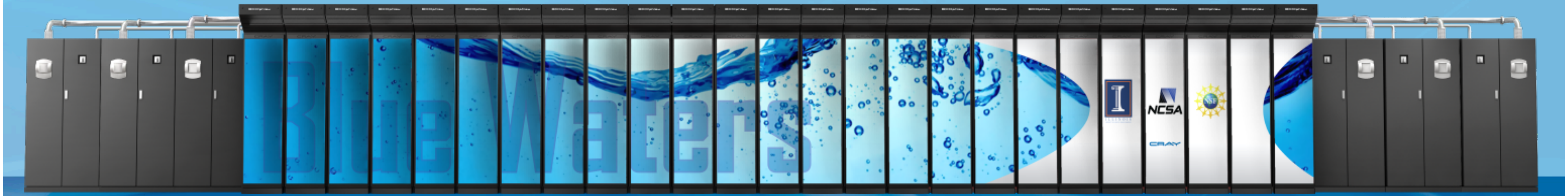
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

4/12/12

Blue Waters Super System

Michelle Butler

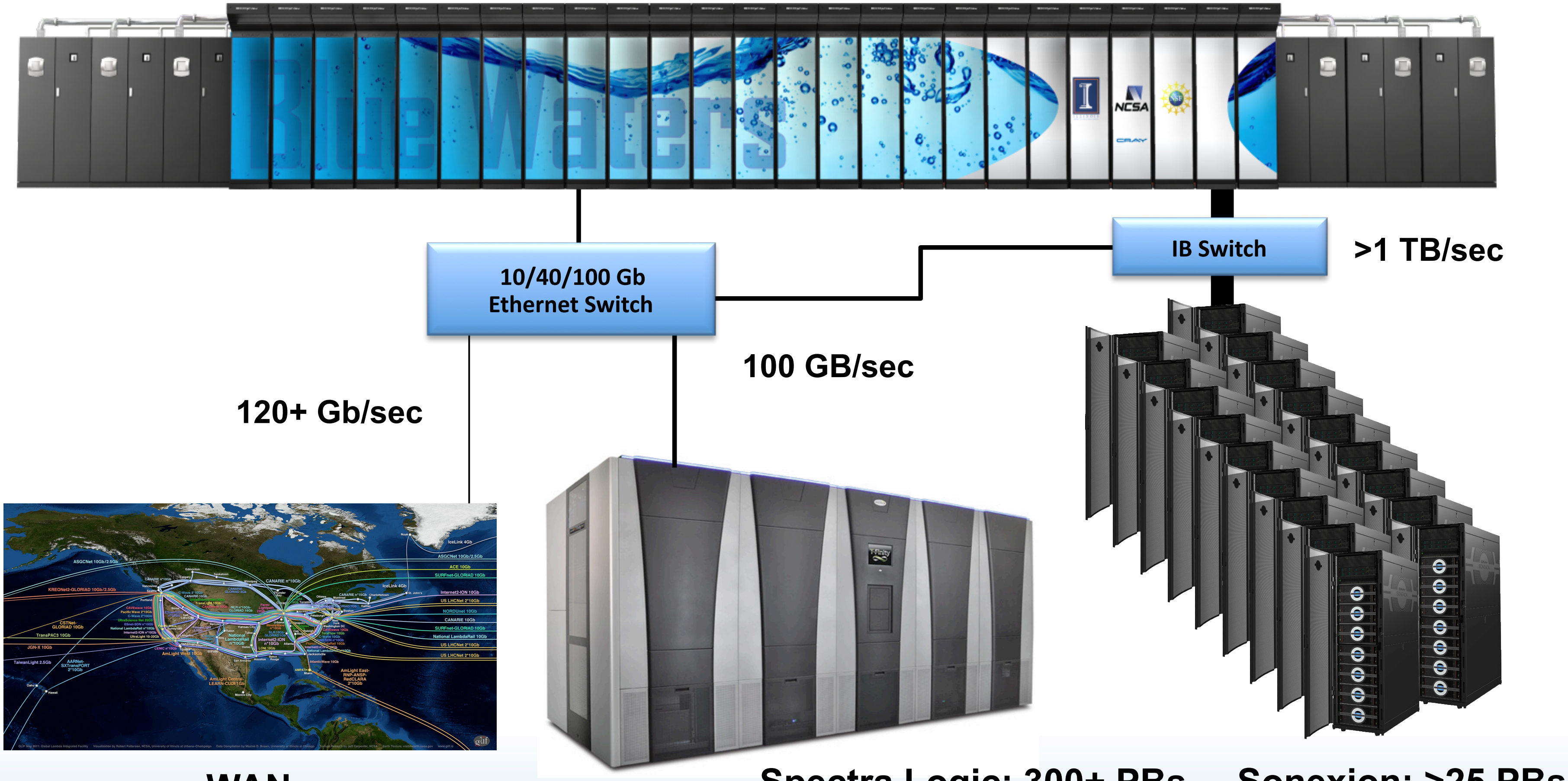


GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

NCSA Has Completed a Grand Challenge

- In August, IBM terminated their contract to deliver the base Blue Waters system
- NSF asked NCSA to propose a change of technology and to adjust the Project Execution Plan for that change
  - Same expectations and goals
  - Same or better schedule
  - Same or lower budget
  - Less Risk
- In September, NCSA proposed a revised plan to NSF and a Peer Review Panel. - 27 Days!!
  - Complete understanding of applications was key to being able to do this
- NSF approved the plan on November 10, 2011
- All previous goals of the project will be met with the new system

