BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Blue Waters Update – June 2013

Dr. William Kramer
National Center for Supercomputing Applications, University of Illinois

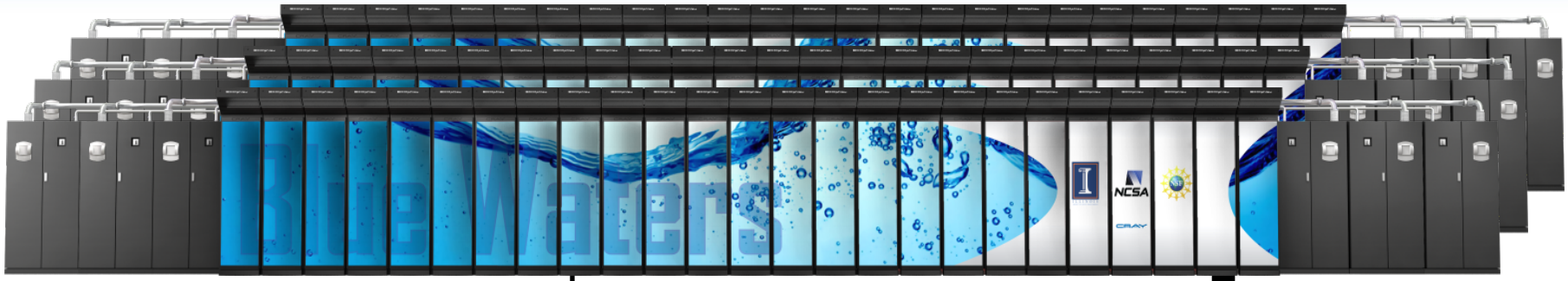
NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Summary

