernovae - Stan Woosley, University of California-Santa Cruz
- Using Petascale Computing Capabilities to Address Climate Change Uncertainties - Donald J. Wuebbles, University of Illinois at Urbana-Champaign
- Petascale Computations for Complex Turbulent Flows at High Reynolds Number - Pui-kuen Yeung, Georgia Institute of Technology
- Breakthrough Petascale Quantum Monte Carlo Calculations - Shiwei Zhang, College of William and Mary

GLCPC S&E Teams

- Scaling the Effects of Intermediate Disturbance and Changes to Small-Scale Ecosystem Structure on Ecosystem, Hydrology, Lake and Weather Interactions to the Scale of the Great-Lakes Region - Gil Bohrer, Ohio State University
- Rotational Turbulence in Magnetized Cylindrical Couette Flow - Fausto Cattaneo, University of Chicago
- Development of New Parameters for SEMO Methods for Transition Metals and Thorium - David Dixon, University of Alabama
- Next-Generation Ab Initio Symmetry-Adapted No-Core Shell Model and Its Impact on Nucleosynthesis - Jerry Draayer, Louisiana State University
- Policy Responses to Climate Change in a Dynamic Stochastic Economy - Lars Hansen, University of Chicago
- Atomistic Modeling of Real World Nanoelectronics Devices - Gerhard Klimeck, Purdue University
- Modeling of Peptide-mediated Ribosome Stalling with Antibiotics - Alexander Mankin, University of Illinois at Chicago
- The Mechanism of the Sarco/Endoplasmic Reticulum ATP-Driven Calcium Pump - Benoit Roux, University of Chicago
- Re-designing Communication and Work Distribution in Scientific Applications for Extreme-scale Heterogeneous Systems - Karen Tomko, Ohio State University
- Implicitly-Parallel Functional Dataflow for Productive Hybrid Programming on Blue Waters - Michael Wilde, University of Chicago

Illinois S&E Teams

- Molecular Mechanism of DNA Exchange - Aleksei Aksimentiev, University of Illinois at Urbana-Champaign
- Quantum Monte Carlo Calculations of Water-Graphene and Water-h-BN Interfaces - Narayana Aluru, University of Illinois at Urbana-Champaign
- Extreme Scale Astronomical Image Composition and Analysis - Robert J. Brunner, University of Illinois at Urbana-Champaign
- The Dynamics of Protein Disorder and its Evolution - Gustavo Caetano-Anolles, University of Illinois at Urbana-Champaign
- Simulations of Cellulosomal Subunits: Components of a Molecular Machinery for Depolymerization of Feedstock for Production of Second Generation Biofuels - Isaac Cann, University of Illinois at Urbana-Champaign
- New Advances in Cloud Modeling: How 3D Radiation Impacts Cloud Dynamics and Properties - Larry Di Girolamo, University of Illinois at Urbana-Champaign
- Semileptonic Kaon Decay Form Factors at the Physical Point - Aida X El-Khadra, 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
- Efficient Scalable Climate Simulations in an Earth System Model via an Adaptive Parallel Runtime System - Atul Jain, University of Illinois at Urbana-Champaign
- Accurate Sequence Alignment Using Distributed Filtering on GPU Clusters - C. Victor Jongeneel, University of Illinois at Urbana-Champaign
- Radio Interferometric Imaging in the Petascale Era: New Opportunities and Challenges - Athol Kemball, University of Illinois at Urbana-Champaign
- 4-D dyanmic evolution of North American continent - Lijun Liu, University of Illinois at Urbana-Champaign
- 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
- Quantum-Classical Path Integral Simulation of Proton and Electron Transfer - Nancy Makri, University of Illinois at Urbana-Champaign
- Variational Multiscale Methods for Non-Newtonian Viscoelastic Blood Flow Modeling: Application to Clot Formation and Dissolution in Patient Specific Models - Arif Masud, University of Illinois at Urbana-Champaign
- Calculations of Single-Scattering Properties of Randomly Oriented Small Atmospheric Ice Crystals to Improve Representations of Ice Clouds in Satellite Retrieval Algorithms and Numerical Models - Greg McFarquhar, University of Illinois at Urbana-Champaign
- Non-Adiabatic Electron-Ion Dynamics and Electronic Stopping - Andre Schleife, University of Illinois at Urbana-Champaign
- Gravitational and Electromagnetic Signatures of Compact Binary Mergers: General Relativistic Simulations at the Petascale - Stuart Shapiro, University of Illinois at Urbana-Champaign
- Parallel Programming Language and Library Research on Multicore Clusters - Marc Snir, University of Illinois at Urbana-Champaign
- Scaling up of a Highly Parallel LBM-based Simulation tool (PRATHAM) for Meso- as well as Large-Scale Laminar and Turbulent Flow and Heat Transfer - Rizwan Uddin, University of Illinois at Urbana-Champaign
- Exploring the Physics of Geological Sequestration of Carbon Dioxide using High-Resolution Pore-Scale Simulation - Albert J. Valocchi, University of Illinois at Urbana-Champaign
- An Extreme-Scale Computational Approach to Redistricting Optimization - Shaowen Wang, University of Illinois at Urbana-Champaign

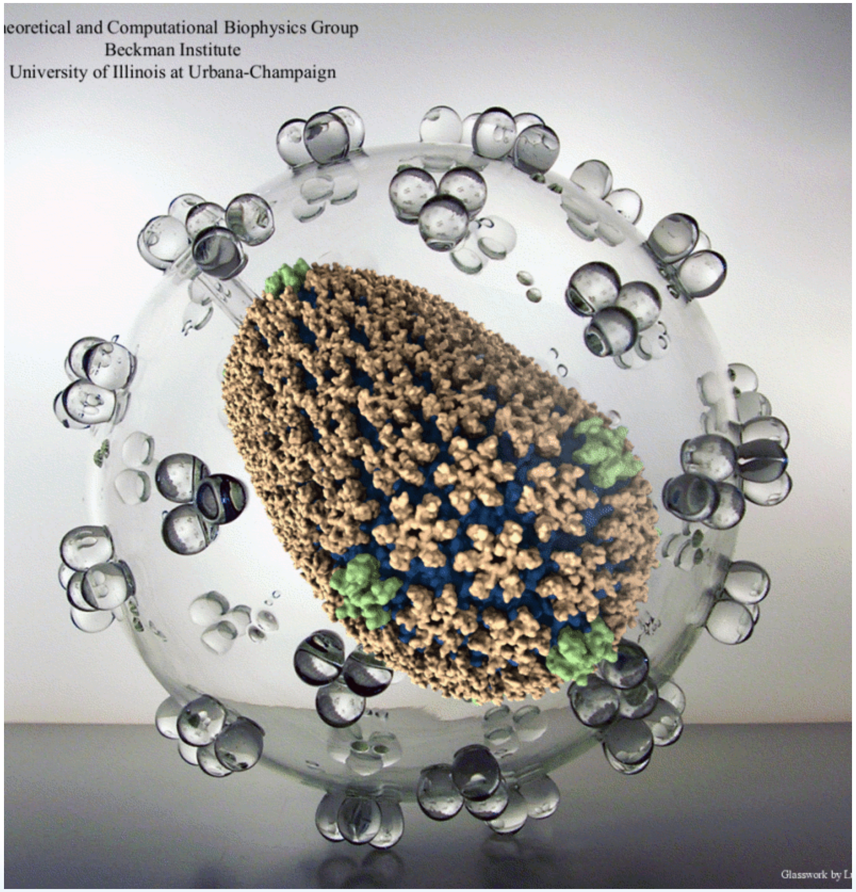
S&E Teams by Institution

- Based on PI at the moment.

Institution	Usage (node*hours)	Usage (core*hours)
Cornell University	7,400	118,406
Georgia Tech University	303,766	4,860,256
Institute of Global Environment and Society	2,171	34,734
Iowa State University	191,915	3,070,643
Lawrence Berkeley National Laboratory	844,636	13,514,171
Louisiana State University	1,291,828	20,669,240
Michigan State University	445,322	7,125,152
North Carolina State University	554,436	8,870,973
Ohio State University	24	381
Purdue University	391,925	6,270,802
Rochester Institute of Technology	229,352	3,669,630
Stanford University	123,219	1,971,506
SURA	2,244	35,896
University of Alabama, Huntsville	169,109	2,705,744
University of California Observatories	683,979	10,943,667
University of California, Davis	910	14,555
University of California, Irvine	163,090	2,609,438
University of California, Los Angeles	1,279,039	20,464,619
University of California, San Diego	634,325	10,149,195
University of California, Santa Barbara	8,762,856	140,205,688
University of Chicago	740,252	11,844,032
University of Florida	179,468	2,871,480
University of Illinois, Urbana-Champaign	4,867,028	77,872,445
University of Illinois, Chicago	0.10	2
University of Minnesota	115,041	1,840,659
University of Nevada, Las Vegas	2,667	42,674
University of Southern California	1,545,326	24,725,208
University of Toronto	-	-
University of Utah	1,696,042	27,136,674
University of Washington	288,893	4,622,280
University of Wyoming	3,154	50,464
Virginia Tech	585,772	9,372,347
Grand Total	26,105,185	417,682,962

Science Area	Number of Teams	Codes	Struct Grids	Unstruct Grids	Dense Matrix	Sparse Matrix	N-Body	Monte Carlo	FFT	PIC	Significant I/O
Climate and Weather	3	CESM, GCRM, CM1/ WRF, HOMME	X	X		X		X			X
Plasmas/Magnetosphere	2	H3D(M),VPIC, OSIRIS, Magtail/ UPIC	X				X		X		X
Stellar Atmospheres and Supernovae	5	PPM, MAESTRO, CASTRO, SEDONA, ChaNGa, MS-FLUKSS	X			X	X	X		X	X
Cosmology	2	Enzo, pGADGET	X			X	X				
Combustion/Turbulence	2	PSDNS, DISTUF	X						X		
General Relativity	2	Cactus, Harm3D, LazEV	X			X					
Molecular Dynamics	4	AMBER, Gromacs, NAMD, LAMMPS			X		X		X		
Quantum Chemistry	2	SIAL, GAMESS, NWChem			X	X	X	X			X
Material Science	3	NEMOS, OMEN, GW, QMCPACK			X	X	X	X			
Earthquakes/Seismology	2	AWP-ODC, HERCULES, PLSQR, SPECFEM3D	X	X			X				X
Quantum Chromo Dynamics	1	Chroma, MILC, USQCD	X		X	X	X		X		
Social Networks	1	EPISIMDEMICS									
Evolution	1	Eve									
Engineering/System of Systems	1	GRIPS,Revisit						X			
Computer Science	1			X	X	X			X		X

