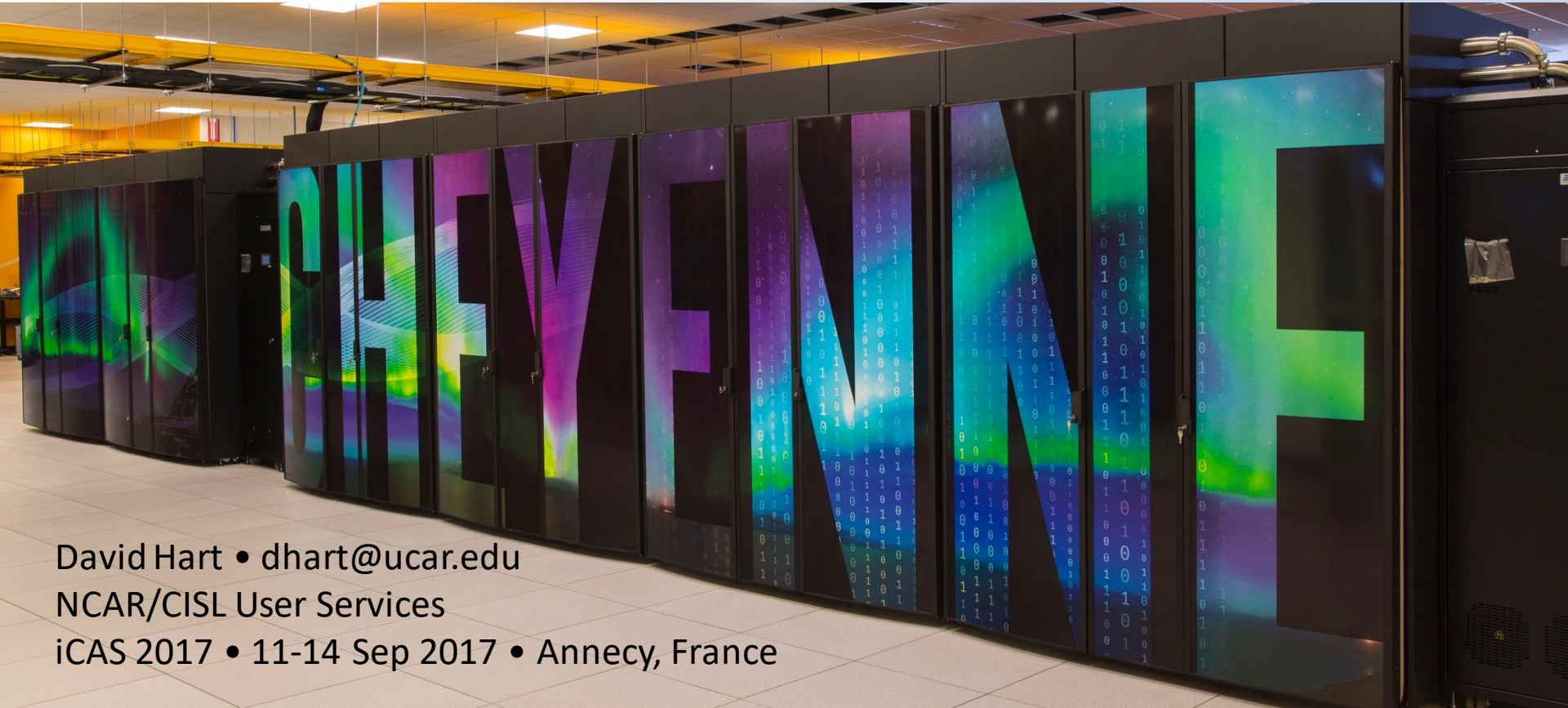
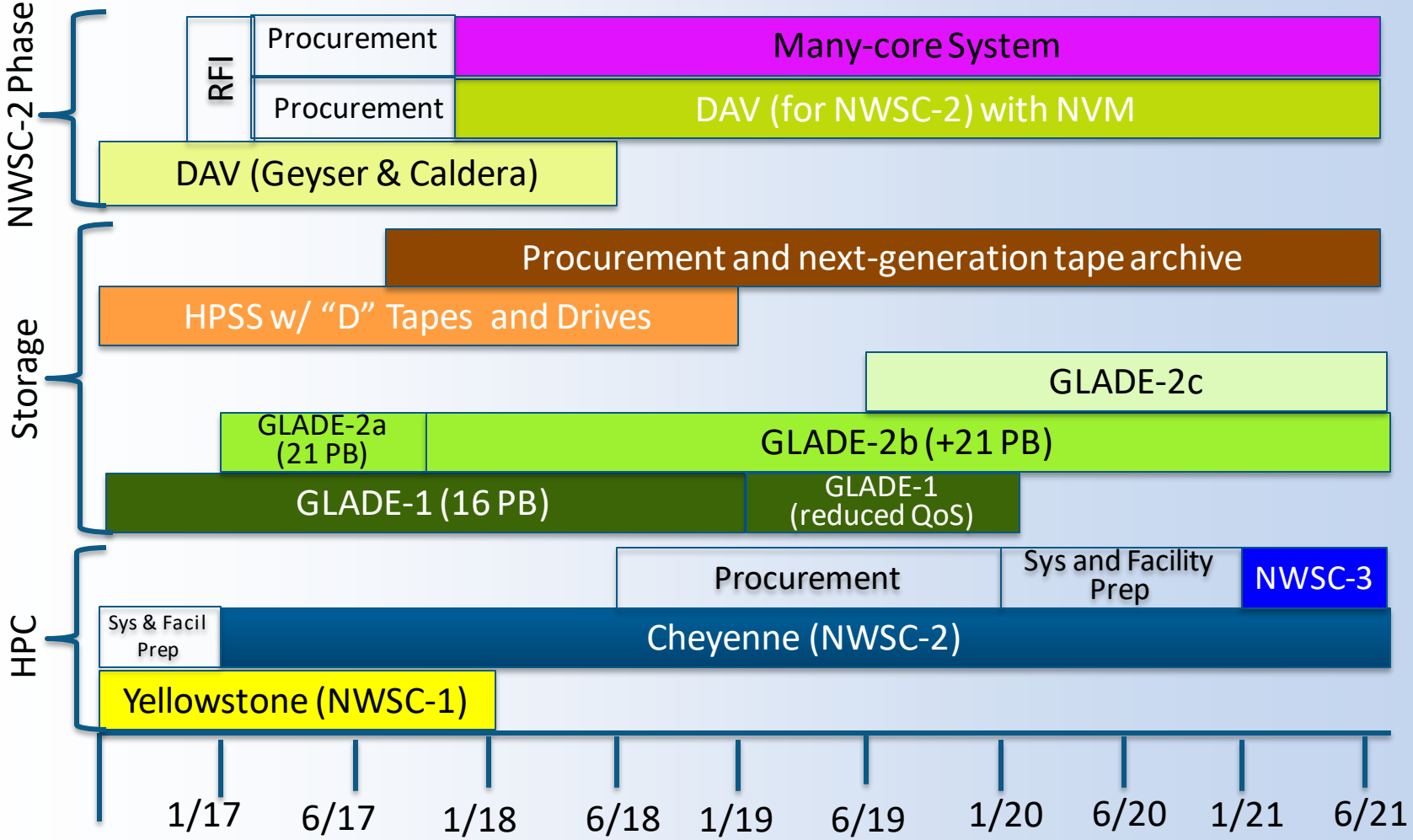
Cheyenne and Beyond: NCAR's Research Computing and Storage Roadmap to 2022