Blue Waters Introduction Video

http://www.youtube.com/watch?v=9jV8-akXjEk&list=UUNZs8P-Bw4liNLNMhcJAEug&index=2

Blue Waters Computing System



Aggregate Memory – 1.5 PB

10/40/100 Gb Ethernet Switch

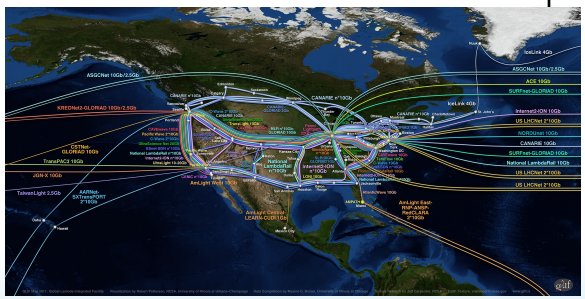
External Servers

IB Switch

>1 TB/sec

120+ Gb/sec

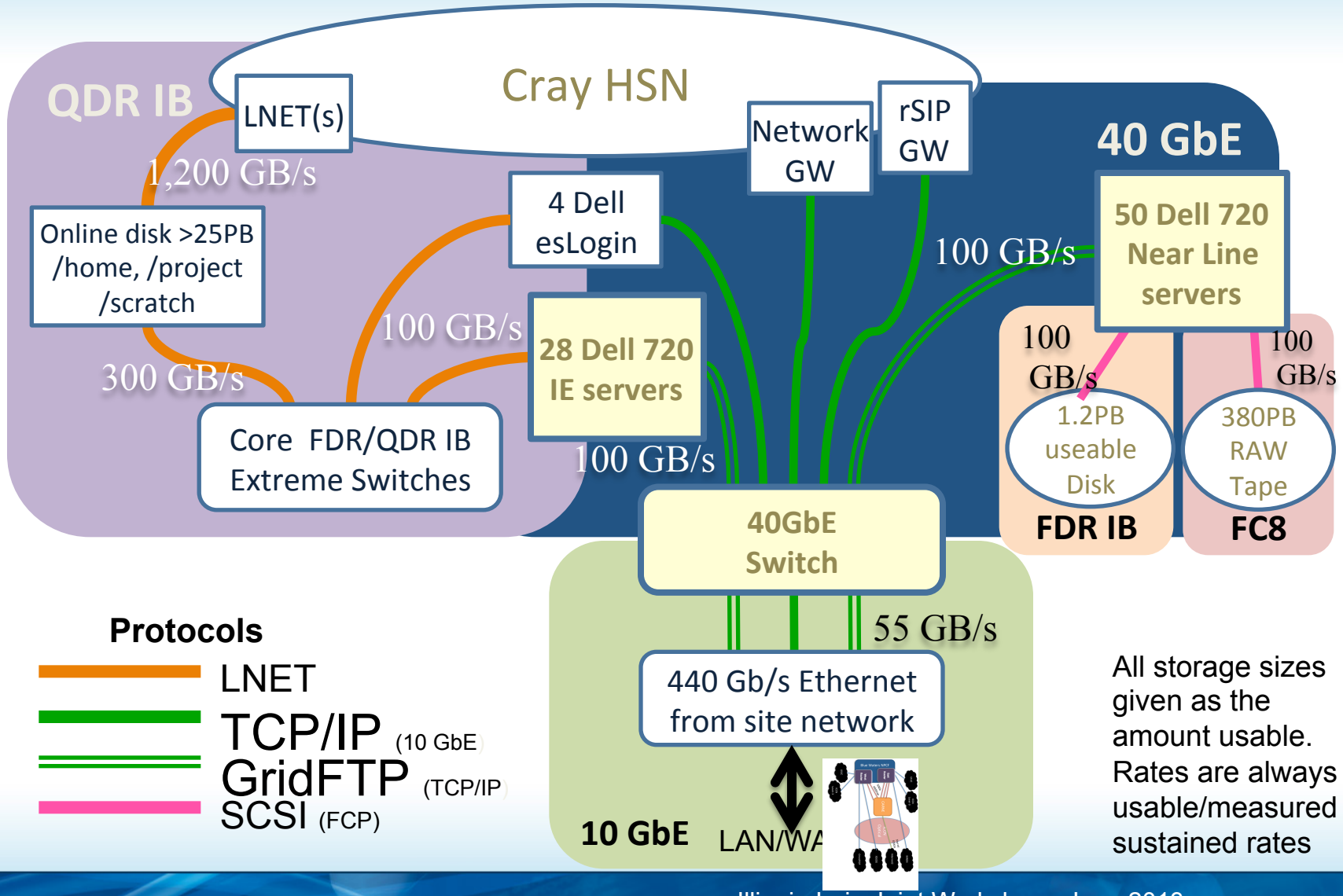
100 GB/sec



100-300 Gbps WAN

Spectra Logic: 300 usable PB

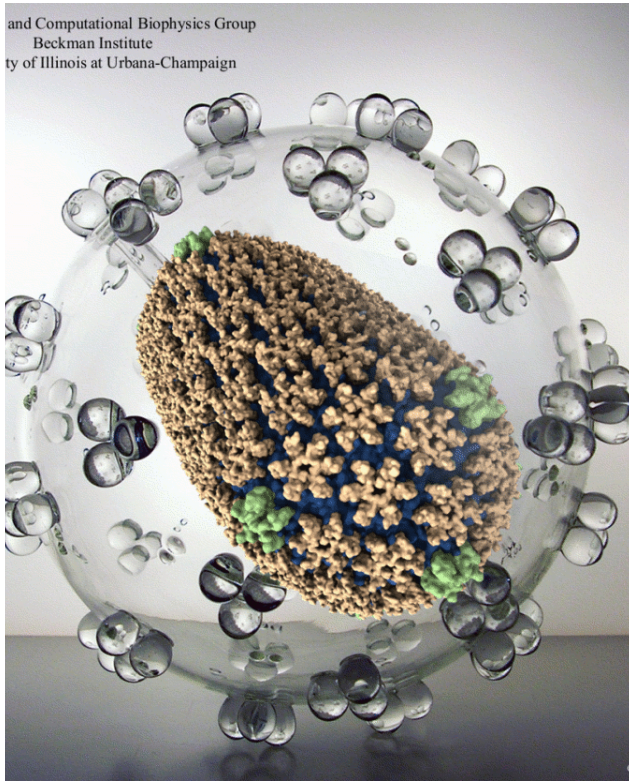
Sonexion: 26 usable PB



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

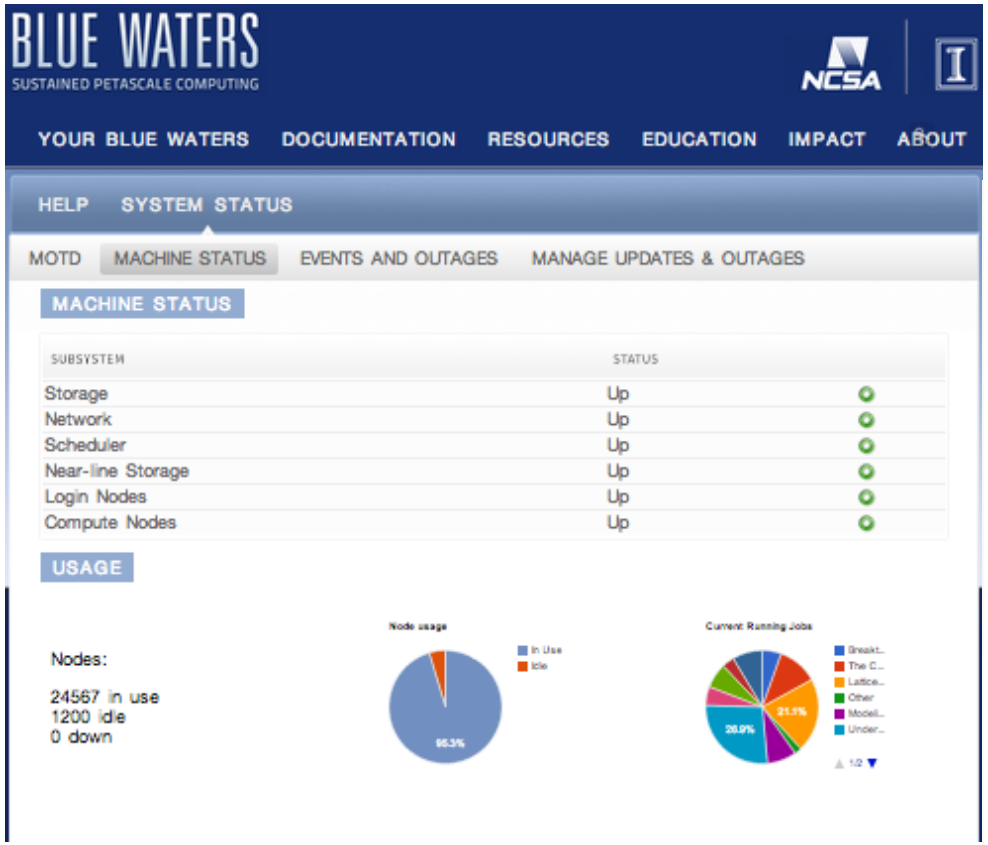
First Unprecedented Result – Computational Microscope