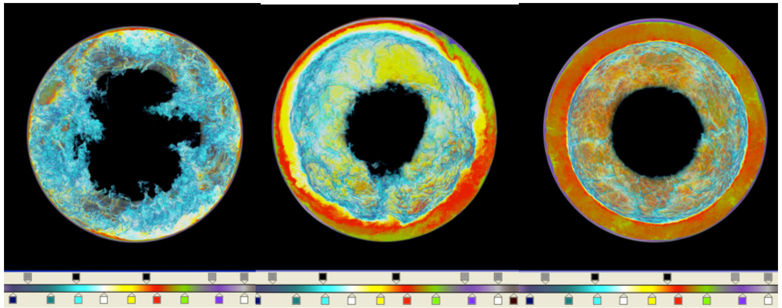
First Unprecedented Result – Computational Microscope

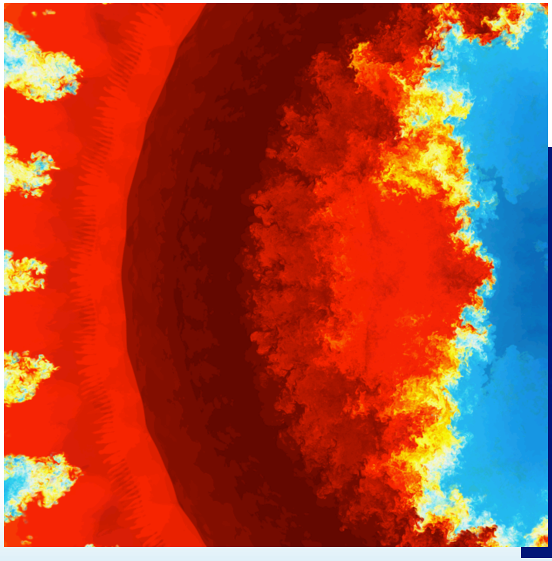


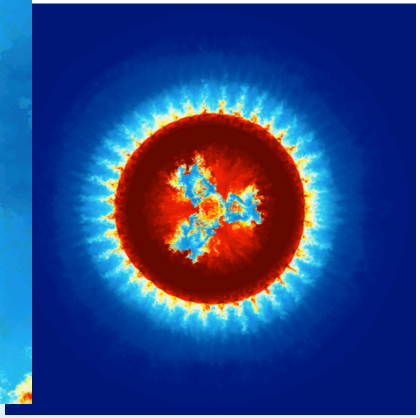
- Klaus Schulten (PI) and the NAMD group - Code NAMD/ Charm++
- Completed the highest resolution study of the mechanism of HIV cellular infection.
- May 30, 2013 Cover of *Nature*
- Orders of magnitude increase in number of atoms – resolution at about 1 angstrom

Petascale Simulation of Turbulent Stellar Hydrodynamics

- Paul Woodward PI – Code PPM
  - 1.5 Pflop/s sustained on Blue Waters
  - 10,5603 grid
  - A Trillion Cell, Multifluid CFD Simulation
  - 21,962 XE nodes; 702,784 interger cores; 1331 I/Os; 11 MW
  - All message passing and all I/O overlapped w. comput.
  - 12% theoretical peak performance sustained 41 hrs
  - 1.02 PB data written and archived; 16.5 TB per dump.
  - Ran over 12 days in 6-hour increments

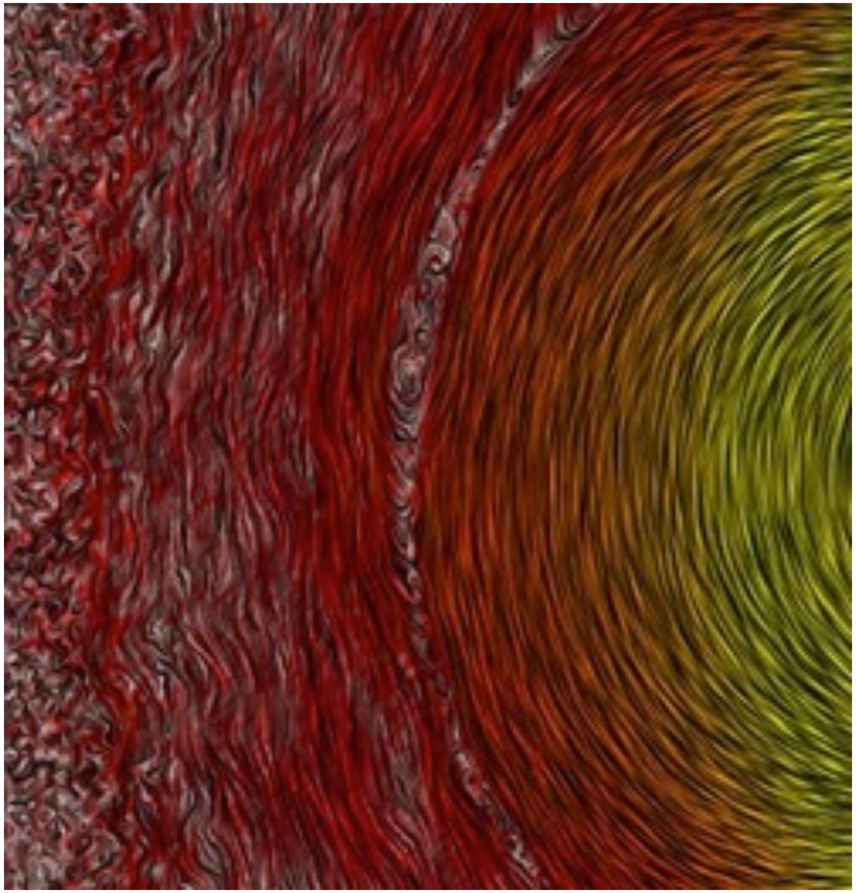






Enabling Breakthrough Kinetic Simulations of the Magnetosphere via Petascale Computing

- Homa Karimabadi PI – Code PPM
  - Possible extreme solar storms could significantly disrupt many modern infrastructure systems
  - This project studies the initiation and transmission of the solar wind



GEOScience PRAC Projects

- Understanding Tornadoes and Their Parent Supercells Through Ultra-High Resolution Simulation/Analysis
  - Robert Wilhelmson
- Using Petascale Computing Capabilities to Address Climate Change Uncertainties
  - Don Wuebbles
- Collaborative Research: Testing Hypotheses about Climate Prediction at Unprecedented Resolutions Using the Blue Waters System
  - Cristiana Stan
- Collaborative Research: Petascale Design and Management of Satellite Assets to Advance Space Based Earth Science
  - Patrick Reed
- Petascale Research in Earthquake System Science on Blue Waters (PressOn)
  - Thomas Jordan
- Enabling Large-Scale, High-Resolution, and Real-Time Earthquake Simulations on Petascale Parallel Computers
  - Liqiang Wang

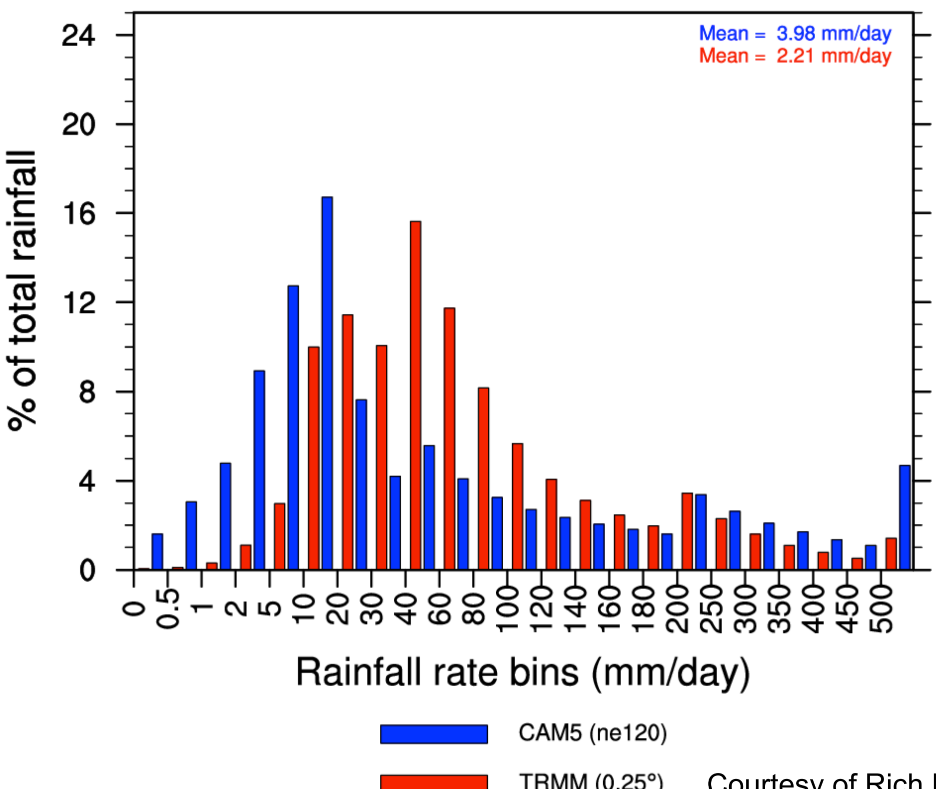
Using Petascale Computing Capabilities to Address Climate Change Uncertainties Goals

- The purpose to address key uncertainties associated with the numerical modeling of the Earth's climate system and the ability to accurately analyze past and projected future changes in climate.
  - A unique aspect of the project is coupling of CESM with the newly built Cloud-Aerosol-Radiation (CAR) ensemble modeling system.
  - Model development which is complementary to the main NCAR/CESM effort
- The specific scientific objectives are:
  - evaluate the model with different model dynamical cores and the effects of going to much higher horizontal and vertical resolution.
  - very high resolution (10-30 km horizontal resolution) in coupled climate models is motivated evidence that increased resolution leads to better simulations both of transient small-scale activity, e.g. eddies and waves, and of large-scale features of the mean climate.
  - New model dynamical cores allow for enhanced resolution as well as the use of the extensive number of the petascale capabilities (i.e., the cores scale to tens-of-thousands of processors).
- Examine uncertainties associated with the representations of processes and interactions occurring between clouds, aerosols, and radiative transfer in these models and how these uncertainties influence the climate sensitivity (see slides 2 and 3).

