NERSC

National Energy Research Scientific Computing Center

Ravi Cheema

Kristy Kallback-Rose

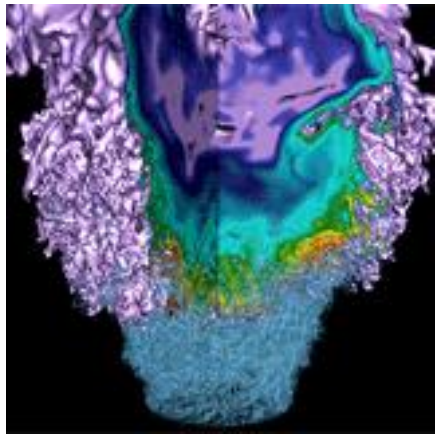
NERSC Storage Systems Team

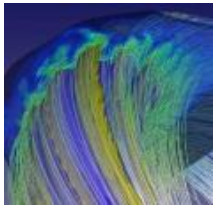
September 24, 2019

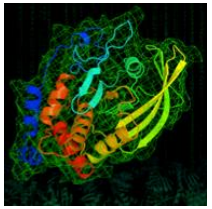
Agenda

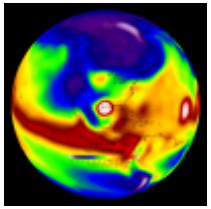
- **General NERSC & Systems Overview**
- **Storage 2020 Strategy & Progress**
- **GHI Testing**
- **Tape Library Update**
- **Futures**

