

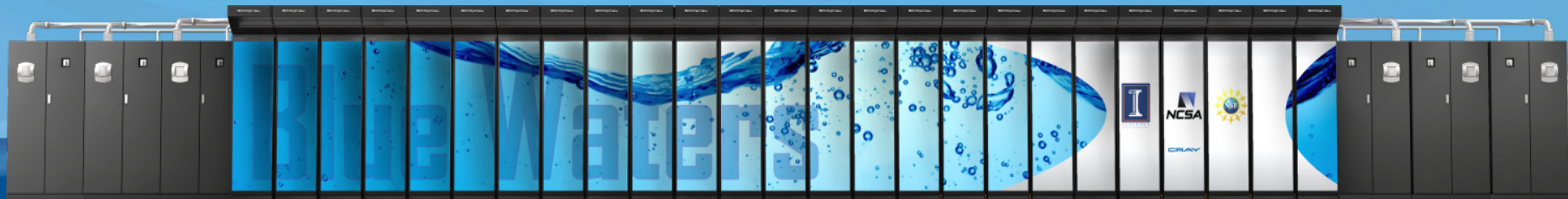
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Blue Waters Update and Performance Report

William Kramer

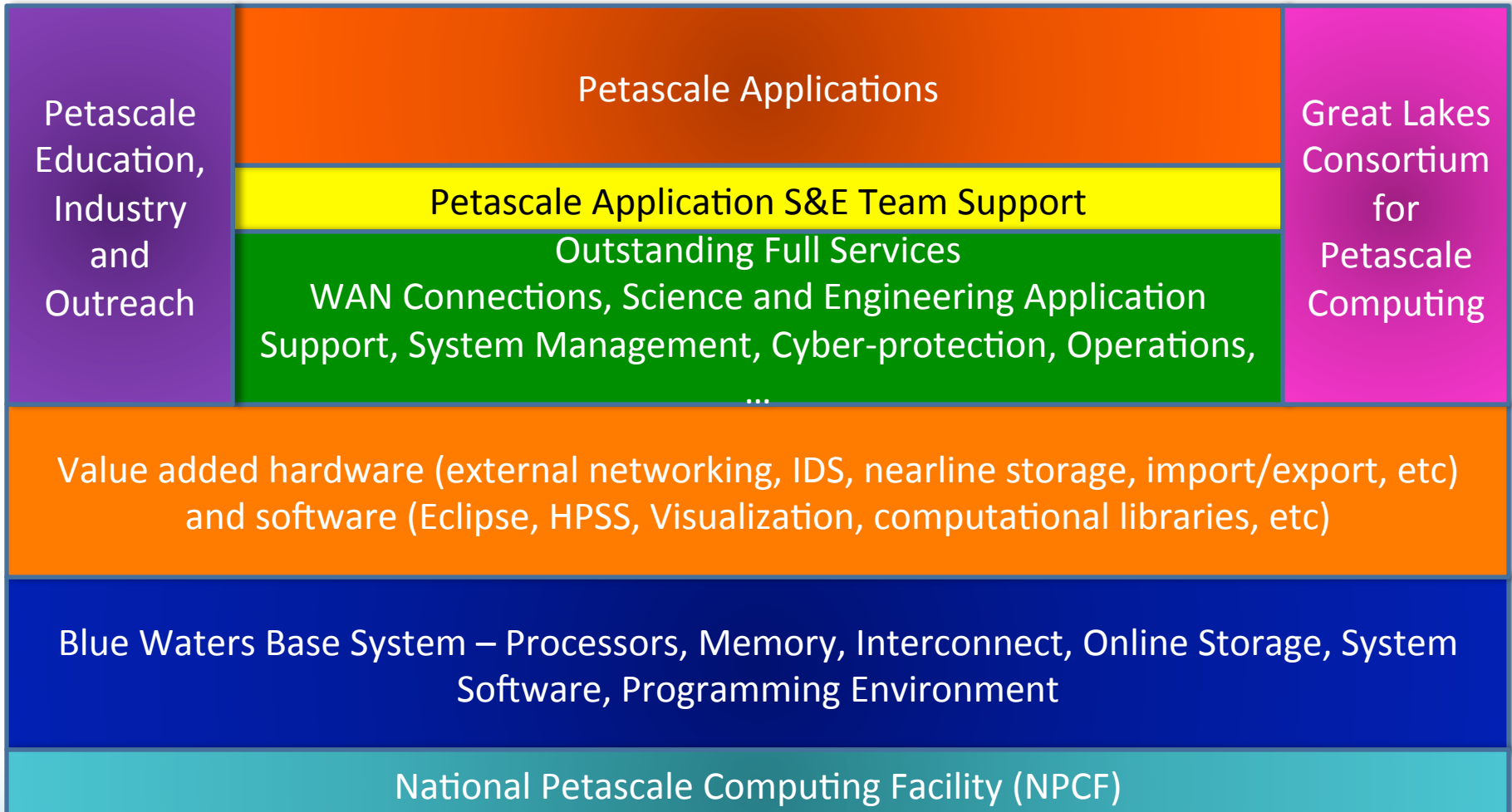
National Center for Supercomputing Applications,
University of Illinois at Urbana-Champaign



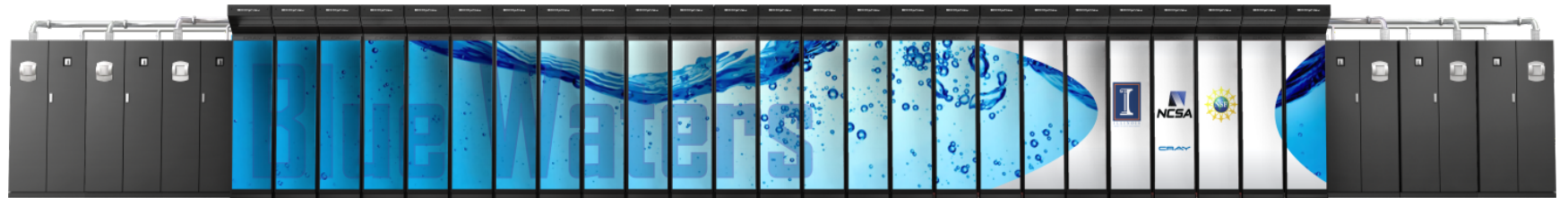
GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY®