Blue Waters Computing Super-system



10/40/100 Gb Ethernet Switch

IB Switch

>1 TB/sec

100 GB/sec

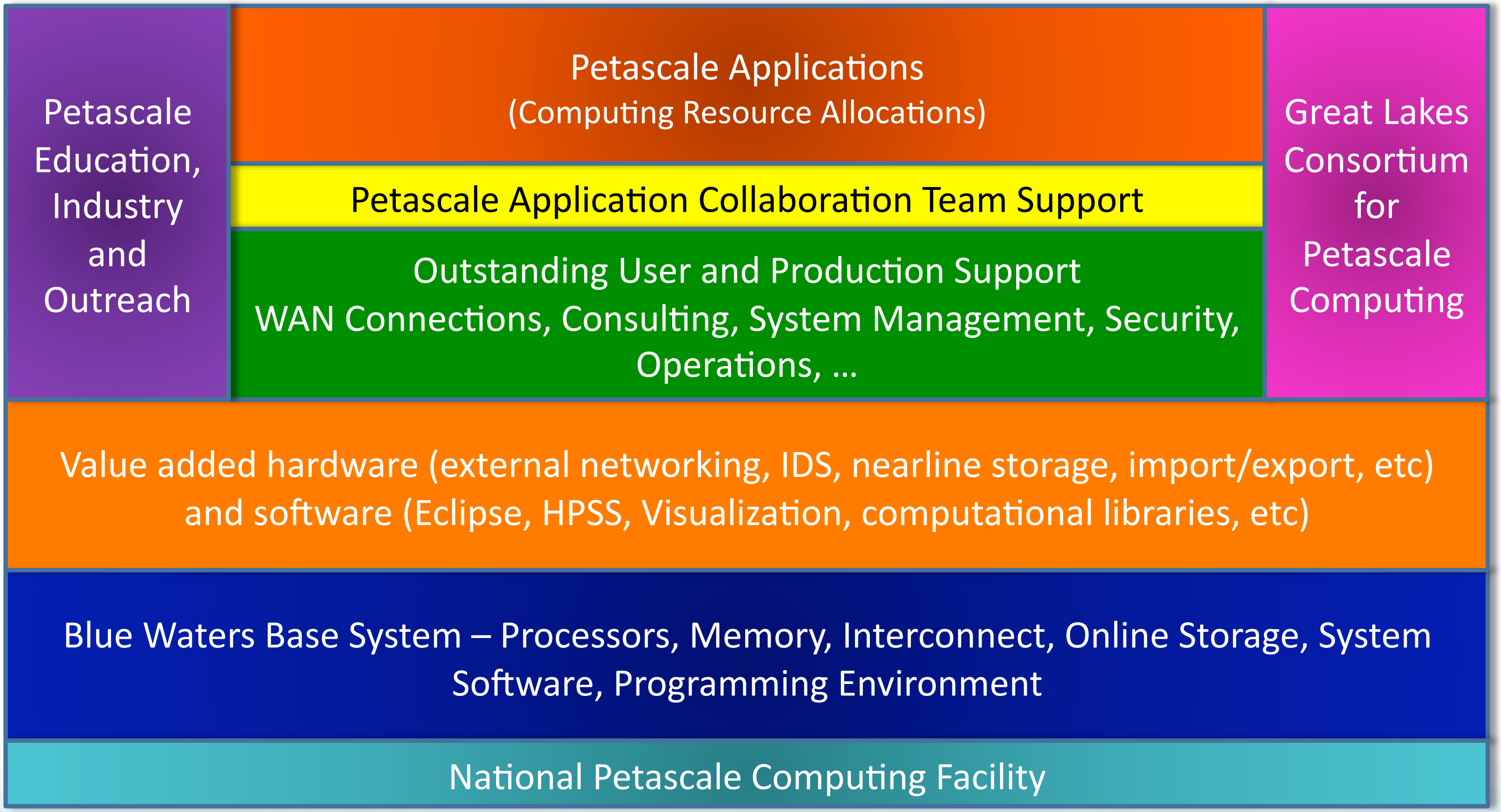
120+ Gb/sec

WAN

Spectra Logic: 300+ PBs

Sonexion: >25 PBs

The Blue Waters EcoSystem



Petascale Education, Industry and Outreach

Petascale Applications
(Computing Resource Allocations)

Petascale Application Collaboration Team Support

Outstanding User and Production Support
WAN Connections, Consulting, System Management, Security, Operations, …

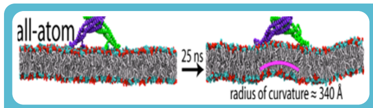
Great Lakes Consortium for Petascale Computing

Value added hardware (external networking, IDS, nearline storage, import/export, etc) and software (Eclipse, HPSS, Visualization, computational libraries, etc)

Blue Waters Base System – Processors, Memory, Interconnect, Online Storage, System Software, Programming Environment

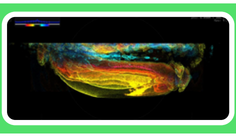
National Petascale Computing Facility

More than 25 PRAC science teams
12 distinct research fields
selected to run on the new Blue Waters
Expect ~10 more major teams

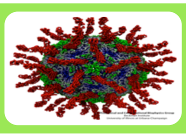


Nanotechnology

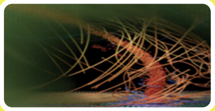
Astronomy/Astrophysics



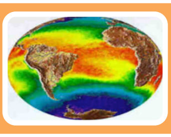
Earthquakes and the damage they cause



Viruses entering cells



Severe storms



Climate change

Additional software environments worked on by BW staff

Project		Project Description
Integrated System Console		Integrated monitoring and analysis system for the Blue Waters system
Storage & Archive Software		-- RAIT for tape technology to reduce cost, maintain reliable long-term data storage -- Integrated wide-area data transfer technology with the Blue Waters system
Workflow		Integrated and enhanced workflow system for Blue Waters super-system
Computational Libraries		Enhanced performance of computational libraries important to Blue Waters Science Teams
Performance Tools (RENCI)	✓	Ported and enhanced performance of scalable performance measurement tools for Blue Waters
Cactus (LSU)	✓	Cactus build system plugin for the Eclipse application development environment
Photran	✓	Full Fortran 2008 syntax support for Eclipse application development environment
Eclipse IADE		Integrated application development environment for use by Science Team
Software Tools		Integrated and enhanced key third party software tools for Blue Waters environment
Visualization Software		Port key visualization software packages needed by Science Teams
Programming Models		Benchmark and functionality test suite for traditional and PGAS programming models
Compiler Benchmarks		Benchmark test suites to evaluate key aspects of Blue Waters compilers

National Petascale Computing Facility

- Modern Data Center
  - 90,000+ ft^2 total
  - 30,000 ft^2 raised floor
    20,000 ft^2 machine room gallery
- Energy Efficiency
  - LEED certified Gold
  - Power Utilization Efficiency, PUE = 1.1–1.2

