NERSC

National Energy Research Scientific Computing Center

Nicholas Balthaser

Kirill Lozinskiy

Melinda Jacobsen

Kristy Kallback-Rose

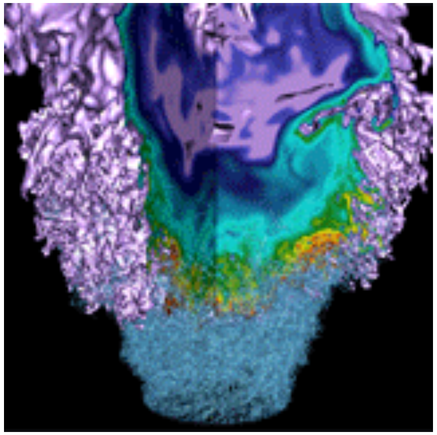
NERSC Storage Systems Team

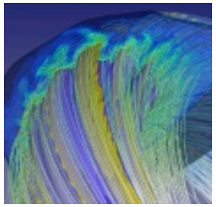
October 16, 2019

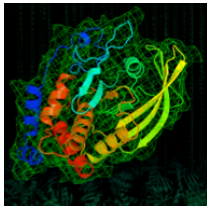
Agenda

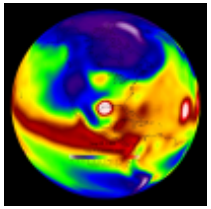
- General NERSC & Systems Overview
- Storage 2020 Strategy & Progress
- GHI Testing
- Tape Library Update
- Futures

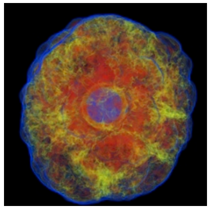
NERSC & Systems Overview

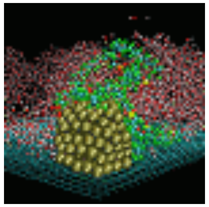






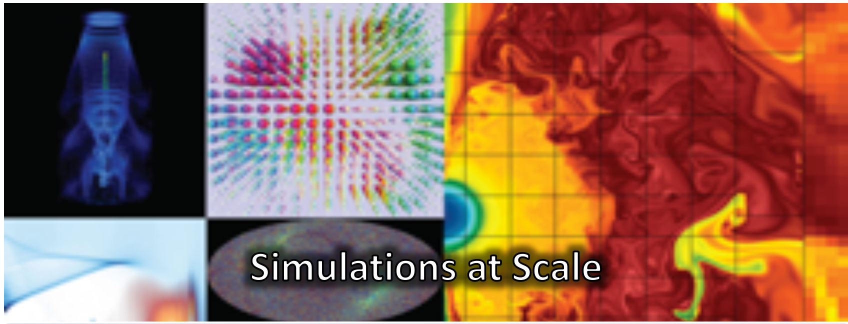




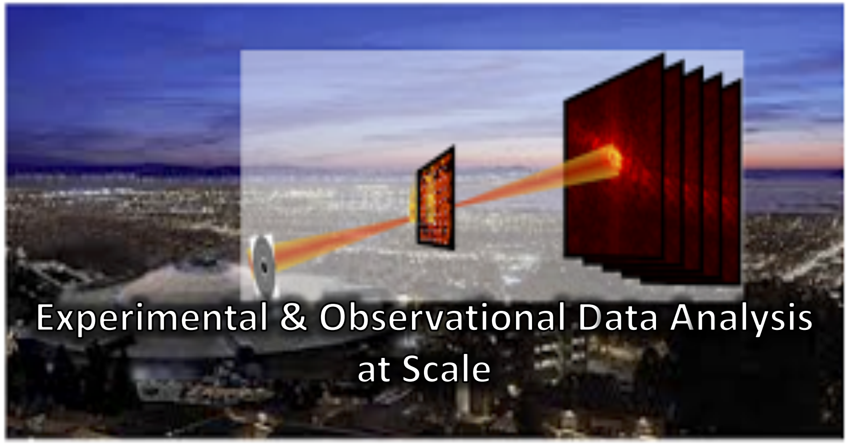


NERSC is the mission HPC computing center for the DOE Office of Science

- **HPC and data systems for the broad Office of Science community**
- **Approximately 7,000 users and 870 projects**
- **Diverse workload type and size**
  - Biology, Environment, Materials, Chemistry, Geophysics, Nuclear Physics, Fusion Energy, Plasma Physics, Computing Research

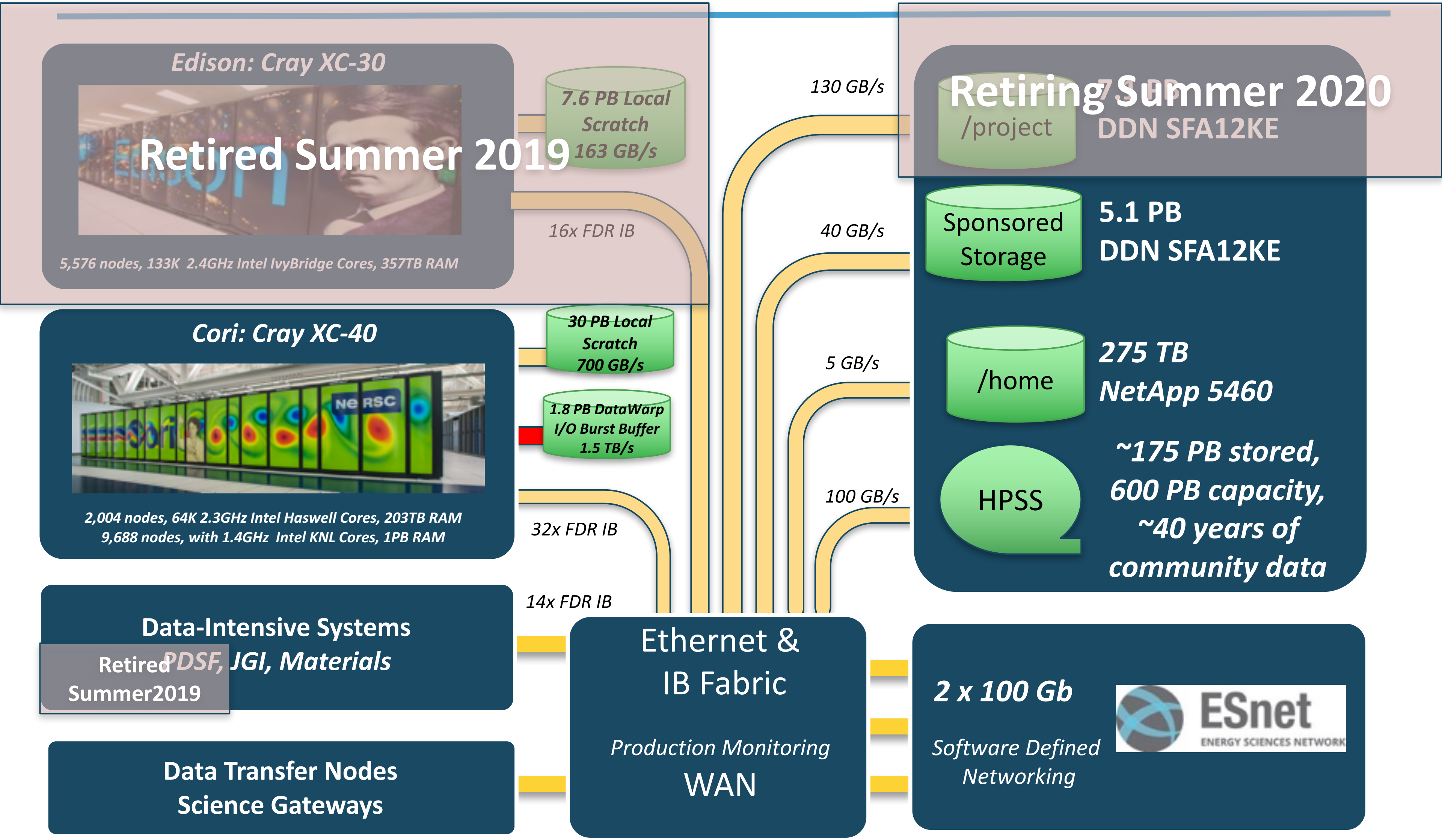


Simulations at Scale



Experimental & Observational Data Analysis at Scale

NERSC - Resources at a Glance 2019



Edison: Cray XC-30

Retired Summer 2019

5,576 nodes, 133K 2.4GHz Intel IvyBridge Cores, 357TB RAM

7.6 PB Local Scratch 163 GB/s

16x FDR IB

Cori: Cray XC-40

2,004 nodes, 64K 2.3GHz Intel Haswell Cores, 203TB RAM
9,688 nodes, with 1.4GHz Intel KNL Cores, 1PB RAM

