NWSC Storage: A look at what users need

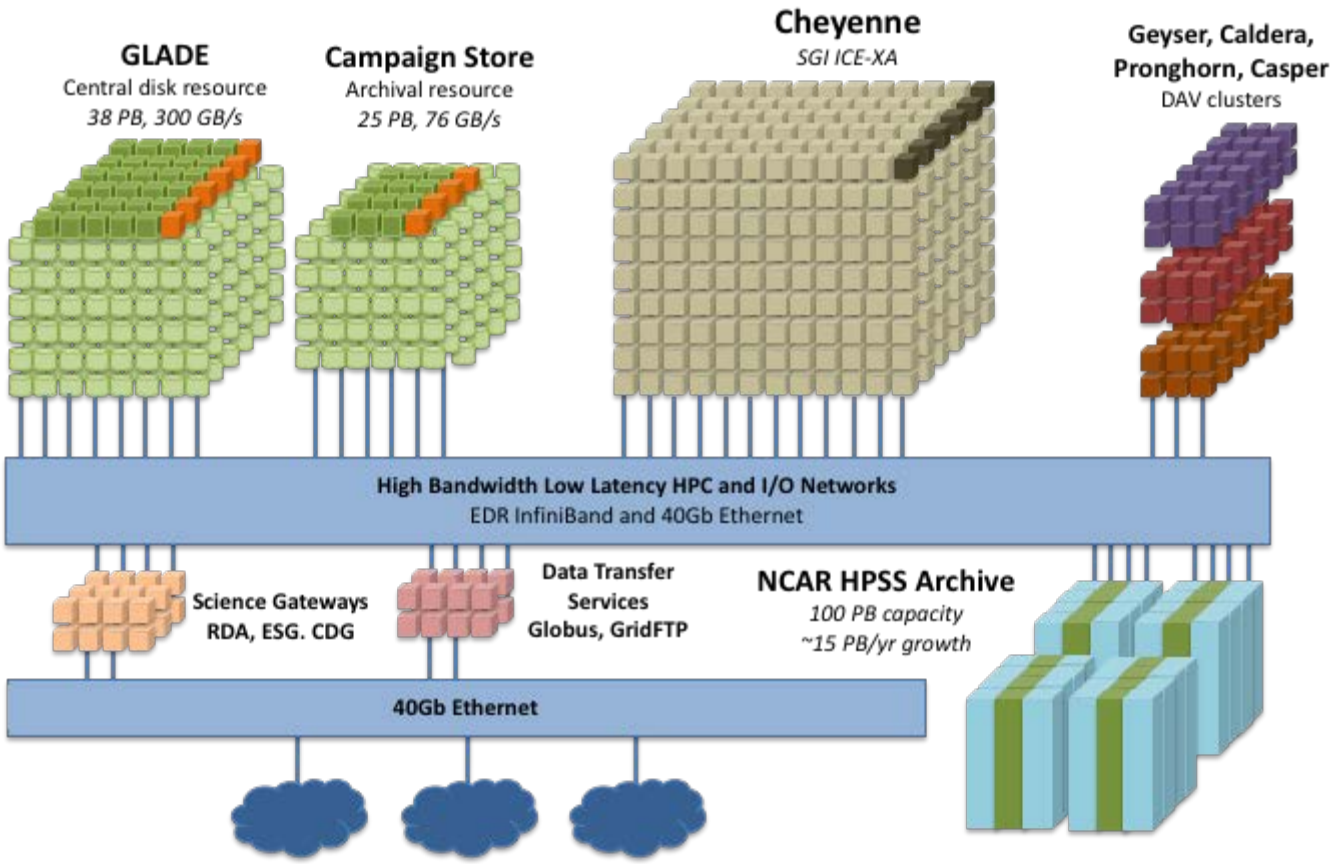
Christopher Hoffman
National Center for Atmospheric Research

35th International Conference on Massive Storage Systems and Technology
2019

Brief History

- NCAR Wyoming Supercomputing Center founded in 2012
  - Funded by NSF and the State of Wyoming
  - Located near Cheyenne, WY
- Previous Data Center located Boulder, CO at Mesa Lab
- 90 minutes to compute center
- 4 MW capacity now, expandable to 8 or to 16MW with additional building