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

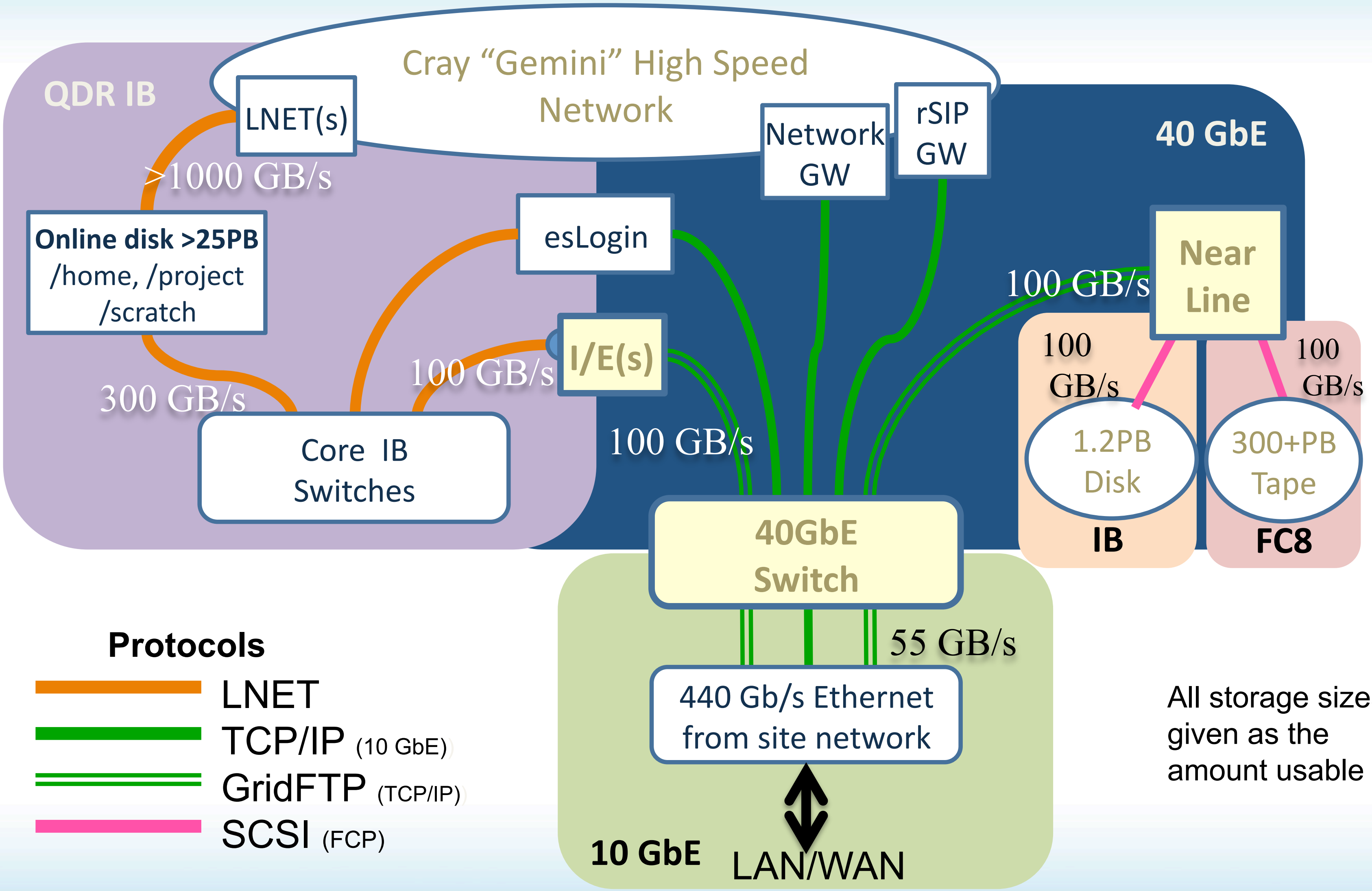
Cray System & Storage cabinets:	• >300
Compute nodes:	• >25,000
Usable Storage Bandwidth:	• >1 TB/s
System Memory:	• >1.5 Petabytes
Memory per core module:	• 4 GB
Gemini Interconnect Topology:	• 3D Torus
Usable Storage:	• >25 Petabytes
Peak performance:	• >11.5 Petaflops
Number of AMD processors:	• >49,000
Number of AMD x86 core module:	• >380,000
Number of NVIDIA GPUs:	• >3,000