David Hart • dhart@ucar.edu
NCAR/CISL User Services
iCAS 2017 • 11-14 Sep 2017 • Annecy, France

The best-laid plans—
NCAR system roadmap, June 2016



CISL Computational & Information Systems Laboratory

NCAR

Cheyenne

Planned production: 2017 – 2021

- **SGI ICE XA cluster**
  - 4,032 dual-socket nodes
  - 18-core, 2.3-GHz Intel Xeon E5-2697v4
  - 145,152 Broadwell cores
  - 313 TB memory (64-GB & 128-GB nodes)
  - 5.34 PFLOPs peak
  - Mellanox EDR InfiniBand
  - 9-D enhanced hypercube
- **>3 Yellowstone equivalents on NCAR Benchmark Suite**

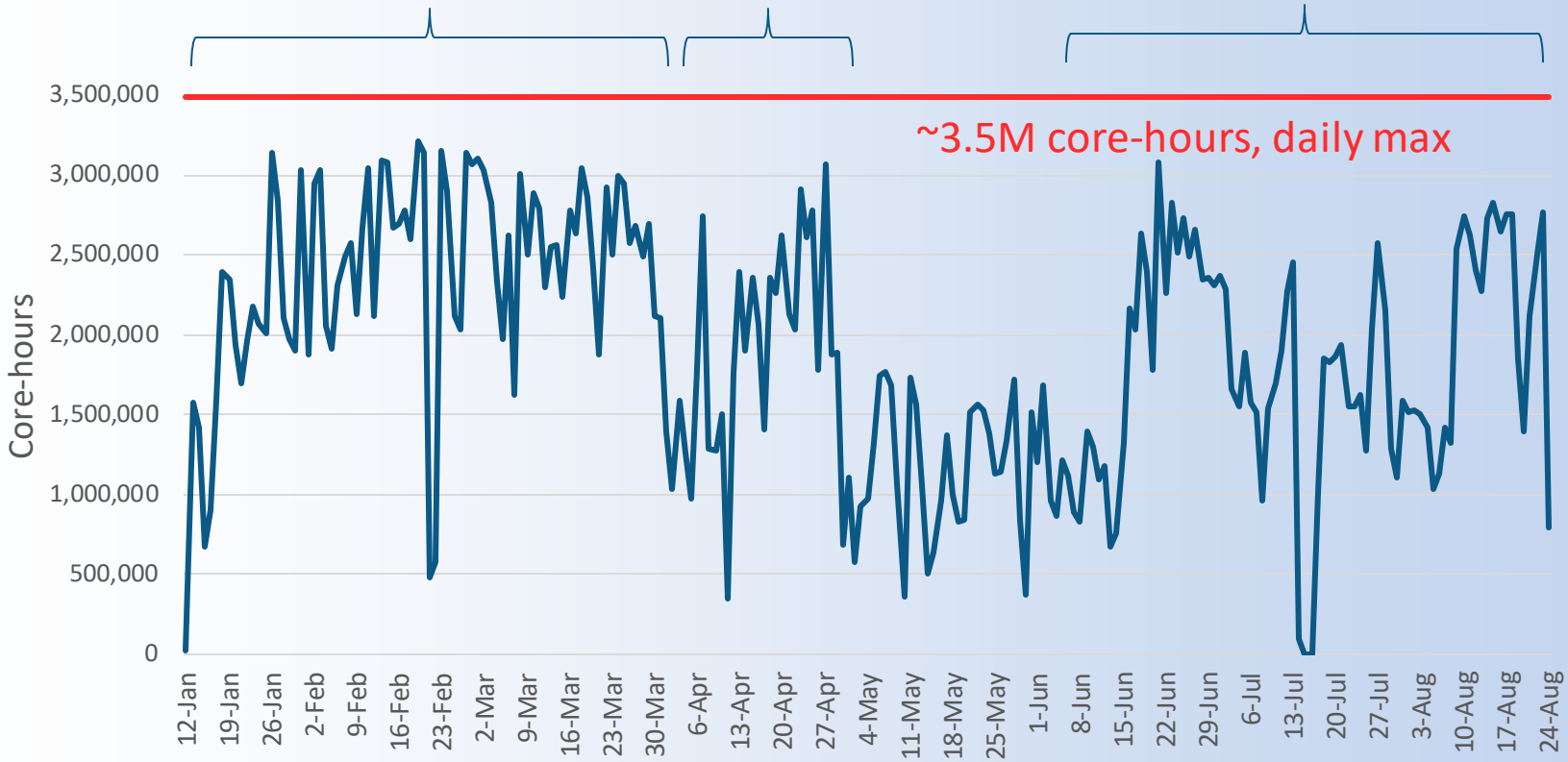
Cheyenne ribbon cutting ceremony held Aug. 17, 2017, as part of the city of Cheyenne's sesquicentennial.