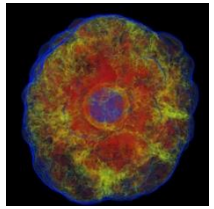
NERSC & Systems Overview

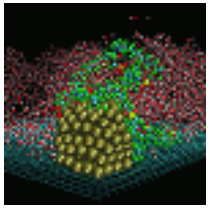






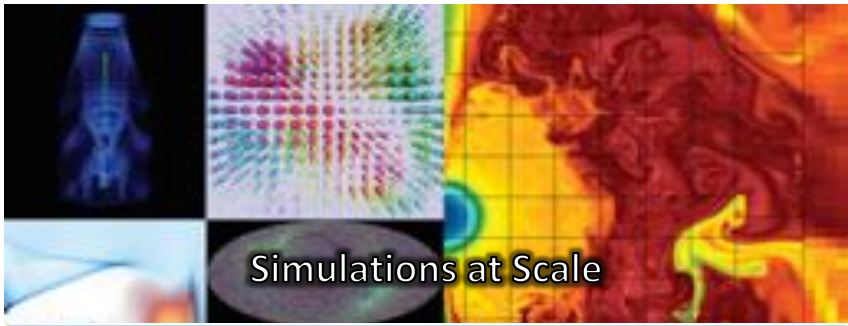




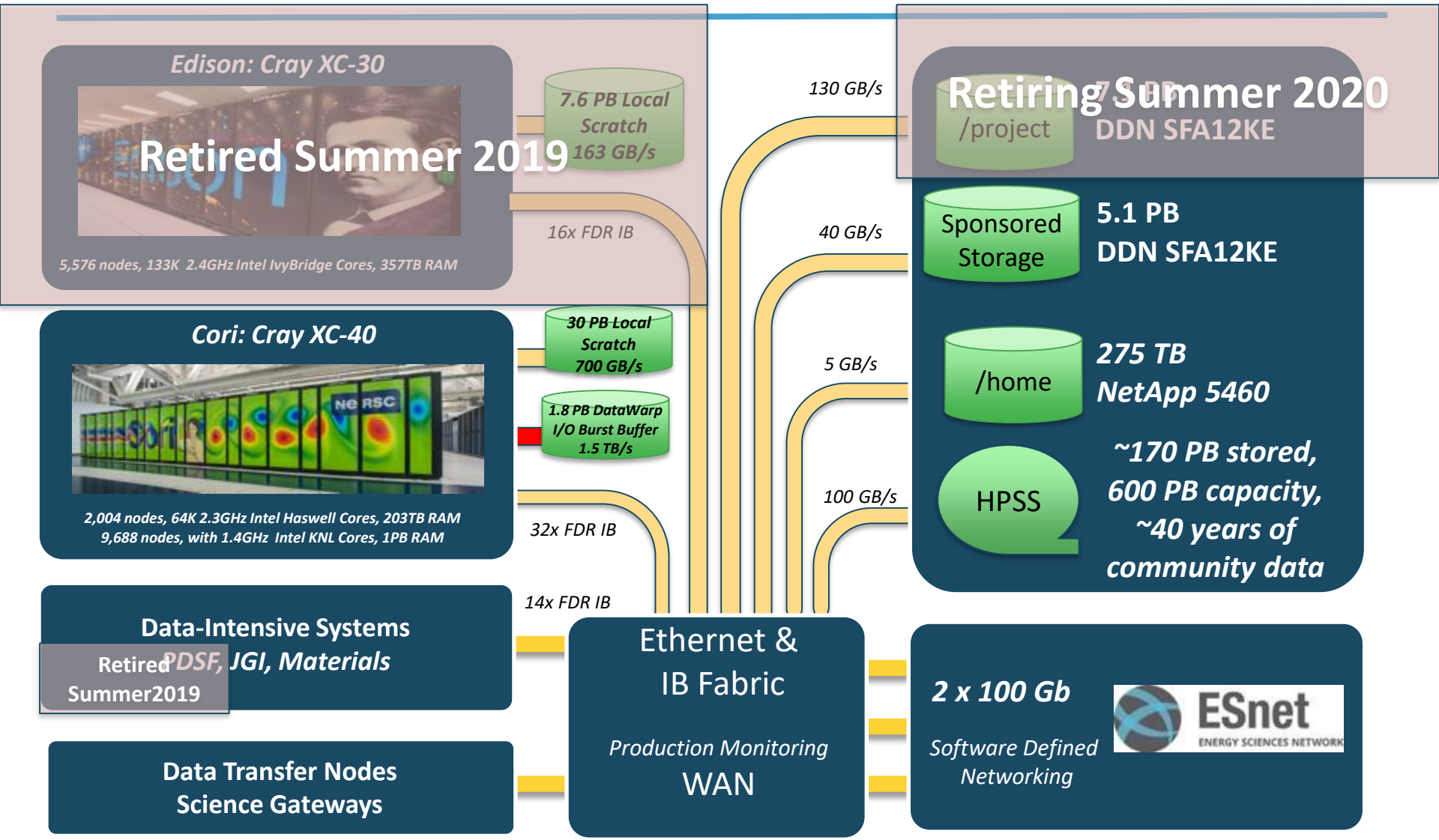


NERSC is the mission HPC computing center for the DOE Office of Science

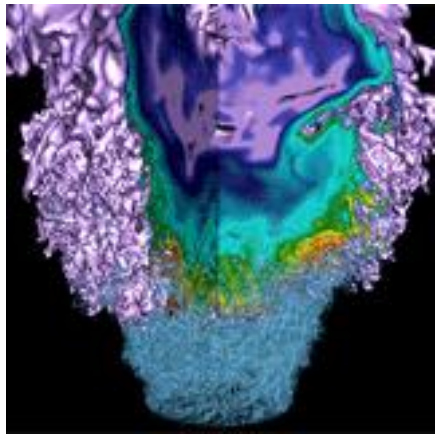
- **HPC and data systems for the broad Office of Science community**
- **Approximately 7,000 users and 870 projects**
- **Diverse workload type and size**
  - Biology, Environment, Materials, Chemistry, Geophysics, Nuclear Physics, Fusion Energy, Plasma Physics, Computing Research