30 PB Local Scratch 700 GB/s

1.8 PB DataWarp I/O Burst Buffer 1.5 TB/s

32x FDR IB

14x FDR IB

Data-Intensive Systems
PDSF, JGI, Materials

Retired Summer2019

Data Transfer Nodes
Science Gateways

Ethernet &
IB Fabric

Production Monitoring

WAN

130 GB/s

Retiring Summer 2020

7.1 PB /project DDN SFA12KE

40 GB/s

Sponsored Storage — **5.1 PB DDN SFA12KE**

5 GB/s

/home — **275 TB NetApp 5460**

100 GB/s

HPSS — **~175 PB stored, 600 PB capacity, ~40 years of community data**

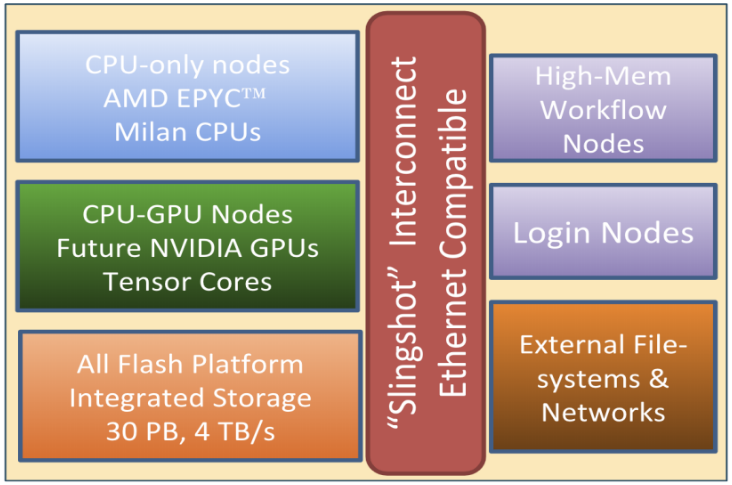
2 x 100 Gb

Software Defined Networking

ESnet
ENERGY SCIENCES NETWORK

NERSC-9 aka Perlmutter

- **Designed for both large scale simulation and data analysis from experimental facilities**
- **Overall 3x to 4x capability of Cori**
- **Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes**
  - >4,000 node CPU-only partition provides (same capability as all of Cori)
  - GPU nodes: 1 AMD Milan CPU + 4 NVIDIA GPUs
- **Slingshot Interconnect**
  - Capable of Terabit connections to/from the system
  - Ethernet compatibility
  - Adaptive Routing/Congestion Control
- **Single Tier, All-Flash Lustre scratch filesystem**
- **Robust readiness program (NERSC Exascale Science Applications Program, NESAP)**
- **Delivery in late 2020**



CPU-only nodes
AMD EPYC™
Milan CPUs

CPU-GPU Nodes
Future NVIDIA GPUs
Tensor Cores

All Flash Platform
Integrated Storage
30 PB, 4 TB/s

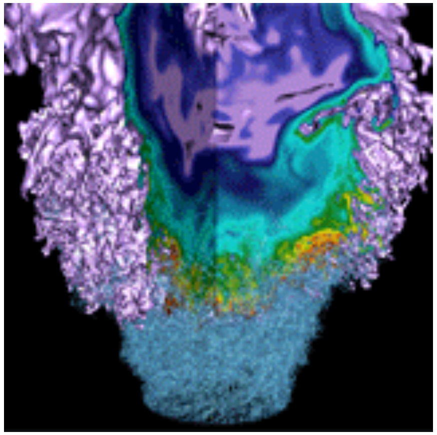
"Slingshot" Interconnect
Ethernet Compatible

