Computational & Information Systems Laboratory

CISL

NCAR's
Data-Centric Supercomputing Environment
Yellowstone

February 29, 2012
David L. Hart, CISL
dhart@ucar.edu

NCAR NSF

Welcome to the Petascale

- **Yellowstone hardware and software**
- **Deployment schedule**
- **Yellowstone allocations opportunities**

Yellowstone

NWSC High-Performance Computing Resource

- **Batch Computation**
  - ***72,288 cores total – 1.5 PFLOPs peak***
  - 4,518 IBM dx360 M4 nodes – 16 cores, 32 GB memory per node
  - Intel Sandy Bridge EP processors with AVX – 2.6 GHz clock
  - 144.6 TB total, DDR3-1600 memory
  - 28.9 Bluefire equivalents
- **High-Performance Interconnect**
  - Mellanox FDR InfiniBand full fat-tree
  - 13.6 GB/s bidirectional bw/node
  - <2.5 µs 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

CISL Computational & Information Systems Laboratory

NCAR NSF

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

CISL Computational & Information Systems Laboratory

NCAR NSF

Research Data Archive

- **Total archive volume over 1.3 PB**
  - Meteorological and oceanographic data
  - E.g., reanalyses, global observations, etc.
- **Access to data from HPC and DAV at NWSC**
  - All data from the HPSS archive
  - Most popular 500 TB of data on GLADE
  - Value-added services provided
    - Command-line metadata access
    - Command-line subset requests

http://dss.ucar.edu/

CISL Computational & Information Systems Laboratory
NCAR NSF

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 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 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

CISL Computational & Information Systems Laboratory

NCAR

Yellowstone Software

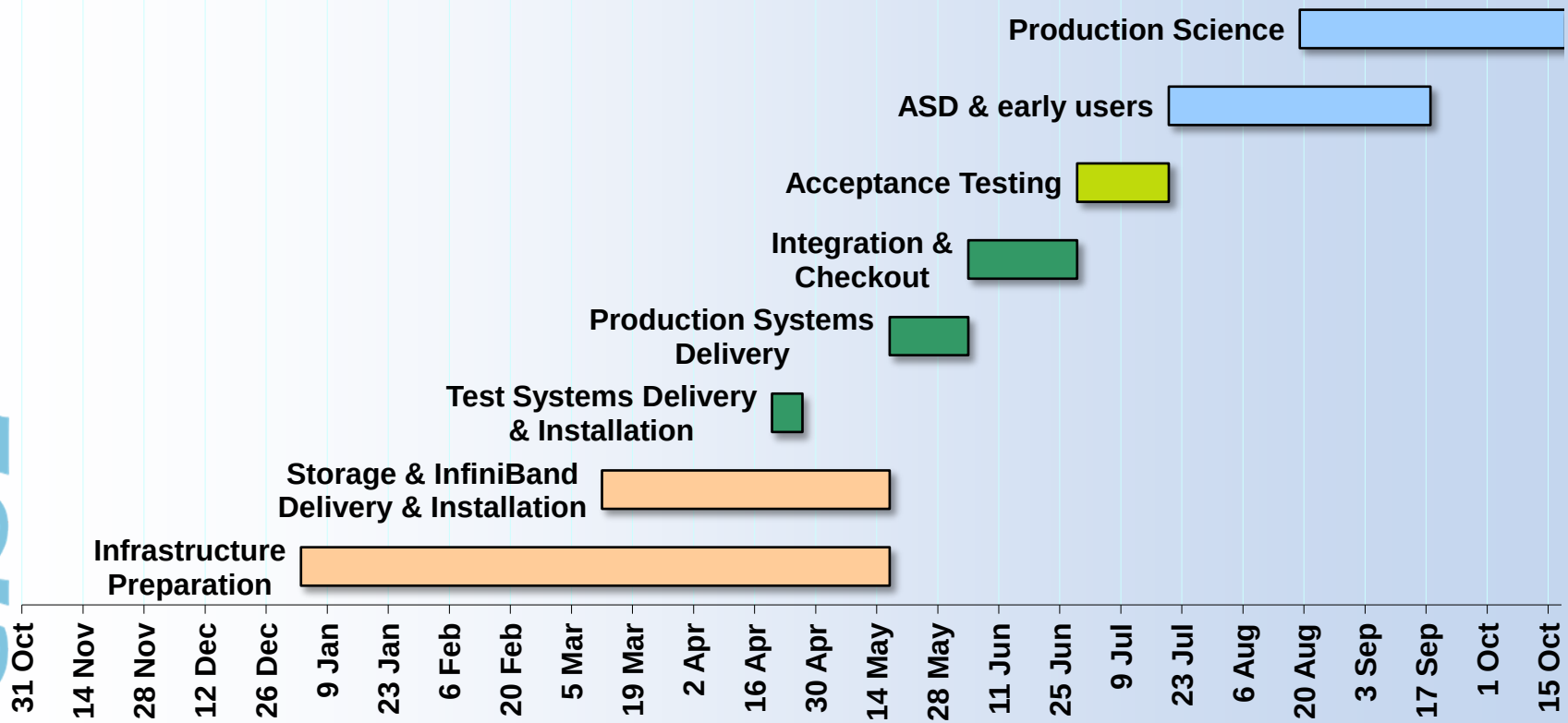
- **Compilers, Libraries, Debugger & Performance Tools**
  - **Intel** Cluster Studio (Fortran, C++, performance & MPI libraries, trace collector & analyzer) 50 concurrent users
  - **Intel** VTune Amplifier XE performance optimizer 2 concurrent users
  - **PGI** CDK (Fortran, C, C++, pgdbg debugger,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