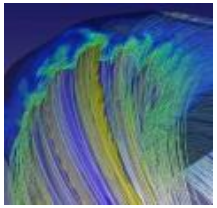


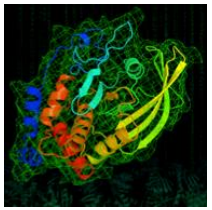
NERSC - Resources at a Glance 2019

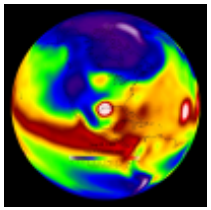


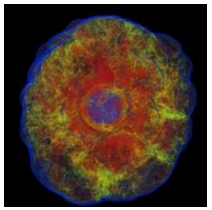
Storage 2020 Strategy

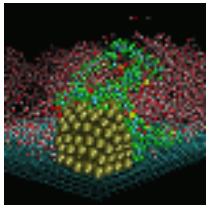




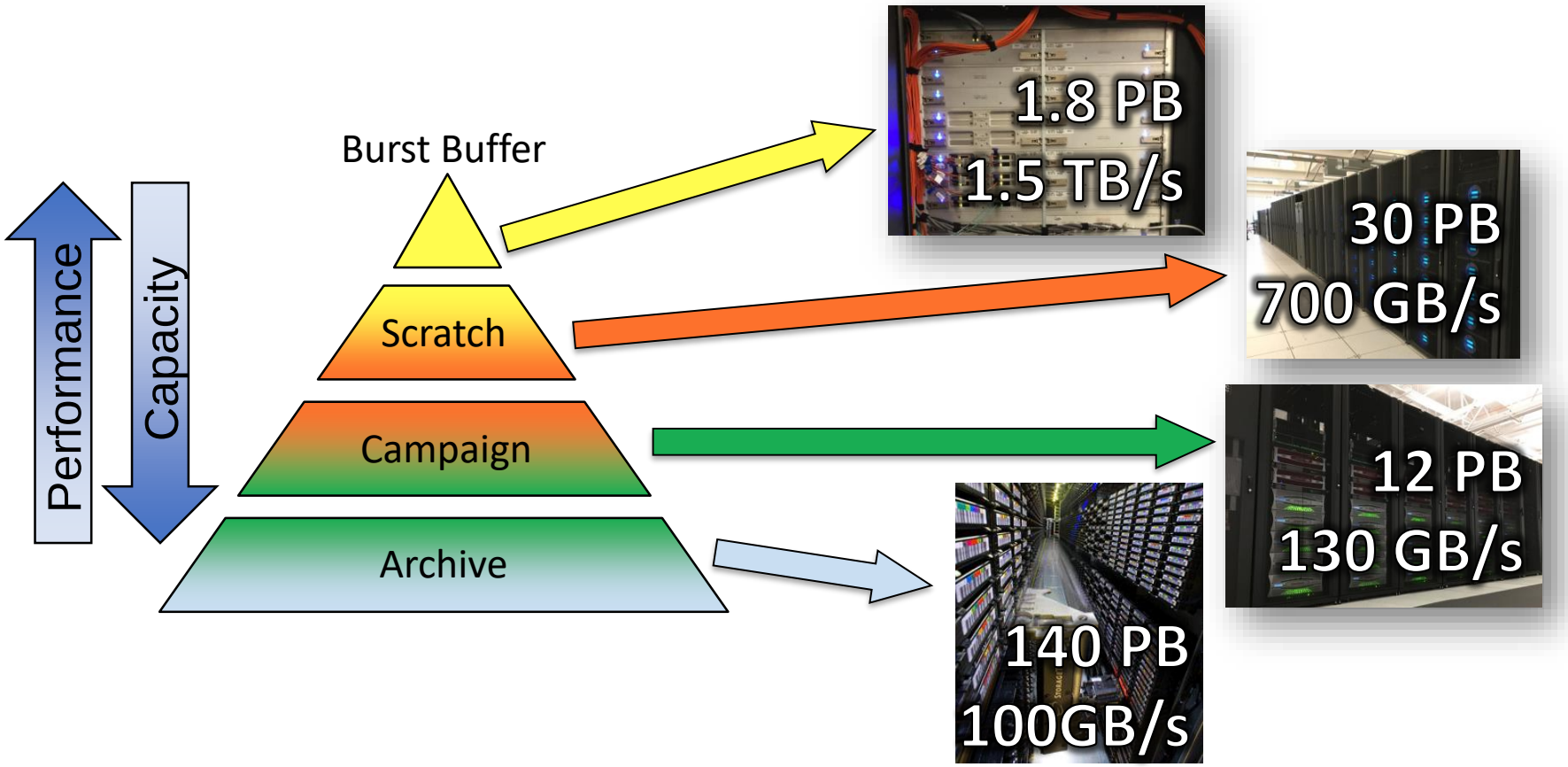






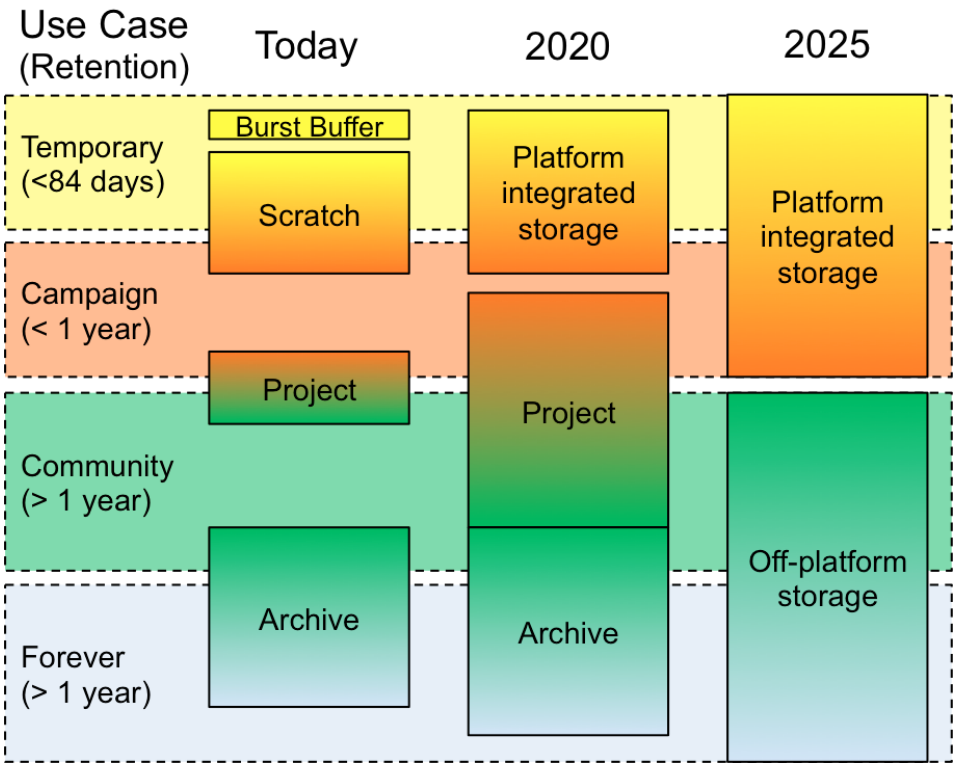


NERSC's storage hierarchy (Current)



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

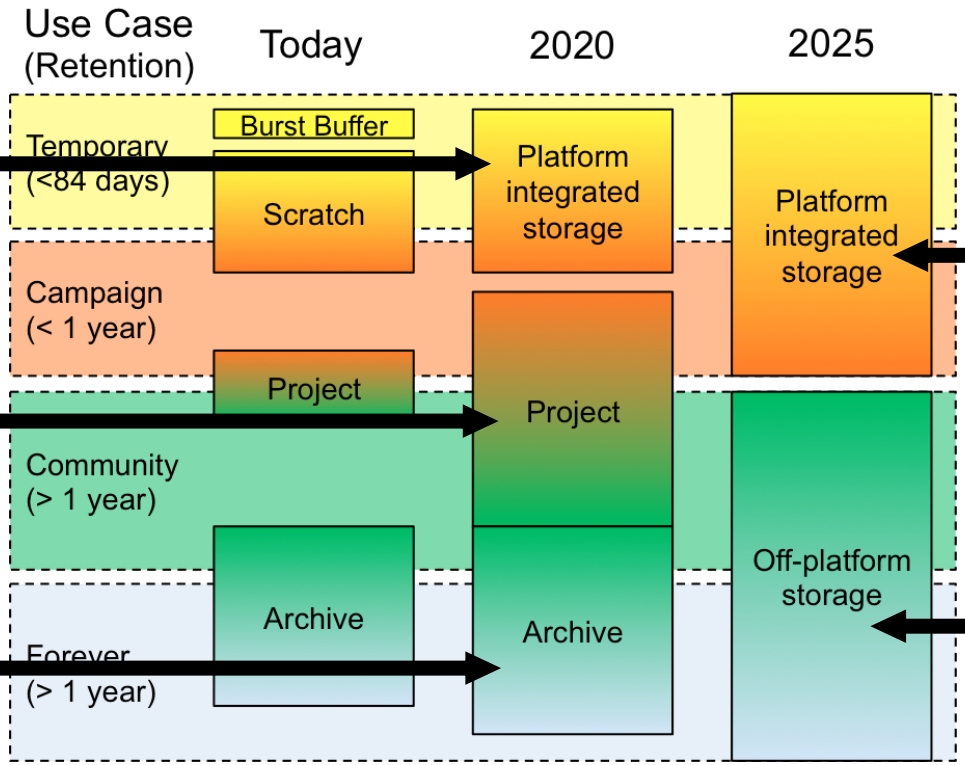


Storage 2020: A Vision for the Future of HPC Storage https://escholarship.org/uc/item/744479dp

