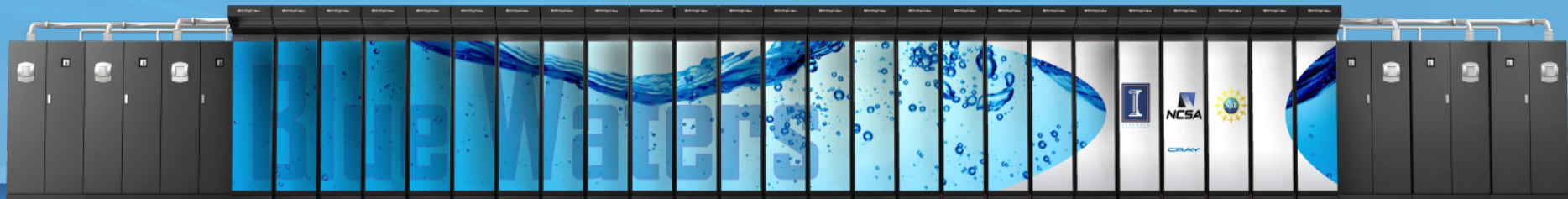


BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Blue Waters System Overview

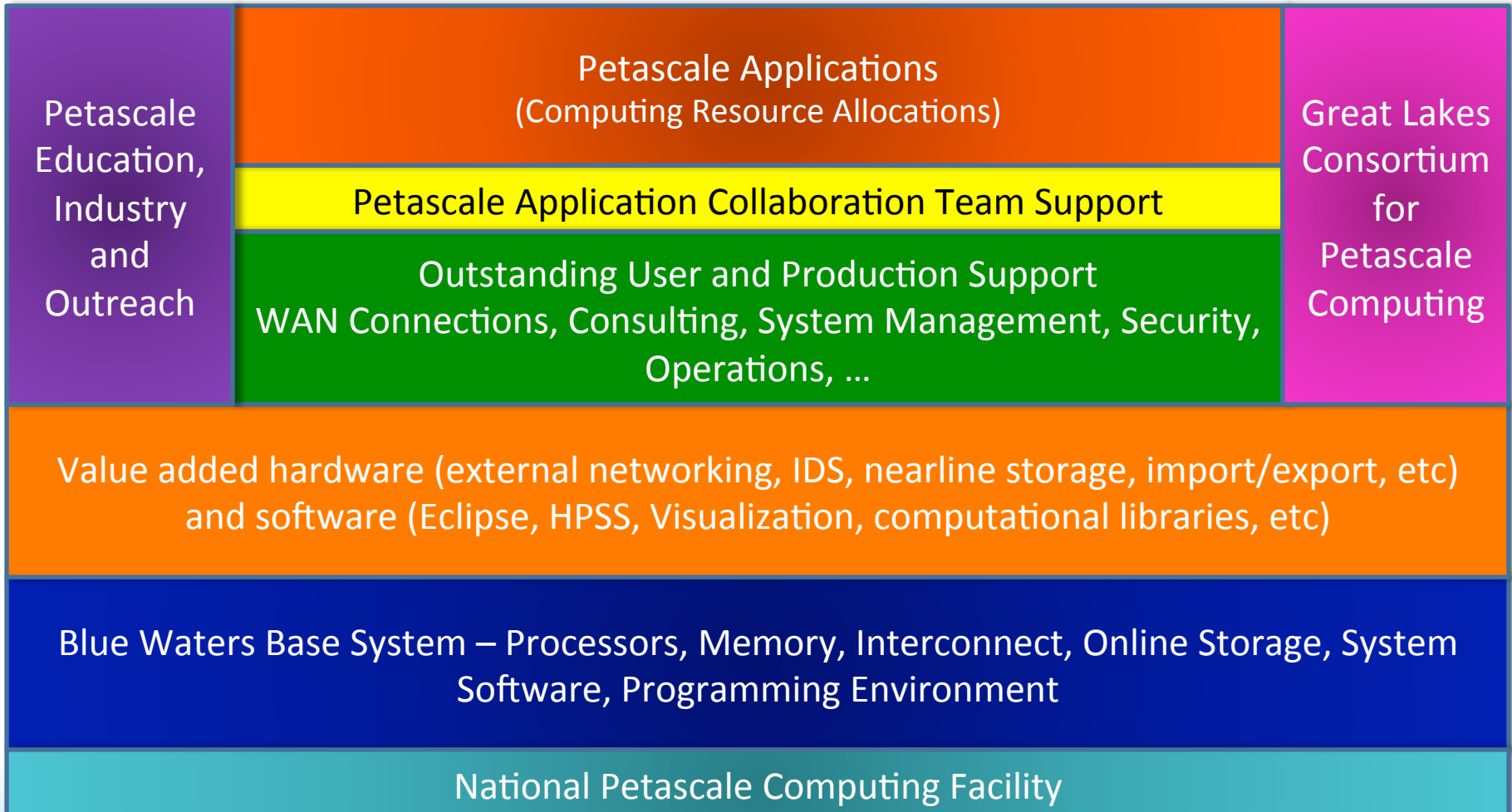
Greg Bauer



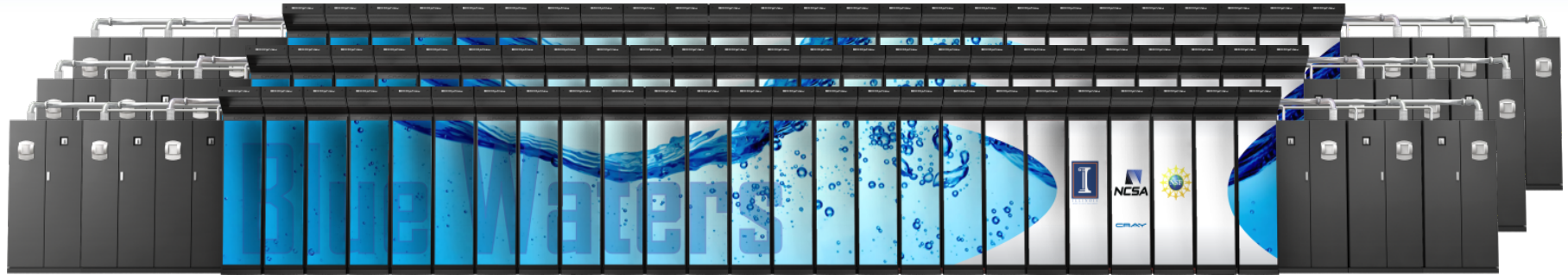
GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY®