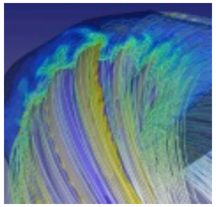
High-Mem
Workflow
Nodes

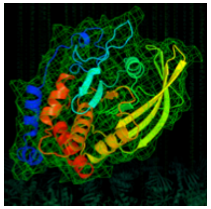
Login Nodes

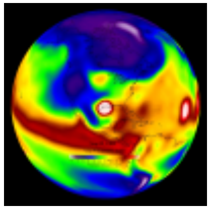
External File-systems &
Networks

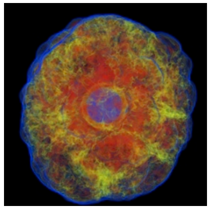
Storage 2020 Strategy

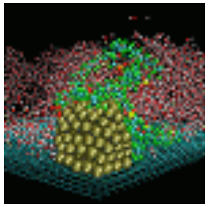




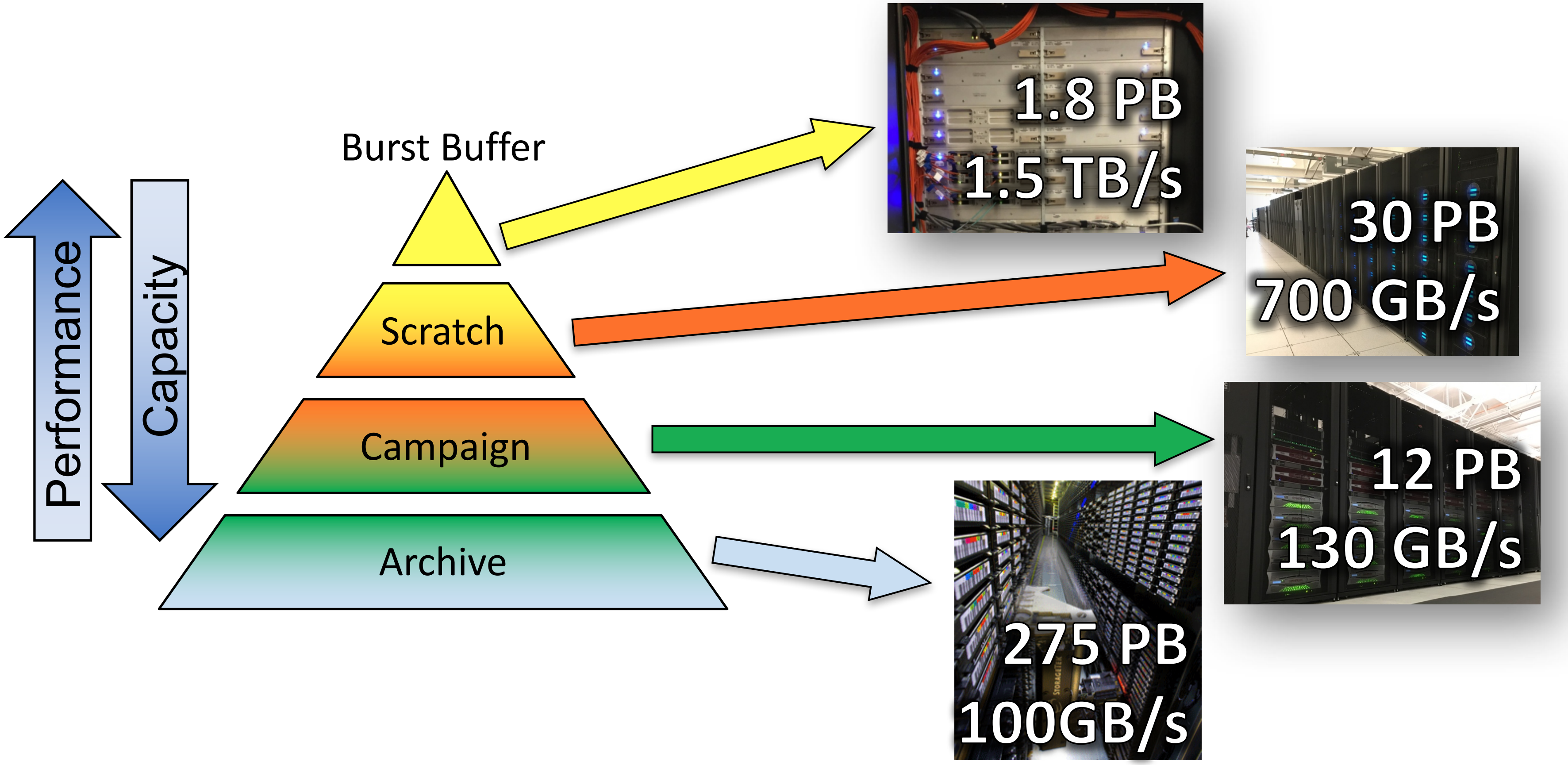




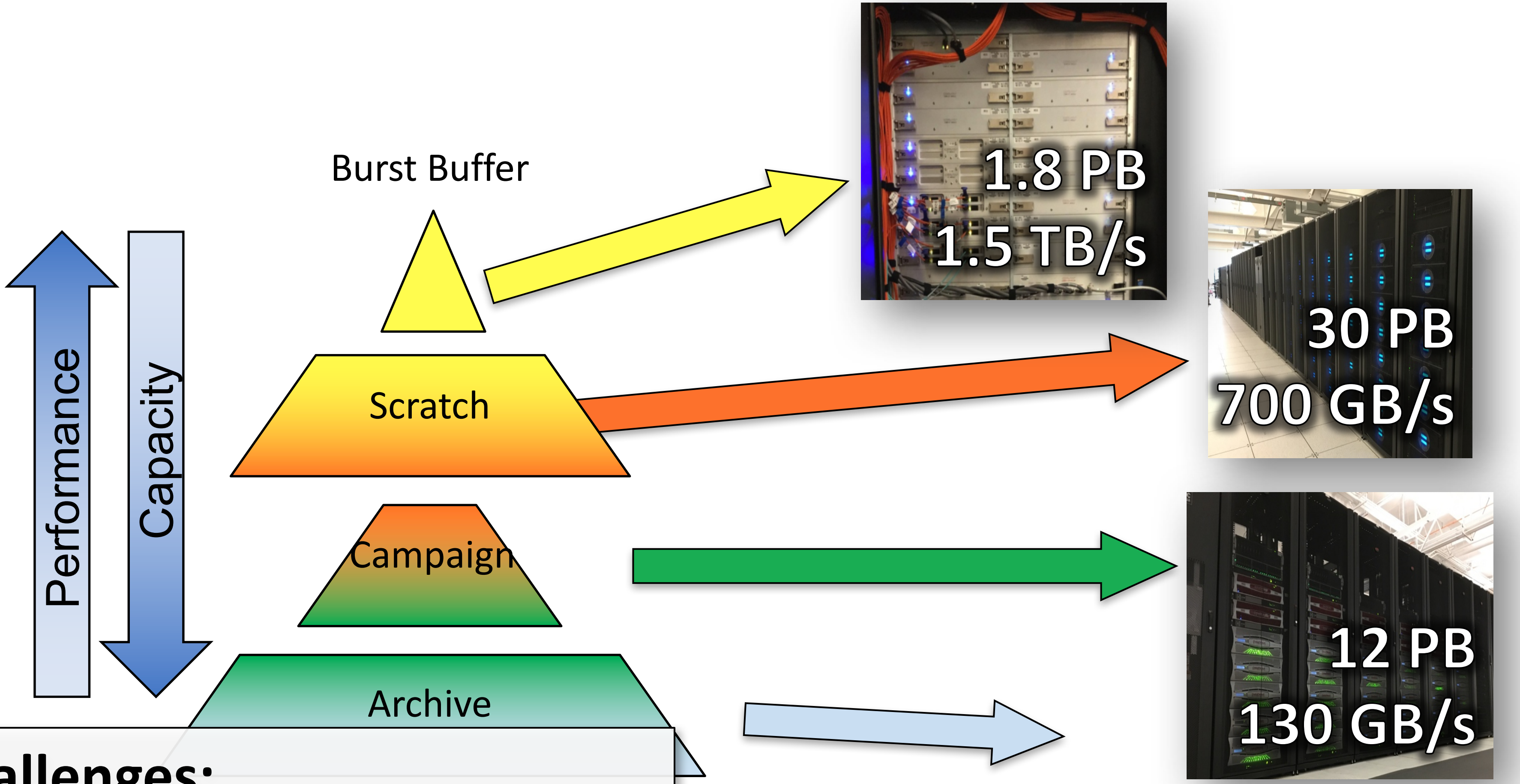




NERSC's storage hierarchy (current)



Beauty in the eye of the …

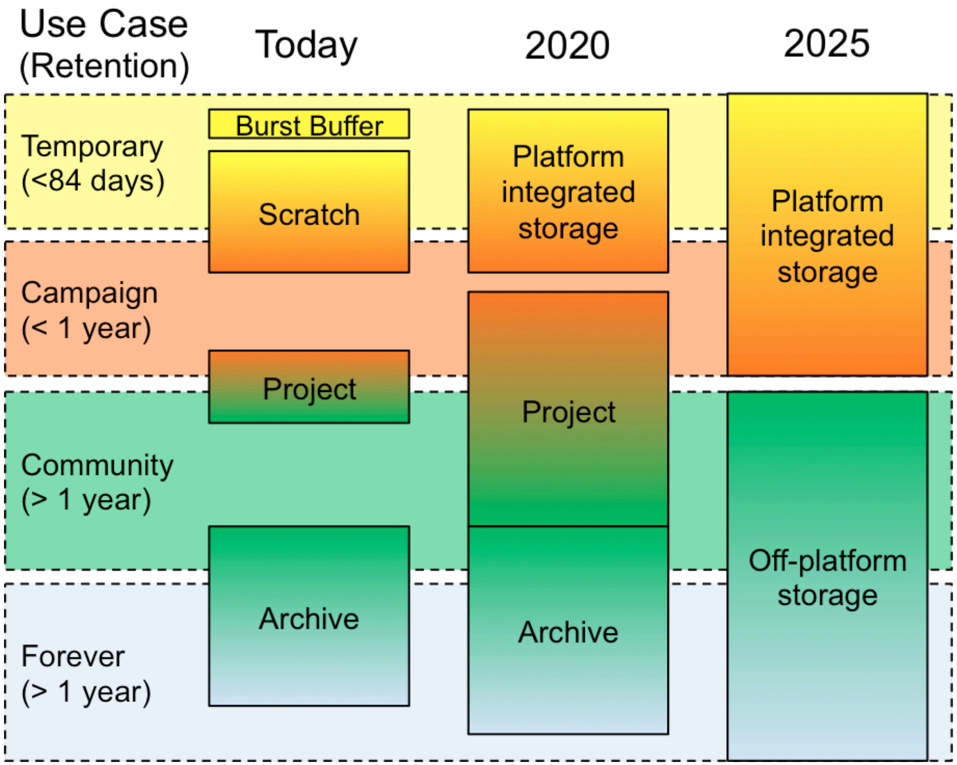


Challenges:

- Inefficient pyramid
- New detectors, experiments
- Exascale == massive data
- HPC tech landscape changing

NERSC Storage 2020: Design goals

- **Target 2020**
  - Collapse burst buffer and scratch into all-flash scratch
  - Invest in large disk tier for capacity
  - Long-term investment in tape to minimize overall costs
- **Target 2025**
  - Use single namespace to manage tiers of SCM and flash for scratch
  - Use single namespace to manage tiers of disk and tape for long-term repository



Storage 2020: A Vision for the Future of HPC Storage https://escholarship.org/uc/item/744479dp