NERSC Storage 2020: Implementation

All-flash parallel file system on Perlmutter

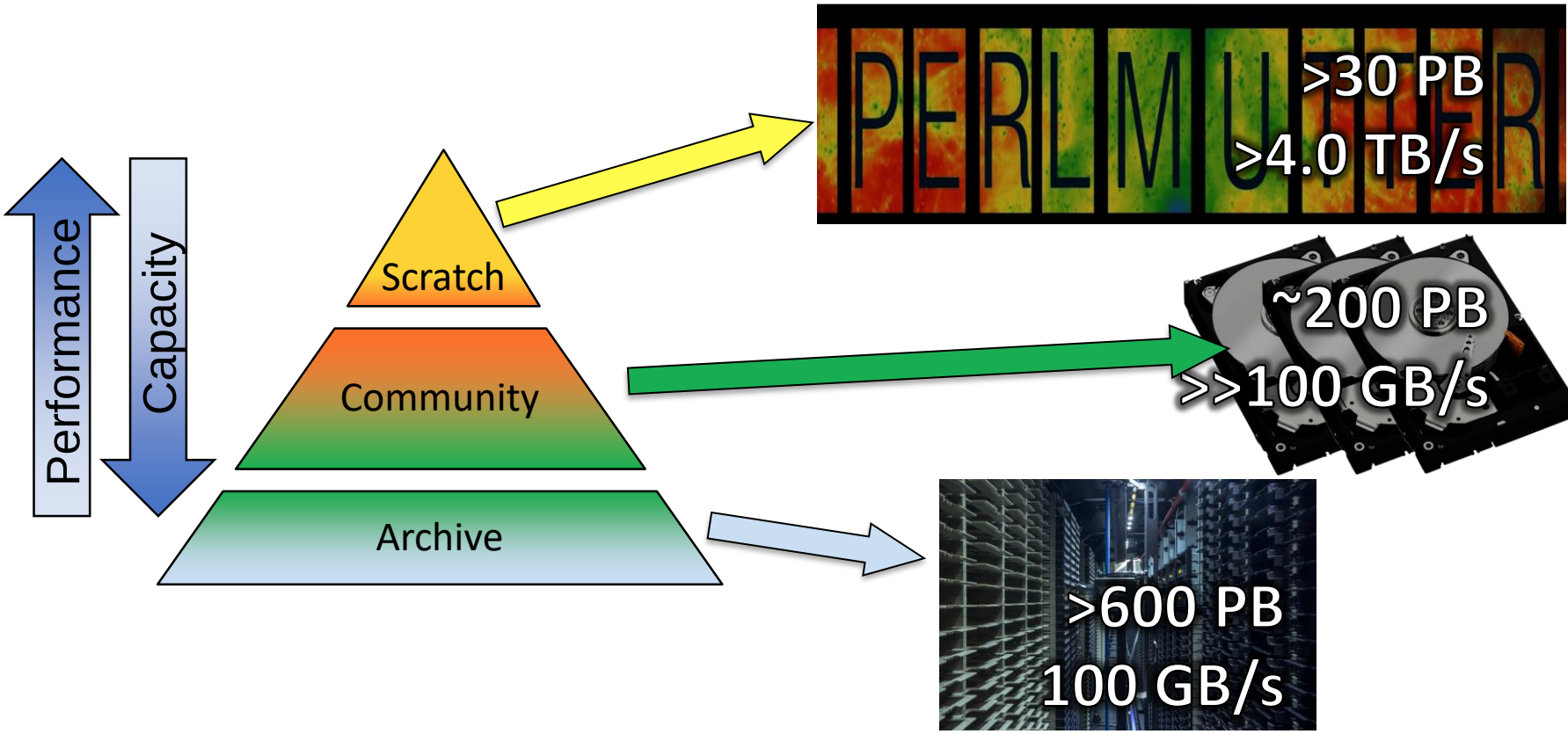
Capacity-focused, disk-based file system
600 PB HPSS archive w/ IBM TS4500
+Integrated Cooling



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

Archival object store w/ HDD+tape (GHI+HPSS? Versity? Others?)

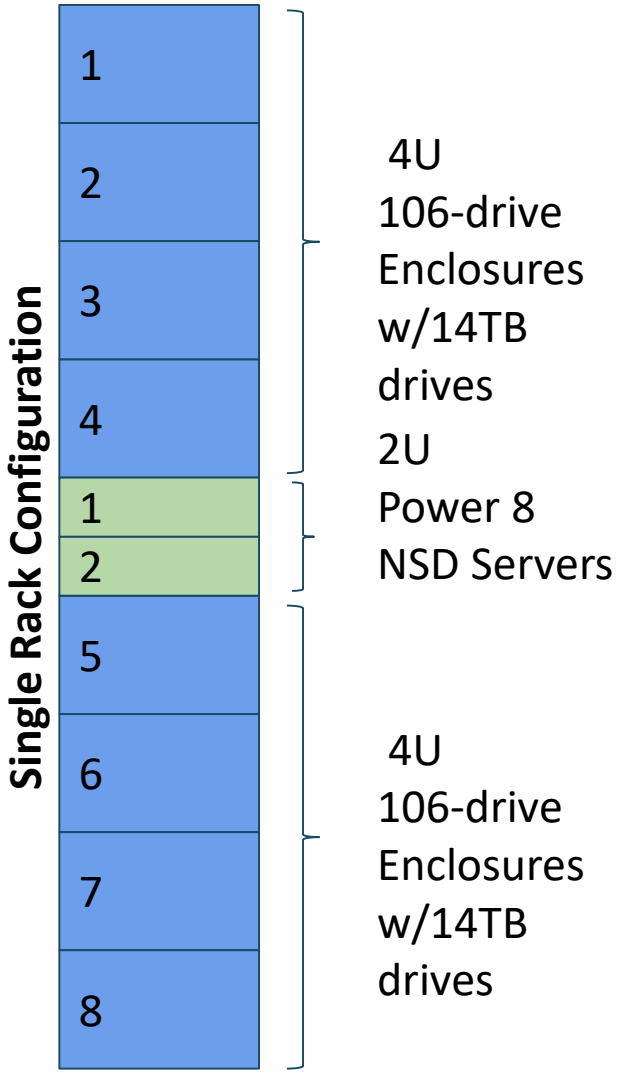
NERSC's storage infrastructure (2020)



Community Filesystem

- **Selected Solution**
  - Technology Details
  - Setup & Configuration
- **Implementation Status**
- **Data Migration Plans**

ESS GL8c Rack Layout



- **Spectrum Scale (aka GPFS): same as /project**
- **57 PB usable space, aggregate bandwidth of 100 GB/s**
- **Denser system ~10PB per rack (project was 1PB per rack), leaves room for expansion, more power efficient**

Production System:

- 7 Racks
  - 11.8 PB per rack
- + ESS Management System (1)

Test System:

- 2U NSD Servers (2)
- 4U Enclosures (2)
- ESS Management System (1)

Deployment Timeline

- Process started in late 2017, team led by Kristy w/other NERSC teams
  - Initial meetings with vendors, and Q&A
- RFP released to vendors September, 2018
  - Vendor responses received and evaluated by team
- IBM proposal selected December, 2018
- Contract negotiations completed July 16, 2019
- First equipment delivered to NERSC July 17, 2019
  - Assembly and vendor testing
- Benchmarking tests August 26, 2019
  - 50 nodes for IOR and MDT tests
- Second round of benchmarks 9/10 - 9/12
- Once benchmark tests pass, one week availability period begins
  - Staff testing

