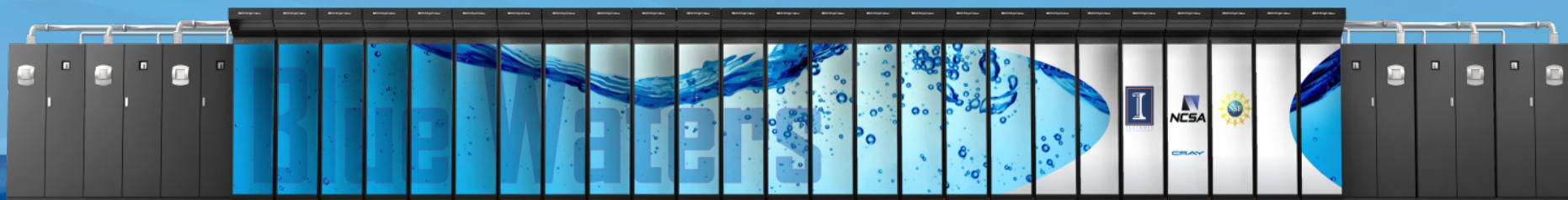


# BLUE WATERS

SUSTAINED PETASCALE COMPUTING

## HPC User Forum – Richmond VA – April 2012

### Merle Giles, NCSA Business & Economic Development



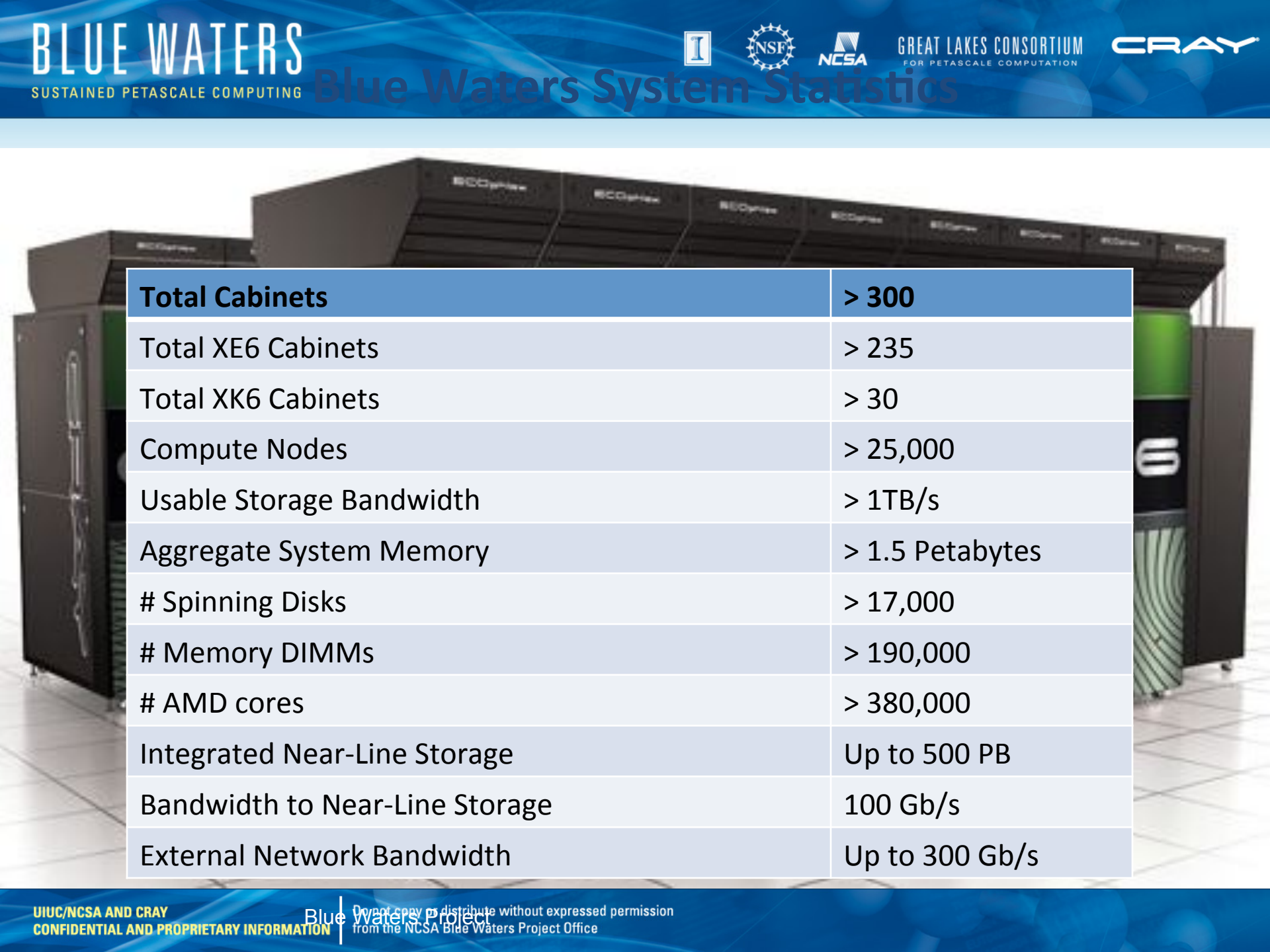
GREAT LAKES CONSORTIUM  
FOR PETASCALE COMPUTATION

CRAY®

UIUC/NCSA AND CRAY  
CONFIDENTIAL

Do not copy or distribute without expressed permission  
from the NCSA Blue Waters Project Office

## Blue Waters System Statistics



<b>Total Cabinets</b>	<b>&gt; 300</b>
Total XE6 Cabinets	> 235
Total XK6 Cabinets	> 30
Compute Nodes	> 25,000
Usable Storage Bandwidth	> 1TB/s
Aggregate System Memory	> 1.5 Petabytes
# Spinning Disks	> 17,000
# Memory DIMMs	> 190,000
# AMD cores	> 380,000
Integrated Near-Line Storage	Up to 500 PB
Bandwidth to Near-Line Storage	100 Gb/s
External Network Bandwidth	Up to 300 Gb/s

## Summary

- Outstanding Computing System
  - Cray's most advanced computing system with technologies from AMD (processor), Xyratex (storage) and NVIDIA (accelerators) plus Spectra Logic (archive)
  - Balanced computing system capable of handling most challenging compute-, memory- and data-intensive problems
- Outstanding Development and Deployment Project
  - Focus on *revolutionizing science and engineering through petascale computing*, not simply fielding a computer
    - Petascale computing super-system, petascale applications porting and optimization, petascale software development, petascale educated scientists and engineers, outreach to underrepresented groups
