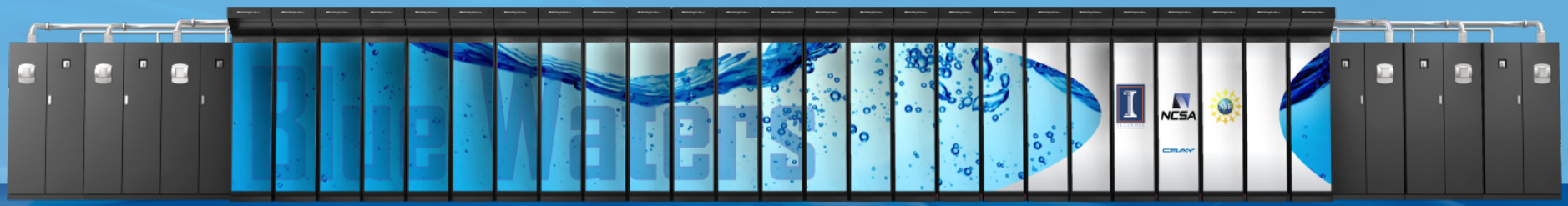
BLUE WATERS

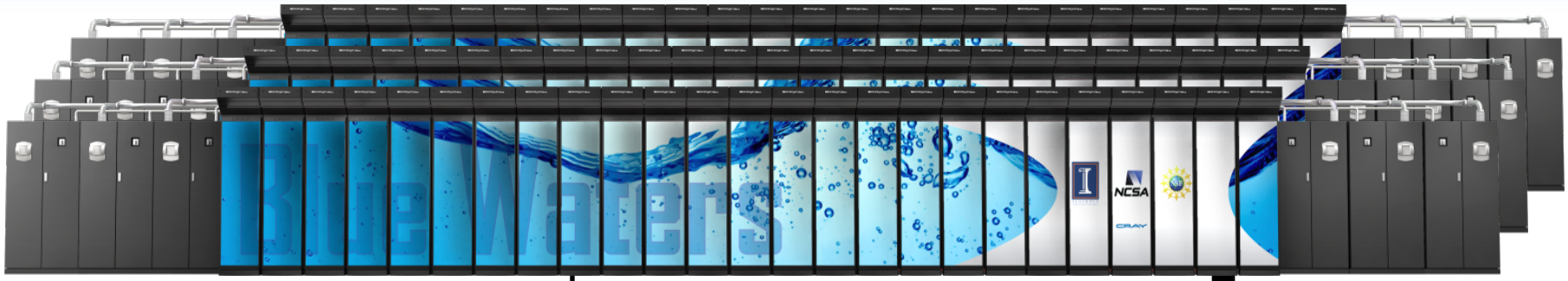
SUSTAINED PETASCALE COMPUTING

Blue Waters System Overview



NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Blue Waters Computing System