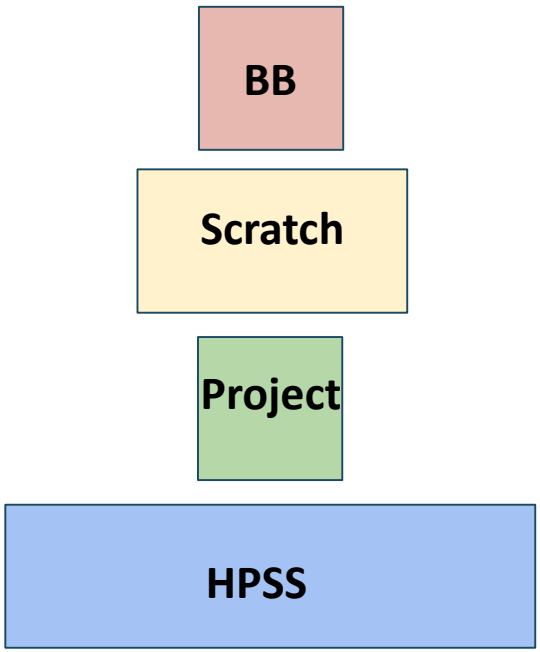
Challenges

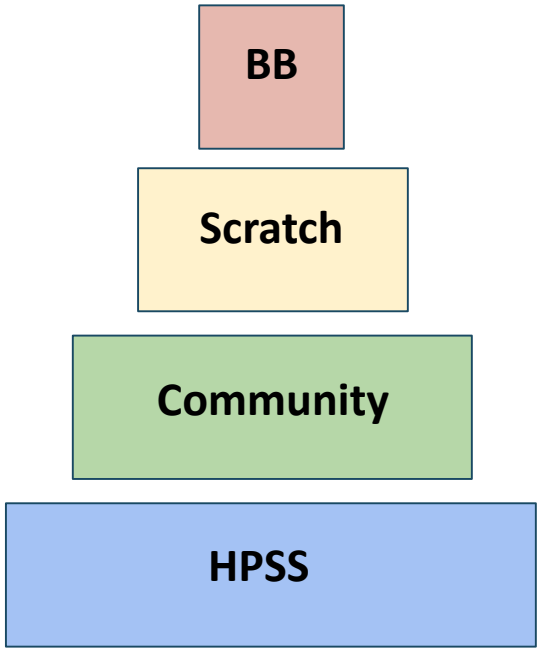
- **Hung process on GPFS client while running IOR/mdtest**
  - running IOR/mdtest using a common directory but on multiple remote clusters, hangs commands like ls in that directory due to unanswered token revoke requests from a node in another cluster, only way out of this was to expel that node or reboot. Fixed in 5.0.3.3
- **IB/network connectivity**
  - Used nsdperf for intial testing of bandwidth between 50 client nodes (1-HCA/node) and 14 ESS I/O nodes (4-HCA/node). Decent performance on one-to-one tests but unable to scale, Continuing to troubleshoot with IBM and mellanox.
- **Rack design**
  - ESS requires a deeper than the a 48 U deep rack. Dimensions and weight of 48U deep rack is seismic certified for NERSC datacenter, certifying any other rack would've required months of design testing/review. Resolved by adding rack extensions to the back of the current rack to accommodate for extended cable arms.

Data Migration

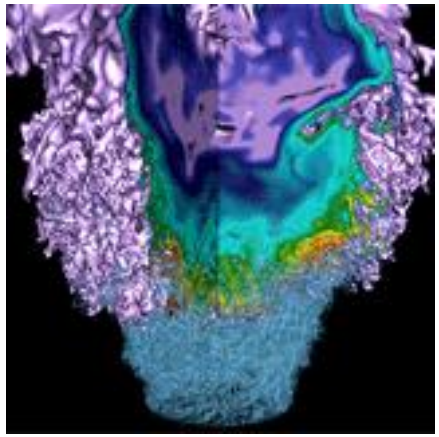
- Chose not to expand the current file system in favor of creating a new one to benefit from new features in SS 5.x layout
  - Larger block size
  - Variant sub-blocks
- This requires a data migration
  - Ruled out AFM
    - we have ~2000 dependent fileset to migrate.
    - Found that AFM wouldn't sync up the timestamp of the directories.
  - Testing rsync/tar using list generated using ILM for initial copy, followed by a final rsync.
  - Testing parallel copy using fpsync/fpart
  - Testing mpifileutils

