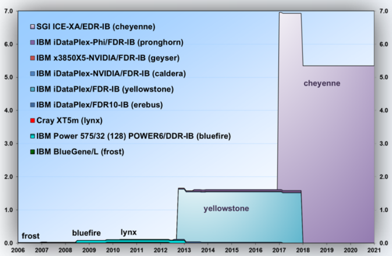
NCAR's Data-Centric Supercomputing Environment — 2017 edition —

David Hart

NCAR/CISL User Services Manager

dhart@ucar.edu



CISL Computational & Information Systems Laboratory

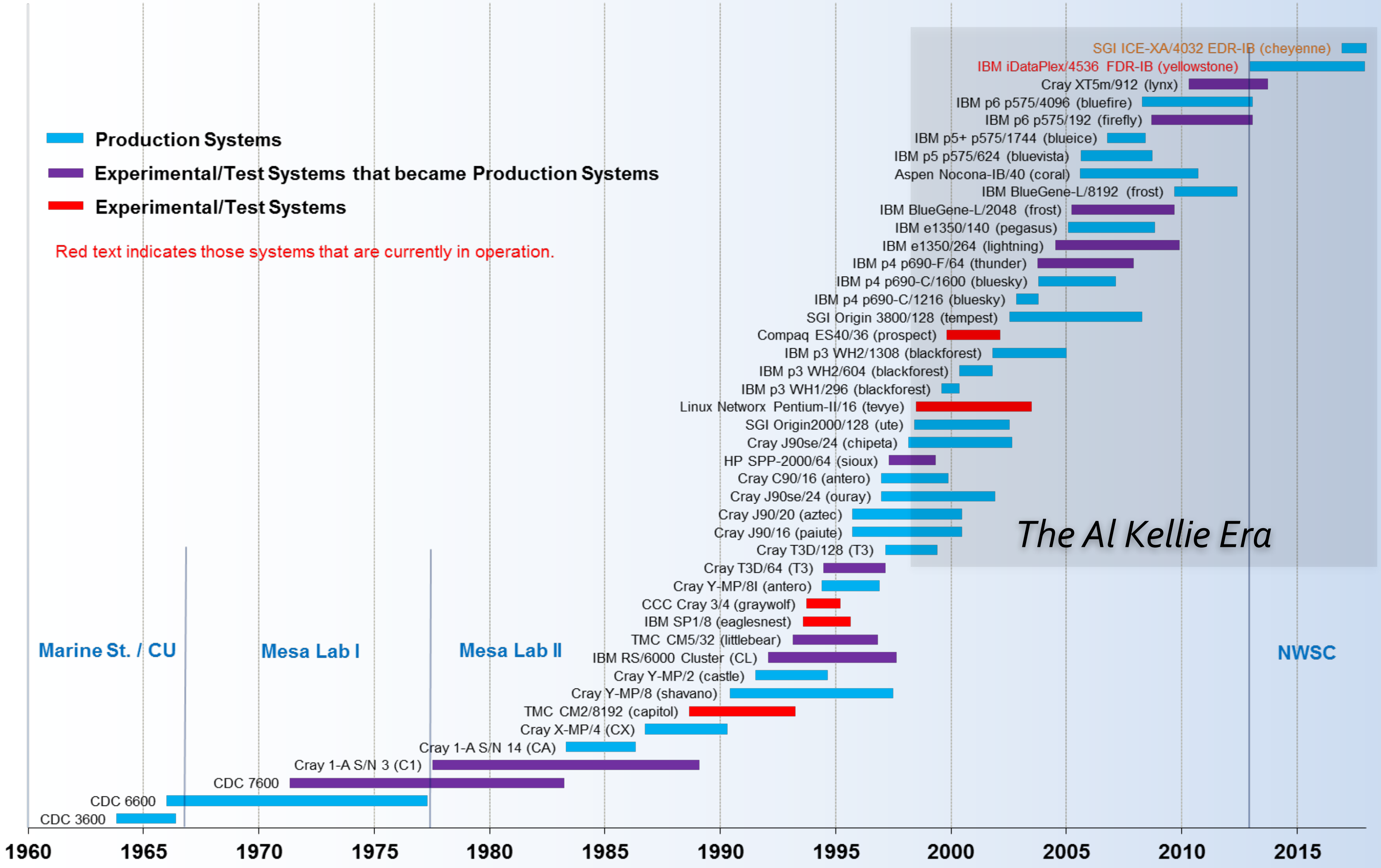
In memoriam — Al Kellie

24 Mar. 1945 – 7 Sep. 2016

- Joined NCAR as 9th Director of Scientific Computing Division (SCD) in 1998.
- Evolved SCD into CISL in 2005.
- Spearheaded NCAR's efforts to construct new data center, from concept to opening, 2003–2012.
- Oversaw procurement of first two petascale systems for NWSC.

