BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Blue Waters – with a focus on science data and networks

William Kramer

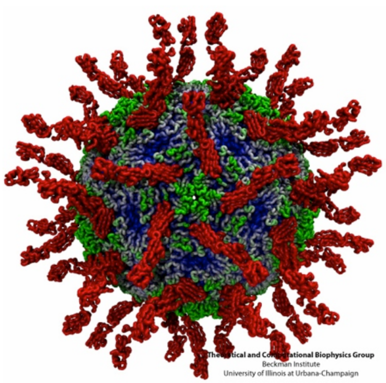
National Center for Supercomputing Applications, University of Illinois

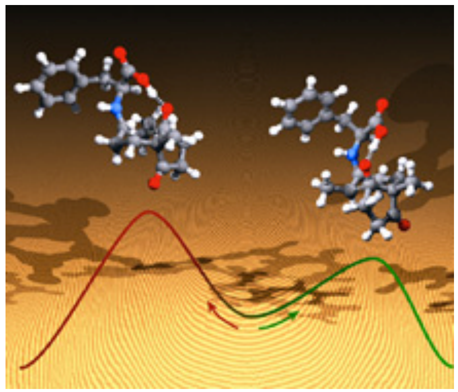
NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Science & Engineering on Blue Waters

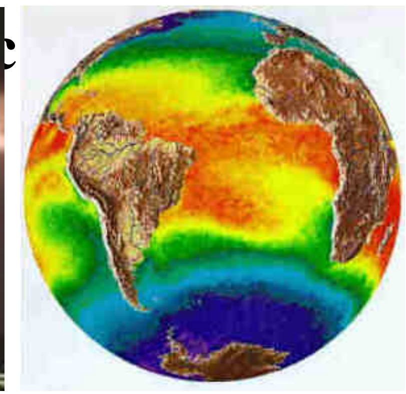
Blue Waters will enable advances in a broad range of science and engineering disciplines. Examples include:

Molecular Science



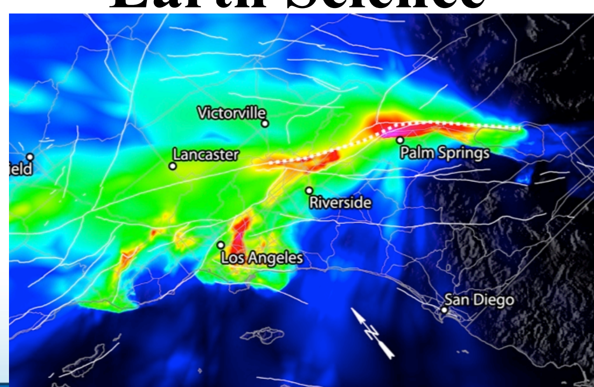


Weather & Climate



Astro*

Earth Science



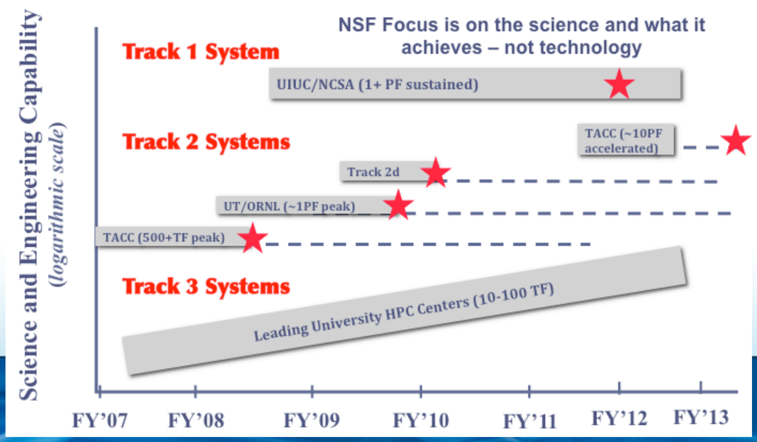
Health

NSF's Track 1 Solicitation

"The petascale HPC environment will enable investigations of computationally challenging problems that require **computing systems** capable of delivering **sustained performance approaching 10^{15} floating point operations per second** (petaflops) **on real applications**, that consume **large amounts of memory**, and/or that work with **very large data sets**."

Leadership-Class System Acquisition - Creating a Petascale Computing Environment for Science and Engineering

NSF 06-573



Blue Waters Project Components

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, …

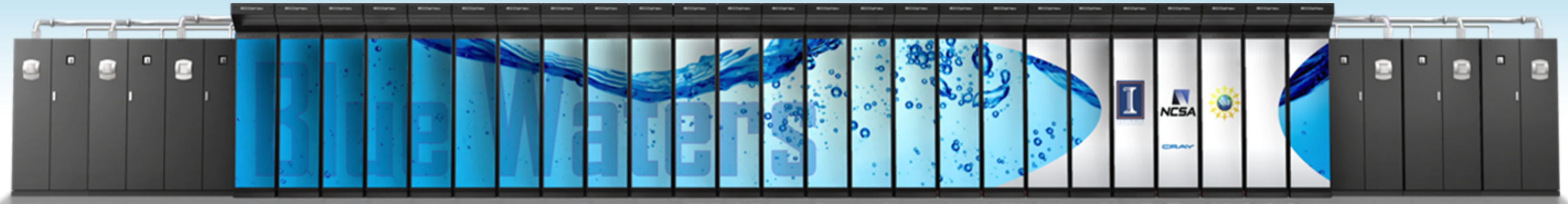
Great Lakes Consortium for Petascale Computing

Value added Software – Collaborations

Value added hardware and software (HPSS, near-line HW, LAN/WAN Networking, etc.)

Blue Waters Base System – Processors, Memory, Interconnect, On-line Storage, System Software, Programming Environment

Petascale Computing Facility



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
Gemin 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 SCIENCE TEAMS