- Outstanding Team
  - UIUC/NCSA—Four decades of leadership in HPC
  - Cray—HPC leadership and commitment with 16 systems in Top50
  - Great Lakes Consortium—Broad range of HPC expertise
- Outstanding Location
  - University campus, combining security with open access

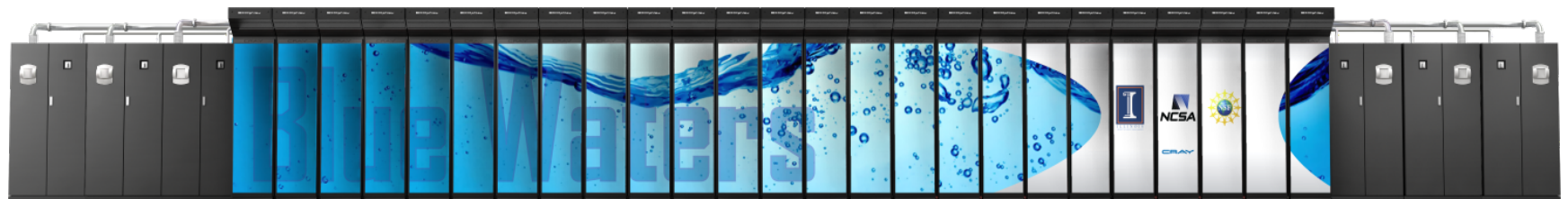


## Blue Waters Goals

- Deploy a capable, balanced system of sustaining more than one petaflops or more for a broad range of applications
  - Base is the Cray XE/XK system using well defined metrics
- Enable the Science Teams to take full advantage of the sustained petascale system
  - Strong and deep partnership with Science Teams, helping them to improve the performance, scalability and heterogeneity of their applications
- Enhance the operation and use of the sustained petascale system
  - Tools, libraries, methods to aid in operation of the system and to help scientists and engineers make very effective use of the system
- Provide a world-class computing environment for the petascale system
  - The Blue Waters *Super-System* - Outstanding balance of primary system, a modern, energy-efficient data center, a rich WAN environment (100-300 Gbps), near-line, automated data storage subsystem (300-400 usable PB) all with advanced cyber-protection.
- Exploit advances in innovative computing technology
  - Balance of general and heterogeneous computing to help the computational community transform to new modes for computational and data-driven science and engineering