History of computing at NCAR

Computational & Information Systems Laboratory



CISL

NCAR

NSF

User communities for CISL HPC

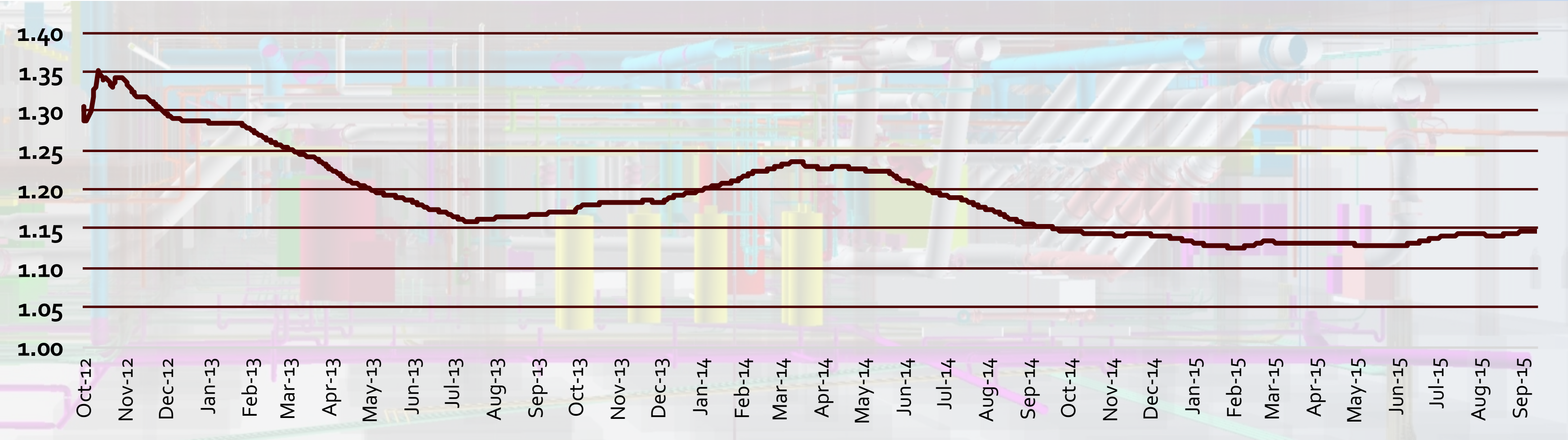
NCAR supports four user communities through policies established via various agreements with NSF or approved by NSF.

- University research
  - For U.S.-based researchers with NSF awards in the atmospheric, geospace or closely related sciences
  - Roughly 100 “large” projects reviewed, approved each year
  - About 250 “small” allocations approved each year
- Climate Simulation Laboratory
  - Supports large-scale, long-running climate simulations
  - Eligibility otherwise similar to University allocations
  - Also supports large annual allocation to CESM Community
- NCAR Lab and NCAR Strategic Capability activities
  - Lab allocations support smaller-scale initiatives, projects
  - Large NSC requests reviewed by internal NCAR panel
- Wyoming-NCAR Alliance
  - Must be led by U Wyoming researcher
  - Must be in the Geosciences or related fields
  - About 20 large requests reviewed, approved each year

NWSC-2
Second petascale procurement at NWSC

NCAR-Wyoming Supercomputing Center

- NWSC regularly achieving 1.11–1.16 PUE at half the assumed 4-MW load
- Expected to hit PUE target of 1.08 with deployment of Cheyenne
- Enhancements for Cheyenne nearing completion
- Ongoing improvements to achieve best operational efficiency



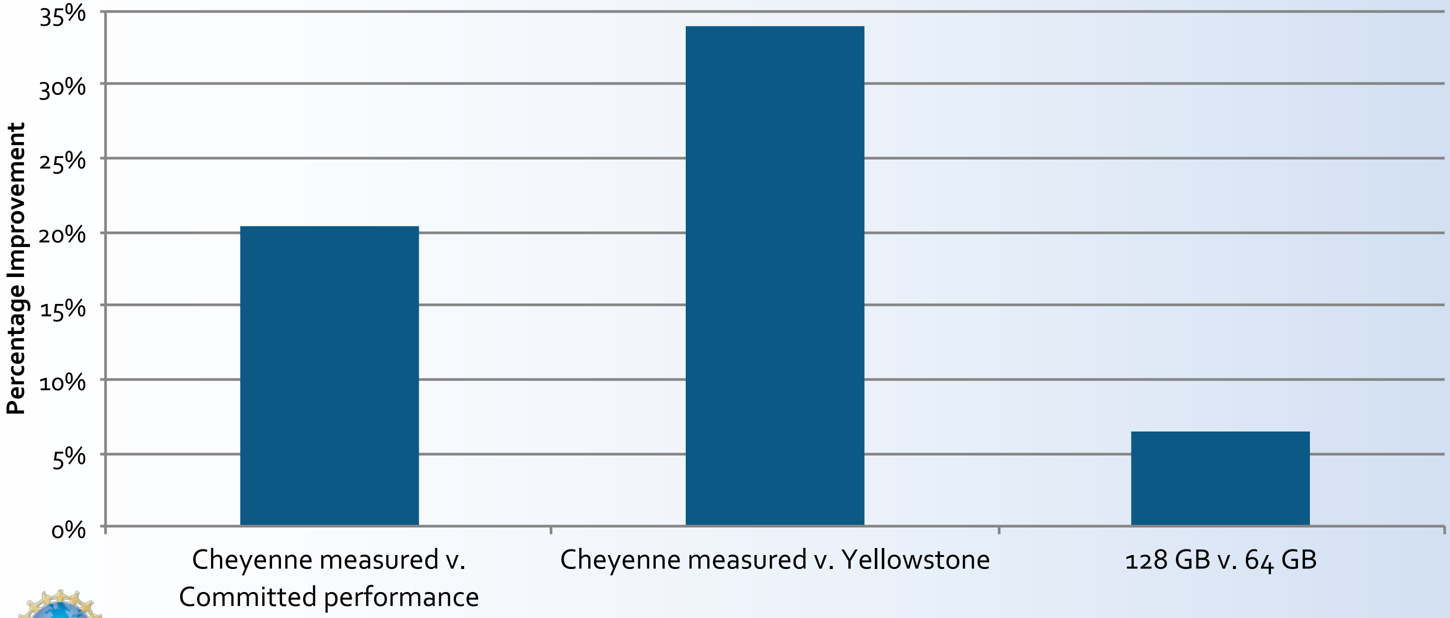
Cheyenne

Planned production: January 2017 – 2021

- Scientific computation nodes
  - SGI ICE XA cluster
  - 4,032 dual-socket nodes
  - 18-core, 2.3-GHz Intel Xeon E5-2697v4 processors
  - 145,152 “Broadwell” cores total
  - 5.34 PFLOPs peak – 1.325 TFLOPs per node
  - 313 TB total memory (64-GB & 128-GB nodes)
  - **>3 Yellowstone equivalents on NCAR Benchmark Suite**
- High-performance interconnect
  - Mellanox EDR InfiniBand
  - 9-D enhanced hypercube topology
  - 224 36-port switches, no director switches
- Login nodes (6) & service nodes (4)
- SuSE Linux OS, Altair PBS Pro scheduler