NERSC Storage 2020: Implementation

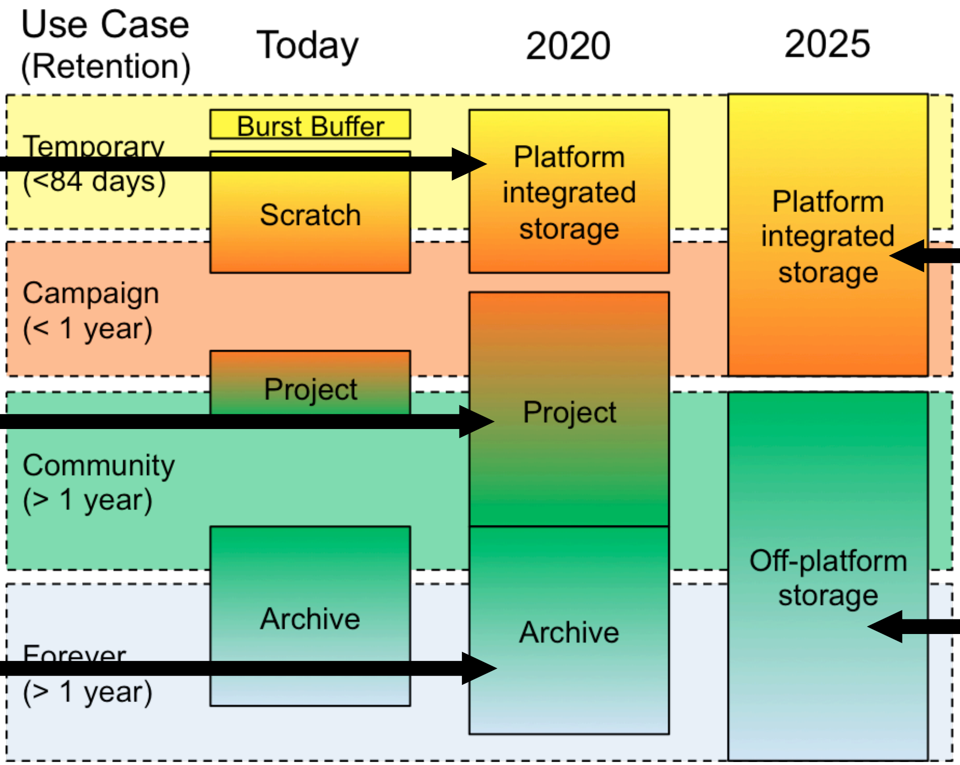
All-flash parallel file system on Perlmutter

Capacity-focused, disk-based file system

600 PB HPSS archive w/

IBM TS4500

+Integrated Cooling

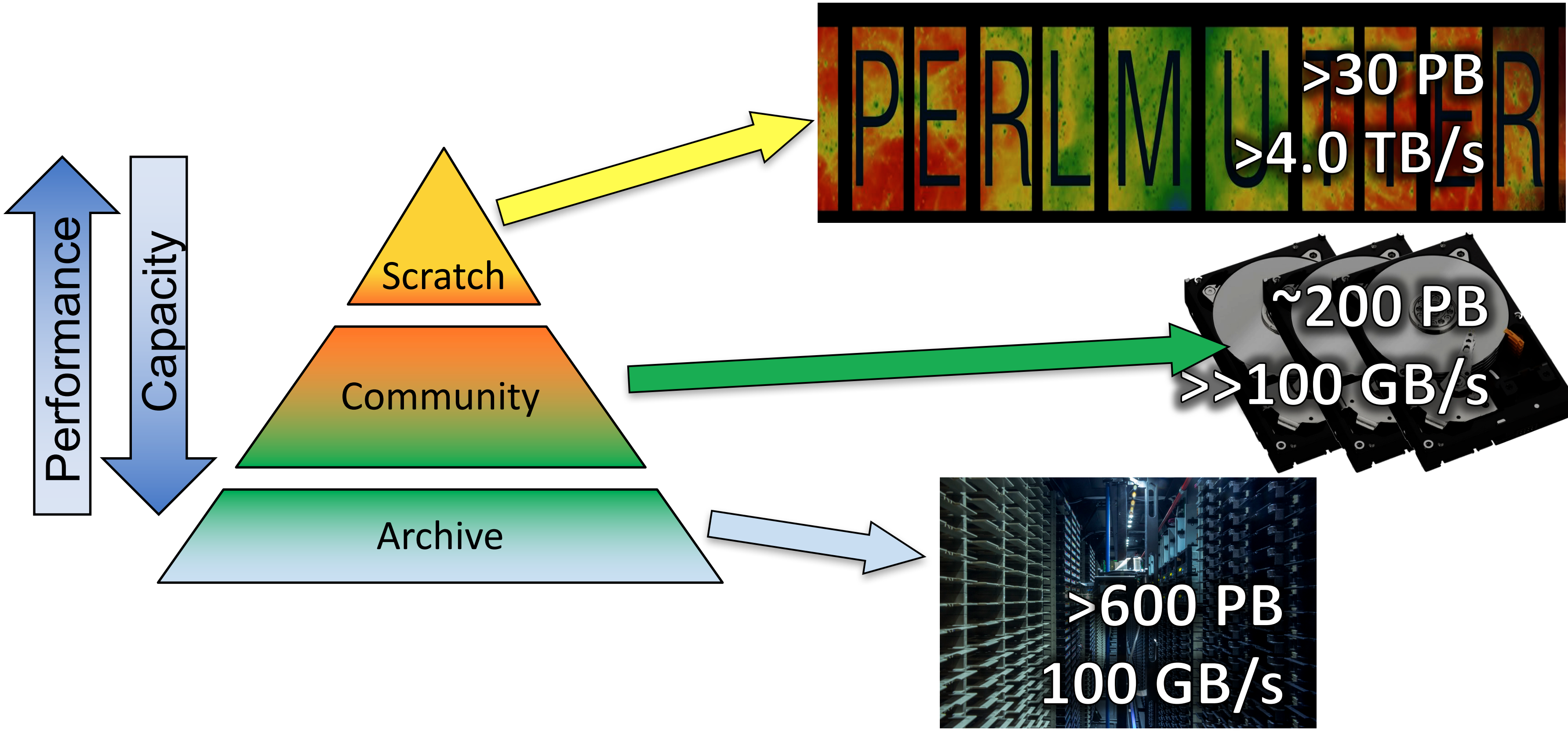


Performance object store w/ SCM+SLC+QLC? on NERSC-10

Archival object store w/ HDD+tape

(GHI+HPSS? Others?)

NERSC’s storage infrastructure (2020)