Production Environment



Cheyenne

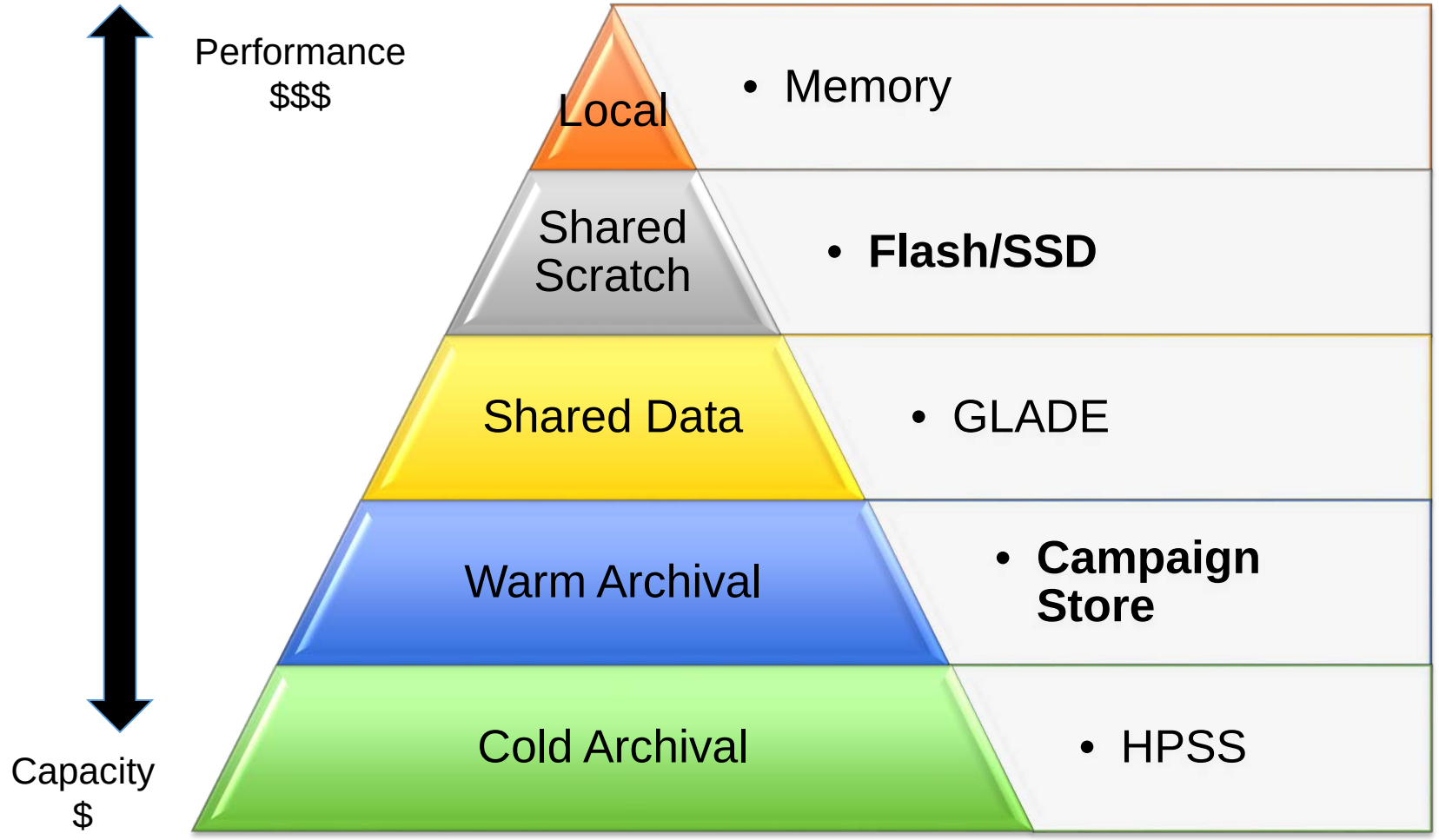
Planned production, January 2017 – December 2021