Initial experience with Cheyenne

Early user period ran Jan. 13 through end of March

"Bonus" time for some early projects

Discount economy queue charging begins



CISL Computational & Information Systems Laboratory

NCAR NSF

Completing the NWSC-2 environment

- **Cheyenne strictly conventional multi-core architecture**
  - Also largely conventional GLADE disk storage
  - Most components low-risk, high-maturity
  - Model readiness a key driver for system choices
- **Cheyenne procured *without* new analysis or visualization clusters**
  - Stretching the lifetime of Geyser (Westmere) and Caldera (Sandy Bridge) clusters
- **After Cheyenne, NCAR planned to take steps to prepare for next-generation system**

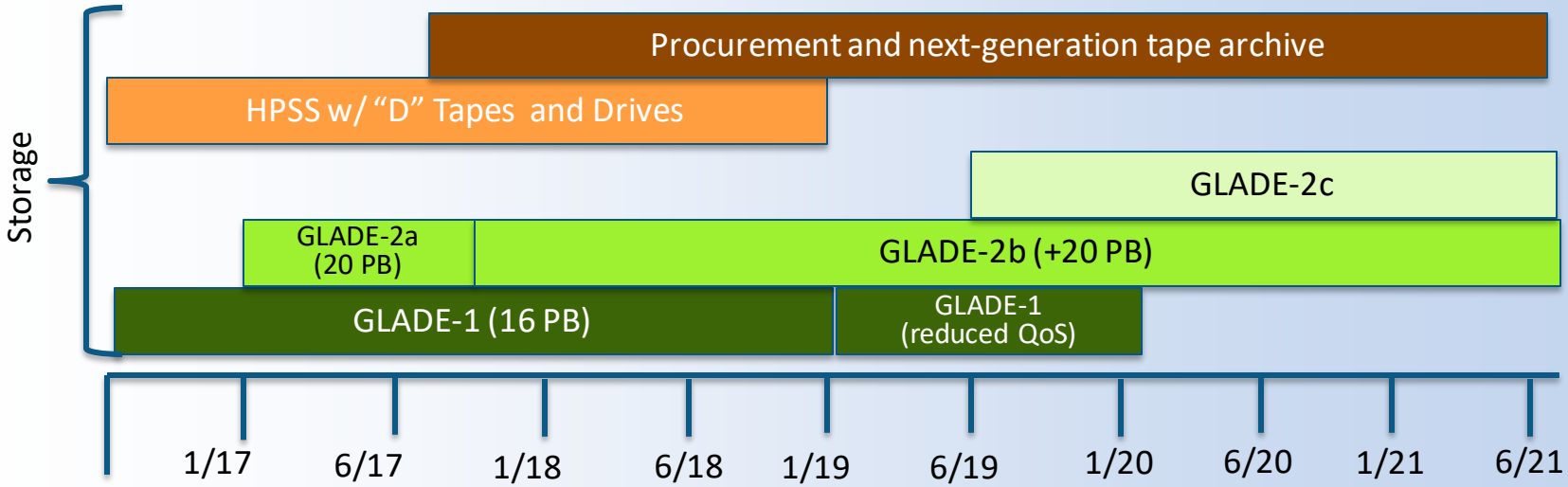
Post-Cheyenne procurements & goals

In first half 2017, NCAR issued a "request for information" on systems to fill out the Cheyenne environment, support science workflows, and look to the future.

- **NWSC-2a: Data analysis and visualization**
  - To replace aging Geyser and Caldera systems
  - Ready for machine learning/deep learning
  - Assessing value/need for SSD (endurance, latency, IOPS)
- **NWSC-2b: Alternate architecture(s) system**
  - Experimental system to evaluate "post-multicore" architecture(s)