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Blue Waters Goals

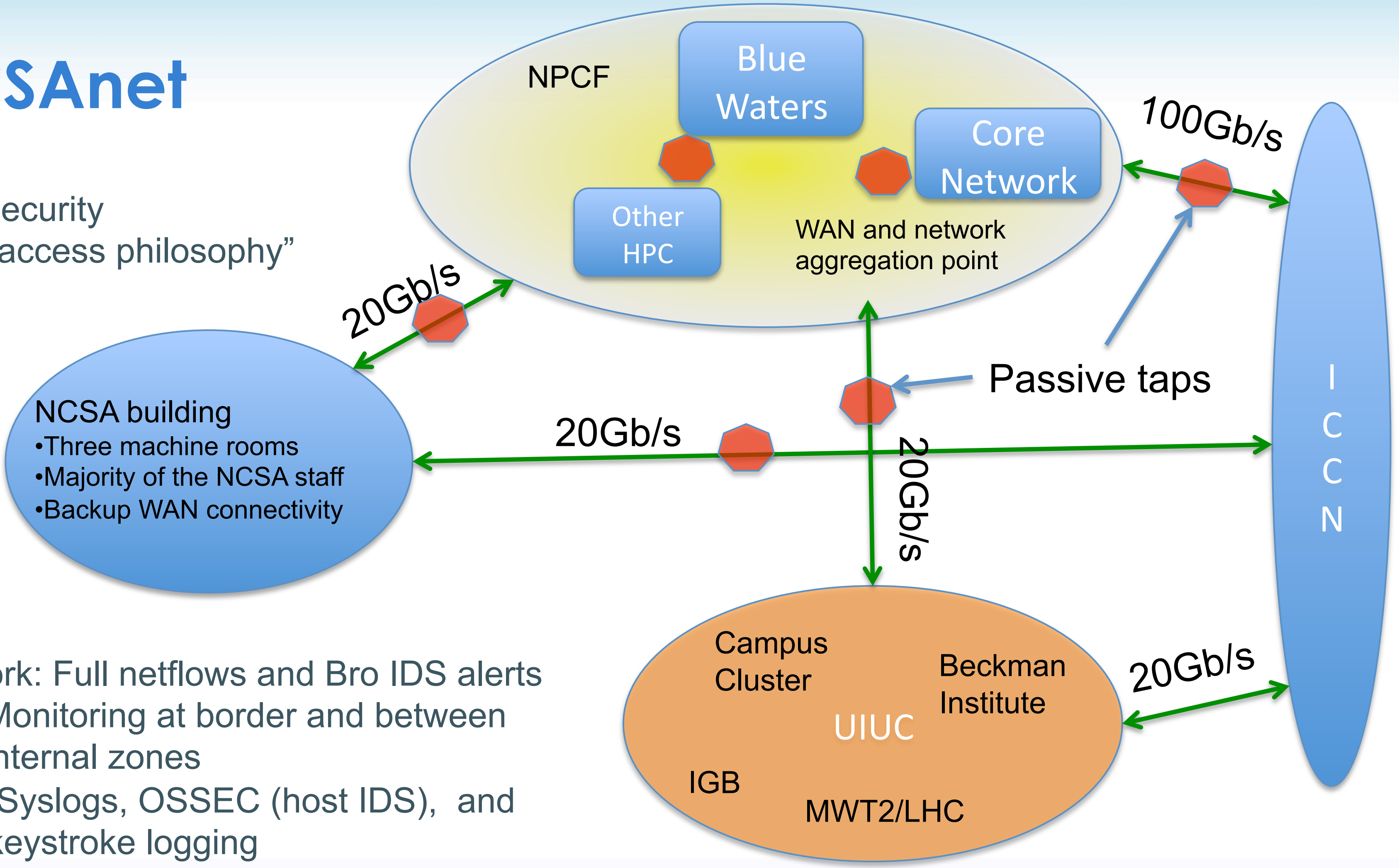
- **To deploy a computing system capable of sustaining one petaflops or more for a broad range of applications**
  - Cray system achieves this goal using a well defined metrics
- **To enable the Science Teams to take full advantage of the sustained petascale computing system**
  - Blue Waters Team has established strong partnership with Science Teams, helping them to improve the performance and scalability of their applications
- **To 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
- **To 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)
- **To 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



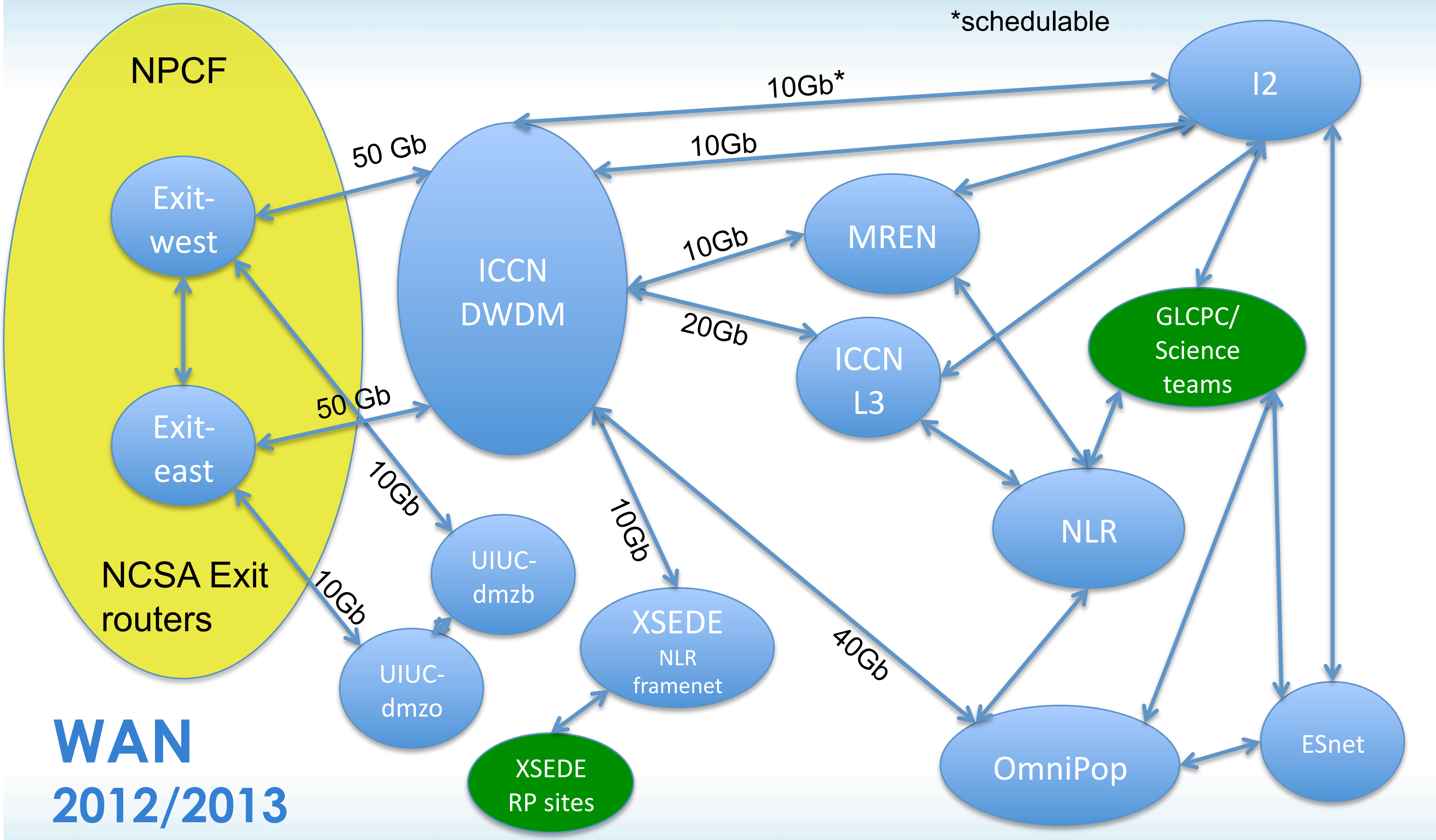
All storage sizes given as the amount usable