The Blue Waters Eco-System



Blue Waters Computing Super-system



10/40/100 Gb
Ethernet Switch

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec



WAN



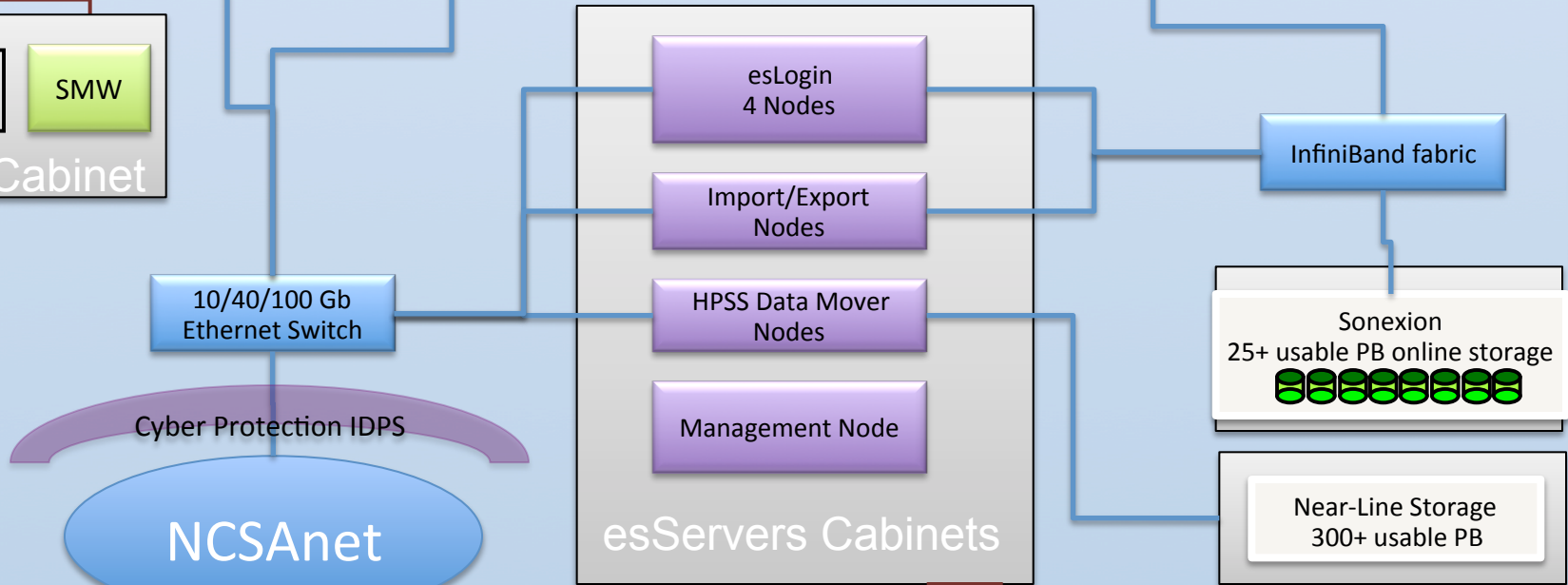
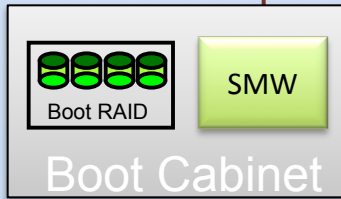
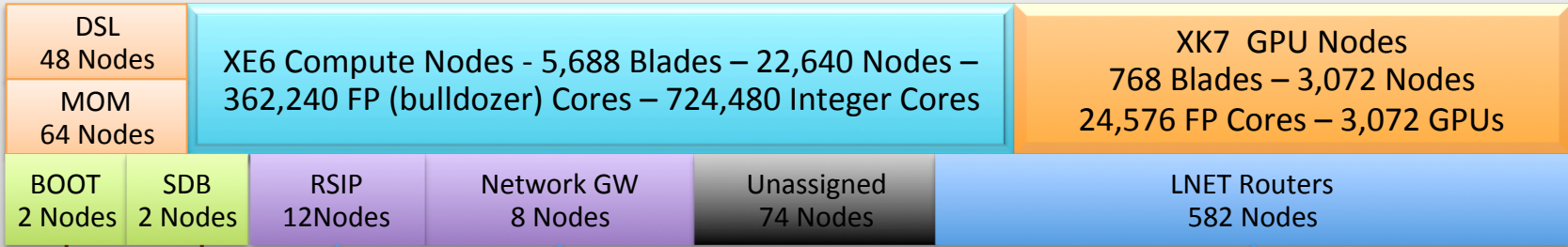
Spectra Logic: 300+ Usable PBs



Sonexion: >25
usable PBs

Gemini Fabric (HSN)