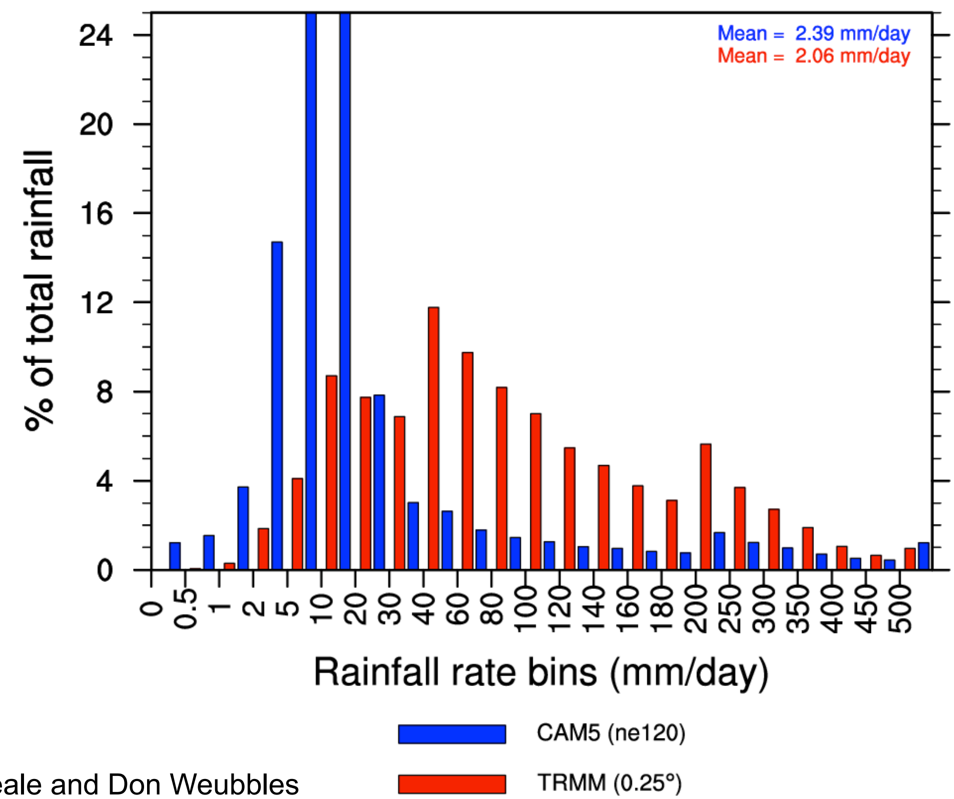
Current Climate Models: Bias in Rainfall Frequency

Common bias for many regions and most/all models: Too much light rainfall, not enough heavy

East Pacific (JFM 2002)

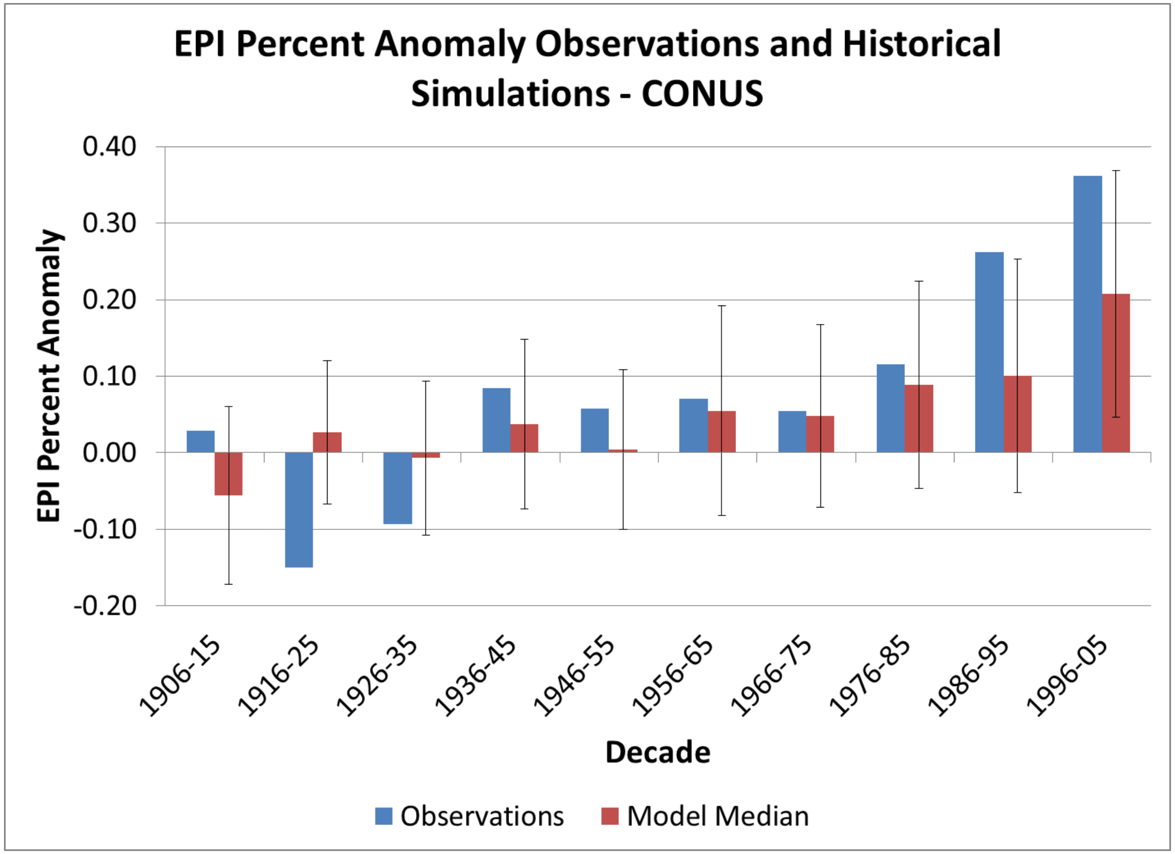


US Great plains (JJA 2002)



Courtesy of Rich Neale and Don Weubbles

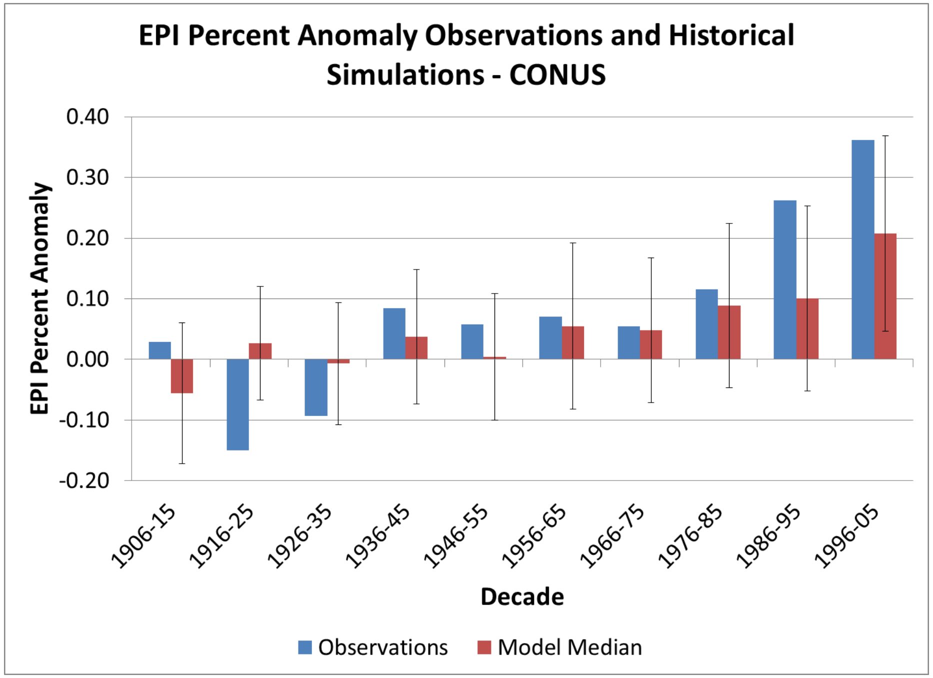
Current Climate Models: Underestimate Trends in Severe Precipitation



- 2 day duration 5 year return
  - EPI calculated annually for 1901-2005,
  - Decadal averages calculated for 1906-2005
- CMIP5: Positive trend in observed EPI anomalies over the past 4 decades
- Multi-model median of CMIP5 simulations shows an increasing trend in EPI anomalies over last 4 decades
  - Smaller than observed
  - Standard deviation between models large

CESM On Blue Waters: Allows Higher Resolution Analyses

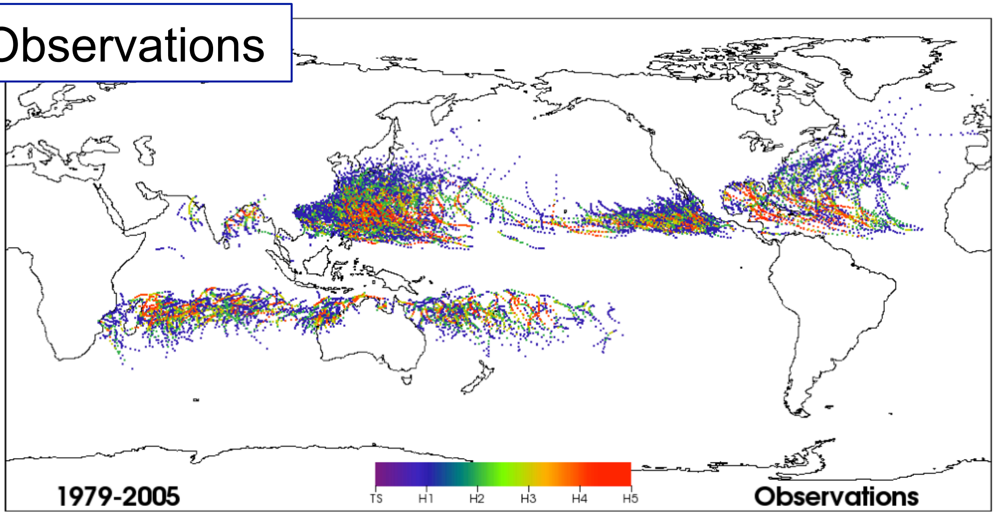
- UIUC/NCAR project with NSF to run CESM1 at 0.25° (~25 km) resolution
- 100 years in past and 100 years future
- Multiple realizations (ensembles)
- Also will be doing uncertainty analyses to enhance understanding of radiative-cloud-aerosol interactions



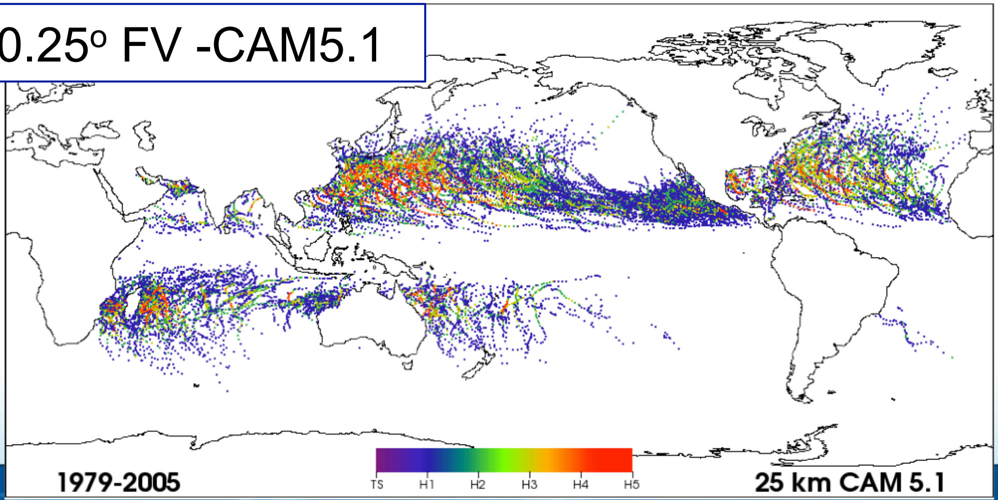
Courtesy of Don Weubbles and colleagues

High Resolution Atmospheric Component of CESM: Projected Changes in Tropical Cyclones