Performance

Capacity

Scratch

Community

Archive

>30 PB

>4.0 TB/s

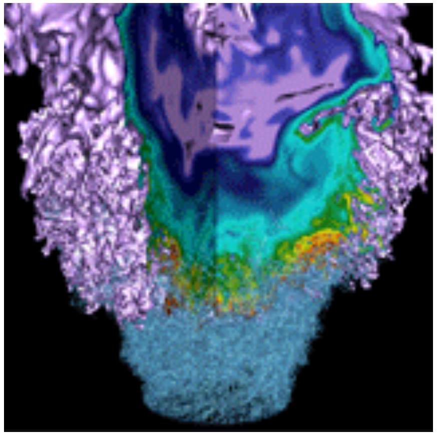
~200 PB

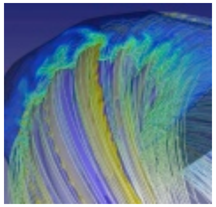
>>100 GB/s

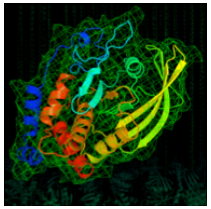
>600 PB

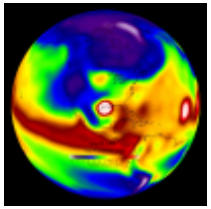
100 GB/s

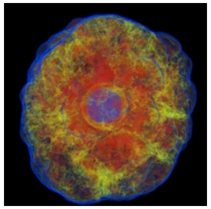
Tape System Migration Update

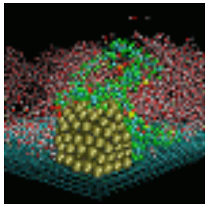






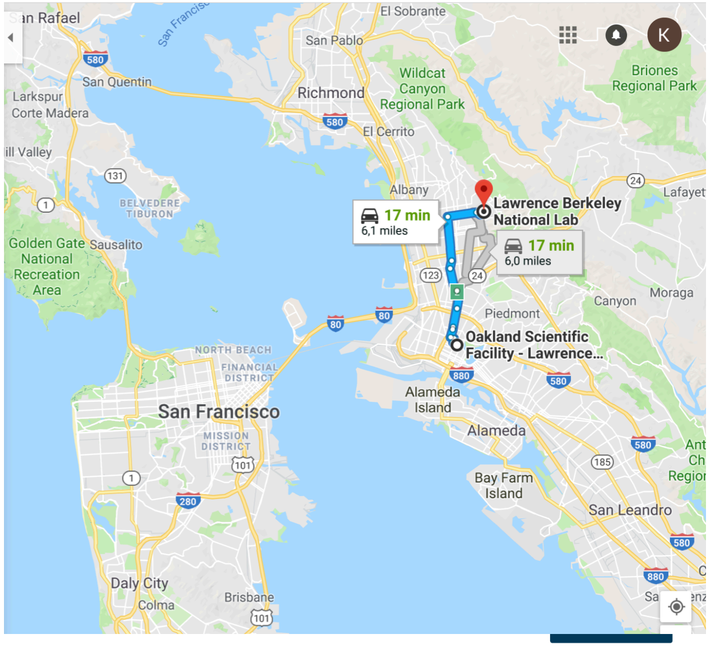






HPSS Archive – Two significant needs

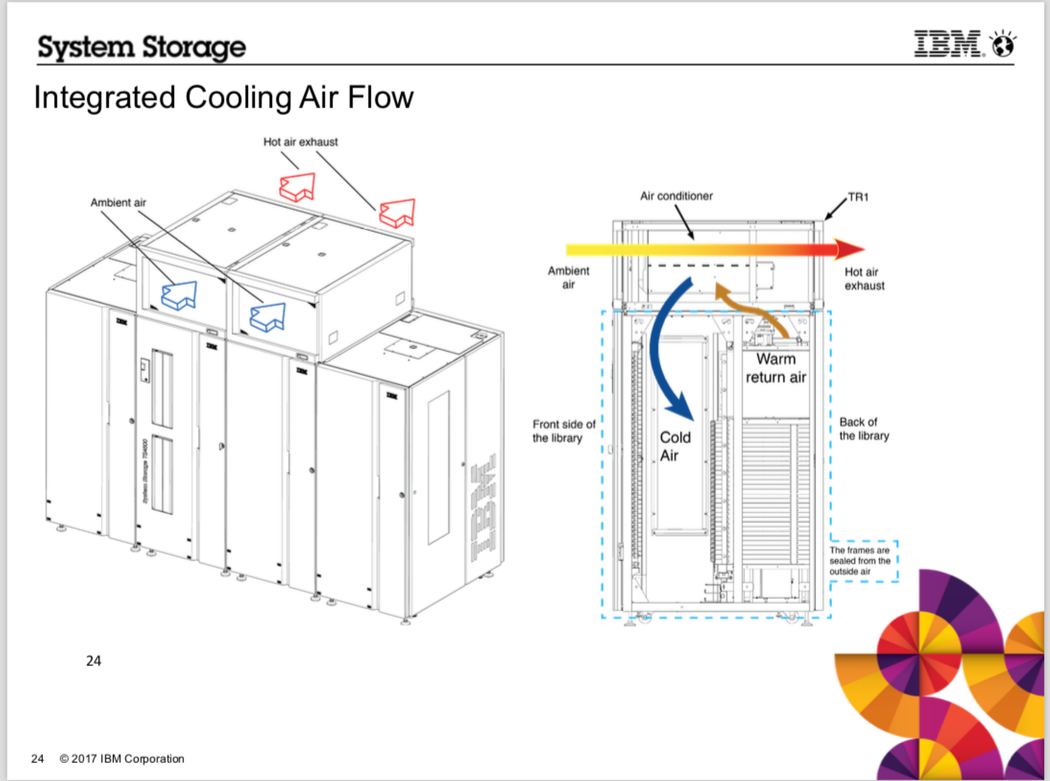
- **Technology decision**
  - Discontinued Oracle Enterprise Tape Drive
    - 4 Fully configured Oracle SL8500 libraries (archive)
    - 60 Oracle T10KC tape drives (archive)
    - 1 IBM TS3500 (mainly system backups)
    - 36 IBM TS1150 tape drives (mainly system backups)
- **Physical move required**
  - Oakland to Berkeley (~6 mi/~9 km)



HPSS Archive – Green Data Center Solution

IBM TS4500 Tape Library *with Integrated Cooling*

- seals off the library from ambient temperature and humidity
- built-in AC units (atop library) keeps tapes and drives within operating spec



System Storage

IBM

Integrated Cooling Air Flow

24

24 © 2017 IBM Corporation

HPSS Archive – Tech Change

- **One "Storage Unit" (my term) [Cooling Zone]**
  - Two S25 frames sandwich, one L25 and one D25 frame
    - S25: High-density frame, tape slots (798-1000)
    - D25: Expansion frame, drive (12-16), tape slots (590-740)
    - L25: Base frame, drive (12-16), tape slots (550-660), I/O station and control electronics (for subsequent libraries no L25)
  - Each one of these storage units considered it's own cooling zone
- **AC units go atop L and D frames**
  - Air recirculated, no special filters
  - Fire suppression a little trickier, but possible

HPSS Archive – Tech Change (CRT)

- **Each library has 4 cooling zones**
  - 16 frames
  - 64 TS1155/3592-55F(FC)/Jag(uar)6 tape drives
  - ~13,000 tape slots
    - JD media @15TB/cartridge
- **We have installed 3 of the above**
- **Thoughts on TS4500 so far**
  - Pro: Integrated cooling and enterprise drives (not LTO)
  - Pro: GUI and CLI are OK but ACSLS (STK) is missed
    - REST API looks promising (testing TBD)
  - Needs work: Some firmware glitches

HPSS Archive – Loose timeline

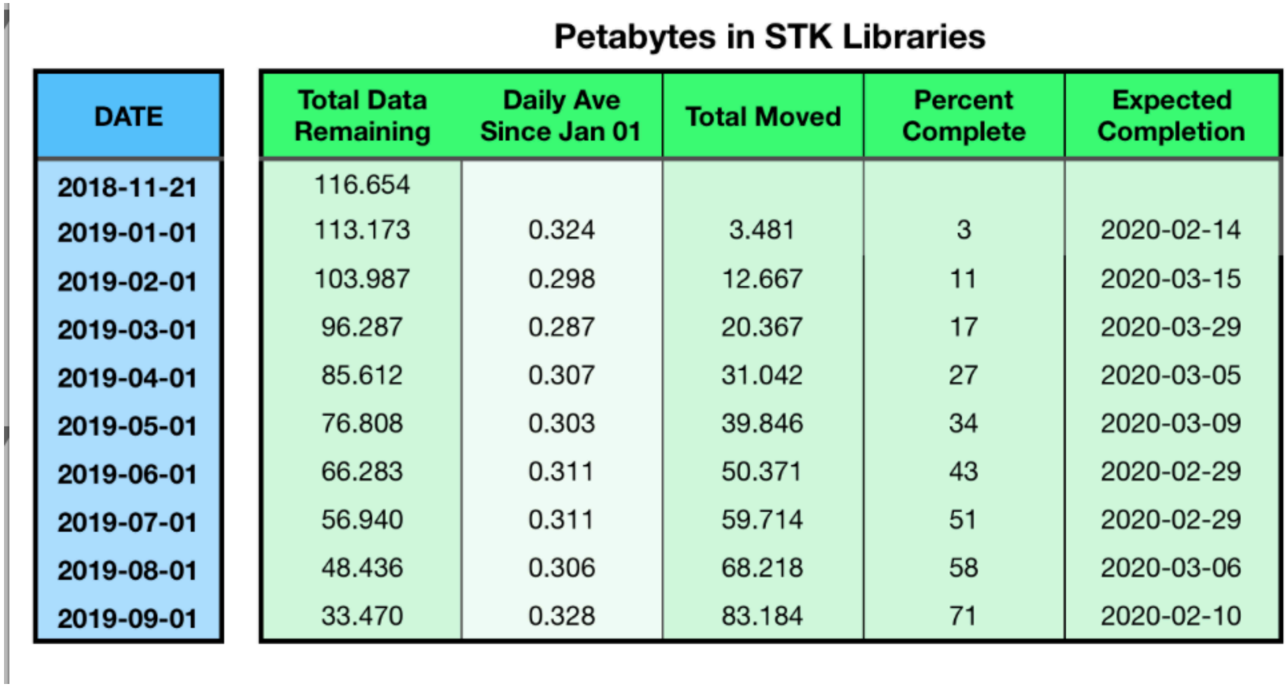
Now:

- **Oakland tapes read-only**
- **Data migrating to BDC via HPSS *repack* functionality**
  - 400Gbps Oakland <-> BDC link
  - >400 TB/day from OSF to CRT (Oracle → IBMmedia)
  - Sneakernet: 30PB IBM media moved out of OSF by truck
- **2020 (or earlier, see next slide) data migration complete**

HPSS Archive – Status as of Sept. 2019

- Data migration stepped up
  - New goal bulk of data moved by Q1 2020
- Tape volumes processed chronologically
  - Later files are larger, better streaming from tape drives, better data rates.
- Smaller data
  - expect higher error rates on this data
  - More labor intensive