- Klaus Schulten and the NAMD group completed the highest resolution study of the mechanism of HIV cellular infection.
- May 30 Cover of Nature
- Orders of magnitude increase in number of atoms – resolution at about 1 angstrom

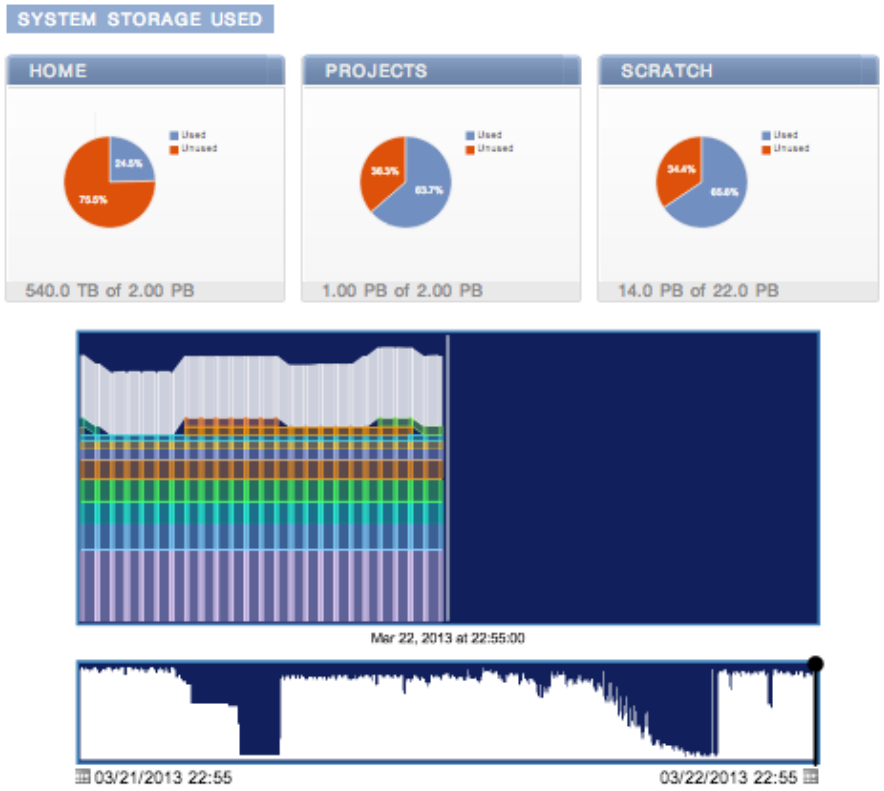


Science Area	Number of Teams	Codes	Struct Grids	Unstruct Grids	Dense Matrix	Sparse Matrix	N-Body	Monte Carlo	FFT	PIC	Significant I/O
Climate and Weather	3	CESM, GCRM, CM1/WRF, HOMME	X	X		X		X			X
Plasmas/Magnetosphere	2	H3D(M),VPIC, OSIRIS, Magtail/ UPIC	X				X		X		X
Stellar Atmospheres and Supernovae	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	1	GRIPS,Revisit						X			