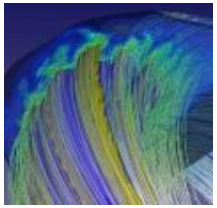
A New Order

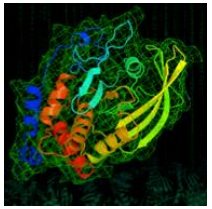


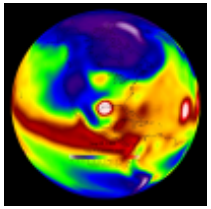


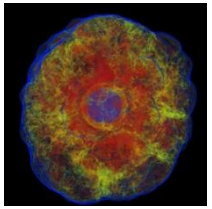
Tape System Migration Update

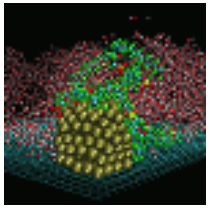






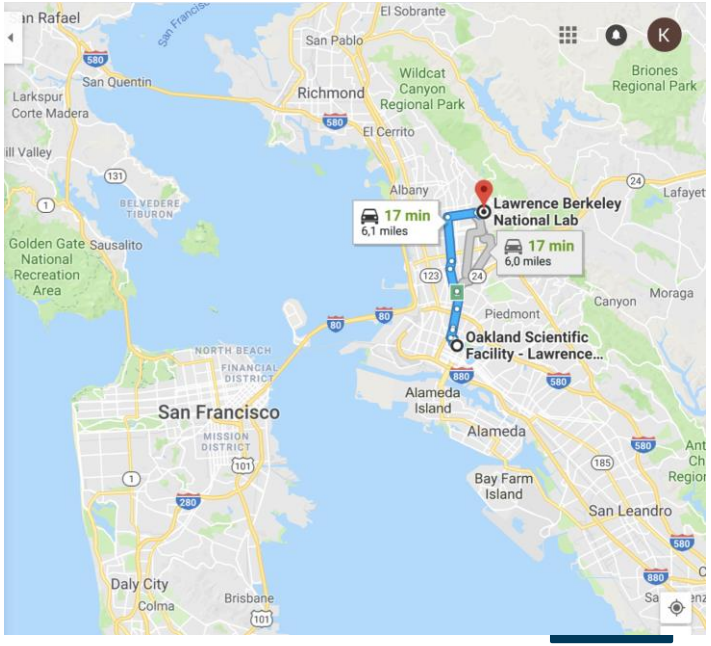






HPSS Archive – Two significant needs

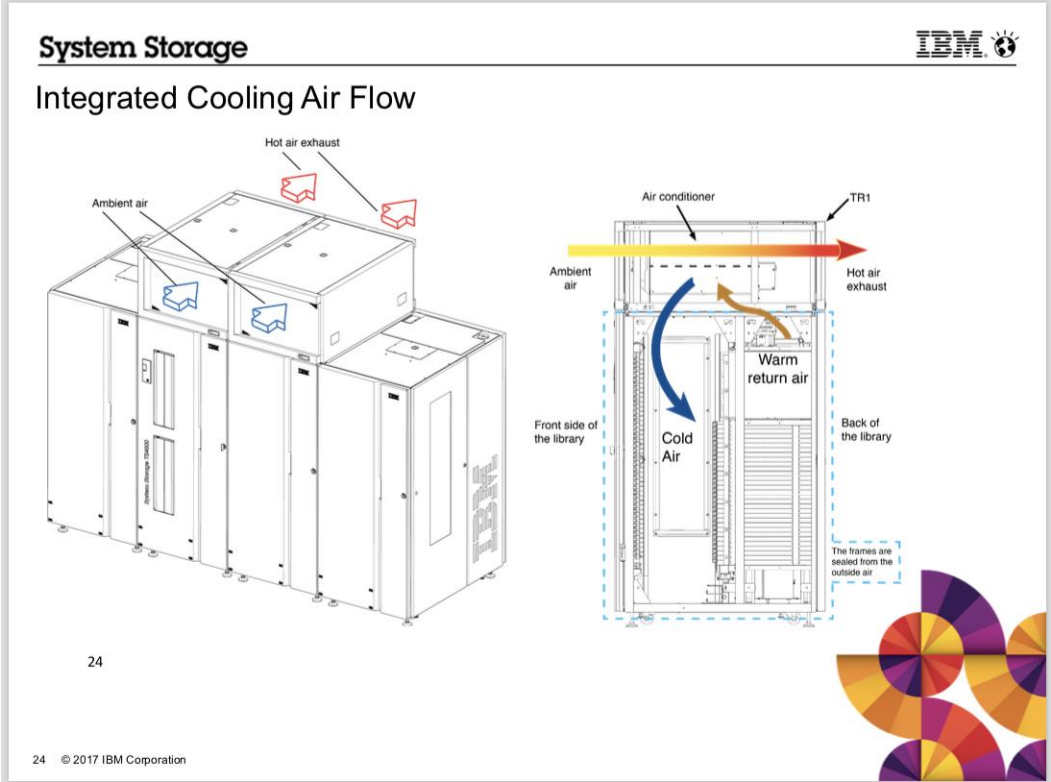
- **Technology decision**
  - Discontinued Oracle Enterprise Tape Drive
    - 4 Fully configured Oracle SL8500 libraries (archive)
    - 60 Oracle T10KC tape drives (archive)
    - 1 IBM TS3500 (mainly system backups)
    - 36 IBM TS1150 tape drives (mainly system backups)
- **Physical move required**
  - 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



U.S. DEPARTMENT OF ENERGY | Office of Science

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

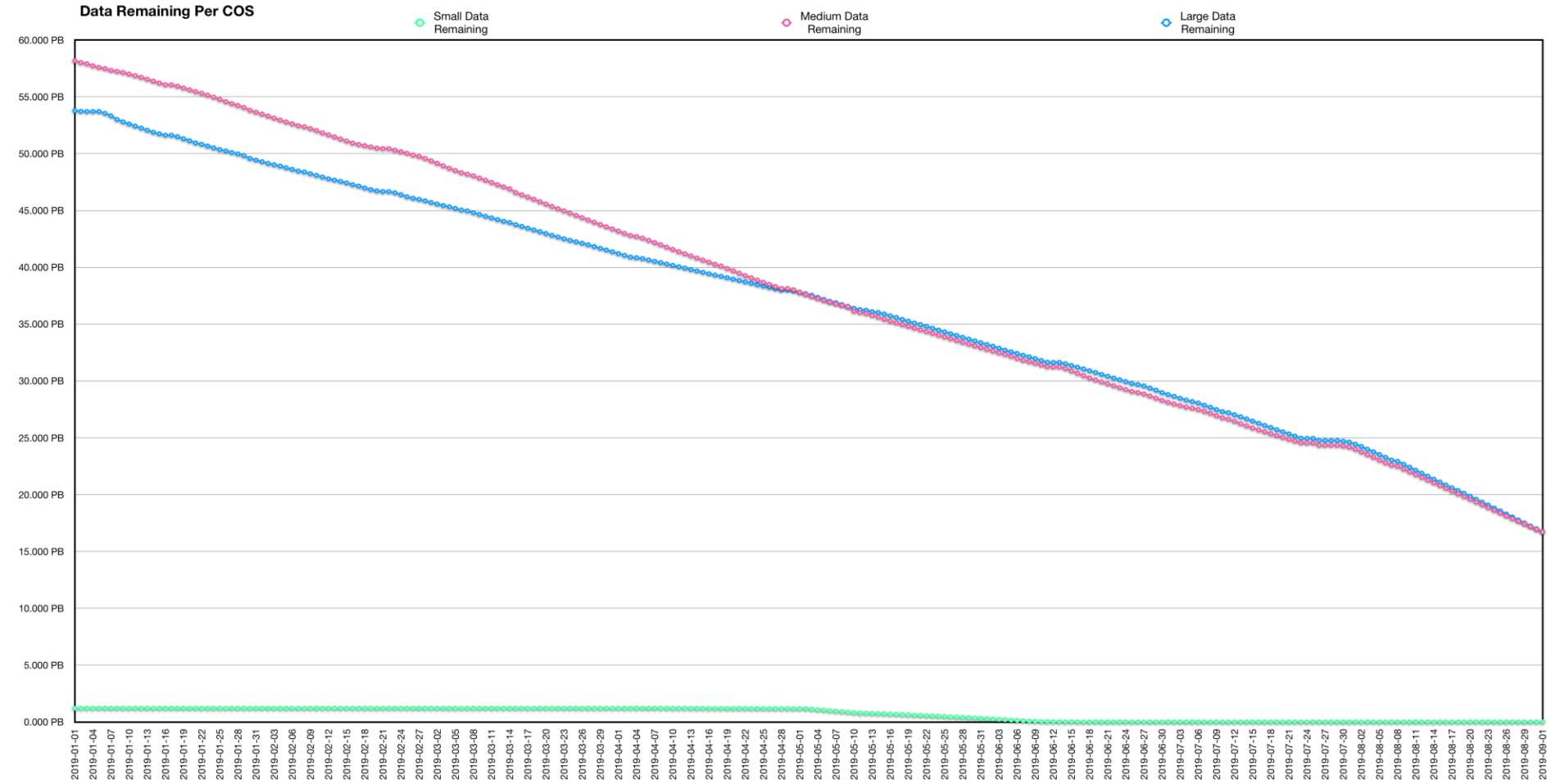
- Data migration stepped up
  - New goal bulk of data moved by Q1 2020
- Tape volumes processed chronologically
  - Later files are larger, better streaming from tape drives, better data rates.
- Smaller data
  - expect higher error rates on this data
  - More labor intensive