- 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
  - 313 TB total memory (3,164 64-GB and 864 128-GB nodes)
- High-Performance Interconnect
  - Mellanox EDR InfiniBand
  - 9-D enhanced hypercube topology
  - 97 Gbps link bandwidth — 0.5 µs latency
  - 224 36-port switches, no director switches
- GLADE — Central file systems and storage
  - 38 PB usable
  - 8x DDN SFA14KXe each with 10x 84-slot drive chassis
    - 32 embedded NSD servers
    - 6,580 8-TB SAS disk drives
    - 160 4-TB SSD drives
  - ~300 GB/s aggregate I/O bandwidth for new capacity
  - Currently 1.5 Billion files
  - IBM Spectrum Scale (GPFS) file system

Problem At Hand

- We must provide more with same budget
- Recent Storage Technology
  - Burst Buffer
  - Campaign Storage
  - Object Stores
  - Flash
- Many different view points
  - Atmospheric Modeling
  - Batch and Grid Computing
  - Data Stewards and Curators
  - Storage Admins
  - Directors, Strategic Planning for Storage
- Must come up with solution that fits best

New Storage Tiers



New Storage Tiers Challenges

- Added Flash and Campaign Store
  - New tiers data migration is needed
  - Current tools aren't sufficient to move PBs of data
- Flash adoption challenges
  - Flash has purge of 14 days versus longer periods on scratch and projects
  - While flash has more IOPs and bandwidth per TB, due to sizing is similar to GLADE
- Campaign Store challenges
  - Space not best utilized
  - Quotas too small
  - Oversubscription possibly needed
  - After nearly a year 32% used or 8PB

