

# NCAR's Data-Centric Supercomputing Environment Yellowstone

**November 28, 2011**  
**David L. Hart, CISL**  
**[dhart@ucar.edu](mailto:dhart@ucar.edu)**



# Welcome to the Petascale

- **Yellowstone hardware and software**
- **Deployment schedule**
- **Allocations opportunities at NWSC**
  - University, CSL, NCAR, and Wyoming-NCAR alliance

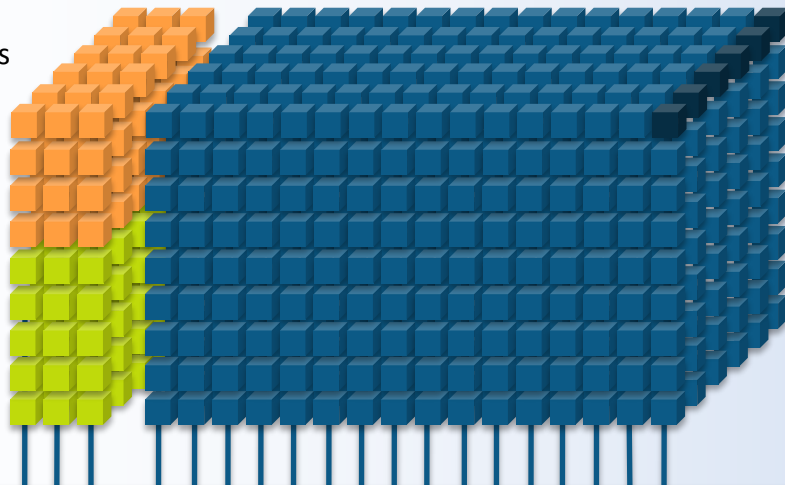
Construction complete!



# Yellowstone Environment

Computational & Information Systems Laboratory

**Geyser & Caldera**  
DAV clusters

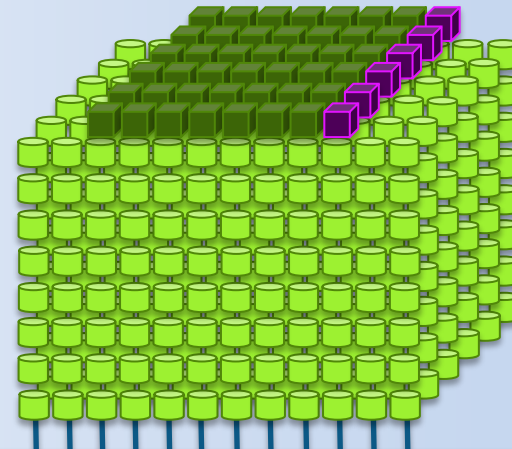


## Yellowstone

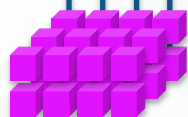
HPC resource, 1.55 PFLOPS peak

## GLADE

Central disk resource  
11 PB (2012), 16.4 PB (2014)



**High Bandwidth Low Latency HPC and I/O Networks**  
FDR InfiniBand and 10Gb Ethernet



## NCAR HPSS Archive

100 PB capacity  
~15 PB/yr growth

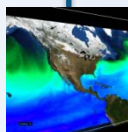


Arrived  
Nov. 4

1Gb/10Gb Ethernet (40Gb+ future)

Science Gateways  
RDA, ESG

Data Transfer  
Services



Remote Vis



Partner Sites



XSEDE Sites



# Yellowstone

## *NWSC High-Performance Computing Resource*

- **Batch Computation**

- 74,592 cores total – 1.552 PFLOPs peak
- 4,662 IBM dx360 M4 nodes – 16 cores, 32 GB memory per node
- Intel Sandy Bridge EP processors with AVX – 2.6 GHz clock
- 149.2 TB total DDR3-1600 memory
- 29.8 Bluefire equivalents

- **High-Performance Interconnect**

- Mellanox FDR InfiniBand full fat-tree
- 13.6 GB/s bidirectional bw/node
- <2.5  $\mu$ s latency (worst case)
- 31.7 TB/s bisection bandwidth