The Blue Waters EcoSystem



Blue Waters Computing System



Aggregate Memory – 1.5 PB

Scuba Subsystem -
Storage Configuration
for User Best Access

120+ Gb/sec

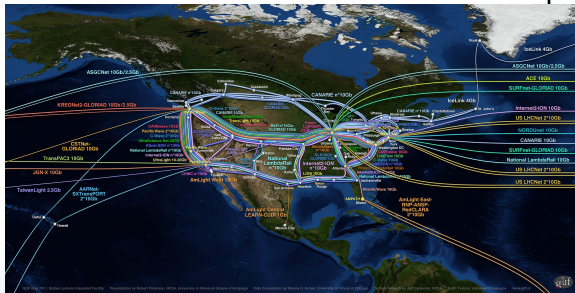
10/40/100 Gb
Ethernet Switch

External Servers

IB Switch

>1 TB/sec

100 GB/sec



100-300 Gbps WAN



Spectra Logic: 300 usable PB



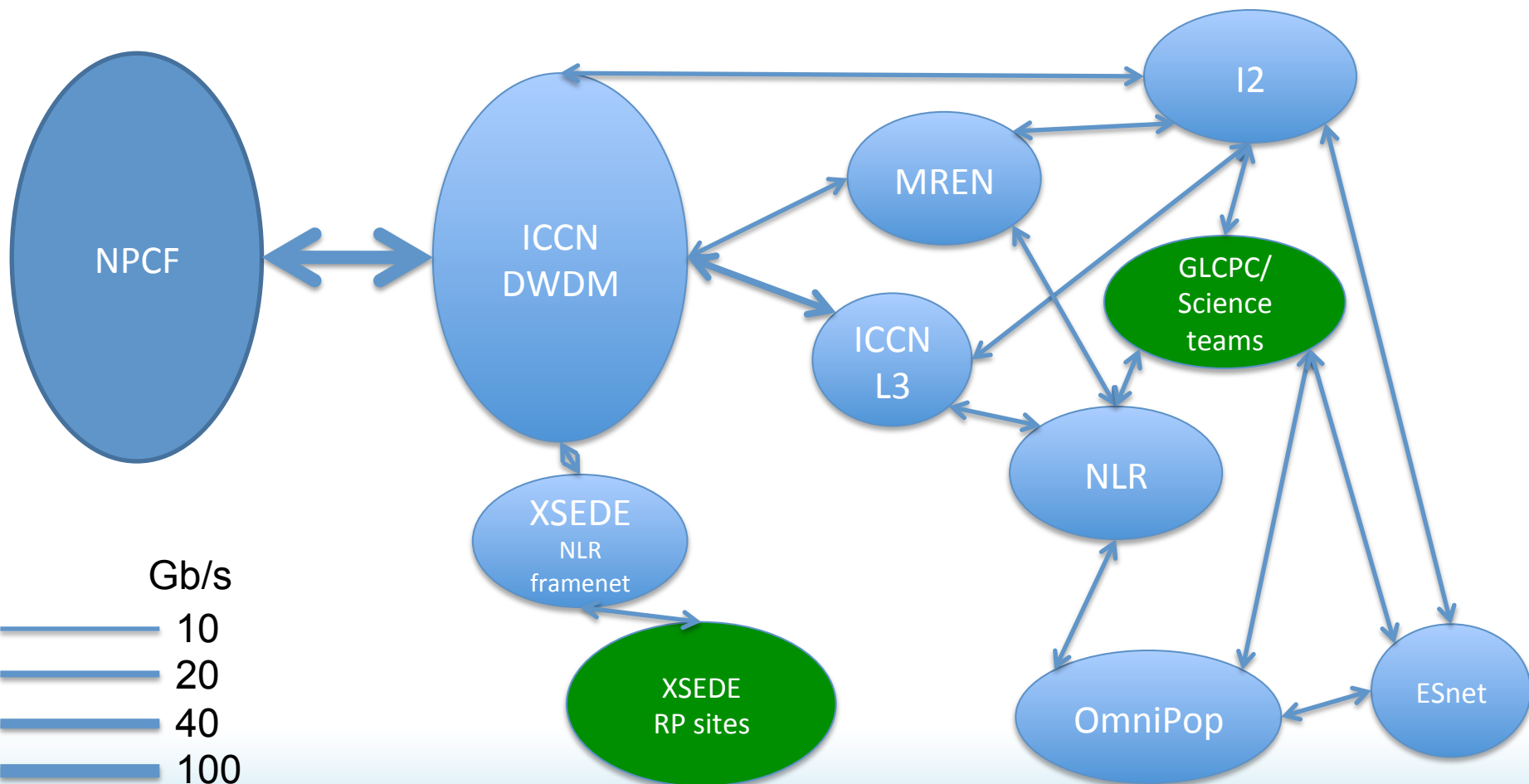
Sonexion: 26 usable PB

