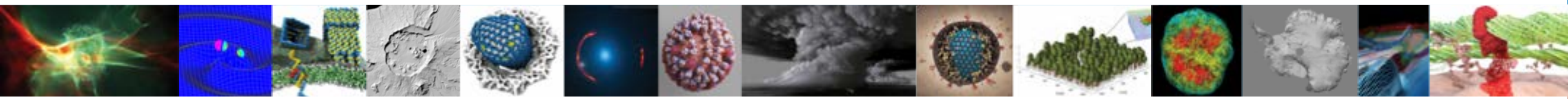
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Blue Waters Overview



NCSA

NSF

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY

Welcome to an overview of Blue Waters

- ❑ Our goal is to introduce you to the **Blue Waters Project** and the **opportunities** to utilize the resources and services that it offers

- ❑ We welcome questions through the live **YouTube chat**, **Slack** as well as **email**

help+bw@ncsa.illinois.edu

https://bluewaters.ncsa.illinois.edu/**blue-waters**

Brett Bode

INTRODUCTION

Blue Waters

- **Most capable** supercomputer on a University campus
- Managed by the **Blue Waters Project** of the **National Center for Supercomputing Applications** at the University of Illinois
- Funded by the **National Science Foundation**

Goal of the project

Ensure researchers and educators can advance discovery in all fields of study

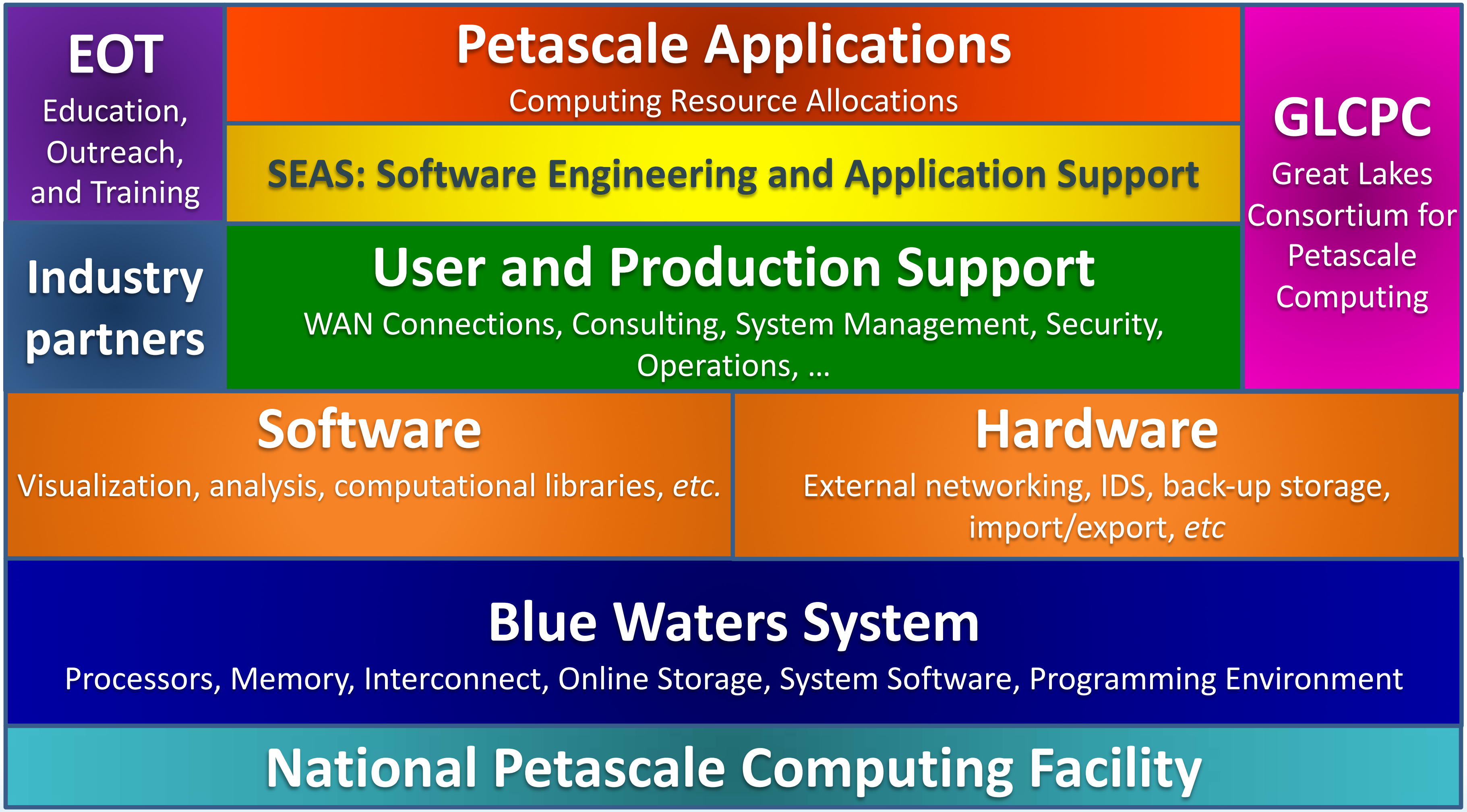
Blue Waters System

Top-ranked system in all aspects of its capabilities

Emphasis on **sustained performance**

- Built by **Cray** (2011 – 2012).
- **45% larger than any other system** Cray has ever built
- By far **the largest NSF GPU resource**
- Ranks among **Top 10** HPC systems in the world in peak performance **despite its age**
- **Largest memory capacity** of any HPC system in the world: **1.66 PB** (PetaBytes)
- One of the **fastest file systems** in the world: more than **1 TB/s** (TeraByte per second)
- **Largest backup system** in the world: more than **250 PB**
- **Fastest external network capability** of any open science site: more than **400 Gb/s** (Gigabit per second)

Blue Waters Ecosystem



EOT

Education, Outreach, and Training

Industry partners

Petascale Applications

Computing Resource Allocations

SEAS: Software Engineering and Application Support

User and Production Support

WAN Connections, Consulting, System Management, Security, Operations, …

GLCPC

Great Lakes Consortium for Petascale Computing

Software

Visualization, analysis, computational libraries, *etc.*

Hardware

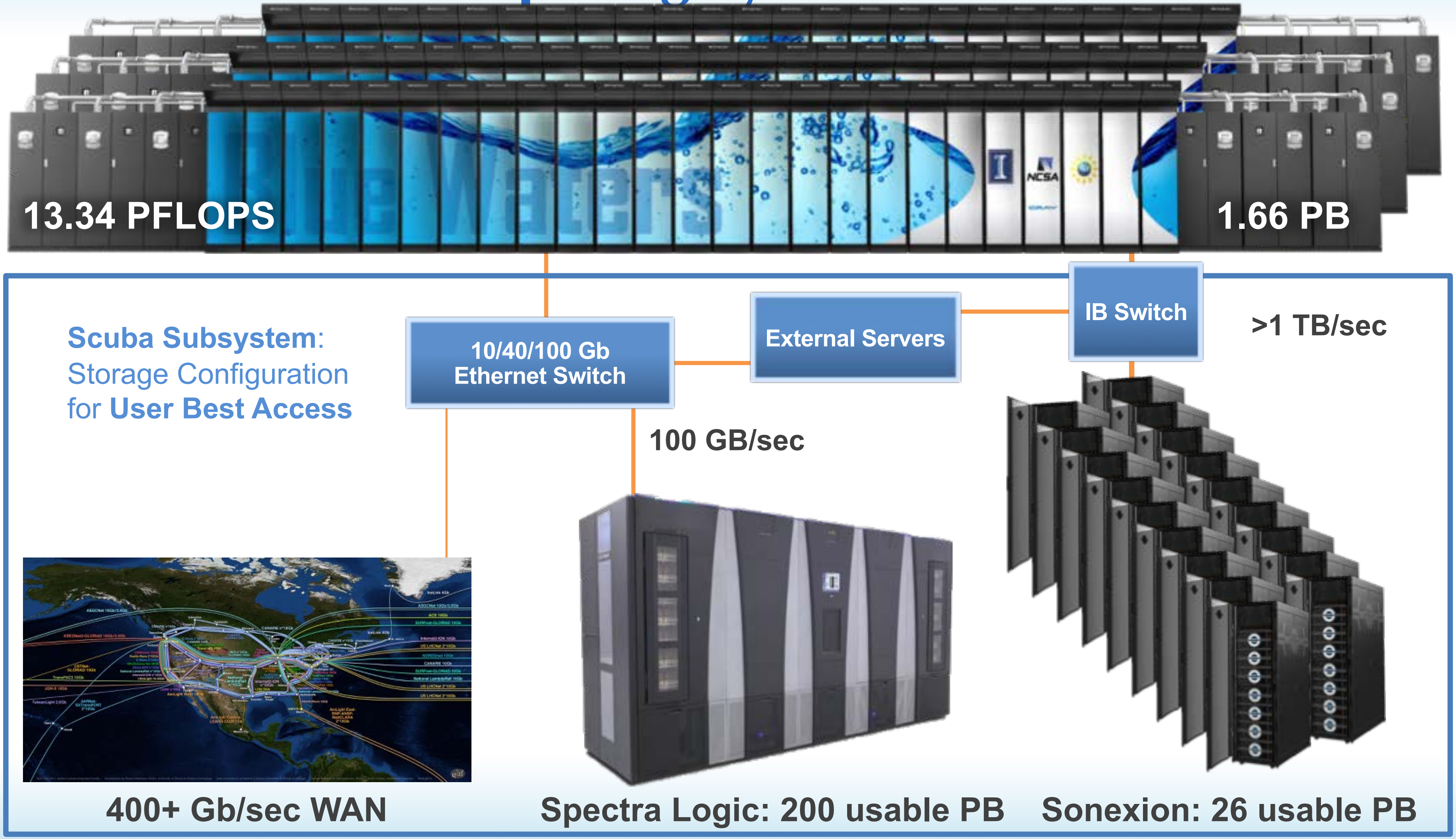
External networking, IDS, back-up storage, import/export, *etc*