Cheyenne performance on NCAR Benchmark Suite

- Cheyenne system performance is approximately 3 YSEP
  - Yellowstone Sustained Equivalent Performance (YSEP) measures total system throughput against Yellowstone performance.
- Summarizes the 34 benchmarks
  - 7 apps/app kernels + 6 I/O benchmarks and micro-benchmarks
- Comparison is on a core-to-core basis



Cheyenne deployment schedule

1

Early 2016 — Facility and networking upgrades at NWSC start

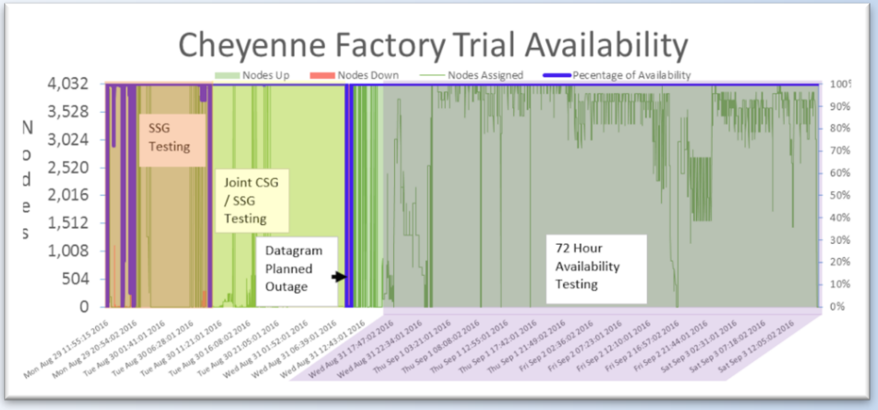
  – Includes 100-GbE WAN link

- July — Test system arrives at NWSC
- July–August — Cheyenne hardware assembled in Chippewa Falls
  – Includes initial HPL run

2

29 Aug. — Factory testing begins

- 31 Aug. – 3 Sep. — 72-hour factory availability test *(99.995% uptime)*
- 12 Sept. — Cheyenne hardware arrived at NWSC on six trucks
- 15 Sept. — Cheyenne racks powered up and all nodes booted
- 21 Sept. — HPL run on Cheyenne *(better results than factory HPL run)*