HPSS Archive – Status as of Sept. 2019

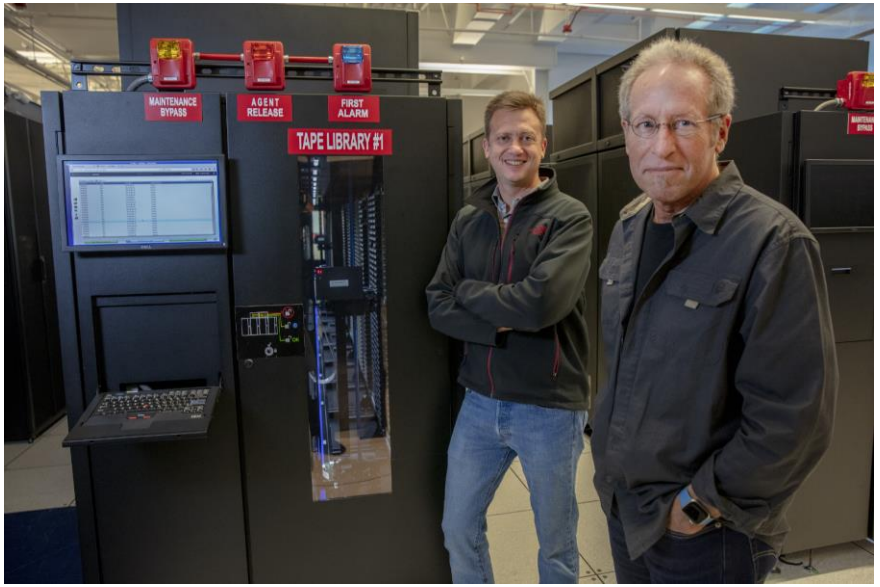
Petabytes in STK Libraries

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



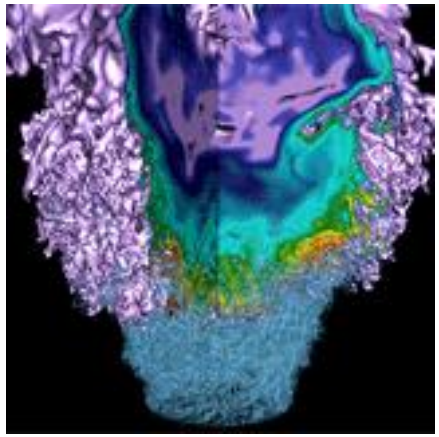
New Tape Libraries at Berkeley

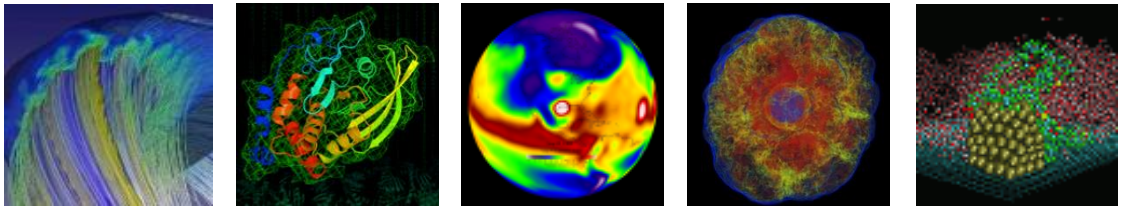


Nice article in HPCWire: https://bit.ly/2OwX24N

Not an “S”, also not an “S”

GPFS-HPSS-Integration (GHI)





GHI – JKL about it? It is:

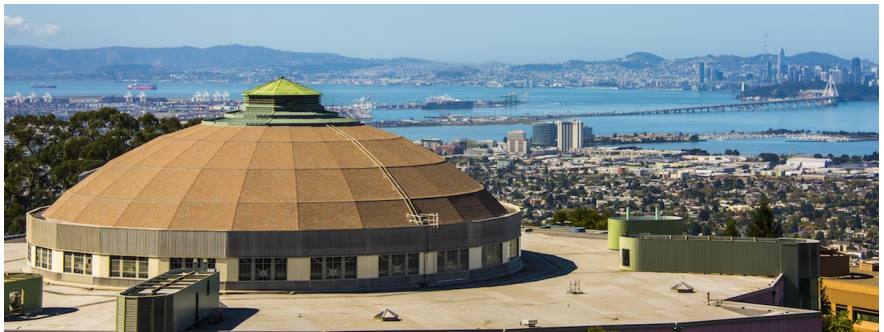
- **Optional piece of HPSS**
  - connects Spectrum Scale/GPFS and HPSS
  - automated data movement between the two
- **GHI primary functions:**
  - Space management (current focus)
    - Migrate
    - Purge
    - Recall
  - Disaster recovery (maybe later)
    - Backup
    - Restore