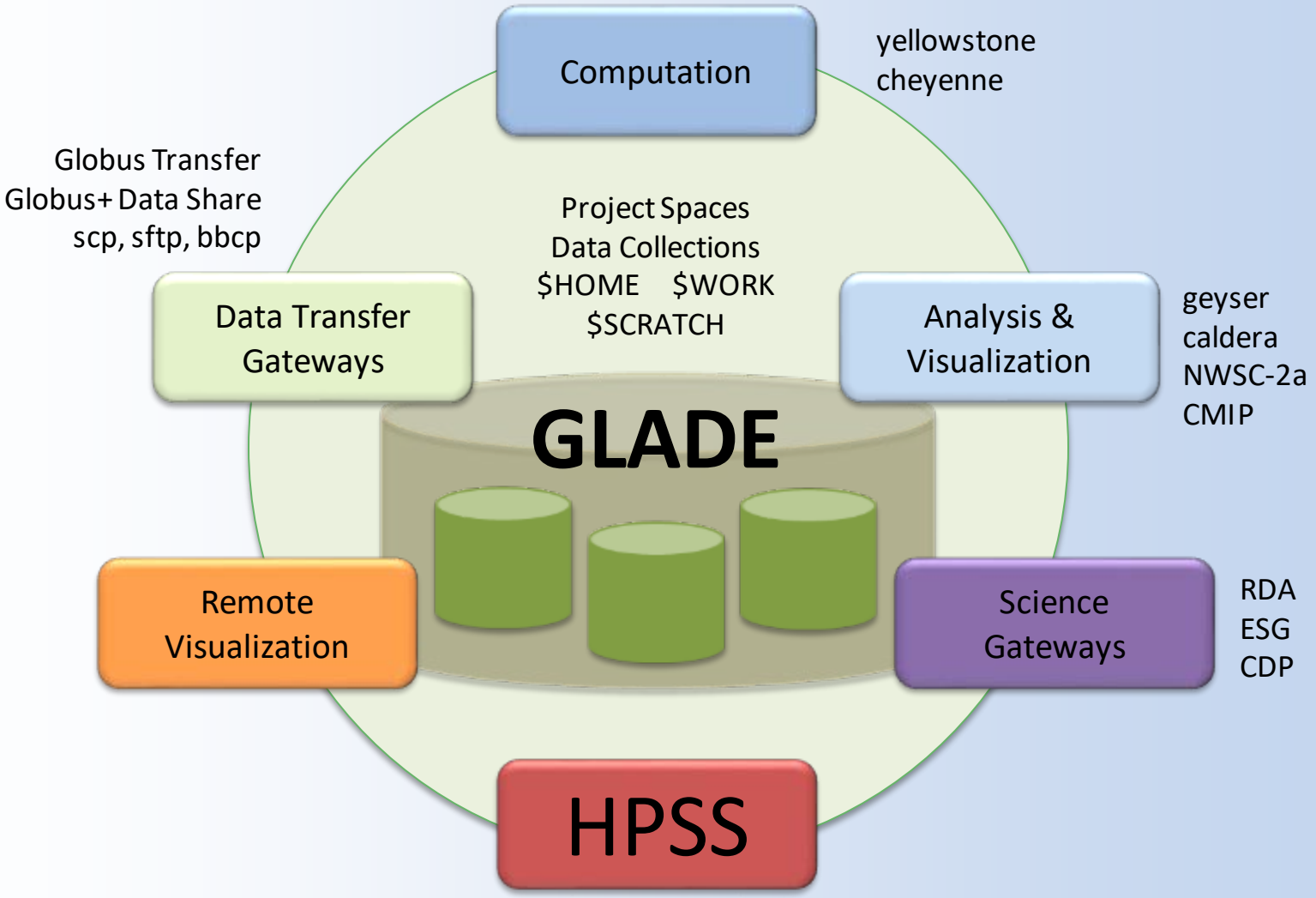
5y0, 1.9MW cluster, free to good home

- **Yellowstone to be decommissioned Dec. 2017**
- **2,900 users served**
- **2.7 billion core-hours delivered**
- **18 million jobs completed**
- **One final experiment planned: FDR vs. 40 GigE showdown**
  - Compare performance on full machine of HPCG, HPL & all-to-all
- **Erebus already decommissioned**
  - Being used as experimental cloud platform



EVOLVING STORAGE FUTURES

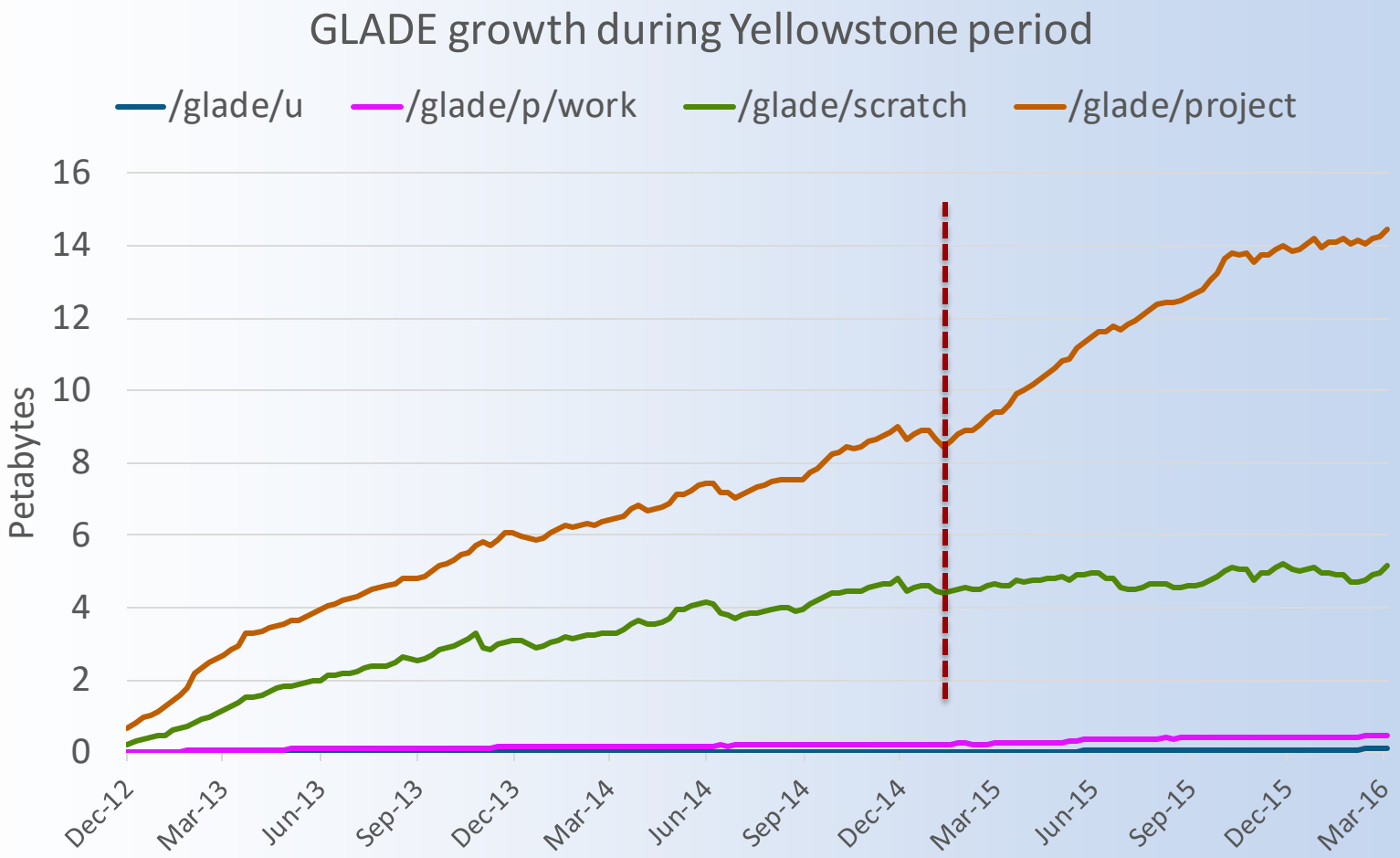
GLADE environment



NWSC-2c — GLADE plus SSDs

- **An early divergence from original plan for NWSC-2a and NWSC-2b systems**
  - Original plan assumed SSDs would be part of NWSC-2a
- **NCAR decided to acquire separate flash/SSD-based shared storage system**
  - Augment existing GLADE with latest and greatest offerings in storage technologies
  - Mounted on Cheyenne, NWSC-2a & future systems