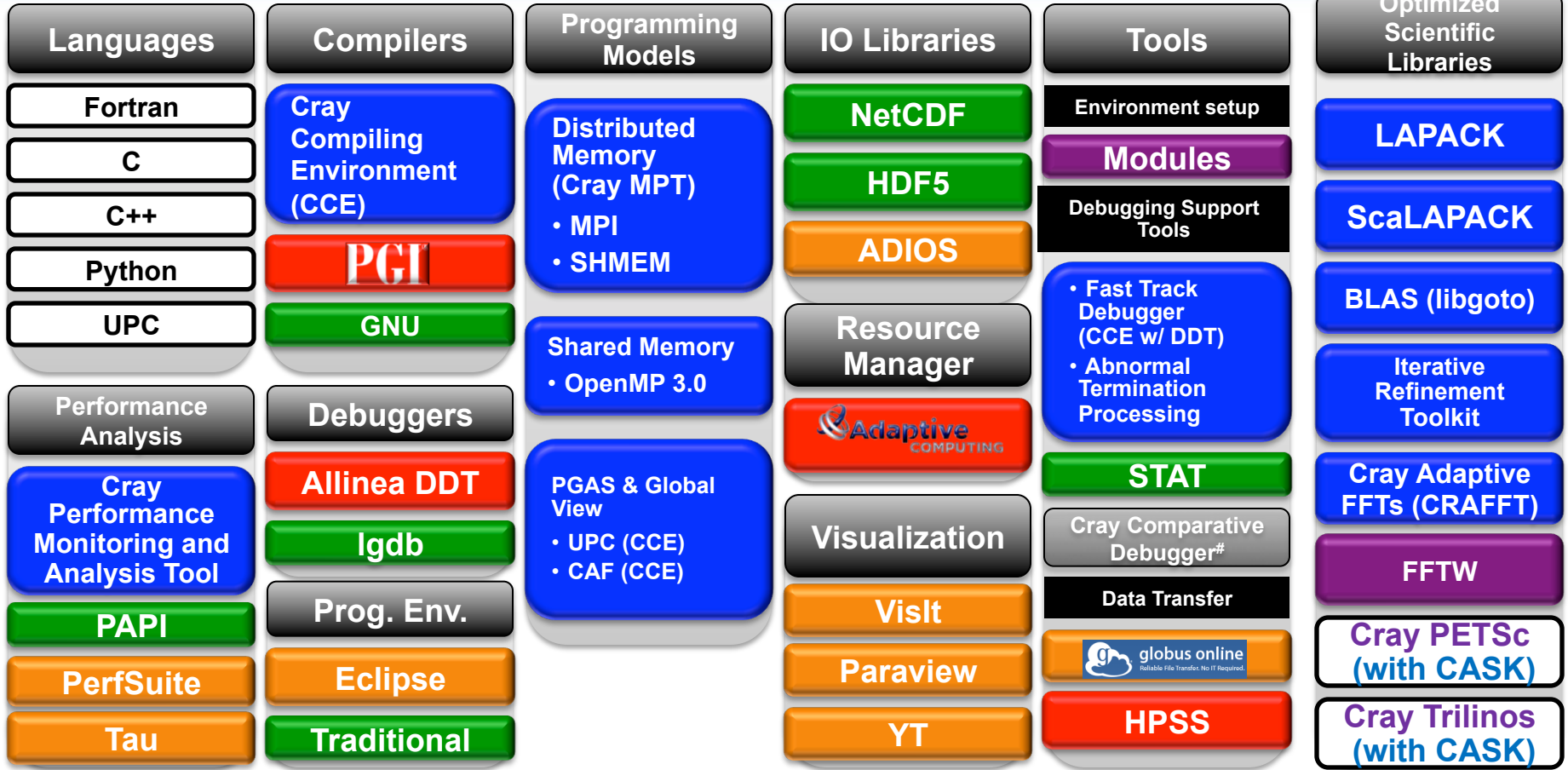
Cray XE6/XK7 - 276 Cabinets



NPCF

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

Blue Waters Software Environment



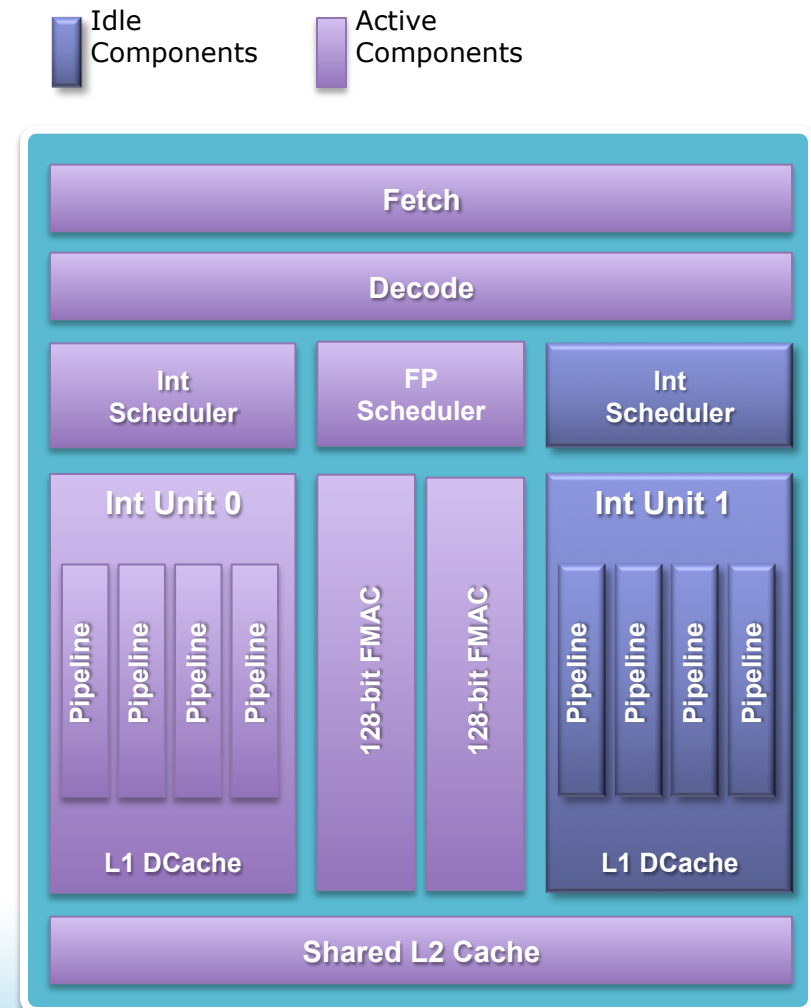
Cray Linux Environment (CLE)/SUSE Linux

Cray developed
Under development
Licensed ISV SW

3rd party packaging
NCSA supported
Cray added value to 3rd party

Defining a Core - AMD Wide AVX mode

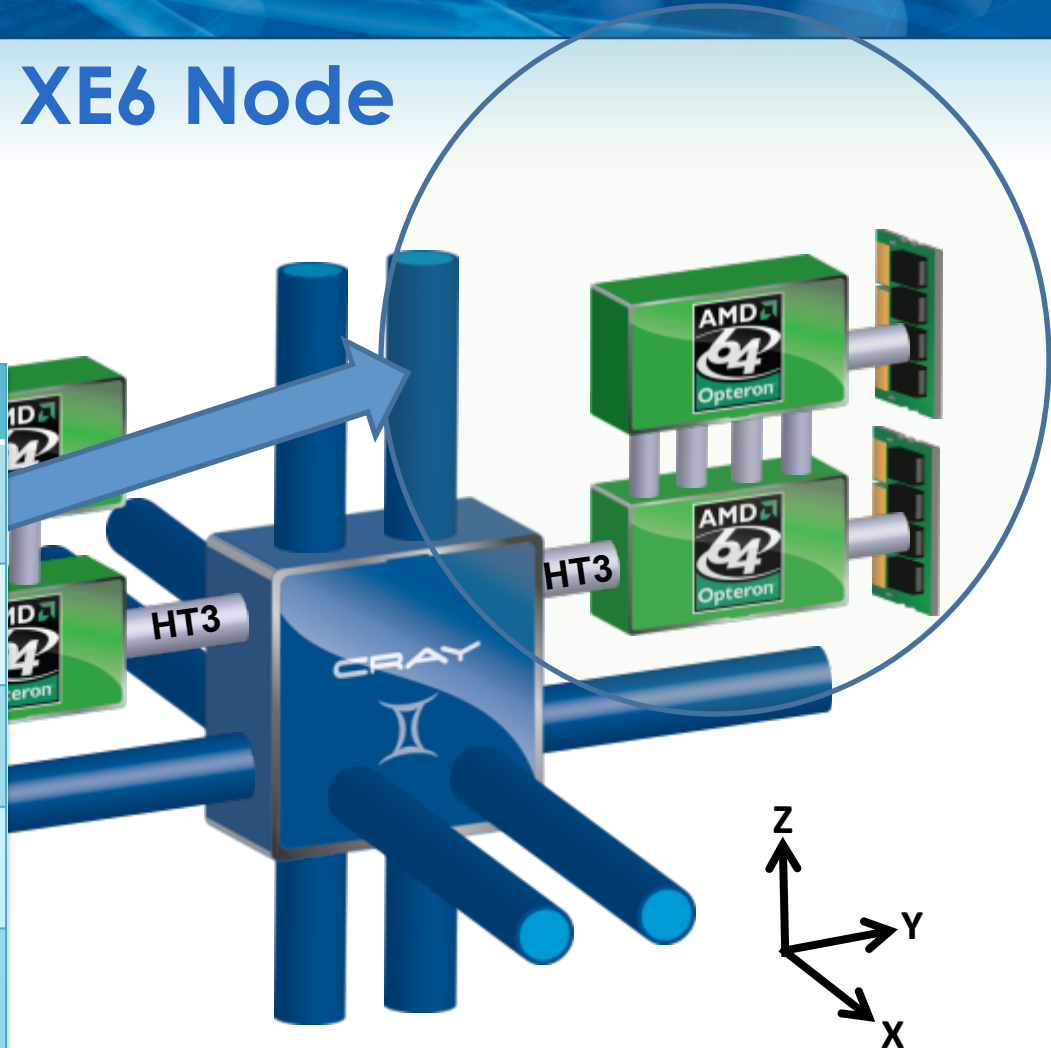
- In this mode, only one integer scheduler unit is used
 - Most common mode for S&E applications
 - Code is Floating Point dominated and makes use of AVX instructions
 - Code needs more memory per MPI rank
- Implications
 - This core has *exclusive* access to the 256-bit FP unit and is capable of 8 FP results per clock cycle
 - The core has twice the memory capacity
 - The core has twice the memory bandwidth
 - The L2 cache is effectively twice as large
 - The peak performance of the chip is not reduced
- AMD refers to this as a “Core Module”