Blue Waters System

Processors, Memory, Interconnect, Online Storage, System Software, Programming Environment

National Petascale Computing Facility

Blue Waters Computing System



13.34 PFLOPS

1.66 PB

Scuba Subsystem: Storage Configuration for **User Best Access**

10/40/100 Gb Ethernet Switch

External Servers

IB Switch

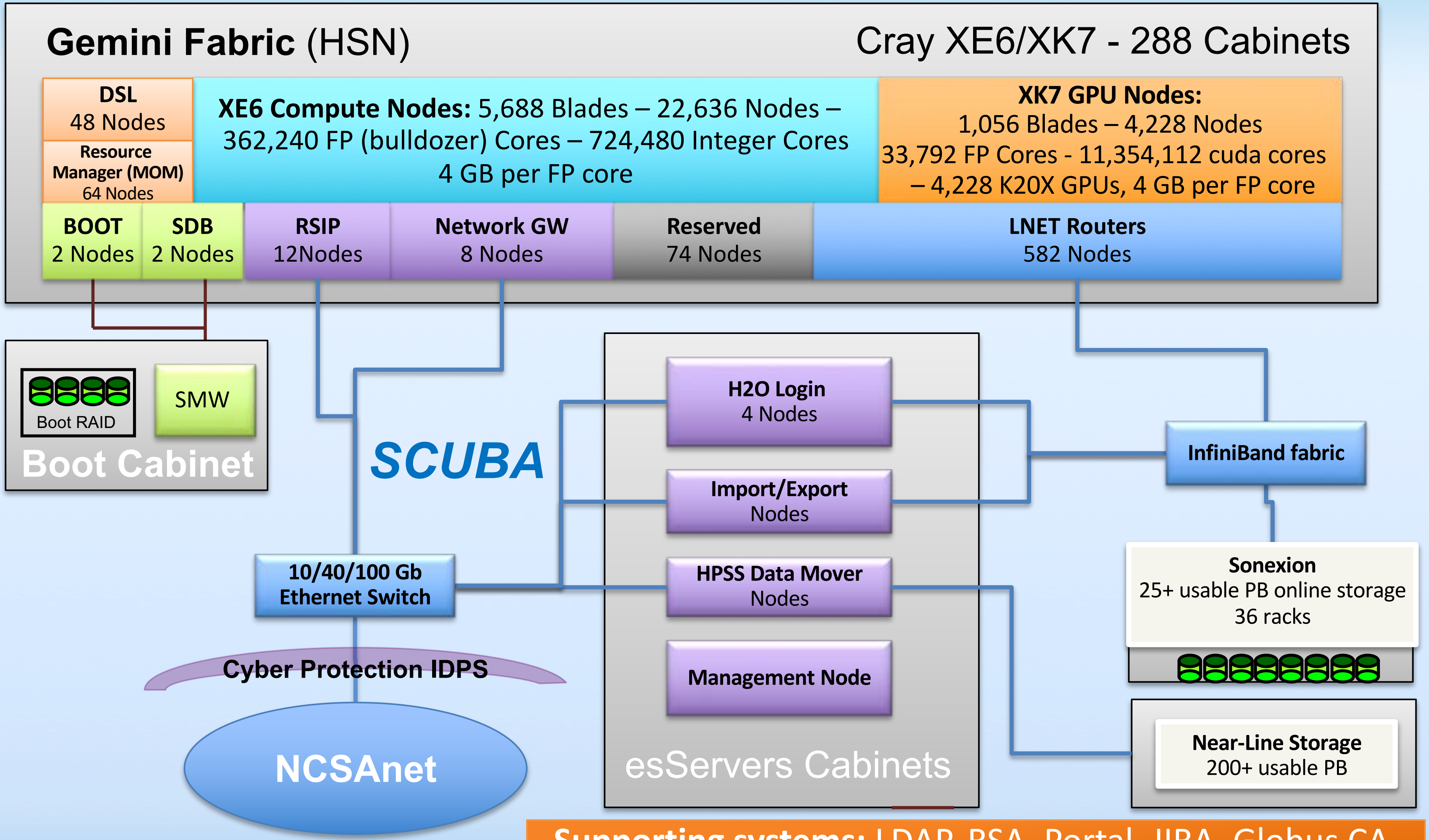
>1 TB/sec

100 GB/sec

400+ Gb/sec WAN

Spectra Logic: 200 usable PB

Sonexion: 26 usable PB



Gemini Fabric (HSN)

Cray XE6/XK7 - 288 Cabinets

DSL
48 Nodes

Resource Manager (MOM)
64 Nodes

XE6 Compute Nodes: 5,688 Blades – 22,636 Nodes – 362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core

XK7 GPU Nodes:
1,056 Blades – 4,228 Nodes
33,792 FP Cores - 11,354,112 cuda cores – 4,228 K20X GPUs, 4 GB per FP core