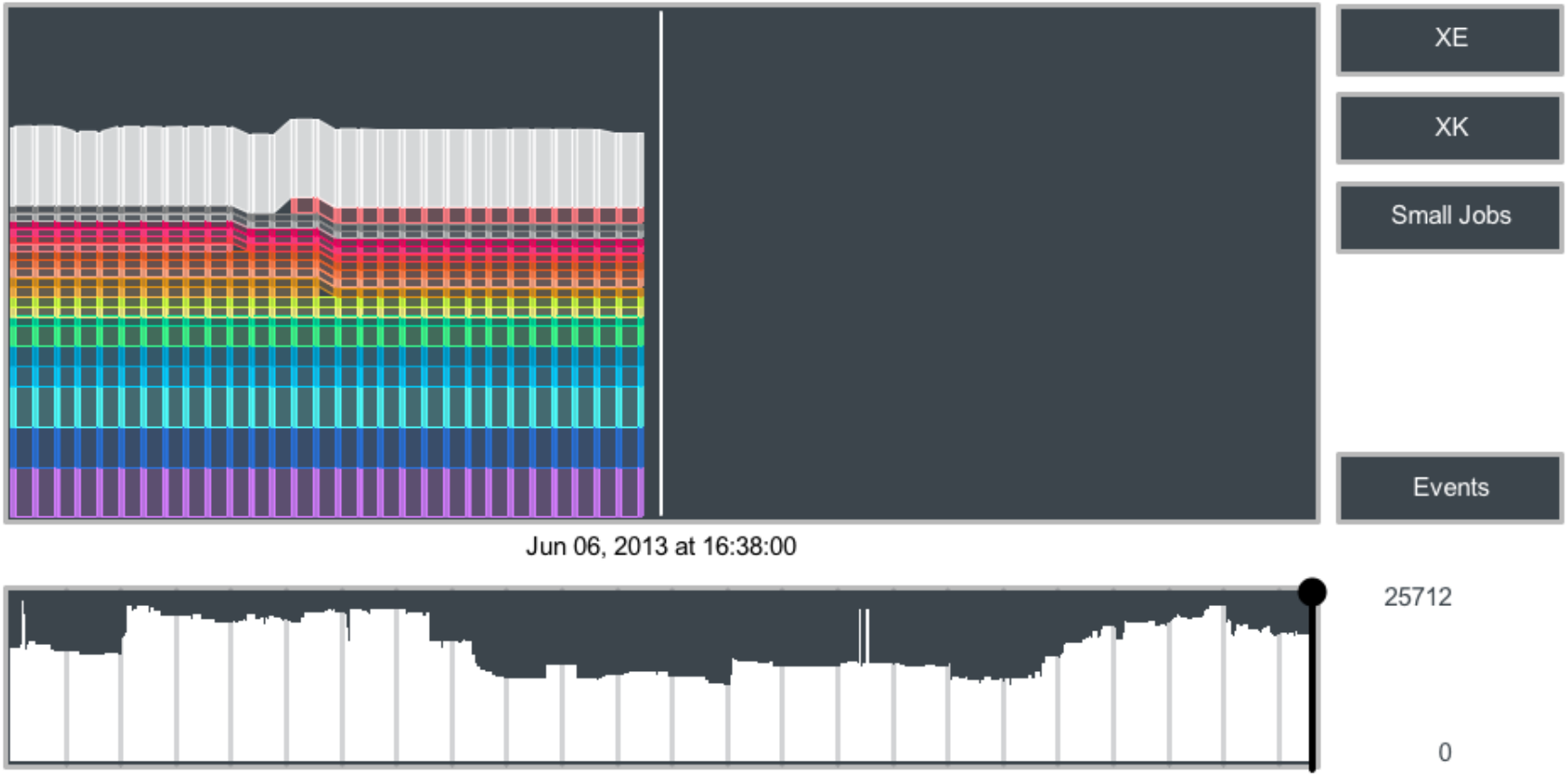
View from the Blue Waters Portal



As of April 2, 2013, Blue Waters has delivered over 1.3 Billion core-hours to S&E Teams