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



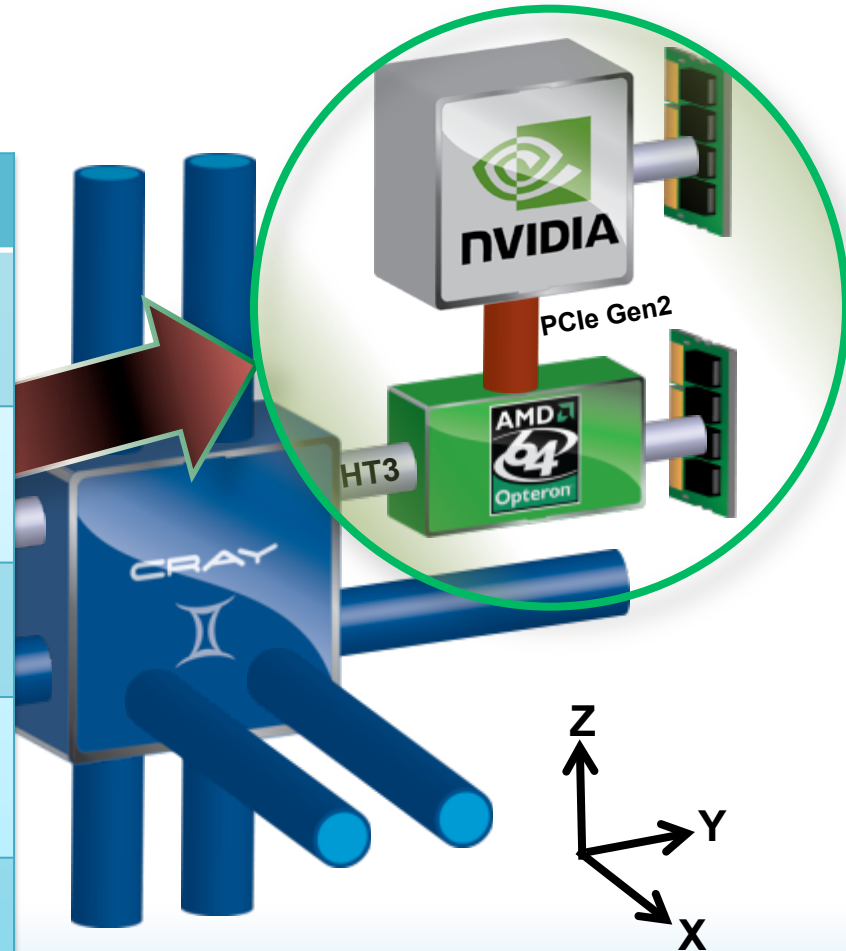
**Each core module includes 1 256-bit wide FP unit and 2 integer units. This is often advertised as 2 cores, leading to a 32 core node.*

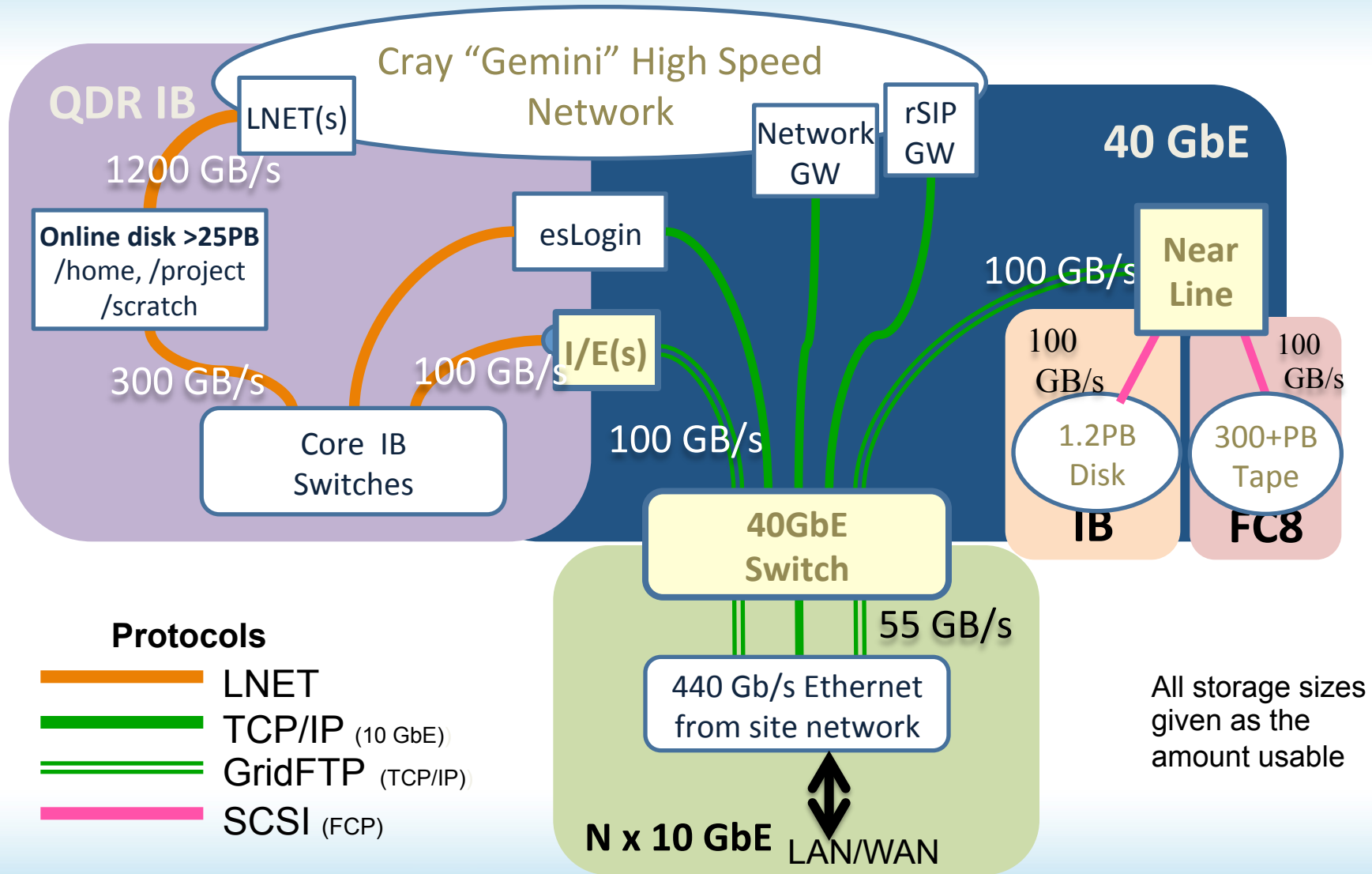
Cray XK7 and a Path to the Future

Blue Waters contains 3,072 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)	1.4 Tflops
Host Memory	32GB 51 GB/sec
Kepler Memory	6GB GDDR5 capacity 180 GB/sec