National Petascale Computing Facility



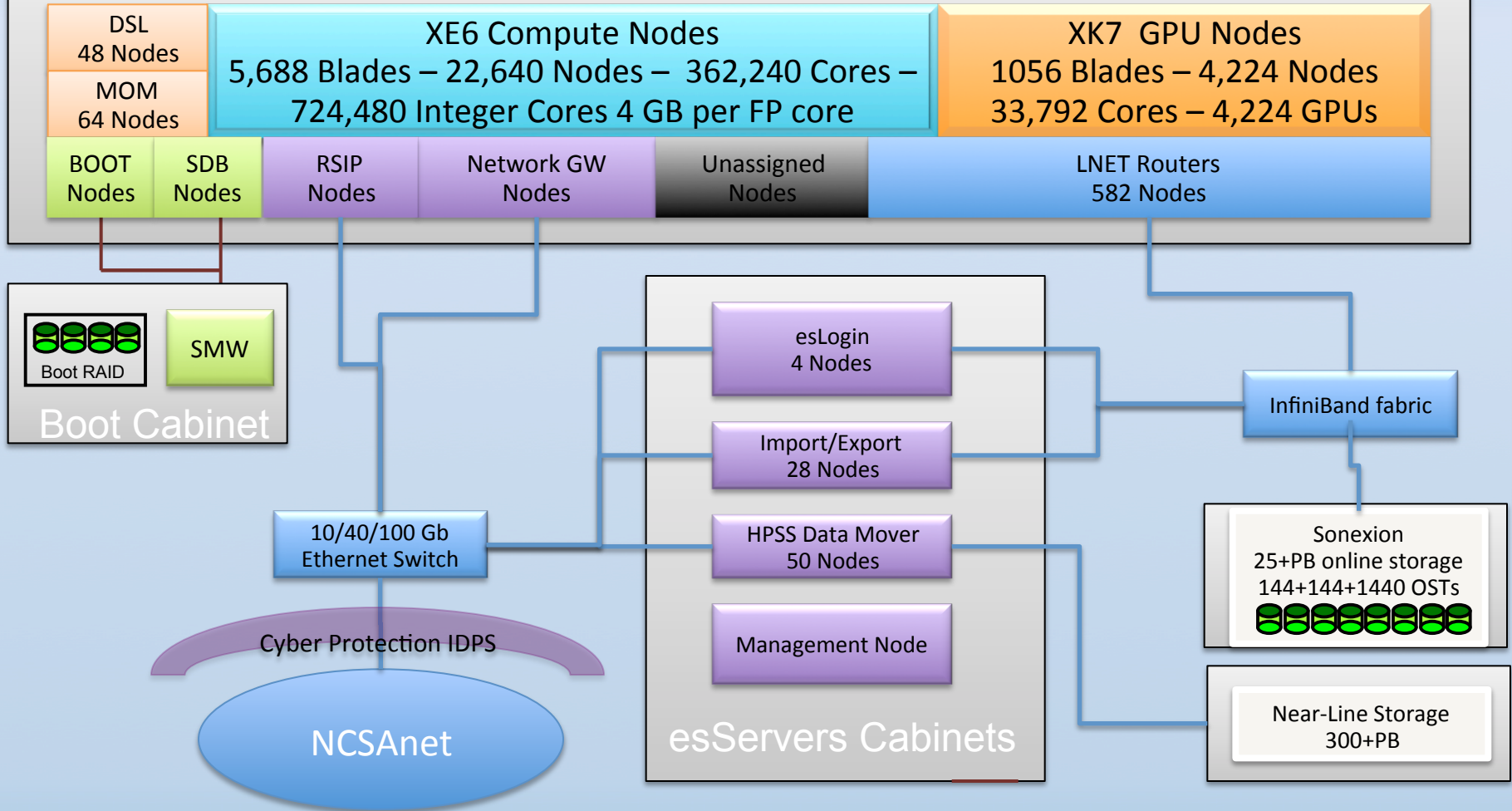
- Modern Data Center
 - 90,000+ ft² total
 - 30,000 ft² 6 foot raised floor
 - 20,000 ft² machine room gallery with no obstructions or structural support elements
- Energy Efficiency
 - LEED certified Gold
 - Power Utilization Efficiency, PUE = 1.1–1.2
 - 24 MW current capacity – expandable
 - Highly instrumented