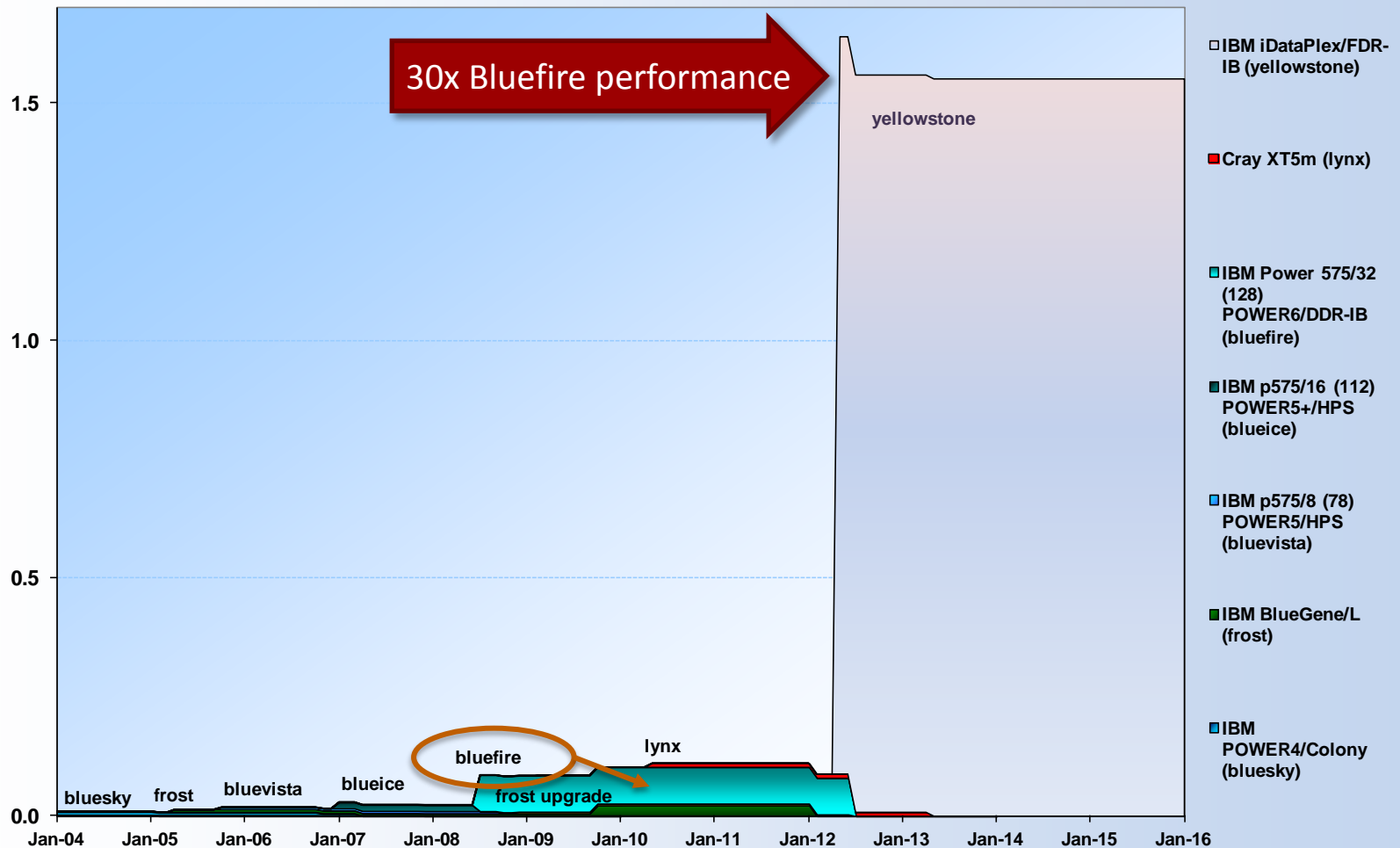
- **Login/Interactive**

- 6 IBM x3650 M4 Nodes; Intel Sandy Bridge EP processors with AVX
- 16 cores & 128 GB memory per node



# NCAR HPC Profile

## Peak PFLOPs at NCAR



# GLADE

- **10.94 PB usable capacity → 16.42 PB usable (1Q2014)**

Estimated initial file system sizes

- **collections** ≈ 2 PB RDA, CMIP5 data
- **projects** ≈ 3 PB long-term, allocated space
- **users** ≈ 1 PB medium-term work space
- **scratch** ≈ 5 PB shared, temporary space

- **Disk Storage Subsystem**

- 76 IBM DCS3700 controllers & expansion drawers
  - 90 2-TB NL-SAS drives/controller
  - add 30 3-TB NL-SAS drives/controller (1Q2014)