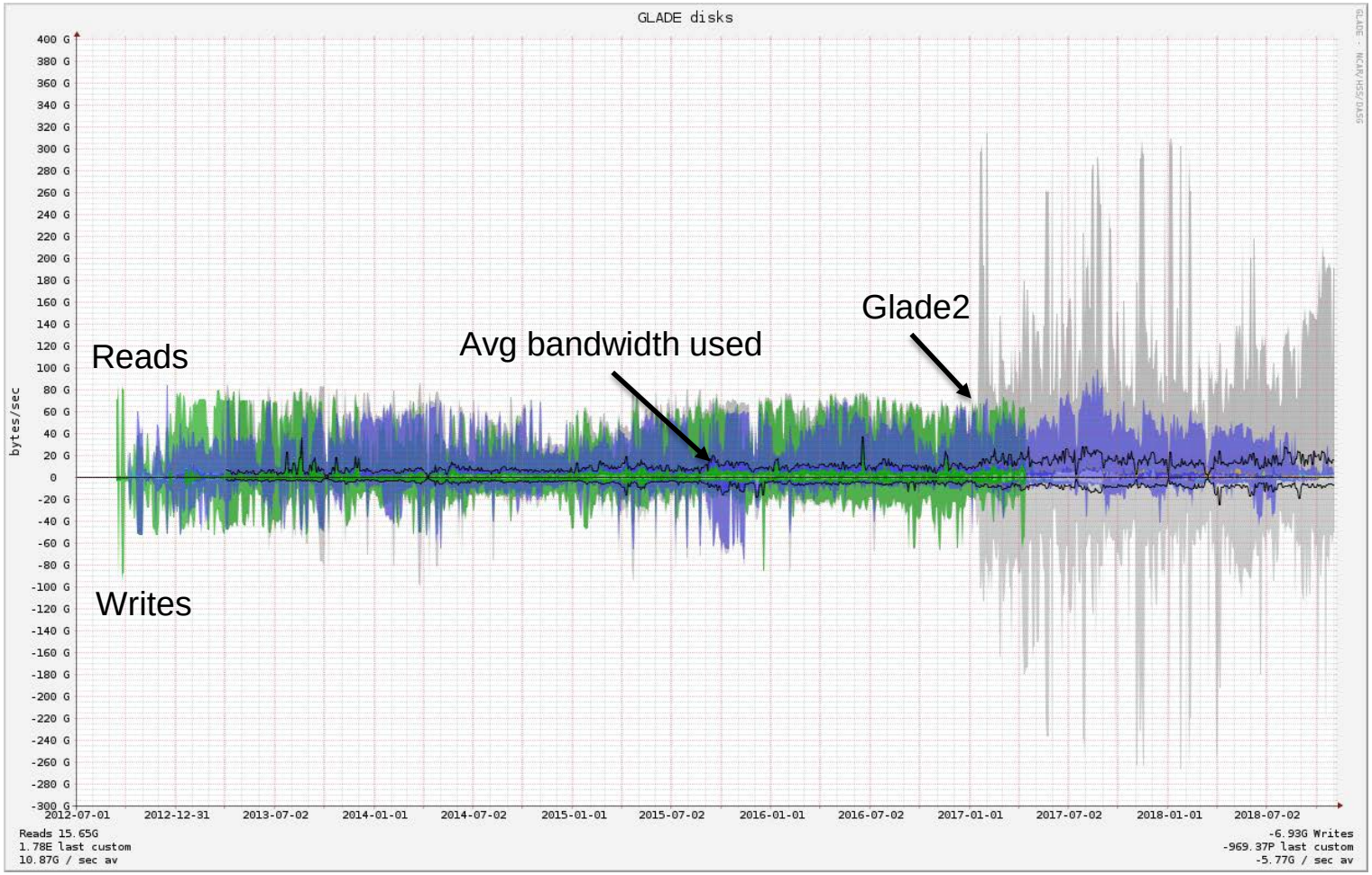
Globus

- Introduced Globus Interface
  - Data between movement between tiers: Flash, GLADE, Campaign Store, HPSS, offsite
- Challenges
  - New interface
  - Authentication
  - Only interface for Campaign Store
  - Timeouts on large directories
    - Web, CLI and REST interface
  - Large directories won't transfer
  - No permission management capability
  - No find capability
    - Cant manually traverse as directory listings timeout
- We noticed moving data around isn't fun for users