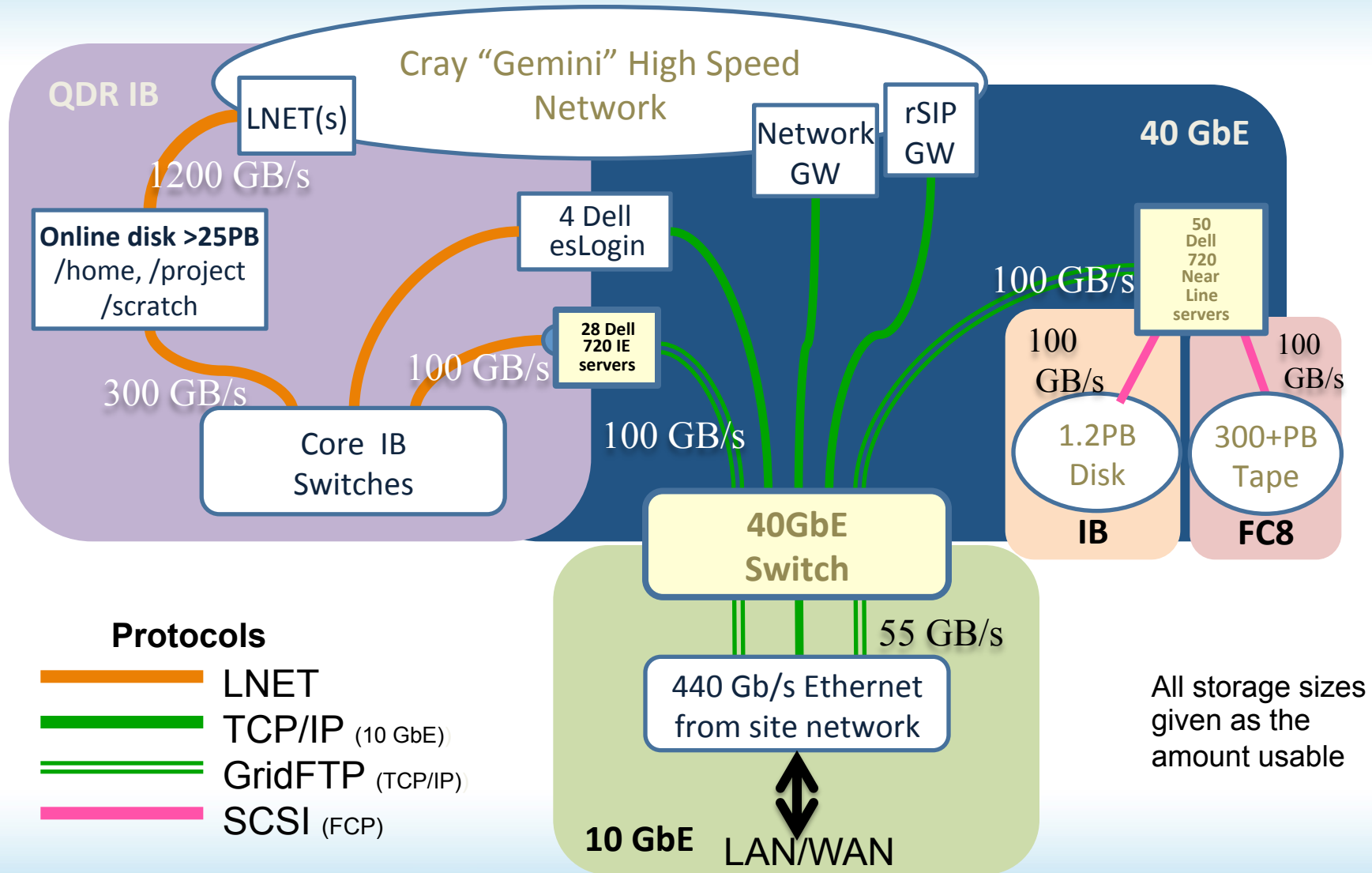
The Movement of Data



Gemini Fabric (HSN)

Cray XE6/XK7 - 276 Cabinets



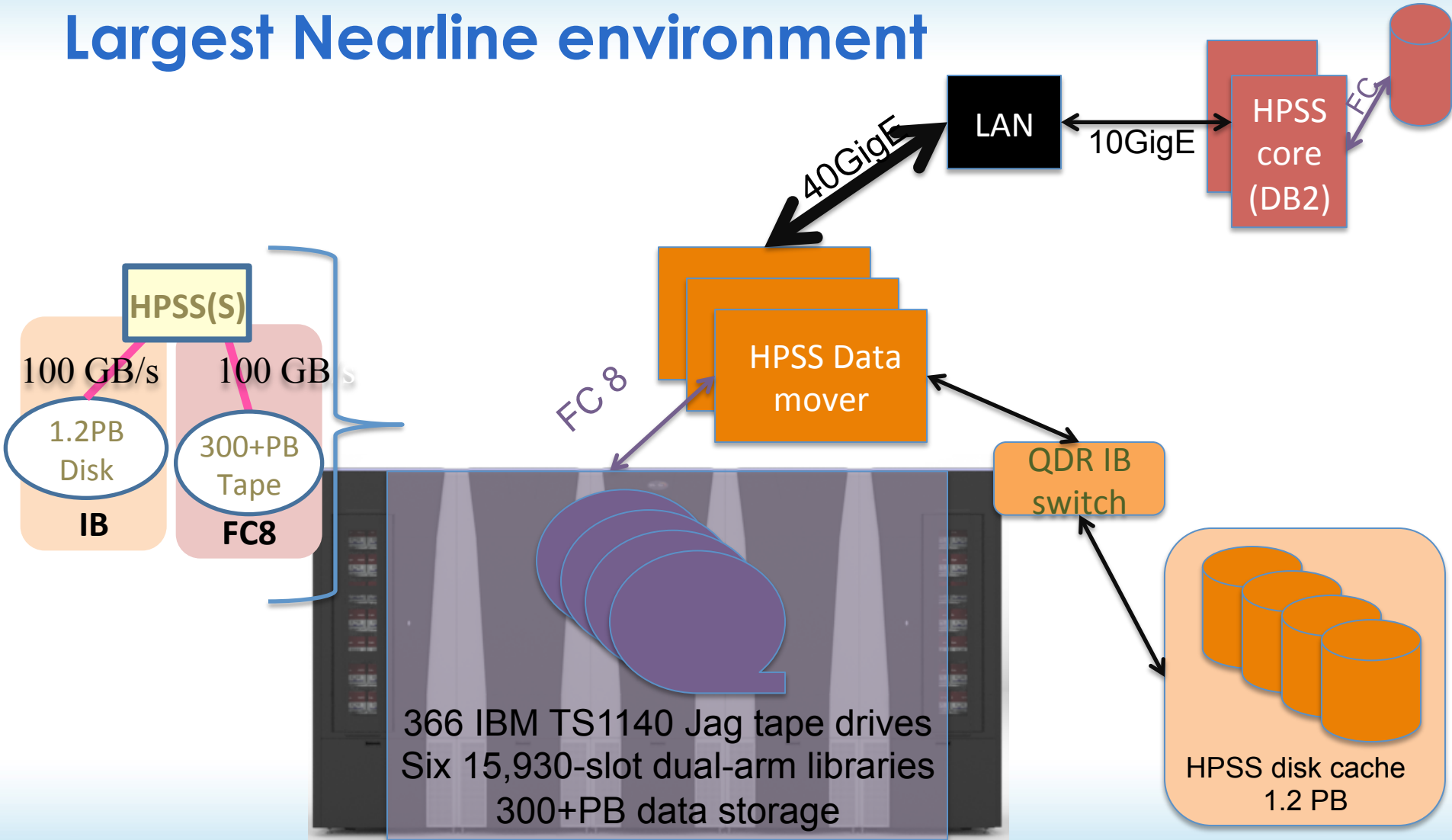


Blue Waters Nearline/Archive System

- 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 Nearline environment



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.
- Home and project file-systems are backed up.