As of July 2012

NSF PRAC Major Science Teams

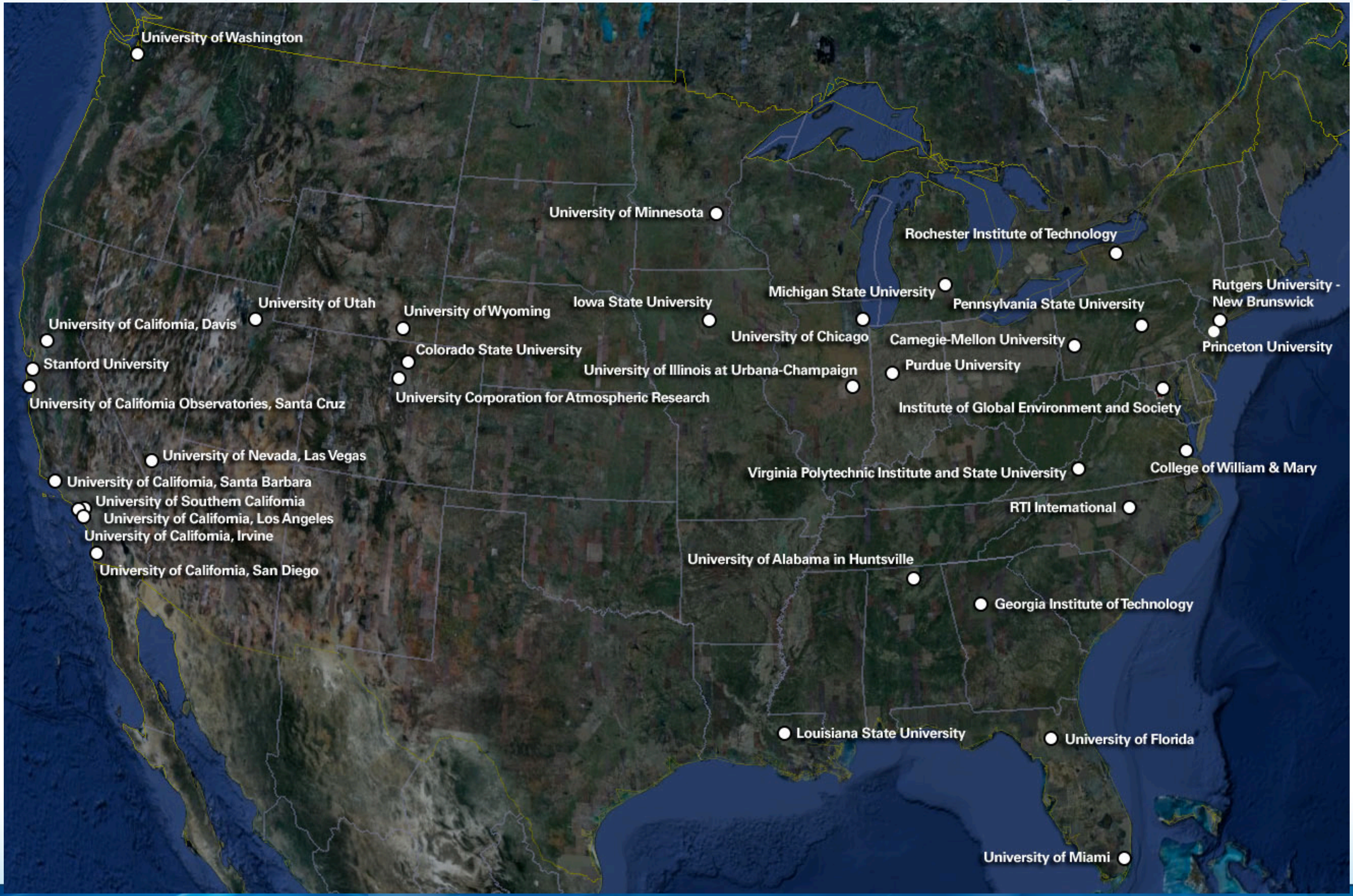
PI	Award Date	Project Title
Sugar	04/15/2009	Lattice QCD on Blue Waters
Bartlett	04/15/2009	Super instruction architecture for petascale computing
Nagamine	04/15/2009	Peta-Cosmology: galaxy formation and virtual astronomy
Bissett	05/01/2009	Simulation of contagion on very large social networks with Blue Waters
O'Shea	05/01/2009	Formation of the First Galaxies: Predictions for the Next Generation of Observatories
Schulten	05/15/2009	The computational microscope
Stan	09/01/2009	Testing hypotheses about climate prediction at unprecedented resolutions on the NSF Blue Waters system
Campanelli	09/15/2009	Computational relativity and gravitation at petascale: Simulating and visualizing astrophysically realistic compact binaries
Yeung	09/15/2009	Petascale computations for complex turbulent flows
Schnetter	09/15/2009	Enabling science at the petascale: From binary systems and stellar core collapse To gamma-ray bursts
Woodward	10/01/2009	Petascale simulation of turbulent stellar hydrodynamics
Tagkopoulos	10/01/2009	Petascale simulations of Complex Biological Behavior in Fluctuating Environments
Wilhelmson	10/01/2009	Understanding tornadoes and their parent supercells through ultra-high resolution simulation/analysis
Wang	10/01/2009	Enabling large-scale, high-resolution, and real-time earthquake simulations on petascale parallel computers
Jordan	10/01/2009	Petascale research in earthquake system science on Blue Waters
Zhang	10/01/2009	Breakthrough peta-scale quantum Monte Carlo calculations
Haule	10/01/2009	Electronic properties of strongly correlated systems using petascale computing
Lamm	10/01/2009	Computational chemistry at the petascale

NSF PRAC Major Science Teams (cont)

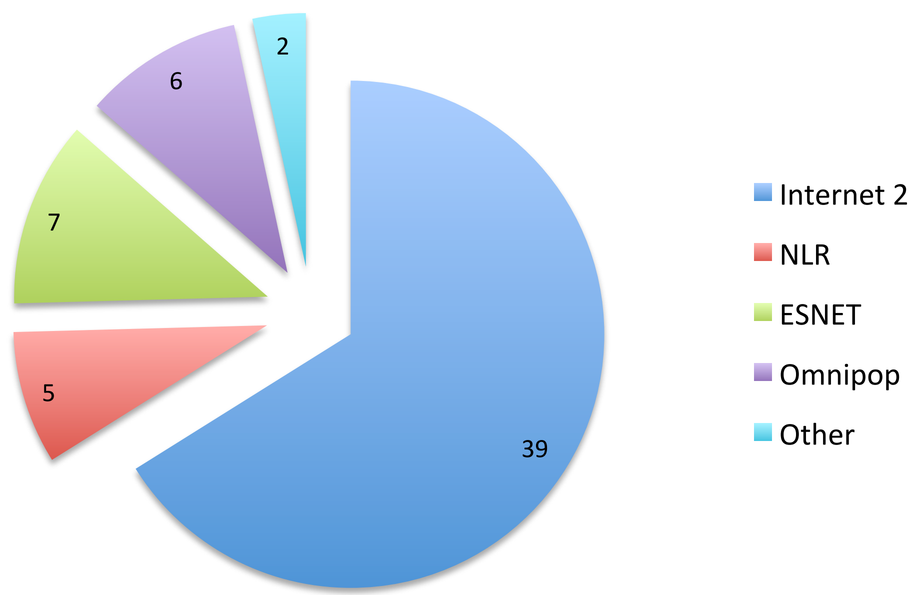
PI	Award Date	Project Title
Karimabadi	11/01/2010	Enabling Breakthrough Kinetic Simulations of the Magnetosphere via Petascale Computing
Mori	01/15/2011	Petascale plasma physics simulations using PIC codes
Voth	02/01/2011	Petascale multiscale simulations of biomolecular systems
Woosley	02/01/2011	Type Ia supernovae
Cheatham	02/01/2011	Hierarchical molecular dynamics sampling for assessing pathways and free energies of RNA catalysis, ligand binding, and conformational change
Wuebbles	04/15/2011	Using petascale computing capabilities to address climate change uncertainties
Gropp	06/01/2011	System software for scalable applications
Klimeck	09/15/2011	Accelerating nano-scale transistor innovation
Pande	09/15/2011	Simulating vesicle fusion on Blue Waters
Elghobashi	05/18/2012	Direct Numerical Simulation of Fully Resolved Vaporizing Droplets in a Turbulent Flow
Quinn	05/18/2012	Evolutions of the Small Galaxy Populations From High Redshift to the Present
Wood/Reed	06/12/2012	Collaborative Research: Petascale Design and Management of Satellite Assets to Advance Space Based Earth Science
Pogorelov	0613/2012	Modeling Heliophysics and Astrophysics Phenomena with a Multi-Scale Fluid Kinetic Simulation Suite
Bernholc	07/15/2012	Petascale quantum simulations of nano systems and biomolucles
Stein	08/01/2012	Ab Initio Models of Solar Activity

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

Blue Waters PRAC PI Institutions – June 2012



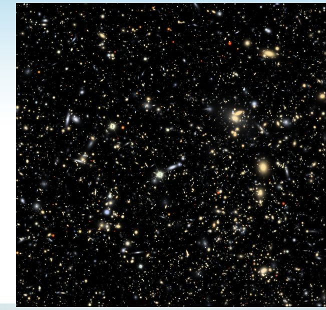
Science Team Backbone distribution

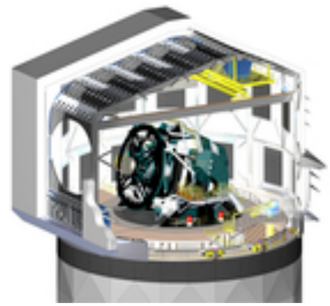


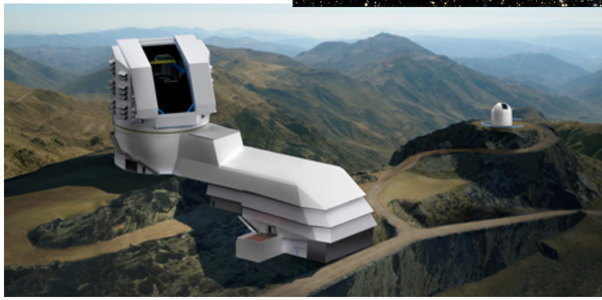
Data based on Science Team home organization

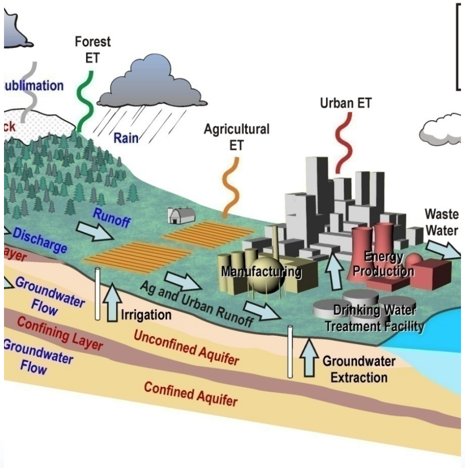
NCSA Many Data Domains

- Blue Waters – most intense data system in the world
- Large Synoptic Survey Telescope
  - 1 PB of final images per year
  - Reprocess the entire repository every year
- Dark Energy Survey
- Genomics
- Personalize Medicine
- Ground Water
- Intelligent Agriculture









EARLY SCIENCE SYSTEM

ESS Use Purpose

- The primary purposes of science team use of the early science time were:
  - To provide a substantial new, interim resource to certain science and engineering teams who have the potential to accomplish a significant science result in an interim period of time.
  - To help the Blue Waters team test and evaluate the early system and prepare for full system testing.