- **GPFS NSD Servers**

- **91.8 GB/s** aggregate I/O bandwidth; 19 IBM x3650 M4 nodes

- **I/O Aggregator Servers (GPFS, GLADE-HPSS connectivity)**

- 10-GbE & FDR interfaces; 4 IBM x3650 M4 nodes

- **High-performance I/O interconnect to HPC & DAV**

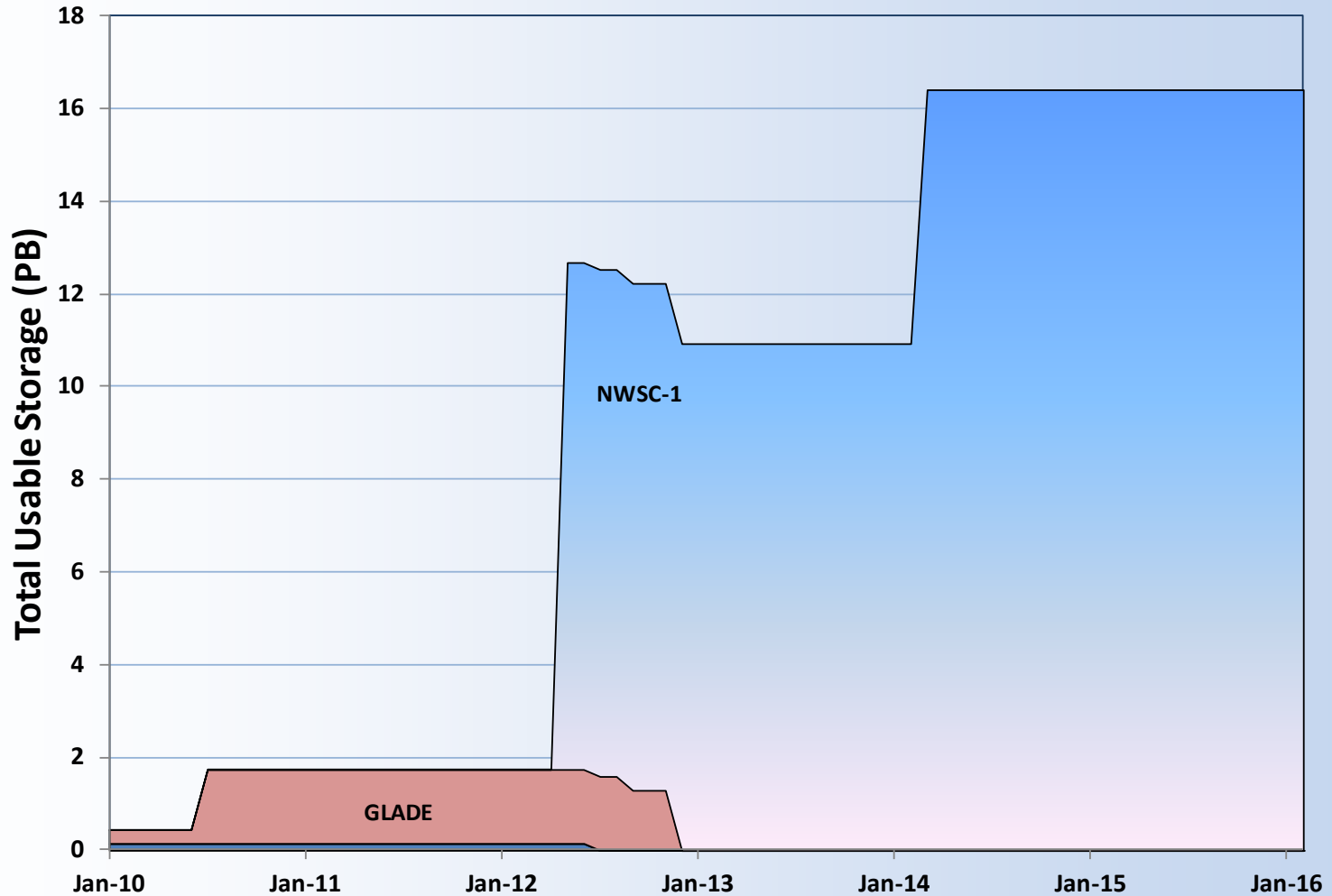
- Mellanox FDR InfiniBand full fat-tree
- 13.6 GB/s bidirectional bandwidth/node