# Blue Waters Computing Super-system



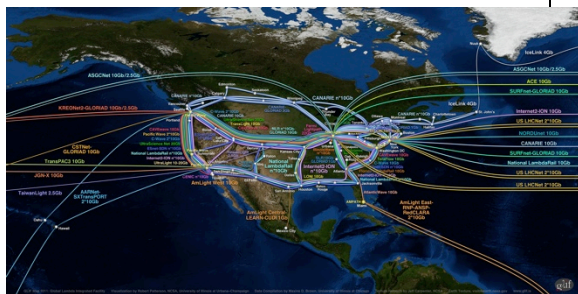
10/40/100 Gb  
Ethernet Switch

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec



**WAN**



**Spectra Logic: 300+ PBs**



**Sonexion: >25 PBs**

# 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://bluwaters.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*

## Early Feedback from Science Teams

- Woosley Science Team

“As an Early Science System user, we have access to the ESS machine which has, thus far, been remarkably stable, especially considering that this system is still very much in an alpha development stage. Any bugs that we’ve found have been resolved by the NCSA staff, usually within a matter of hours.”

- Schulten Science Team

“What an unbelievable first day on Blue Waters! ... In fact, the start was so great that we could begin immediately long-waiting science projects ... Blue Waters made us wait a long time, but it surely was worth waiting for!”

- Sugar Science Team

“The machine has performed well above our expectations in the first week of running. ... Thanks to NCSA staff, we have also made excellent progress in improving the efficiency of our codes for the new architecture. We expect to be able to achieve a performance of more than 3 Tflops on 3072 cores (192 nodes).”



## Industry S&E Teams



- Industry can participate in the NSF PRAC process
- 5+% allocation can be dedicated to industrial use
  - Specialized support by the NCSA Private Sector Program (PSP) staff
  - Blue Waters staff will support the PSP staff as needed
  - Potential to provide specialized services within Service Level Agreements parameters
    - E.g. throughput, extra expertise, specialized storage provisions, etc.

High interest shared by partner companies in the following:

- Scaling capability of a well-known and validated CFD code
- Temporal and transient modeling techniques and understanding.

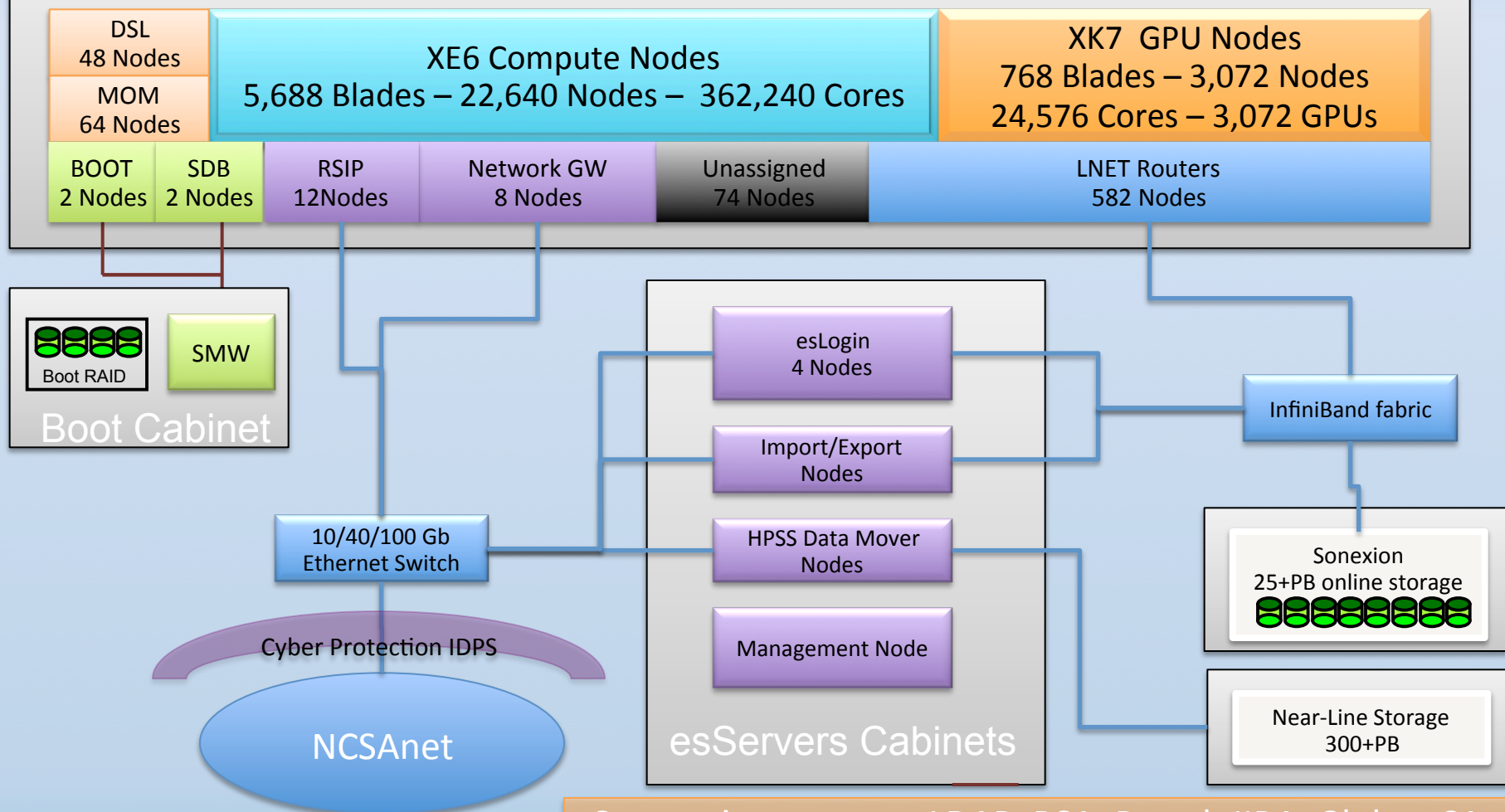
Two example cases under discussion:

- NASA OVERFLOW at scale for CFD flows
- Temporal modeling techniques using the freezing of H<sub>2</sub>O molecules as a use case and as a reason to conduct both large-scale single runs and to gain significant insight by reducing uncertainty.

Science Area	Number of Teams	Codes	Structured Grids	Unstructured Grids	Dense Matrix	Sparse Matrix	N-Body	Monte Carlo	FFT	Significant I/O
Climate and Weather	3	CESM, GCRM, CM1, HOMME	<b>X</b>	<b>X</b>		<b>X</b>		<b>X</b>		
Plasmas/ Magnetosphere	2	H3D(M), OSIRIS, Magtail/UPIC	<b>X</b>				<b>X</b>		<b>X</b>	<b>X</b>
Stellar Atmospheres and Supernovae	2	PPM, MAESTRO, CASTRO, SEDONA	<b>X</b>			<b>X</b>		<b>X</b>		<b>X</b>
Cosmology	2	Enzo, pGADGET	<b>X</b>			<b>X</b>	<b>X</b>			
Combustion/ Turbulence	1	PSDNS	<b>X</b>						<b>X</b>	
General Relativity	2	Cactus, Harm3D, LazEV	<b>X</b>			<b>X</b>				
Molecular Dynamics	4	AMBER, Gromacs, NAMD, LAMMPS			<b>X</b>		<b>X</b>		<b>X</b>	
Quantum Chemistry	2	SIAL, GAMESS, NWChem			<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>		<b>X</b>
Material Science	3	NEMOS, OMEN, GW, QMCPACK			<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>		
Earthquakes/ Seismology	2	AWP-ODC, HERCULES, PLSQR, SPECFEM3D	<b>X</b>	<b>X</b>			<b>X</b>			<b>X</b>
Quantum Chromo Dynamics	1	Chroma, MILD, USQCD	<b>X</b>		<b>X</b>	<b>X</b>	<b>X</b>		<b>X</b>	
Social Networks	1	EPISIMDEMICS								
Evolution	1	Eve								
Computer Science	1			<b>X</b>	<b>X</b>	<b>X</b>			<b>X</b>	<b>X</b> <sup>9</sup>

## Gemini Fabric (HSN)

## Cray XE6/XK7 - 276 Cabinets



NPCF

Supporting systems: LDAP, RSA, Portal, JIRA, Globus CA, Bro, test systems, Accounts/Allocations, CVS, Wiki