BOOT
2 Nodes

SDB
2 Nodes

RSIP
12Nodes

Network GW
8 Nodes

Reserved
74 Nodes

LNET Routers
582 Nodes

Boot RAID

SMW

Boot Cabinet

SCUBA

10/40/100 Gb Ethernet Switch

Cyber Protection IDPS

NCSAnet

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

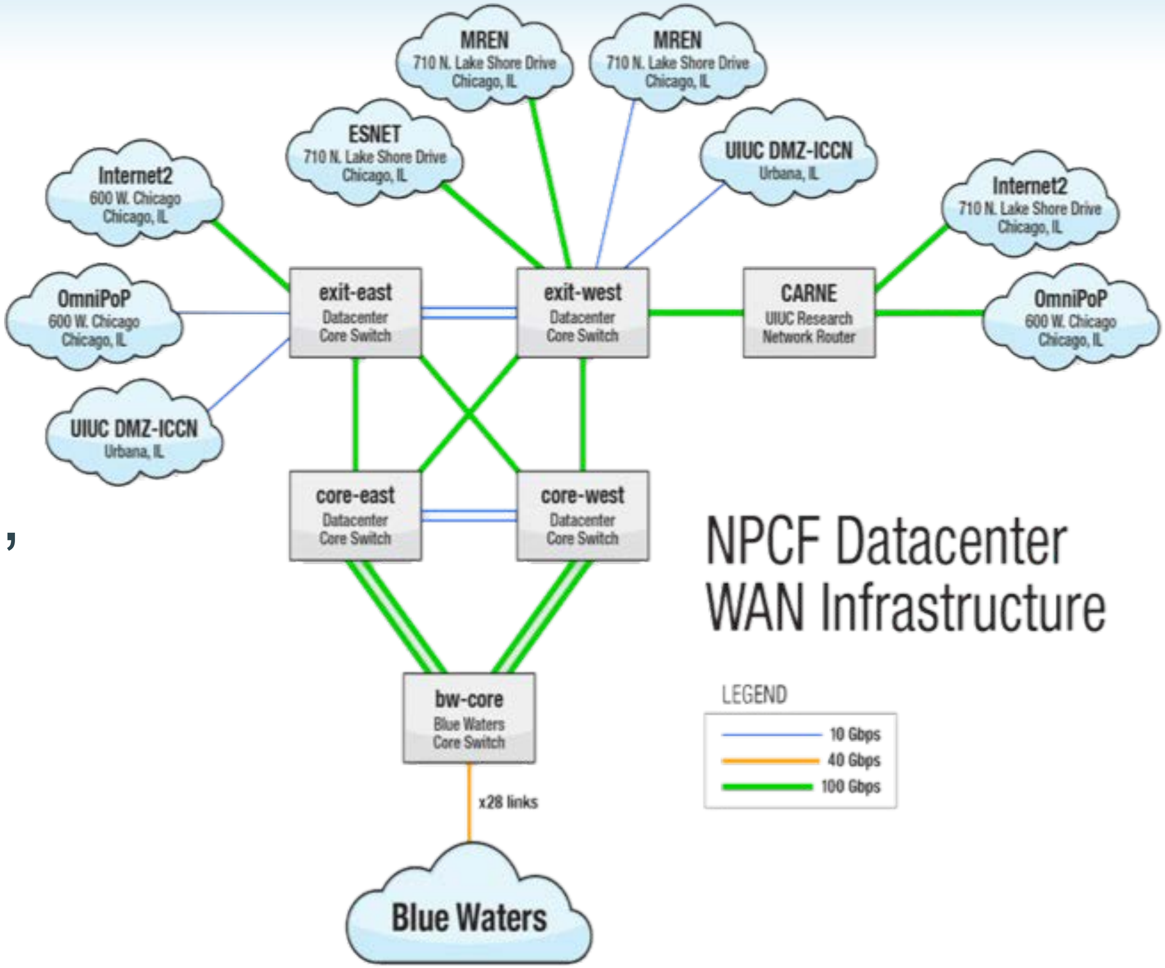
Near-Line Storage
200+ usable PB

NPCF

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

Connectivity

- Blue Waters is well connected.
- Ample bandwidth to other networks, HPC centers, universities.



Blue Waters Allocations: ~600 Active Users

NSF PRAC, 80%

- 30 – 40 teams, annual request for proposals (RFP) coordinated by NSF
- Blue Waters project does not participate in the review process

Illinois, 7%

- 30 – 40 teams, biannual RFP

GLCPC, 2%

- 10 teams, annual RFP

Education, 1%

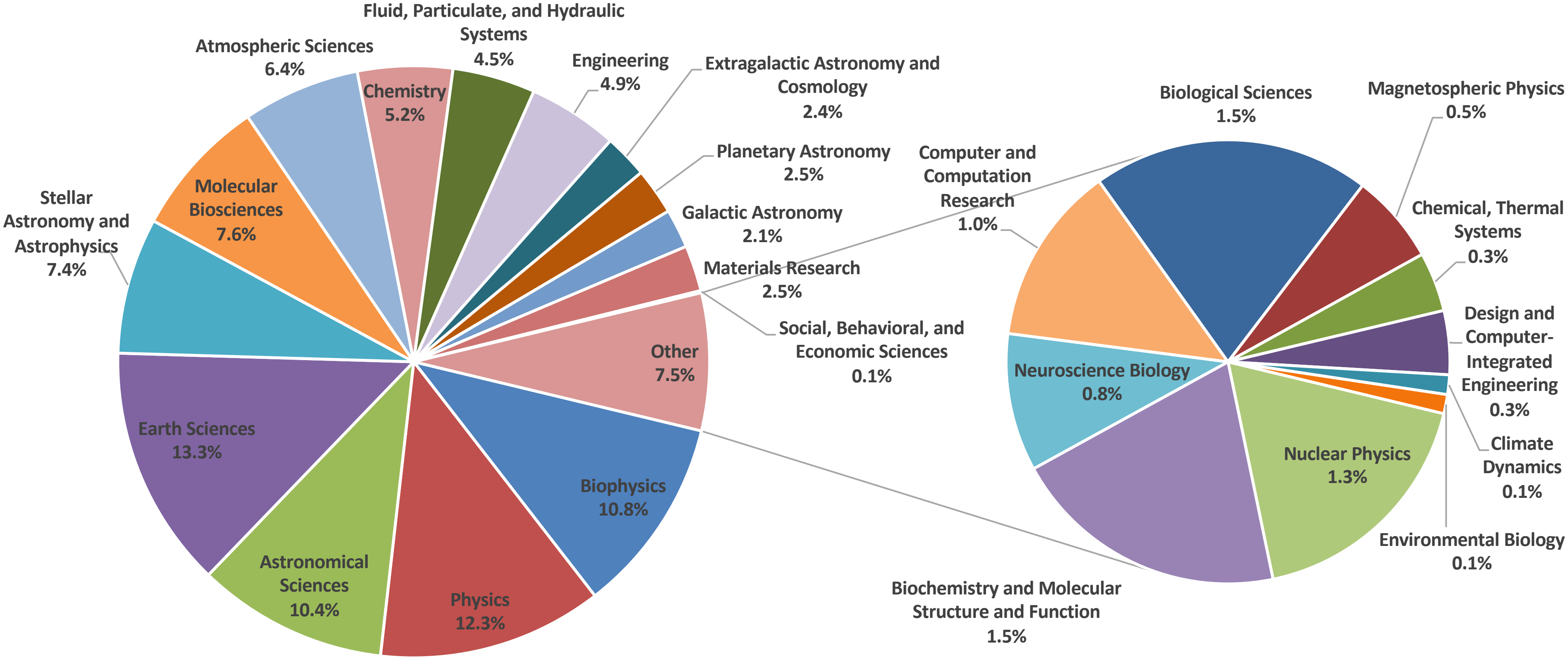
- Classes, workshops, training events, fellowships. Continuous RFP.

Industry

Innovation and Exploration, 5%

Broadening Participation, a new category for **underrepresented** communities

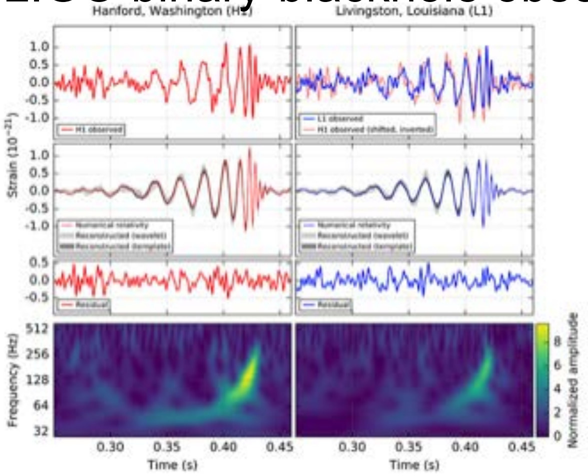
Usage by Discipline and User



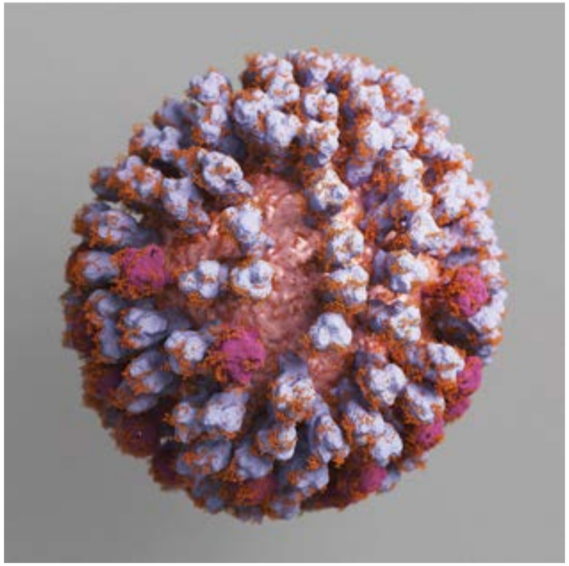
Data From Blue Waters 2016-2017 Annual Report

Recent Science Highlights

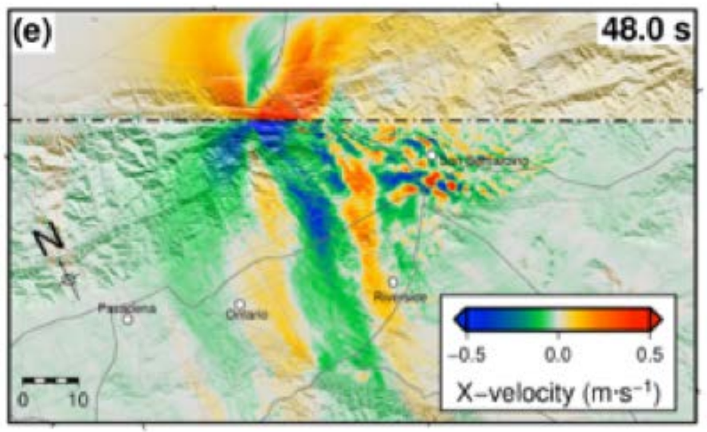
LIGO binary-blackhole observation verification