NCSAnet

Cybersecurity
“Open access philosophy”



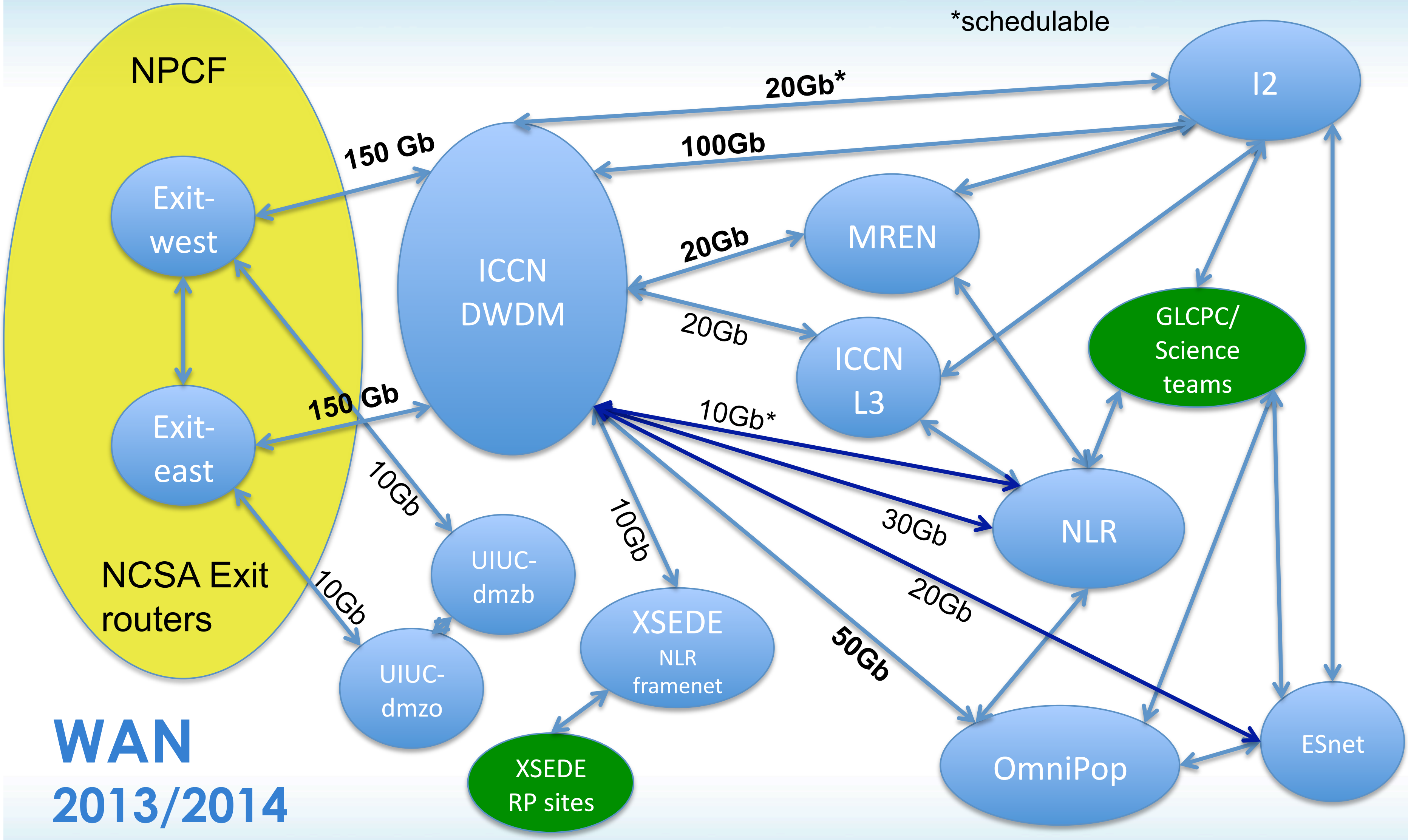
Network: Full netflows and Bro IDS alerts
Monitoring at border and between internal zones
Host: Syslogs, OSSEC (host IDS), and SSH keystroke logging