Blue Waters Early Science System

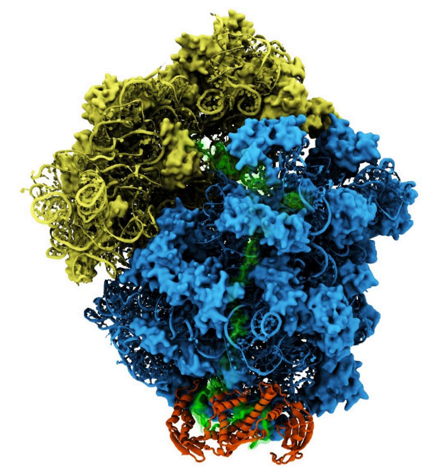
- BW-ESS Configuration
  - 1.4+ PFs (peak)
  - 48 cabinets, 4,512 XE6 compute nodes, 96 service nodes ~18%
  - 2 PBs Sonexion Lustre storage appliance ~5%

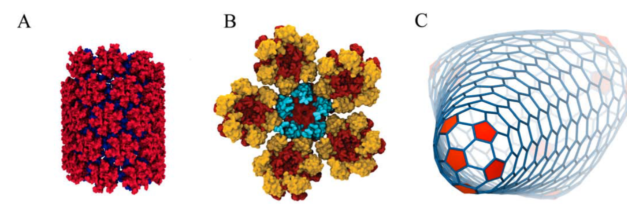
- Current Projects
  - **Biomolecular Physics**—K. Schulten, University of Illinois at Urbana-Champaign
  - **Cosmology**—B. O'Shea, Michigan State University
  - **Climate Change**—D. Wuebbles, University of Illinois at Urbana-Champaign
  - **Lattice QCD**—R. Sugar, University of California, Santa Barbara
  - **Plasma Physics**—H. Karimabadi, University of California, San Diego
  - **Supernovae**—S. Woosley, University of California Observatories
  - **Severe Weather** –R Wilhelmson, University of Illinois
  - **High Resolution/Fidelity Climate** – C Stan, Center for Ocean-Land-Atmospheric Studies (COLA)
  - **Complex Turbulence** – P.K. Yeung, Georgia Tech
  - **Turbulent Stellar Hydrodynamics** – P Woodward, University of Minnesota

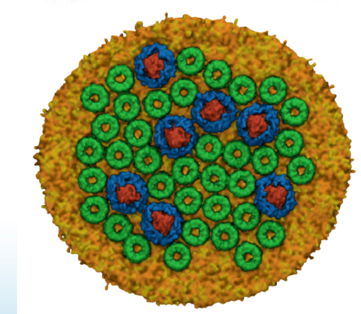
The Computational Microscope: NAMD

- “Not in our wildest dreams could we have imagined the greatness” of Blue Waters

1. **Simulated flexibility of ribosome trigger factor complex at full length** and obtained better starting configuration of trigger factor model (simulated to 80ns)
2. **100ns simulation of cylindrical HIV 'capsule’ of CA proteins** revealed it is stabilized by hydrophobic interactions between CA hexamers; maturation involves detailed remodeling rather than disassembly/re-assembly of CA lattice, as had been proposed.
    - 200ns simulation of CA pentamer surrounded by CA hexamers suggested interfaces in hexamer-hexamer and hexamer-pentamer pairings involve different patterns of interactions
3. **Simulated photosynthetic membrane** of a chromatophore in bacterium Rps. photometricum for 20 ns -- simulation of a few hundred nanoseconds will be needed



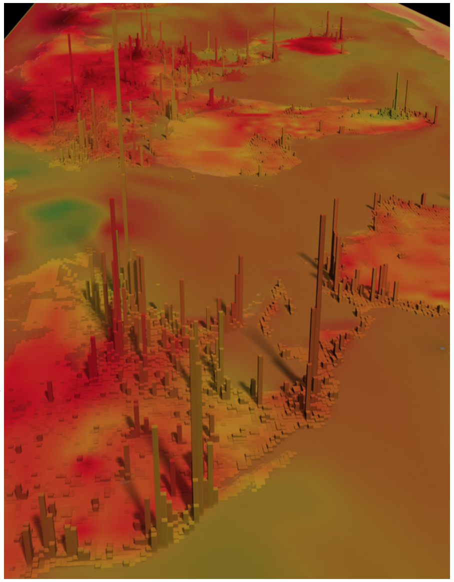




PI: Klaus Schulten, University of Illinois at Urbana-Champaign

Climate Change Uncertainties

- Ported and validated Community Earth System Model
- Performed initial tests of CAR system in CESM using low-resolution version
- Conducted **one-year test of CESM with 0.25° finite-volume dynamical core**. Validated by NCAR
- ESS time helped identify issues with CAM5-PROG at high resolution and I/O bug
- Results from the new prognostic aerosol run (CAM5-PROG) show significant differences in cloud radiative forcing compared to an earlier run with BAM prescribed aerosols, particularly in high latitudes. **Results clearly demonstrate the value of running CAM5-PROG at high resolution.**



PIs: Donald Wuebbles and Xin-Zhong Liang, University of Illinois at Urbana-Champaign

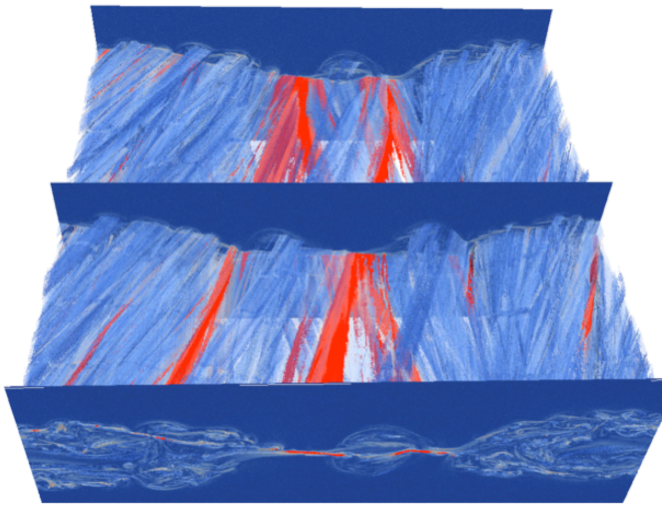
Turbulent Stellar Hydrodynamics

- Obtained **12.02% of peak**, single precision flops for a very large (72 billion-cell) problem. 312 Tflops/s on ESS was counted
- Tested new inertial confinement fusion w. performance enhancements resulting from collaboration with Cray

PI: Paul Woodward, University of Minnesota-Twin Cities

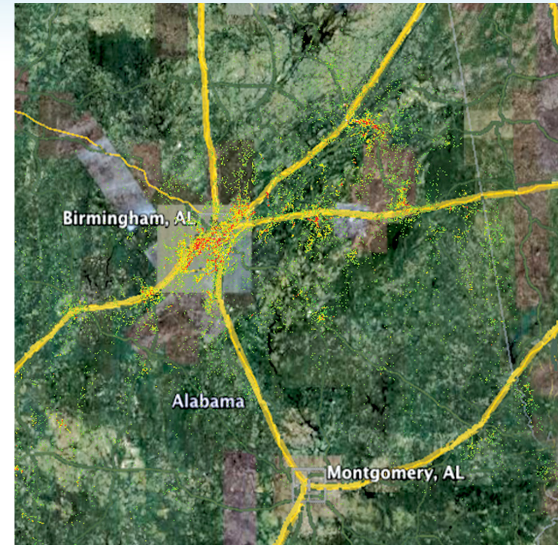
Kinetic Simulations of the Magnetosphere