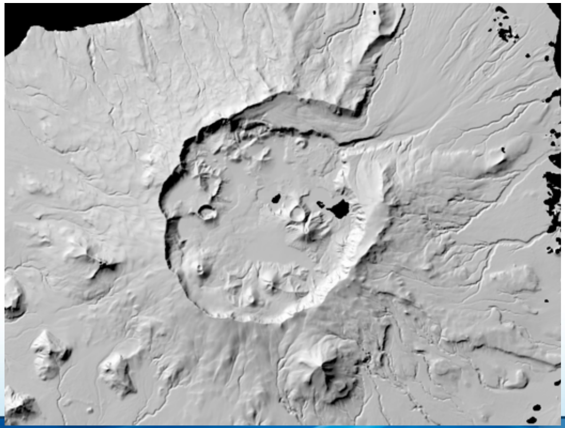


160-million-atom flu virus



Earthquake rupture





Arctic Elevation Maps

EF5 Tornado Simulation

Blue Waters Symposium

Goal Build an extreme scale community of practice among researchers, developers, educators, and practitioners

Unique annual event in June 2018 bringing together a diverse mix of people from multiple domains, institutions, and organizations

Strong Technical Program

- Over **150** people attend annually, over **50** PIs
- Over **70** talks on research achievements
- Invited plenary presentations by leaders in the field
- Technology updates and workshops by BW support team
- Posters by more than a dozen graduate students, fellows, and interns

Blue Waters Portal

https://bluewaters.ncsa.illinois.edu

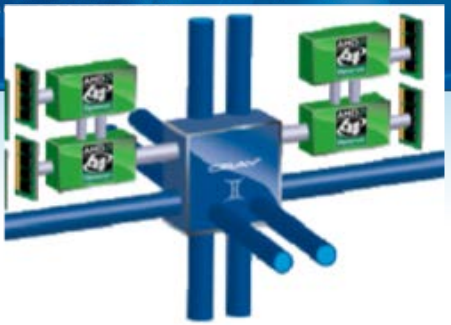
- **Allocations**
  https://bluewaters.ncsa.illinois.edu/**aboutallocations**
- **Documentation**
  https://bluewaters.ncsa.illinois.edu/**documentation**
- **User Support**
  https://bluewaters.ncsa.illinois.edu/**user-support**
- **Blue Waters Symposium**
  https://bluewaters.ncsa.illinois.edu/**blue-waters-symposium**

NSF Plans for a Follow-on System

- The funding for a follow-on machine to Blue Waters is currently under review at NSF.
- ***"Towards a Leadership-Class Computing Facility"***
  - **https://www.nsf.gov/pubs/2017/nsf17558/nsf17558.htm**
  - To deploy a system with **2–3x** the performance of Blue Waters entering service by 9/30/2019.
  - NSF PRAC allocation mechanism to remain the same, the remaining 20% TBD by the winning proposal.

Greg Bauer

BLUE WATERS SYSTEM ARCHITECTURE



Blue Waters Compute System

- Blue Waters' distributed computing system has two types of nodes (CPU and GPU) interconnected by a high-speed network.
- Low latency network for strong scaling of MPI or PGAS codes. MPI-3 support and lower level access.
- Weak scaling supported by high aggregate bandwidth of 3D torus network topology.

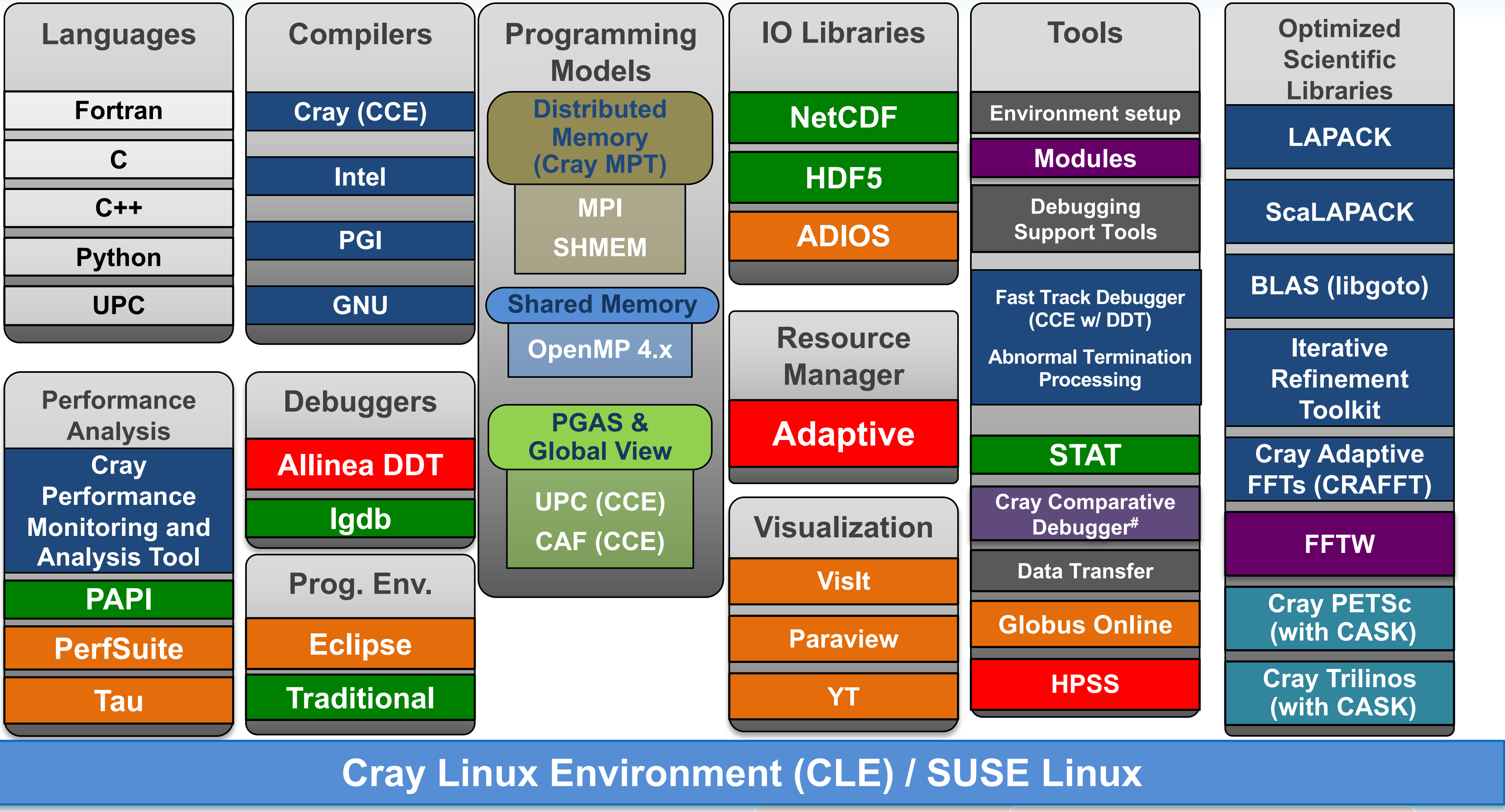
XE CPU Node Features

- Dual socket AMD “Interlagos” CPUs
- 16 floating point units and 32 cores per node.
- 64 GB RAM per node typical, 96 nodes at 128 GB.
- 102 GB/s memory bandwidth per node.
- Low OS noise for strong scaling.
- Support for MPI, OpenMP, threads, etc.

XK GPU Node Features

- One AMD CPU and one NVIDIA K20x GPU per node.
- 32 GB RAM per node typical, 96 nodes at 64 GB.
- Support for OpenCL, OpenACC and CUDA (7.5).
- CUDA MultiProcessService supported.
- RDMA message pipelining from GPU.
- Support for GPU enabled ML and visualization.