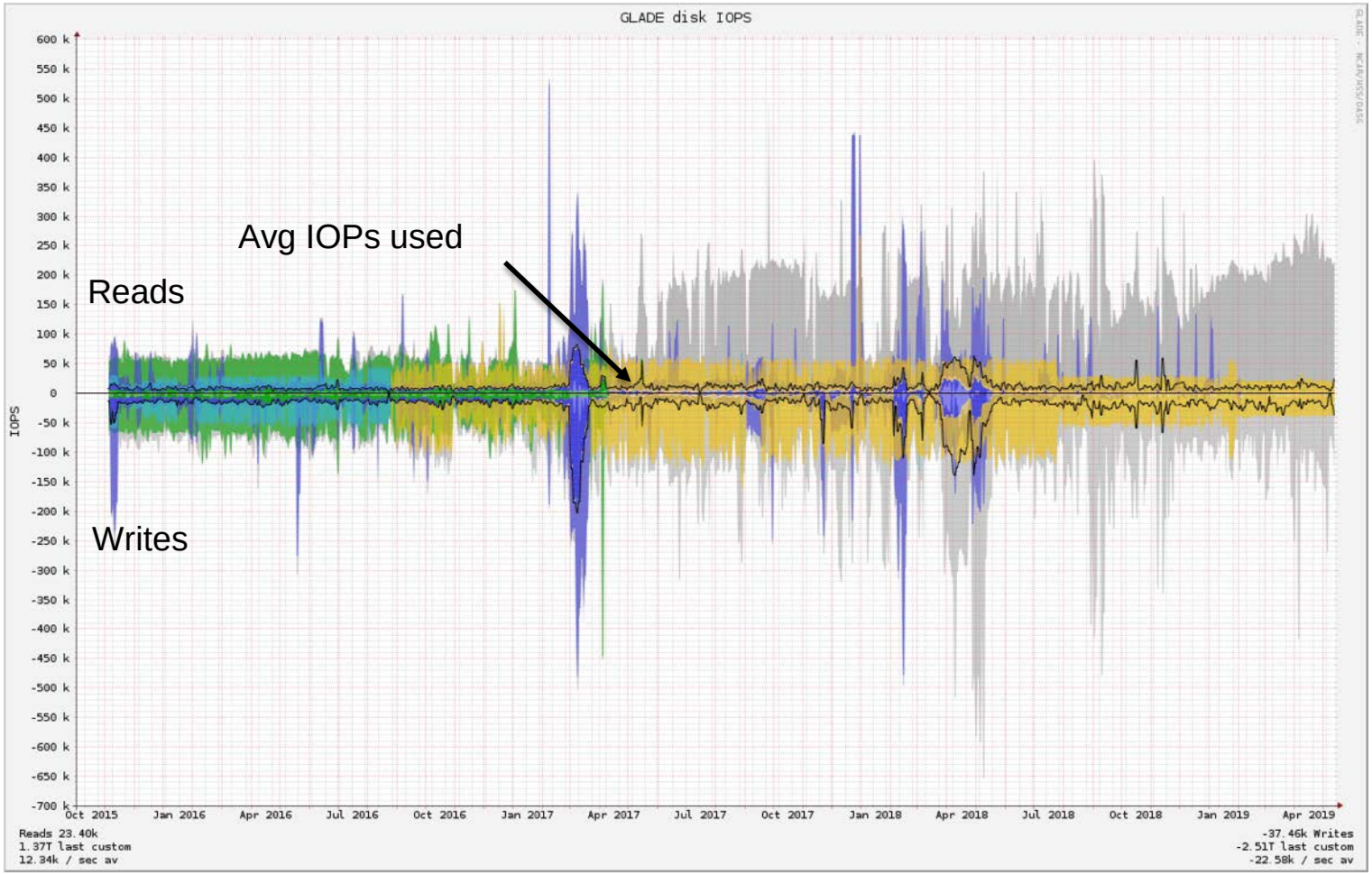
The response

- Changing user habits
  - Tools, scripts and workflows
    - Changing existing workflows and tools that just dump everything to HPSS
    - Creating new tools to utilize Globus
  - Tutorials
  - Documentation
- Users want more reliability, availability, and capacity
- Hardware configuration exploration
  - Segregated compute and storage networks
  - Introduce more reliable interconnects such as Ethernet
  - Some users want no purge, some want longer purge times
  - Historically compute goes up 2-4x each cycle, storage 1.3—1.5x
  - Buy more commodity storage for Campaign Store
    - Cost per bit of NL storage is finally coming down, MAMR and HAMR coming soon
    - Utilize tape in this model

GLADE Performance – Past 6 years



GLADE IOPs


GLADE disk IOPS
IOPS
600 k
550 k
500 k
450 k
400 k
350 k
300 k
250 k
200 k
150 k
100 k
50 k
0
-50 k
-100 k
-150 k
-200 k
-250 k
-300 k
-350 k
-400 k
-450 k
-500 k
-550 k
-600 k
-650 k
-700 k
Oct 2015
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
Oct 2017
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
Apr 2019
Avg IOPs used
Reads
Writes
Reads 23.40k
1.37T last custom
12.34k / sec av
-37.46k Writes
-2.51T last custom
-22.58k / sec av

Flash Tier IOPs Performance

Reduce Archive Footprint

- HPSS Cool down project
  - This project will lower growth and footprint of archive
    - Important for migration of data from Oracle drives and tape
  - Phase 1
    - Goal: HPSS to be a true cold archive and not active archive
    - Historically growing 1-2PB/month
    - Slowed down to 200-300TB/mo
    - Actively engaging power users in justification
    - Pushback due to challenges related to movement and data policies
  - Phase 2
    - Goal: Inventory and do clean up on existing 90PB+ of data
    - Keep archival data, move active data to Campaign Store
    - Identify data sets with active UCAR/NCAR employees
    - Identify data sets with no active users, correspond to PI or Management
- Common theme we are seeing is data management.

Data Management

- Users and Data Managers don’t know what they have
- How do I search files to see what I have
  - Globus has timeouts
  - Inode trees are costly to crawl for every query
- We looked at various market tools available and decided to go with OSS
  - We wanted something that is easy to understand and troubleshoot
  - We wanted to be able to add features
  - Low barrier to testing
- NCAR will pilot Grand Unified File Index
  - Open source project from LANL and has vendor interest
  - GLADE home spaces will be first
  - Data Stewards will be first to test

Other misc things to consider

- Which of the tiers are really needed?
- How to best utilize tiers
  - Data Management with automated moving?
- We can implement new tiers before our tech to use them is there
  - Less $$ for other requirements
- Object Store
  - Rework of Gateway Portal codes , Climate Codes, etc
  - Some POSIX capability required
  - Instead of putting POSIX on top of OS, why not other way around
    - S3 application servers on top of POSIX

What we learned

- Campaign Store has been overall positive
- Flash adoption continues to be low
- Changing user habits is a hard and slow process
- User engagement is crucial to driving requirements
  - Same faces give feedback
  - Need more engagement
- We need more of overall budget for storage
- Explore options to minimize the amount of logical tiers
- Coordinating user needs and vendor offerings is a challenging task

Questions