- Objective: understand 3D evolution of force-free current layers using recent theory describing tearing modes (plasma instability that produces magnetic reconnection while giving rise to topological changes in magnetic field)
- Performed 3 3D simulations with varied rotation in the magnetic field across the initial layer
  - 2 runs: 2048 x 2048 x 1024 = 4.3 billion cells / 1 trillion particles / run on 65536 cores
  - 3rd run: 2048 x 2048 x 1536 = 6.4 billion cells / 1.5 trillion particles / run on 98304 cores
  - Each run generated ~25-30 TB of grid-based data, and another 32TB of particle data
  - 8.3 million particle pushes per second for each core. Corresponds to ~0.2 petaflops for large run
- **New results are dramatically different than previous 2D simulations**
- Hope to have paper completed by end of summer
- “Both the stability and performance of the ESS were outstanding”



PIs: Homayoun Karimabadi, Kevin Quest, Amitava Majumdar, University of California-San Diego

Simulation of Contagion: EpiSimdemics

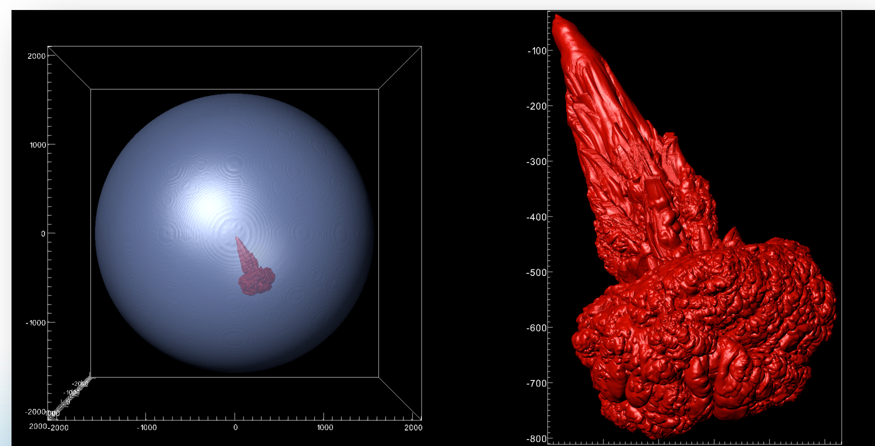


- Measured scaling w. **2 datasets: Michigan (pop. 9.M) and North Carolina/Tennessee/Texas (32.7M)**
- **Should efficiently run on 20k-30k cores**
- On full BW, plan to simulate spread of influenza across U.S., comparing intervention combinations. Problem has not been simulated with this level of detail and at this scale.

PIs: Keith Bisset, Virginia Tech; Shawn Brown, Carnegie-Mellon University; Douglas Roberts, Research Triangle Institute

Modeling Type 1a Supernovae

- **Off-center ignition of Type 1a Supernovae, 1 second duration**
- Codes: MAESTRO and CASTRO
- Used 68 million core hours
- Produced 45 TBs of data



PIs: PI: S. Woosley, University of California Observatories

Lattice Gauge Theory on Blue Waters

- Calculation of the spectroscopy of charmonium, the positronium-like states of a charm quark and an anticharm quark.
- Also spent a limited amount of time preparing for the two large projects we hope to run on Blue Waters when the full system is available for production work.
- Able to reproduce these **mass splittings at a record precision of a couple MeV** is an impressive test of our methodology,
- Gives us confidence in our ability to make predictions of other levels where the experimental values are not known or the classification of the states is not understood.

PI: Robert Sugar, U C Santa Barbara

Simulations of Homogeneous Turbulence

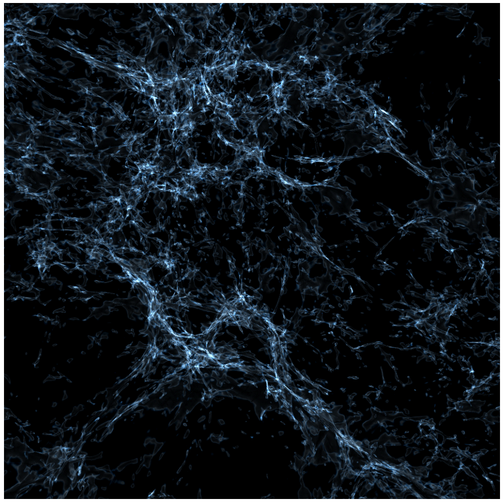
- Performed **40963 simulation, run on 16k or 32k cores**, 20 sets of restart files were transferred to NICS
- Code performance obtained on BW-ESS was generally slightly better than on Jaguarpf at NCCS.
- Performance improvement using Co-Array Fortran (CAF) in place of mpi alltoall(v) with the help of Cray Team.
- In decent shape in terms of the viability of running an Reτ ~ 5000 calculation (the next-generation channel flow target) on the full-scale Blue Waters.

PI: P. K. Yeung, Georgia Institute of Technology

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Formations of galaxy complexes at high redshift



- Performed **8 cosmological simulations to understand how galaxies in the early Universe** (the first billion years or so) grow and evolve, in several statistically-‐dissimilar environments
- **Larger than any other AMR cosmological simulations ever done**.
- BW's high memory per core was crucial to the success.
- Happy with the performance of the system except I/O subsystem.
- System performed brilliantly and technical support was prompt and of exceptional quality.
- **Transferred close to 800 TB of data from NPCF after runs** http://galactica.pa.msu.edu/~bwoshea/data/BlueWaters/ESSmovies/

PI: Brian O'Shea, Michigan State University, Co-PI: Michael Norman, UC San Diego/SDSC

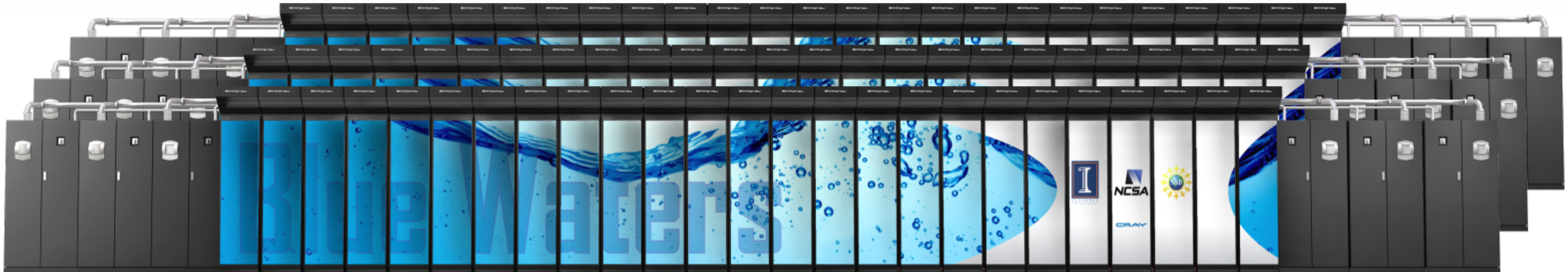
Computational Chemistry at the Petascale

- Ported and obtained good initial timings for the Gamess code
- An energy + gradient calculation on a cluster of 512 water molecules was run using FMO at the MP2 level of theory using the aug-cc-pVDZ basis set.
- Timings for energy + gradient without the fully analytic gradient : **BW calculation on 4096 ~11.7 minutes, while the analogous BG/P calculation took 28.9 minutes on 8192 cores.**