Blue Waters Software Environment



Languages
- Fortran
- C
- C++
- Python
- UPC

Compilers
- Cray (CCE)
- Intel
- PGI
- GNU

Programming Models
- Distributed Memory (Cray MPT)
  - MPI
  - SHMEM
- Shared Memory
  - OpenMP 4.x
- PGAS & Global View
  - UPC (CCE)
  - CAF (CCE)

IO Libraries
- NetCDF
- HDF5
- ADIOS

Tools
- Environment setup
- Modules
- Debugging Support Tools
- Fast Track Debugger (CCE w/ DDT)
- Abnormal Termination Processing
- STAT
- Cray Comparative Debugger[#]
- Data Transfer
- Globus Online
- HPSS

Optimized Scientific Libraries
- LAPACK
- ScaLAPACK
- BLAS (libgoto)
- Iterative Refinement Toolkit
- Cray Adaptive FFTs (CRAFFT)
- FFTW
- Cray PETSc (with CASK)
- Cray Trilinos (with CASK)

Performance Analysis
- Cray Performance Monitoring and Analysis Tool
- PAPI
- PerfSuite
- Tau

Debuggers
- Allinea DDT
- lgdb

Prog. Env.
- Eclipse
- Traditional

Resource Manager
- Adaptive

Visualization
- VisIt
- Paraview
- YT

Cray Linux Environment (CLE) / SUSE Linux

Cray developed
Under development
Licensed ISV SW

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

Support for Python and Containers

- Approx. 20% of Blue Waters users use **Python**.
- We provide over **260 Python packages** and two Python versions.
- **Support** for **GPUs**, ML/DL, etc.

- Support for "Docker-like" **containers** using Shifter.
- **MPI across nodes** with access to native driver.
- **Access to GPU** from container.
- **Support** for **Singularity** coming.

Data Science and Machine Learning

Currently available libraries

- TensorFlow 1.3.0

In the Pipeline

- TensorFlow 1.4.x
- PyTorch
- Caffe2
- Cray ML Acceleration

Data challenge: large training datasets

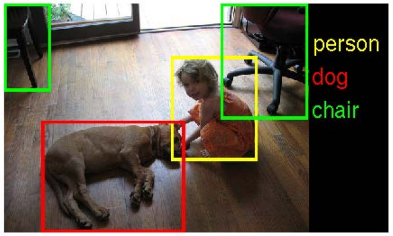
- Example/Research Data on BW
  - ImageNet
- Seeking Datasets for:
  - Natural Language Processing
    - Still looking for data set large enough
  - Biomedical dataset
    - biobank http://www.ukbiobank.ac.uk
- Seeking users interests

TensorFlow

PYTORCH

Caffe2

IMAGENET



biobank uk
Improving the health of future generations

Blue Waters Support Model

Blue Waters Partner Consulting

- Assistance with porting, debugging, allocation issues, and software requests.

Advanced Application Support for projects

- Requests are reviewed and evaluated for breadth, reach and impact.

Point of Contact (PoC)

- Major Science teams (such as NSF PRAC awards).
- Tuning, modeling, IO, optimizing application codes.
- Code restructuring, re-engineering or redesign.
- Work plans are reviewed by the Blue Waters project office.

Support for workflows, data movement, visualization.

Blue Waters Staff Expertise

Domain expertise

- Bioinformatics
- CFD (Finite Difference and Finite Element Methods)
- Computational Chemistry (NWCHEM, GAMESS US, CHARMM)
- Molecular Dynamics (NAMD, GROMACS, etc.)
- Numerical Methods
- Astrophysics

Computational expertise

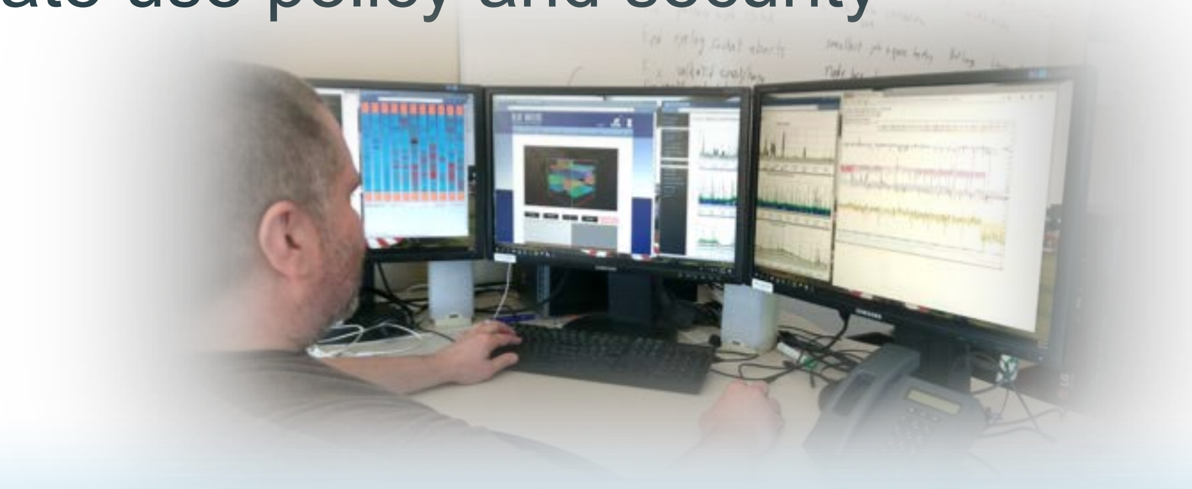
- Runtimes
- Charm++
- Einstein Toolkit
- Performance analysis
- Programming models: MPI+X

Jeremy Enos

OPERATIONS

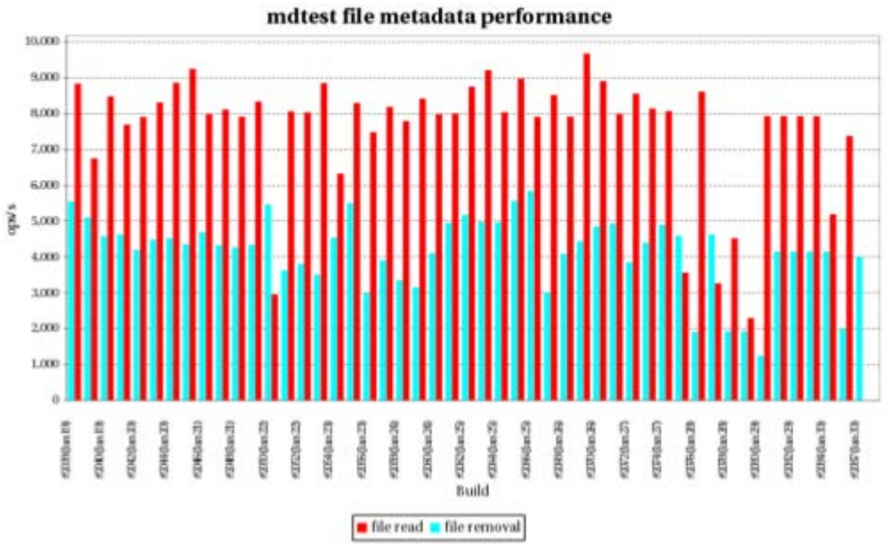
Operational Goals

- High performance, high availability
- Job scheduling policy
- Ensure best system utilization
- Enforce appropriate use policy and security



Performance and Availability

- **Regression tests** done for software and hardware, performance and function
- Aggressive monitoring and **anomaly investigation**
- Minimize interference between users
- **24/365** on-call staff to service machine
- 7+ day **advance notice** of scheduled outages

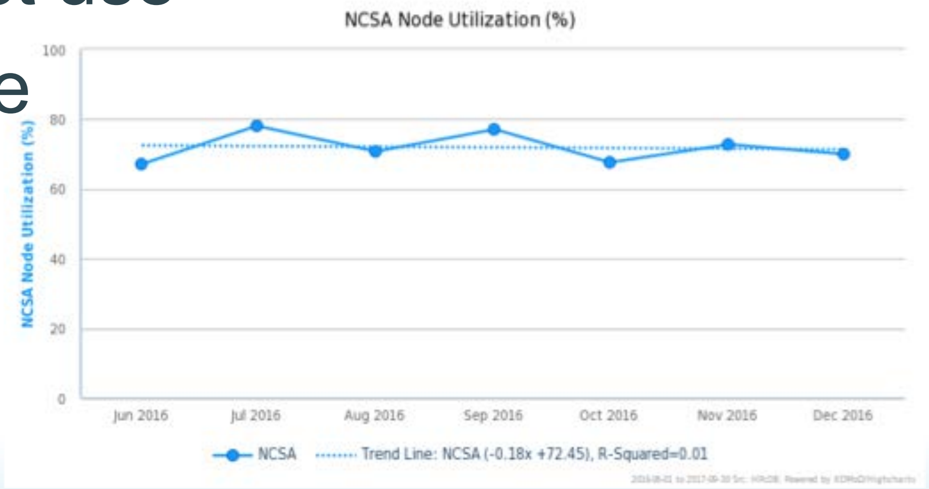


Job Scheduling

- Retain maximum job submission flexibility
- General scheduling policy favors large jobs
- High, normal, low, and debug queue priority options
- Fairness measures within general policy
- Minimize job turnaround time
- Minimum chargeable unit = 1 node
- GPU and CPU nodes have same charge
- Maximum runtime allowed = 48 hours
- Special requests (longer runtimes, advance reservations, courses, deadlines, etc.)