Nearline Storage (HPSS)



home: 5 TB quota



projects: 50 TB group quota

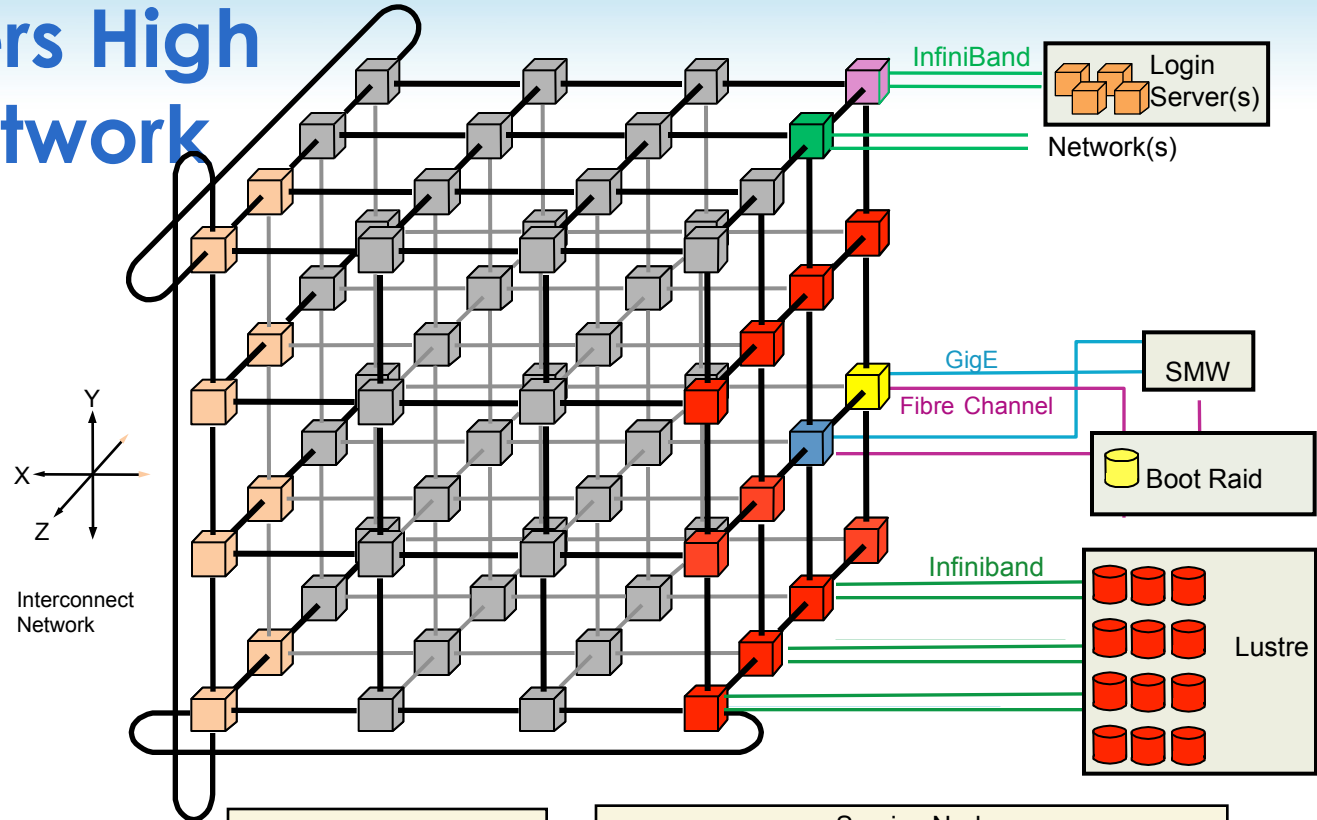
- IBM HPSS + DDN + Spectra Logic.
- Accessed via GO or globus-url-copy.

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

Compute Nodes

- Grey cube: Cray XE6 Compute
- Orange cube: Cray XK7 Accelerator

Service Nodes

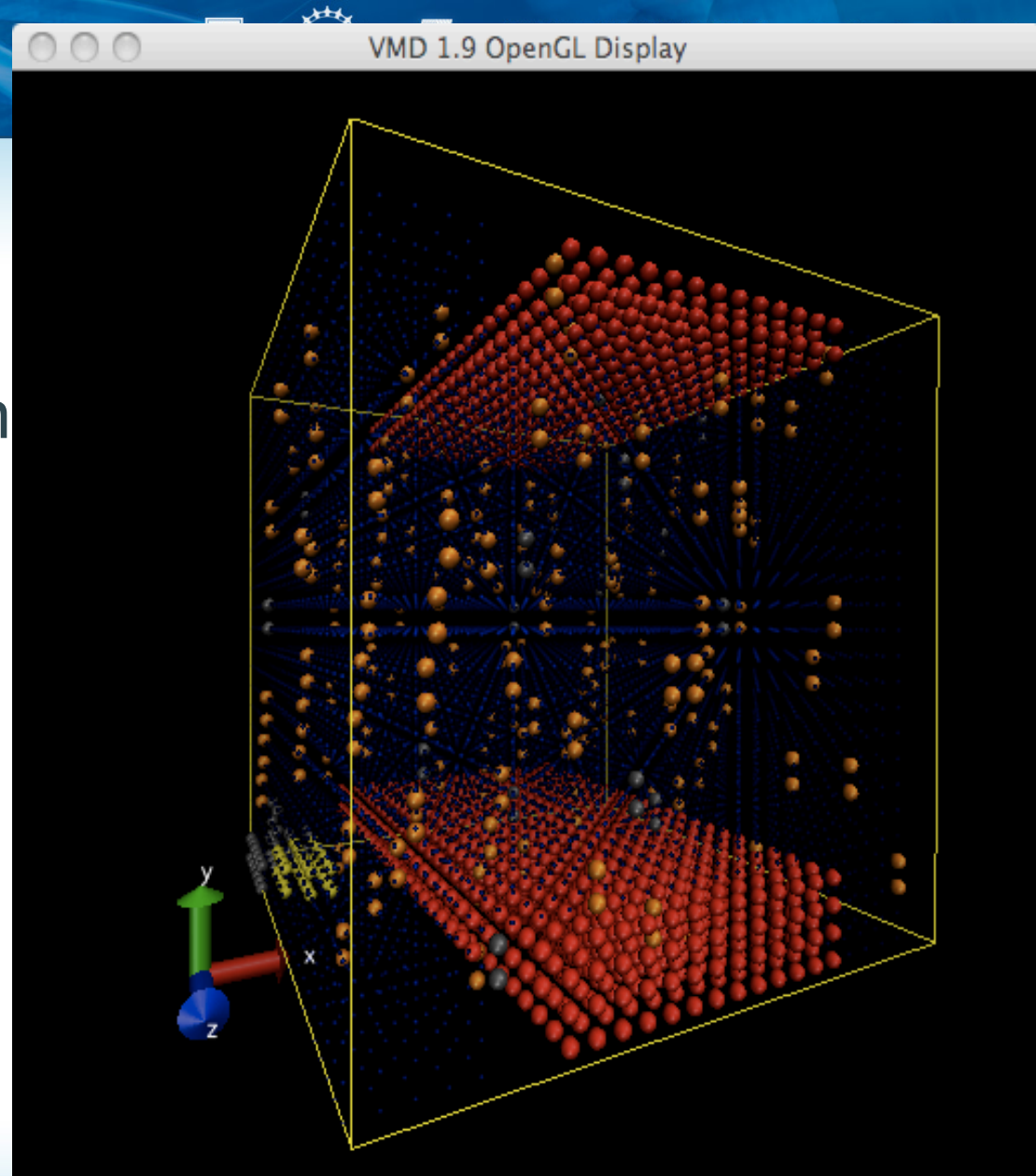
Operating System	Login/Network
Blue cube: Boot	Pink cube: Login Gateways
Yellow cube: System Database	Green cube: Network
Lustre File System	
Red cube: LNET Routers	