# NCAR Disk Capacity Profile

## Total Centralized Filesystem Storage (PB)

NWSC-1 GLADE bluefire



# Geyser and Caldera

## NWSC Data Analysis & Visualization Resource

- **Geyser: Large-memory system**
  - 16 IBM x3850 nodes – Intel Westmere-EX processors
  - 40 cores, **1 TB memory**, 1 NVIDIA Kepler Q13H-3 GPU *per node*
  - Mellanox FDR full fat-tree interconnect
- **Caldera: GPU computation/visualization system**
  - 16 IBM x360 M4 nodes – Intel Sandy Bridge EP/AVX
  - 16 cores, 64 GB memory per node
  - 2 NVIDIA Kepler Q13H-3 GPUs per node
  - Mellanox FDR full fat-tree interconnect
- **Knights Corner system (November 2012 delivery)**
  - Intel Many Integrated Core (MIC) architecture
  - 16 IBM Knights Corner nodes
  - 16 Sandy Bridge EP/AVX cores, 64 GB memory
  - 1 Knights Corner adapter per node
  - Mellanox FDR full fat-tree interconnect



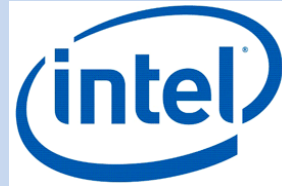
# Yellowstone Software

- **Compilers, Libraries, Debugger & Performance Tools**

- **Intel** Cluster Studio (Fortran, C++, performance & MPI libraries, trace collector & analyzer) 50 concurrent users
- **Intel** VTune Amplifier XE performance optimizer 2 concurrent users
- **PGI** CDK (Fortran, C, C++, pgdbg debugger, pgprof) 50 conc. users
- **PGI** CDK GPU Version (Fortran, C, C++, pgdbg debugger, pgprof) for DAV systems only, 2 concurrent users
- **PathScale** EccoPath (Fortran C, C++, PathDB debugger) 20 concurrent users
- Rogue Wave **TotalView** debugger 8,192 floating tokens
- **IBM** Parallel Environment (POE), including IBM HPC Toolkit