Ensure Best System Utilization

- Discounts for job submission designed to complement idle system portions
- Job placement by communication profile
- Provide guidance for best use
- Investigation of disruptive workflows
- Investigation of inconsistent runtimes



Security and Appropriate Use Policy

- Perfect, **zero compromise** track record
- **State-of-the-art IDS**, keystroke logging
- Two-factor authentication
- Hierarchical, **unidirectional privilege model**
- Security team also monitors for appropriate use for scientific purpose
- Extreme priority placed on security patches

Michelle Butler

DATA STORAGE AND MANAGEMENT

Online Storage

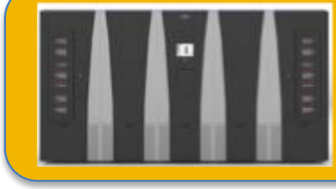
home : 2.2 PB useable : 1 TB user quota

projects: 2.2 PB useable : 5 TB group quota

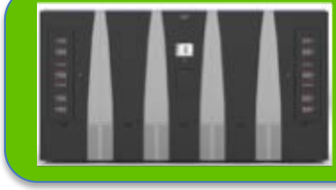
scratch: 22 PB useable : 500 TB group quota

- 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.

Nearline Storage (HPSS)



home: 5 TB quota



projects: 50 TB group quota

- **200 PB** of **usable storage space**.
- Accessed *via* **Globus Online** graphical or command line interfaces.
- Preserves **projects** *vs.* **home distinction**

Easy to Move Data to/from Blue Waters

Globus Online

- GUI, API and command line interfaces

Globus Connect Servers

- Very high bandwidth
- Asynchronous
- Very parallel
- Specialized resources for endpoints

Globus Connect Personal

- For local resources (laptop, workstation) that don’t have server running.

Rob Sisneros

SCIENTIFIC VISUALIZATION

Supporting Science on Blue Waters

Software

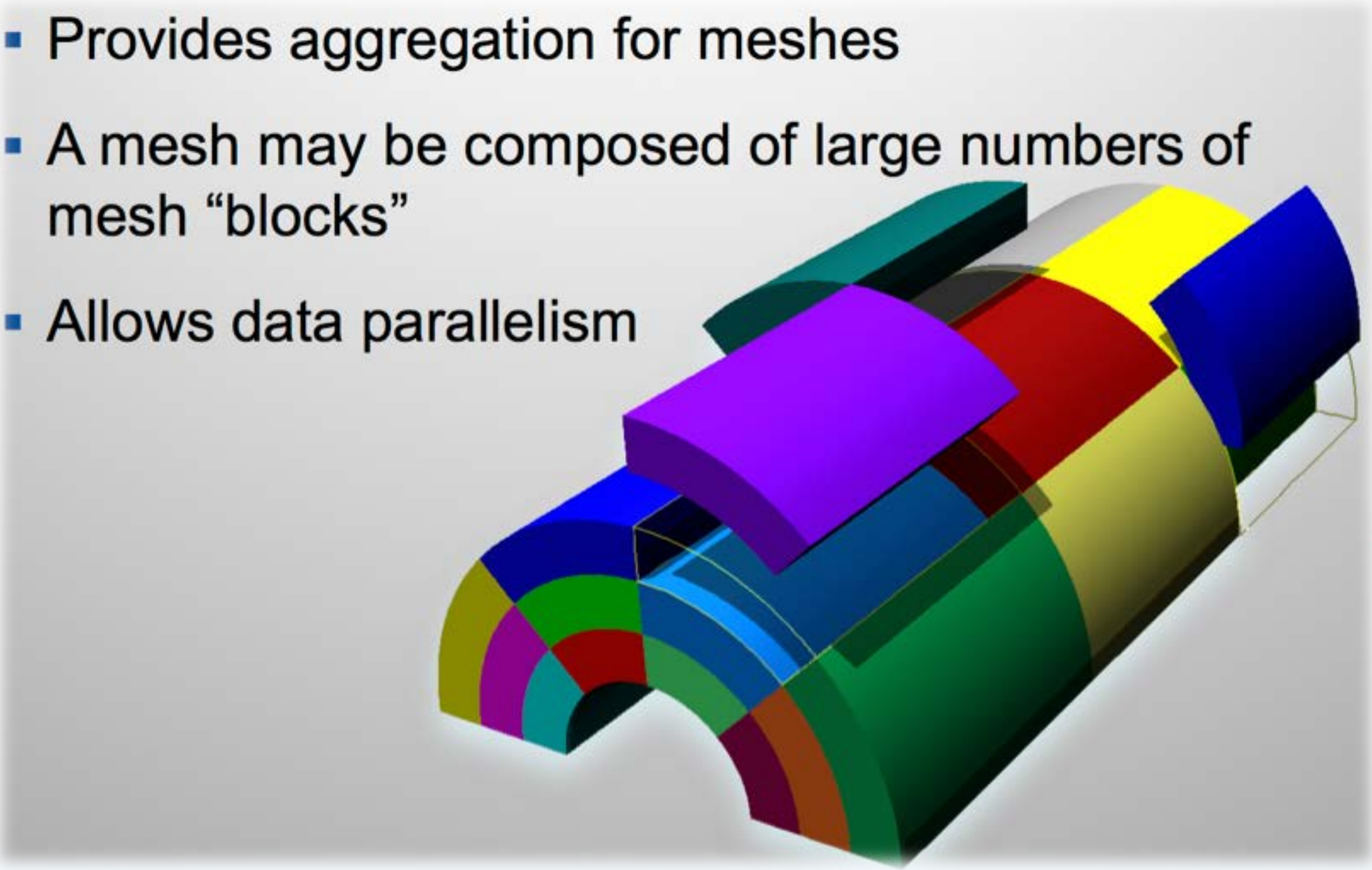
- Installation + maintenance
- Data preparation
- Usage/Training

Research

- Is this in my data?
- This is complex, can I show it?
- Visualization for HPC

Outreach: Getting data out there

How to Analyze in Parallel



- Provides aggregation for meshes
- A mesh may be composed of large numbers of mesh “blocks”
- Allows data parallelism