Aggregate Memory – 1.5 PB

Scuba Subsystem -
Storage Configuration for User Best Access

10/40/100 Gb Ethernet Switch

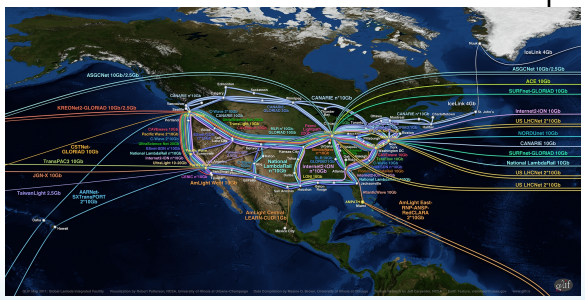
External Servers

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec



100-300 Gbps WAN

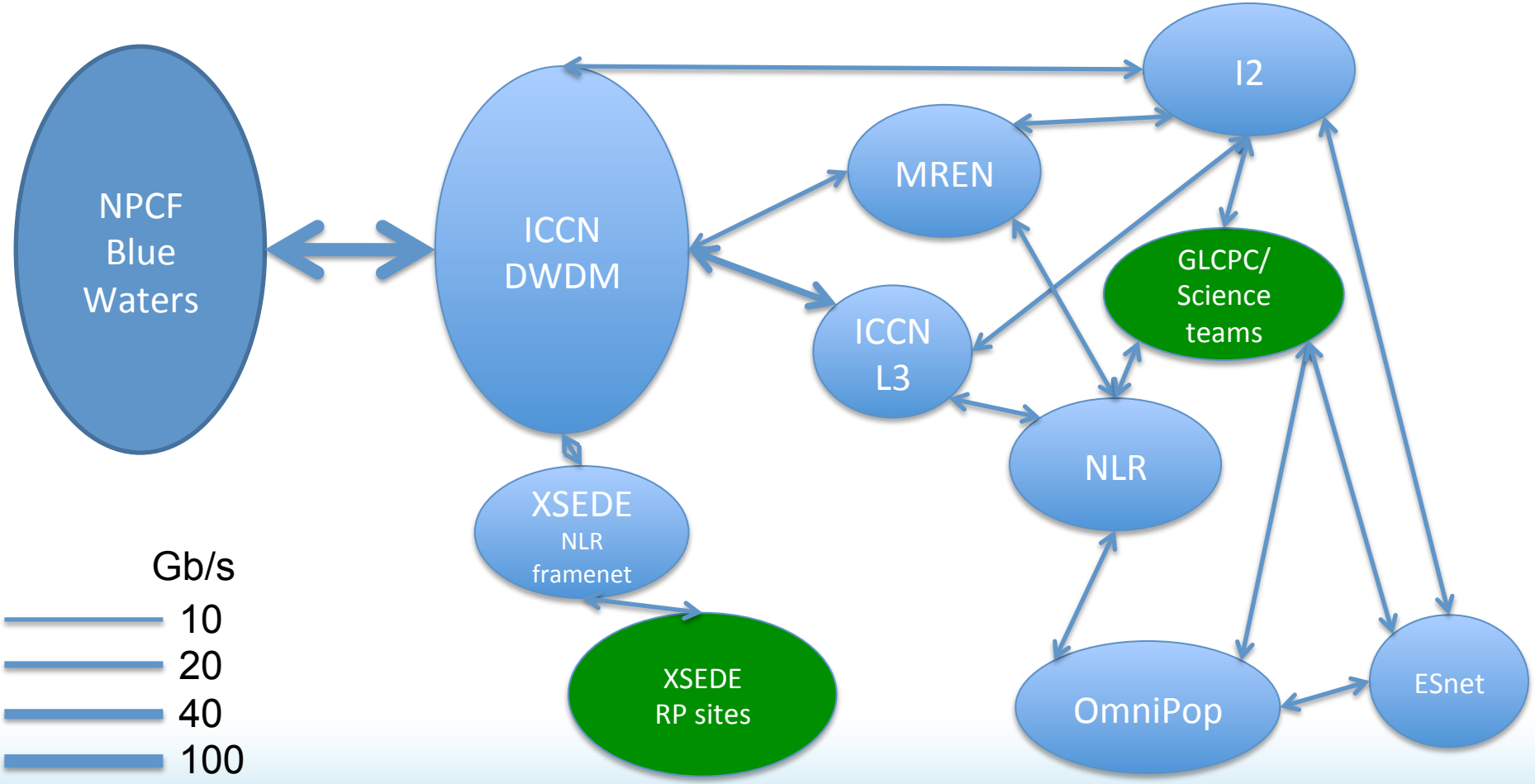
Spectra Logic: 300 usable PB

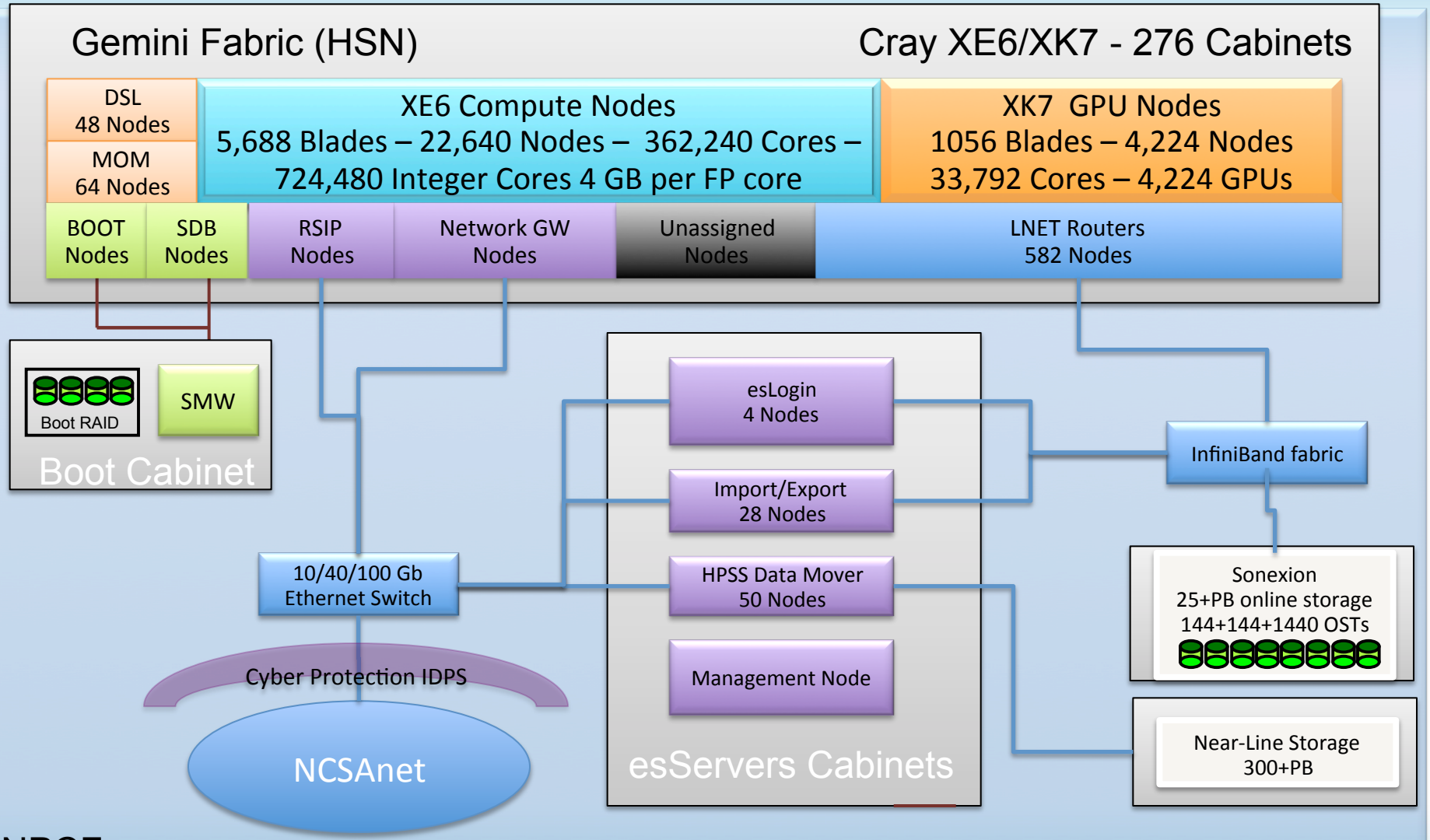
Sonexion: 26 usable PB

National Petascale Computing Facility

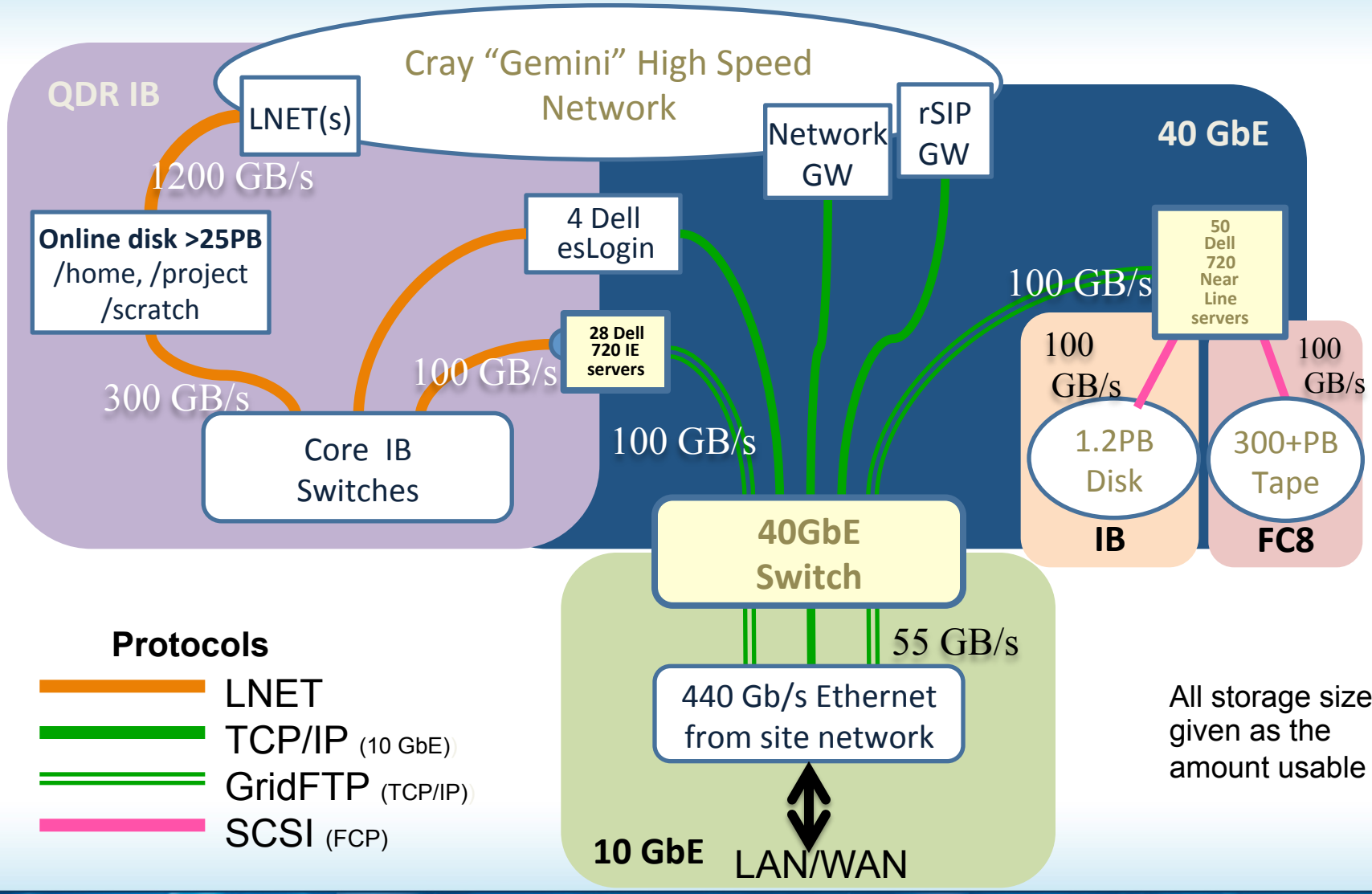
- Modern Data Center
  - 90,000+ ft^2 total
  - 30,000 ft^2 6 foot raised floor
    20,000 ft^2 machine room gallery with no obstructions or structural support elements
- Energy Efficiency
  - LEED certified Gold
  - Power Utilization Efficiency, PUE = 1.1–1.2
  - 24 MW current capacity – expandable
  - Highly instrumented

The Movement of Data




Gemini Fabric (HSN)
Cray XE6/XK7 - 276 Cabinets
DSL
48 Nodes
MOM
64 Nodes
XE6 Compute Nodes
5,688 Blades – 22,640 Nodes – 362,240 Cores –
724,480 Integer Cores 4 GB per FP core
XK7 GPU Nodes
1056 Blades – 4,224 Nodes
33,792 Cores – 4,224 GPUs
BOOT Nodes
SDB Nodes
RSIP Nodes
Network GW Nodes
Unassigned Nodes
LNET Routers