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
- **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

ESS Basic Testing

Code/Test	Measured	
Non-PS MILC (2048)	1070 sec	↑↑
Non-PS Paratec (256)	378 sec	↑↑
Non-PS WRF (512)	1448 sec	↑↑
Turbulence	406 sec	↑↑
NAMD	10.72 ms/step	↑↑
G-HPL	.981 PF (untuned)	↑↑
G-Random Access	34 GUPs	↑↑↑
G-PTRANS	1.15 TB/s	↑↑↑
G-STREAM TRIAD	0.302 PB/s	↑↑
G-FFTE	2.99 TF	↑↑
EP-DGEMMN	1.17 PF	↑↑
RandomRing Latency	3.91 μsec	↑↑↑
RandomRing Bandwidth	0.042 GB/s	↑↑
IOR Optimal Configuration	58.8 GB/s	↑↑

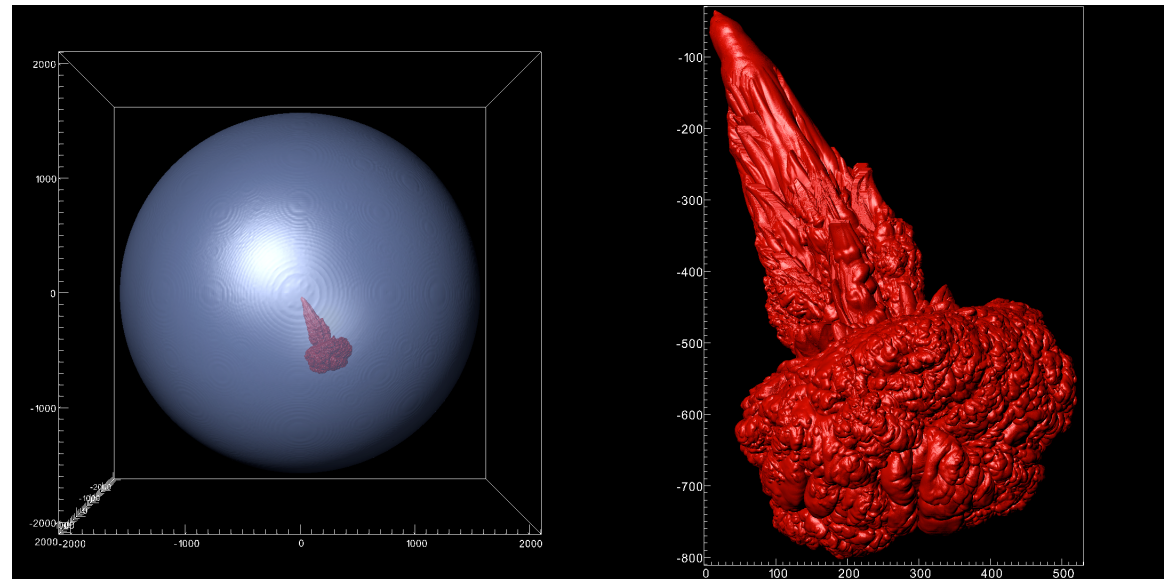
Early Science Use

PI Group	Project Title	#Jobs	#XK_Jobs	Node*hrs	Core Module*hours
Bisset	Simulation of contagion on very large social networks with Blue Waters	382	-	8,559	68,471
Campanelli	Computational relativity and gravitation at petascale: Simulating and visualizing astrophysically realistic compact	580	-	30,651	245,206
Karimabadi	Enabling Breakthrough Kinetic Simulations of the Magnetosphere via Petascale Computing	280	-	1,147,696	9,181,570
Lamm	Computational chemistry at the petascale	38	-	9,507	76,058
Nagamine	Peta-Cosmology: galaxy formation and virtual astronomy	3	-	16	129
O'Shea	Formation of the First Galaxies: Predictions for the Next Generation of Observatories	729	-	427,701	3,421,606
Pande	Simulating vesicle fusion on Blue Waters	120	-	5,404	43,230
Schulten	The computational microscope	4,459	281	1,540,632	12,325,059
<i>Stan</i>	<i>Testing hypotheses about climate prediction at unprecedented resolutions on the NSF Blue Waters system</i>	63	-	821	6,571
Sugar	Lattice QCD on Blue Waters	3,744	-	1,540,631	12,325,046
<i>Wilhelmson</i>	<i>Understanding tornadoes and their parent supercells through ultra-high resolution simulation/analysis</i>	151	-	180,267	1,442,138
<i>Woodward</i>	<i>Petascale simulation of turbulent stellar hydrodynamics</i>	96	-	309,748	2,477,981
Woosley	Type Ia supernovae	579	-	834,148	6,673,181
Wuebbles	Using petascale computing capabilities to address climate change uncertainties	372	3	482,427	3,859,413
<i>Yeung</i>	<i>Petascale computations for complex turbulent flows</i>	287	0	52678.6	421428.8

Bold - Initial 6 ESS teams, Italics - next 4 ESS teams

Understanding a Type Ia Supernova Explosion

- Off center ignition (~41 km) at ~1 second duration
- MAESTRO code and CASTRO finite volume AMR code
- 16K to 32K core modules
- About 850K nodes hours (68M core hours)
- 130m/zone resolution - ~ 1B zones at highest resolution
- 350 GB restart files
- 6 simulations totals 45 TB of data