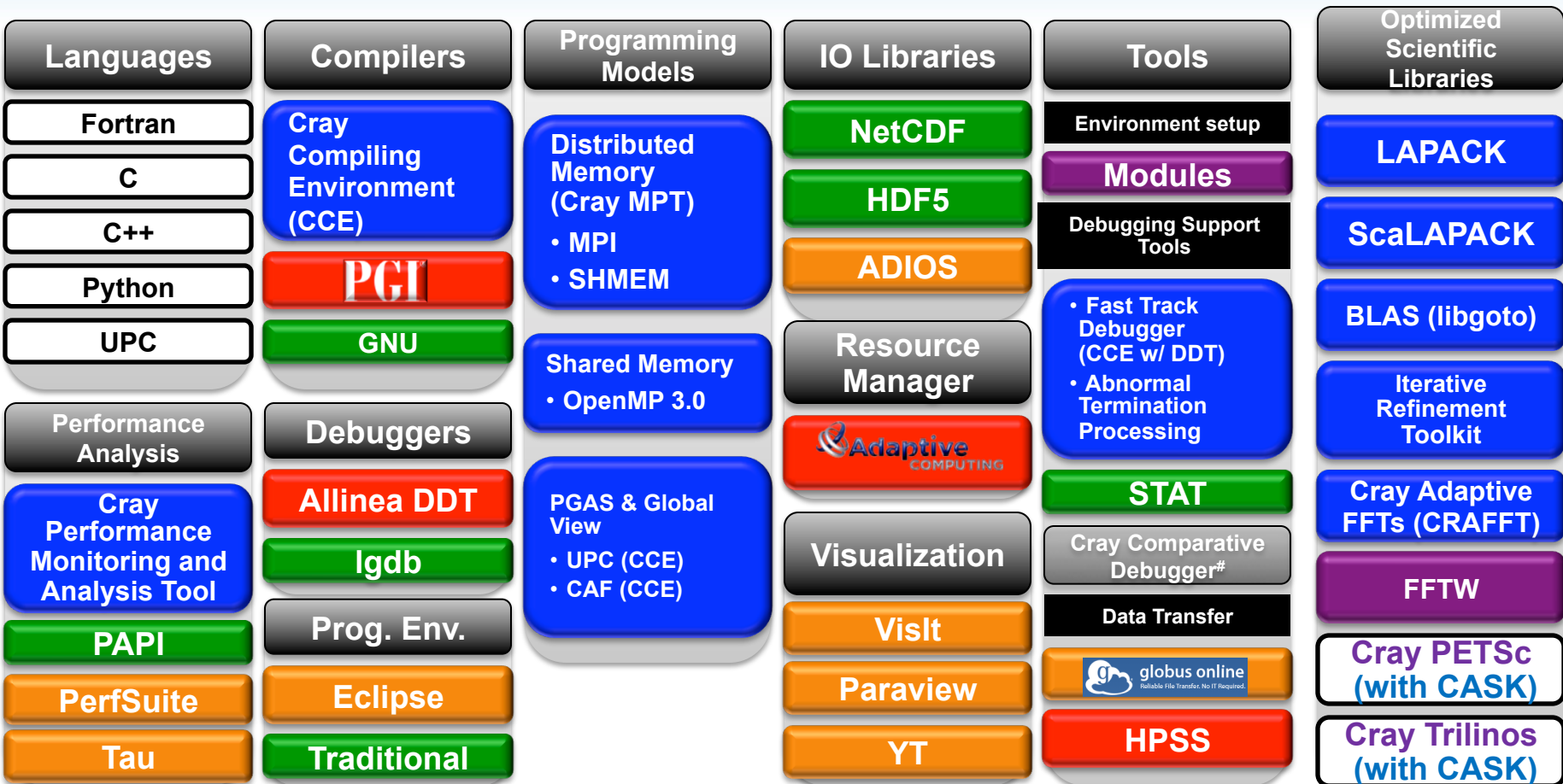




## RAIT

- 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

## Blue Waters Software Environment

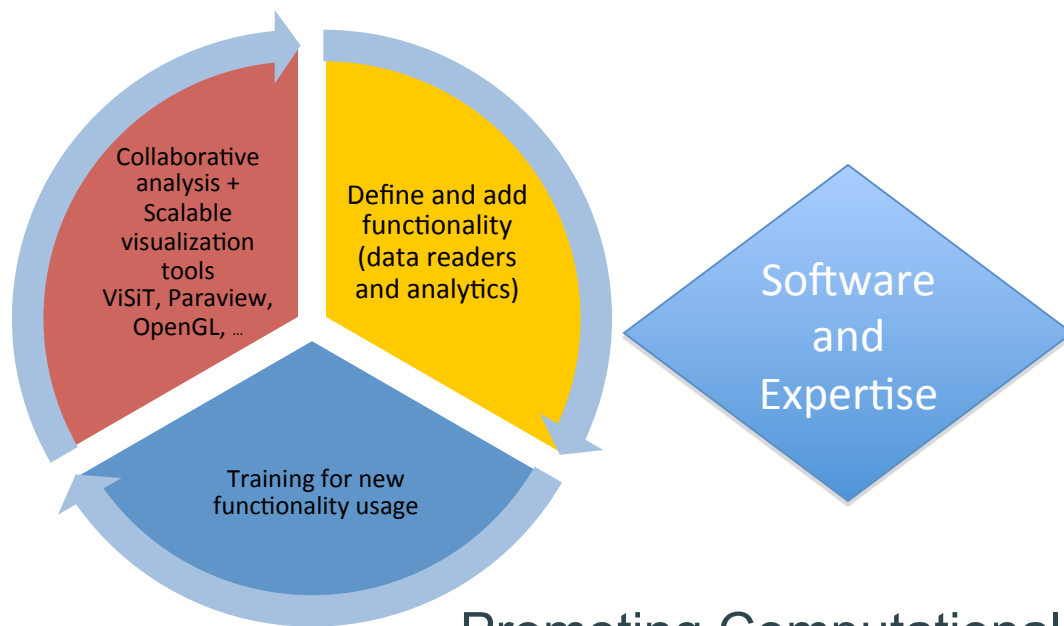


Cray Linux Environment (CLE)/SUSE Linux



# Visualization and Analysis

Enabling Knowledge Discovery



Enhancing Collaboration

**Remote Large Format Displays**

- Well suited for viewing large data
- Accommodate groups of viewers

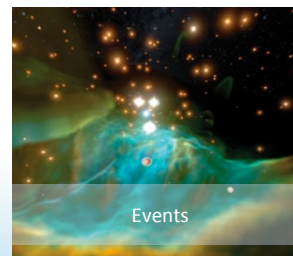
**Inline visualization for Simulations**

- Monitor health of running simulations
- Distributed collab. via multiple viewers

**Web-based Visualization**

- Web access to visualization tools
- Shared repository of visualization products

Promoting Computational Science



# State-of-the-Art Cyber-Protection

Top Down – Bottom Up **Assessment** of risks, threats and vulnerabilities

Zoned approach for network monitoring

Emphasis on:

Prevention (configuration management, timely and expert system management...)

Monitoring (host/network/usage)