582 Nodes
Boot RAID
SMW
Boot Cabinet
esLogin
4 Nodes
Import/Export
28 Nodes
HPSS Data Mover
50 Nodes
Management Node
esServers Cabinets
InfiniBand fabric
10/40/100 Gb
Ethernet Switch
Cyber Protection IDPS
NCSAnet
Sonexion
25+PB online storage
144+144+1440 OSTs
Near-Line Storage
300+PB

NPCF



All storage sizes given as the amount usable

Blue Waters Nearline/Archive System

- Spectra Logic T-Finity
  - Dual-arm robotic tape libraries
  - High availability and reliability, with built-in redundancy
- Blue Waters Archive
  - Capacity: 380 PBs (*raw*), 300 PBs (*usable*)
  - Bandwidth: 100 GB/sec (*sustained*)
  - Redundant arrays of independent tapes RAIT for increased reliability.
  - Largest HPSS open production system.

Online Storage

home : 144 OSTs : 2.2 PB useable : 1 TB quota

projects: 144 OSTs : 2.2 PB useable : 5 TB group quota

scratch: 1440 OSTs : 22 PB useable : 500 TB group quota

- Cray Sonexion with Lustre for all file-systems.
- All visible from compute nodes.
- Scratch has 30 day purge policy in effect for both files and directories. Not backed up.
- ONLY home and project file-systems are backed up.