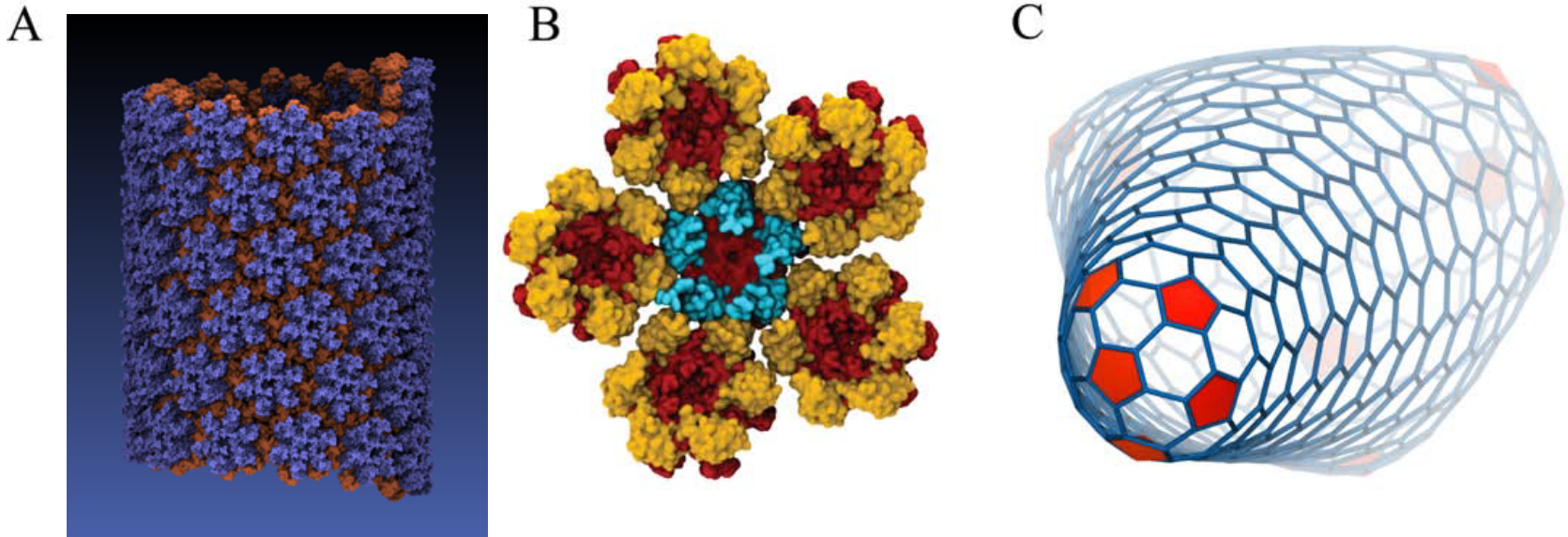


A snapshot of the evolution of the buoyant flame after about 0.6 seconds of evolution. The left panel show the stellar surface in blue and the surface of the flame in read. The right panel is a zoom-in of the flame surface, which exhibits significant small scale structure. The unit of the axes are in kilometers.

Image courtesy of Woosley team

NAMD - HIV Virus Infecting a Cell

Images courtesy of NAMD team



(A) All-atom model of the hexameric form of the CA protein as found in cylindrical assemblies of HIV capsids in vitro. This model has been determined using Blue Waters, and experimental measurements (x-ray and cryo-EM experiments). (B) The structure of a system containing one CA-pentamer surrounded by five CA-hexamers was also successfully determined and constitutes the first step towards assembly of a native HIV capsid. (C) Schematic diagram of a native HIV capsid showing a combination of pentameric (red) and hexameric (blue) centers.

NAMD - Chromosphere

Snapshot from the simulation of a 20-million atom photosynthetic membrane. The blue and green proteins absorb sunlight and transport its energy to the red proteins, which turn the energy into an electrical voltage across the membrane. The membrane is made of, besides the proteins mentioned, lipids, shown in brown. The simulation is the largest realistic all-atom molecular dynamics simulation ever carried out and answered how the initial step of photosynthesis, namely the harvesting of sunlight, is organized. This simulation was the subject of the original Blue Waters benchmark test and was carried out to complete the early tests. Computer resources used are those from the Japanese Tsubame GPU accelerated computer, DOE computers at ORNL, and Blue Waters.

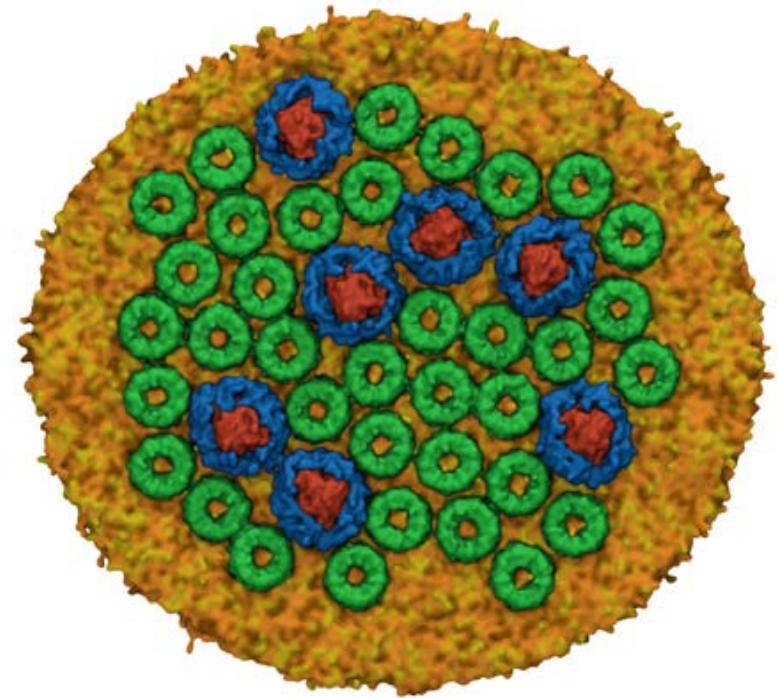


Image courtesy of NAMD team

NAMD - Ribosome Making a Protein

Preliminary structural model of a ribosome-trigger factor complex in which a nascent protein (green), emerging from the ribosome (yellow and cyan), is protected by the bound trigger factor (orange).

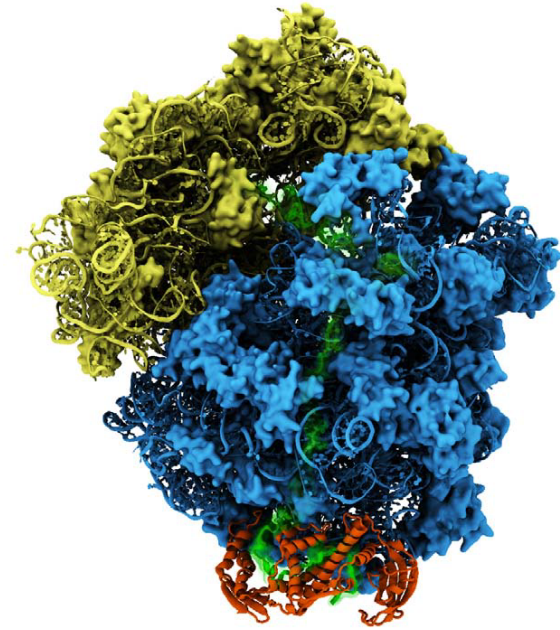
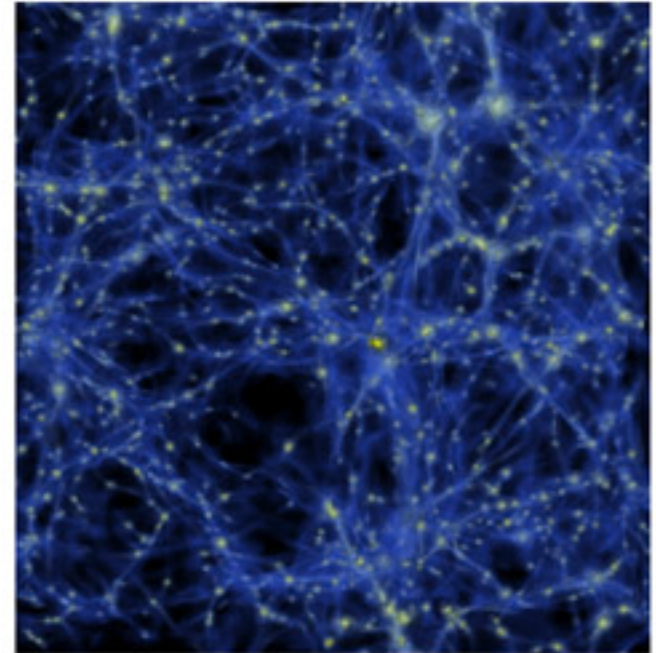