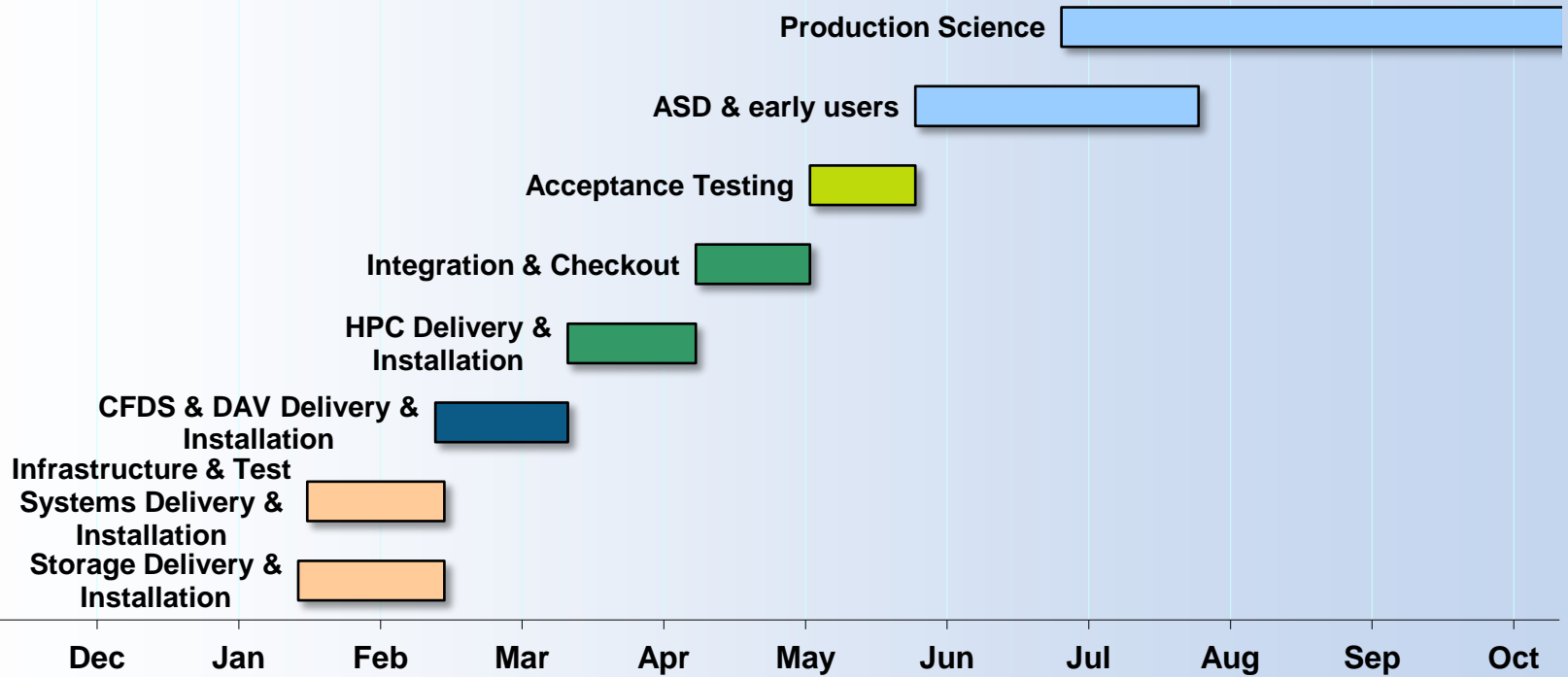
- **System Software**

- **LSF-HPC** Batch Subsystem / Resource Manager
  - IBM has purchased Platform Computing, Inc., developers of LSF-HPC
- Red Hat Enterprise **Linux** (RHEL) Version 6
- IBM General Parallel Filesystem (**GPFS**)
- Mellanox Universal Fabric Manager
- IBM xCAT cluster administration toolkit