PIs: Monica Lamm, Iowa State University

THE DATA SIDE OF BLUE WATERS

Blue Waters Computing System



Aggregate Memory - Logic: 1,500 PB

10/40/100 Gb Ethernet Switch

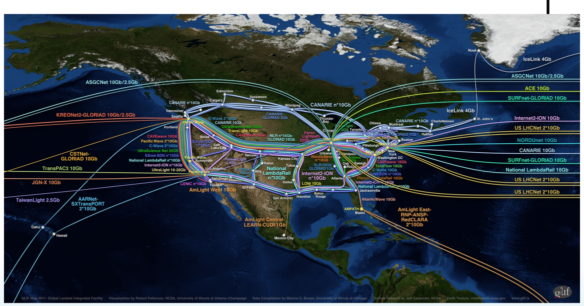
External Servers

IB Switch

>1 TB/sec

120+ Gb/sec

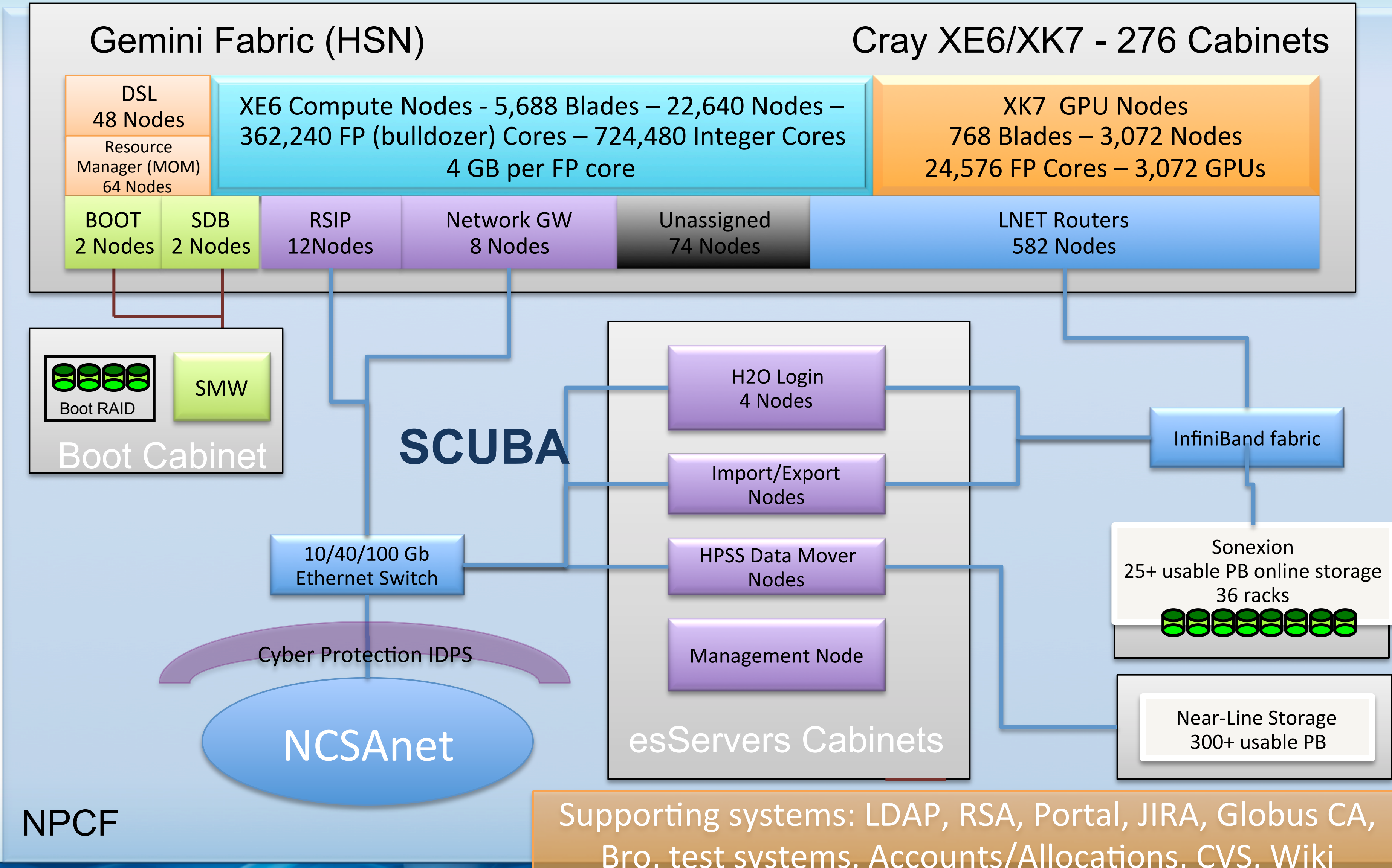
100 GB/sec

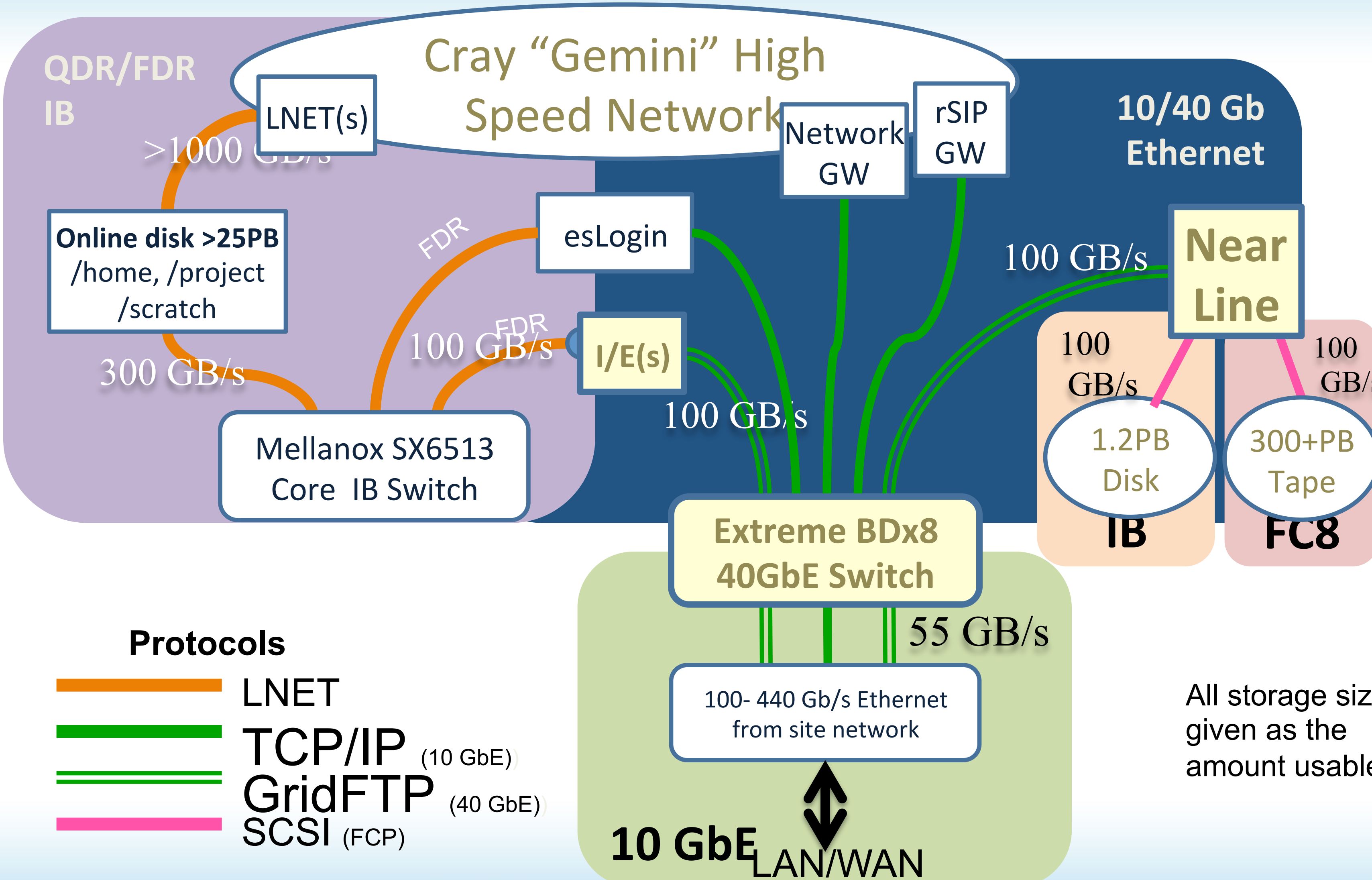


100-300 Gbps WAN

Spectra Logic: 300 PB

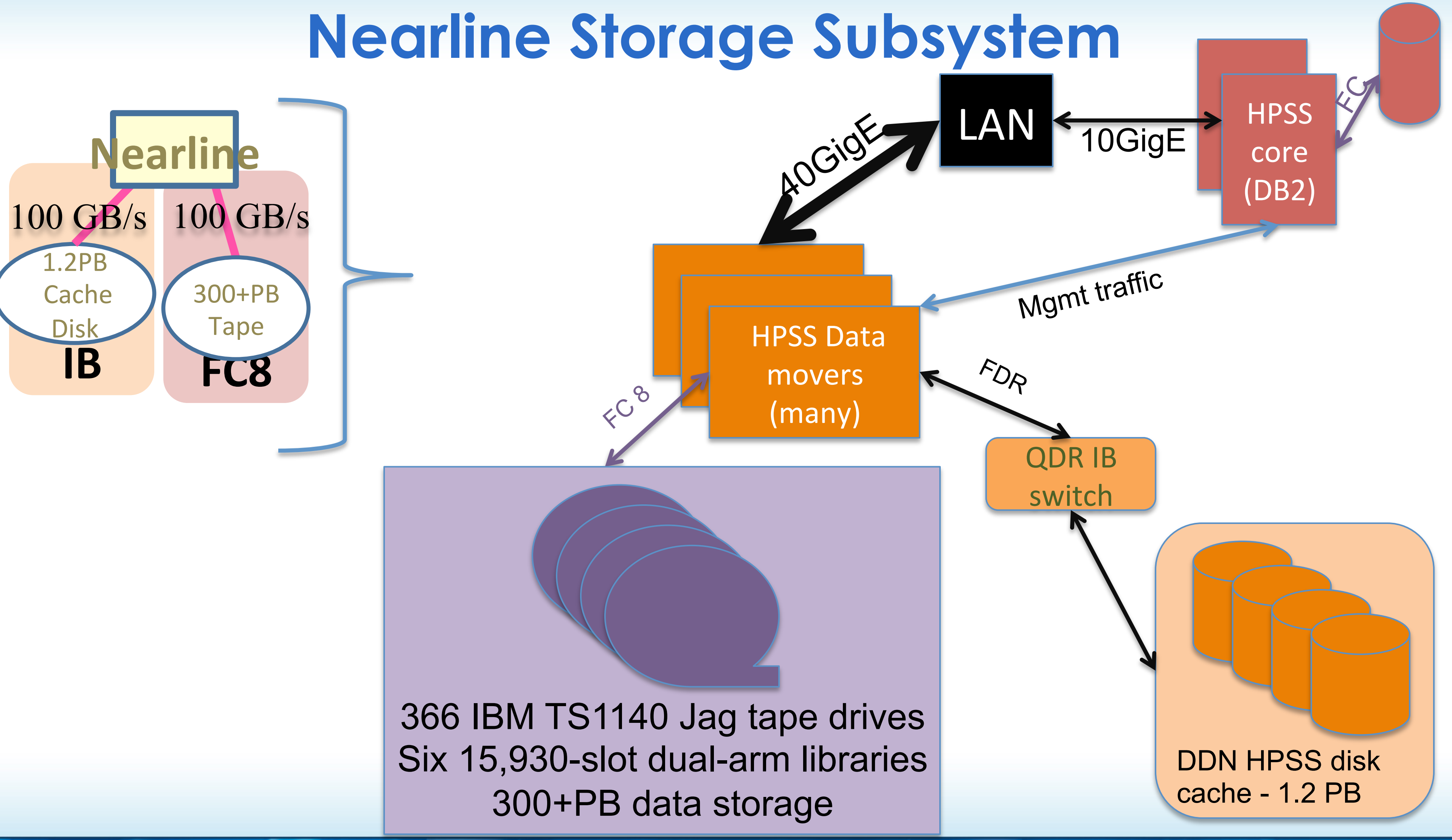
Sonexion: 26 PB


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 Nodes
24,576 FP Cores – 3,072 GPUs
BOOT
2 Nodes
SDB
2 Nodes
RSIP
12Nodes
Network GW
8 Nodes
Unassigned
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