Image courtesy of NAMD team



Simulation of the formation and evolution of galaxies formed shortly after the Big Bang. Data courtesy the Brian O'Shea PRAC team. Visualization courtesy of Blue Waters visualization staff, Rob Sisneros and Dave Semeraro.

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

Installation



BW Sustained Petascale Performance Measures

- Original NSF Benchmarks
 - Full Size – QCD (MILC), Turbulence (PNSDNS), Molecular Dynamics (NAMD)
 - Modest Size – MILC, Paratec, WRF
- SPP – is a time to solution metric that is using the planned applications on representative parts of the Science team problems
 - Represents end to end problem run including I/O, pre and post phases, etc.
 - Coverage for science areas, algorithmic methods, scale
- SPP Application Mix (details and method available)
 - NAMD – molecular dynamics
 - MILC, Chroma – Lattice Quantum Chromodynamics
 - VPIC, SPECFEM3D – Geophysical Science
 - WRF – Atmospheric Science
 - PPM – Astrophysics
 - NWCHEM, GAMESS – Computational Chemistry
 - QMCPACK – Materials Science
- At least three SPP benchmarks run at full scale
- XK nodes have to add 15% more SPP

Effective Use of Complex Systems

- Additional effort by the science and engineering teams needed to achieve full potential of Blue Waters (and similar) systems, with all its advanced technology
- Increasing performance requires dramatic increases in parallelism that then generates complexity challenges for science and engineering teams
 - **Scaling applications to large core counts on general-purpose CPU nodes**
 - **Effectively using accelerators and “many-core” processors**
 - **Using homogeneous general purpose and heterogeneous (accelerated) nodes in a single, coordinated simulation**
 - **Effectively using parallel IO systems for data-intensive applications and innovative storage and data paradigms**
 - **Effectively using limited bandwidth of interprocessor network**
 - **Enhancing application flexibility to increasing effective, efficient use of systems**

Performance and Scalability

- The problem is fewer applications are able to scale in the face of limited bandwidths. Hence the need to work with science teams and technology providers to
 - Develop better process-to-node mapping using for graph analysis to determine MPI behavior and usage patterns.
 - Topology Awareness in Applications and in Resource Management
 - Improve use of the available bandwidth (MPI implementations, lower level communication, etc.).
 - Consider new algorithmic methods
 - Considering alternative programming models that improve efficiency of calculations

Performance and Scalability

- Use of heterogeneous computational units
 - While more than $\frac{1}{2}$ of the science has some GPU based investigations, only a few are using GPUs in production science
 - Many applications are GPUized only in a very limited way
 - Few are using GPUs at scale (more GPU resources are relatively small scale with limited networks)
 - Help the science teams to make more effective use of GPUs consists of two major components.
 - Introduce compiler and library capabilities into the science team workflow to significantly reduce the programming effort and impact on code maintainability.
 - OpenACC support is the major path to more general acceptance
 - Load balancing at scale
- Storage Productivity
 - Interface with improved libraries and middle ware
 - Modeling of I/O
 - On-line and Near-line transparent interfaces

Application Flexibility

- Using both XE and XK nodes in single applications
 - For multi-physics applications that provide a natural decomposition into modules is to deploy the most appropriate module(s) different computational units.
 - For applications use the Charm++ adaptive runtime system, heterogeneity can be handled without significant changes to the application itself.
 - Some applications naturally involve assigning multiple blocks to individual processors include multiblock codes (typically in fluid dynamics), and the codes based on structured adaptive mesh refinement.
 - The application-level load balancing algorithms can be modified to deal with the performance heterogeneity created by the mix of nodes.
- Malleability
 - Understanding topology given and maximizing effectiveness
 - Being able to express desired topology based on algorithms
 - Mid ware support

Flexibility - Application Based Resiliency

- Multiple layers of Software and Hardware have to coordinate information and reaction
- Analysis and understanding is needed before action
- Correct and actionable messages need to flow up and down the stack to the applications so they can take the proper action with correct information
- Application Situational Awareness - need to understand circumstances and take action
- Flexible resource provisioning needed in real time
- Interaction with other constraints so sub-optimization does not adversely impact overall system optimization

Current State of Enhanced Approaches

- Blue Waters has redirected up to 6% of its hardware funding to support enhanced intellectual services
- Blue Waters is providing direct funding to current science teams to improve their applications in ways that would not normally take place
 - Some are dramatically re-engineering their applications
 - Complete - ~\$1.5M
- Expanding educational and outreach efforts
 - Expanded virtual school classes, internships and fellowships
 - Expanding community engagements
 - Near approval completion - ~\$3.8M
- Matching technology providers with technology consumers
 - New methods and techniques that directly improve the sustained performance of one or more science teams
 - Plan is in development – expect to complete by end of summer

Observations on Co-Design

- Blue Waters has been doing co-design before co-design was a term
- Focus on Interconnect and SW – not processors
- Much different process for vendors – many vendor staff do not understand co-design
- Independent co-design teams is counter productive
 - Competing inputs – now resolution path
 - Many implicit requirements for “productization”
- Transparency in cost tradeoffs has to be much better
- Much better and more application modeling
- Resourcing and risk management is very key
- Most of the burden of re-design is still on applications not the system providers
- Need to explicitly fund application teams for co-design and modeling or they will not be able to pay enough attention

Blue Waters Sustaining Goals

- Deploy a capable, balanced system of sustaining more than one petaflops or more for a broad range of applications
- Enable Science Teams to take full advantage of sustained petascale systems
- Enhance the operation and use of the sustained petascale system
- Provide a world-class computing environment for the petascale system
- Exploit advances in innovative computing technology
- Provide National Leadership