Observations



0.25° FV -CAM5.1



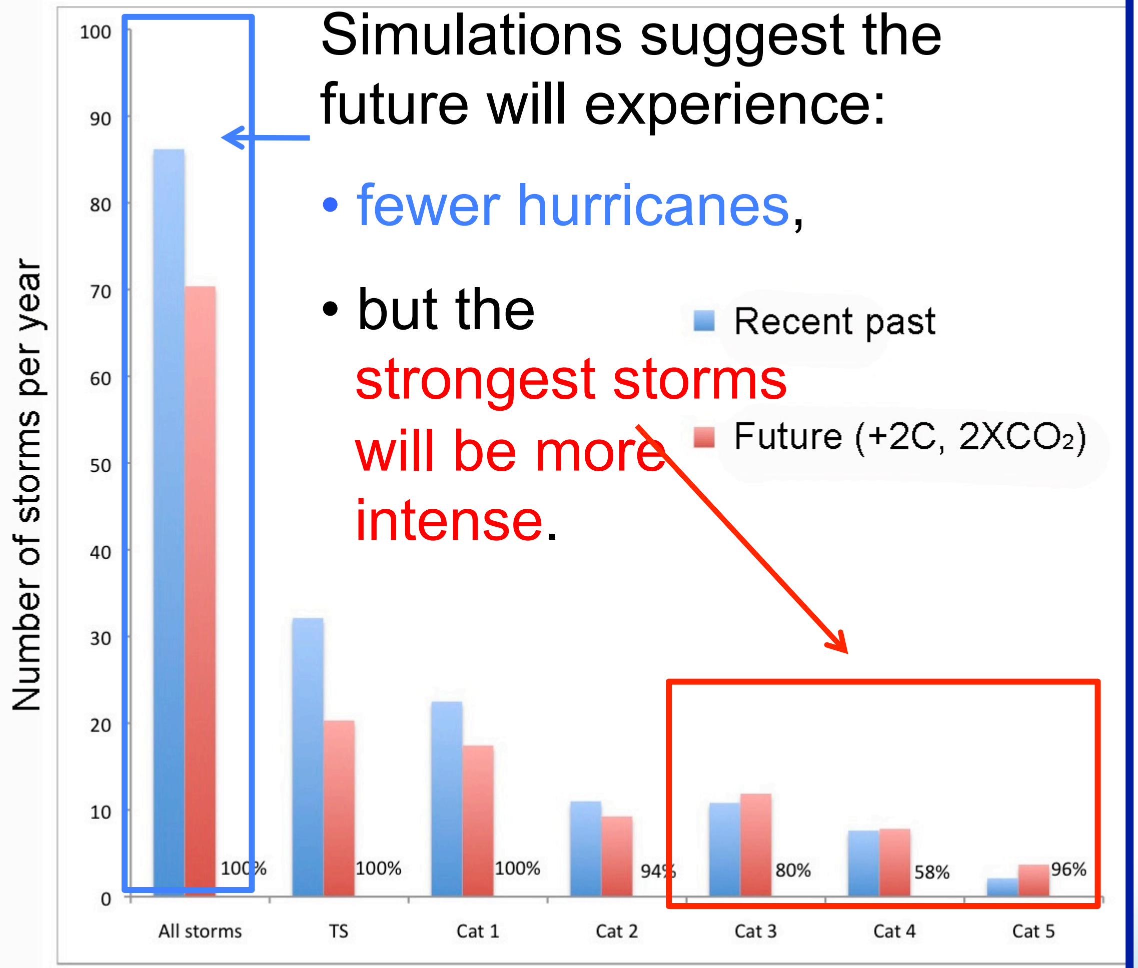
High resolution (0.25°) atmosphere simulations produce an excellent global hurricane climatology

Courtesy of
Michael Wehner, LBNL

High Resolution Atmospheric Component of CESM: Projected Changes in Tropical Cyclones

Simulations suggest the future will experience:

- fewer hurricanes,
- but the strongest storms will be more intense.



High resolution (0.25°) atmosphere simulations produce an excellent global hurricane climatology

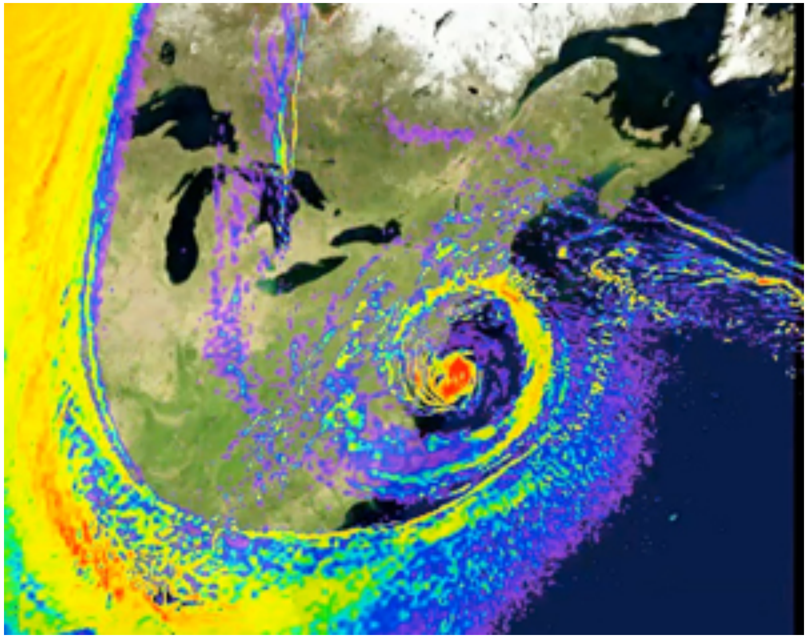
Courtesy of
Michael Wehner, LBNL

Why Blue Waters

- The specific resource needs are:
  - allows for rapid development of CESM parameterizations at unprecedented high resolution to meet the objectives.
  - Access to a near-dedicated computing environment is the key to creating a suitable and valid high resolution version of CESM within a reasonable amount of time. Tuning of CESM with new, high-resolution, representations of the processes described above requires multiple, decade-to-century, ensemble simulations of CESM (slide 4). At significant cost (order of 20M core-hours per simulated century), the tuning effort requires an effort and resource where the end product becomes the mainstream tool for future work. Thus, the return on this investment is extremely important to the next-generation climate modeling efforts.
  - A computing, data analysis, and data storage environment which allows unobstructed (online) access to data for analysis by a small number of team members. During the tuning process, a century of data at the target resolution requires on the order of 100 TB which must be readily accessible in order for intercomparison of experiments of similar size.
- Advantages of Bluewaters
  - A petascale environment supplements the CESM core effort by ability to perform application development requiring tens-of-millions of hours of compute resource which they otherwise would not have access. This feature alone represents a climate change research arena which is beyond business-as-usual.
  - Rapid turnaround afforded by (a) the ability to scale CESM to tens-of-thousands of processors on the Cray architecture, and (b) a system access design which allows priority use of large resources.
  - A computing environment within which large datasets can be easily accessed, manipulated, and analyzed with open source tools over wide-area high speed networks, and is supported by an experienced and knowledgeable staff.

Hi Resolution Study of Hurricane Sandy

- Computer forecast models for Sandy did not reach strong agreement on Sandy's unusual track until several days before landfall.
    - One factor hindering prediction models is most track the atmosphere at points separated by 10 km (6 miles) or more.
    - Too coarse to directly simulate the showers and thunderstorms that help drive hurricane behavior. Increased computer power now supports some models with resolution as fine as 3 km (2 mi).
- Blue Waters used to complete one of the most detailed simulations of a weather event ever conducted
- Utilized ARW-WRF
- Simulated a large region, from eastern North America to the western North Atlantic, at a horizontal resolution of 500 meters (1,640 feet), with 150 vertical levels.
- Tracked every second over a four-day period at more than 4 billion points.
- The results capture key aspects of Sandy in detail, including the intensification of winds just before landfall and the bands of heavy snowfall produced by Sandy across the Appalachians.
- Work done by a team from NCAR, NCSA and Cray based on the performance optimization done for WRF during Blue Waters Acceptance
- Paper being presented at SC 13



Steps to Scale WRF

- Analysis
  - Seemingly minor I/O quickly became an impediment to rapid testing – each MPI rank was writing redundant diagnostic information; had to make minor source code mods. Very problematic to "ls" or "rm" 100,000 files…
  - Hybrid MPI/OpenMP code – optimal balance of communication and cache – experimented with MPI task layout and number of OMP tiles
  - Additional compiler options do not help WRF substantially
    - Cray compiler has fairly aggressive defaults; -lfast_mv (module load libfast) only a few %
  - Nearest neighbor 2-D communication pattern –
    - Benefits from grid_order utility (setenv MPICH_RANK_REORDER_METHOD 3) by up to 20%
- The WRF version 3.3.1 modified for concerns for I/O burden per MPI task.
- WRF is a hybrid MPI/OpenMP code; it has been empirically determined to give better performance if nproc_x << nproc_y, because this leads to longer vectors on the inner compute loops.
- WRF allows for several types of parallel I/O, including use of the parallel netCDF library, and multi-file (one file per MPI task). We used the former for output, and the latter for input. Used Lustre striping over 128 OSTs.
- Used WRF's auxiliary history output options to select only the output fields of greatest interest, thus reducing the volume of output considerably.
- **Cray-specific topology-aware task placement tools which we found to benefit WRF: rank reordering had greatest impact.**
- Although weather simulations typically exhibit some load imbalance, with this hurricane simulation we observed little load imbalance because of the large rain and extensive cloud shield.

WRF Scale Insights

- At larger scale (>10,000 cores), we did see periodic increases of up to 50% wall clock time in regular, periodic groups of integration steps. Employed various jitter-reducing methodologies (e.g. core specialization) to no benefit;
  - Concluded that the regular (every 75s of wallclock time) spikes in step times were most likely due to Lustre *ping* effect.
- Overall, we noted only a little over 1% performance increase between batch and dedicated runs. This indicates that sharing links of the torus with other running jobs had minimal impact on performance.
- Used the Cray *grid_order* perl script to generate improved placement of the ranks for the nearest neighbor halo exchanges. Reducing the number of neighbors communicating off-node is the primary goal.
- Most effective way to run WRF on the AMD Bulldozer core-modules was to use MPI/OpenMP hybrid mode with 2 OpenMP threads per MPI rank. This puts 16 MPI ranks on each XE6 node.
- Using an alternate placement via grid ordering allows us to get 3 communication partners for most MPI ranks on the same node. At very high scales, this strategy improves overall WRF performance by 18% or more.
- Optimized placement also has the benefit of sending smaller east-west direction exchanges off-node and keeping as many larger north-south messages on-node as possible – 75% fewer bytes are sent over the network.
- 13,680 node (437,760 integer cores) run assuming 42 halo exchanges occurring for each WRF integration step. Over 12 million off-node halo exchange messages totaling 280 Gbytes every WRF time step.