# Yellowstone Schedule

## Delivery, Installation, Acceptance & Production



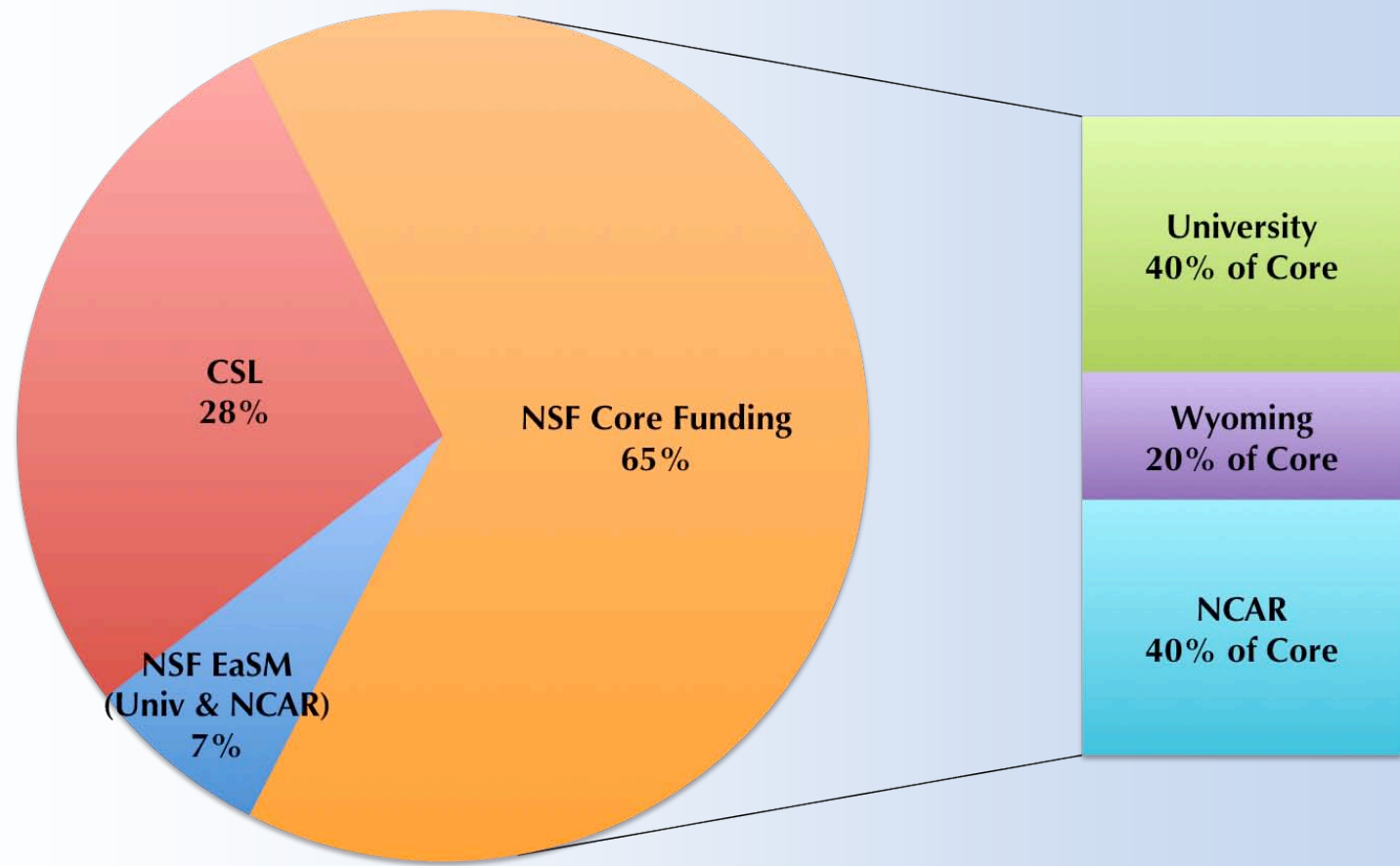
# Janus: Available now

- **Janus Dell Linux cluster**
  - 16,416 cores total – 184 TFLOPs peak
  - 1,368 nodes – 12 cores, 24 GB memory per node
  - Intel Westmere processors – 2.8 GHz clock
  - 32.8 TB total memory
  - QDR InfiniBand interconnect
  - Red Hat Linux, Intel compilers (PGI coming)
- **Deployed by CU in collaboration with NCAR**
  - ~10% of the system allocated by NCAR
- ***Available for Small allocations to university, NCAR users***
  - CESM, WRF already ported and running
  - Key elements of NCAR software stack already installed
- **[www2.cisl.ucar.edu/docs/janus-cluster](http://www2.cisl.ucar.edu/docs/janus-cluster)**