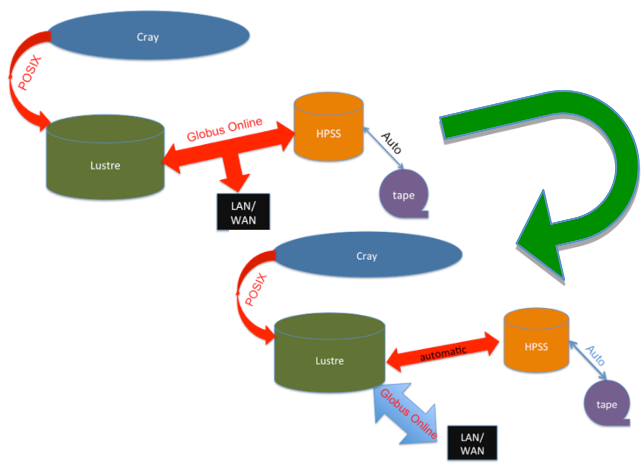


All storage sizes given as the amount usable

Nearline Storage Subsystem

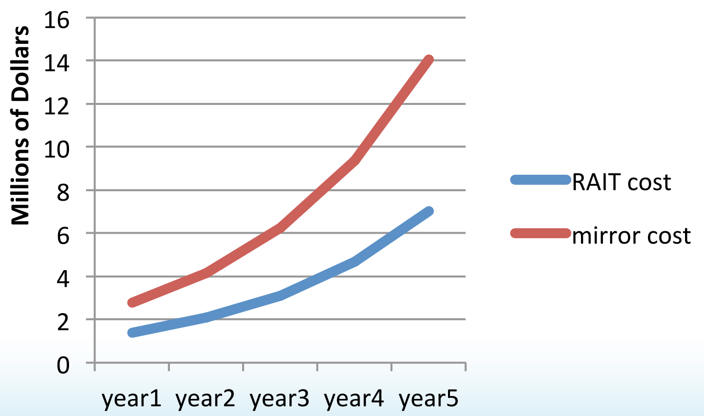


Near-Line Storage



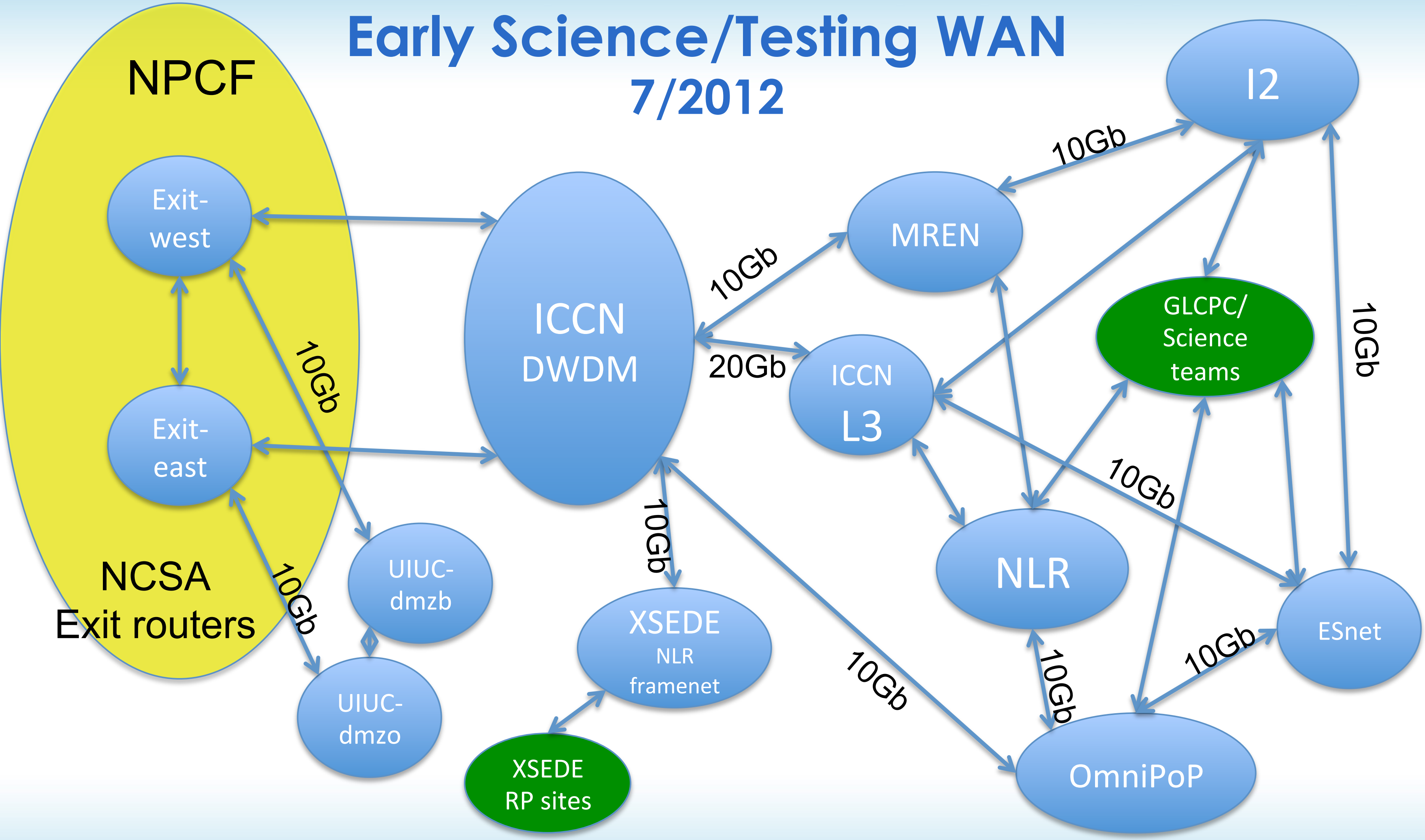
- ***Have the right data at the right place at the right time***
- ***Eliminate Partner Data Pain***
- Cost Efficient
  - RAIT
  - Managing data (limits, transparent movement, consolidation, etc.)
- Import/Export server management and support
- Community Leadership

- **Most balanced and intense storage implementation in open science**
  - Scale and Performance
- Advanced Technologies
  - RAIT, Lustre-HPSS Interface, ILM, etc.
- Maintain storage related software packages
- Maintain and improve BW developed SW
- Performance testing and tuning
- Import/export facility maintenance and service request management