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

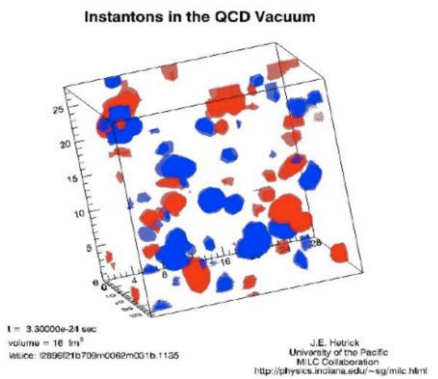
GHI Use Case: GHI Use Case: ALS

- 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

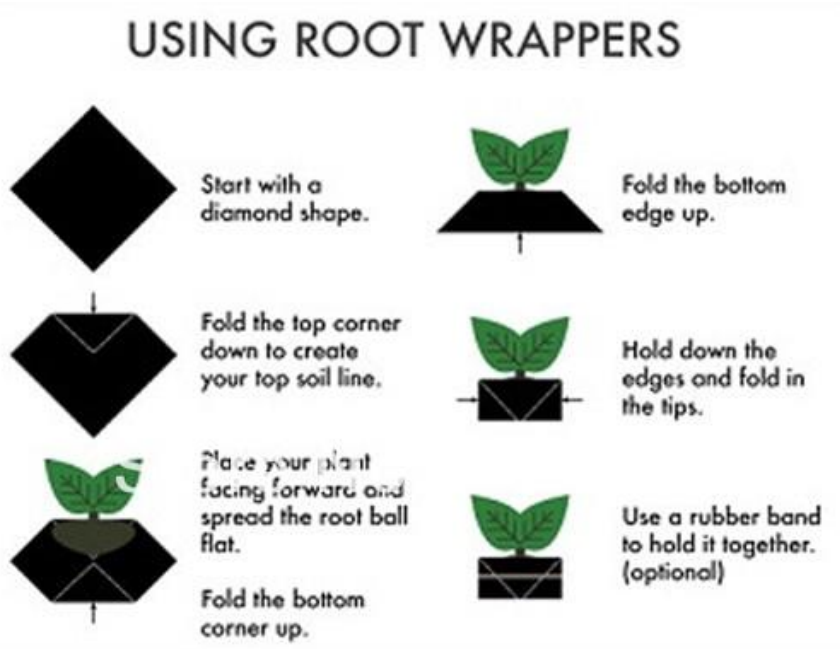
- Collecting QCD simulation data and serving it to scientists (along with descriptive metadata). Currently serves data out of HPSS via FTP, which limits the size of datasets they can serve
- *GHI will let them store TB-size datasets in HPSS and serve them out via Globus Sharing*
  - For large datasets,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



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!



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

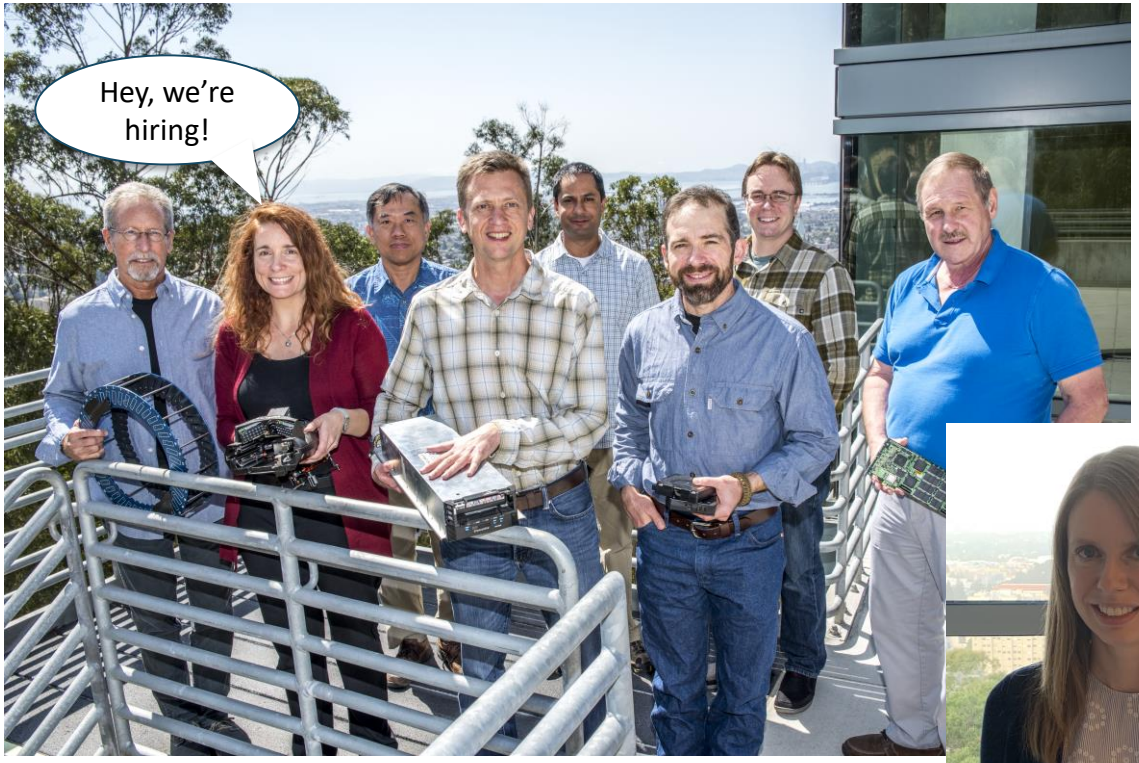
Other Projects – Future Talks?

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

Data Migration Results– Orchids and Onions (i.e. what went well, what, um, didn't)

--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?

National Energy Research Scientific Computing Center