Tracking and predicting data growth



And plans started going awry

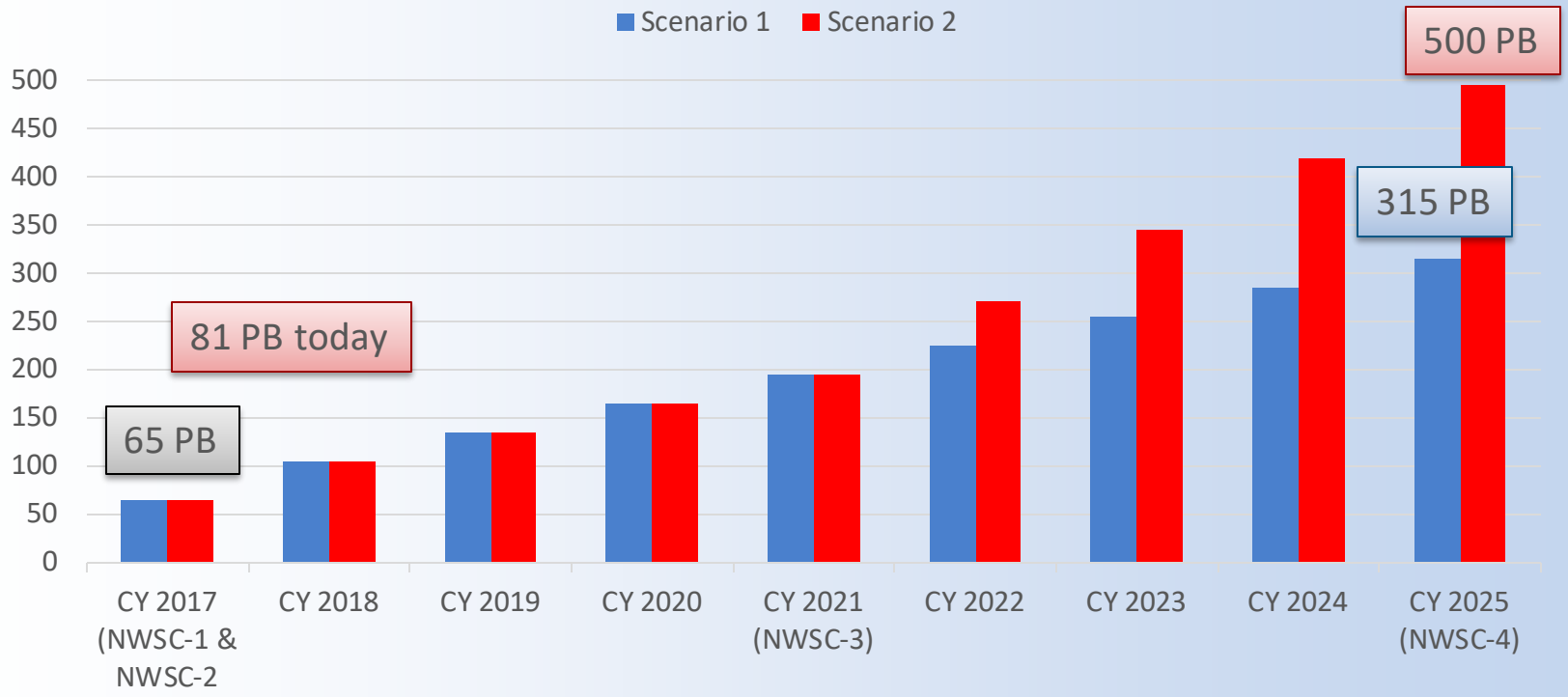
- **The original 16-PB GLADE reaching and exceeding capacity coincided with Cheyenne's arrival**
- **In early 2017, HPSS tape writes surged to 2+ PB per month**
- **First reaction was that Cheyenne caused the spike —but Cheyenne had not been opened to general users**
- **…and then Oracle dropped "E" generation tapes and drives from their roadmap**

Dwindling tape supplies

date	empty tapes	data written to tape in prior week (PB)	tape depletion date
02Jan2017	2,027	0.216	Jun 2018
09Jan2017	1,999	0.224	May 2018
16Jan2017	1,962	0.296	Jan 2018
23Jan2017	1,910	0.416	Sep 2017
30Jan2017	1,850	0.480	Aug 2017
06Feb2017	1,790	0.480	Aug 2017
13Feb2017	1,708	0.656	Jul 2017
20Feb2017	1,640	0.544	Jul 2017