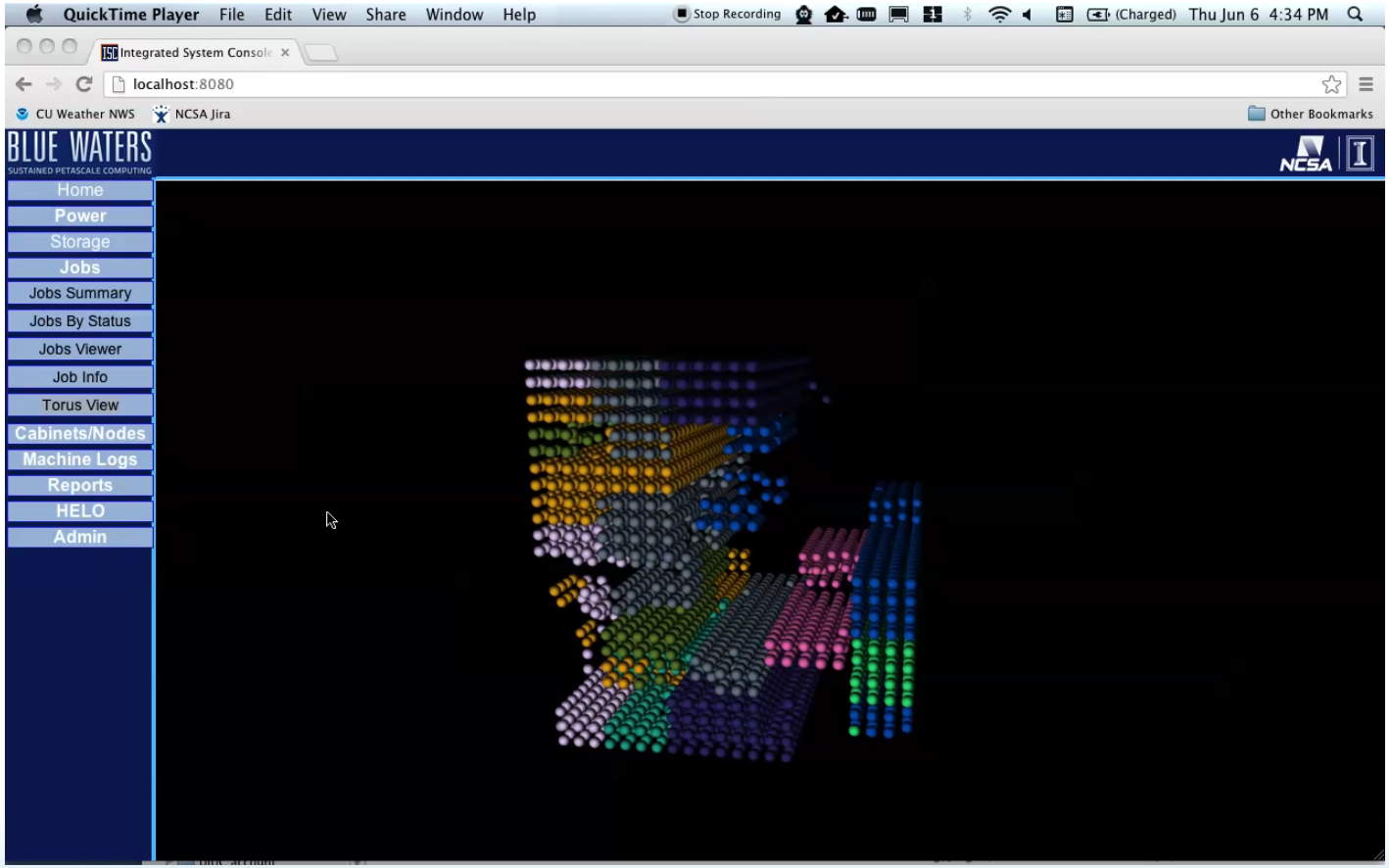


Ribbon View and Utilization



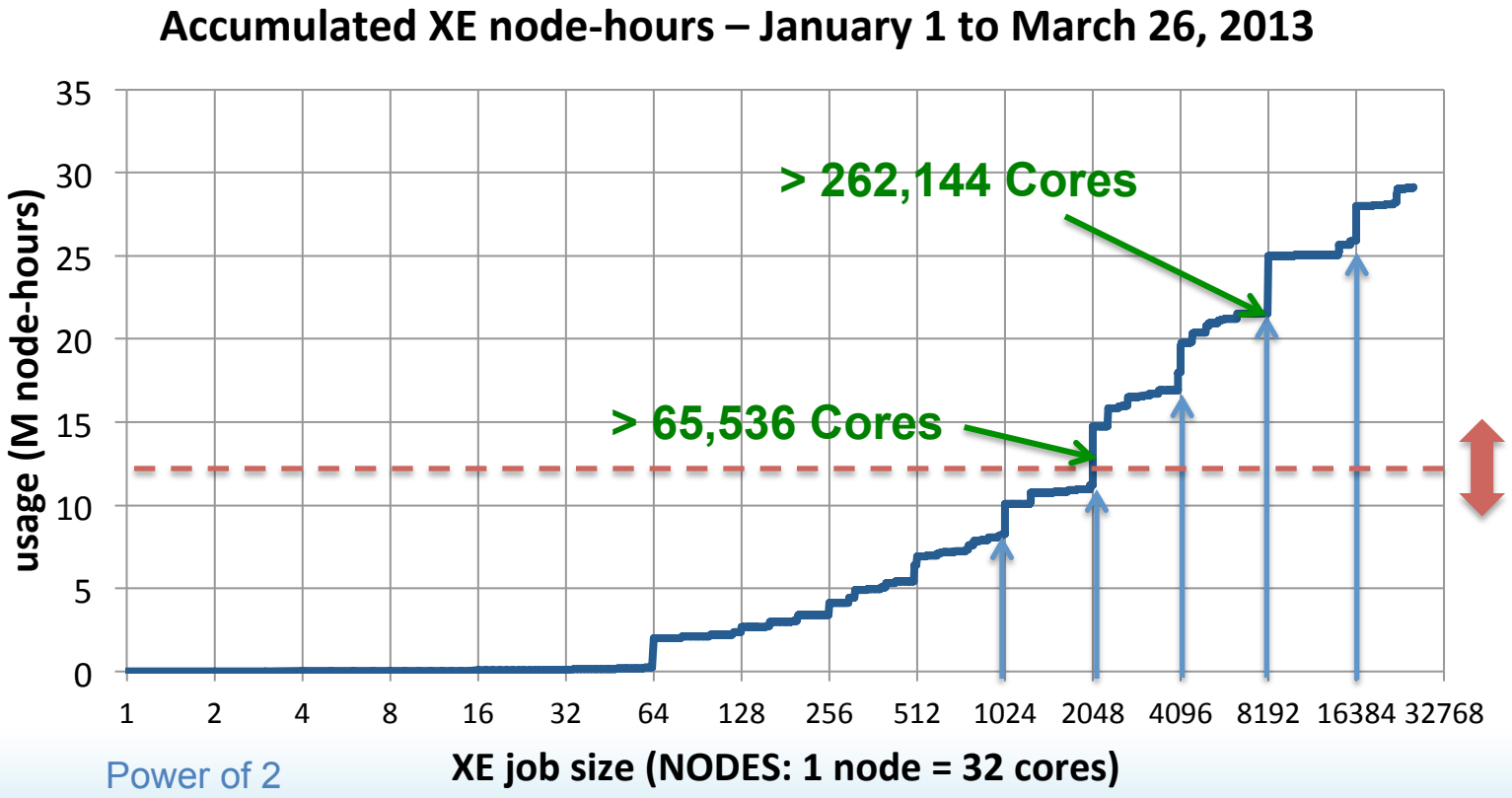
Torus View