WAN 2012/2013

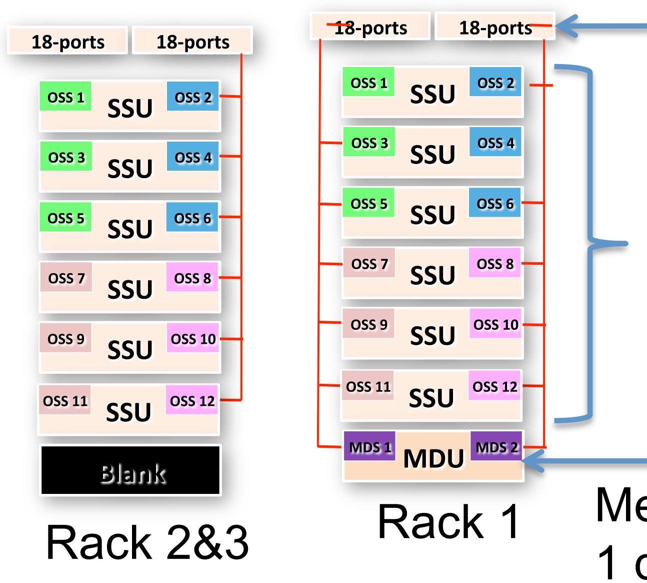
WAN 2013/2014

*schedulable



Storage environments specific

SonExion(Xyratek) Lustre 2.1 Environment



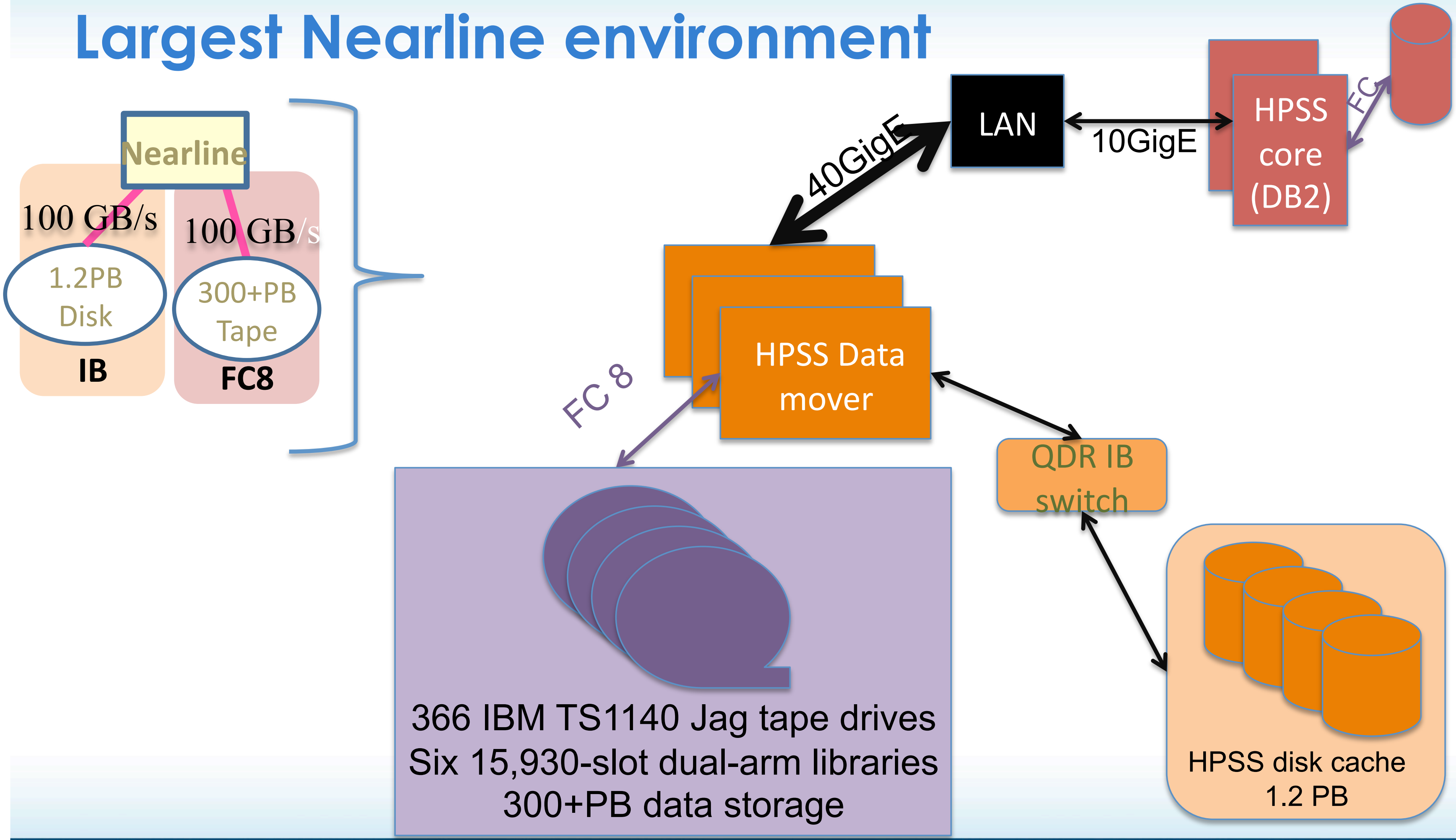
IB switches

6 drawers of 80 2TB disks each
12 Lustre servers; 2 for each drawer or SSU
Failover works!

Metadata unit with Lustre MDS and MGS
1 drawer of 24 drives

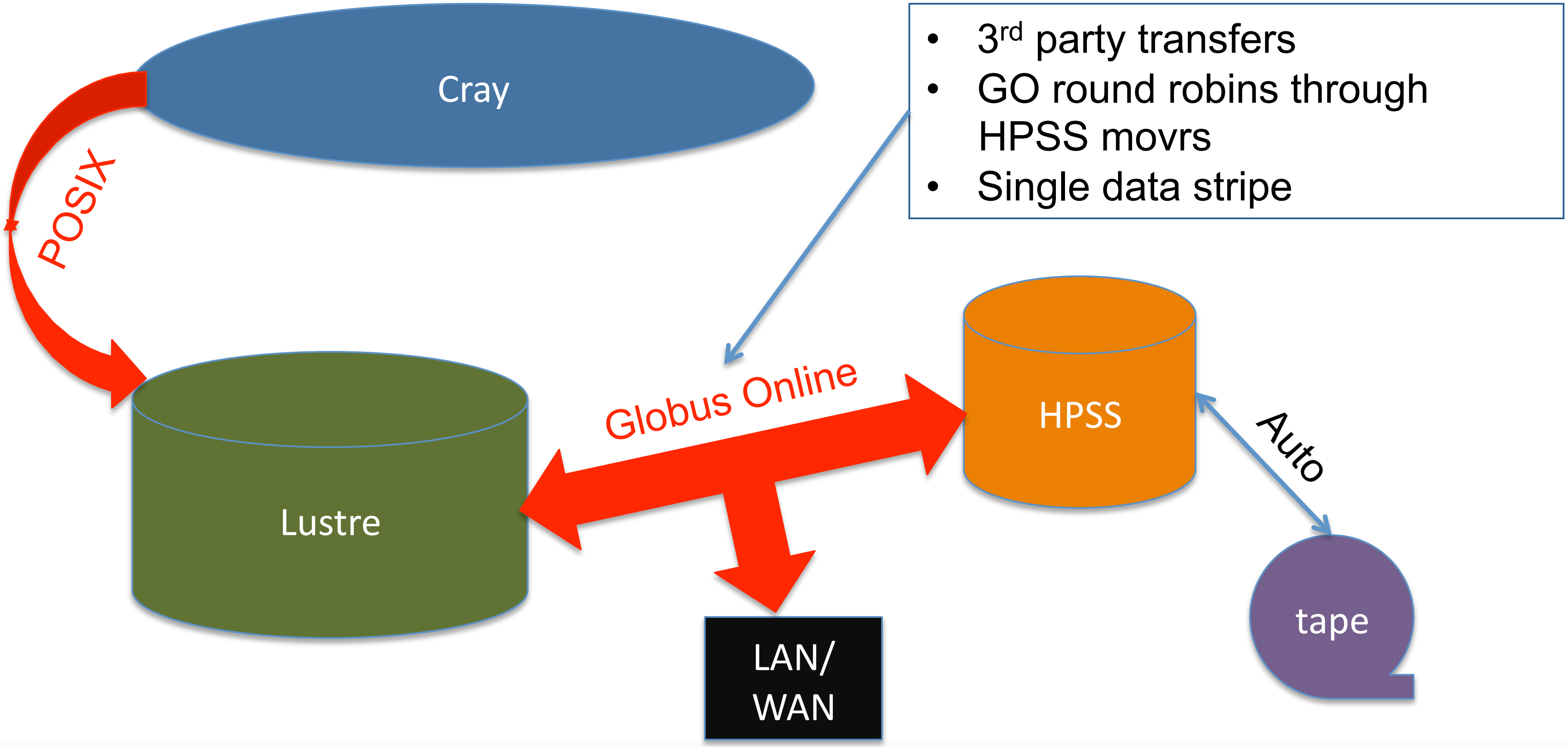
/home = 2PB 3 racks , /project=2PB 3 racks , /scratch=21PB 30 racks