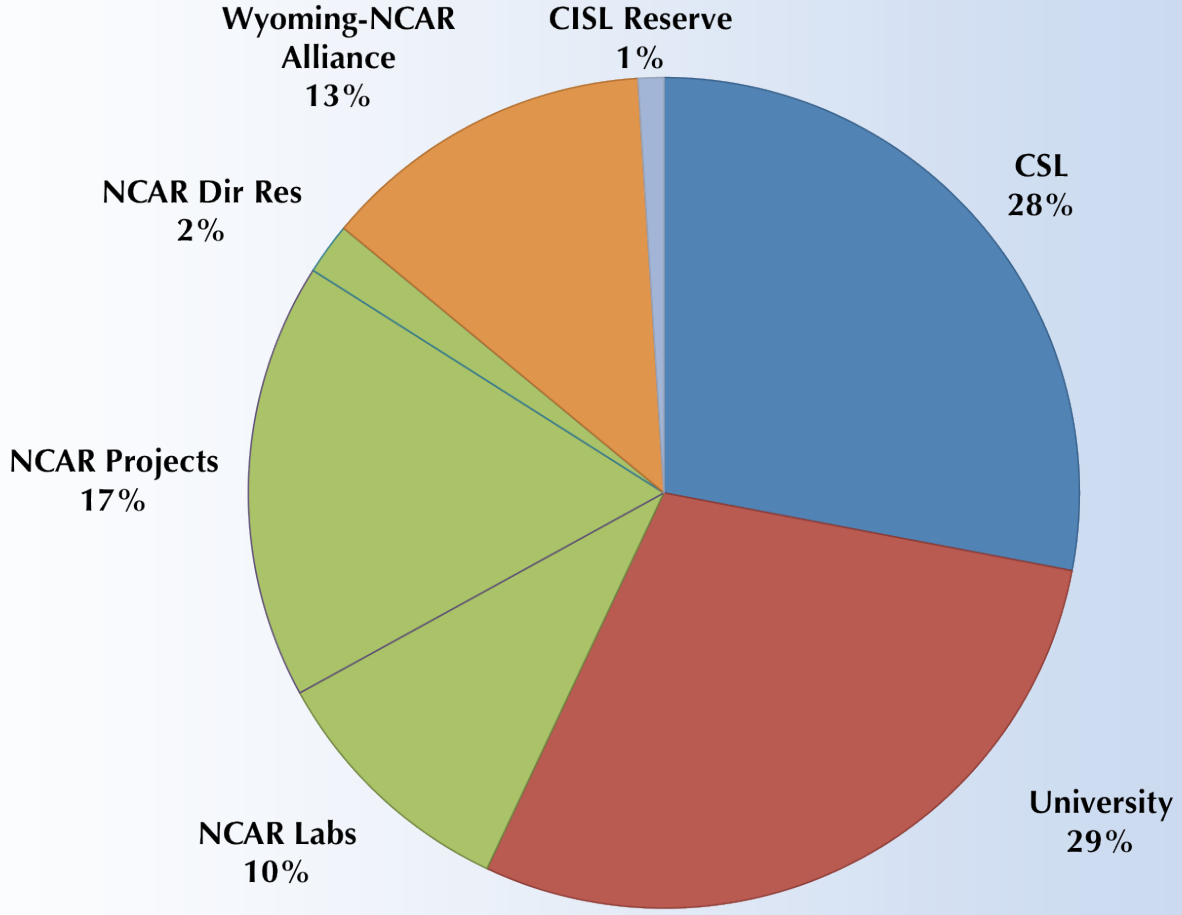


# Yellowstone allocation opportunities



### Yellowstone funding sources

**Yellowstone will be capable of *653 million core-hours per year*, compared to 34 million for Bluefire, and each Yellowstone core-hour is equivalent to 1.53 Bluefire core-hours.**



### Yellowstone allocation opportunities

The segments for CSL, University and NCAR users each represent about *170 million core-hours per year* on Yellowstone (compared to less than 10 million per year on Bluefire) plus a similar portion of DAV and GLADE resources.

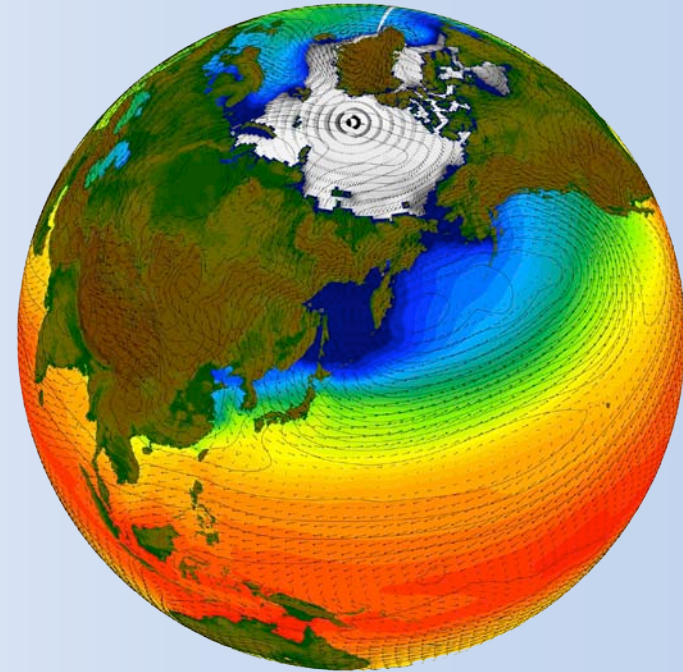
Early-use opportunity:

# Accelerated Scientific Discovery

- **Deadline: January 13, 2012**
- **Targeting a small number of rapid-turnaround, large-scale projects**
  - *Minimum* HPC request of 5 million core hours
  - Roughly May-July, with access to DAV systems beyond that point through final report deadline, February 2013
- **Approximately 140 million core-hours, in two parts**
  - University-led projects with NSF awards in the geosciences will be allocated **70 million core-hours**
  - NCAR-led projects will make up the other half, selected from NCAR Strategic Capability requests that designate themselves “ASD-ready.”
- **Particularly looking for projects that contribute to NWSC Community Science Objectives**
  - High bar for production readiness, including availability of staff time
- **[www2.cisl.ucar.edu/docs/allocations/asd](http://www2.cisl.ucar.edu/docs/allocations/asd)**

# Climate Simulation Laboratory

- ***Deadline: February 15, 2012***
- **Targets large-scale, long-running simulations of the Earth's climate**
  - Dedicated facility supported by the U.S. Global Change Research Program
  - Must be climate-related work, but support may be from any agency
- ***Minimum request and award size***
  - Typically 18-month allocation period
  - Approx. 250 million core-hours to be allocated
  - Estimated minimum request size: ~10 million core-hours
- **Preference given to large, collective group efforts, preferably interdisciplinary teams**