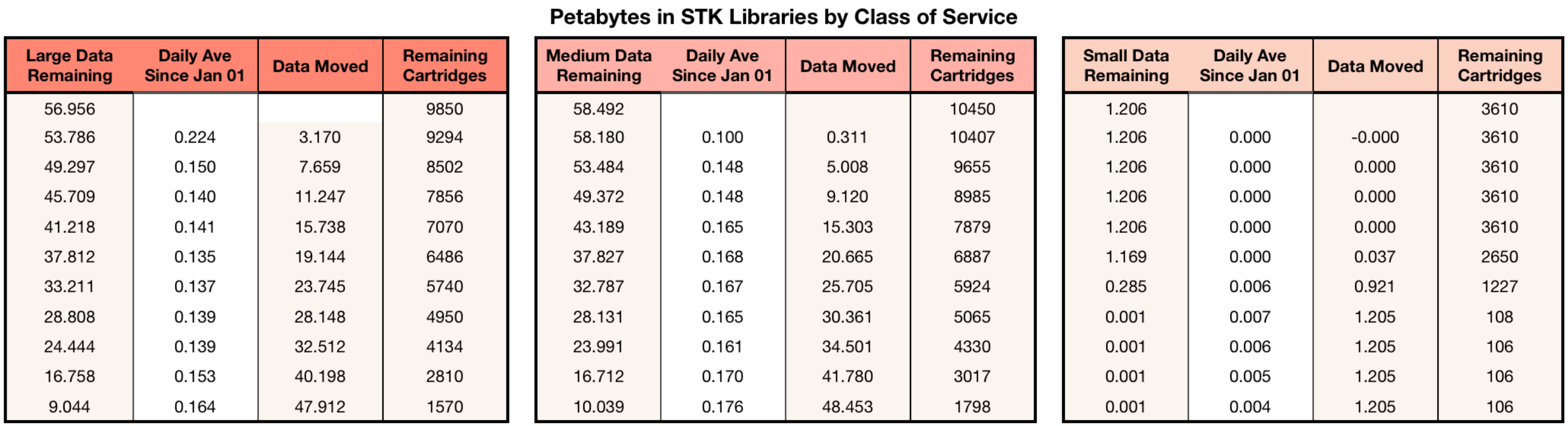
HPSS Archive – Status as of Sept. 2019



Petabytes in STK Libraries

DATE	Total Data Remaining	Daily Ave Since Jan 01	Total Moved	Percent Complete	Expected Completion
2018-11-21	116.654				
2019-01-01	113.173	0.324	3.481	3	2020-02-14
2019-02-01	103.987	0.298	12.667	11	2020-03-15
2019-03-01	96.287	0.287	20.367	17	2020-03-29
2019-04-01	85.612	0.307	31.042	27	2020-03-05
2019-05-01	76.808	0.303	39.846	34	2020-03-09
2019-06-01	66.283	0.311	50.371	43	2020-02-29
2019-07-01	56.940	0.311	59.714	51	2020-02-29
2019-08-01	48.436	0.306	68.218	58	2020-03-06
2019-09-01	33.470	0.328	83.184	71	2020-02-10

HPSS Archive – Status as of Sept. 2019



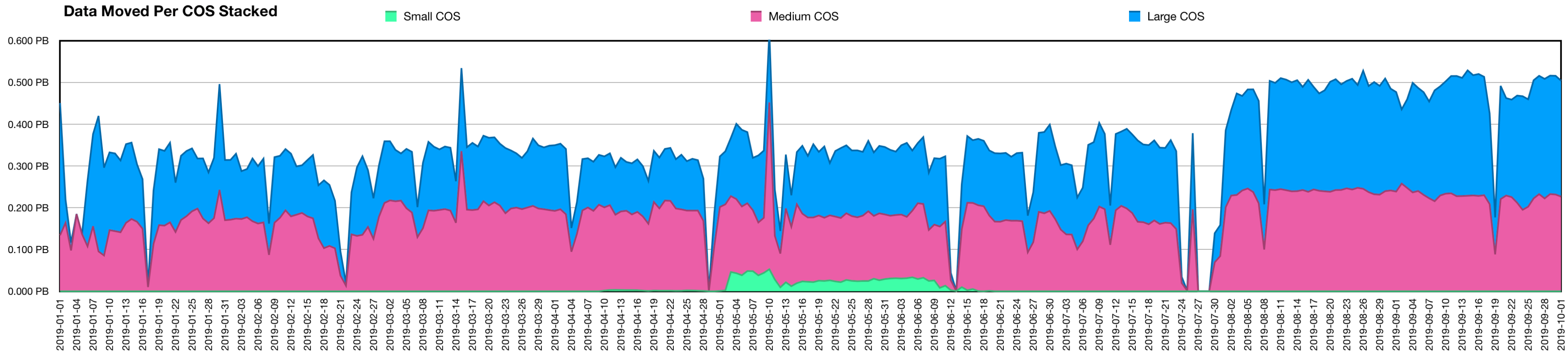
Petabytes in STK Libraries by Class of Service

Large Data Remaining	Daily Ave Since Jan 01	Data Moved	Remaining Cartridges
56.956			9850
53.786	0.224	3.170	9294
49.297	0.150	7.659	8502
45.709	0.140	11.247	7856
41.218	0.141	15.738	7070
37.812	0.135	19.144	6486
33.211	0.137	23.745	5740
28.808	0.139	28.148	4950
24.444	0.139	32.512	4134
16.758	0.153	40.198	2810
9.044	0.164	47.912	1570

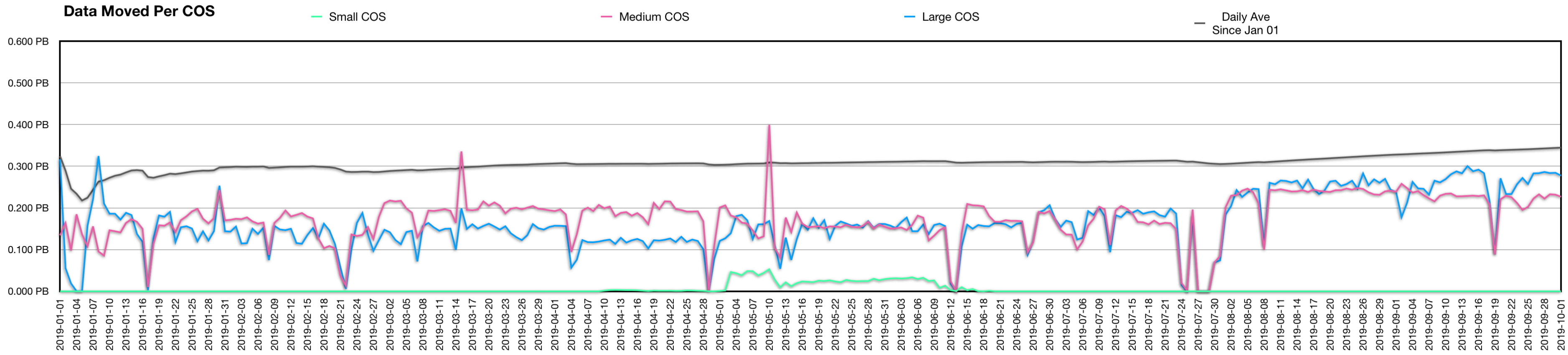
Medium Data Remaining	Daily Ave Since Jan 01	Data Moved	Remaining Cartridges
58.492			10450
58.180	0.100	0.311	10407
53.484	0.148	5.008	9655
49.372	0.148	9.120	8985
43.189	0.165	15.303	7879
37.827	0.168	20.665	6887
32.787	0.167	25.705	5924
28.131	0.165	30.361	5065
23.991	0.161	34.501	4330
16.712	0.170	41.780	3017
10.039	0.176	48.453	1798

Small Data Remaining	Daily Ave Since Jan 01	Data Moved	Remaining Cartridges
1.206			3610
1.206	0.000	-0.000	3610
1.206	0.000	0.000	3610
1.206	0.000	0.000	3610
1.206	0.000	0.000	3610
1.169	0.000	0.037	2650
0.285	0.006	0.921	1227
0.001	0.007	1.205	108
0.001	0.006	1.205	106
0.001	0.005	1.205	106
0.001	0.004	1.205	106