Nearline Storage (HPSS)

home: 5 TB quota

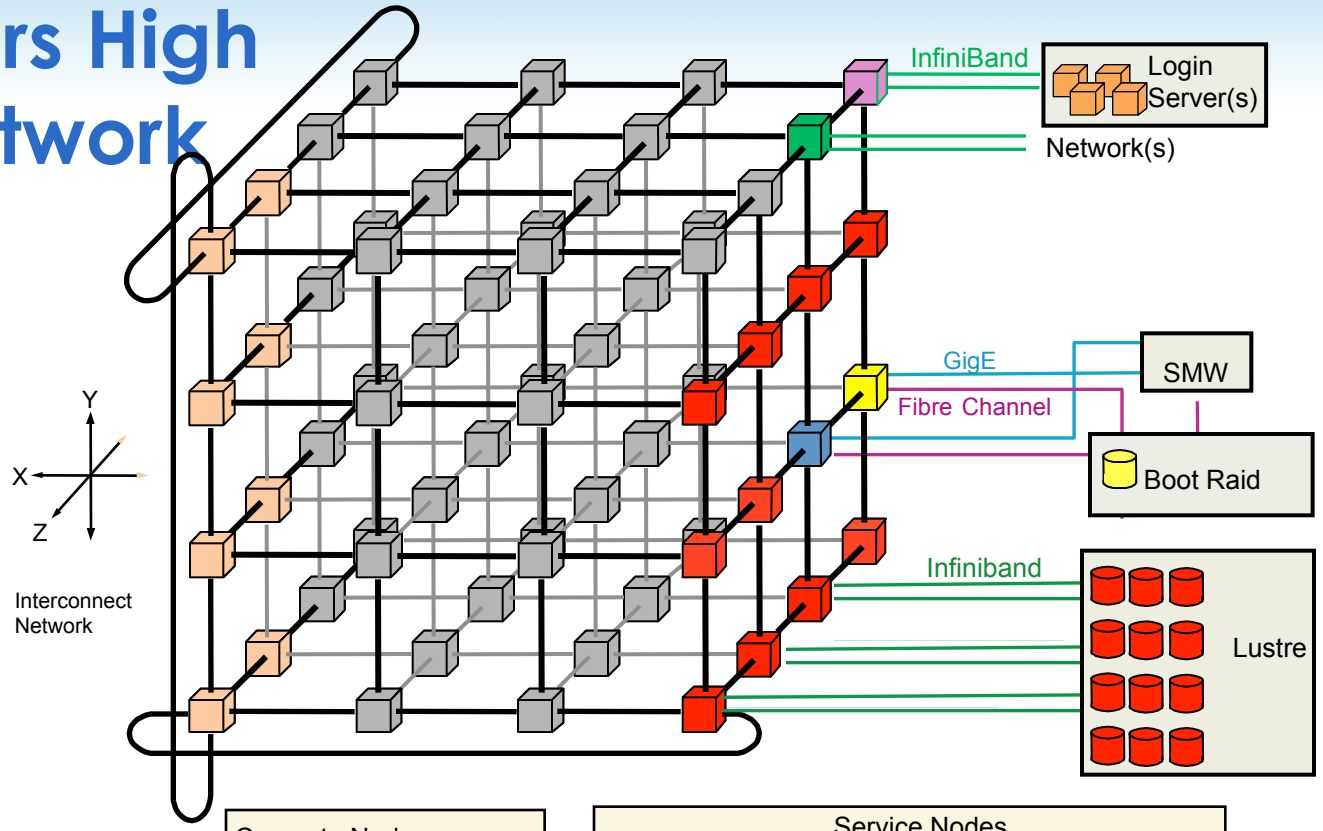
projects: 50 TB group quota

- IBM HPSS + DDN + Spectra Logic.
- Accessed via GO or globus-url-copy.

GO with Globus Online

- GridFTP client development for IE and HPSS nodes.
- Enabled data striping with GridFTP.
- Managed file transfers.
- Command line interface.
- Globus connect for sites without GridFTP endpoints.