# Overall process familiar

- **Submission format similar**
  - Modified to encompass storage resource needs
  - Should get CESM pre-approval to deviate from standard 5-page format
- **Review of requests by CSL Allocation Panel**
  - CSLAP meets late March
  - Awards announced by early May
- **Early access for new CSL awardees in June**
  - Limited scale, lower priority than ASD
  - To prepare for running at scale in August, after ASD period
- **Current CSL allocations will run to completion on Bluefire**

# Allocation changes

- **Not just HPC, but DAV, HPSS, GLADE allocations**
  - Non-HPC resources  $\approx$  1/3 procurement cost
  - Ensure that use of scarce and costly resources are directed to the most meritorious projects
  - Ensure that HPC output doesn't swamp the storage resources
- **Want to identify projects contributing to the NWSC Community Scientific Objectives**
  - [www2.cisl.ucar.edu/resources/yellowstone/science](http://www2.cisl.ucar.edu/resources/yellowstone/science)
- **Balance between the time to prepare and review requests and the resources provided**
  - Minimize user hurdles and reviewer burden
  - Build on familiar process for requesting HPC allocations
- **All new, redesigned accounting system (SAM)**
  - Separate, easier to understand allocations
  - Switchable 30/90 option, per project, as an operational control
    - ("30/90" familiar to NCAR labs and CSL awardees)

# Justifying resource needs

- **HPC — similar to current practice**
  - Cost of runs necessary to carry out experiment, supported by benchmark runs or published data
- **DAV — will be allocated, similar to HPC practice**
  - A “small” allocation will be granted upon request
  - Allocation review to focus on larger needs associated with batch use
  - Memory and GPU charging to be considered
- **HPSS — focus on storage needs above a threshold**
  - 20-TB default threshold initially
    - Perhaps lower default for “small” allocations”
  - CISL to evaluate threshold regularly to balance requester/reviewer burden with demand on resources
  - Simplified request/charging formula
- **GLADE — project (long-term) spaces will be reviewed and allocated**
  - scratch, user spaces not allocated

# GLADE resource requests

- **Only for project space**
  - No need to detail use of scratch, user spaces
- **Describe why project space is essential**
  - That is, why scratch or user space insufficient
    - Show that you are aware of the differences
  - Shared data, frequently used, not available on disk from RDA, ESG (collections space)
- **Relate the storage use to your workflow and computational plan**
  - Projects with data-intensive workflows should show they are using resources efficiently

# HPSS resource requests

- **CESM-P has 2.25 PB stored in HPSS**
  - Plus 450 TB for IPCC, 72 TB by Development
- **Goal: Demonstrate that HPSS use is efficient and appropriate**
  - Not fire and forget into a “data coffin”
  - Not using as a temporary file system
- **Explain new data to be generated**
  - Relate to computational experiments proposed
  - Describe scientific value/need for data stored
- **Justify existing stored data**
  - Reasons for keeping, timeline for deletion
- **Data management plan: Supplementary information**
  - Additional details on the plans and intents for sharing, managing, analyzing the data

# Calculating storage use

- In terms of real cost, the storage analog of computing core-hours is storage “TB-hours”

*Usage = space \* time \* QOS*  
becomes

*Usage = TB stored \* hours \* COS*

- **Thus, holding 2 PB for 18 months:**
  - 2,000 TB \* 12,960 = 25,920,000 TB-hours
- **For comparison, a default small HPSS allocation of 20 TB:**
  - 20 TB \* 8,760 = 175,200 TB-hours
- ***Could also be calculated in TB-years. Preference?***
- **For now on HPSS, COS (Class of Service) = # of copies**
  - For GLADE, COS is not applicable

# To GAU or not to GAU?

**A question for you, as users:**

- **Do you think about computing use in terms of “GAUs” or “core-hours” or some other units?**
  - More to the point, would you hate to see GAUs fade away, in favor of core-hours?
  - The machine factor for Yellowstone will be 2.15, meaning 1 core-hour = 2.15 GAUs

<http://www2.cisl.ucar.edu/resources/yellowstone>

<http://www2.cisl.ucar.edu/docs/allocations>



**QUESTIONS?**