HSN View

Gemini-node distinction

- Red – XK
- Orange – LNET
- Yellow – MOM
- Gray – Service
- Blue – XE

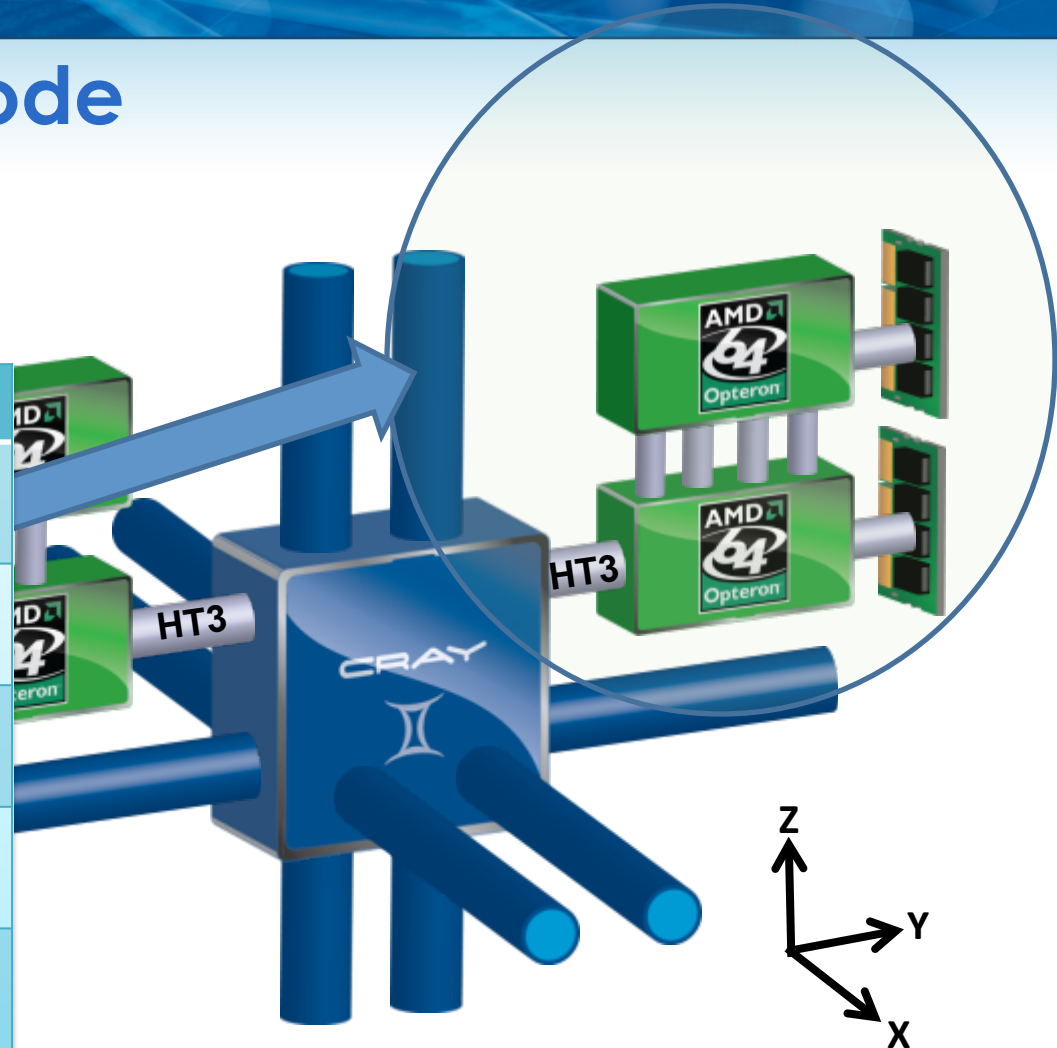
Complicated Topology



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.*

CPU Node Comparison

- Two AMD 6276 Interlagos processors per node. Eight 256-bit SSE FPUs or Bulldozer modules per CPU.
- Each FPU is shared by two (2) "integer" cores.
- The OS and the batch job resource manager currently see each "integer" core as a "processor".
- Allocations are in node-hours to avoid complications. No convertible SU.
- Yes, we did run HPL as part of acceptance.