Largest Nearline environment



366 IBM TS1140 Jag tape drives
Six 15,930-slot dual-arm libraries
300+PB data storage

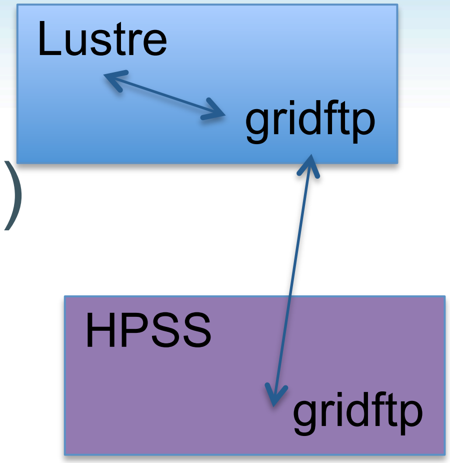
Data Movement today



- 3rd party transfers
- GO round robins through HPSS movrs
- Single data stripe

Problem: No network affinity. Data hops between HPSS nodes before going out

Future Gridftp server



Lustre

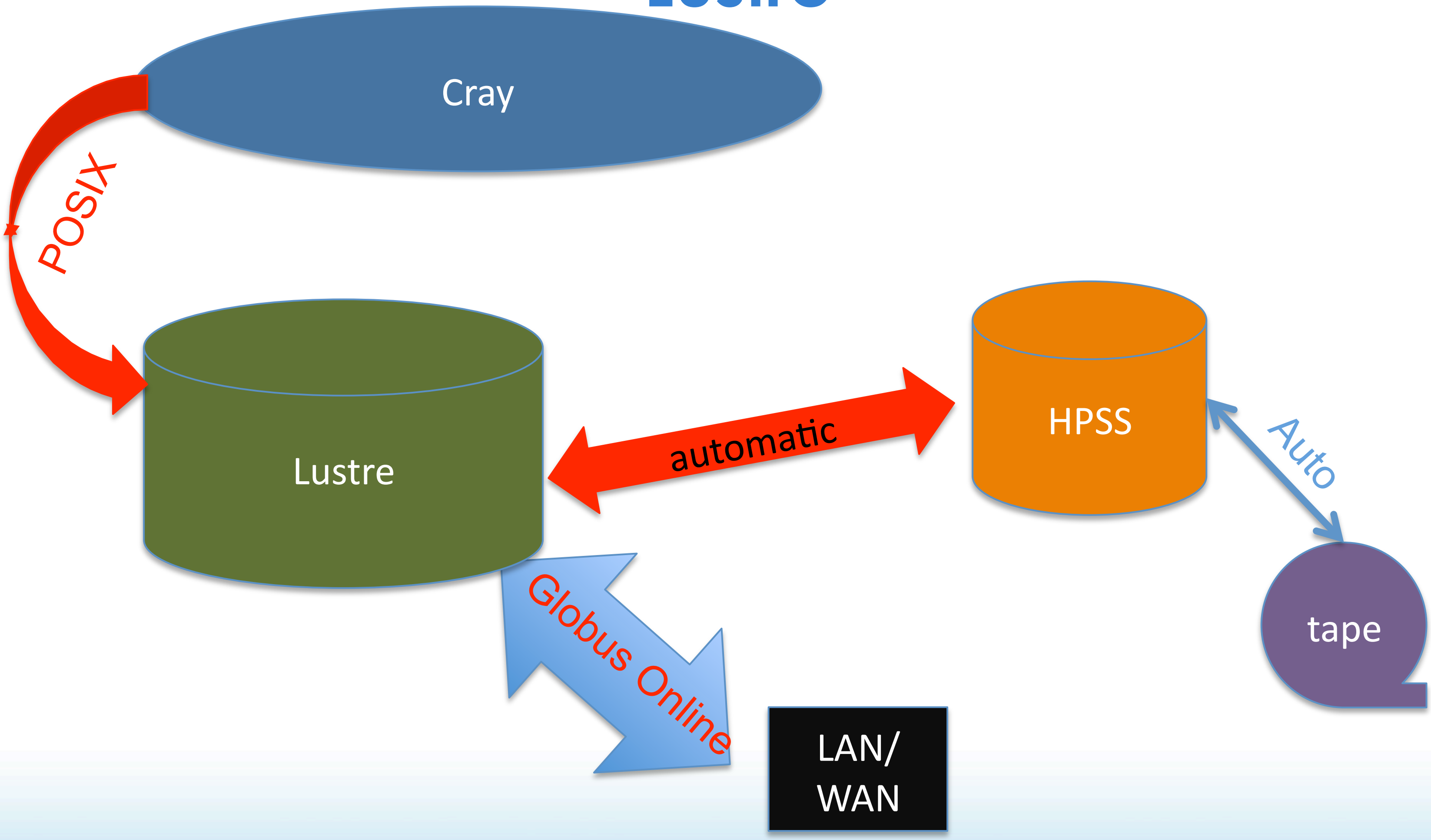
gridftp

HPSS

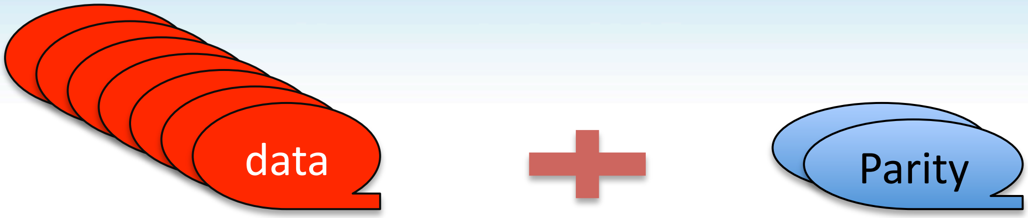
gridftp

- Data Storage Interface within gridftp (DSI)
  - HPSS talks directly to Lustre gridftp
    - Eliminates the affinity issue for intra-BW transfers
  - Striped gridftp to match HPSS stripe COS (RAIT)
  - Affinity for gridftp/HPSS for external transfers
    - Can't always know location of movr process before hand
  - RAIT Engine Affinity
  - Aggregation of data

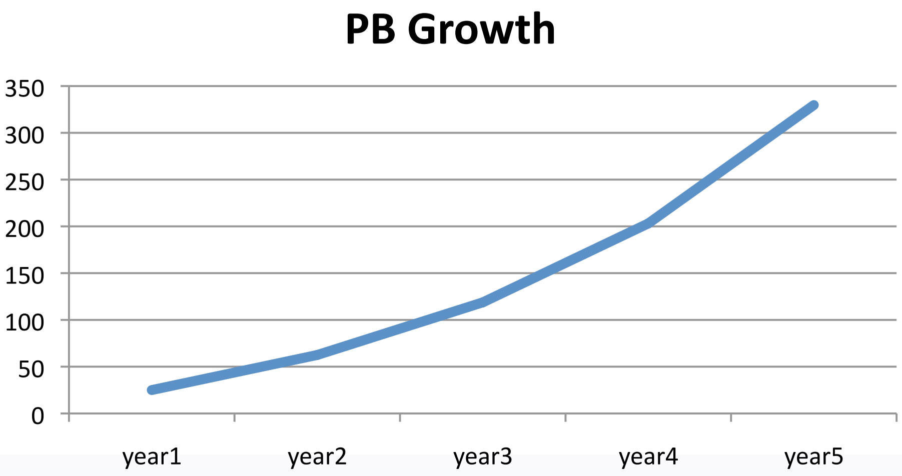
Data Movement Future All Namespace in Lustre

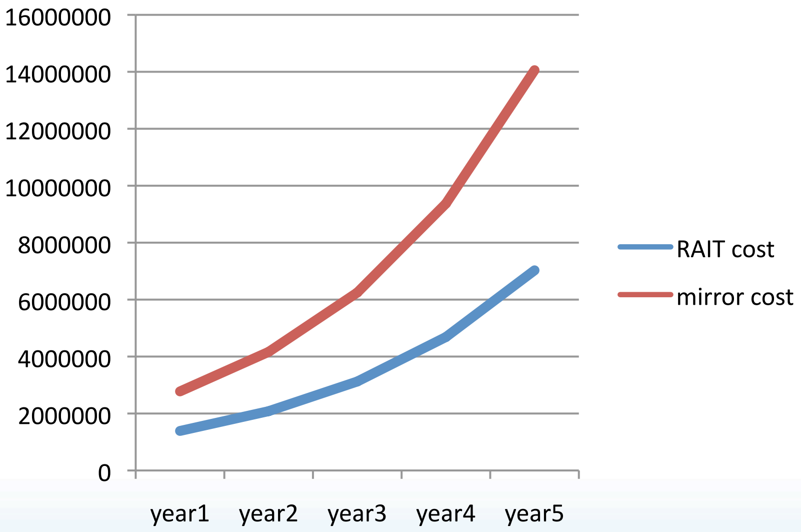


Data integrity



- Archive HPSS - RAIT on tape
  - Striped data written to tape with parity stripe to ensure that media loss is not user data loss.





RAIT Requirements

- Multiple data classes depending on file size to take advantage of tape writing strategies
    - 4+1, 7+2, 10+2, etc
    - Trade off latency vs performance – Depends on the size of the data
- Multiple levels of RAIT part of the requirements
    - Parity can not be wider than the data.
    - Can not have more than 8 levels of parity
    - If a tape failure occurs the read will continue and then will be flagged for repack.