HPSS Archive – Status as of Sept. 2019

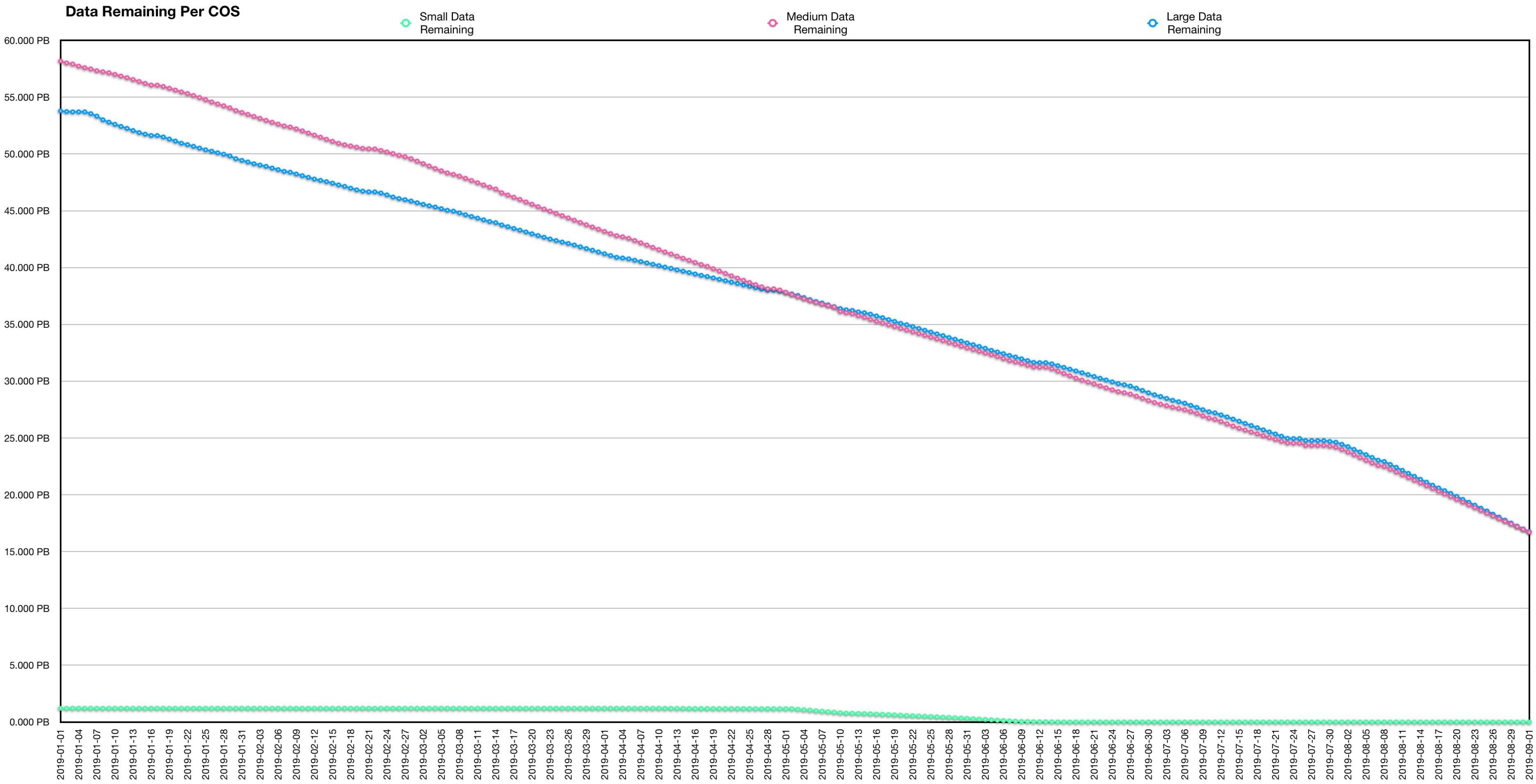


Data Moved Per COS Stacked

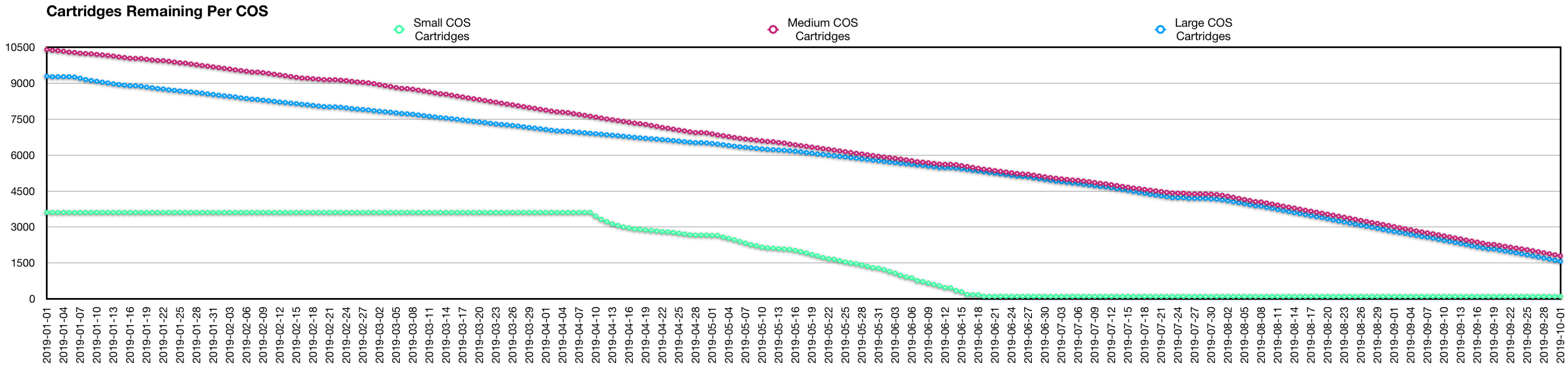


Data Moved Per COS

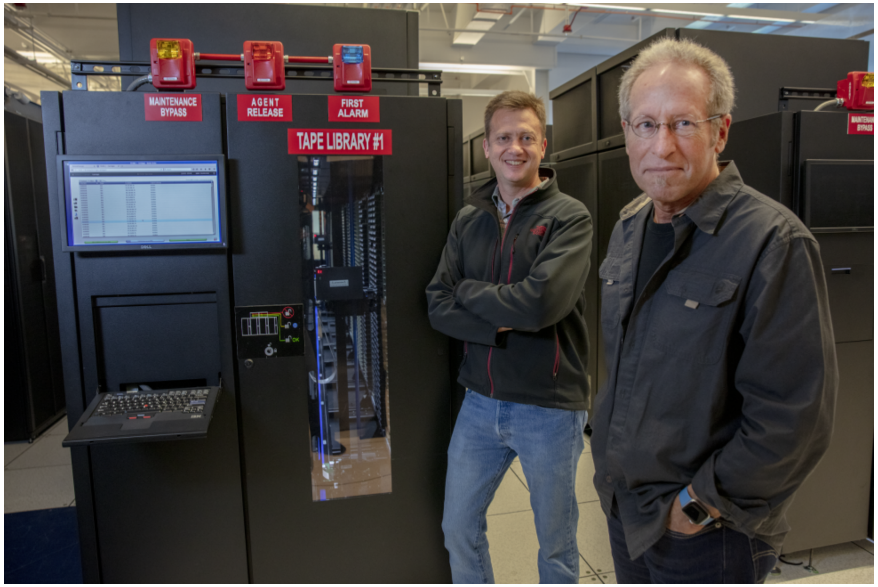
HPSS Archive – Status as of Sept. 2019



HPSS Archive – Status as of Sept. 2019



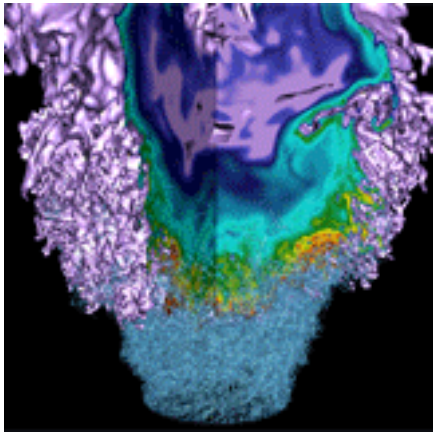
New Tape Libraries at Berkeley

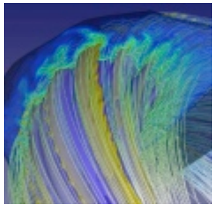


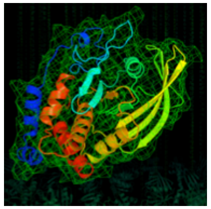
Nice article in HPCWire: https://bit.ly/2OwX24N