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

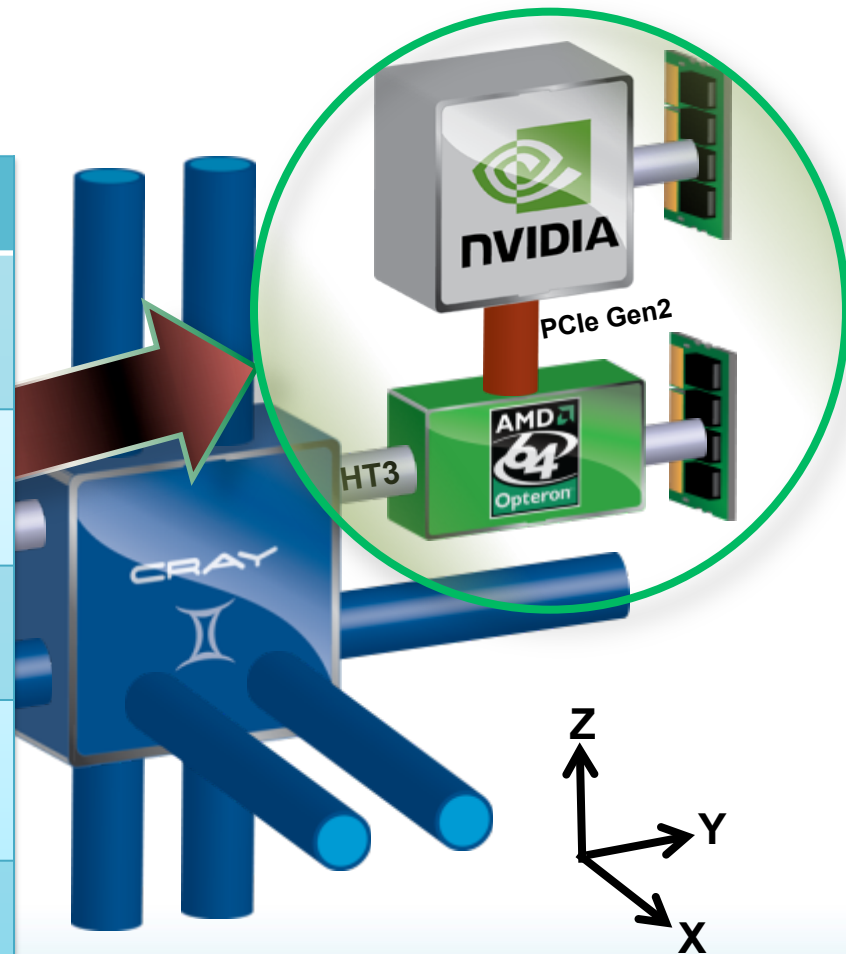
An * indicates processors with 8 flops per clock period.

Cray XK7

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

XK7 Compute Node Characteristics

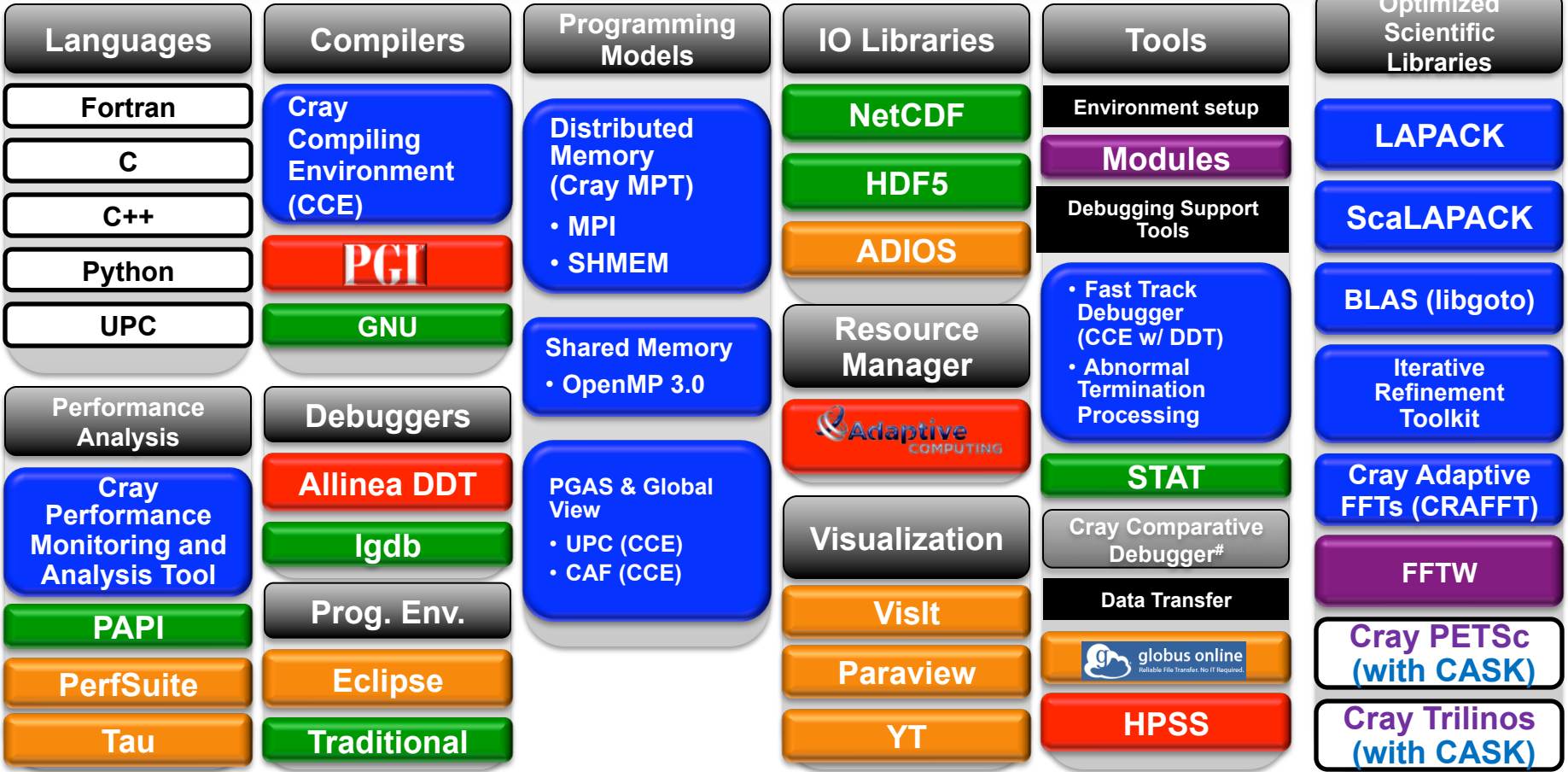
Host Processor	AMD Series 6200 (Interlagos)
Host Processor Performance	156.8 Gflops
Kepler Peak (DP floating point)	1.32 Tflops
Host Memory	32GB 51 GB/sec
Kepler Memory	6GB GDDR5 capacity > 180 GB/sec