At 2 PB per month, NCAR's HPSS would have run out of tapes in a matter of months

Projecting HPSS growth

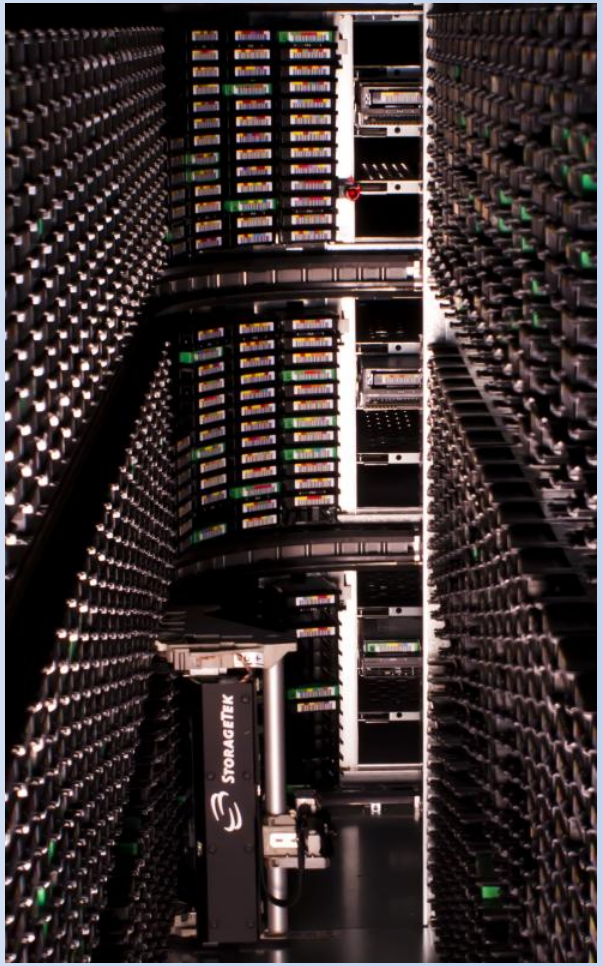


Scenario 1—4 PB/month while both Yellowstone and Cheyenne in production during 2017, then 2.5 PB/month for next seven years.

Scenario 2 — 4 PB/month while both Yellowstone and Cheyenne in production during 2017, then 2.5 PB/month for next three years, and 6.25 PB/month for the four years of NWSC-3.

Challenges facing NCAR services

- **Tape is best for certain use cases**
  - Long-term storage, infrequent use
  - Well-organized data for efficient reads
  - Preservation of data that can't be reproduced
- **"Disk" is not a one-size-fits-all solution**
- **Users have other needs on a scientific timescale**



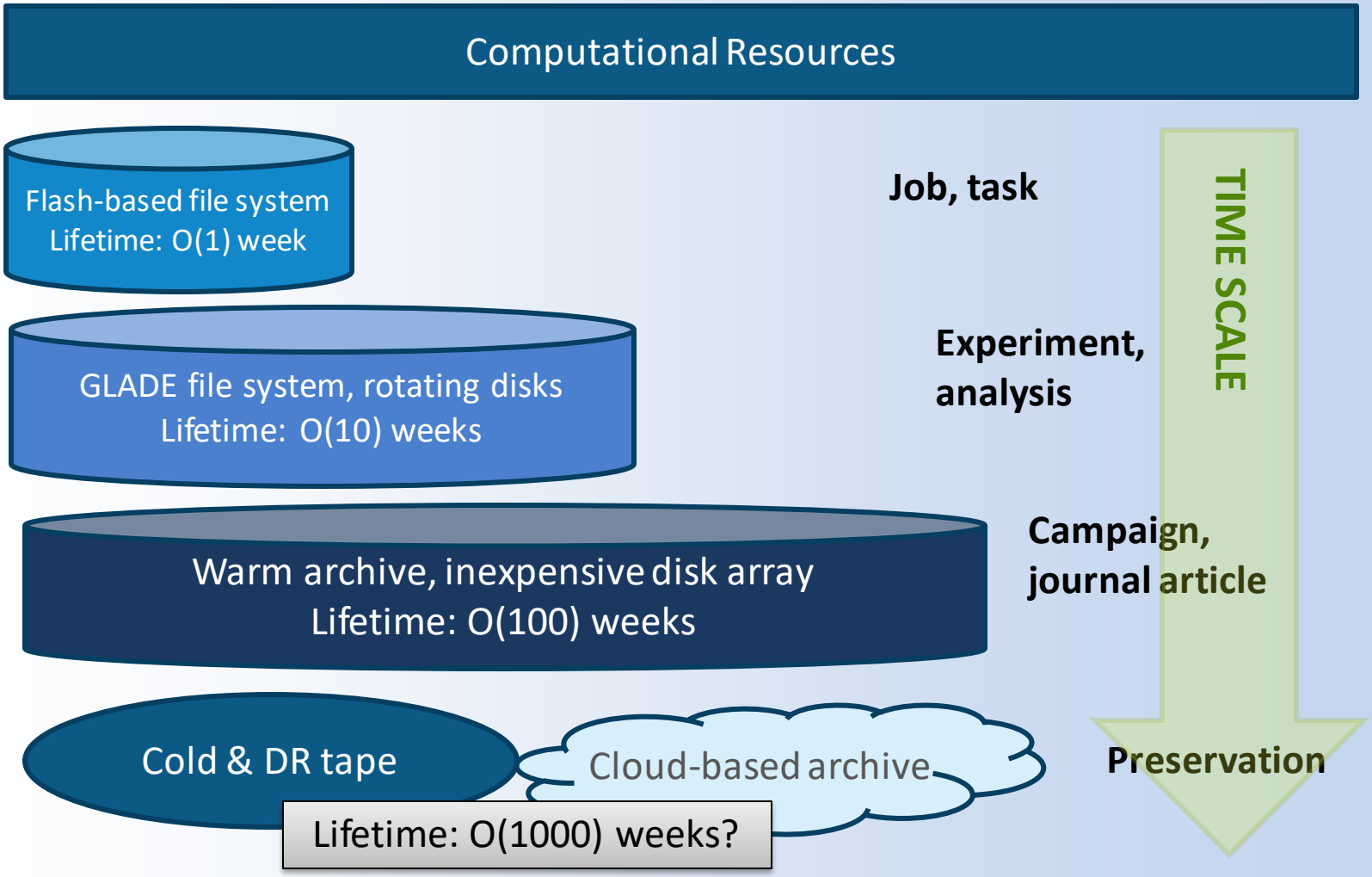
Challenges facing users

- Users can produce more data than NCAR can afford to store
- They can generate data faster than they can—or want to—annotate, catalog, and organize
- How do they store and find the data they need?
- Do they have a good scientific sense of the data they have archived?
- Should they reproduce a data set rather than store the data?
- What is the useful lifetime of data collections?
- Unclear technology risk/cost/data value trade-offs