EARLY USAGE INFORMATION

Early Science – Jan 2012 to May 2012

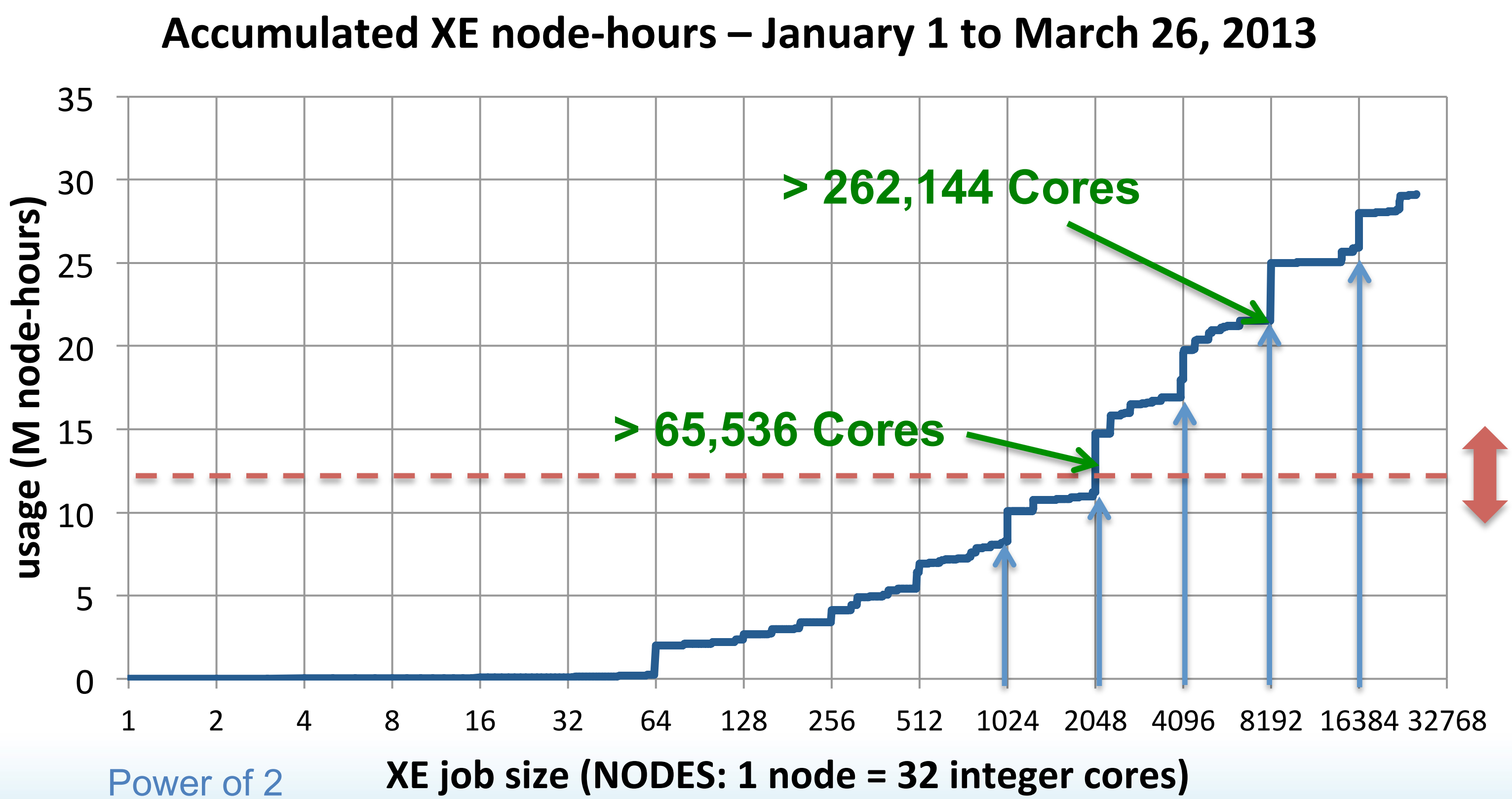
Friendly User – December 21, 2012 to April 1, 2013 and August 15, 2013 to August 31, 2013

Interim Full Service – April 2, 2013 to July 14, 2013

Full Service – September 1, 2013 onward

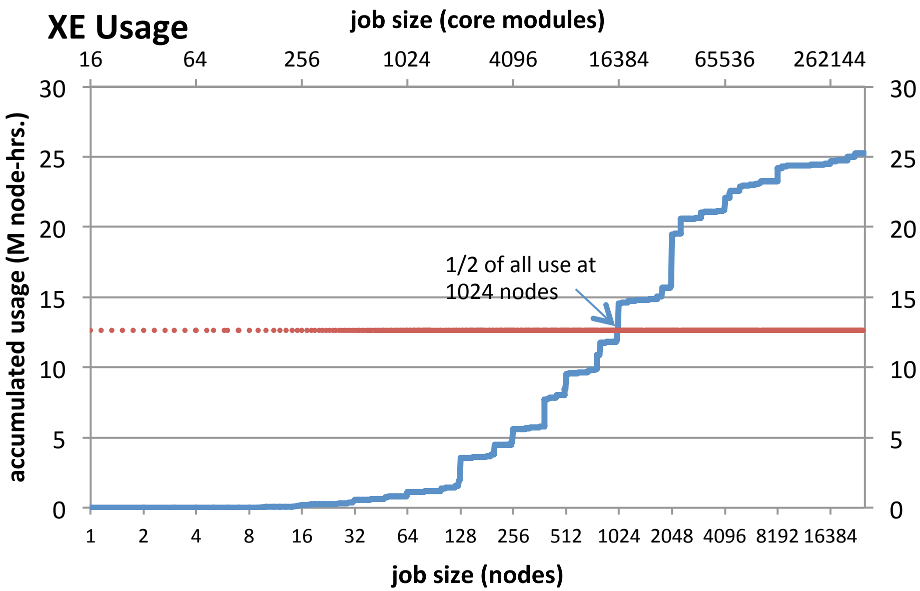
Usage Breakdown – Jan 1 to Mar 26, 2013

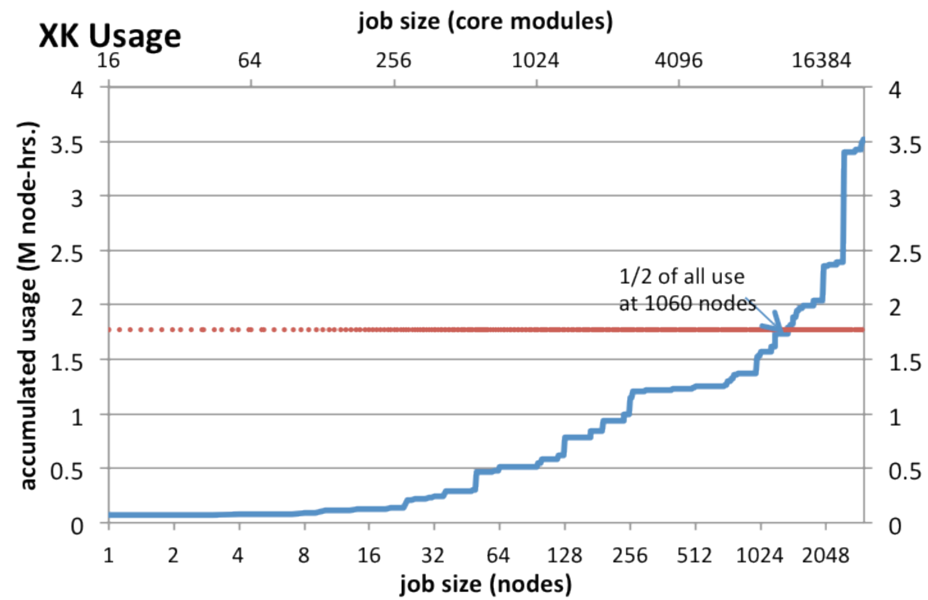
- Torque log accounting (NCSA, Mike Showerman)



Power of 2

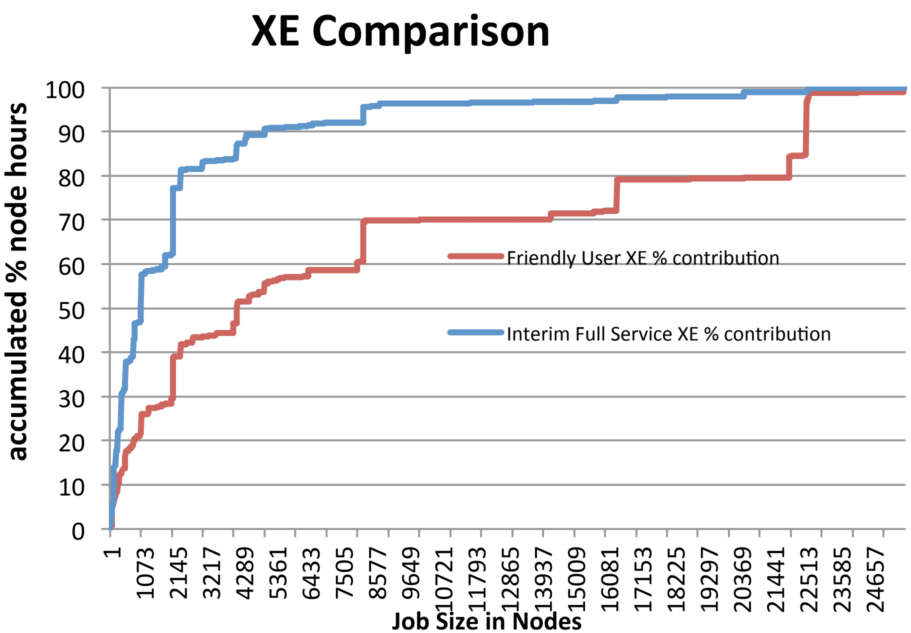
Interim Full Service Usage

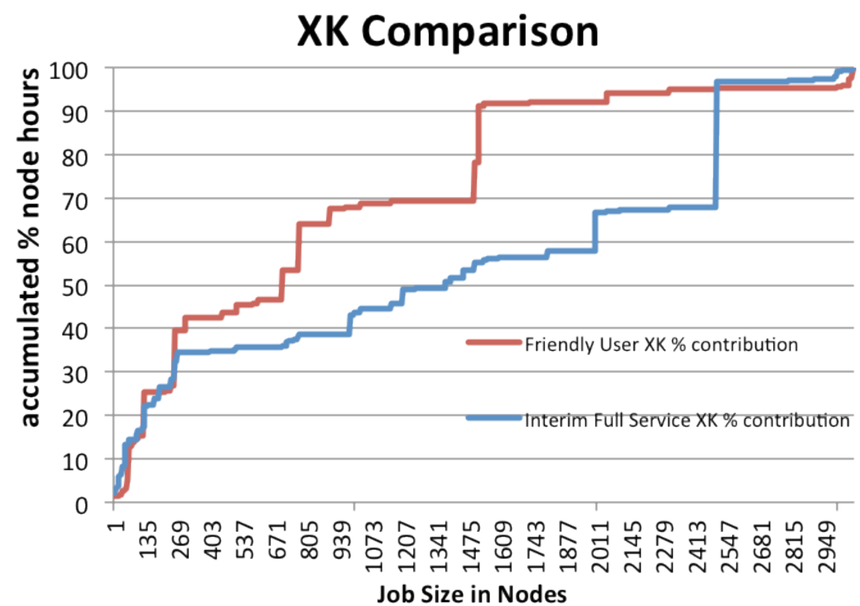




- Half the XE usage for jobs with 16K core-modules.
  - Contributions at 32K and 128K core-modules.
- Half the XK usage for jobs with 1,060 GPUs (nodes).
  - Note contribution from jobs at 2,500 nodes.