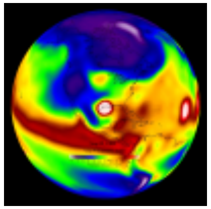
Not an “S”, also not an “S”

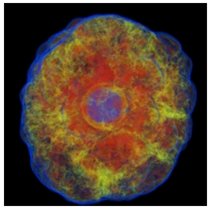
GPFS-HPSS-Integration (GHI)

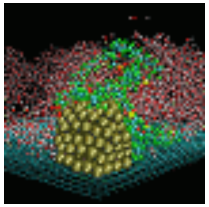












GHI – JKL about it? It is:

- **Optional piece of HPSS**
  - connects Spectrum Scale/GPFS and HPSS
  - automated data movement between the two
- **GHI primary functions:**
  - Space management (current focus)
    - Migrate
    - Purge
    - Recall
  - Disaster recovery (maybe later)
    - Backup
    - Restore

GHI – Functionality

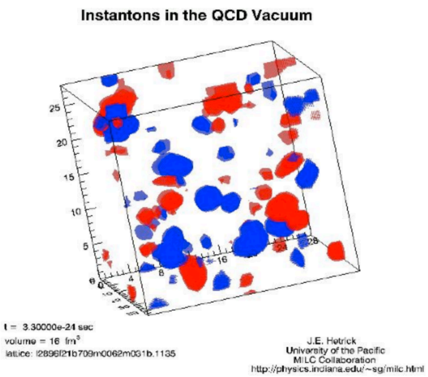
- **GPFS HSM space management / file migrations**
  - GPFS Data Management API (DMAPI) notifies GHI of events
  - HPSS references are stored as GPFS extended attributes
  - GPFS ILM scans and policies
    - ILM scans billions of files in minutes
    - Files are continuously identified and migrated/purged/recalled to/from HPSS per policy
  - If GPFS reaches a space threshold, candidates are purged (stubbed out)
  - When a user requests a file in HPSS, GHI stages it back
  - Small files are aggregated with a tar-like utility to improve performance
  - Policy rules provide robust data management solutions
  - GHI uses the HPSS Parallel I/O (PIO) for parallel access to files stored in HPSS

GHI Use Case: GHI Use Case: ALS

- Advanced Light Source: Beamline of X-ray light used to examine the atomic and electronic structure of matter
- Data from the beamline streams to NERSC, gets analyzed, and a copy gets put into HPSS, beamline users download their data via Globus Sharing: 400TB on spinning disk, 3 PB in HPSS
- *Want to use GHI to automatically store in HPSS while still maintaining their directory structure and to free up space on spinning disk for active analysis*

GHI Use Case: QCD Library

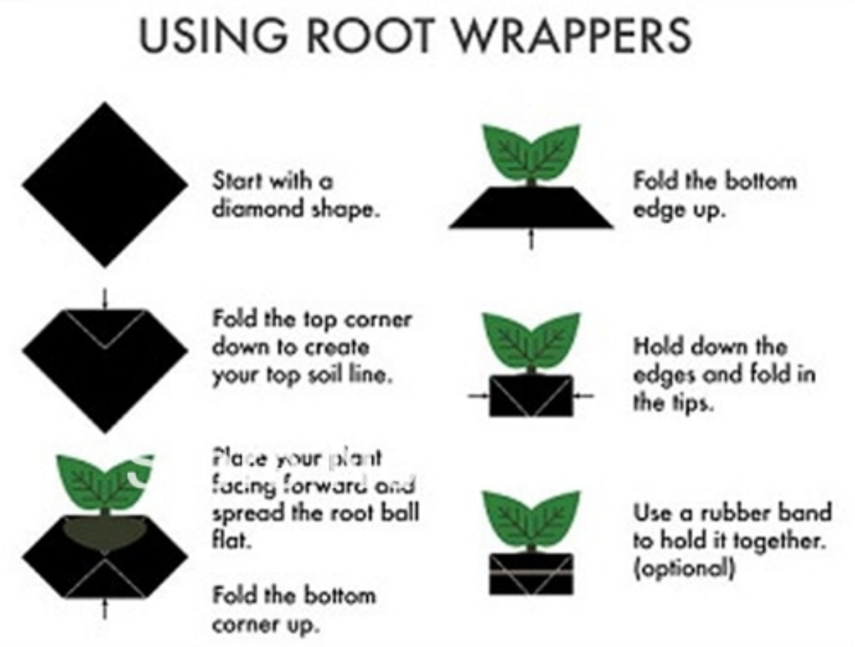
- Collecting QCD simulation data and serving it to scientists (along with descriptive metadata). Currently serves data out of HPSS via FTP, which limits the size of datasets they can serve
- *GHI will let them store TB-size datasets in HPSS and serve them out via Globus Sharing*
  - For large datasets, the time to stage a file is offset by the speedup offered by Globus


Instantons in the QCD Vacuum
25
20
15
10
5
0
4
8
16
24
t = 3.30000e-24 sec
volume = 16 fm³
lattice: l2896f21b709m0062m031b.1135
J.E. Hetrick
University of the Pacific
MILC Collaboration
http://physics.indiana.edu/~sg/milc.html

GHI – NERSC implementation tweaks

- Wrapper scripts for *user* access to ghi operations
- NERSC client systems can only access GPFS systems via remote cluster mounts.
  - So, user access is only via remote cluster mounts

- As it works today, GHI commands are only available on GHI-enabled ***owning*** clusters
  - ***automatic retrieval*** on open available and works on ***remote*** clusters
  - With few exceptions, GHI commands must be run by root.
    -- so… no file access validation.
- Root wrappers to the rescue!


USING ROOT WRAPPERS
Start with a diamond shape.
Fold the top corner down to create your top soil line.
Place your plant facing forward and spread the root ball flat.
Fold the bottom corner up.
Fold the bottom edge up.
Hold down the edges and fold in the tips.
Use a rubber band to hold it together. (optional)

GHI – NERSC implementation tweaks

- 5 GHI command wrappers under development
  - can be run by users on remote clusters
  - run as the user and validate user access and operation permission
  - communicate via sockets to proxy running on the GHI owning cluster
  - validated files and operations are passed to the proxy for execution.

1. ghi_ls: for ghi_ls to list files
2. ghi_pin: for the ghi_pin command
3. ghi_put: for a policy engine run to migrate file data to HPSS
4. ghi_punch: for a policy engine run to punch holes in files
5. ghi_stage: for ghi_stage to retrieve file data from HPSS

GHI – Wish List/Issues

- **GHI understanding of user access permissions & GHI commands to work with remote clusters.**
  - – want ghi activities to be user driven and not administrator driven.
  - – would do a way with the need for wrappers
- **When too few files are selected to form an aggregate**
  - – they just get dropped and left in limbo
- **Ability of htar/ishtar to process encoded characters**
  - – users are ingenious in their ability to generate mangled directory and file names

Other Projects – Future Talks?

GPFS HPSS Integration (GHI) – Where are we now?

Data Migration - Data Migration/Orchestration will be important with Perlmutter as data flows between flash, disk and tape tiers.

NERSC Storage Team & Fellow Contributors

Right to Left:

Greg Butler

Kirill Lozinskiy

Nick Balthaser

Ravi Cheema

Damian Hazen *(Group Lead)*

Rei Lee

Kristy Kallback-Rose

Wayne Hurlbert

+ Melinda Jacobsen

(recently joined the team)

Thank you. Questions?

NERSC

National Energy Research Scientific Computing Center

Tape Archive - HPSS

- **High Performance Storage System (HPSS)**
  - Developed over >20 years of collaboration among five Department of Energy laboratories and IBM, with significant contributions by universities and other laboratories worldwide.
  - archival storage system for long term data retention since 1998
  - Tiered storage system with a disk cache in front of a pool of tapes
    - On tape: ~140PB PB
    - Disk Cache: 4PB
  - Contains 40 years of data archived by the scientific community
- **Data Transfers via transfer client - there is no direct file system interface**
  - We provide numerous clients: HSI/HTAR (proprietary tools), FTP, pFTP, gridFTP, Globus Online, etc. [VFS is an option which we don't use]