Blue Waters High Speed Network



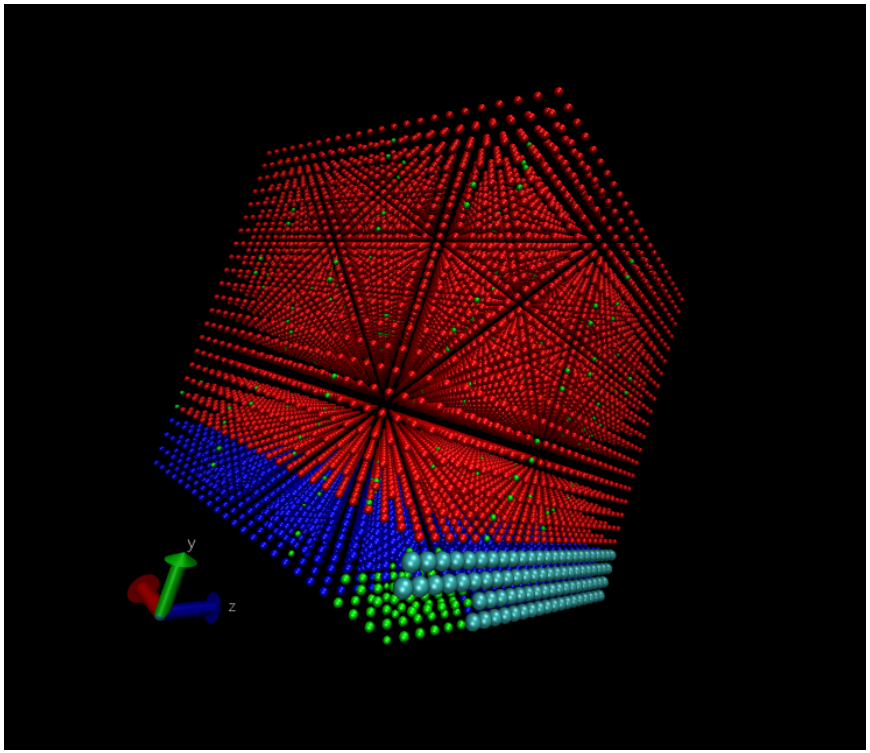
Blue Waters 3D Torus Size
24 x 24 x 24

HSN View

Gemini-node distinction

- Red – XE
- Green– Service, MOM, LNET
- Blue – XK

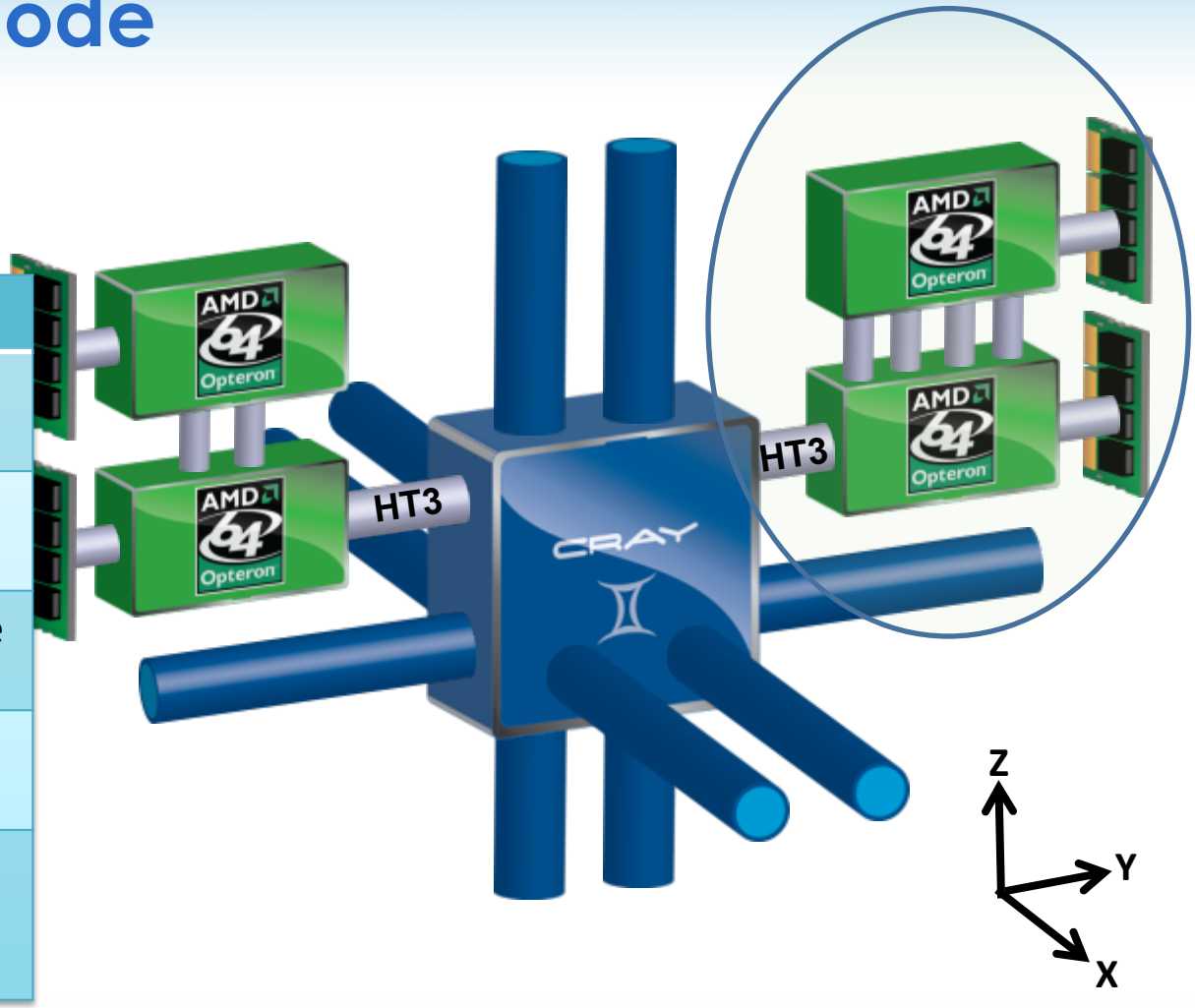
Complicated Topology



Blue Waters XE6 Node

Blue Waters contains 22,640 XE6 compute nodes

Node Characteristics	
Number of Core Modules*	16
Peak Performance	313 Gflops/sec
Memory Size	64 GB per node
Memory Bandwidth (Peak)	102 GB/sec
Interconnect Injection Bandwidth (Peak)	9.6 GB/sec per direction



Each core module includes 1 256-bit wide FP unit and 2 integer units. This is often advertised as 2 cores, leading to a 32 core node.

XE Node NUMA and core complexity

- 2 sockets per XE node.
- 2 NUMA domains per socket.
- 4 Bulldozer FP units per NUMA domain.
- 2 integer units per FP unit.