Adapting NCAR's storage profile

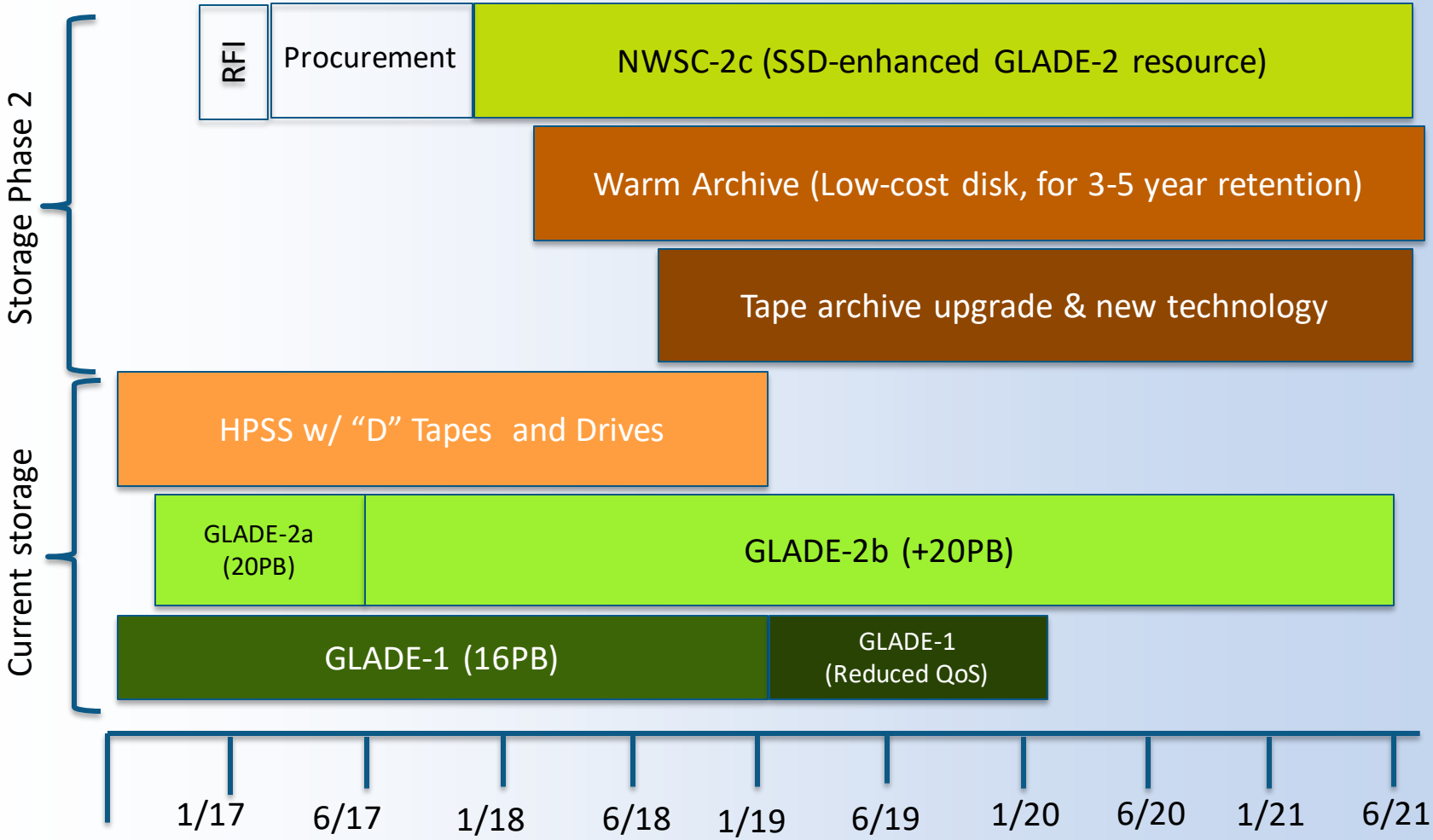
- **Consulted with users on their needs**
- **NCAR missing a medium-term storage service**
- **Options available:**
  - Expand HPSS disk cache, system-managed migration to tape
  - Expand POSIX-accessible GLADE file systems
  - **"Warm-archive" tier between GLADE and HPSS**
- **NCAR disk capacity needs to expand annually**
  - Buying only more *tape* each year does not address all the users' needs

Storage for data lifecycles



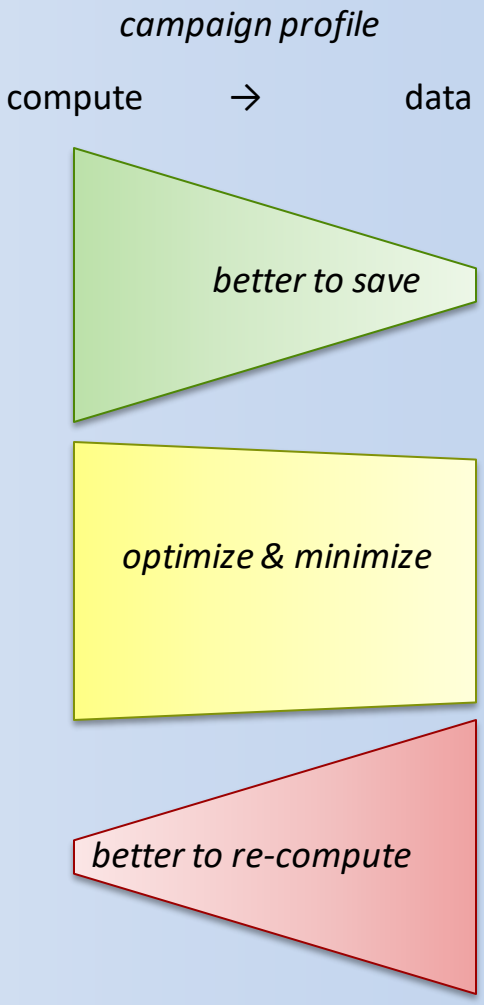
See also "APEX Workflows" whitepaper by LANL, SNL, NERSC