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



Cray Linux Environment (CLE)/SUSE Linux

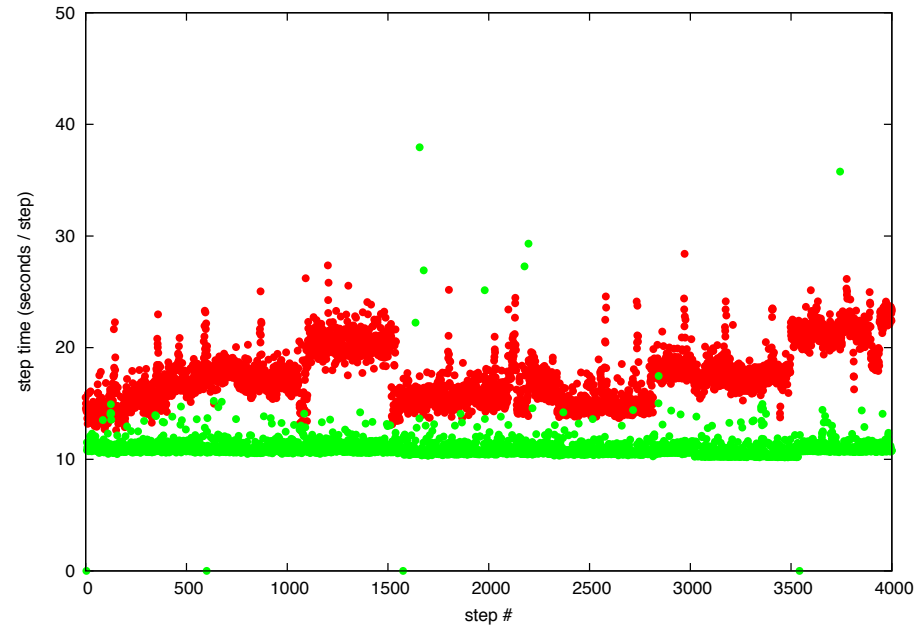
Cray developed	3rd party packaging
Under development	NCSA supported
Licensed ISV SW	Cray added value to 3rd party

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).

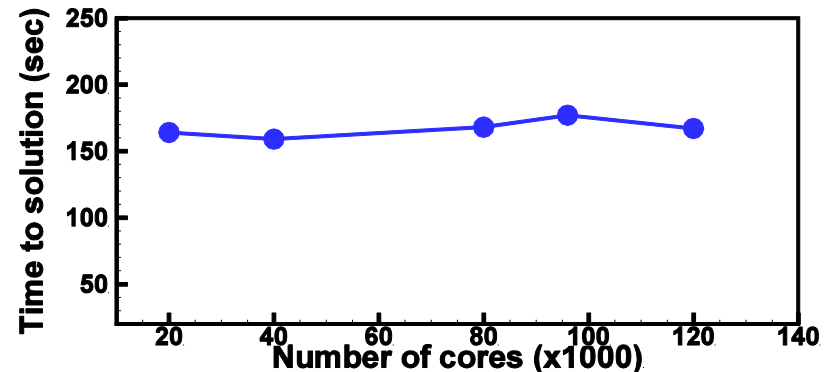
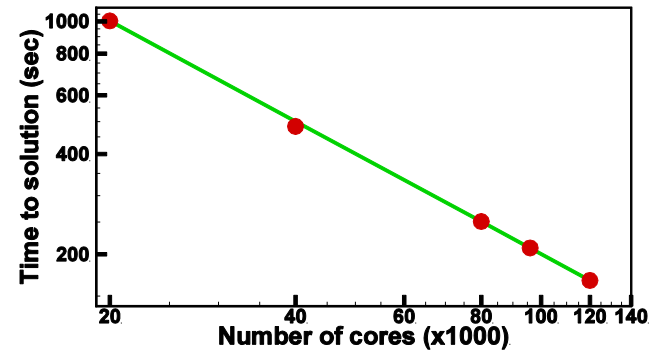
Consistency

- Green job shows consistent step times.
- Red job shows step times impacted by other workloads.
- Both jobs were 8,192 XE nodes.
- Currently requires use of special nodesets. Working with Adaptive to have general, topology-aware scheduling.



Scalability

- Support for both modes.
- Strong scaling
 - Low latency HSN.
 - Low noise OS.
- Weak scaling
 - 4 GB per FP core.
 - Aggregate bandwidth.



Sustained Performance

- 4 applications and 2 benchmarks ran on entire system during acceptance achieving a sustained 1 PF/s.
- Several teams have run production or benchmark workloads on the full XE system.
- Real application IO performance 100-150 GB/s sustained. Benchmarks can hit 800-1,000 GB/s.

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.