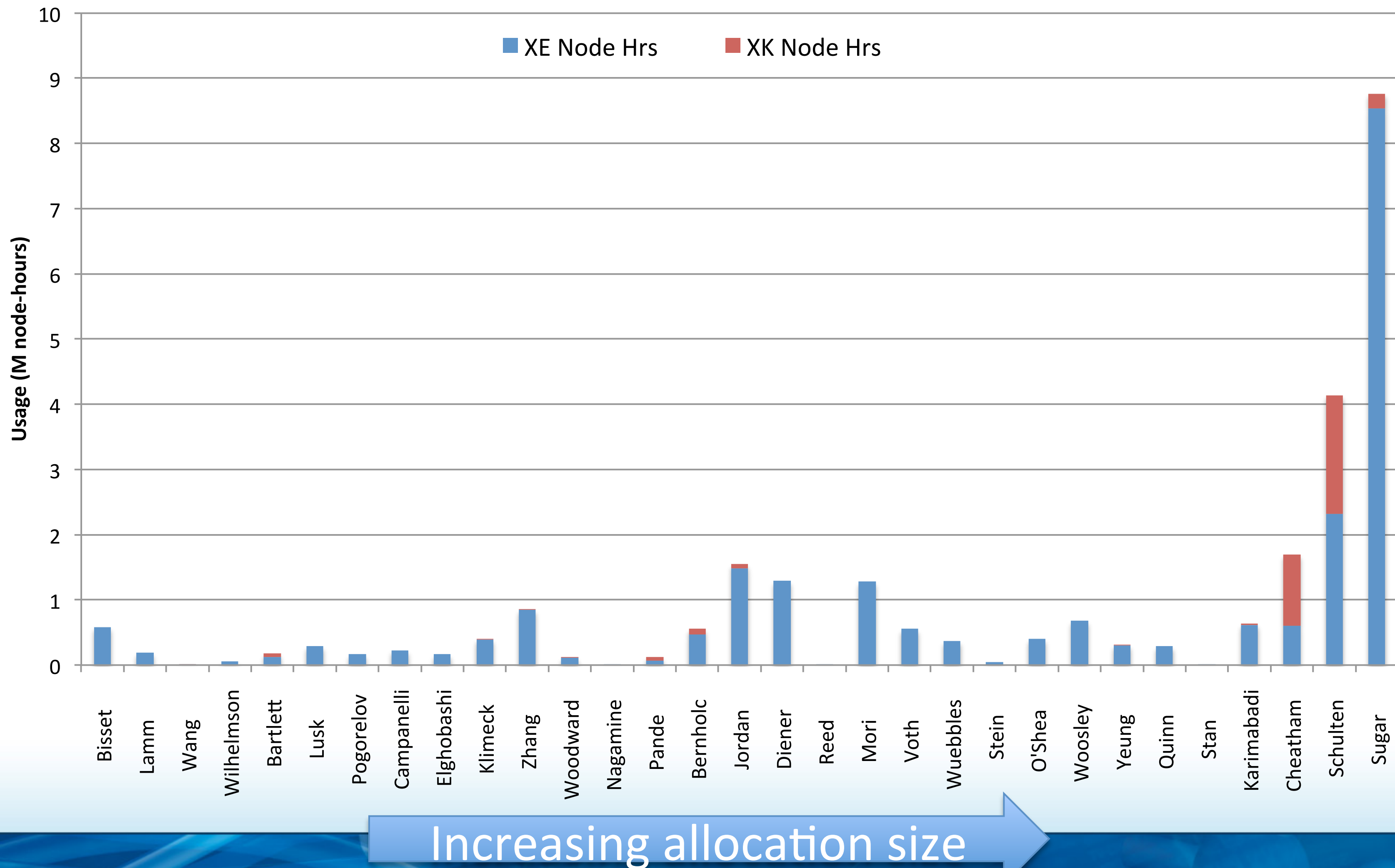
Normalized Usage Comparison





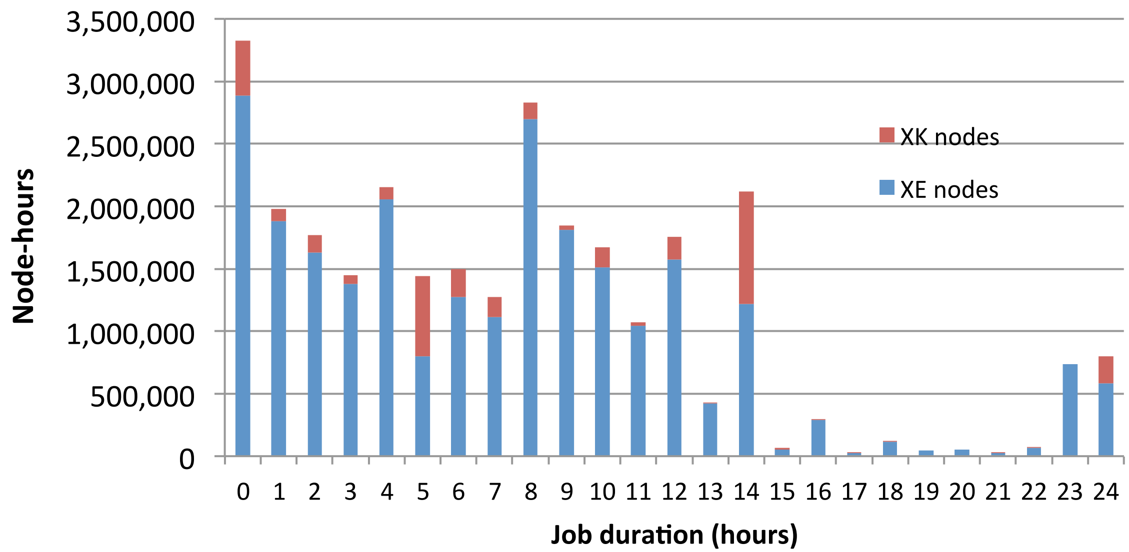
- “Friendly User” – no charging – Dec 2012 to April 1, 2013
- “Interim Full Service” – full charging – April 2, 2013 to July 14, 2013
  - Change in XE workload to smaller jobs.
    - Cautious husbanding of allocation. Some teams still analyzing Friendly User data.
  - Change in XK workload to large jobs.
    - GPU applications maturing.

Usage by PRAC team



Interim Full Service Job Characteristics

- April-July Interval
- Change in queue parameters in early June to have 24 hr. wall clock limit from 14 hr. prior.
- Large job expansion factor well under target of 10.
- 1+(time in queue/time requested)

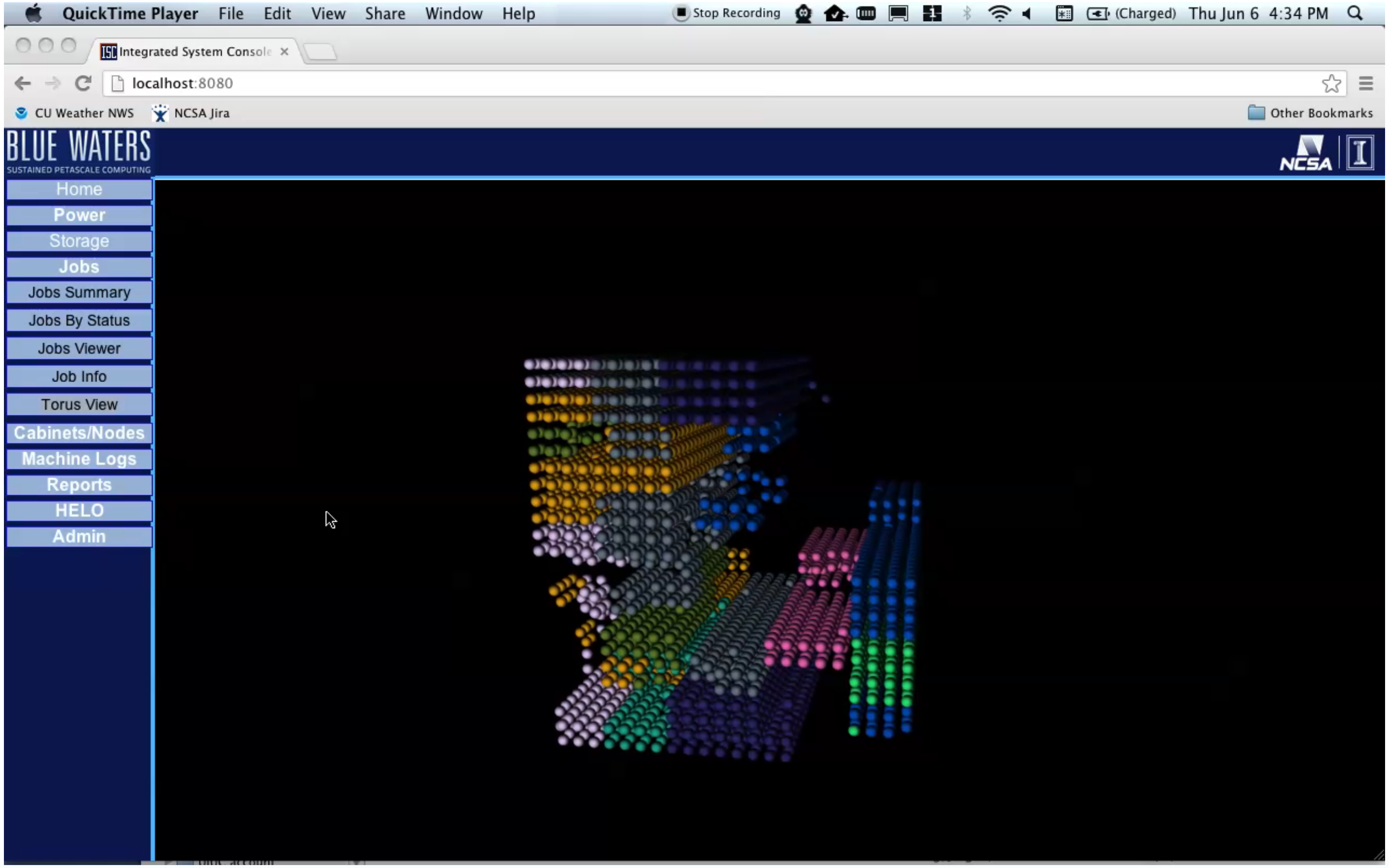


	Small	Medium	Large
XE nodes	1- 1,132 nodes 32-36,224 int cores	1,133 - 4,528 nodes 36,225 – 144,896 int cores	4,529 - 25,712 nodes 114,897 – 822,784 int cores
XK nodes	1 - 16 nodes	17 - 256 nodes	257 - 3,072 nodes