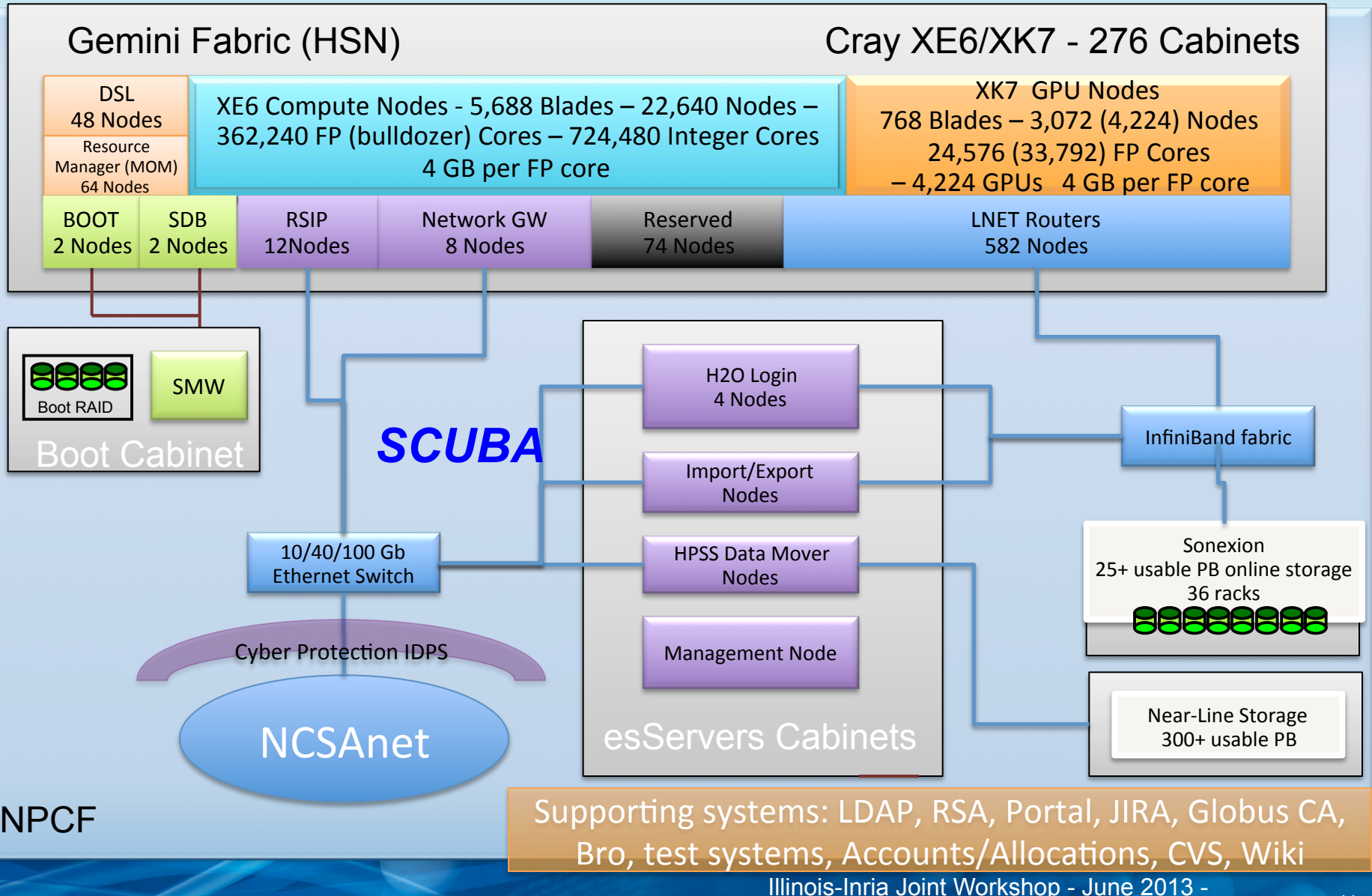
Jobs with node counts greater than 500 nodes (16,000 integer cores) are shown.