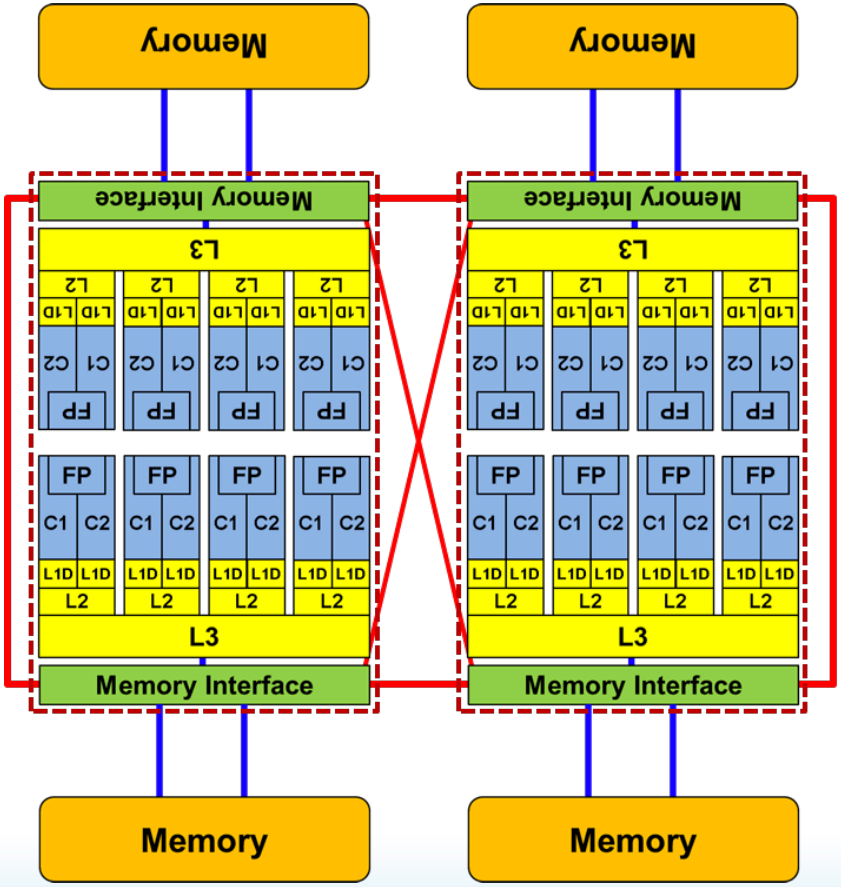


Image courtesy of Georg Hager http://blogs.fau.de/

CPU Node Comparison

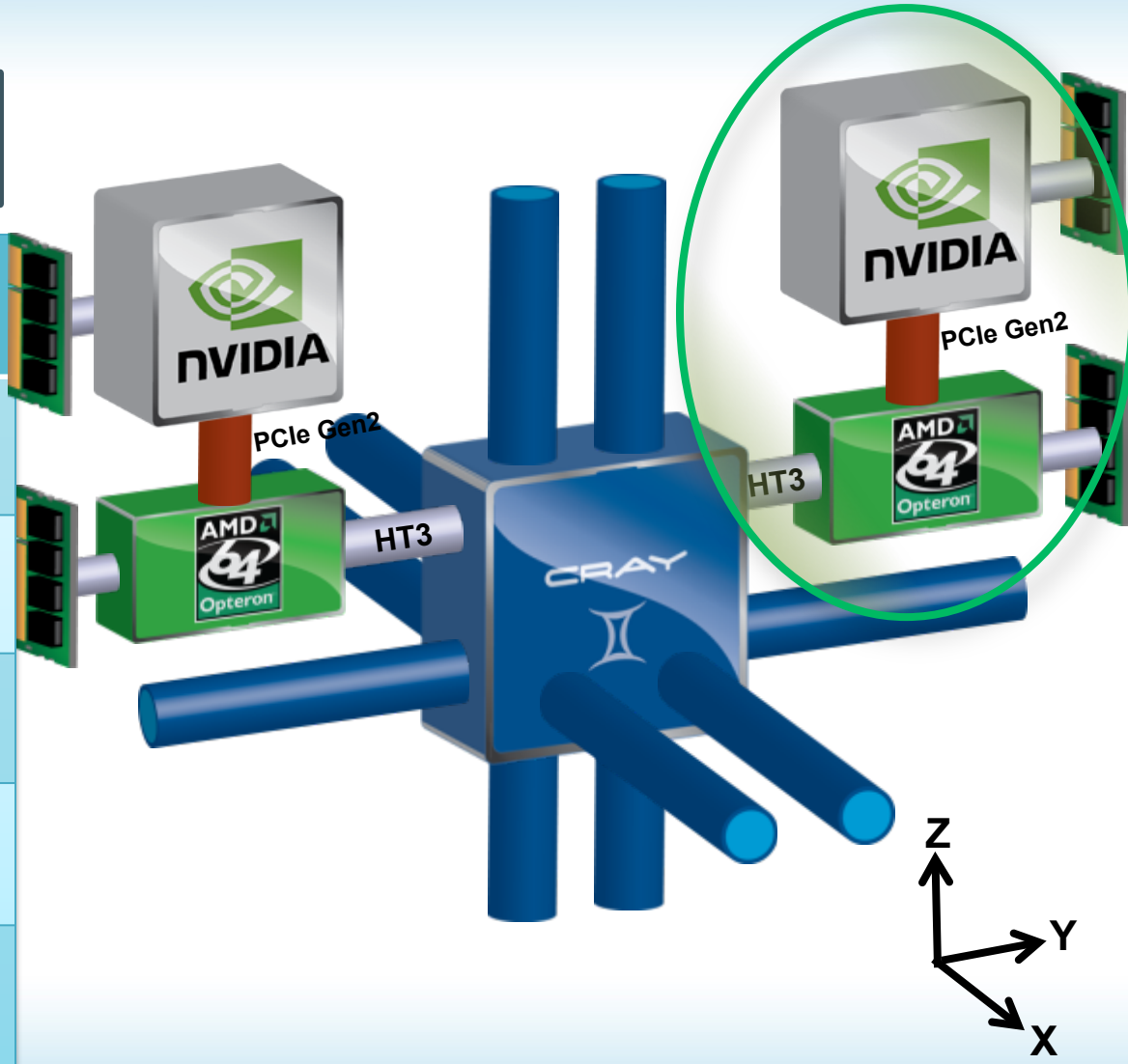
Node	Processor type	Nominal Clock Freq. (GHz)	FPU cores	Peak GF/s	Peak GB/s
Blue Waters Cray XE	AMD 6276 Interlagos	2.45	16*	313	102
NICS Kraken Cray XT	AMD Istanbul	2.6	12	125	25.6
NERSC Hopper XE	AMD 6172 MagnyCours	2.1	24	202	85.3
ANL IBM BG/P	POWERPC 450	0.85	4	13.6	13.6
ANL IBM BG/Q	IBM A2	1.6	16*	205	42.6
NCAR Yellowstone	Intel E5-2670 Sandy Bridge	2.6	16*	333	102
NICS Darter Cray XC30	Intel E5-2600 Sandy Bridge	2.6	16*	333	102

An * indicates processors with 8 flops per clock period.