Expansion factor	Large jobs	Medium jobs	Small jobs
XK nodes	2.53	1.29	1.32
XE nodes	3.50	1.40	1.21

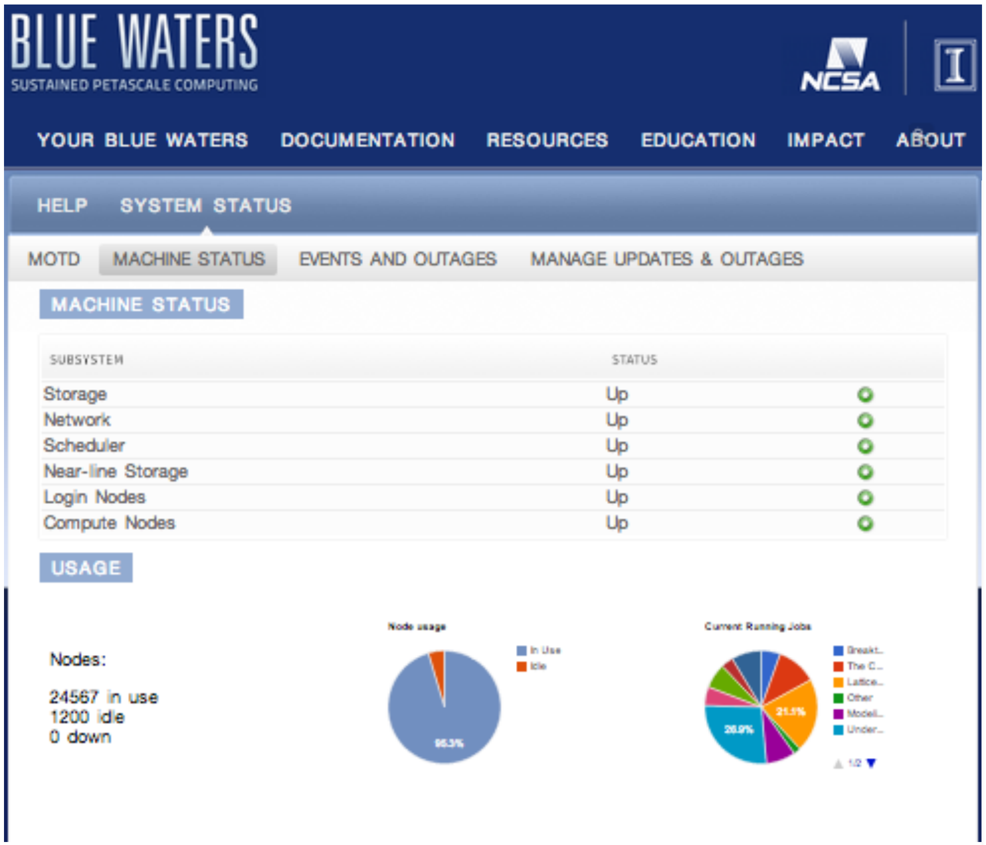
INSIGHTS FOR CURRENT AND FUTURE SYSTEMS

Need New Tools to Understand - Torus View

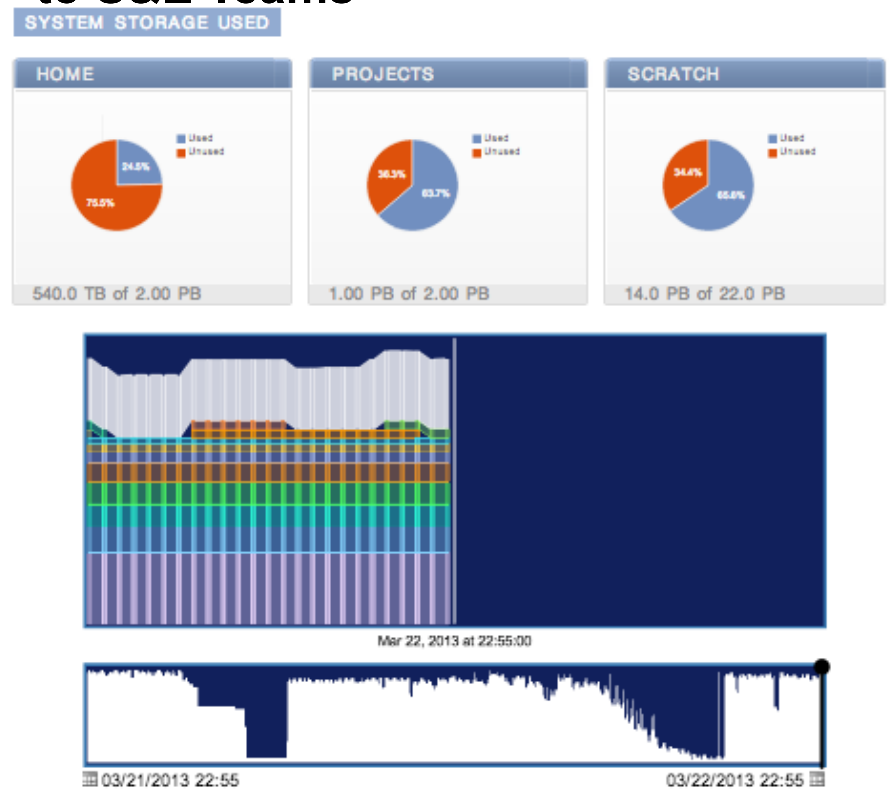


Jobs with node counts greater than 500 nodes (16,000 integer cores) are shown.

View from the Blue Waters Portal



As of Dec 21, 2012, Blue Waters has made available over 1 Billions Node hours (30 Billion integer core-hours) to S&E Teams



Exponential Growth in Data Rates for Event Processing – Much is Unstructured

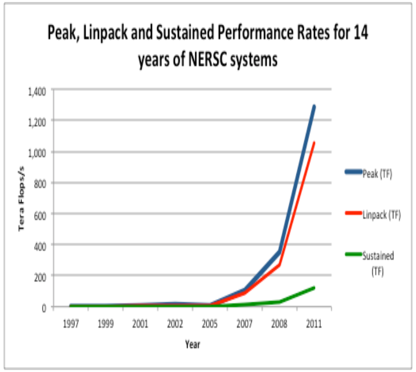
System	Sub-system	Size (Events/Day)	Total Event Types Identified
Blue Waters	Syslog	8 GB (50 million)	3,852
	HPSS	0.001 GB (900,000)	385
	Sonexion/on-line	3.5 GB (10 million)	3,112
	Resource Management	0.5 GB (15 million)	725
	ESMS	3 GB (12 million)	2,452
	Detailed compute node	TBD	TBD
	Detailed Gemini	TBD	TBD
	Detailed storage device level	TBD	TBD
	Total	>15 GB (>>88 million)	10,499

Totals System	Size (Events/Day)	Total Event Types Identified
Blue Waters	**15 GB (88 million)**	**10,499**
Blue gene/L	5.76 MB (25,000)	385
Blue Gene/P	8.12 MB (120,000)	3,112
NCSA Mercury Cluster	152.2 MB (1.5 million)	725
LANL Public Logs**	0.032 MB (432 log entries)	

* Depends on whether we have similar architectures or novel

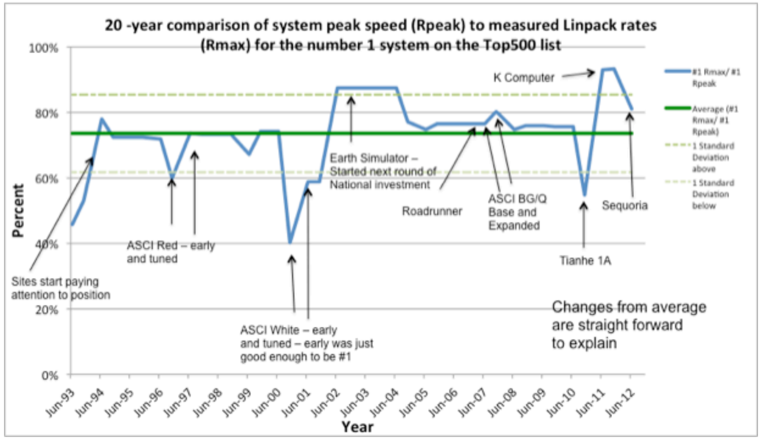
** Between 31 July 2003 11:04 and 30 April 2006 11:28 there are 433,490 messages in the syslogs. This represents 431.75 log entries/day or 32KB/day.

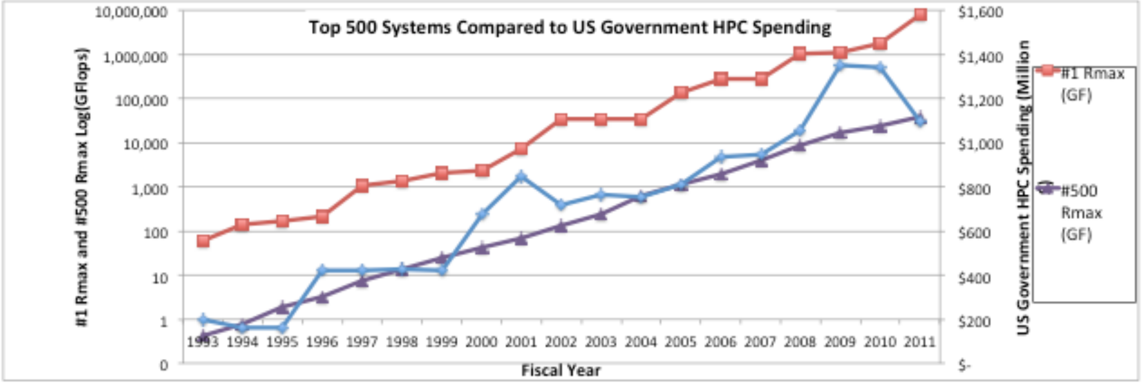
Very Positive Response to our Top500 List Decision