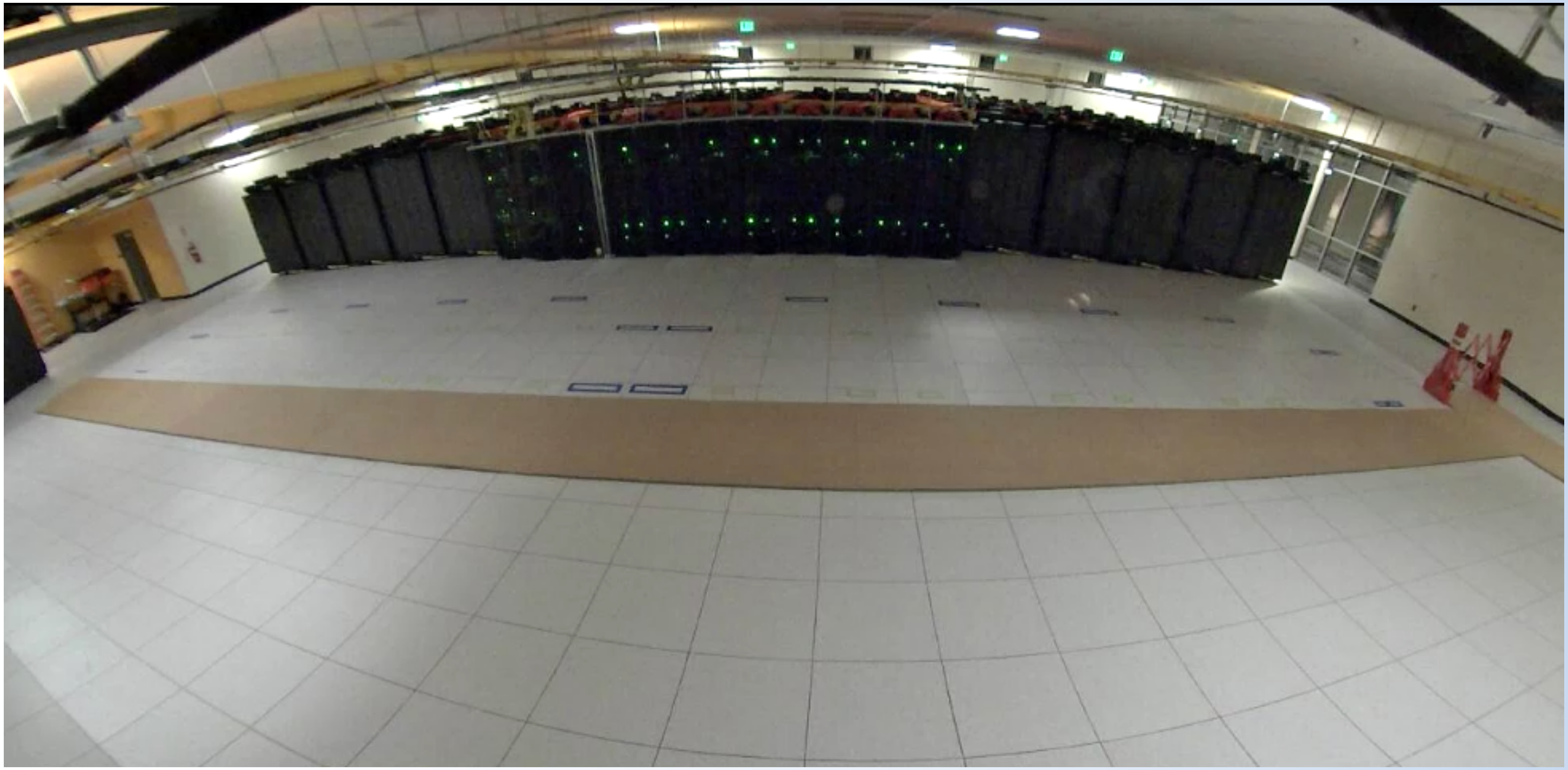


3

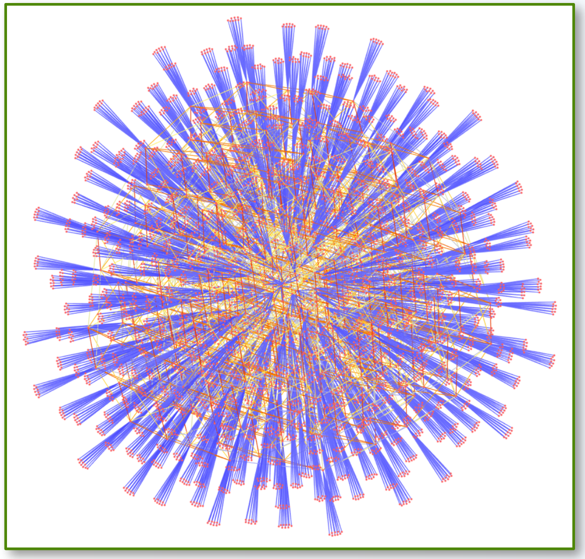
Cluster is being integrated with storage system and going through acceptance testing

- Targeting NCAR acceptance by end of December
- January 2017 — Start of production
  – Yellowstone continues through December 2017
  – City of Cheyenne celebrates its 150th anniversary in 2017

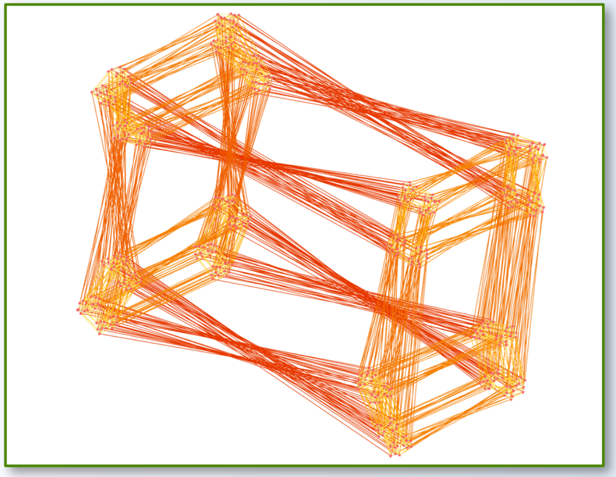
Four days in 30 seconds



InfiniBand rendering with Tulip



Cheyenne's hypercube representation in circular shape without a true center



Cheyenne hypercube projected from 9D to 3D. Imbalance shown by red vs. orange connections.

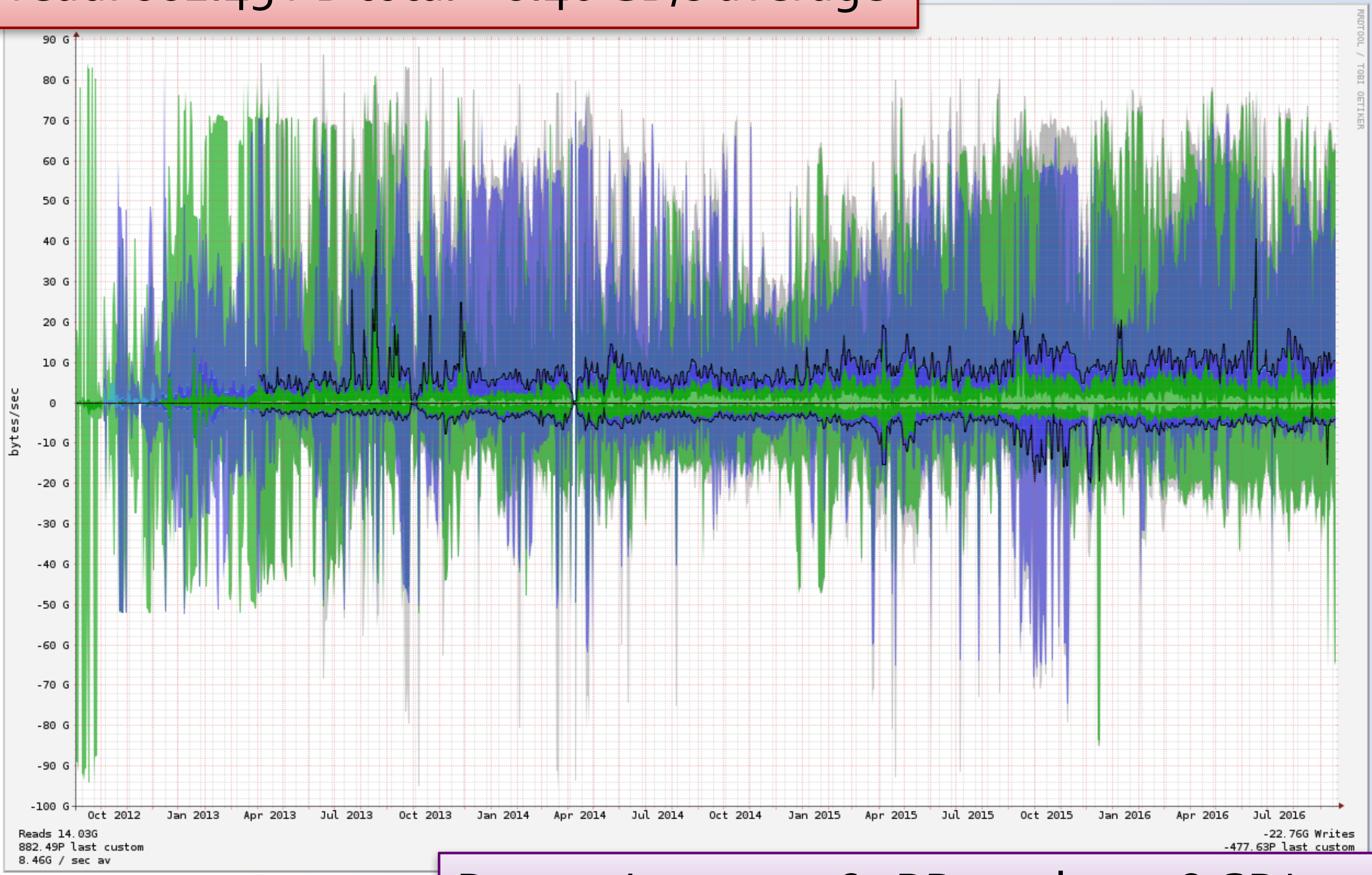
- CISL developed plug-in for Tulip data visualization software for viewing structure and traffic on InfiniBand fabrics.
- Nodes are IB ports
- Cables represented by two directional edges
- Edge colors used to produce heat maps