Supported Visualization Software

Specialized

- yt

General, scalable

- Paraview and VisIt

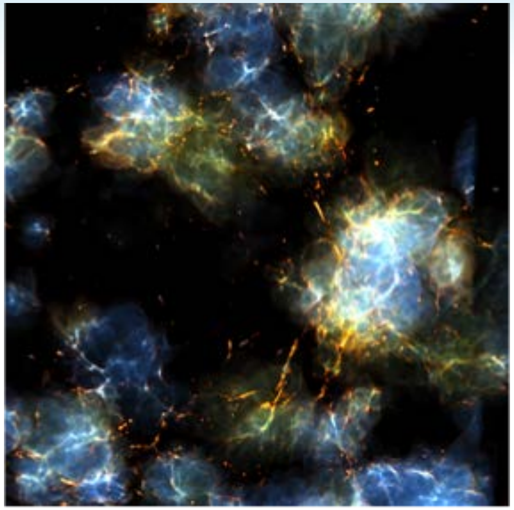
Other

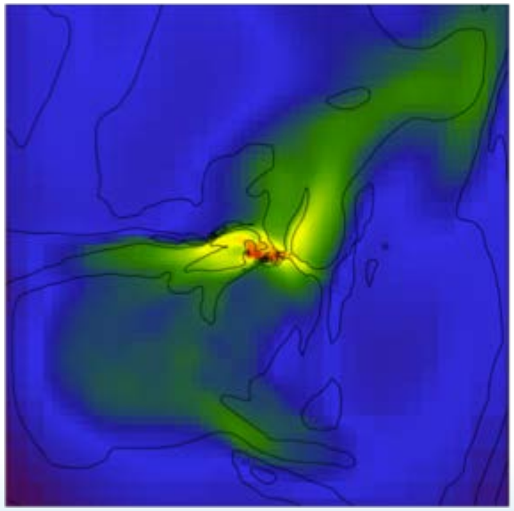
- IDL, imagemagick, other

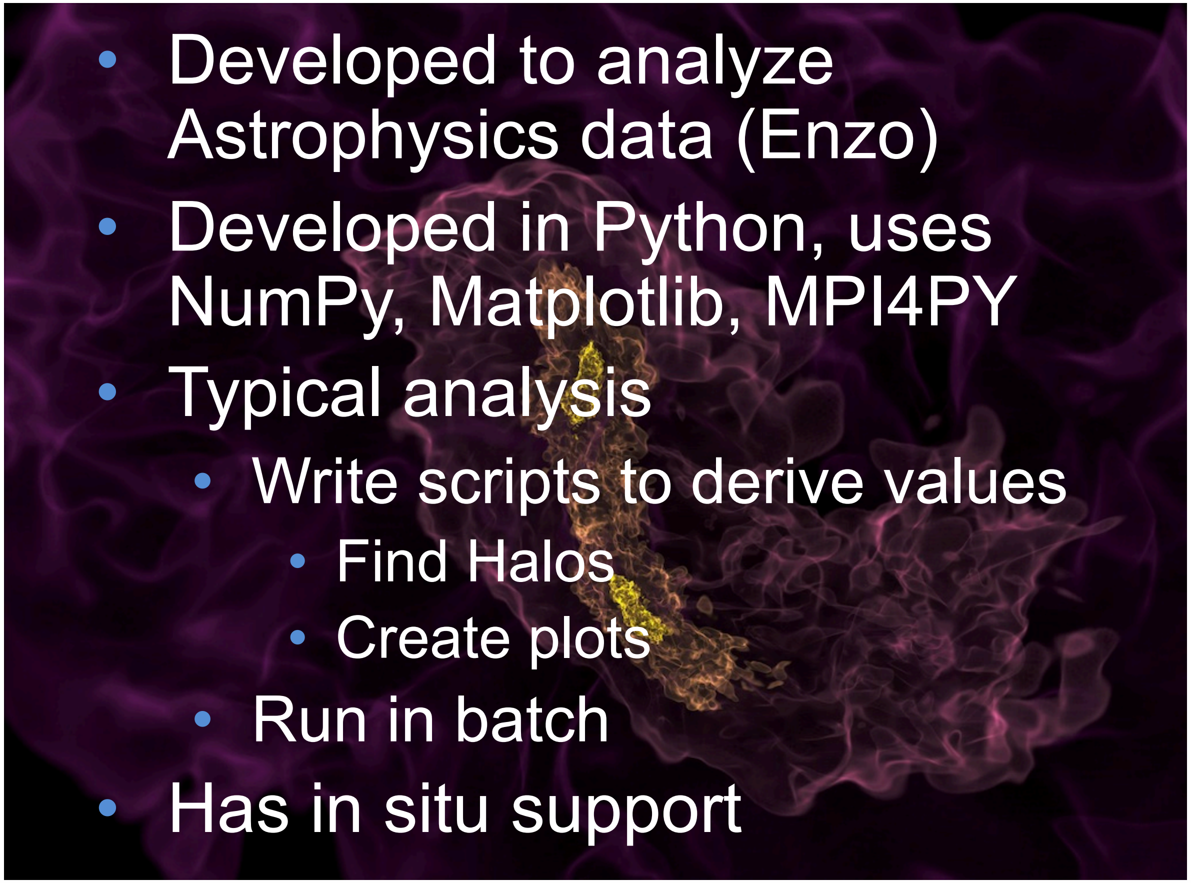
Visualization webinars available on YouTube

- Blue Waters webinar on ***yt*** on February 28

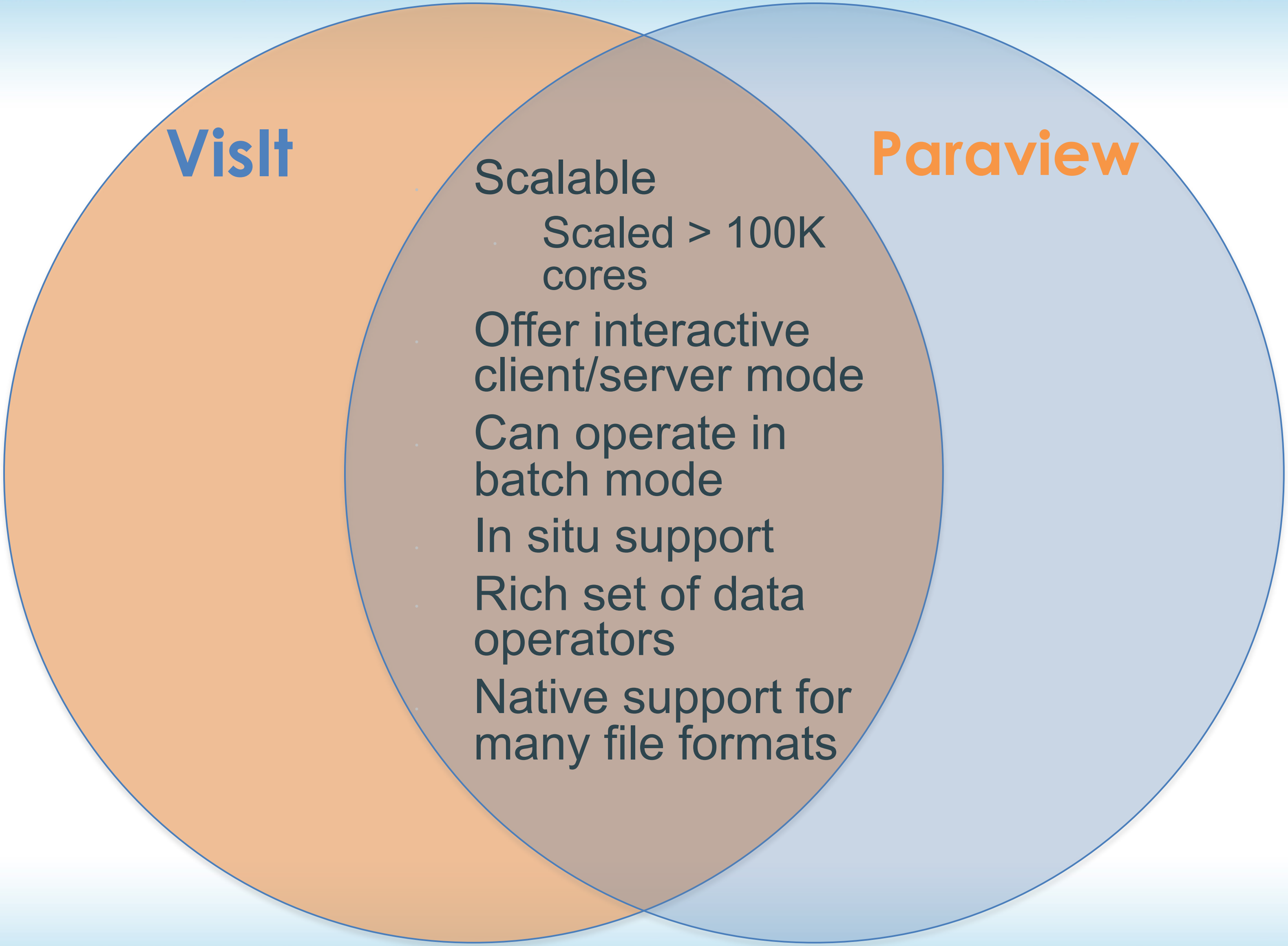
yt







- Developed to analyze Astrophysics data (Enzo)
- Developed in Python, uses NumPy, Matplotlib, MPI4PY
- Typical analysis
  - Write scripts to derive values
    - Find Halos
    - Create plots
  - Run in batch
- Has in situ support

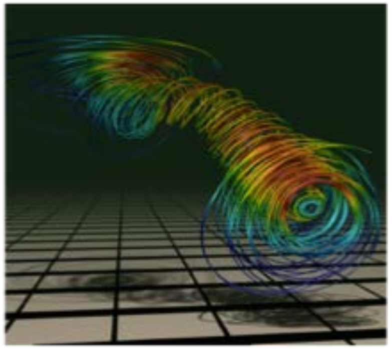


VisIt

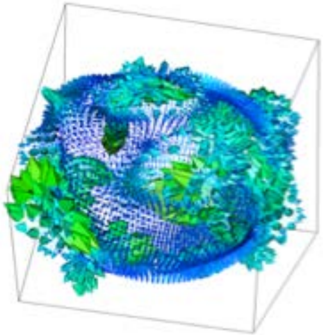
Paraview

- Scalable
  - Scaled > 100K cores
- Offer interactive client/server mode
- Can operate in batch mode
- In situ support
- Rich set of data operators
- Native support for many file formats