Usage Breakdown – Jan 1 to Mar 26, 2013

- Torque log accounting (NCSA, Mike Showerman)




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
768 Blades – 3,072 (4,224) Nodes
24,576 (33,792) FP Cores
– 4,224 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
10/40/100 Gb Ethernet Switch
Cyber Protection IDPS
NCSAnet
Near-Line Storage
300+ usable PB
NPCF
Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA, Bro, test systems, Accounts/Allocations, CVS, Wiki

Sustained Petascale Performance (SPP)

- SPP is an instance of the Sustained System Performance (SSP) **Method** of Evaluating systems
  - Method means there 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
  - The method was formally defined and expanded at Berkeley so it is generalized to cover any scale, any workload 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 workload, systems, etc.
- SPP is the Blue Waters/NSF implementation of the SSP Method

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

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*

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

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
- Added Criteria to the Method
  - Runs at full scale of the SPP codes
  - Comparing application 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* $(P_{XE6,i})$ × N_{XE6} + *Geometric Mean for all j* $(P_{XK7,j})$ × N_{XK7}

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.

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

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
- At least three SPP benchmarks run at full scale
- Kepler processors have to add at least 13% more above the x86 SPP

BW X86 SPP Test Components

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

BW Kepler SPP Test Components

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 SPP Results

- Composite x86 SPP Contribution
  - Before Upgrade – 1.08 PF
  - After Upgrade of 12 XK racks – 1.10 PF
- Composite Kepler SPP Contribution
  - Before Upgrade - 0.16 PF
  - After Upgrade of 12 XK racks – 0.21 PF
- Composite System SPP
  - Before upgrade – 1.24 PF
  - After Upgrade of 12 XK racks – 1.31 PF

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 3.1-49x 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
- See Celso Mendes' talk at 1 pm for more details

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 significant improve performance projections
- Balance run times with optimal performance (need to have ability to do tuning and improvement)

Remember

- SPP is an implementation of SSP that attempts to represent some part of the NSF workload.
- The Method is General, but the Implementation is specific