GLADE-2a file system resource

- 21-PB DDN Storage Fusion Architecture (SFA) system
  - 8x SFA14KXE units
  - 8x10 84-slot drive chassis
  - 3,360x 8-TB NL SAS drives (2x expandable)
  - 26.8 PB raw capacity
- 220 GB/s aggregate I/O bandwidth
  - 48x 800-GB, mixed-use SSDs for metadata
  - 32x embedded NSD servers
  - EDR InfiniBand and 40-Gig Ethernet connections
- Total integrated capacity: 37 PB
  - Integrates with existing 16-PB file system
  - Expandable to 42-PB (58-PB total) by adding drives
- IBM Spectrum Scale software
  - Formerly GPFS
  - RedHat Enterprise Linux OS

GLADE utilization since 2012

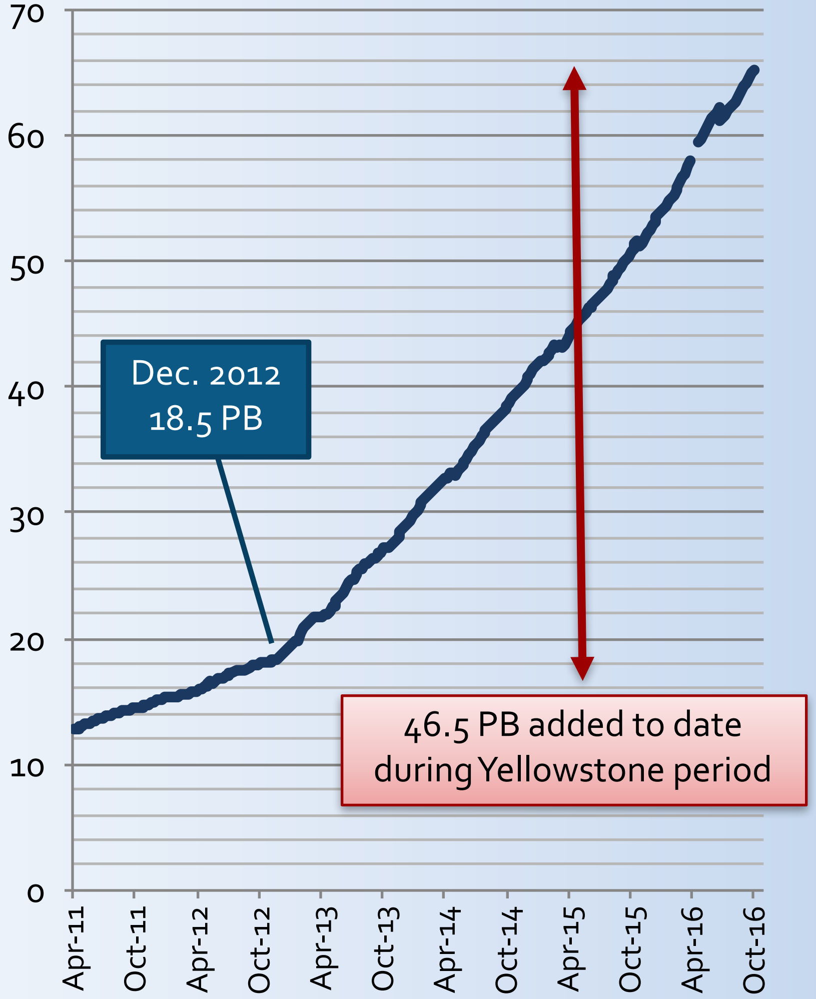
Data read: 882.45 PB total – 8.46 GB/s average



Data written: 477.63 PB total – 4.58 GB/s average

NCAR HPSS archive resource

- Current holdings: 65 PB
- Current growth: 14 PB/yr
- NWSC (Cheyenne, WY)
  - Four SL8500 robotic libraries
  - 46 T10000C tape drives
  - 46 T10000D tape drives
  - 320 PB capacity
- Mesa Lab (Boulder, CO)
  - Two SL8500 robotic libraries
  - 15-PB capacity for disaster recovery data
- Upgrade planned for late 2017