NCAR storage roadmap—2017



Users need to change behavior

- **Data campaigns, not computational campaigns**
  - Bytes, not flops, are often the more scarce resource
  - Computational plans should *follow from* data management plans, not vice versa
- ***Data output, workflows, and management plan* should be optimized**
  - "Run all, analyze later" no longer a feasible default workflow…certainly not for large data campaigns
- **Sometimes, it may be better to re-compute!**
  - Repeated computing may be better than saving extra data "just in case"
  - Re-computing also has human costs, raises questions of reproducibility
- **Users, not technology, have to make choices about the scientific value of data**
  - And help decide how to meet data access mandates from funding agencies and journals



Changing hearts and minds

- **Business as usual is not an option**
  - We need data management plans, not data storage calculations
- **NCAR science labs being handed greater responsibility for their storage footprints, as well as computing needs**
  - Science decisions must be made
  - Some risk of personal storage systems springing up in unused closets
- **Integrate storage services into NCAR-defined policies and processes for data management**
  - Policy becomes much less hypothetical when combined with a finite storage budget

NCAR's Digital Asset Services Hub

Services and resources focused on supporting digital assets across NCAR to make them available to the broader scientific community.
Leveraging DASH policies to frame storage system policies.

Tier 0

- **Unpublished**
- Normal allocations
- Owner managed
- Disk, tape as available to labs
- ***Limited availability***

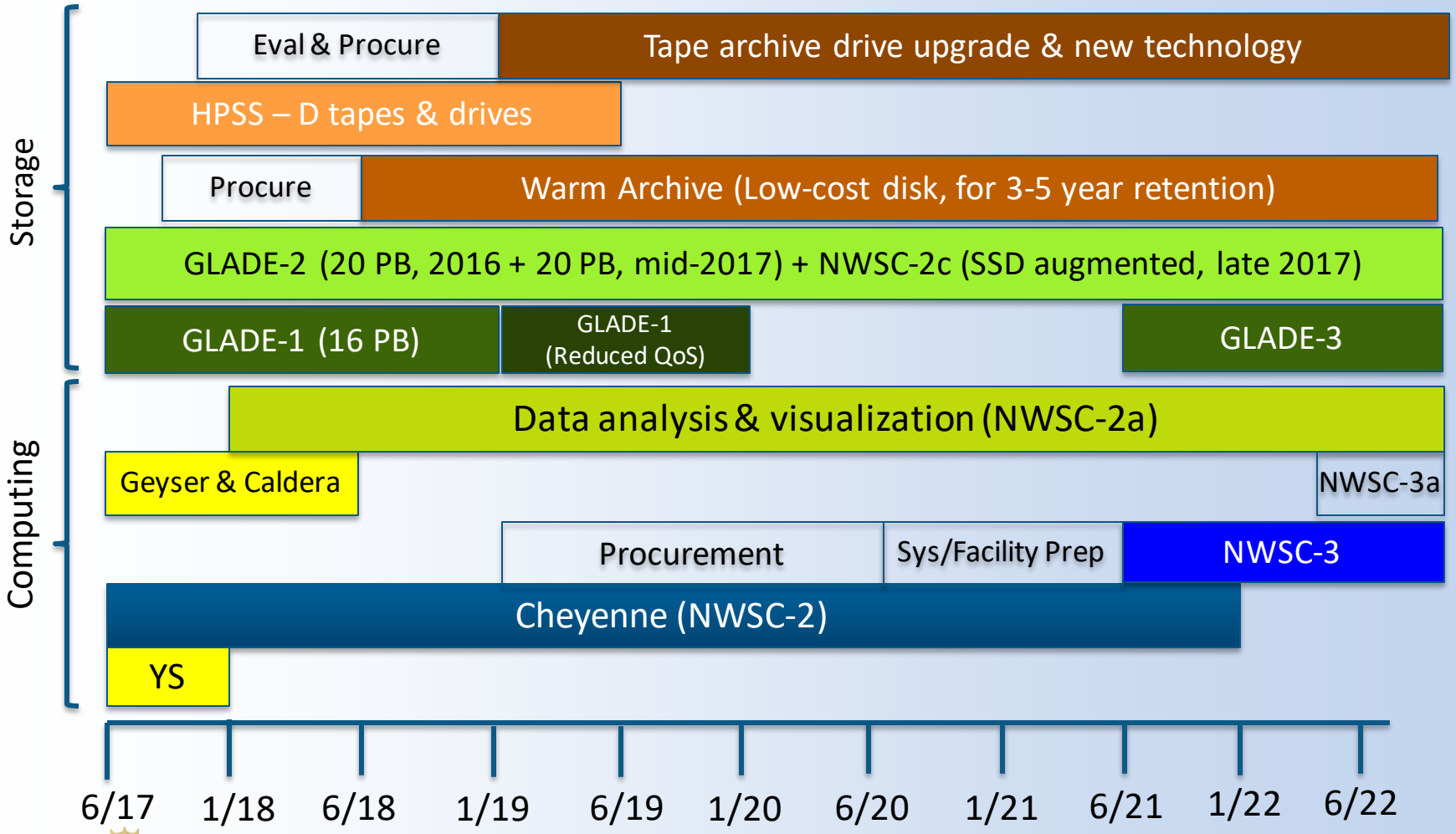
Tier 1

- **Owner-published**
- Metadata Requirements
- Open Access allocations
- Metadata & archiving review by DASH team
- Archiving and web access by owner
- Disk and limited term archiving
- Preservation— 3-5 years

Tier 2

- **DASH Repository**
- Metadata Requirements
- Data format standard
- Web (public) access provided by DASH
- Preservation— 5-10 years
  - renewable with justifying metrics
- Possible disaster recovery copies

The best-laid plans—
NCAR system roadmap, late 2017



CISL Computational & Information Systems Laboratory

NCAR

QUESTIONS?

Thanks to

- Irfan Elahi, Pam Hill, and Erich Thanhardt of CISL's High-End Services Section
- Steve Worley and Sophie Hou of NCAR's Data Stewardship and Engineering Team (DSET) and DASH
- And many others