Vetting/auditing/scanning – periodic and with changes

Adaptable process & procedures

Incident response

## Monitoring

Network:

Full Netflows & Bro IDS alerts

Host:

Syslogs, OSSEC (host IDS)

SSH keystroke logging

Log Management & Correlation:

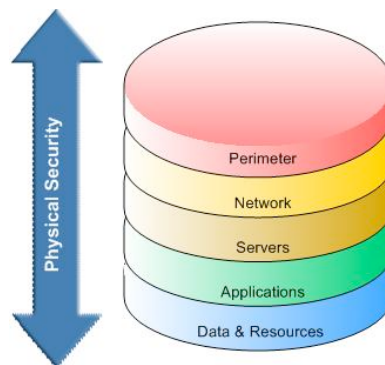
Off system log repository and backups

Customized log correlation tools

Customized Event Monitoring (e.g., SSH User Profiler)



**Rapid, Expert Response**



- **Balance least restrictive environment with appropriate protection**
- Large(st) 80 node **Bro** cluster deployed which monitors:
  - Many 10G WAN links
  - Zone borders with different trust levels
- Active IDPS eliminates issues with firewalls
  - Network performance, fragmentation, delay
  - Static or limited adaptability
  - Restrictions
- Significant effort devoted to maintain, tune, & improve
- Feedback improves Bro rules:
  - NCSA & Berkeley Bro IDS collaboration
  - Novel incidents generate new Bro rules
  - Detect attack sooner next time







# The Blue Waters Team (partially)





# Thank you!