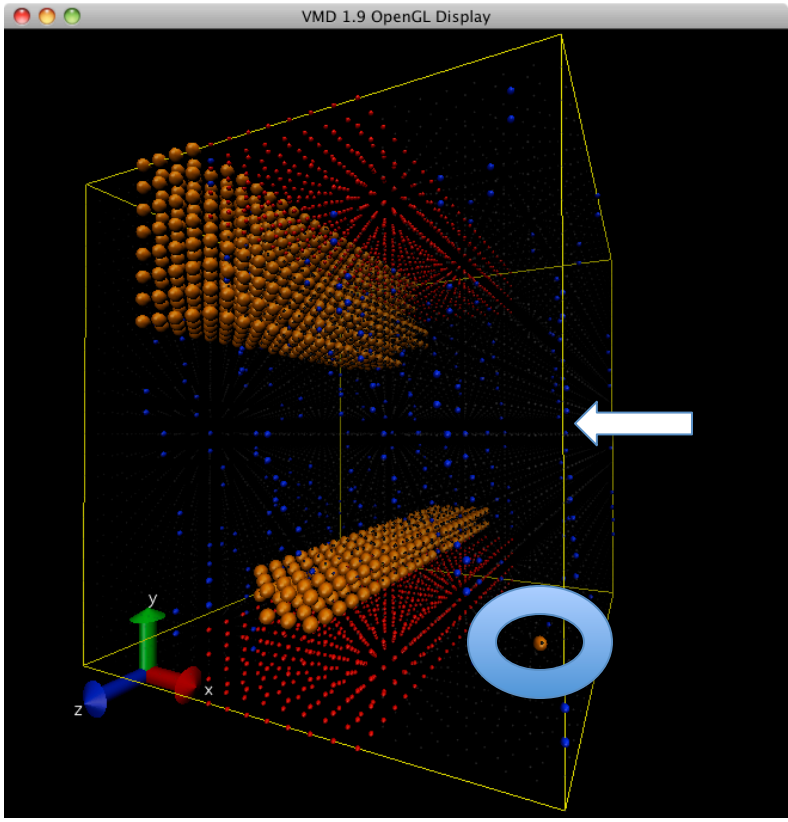
Blue Waters & Titan Computing Systems

System Attribute	UIUC/NCSA *Blue Waters*	DOE/ORNL *Titan*
Vendor(s)	Cray/AMD/NVIDIA	Cray/AMD/NVIDIA
Processors	Interlagos/Kepler	Interlagos/Kepler
Total Peak Performance (PF/s)	(11.6) 13.1	27.11
Total Peak Performance (CPU/GPU)	7.6/5.5	2.63/24.5
Number of CPU Modules (8 cores/Module)	49,504	18,688
Number of GPU Chips	(3,072) 4,224	18,688
Clock Speed CPU/GPU (Ghz)	2.3/.732	2.1/.732
SPP Sustained Performance (PF/s) (w/o clock diff)	**1.31**	**0.64 (est)**
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
2013planned upgrade		32
Sustained Disk Transfer (Average/Highest) (TB/sec)	1.2/1.4	0.245
2013 planned upgrade		~1
Amount of Near-line/Archival Storage (Usable/Maximum) (PB)	300/400	125/250
2013 planned upgrade		150/300
Protection from single point of tape failure	Yes	No
Sustained Tape Transfer (GB/sec)	88	18
Wide Areas Bandwidth (Gbps)	120	
Upgrade in 2014	300	

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

Area of More Focus Topology

- Impact - 1 poorly placed node out of 4116 (0.02%) can slow an application by >30%
- On a dedicated system!
- It is hard to get an optimal topology assignments, especial in non-dedicated use, but is should be easy to avoid really detrimental topology assignments.
- NCSA now has a development effort with Cray and Adaptive to improve the abilities to do topology aware scheduling and layouts



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

Use of Blue Waters

- All requests for time goes through a peer review
  - Science Goals, Readiness, Experience, Need to Unique Aspects of Blue Waters,…
  - Expect 10x to 20x fewer projects and users than XSEDE
    - That makes for projects that are very challenging
- Allocation types
  - National Science Foundation – PRAC Process ≥ 80%
  - Illinois – Illinois Process - ≤ 7%
    - Advancing areas of scholarship across Illinois that are dependent on compute-, memory- or data-intensive computing for progress.
    - Enhancing the University's position for competitive proposals where compute-, memory- or data-intensive computing is a critical factor.
    - Encouraging broad participation in the development, deployment, and use of petascale computing.
    - Developing software to make effective use of petascale systems for a broad range of scholarly applications.
    - Enriching the educational experiences of undergraduate students and graduate students throughout the University with an emphasis on petascale computing.
    - Stimulating economic growth through partnerships with business, industry, and government

Use of Blue Waters

- Allocation Types
  - Industry - ≤ 5%
  - Great Lakes Consortium for Petascale Computing – GLCPC Process - ≤ 2%
    - Research and education within the region, …. available to partners in the states of other members of the Great Lakes Consortium that have materially contributed to the success of the Blue Waters project
  - Education - ≤ 1%
    - integrate research and education in the national science and engineering community, or broaden participation of underrepresented demographic groups in science and engineering
      - upper level classes and workshops, awards, etc
    - Process to be announced soon
  - Principle Investigator Discretion - ≤ 5%

    Further research or education in the national science and engineering community and to broaden participation in high-performance computing [and] … to foster discovery and innovation.

What is needed to request allocation

- Pay attention to announcements
- All allocations will require a good proposal – scaled to the amount of resources requested
  - Goals, plan, timeline, milestones, ..
  - 2-3 pages to 30+ pages
- Why use of Blue Waters is necessary for the to solve the problem – not just for more cycles or accelerators
  - Unique use of characteristics – memory, scale, storage…
  - Use of information from BW

Near Term Activities

- Upgrades in plan for Blue Waters
  - Adding ~ 40% more XK nodes this summer
    - Makes the Torus a 24x24x24 topology
  - Upgrading resource management Software for better topology placement and use
    - Late this year for testing
  - Expanding storage and data support
    - Integrate Storage Hierarchies – 2014/2015
    - Innovative use of highly parallel file systems

Near Term Activities

- NCSA Enhanced Intellectual Services for Petascale Processing – NEIS-P^2
  - Education and Outreach
    - Support for
      - Workshops and classes in petascale+ topics
      - Blue Waters Fellowships and Internships
      - Blue Waters Symposia
  - Technology Insertion
    - Application and System Flexibility
      - Topology, resilience, runtimes, …
      - Heterogeneous use (XE and XK nodes)
      - New applications using many core methods
    - Storage and I/O
      - Performance, communication avoiding
      - Innovative frameworks on highly parallel file systems
    - Communication Sensitive methods and libraries
    - Other areas

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.