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 Contributions From

- Thom Dunning, Marc Snir, Wen-mei Hwu, Bill Gropp
- Cristina Beldica, Brett Bode, Michelle Butler, Greg Bauer, Mike Showerman, John Melchi, Scott Lathrop, Merle Giles
- Sanjay Kale, Ravi Iyer, David Padua
- The Blue Waters Project Team and our partners
- NSF/OCI
- Cray, Inc, AMD, NVIDIA, Xyratex, Adaptive, Allinea

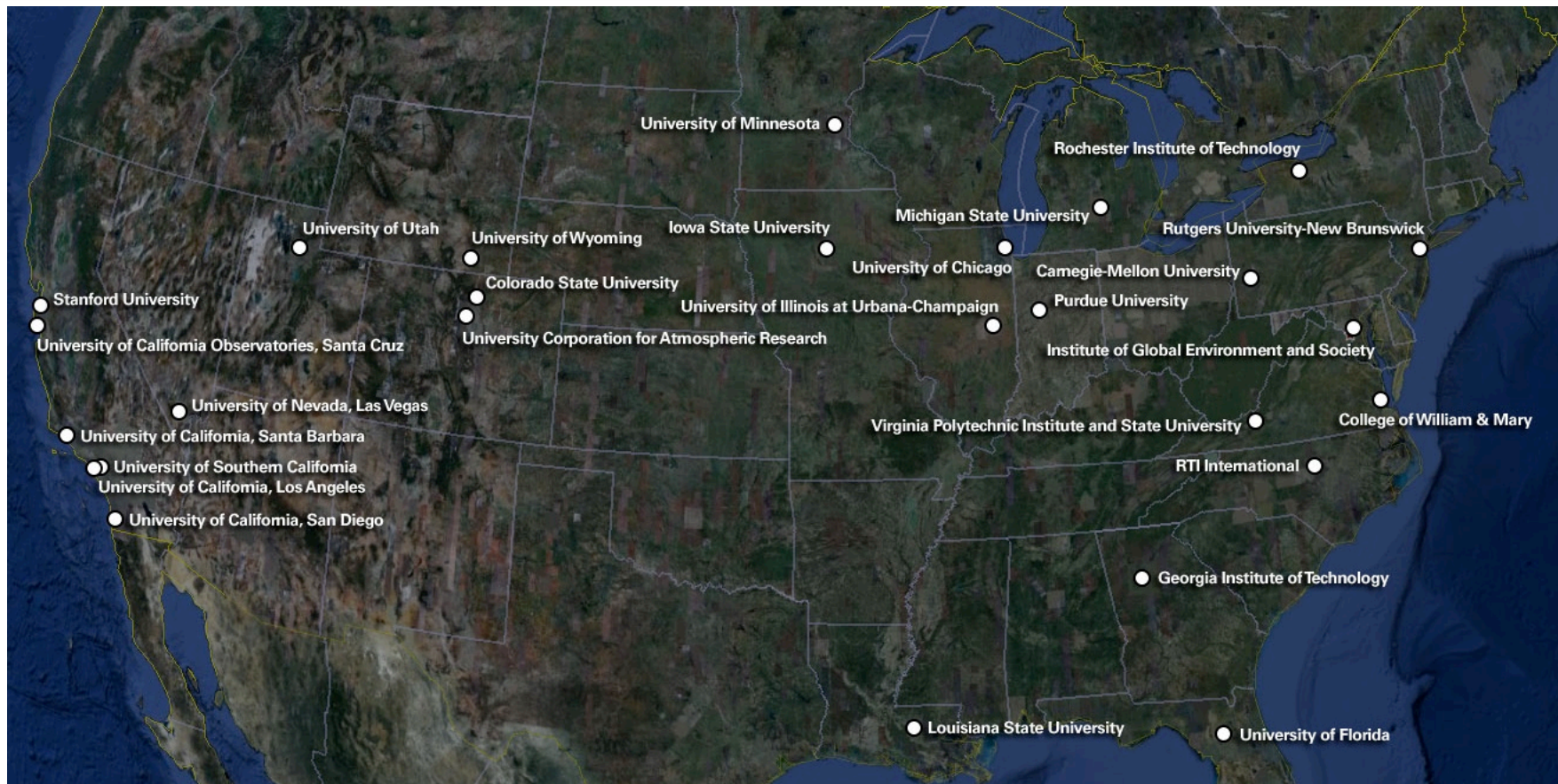
PI	Award Date	Project Title
Sugar	04/15/2009	Lattice QCD on Blue Waters
Bartlett	04/15/2009	Super instruction architecture for petascale computing
Nagamine	04/15/2009	Peta-Cosmology: galaxy formation and virtual astronomy
Bissett	05/01/2009	Simulation of contagion on very large social networks with Blue 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

PI	Award Date	Project Title
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
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

PI	Award Date	Project Title
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

PI	Award Date	Project Title
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

PRAC PI Institutions

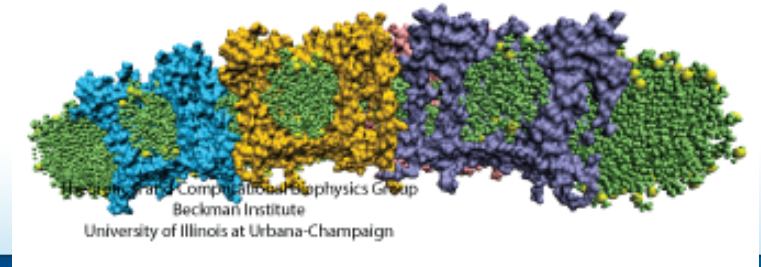
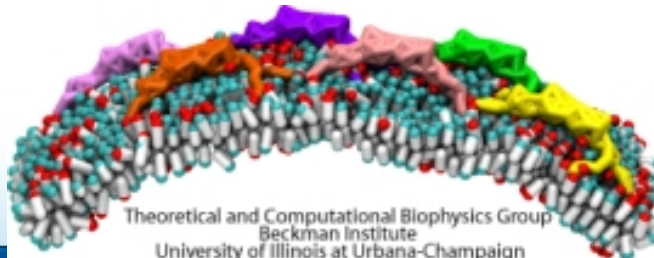


Note – UIUC is only institution leading more than one PRAC (4)
UIUC has Co-PIs on several others

Computational Microscopy

Klaus Schulten (UIUC)

- Simulations of 4 different biological systems
 - Protein elongation in the ribosome
 - Structural transitions in poliovirus entry
 - Sculpting cellular membranes by BAR domains
 - Energy conversion by the chromatophore organelle
- NAMD molecular dynamics code
 - System size will exceed 100M atoms for the first time





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
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

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	X	X		X		X		
Plasmas/ Magnetosphere	2	H3D(M), OSIRIS, Magtail/UPIC	X				X		X	X
Stellar Atmospheres and Supernovae	2	PPM, MAESTRO, CASTRO, SEDONA	X			X		X		X
Cosmology	2	Enzo, pGADGET	X			X	X			
Combustion/ Turbulence	1	PSDNS	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, MILD, USQCD	X		X	X	X		X	
Social Networks	1	EPISIMDEMICS								
Evolution	1	Eve								
Computer Science	1			X	X	X			X	X