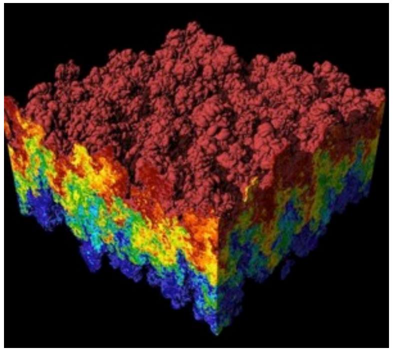
Visualization with VisIt



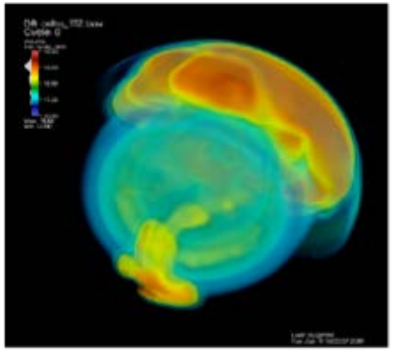
Streamlines



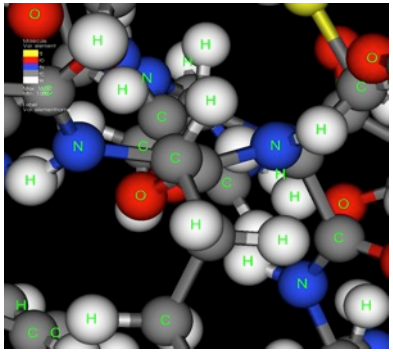
Vector / Tensor Glyphs



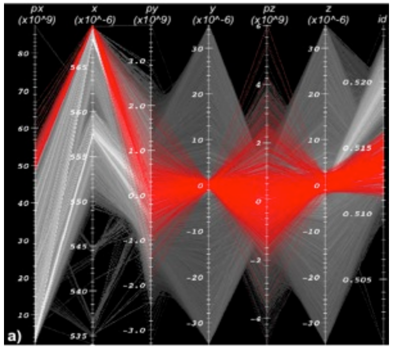
Pseudocolor Rendering



Volume Rendering

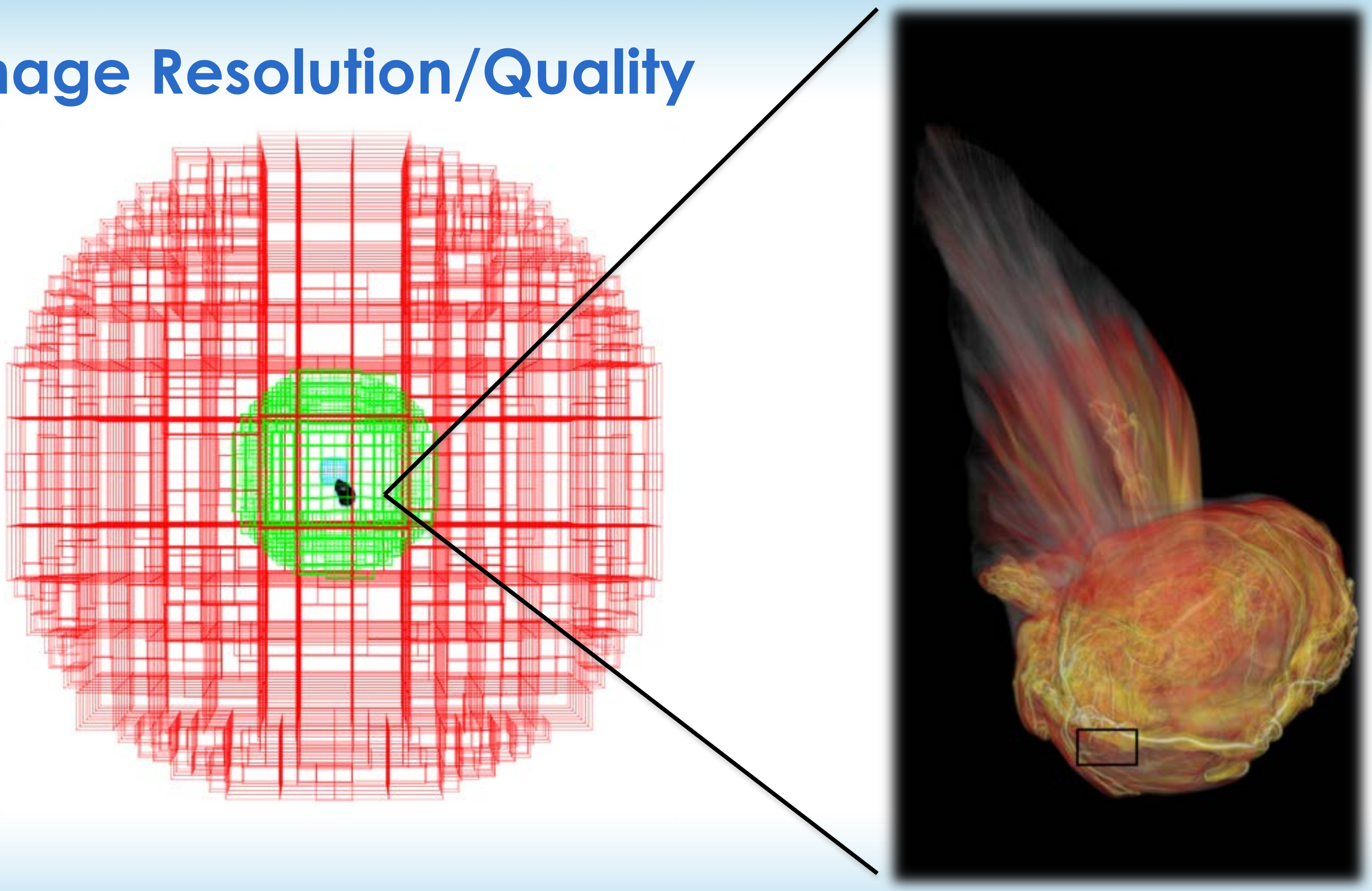


Molecular Visualization



Parallel Coordinates

Image Resolution/Quality



Maxim Belkin

BLUE WATERS TRAINING

Target Audience

Current and future **Blue Waters** users and partners

Training Goals

- Train **new users** on how to better utilize **Blue Waters** resources
- Train advanced users on **new** and **emerging technologies** (HPC container solutions, data analytics, heterogeneous programming, *etc*.)

Blue Waters Training

Webinars

- Applied and general topics
- Informational and **hands-on** sessions
- Feel free to request or suggest a topic!
- Great opportunity to get publicity!

https://bluewaters.ncsa.Illinois.edu/webinars

We support partners' training sessions and events

- Hackathons
- Distributed classrooms

Let us know your needs: **bw-eot@ncsa.Illinois.edu**

Blue Waters Training

Upcoming (hands-on) workshops and events

- Machine Learning in HPC
- Containers in HPC
- GPU Hackathon (August)
- Python in HPC (planned)

Let us know your needs: **bw-eot@ncsa.Illinois.edu**

Scott Lathrop

EDUCATION AND BROADENING PARTICIPATION ALLOCATIONS

Education Allocations

- Support the preparation of the **national workforce** with expertise in **petascale computing**.
- Projects may be requested for **up to one year**, although many will typically cover a one- to two-week period or a semester.
- Please apply at least one month before the allocation is needed.
- Requests are generally limited to at most **25,000 node-hours**
- Possible projects:
  - Focus on **large-scale datasets** and **optimization of I/O operations**.
  - Developing and testing of codes that use **advanced methods**, languages and tools
  - **Optimizing** and **scaling** of a community code to a large-scale simulation.
  - **Optimizing libraries** and tools that leverage architecture features.
  - Focusing on the **unique scale** and scope of the **Blue Waters system**.
  - Use of **large-scale computation** and **data analytics**.

Broadening Participation Allocations

- This is a **new category** open to faculty and research staff at **US academic institutions** who have not previously had a Blue Waters allocations and who are among **underrepresented communities**
- This is **a new initiative** being presented to NSF as a "prototype" program that we hope will be sustained on future NSF-supported systems.
- The guidelines for submissions will be announced in near future.

Broadening Participation Allocations

- Minority Serving Institutions
- Institutions within EPSCoR jurisdictions
- PIs who are women, underrepresented minorities, or people with disabilities
- Fields of study that are traditionally underrepresented in HPC, such as humanities, arts, and social sciences
- Graduate or undergraduate students are not eligible
- Co-PIs and collaborators from other institutions
- First time Blue Waters Allocations PIs

Broadening Participation Allocations

- Requests may be **up to 200,000 node-hours** for one year.
- Projects will be judged based on
  - **scientific merit**
  - **suitability** for Blue Waters
  - **demonstrated need** for the capabilities of Blue Waters.
- Progress reports will be required for all awards

SUMMARY

Blue Waters 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

NCSA is a leader in developing and deploying these technologies as well as contributing to community efforts.

Questions

- **General information** about Blue Waters: https://bluewaters.ncsa.illinois.edu/**blue-waters**
- For assistance with **technical questions** about the computing system, send **email** to **help+bw@ncsa.illinois.edu**
- We look forward to your participation in utilizing the **Blue Waters resources** and **services**.