Cray XK7

Blue Waters contains 4,224 NVIDIA K20x (GK110) GPUs

XK7 Compute Node Characteristics	
Host Processor	AMD Series 6200 (Interlagos)
Host Processor Performance	156.8 Gflops
K20x Peak (DP floating point)	1.32 Tflops
Host Memory	32GB 51 GB/sec
K20x Memory	6GB GDDR5 capacity 235GB/sec ECC



NVIDIA K20x

- Compute complexity.
- Additional memory hierarchy and types.

14 Multiprocessors, 192 CUDA Cores/MP: 2688 CUDA Cores
GPU Clock rate: 732 MHz (0.73 GHz)
Memory Clock rate: 2600 Mhz
Memory Bus Width: 384-bit
L2 Cache Size: 1572864 bytes
Maximum Texture Dimension Size (x,y,z) 1D=(65536), 2D=(65536, 65536), 3D=(4096, 4096, 4096)
Maximum Layered 1D Texture Size, (num) layers 1D=(16384), 2048 layers
Maximum Layered 2D Texture Size, (num) layers 2D=(16384, 16384), 2048 layers
Total amount of constant memory: 65536 bytes
Total amount of shared memory per block: 49152 bytes
Total number of registers available per block: 65536
Warp size: 32
Maximum number of threads per multiprocessor: 2048
Maximum number of threads per block: 1024

XK Features

- Hardware accelerated OpenGL with an X11 server. Not standard support by vendor.
- GPU operation mode flipped to allow display functionality (was compute only).
- X server enabled/disabled at job start/end when specified by user.
- Several teams use XK nodes for visualization to avoid transferring large amounts of data, shortening workflow.

Blue Waters Software Environment

Languages
- Fortran
- C
- C++
- Python
- UPC

Compilers
- Cray Compiling Environment (CCE)
- PGI
- GNU

Programming Models
- Distributed Memory (Cray MPT)
  - MPI
  - SHMEM
- Shared Memory
  - OpenMP 3.0
- PGAS & Global View
  - UPC (CCE)
  - CAF (CCE)

IO Libraries
- NetCDF
- HDF5
- ADIOS
- Resource Manager
- Adaptive Computing
- Visualization
- VisIt
- Paraview
- YT

Tools
- Environment setup
- Modules
- Debugging Support Tools
- Fast Track Debugger (CCE w/ DDT)
- Abnormal Termination Processing
- STAT
- Cray Comparative Debugger#
- Data Transfer
- globus online (Reliable File Transfer. No IT Required.)
- HPSS

Optimized Scientific Libraries
- LAPACK
- ScaLAPACK
- BLAS (libgoto)
- Iterative Refinement Toolkit
- Cray Adaptive FFTs (CRAFFT)
- FFTW
- Cray PETSc (with CASK)
- Cray Trilinos (with CASK)

Performance Analysis
- Cray Performance Monitoring and Analysis Tool
- PAPI
- PerfSuite
- Tau

Debuggers
- Allinea DDT
- Igdb

Prog. Env.
- Eclipse
- Traditional

Cray Linux Environment (CLE)/SUSE Linux

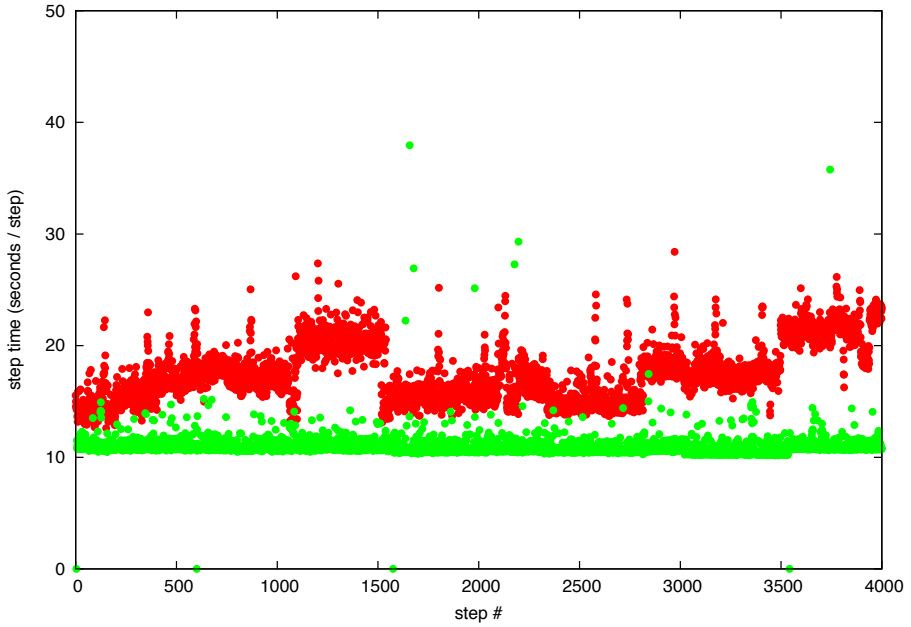
Cray developed	3rd party packaging
Under development	NCSA supported
Licensed ISV SW	Cray added value to 3rd party