    - If a write fails, it can continue (configurable) based on site policy and then flagged for repack

Tentative Schedule

- Phase 0 - Test and Development Rack delivered December 1, 2011
- Science Team Workshop December 13-16, 2011
- Phase 1 - 48 XE racks + 2 PB of storage arrive in late January
  - Expect PRAC early science access starting in early March – limited number of teams selected from the PRAC set
- Phase 2 – All racks and storage and software installed
  - Kepler accelerator modules may not yet be installed.
  - Expect Limited Use access for all science teams in mid-late summer (July/Aug??)
- Phase 3 – All components installed and accepted – Full Service for all teams
  - Early-Mid Fall 2011

NOTE – These are internal targets – official project schedule is to complete the deployment by March 2013

Blue Waters Early Science System

- BW-ESS Configuration
  - 48 cabinets, 4,512 XE6 compute nodes, 96 service nodes
  - 2 PBs Sonexion Lustre storage appliance
- Access through Blue Waters Portal
  - https://bluewaters.ncsa.illinois.edu/

- Current Projects
  - **Biomolecular Physics**—K. Schulten, University of Illinois at Urbana-Champaign
  - **Cosmology**—B. O'Shea, Michigan State University
  - **Climate Change**—D. Wuebbles, University of Illinois at Urbana-Champaign
  - **Lattice QCD**—R. Sugar, University of California, Santa Barbara
  - **Plasma Physics**—H. Karimabadi, University of California, San Diego
  - **Supernovae**—S. Woosley, University of California Observatories
  - *Three more projects held in reserve*

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.