CISL Computational & Information Systems Laboratory

NCAR NSF

Yellowstone Schedule

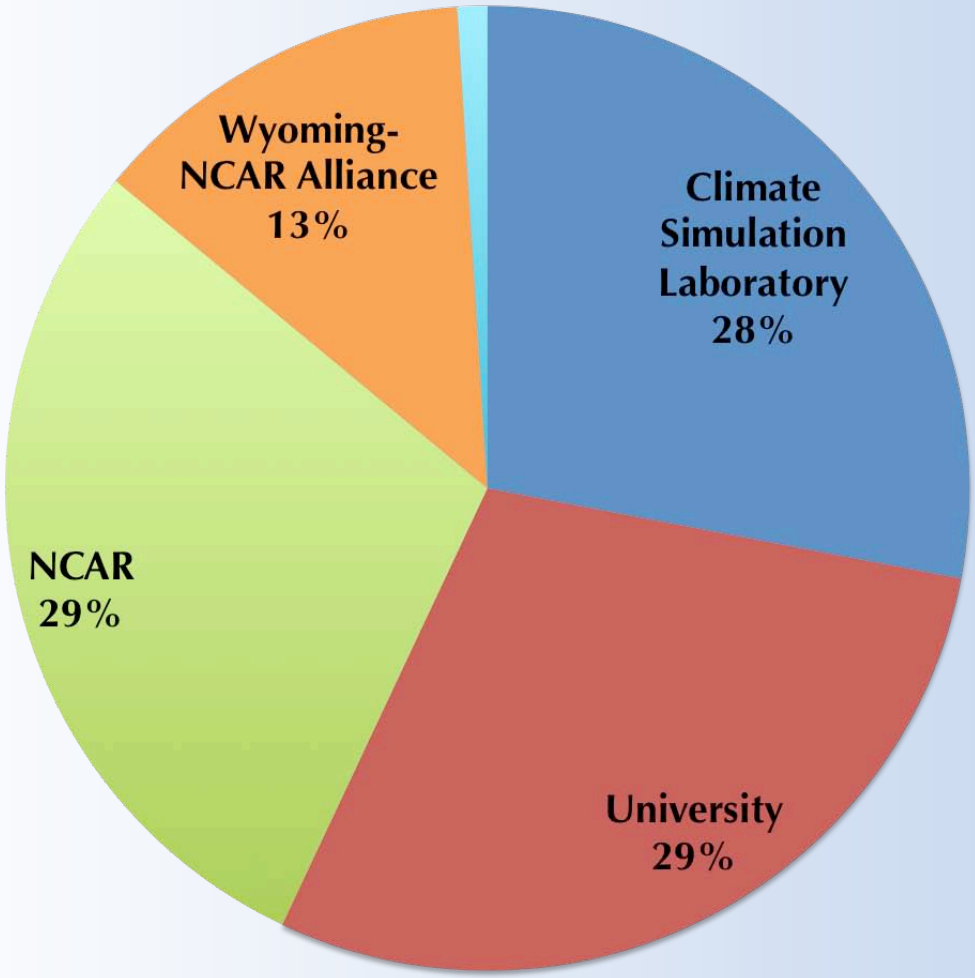
Delivery, Installation, Acceptance and Production



Computational & Information Systems Laboratory
CISL
NCAR
NSF

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 and PGI compilers
- **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**



Yellowstone allocations 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.

Computational & Information Systems Laboratory

CISL

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

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

cislhelp@ucar.edu *or* dhart@ucar.edu

QUESTIONS?

NCAR

BACKUP SLIDES

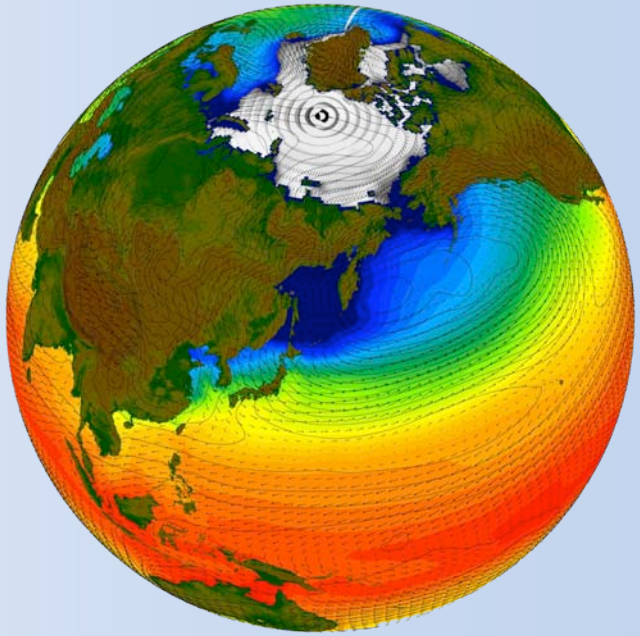
Early-use opportunity:

Accelerated Scientific Discovery

- ***Deadline: March 5, 2012***
- **Targeting a small number of rapid-turnaround, large-scale projects**
  - *Minimum* HPC request of 5 million core hours
  - Roughly August-October, with access to DAV systems beyond that point through final report deadline, April 2013
- **Approximately 120 million core-hours, in two parts**
  - University-led projects with NSF awards in the geosciences will be allocated ***60 million core-hours***
  - NCAR-led projects will comprise the other half
- **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***

Climate Simulation Laboratory

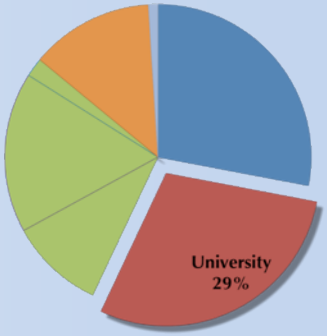
- ***Deadline: February 20, 2012***
- **Targets large-scale, long-running simulations of the Earth's climate**
  - Dedicated facility supported by the U.S. Global Change Research Program
  - Must be climate-related work, but support may be from any agency
- ***Minimum* request and award size**
  - 18-month allocation period
  - Approx. 250 million core-hours to be allocated
  - Minimum request size: 10 million core-hours
- **Preference given to large, collective group efforts, preferably interdisciplinary teams**



CISL Computational & Information Systems Laboratory

NCAR NSF

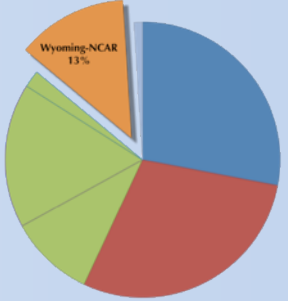
University allocations



- ***Next deadline: March 26, 2012***
- **Large allocations will continue to be reviewed and awarded twice per year**
  - Deadlines in March and September
  - Approx. 85 million core-hours to be allocated at each opportunity
- **Small allocations will also be available once system enters full production**
  - "Small" allocation will be up to 200,000 core-hours
  - Small allocations for researchers with NSF award—appropriate for benchmarking, preparation for large request
  - Small, one-time allocations for grad students, post-docs, new faculty without NSF award
  - Classroom allocations for instructional use
- ***www2.cisl.ucar.edu/docs/allocations/university***

NEW!

Wyoming-NCAR Alliance



- ***Deadline: March 26, 2012***
- **13% of Yellowstone resources**
  - 75 million core-hours per year
  - U Wyoming managed process
- **Activities must have substantial U Wyoming involvement**
  - Allocated projects must have Wyoming lead
  - Extended list of eligible fields of science
  - Eligible funding sources not limited to NSF
- **Actively seeking to increase collaborations with NCAR and with other EPSCoR states.**
- **Otherwise, process modeled on University allocations, with panel review of large requests.**
- ***www.uwyo.edu/nwsc***

CISL Computational & Information Systems Laboratory
NCAR NSF

ALLOCATION REQUESTS

Allocation changes in store

- **Not just HPC, but DAV, HPSS, GLADE allocations**
  - Non-HPC resources ≈ 1/3 procurement cost
  - Ensure that use of scarce and costly resources are directed to the most meritorious projects
- **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
- **Want to identify projects contributing to the NWSC Community Scientific Objectives**
  - *www2.cisl.ucar.edu/resources/yellowstone/science*
- **All new, redesigned accounting system (SAM)**
  - Separate, easier to understand allocations

General submission format

- ***Please see specific opportunities for detailed guidelines!***
- **Five-page request**
  A. Project information (title, lead, etc.)
  B. Project overview and strategic linkages
  C. Science objectives
  D. Computational experiments and resource requirements (HPC, DAV, and storage)
- **Supporting information**
  E. Multi-year plan (if applicable)
  F. Data management plan
  G. Accomplishment report
  H. References and additional figures

CISL Computational & Information Systems Laboratory
NCAR NSF

Tips and advice

- **Remember your audience: computational geoscientists from national labs, universities and NCAR**
  - Don't assume they are experts in *your* specialty
- **Be sure to articulate relevance and linkages**
  - Between funding award, computing project, eligibility criteria, and NWSC science objectives (as appropriate)
- ***Don't submit a science proposal***
  - Describe the science in detail sufficient to justify the computational experiments proposed
- **Most of the request should focus on computational experiments and resource needs**
  - Effective methodology
  - Appropriateness of experiments
  - Efficiency of resource use

CISL Computational & Information Systems Laboratory
NCAR NSF

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: # users x 5,000 core-hours
  - 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 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

- **Goal: Demonstrate that HPSS use is efficient and appropriate**
  - Not store-everything-forever in 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, holding the data