Reliability

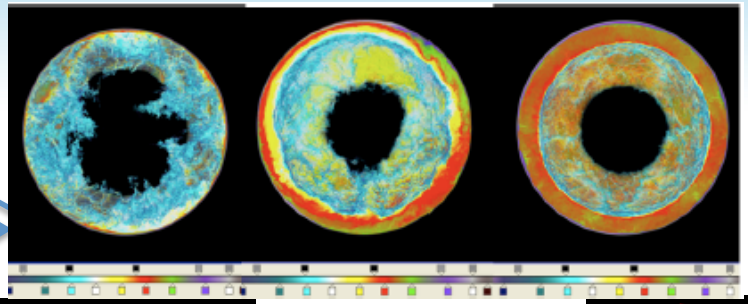
- We provide to the user a checkpoint interval calculator based on the work of J. Daly, using recent node and system interrupt data. User inputs number of XE and/or XK nodes, and the time to write a checkpoint file.
- September data
  - 22,640 XE nodes MTTI ~ 14 hrs.
  - 4,224 XK nodes MTTI ~ 32 hrs.
  - System interrupts MTTI ~ 100 hrs.
- Checkpoint intervals on the order of 4 – 6 hrs. at full system (depending on time to write checkpoint).

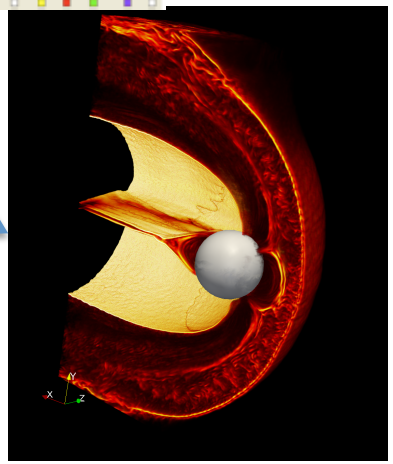
Consistency

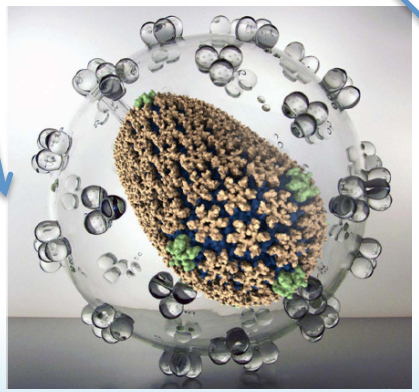
- Green job shows consistent step times.
- Red job shows step times impacted by other workloads.
- Both jobs were 8,192 XE nodes.
- Topology-aware scheduling provides convex shapes to confine communication. Only IO traffic collides on HSN.

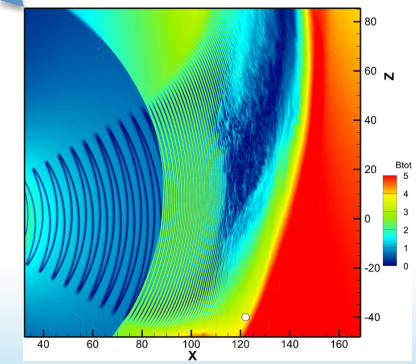


Science Area	Number of Teams	Codes
Climate and Weather	3	CESM, GCRM, CM1/WRF, HOMME
Plasmas/Magnetosphere	2	H3D(M),VPIC, OSIRIS, Magtail/UPIC
Stellar Atmospheres and Supernovae	5	PPM, MAESTRO, CASTRO, SEDONA, ChaNGa, MS-FLUKSS
Cosmology	2	Enzo, pGADGET
Combustion/Turbulence	2	PSDNS, DISTUF
General Relativity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









CADENS – Solar SuperStorm

- http://www.ncsa.illinois.edu/enabling/vis/cadens/documentary/solar_superstorms

Allocations on Blue Waters

- National Science Foundation
  - At least 80 percent of the capacity of Blue Waters--about 150 million node-hours each year--is available to scientists and engineers across the country through the National Science Foundation's Petascale Computing Resource Allocation program. Next proposal due date: Nov. 13, 2015.
- University of Illinois at Urbana-Champaign
  - Up to 7 percent of the computing capacity of Blue Waters—about 13 million node-hours each year—is reserved for faculty and staff at the University of Illinois at Urbana-Champaign. For application detiails, visit https://bluewaters.ncsa.illinois.edu/illinois-allocations. NOTE: Proposals are due September 15, 2015.
  - A portion of the University of Illinois Blue Waters allocation is being made available to the Blue Waters professors. For more details see https://bluewaters.ncsa.illinois.edu/bw-professors
- Great Lakes Consortium for Petascale Computation
  - Up to 2 percent is available to researchers whose institutions are members of the Great Lakes Consortium for Petascale Computation. For more information visit: https://bluewaters.ncsa.illinois.edu/glcpc-allocations or the GLCPC website.
- Education
  - Up to 1 percent of the Blue Waters compute capacity—or 1.8 million node-hours per year—is available for educators and students. For application details, visit https://bluewaters.ncsa.illinois.edu/education-allocations
- Industry
  - NCSA's industry partners and industry partners of any of the Great Lakes Consortium for Petascale Computation institutions also have opportunities to use Blue Waters. To apply for a Blue Waters industry allocation visit, https://bluewaters.ncsa.illinois.edu/new_account/bwspecialprojects/psp/.
- Innovation and Exploration
  - Up to 5 percent is available for innovation and exploration of new uses for high performance computation and data analysis. For application details, visit https://bluewaters.ncsa.illinois.edu/innovation-allocations.

Summary

- Outstanding Computing System
  - The largest installation of Cray's most advanced technology
  - Extreme-scale Lustre file system with advances in reliability/ maintainability
  - Extreme-scale archive with advanced RAIT capability
- Most balanced system in the open community
  - Blue Waters is capable of addressing science problems that are memory, storage, compute, or network intensive or any combination.
  - Use of innovative technologies provides a path to future systems
- Illinois/NCSA is a leader in developing and deploying these technologies as well as contributing to community efforts.