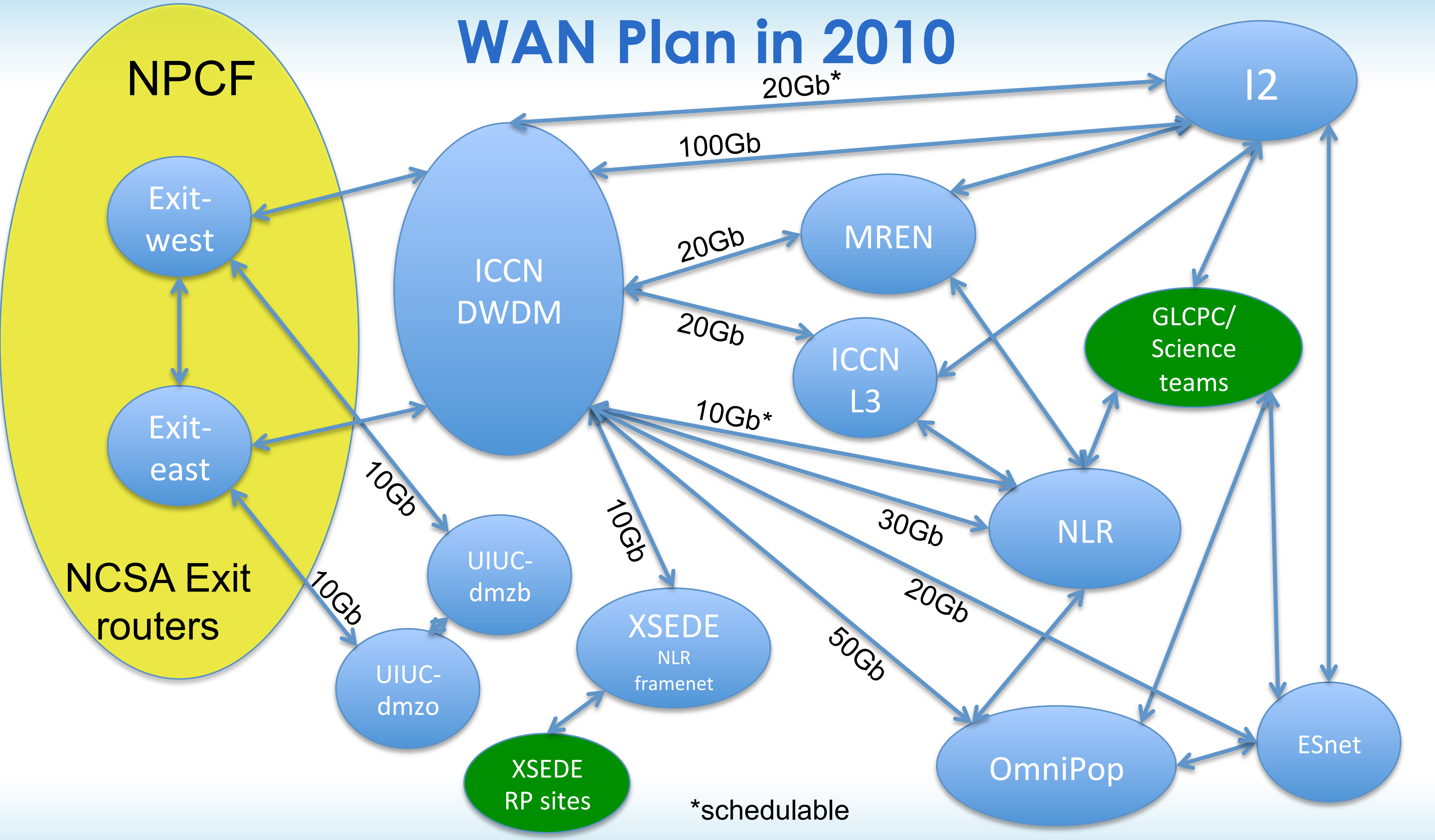


AND NOW – THE WIDE AREA NETWORK

Early Science/Testing WAN
7/2012



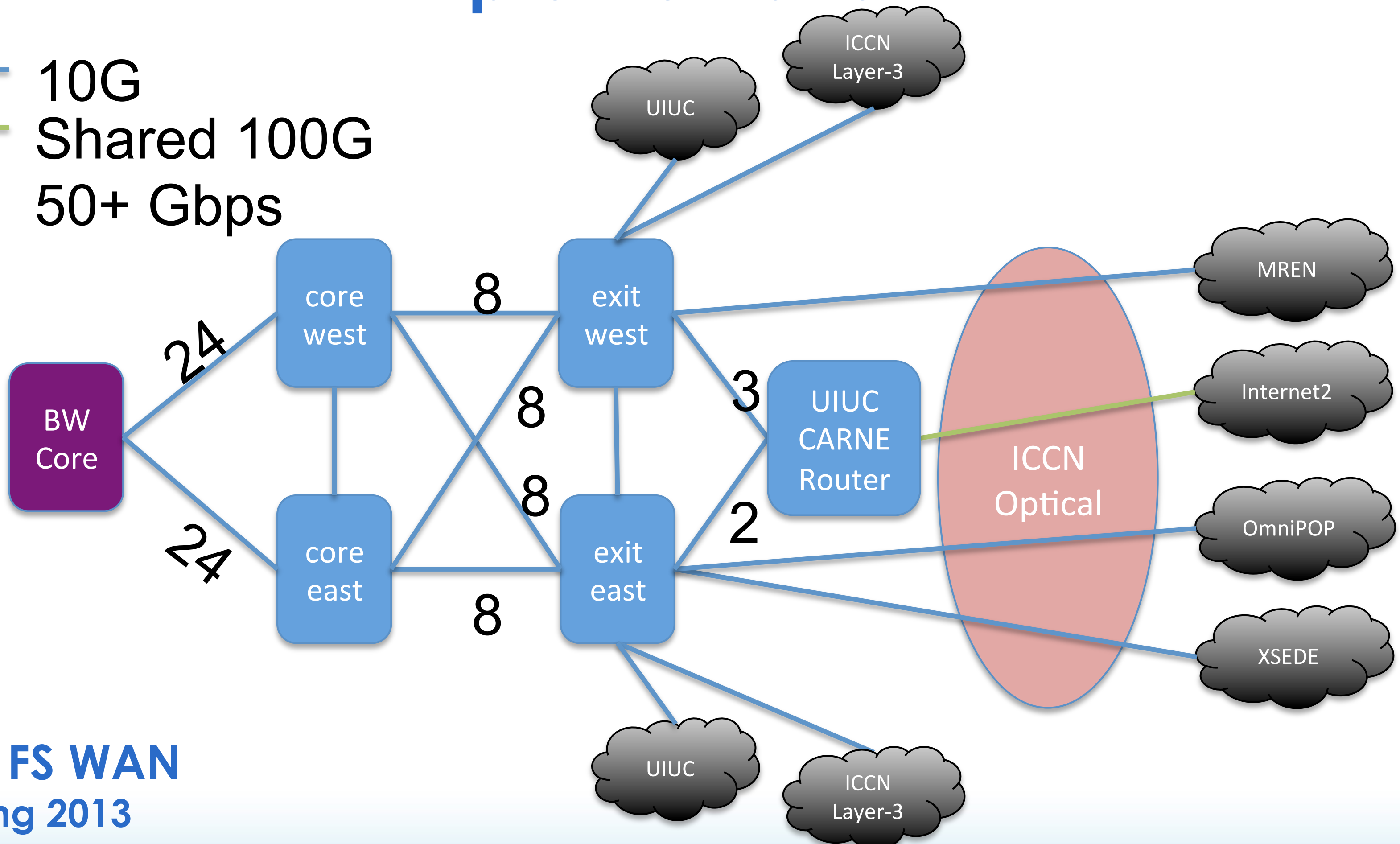
WAN Plan in 2010



*schedulable

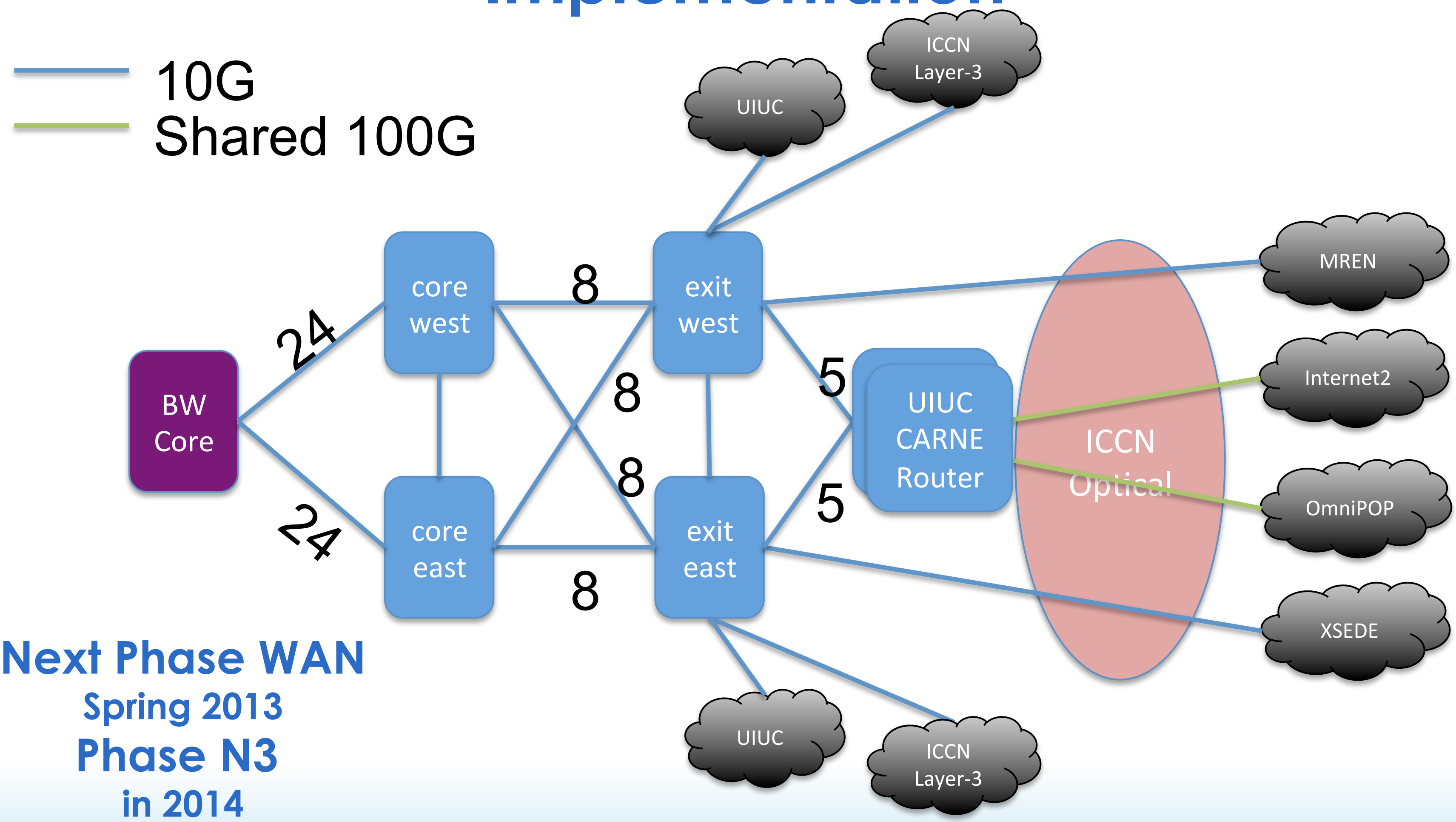
Potential Phase N2a Full Service Implementation

10G
Shared 100G
50+ Gbps



Initial FS WAN
Spring 2013

Potential Phase N2b Full Service Implementation



Next Phase WAN
Spring 2013
Phase N3
in 2014

Network Intellectual Services

- End to End tuning
  - Optimize WAN connectivity in the beginning to create direct peerings with strategic endpoints when possible.
    - New peerings can be brought up fairly quickly
  - Example - Utilize the growing network of Perfsonar network measurement devices
    - Augment NCSA's existing perfsonar servers with a unit on the BW network. Encourage sites to install a perfsonar node.
    - Provides end user accessible tools to characterize paths and test end hosts for tuning issues.
    - Provides a remote testing point for use in eliminating the "easy" problems first. Retains test history creating a performance baseline to other personar nodes.

- Helping Science Teams
  - Understand the issue, hosts, and applications involved.
    - Obtain contact info for remote site network engineer
    - Coordinate testing, debugging, and analysis
    - Implement or assist with implementing solutions.
  - Example - working with UMN to characterize the path from NCSA to the Minnesota Supercomputing Institute (MSI).
    - They are installing a perfsonar node.
    - Will have a history of performance data to use to set performance expectations.
  - When appropriate we will assist partner sites in facilitating acquiring additional connectivity or optimizing existing connectivity.

Summary

- Petascale presents significant challenges for performance, flexibility and data investment tradeoffs
  - One dimensional optimization underserves many communities
- Blue Waters is an exceptional computational resource
- Blue Waters is the most intense data focused environment in the open scientific community

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 Blue Waters Project.

Individual Help From

- Thom Dunning, Marc Snir, Wen-mei Hwu, Bill Gropp
- Cristina Beldica, Brett Bode, Michelle Butler, Paul Wefel, Tim Boemer, Greg Bauer, Mike Showerman, John Melchi, Scott Lathrop, Irene Qualters, Sanjay Kale
- The Blue Waters Project Team and our partners
- NSF/OCI
- Cray, Inc, AMD, NVIDIA, Xyratex, Adaptive, Allinea