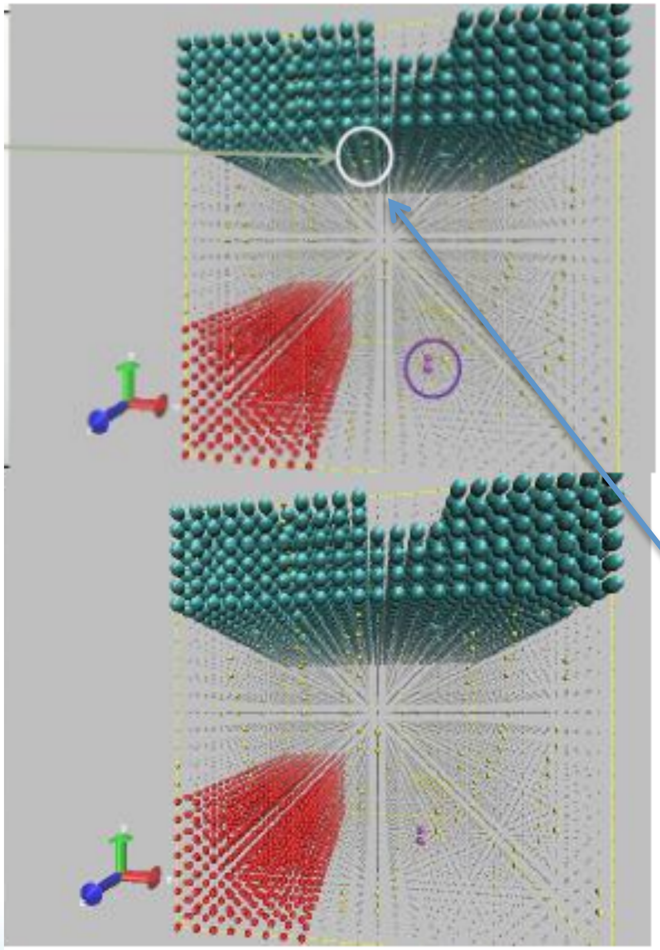
Top500 values do not correlate with vs measured System Sustained Performance - 13 years of systems at NERSC show this trend

TOP500 is dominated by who has the most money to spend–not what system is the best.





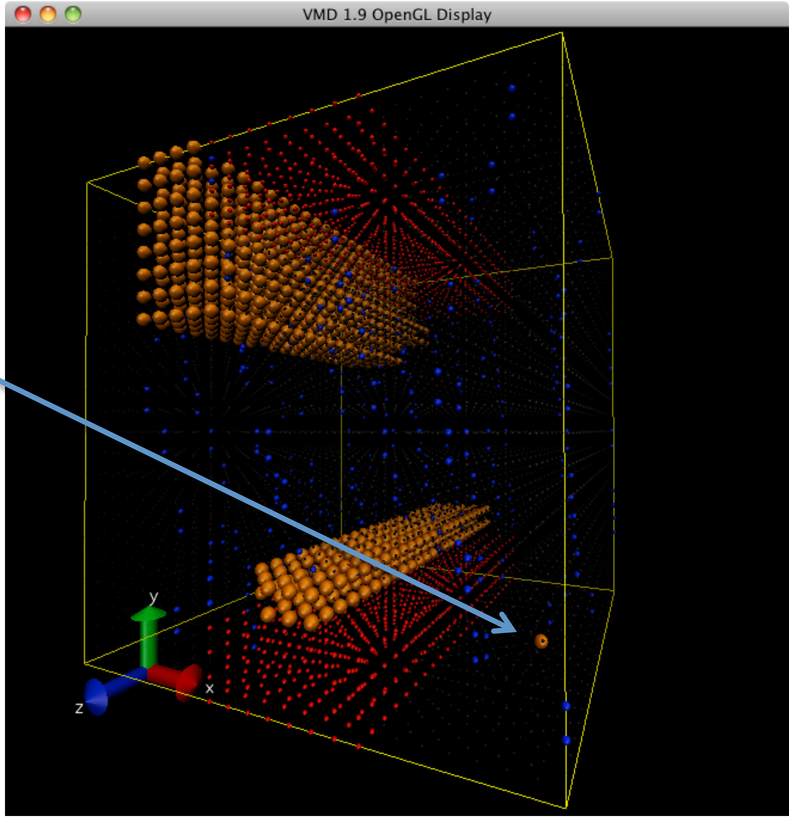
Topology Matters



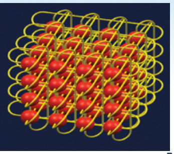
Much of the Benchmark tuning was topology based

1 poorly placed node out of 4116 (0.02%) can slow an application by >30% (on dedicated system)

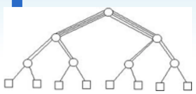
Just 1 of 3057 gemini down out in the wrong place of 6114 can slow an application by >20% (P3DNS – 6114 Nodes)

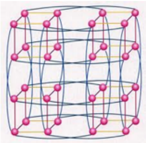


Appears in all system and many applications, but scale makes it clear

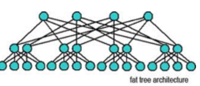


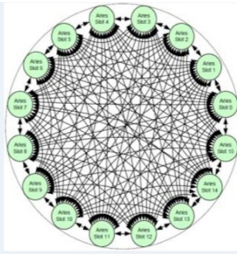
Performance and Scalability through Flexibility

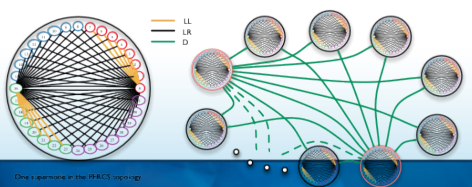




- Applies to all systems and topologies
- Need a system and application partnership to do the best
- Cray developed new management and tuning functions
  - Bandwidth Injection and Congestion Protection features – helps all systems
- BW works with science teams and technology providers to
  - Understand and develop better process-to-node mapping analysis to determine behavior and usage patterns.
  - Better instrumentation of what the network is really doing
  - Topology aware resource and systems management that enable and reward topology aware applications
  - Malleability – for applications and systems
    - Understanding topology given and maximizing effectiveness
    - Being able to express desired topology based on algorithms
    - Mid ware support
- Even if applications scale, consistency becomes an increasing issue for systems applications
- This will only get worse in future systems





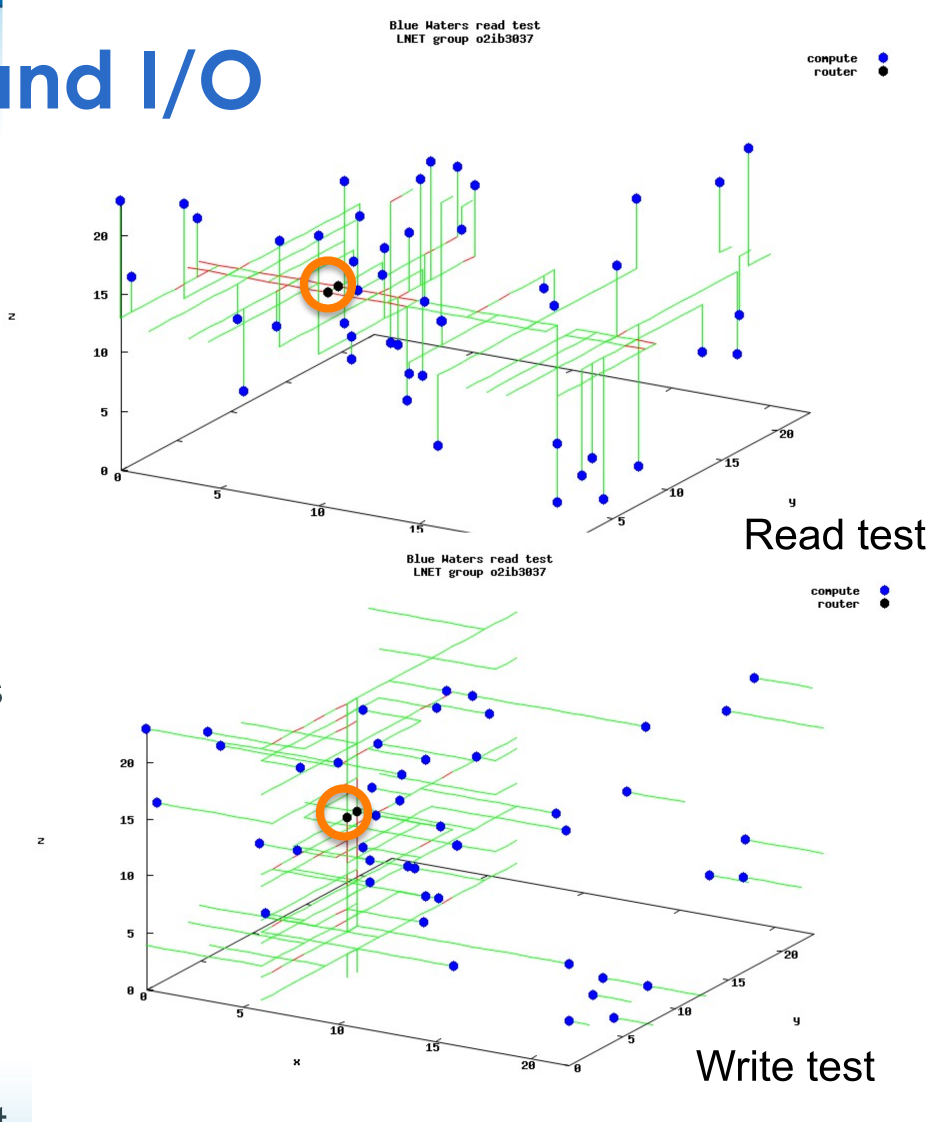


Flexible Resiliency Modes

- Run again 
- Defensive I/O (traditional Checkpoint/resta
  - Expensive and disruptive 
    - Extra overhead for application and system
  - Intrusive 
    - I/O infrastructure share across all jobs
- New C/R (node memory copy, SSD, Journaling,…)
- Spare nodes in job requests to rebalance work if a single point of failure 
  - Wastes resources
  - Run times do not support well yet (but can do it)
- Redistribute work within remaining nodes 
  - Charm++ , some MPI implementations
  - Takes longer
- Add spare nodes from system pool to job 
  - Job scheduler and resource manager and runtime all have to made more flexible

Storage and I/O

- On initial study 25% of applications had scale limited by I/O
- Parallel Storage and I/O significant challenges
  - Scale
  - Interaction with messages
  - Write/Read Asymmetry
- Middle ware becoming in important
  - Helps simplify complex storage systems
  - Meta organization
  - Portability
- Near-line storage
  - RAIT effective
  - Globus On-line Effective but needs feature improvements
- In-line OpenGL visualization is a benefit



Test results courtesy of Mark Swann @ Cray

Summary

- Blue Waters is the most intense computational and data focused system in the world at the moment
  - Computational and analytic resources
  - Storage and Data resources
  - Transfer rates
- BW already is producing unprecedented results
- More coming

Acknowledgements

This work is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (award number OCI 07-25070) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign, its National Center for Supercomputing Applications, Cray, and the Great Lakes Consortium for Petascale Computation.

The work described is achievable through the efforts of the many other on different teams.