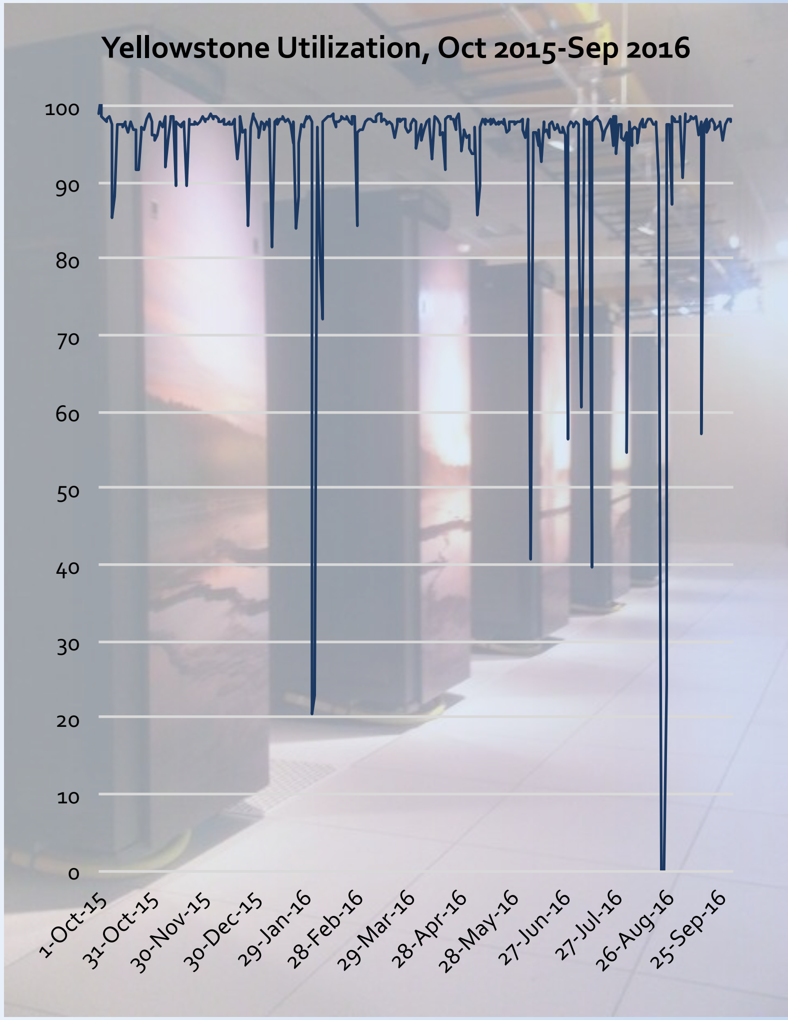
Dec. 2012
18.5 PB

46.5 PB added to date during Yellowstone period

Yellowstone activity

December 2012 through October 2016

- NCAR's first petascale system
  - 1.5 PFLOPS, 4,536 nodes
- More than 2,500 users
- 13.8M user jobs
- 2.2B core-hours delivered
- 98.02% FY16 average user availability
- 95.3% FY16 average utilization
- 14.9 PB and 1.1B files on GLADE



CISL Computational & Information Systems Laboratory

NCAR NSF

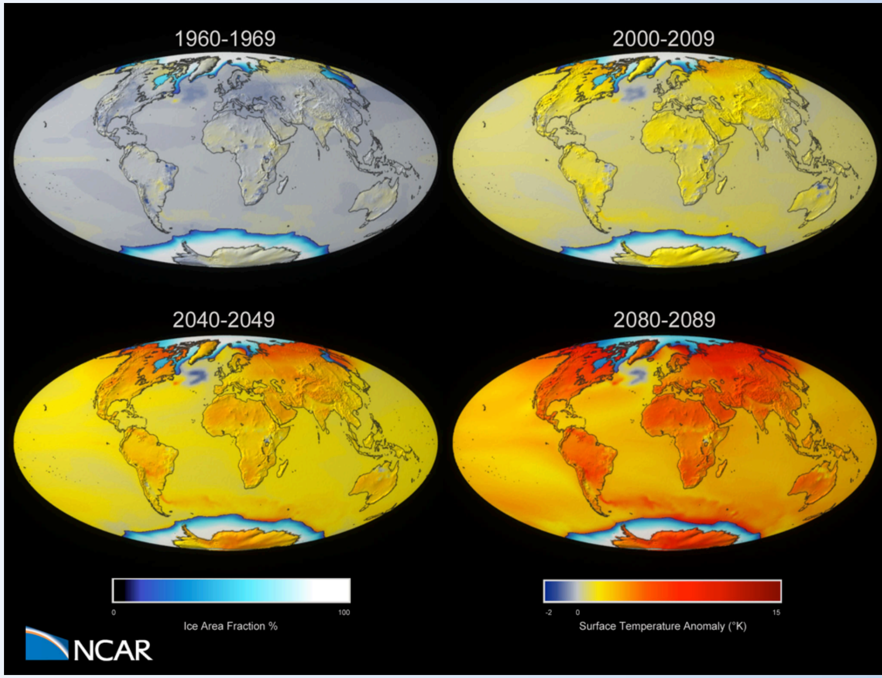
Geyser and Caldera

Data Analysis & Visualization Resources

- More than 1,800 users of these clusters since 2012
- Geyser: Large Memory
  - 7.8 million jobs since Dec. 2012
  - 16 nodes, 1 TB memory per node
  - 40 Intel Westmere cores per node
- Caldera: GPU/Visualization
  - 2.9 million jobs since Dec. 2012
  - 16 nodes, 2 Tesla K20X GPUs per node
  - 16 Intel Sandy Bridge cores per node

CMIP Analysis Platform

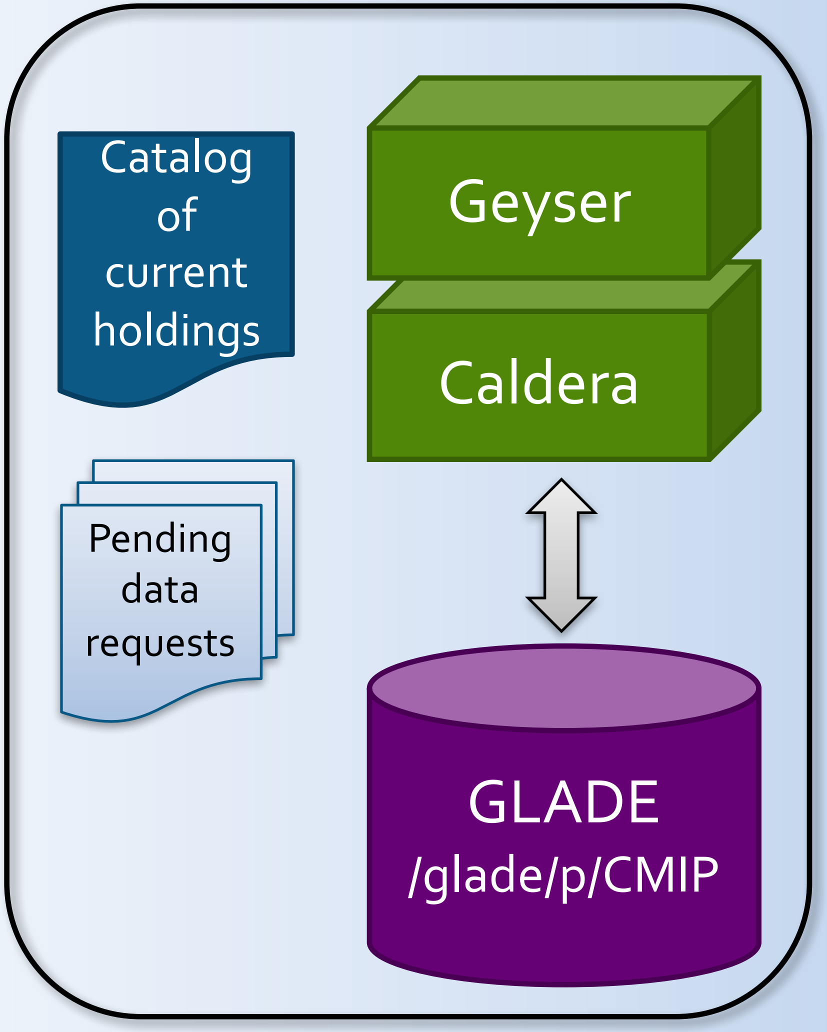
- New NCAR service to address the Big Data problems associated with CMIP analyses
  - Intercomparison requires having globally housed data in a single location
- A prototype available now with CMIP5 data and preparing to scale up for CMIP6
- An overlay of existing CISL resources
  - GLADE disk storage
  - 200 TB of NCAR CMIP5 data
  - Geyser/Caldera analysis clusters
  - User support services



Comparison of four decadal averages of temperature anomalies and ice area fraction. Data from the ensemble average of the CCSM4 monthly surface temperature anomaly (relative to 1850-1899 average for each month) from Jan 1850 to Dec 2100, from CMIP5 historical + RCP8.5 scenario runs. Data provided by Gary Strand. Visualization by Tim Scheitlin and Mary Haley.

CMIP Analysis Platform in operation

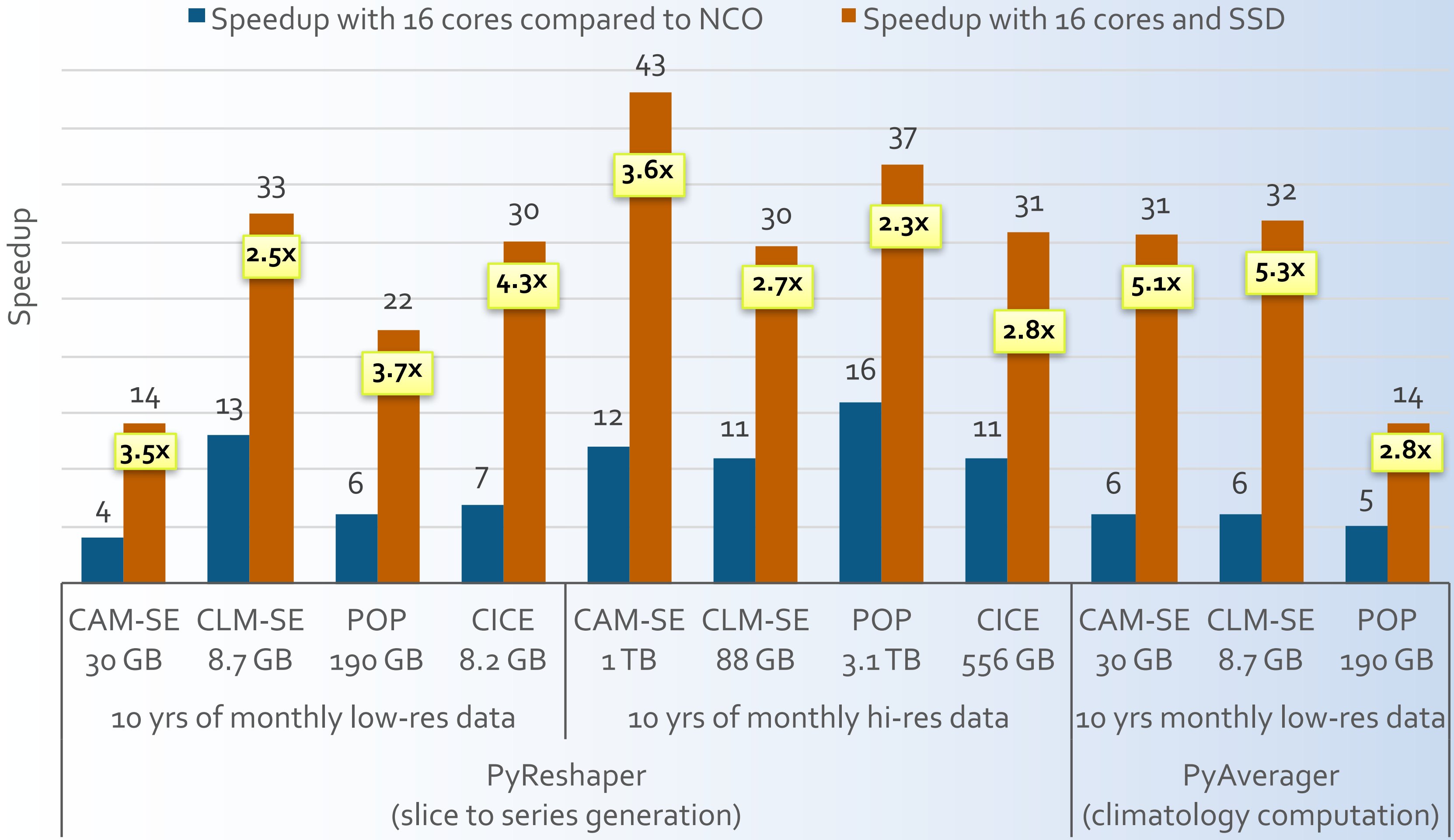
- GLADE disk space at NCAR set aside for the "interlibrary loan" of non-NCAR CMIP5 data sets.
  - In addition to NCAR's CMIP5 published data already on GLADE.
- CMIP Analysis Platform allocation required to request a data set be added.
- Users can request data sets to be added to the CMIP space.
  - CISL staff seek out, acquire, and ingest the data from the host site(s).
- Geyser and Caldera clusters provide analysis capability
- Data also accessible to any project with a Geyser/Caldera or Yellowstone allocation.



Path forward initiatives

- HPC Futures Lab
  - CISL infrastructure test bed for evaluating new hardware and software
- SPOC (Strategic Parallelization and Optimization of Computing)
  - Strategic CISL effort to work with NCAR model developers to improve performance of models, prepare them for future architectures
- IPCC-WACS
  - Intel Parallel Computing Center working on Xeon Phi performance and development for weather and climate systems
  - NCAR, Intel, U Colorado-Boulder, Indian Institute of Science
- SGI-NCAR Joint Center of Excellence
  - Partnership to optimize system and application performance for NCAR models on Cheyenne and future architectures
- NVIDIA GPU Research Center
  - Partnership to apply GPU technology to atmospheric model needs
  - U Wyoming, NCAR, NVIDIA, KISTI

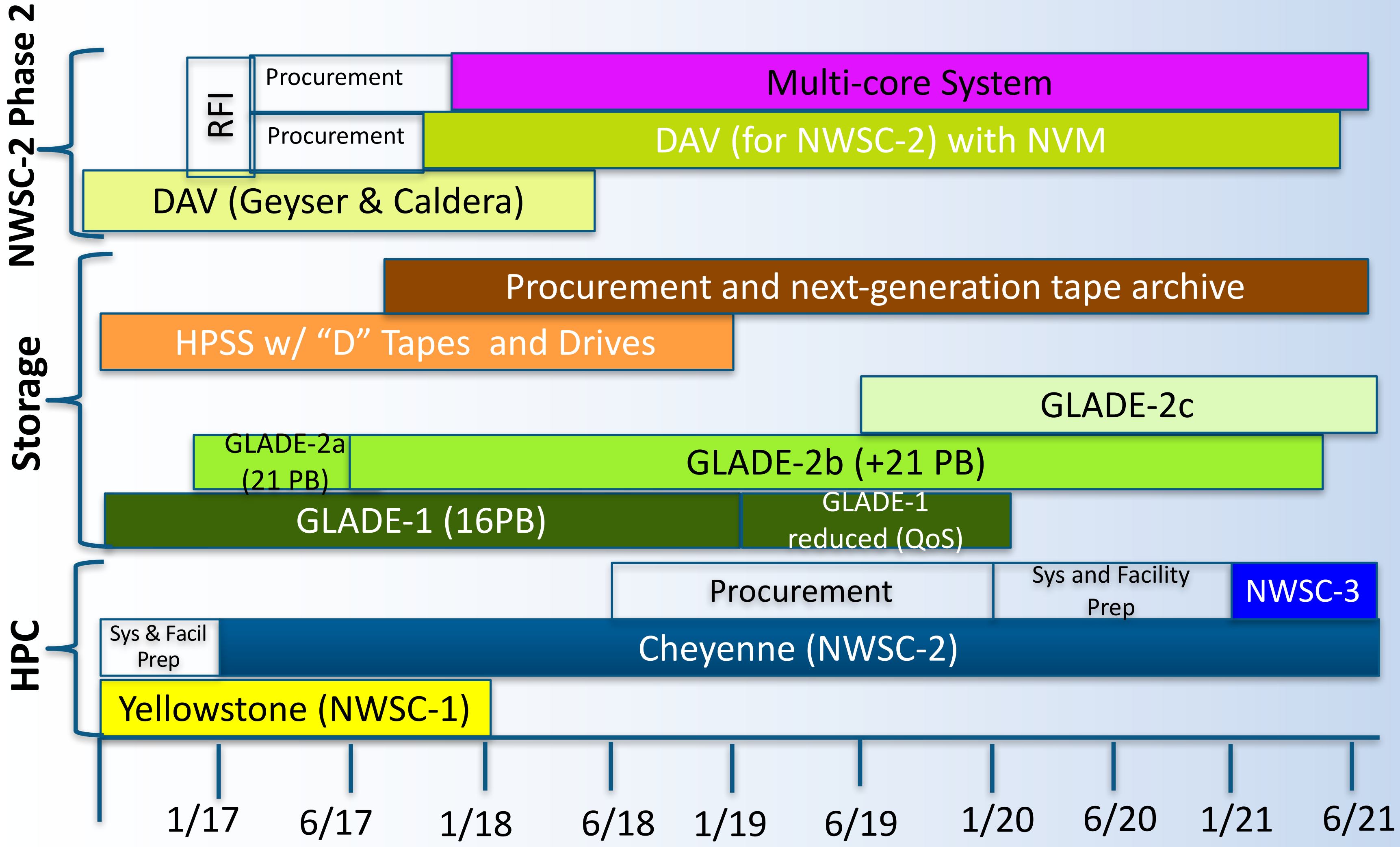
Parallelized workflow and SGI UV300 evaluation



NWSC-2a & NWSC-2b procurements

- NWSC-2a: Data Analysis and Visualization system
  - Geyser and Caldera replacement
  - User requirement evaluation underway
  - New science drivers and workflow
  - Experimentation and testing underway in HPC Future's Lab
  - Assessing value/need for SSD (endurance, latency, IOPS)
    - Looks like it will help climate analytics by 3x-5x
  - RFI planned in Q1 CY2017
- NWSC-2b: Many-core system
  - Experimental system, but likely will be phased into production
  - RFI planned in Q1 CY2017
  - Experimentation and testing underway in HPC Futures Lab

Supercomputing and Storage Roadmap



QUESTIONS?

Thanks to the many CISL staff who contributed to these slides and all those working on the installation of Cheyenne.