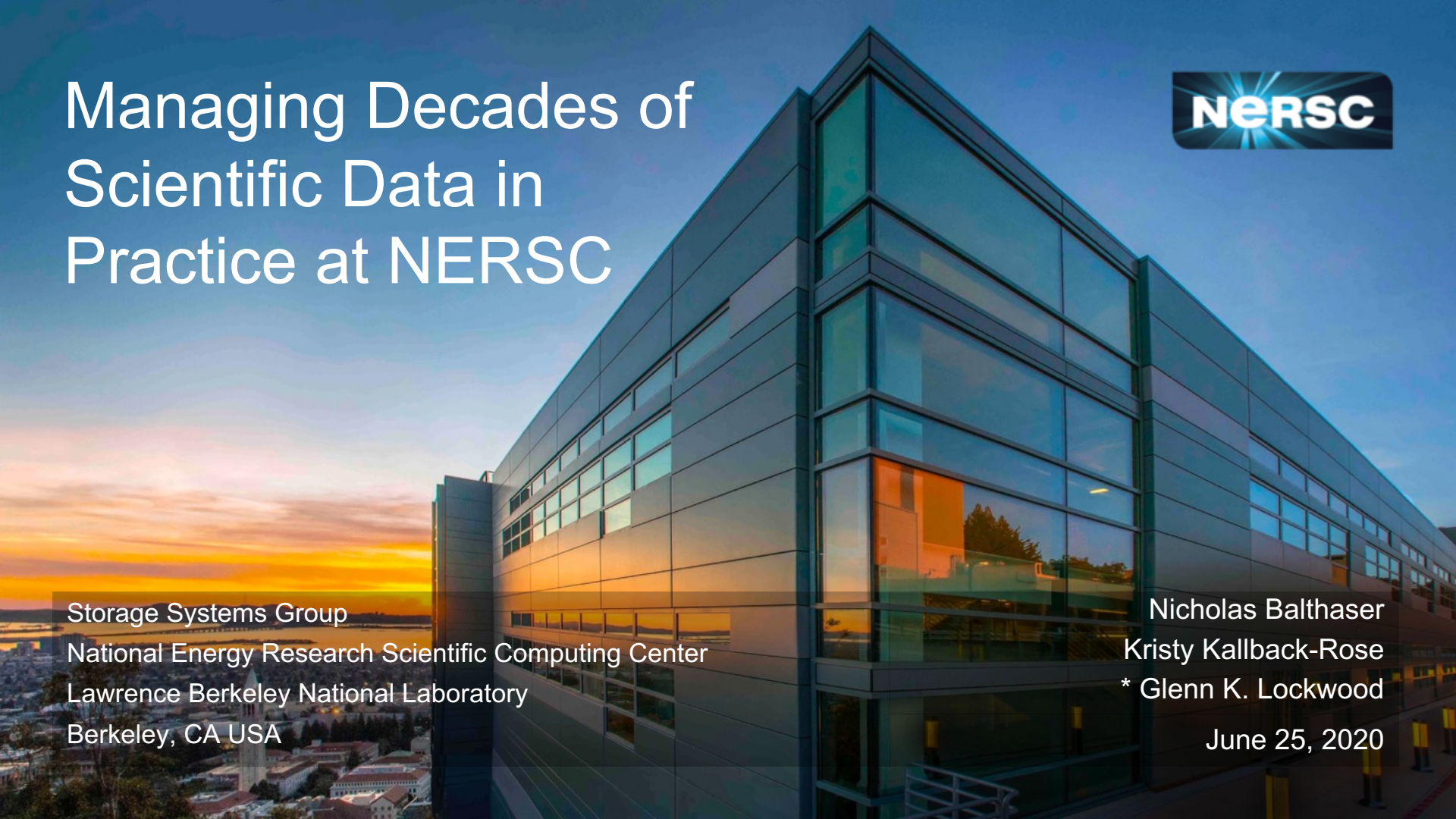
Managing Decades of Scientific Data in Practice at NERSC



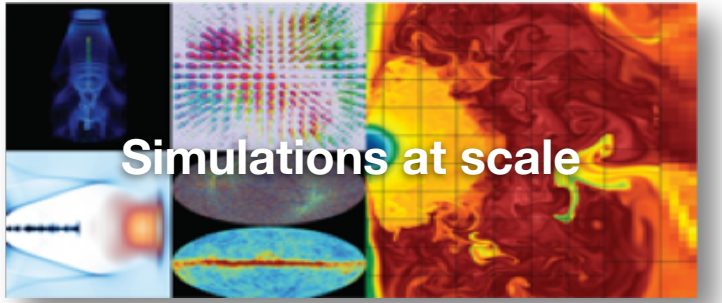
Storage Systems Group
National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
Berkeley, CA USA

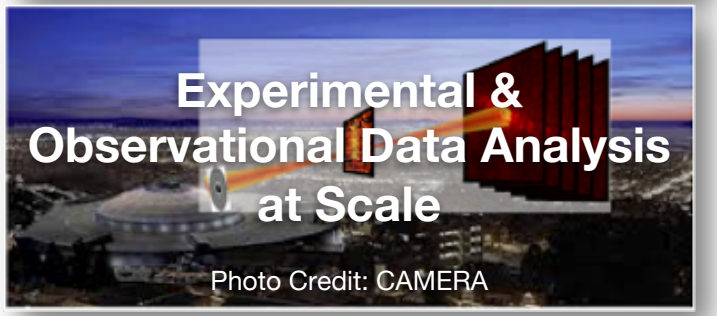
Nicholas Balthaser
Kristy Kallback-Rose
* Glenn K. Lockwood

June 25, 2020

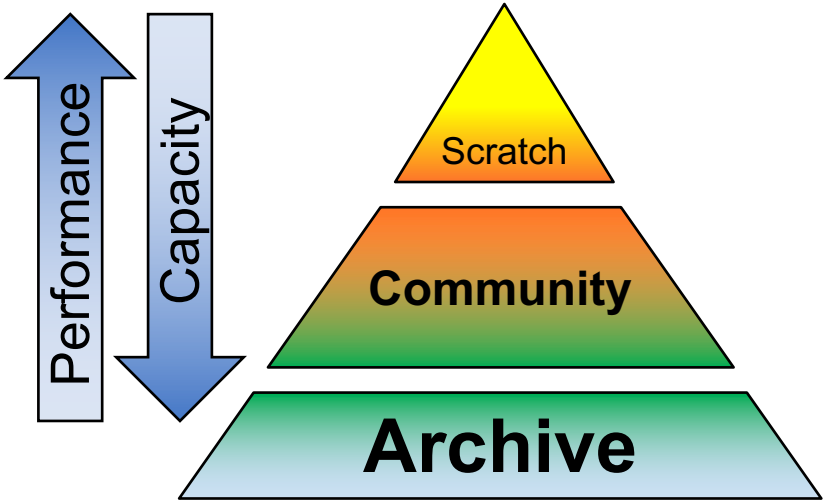
NERSC is the mission HPC facility
for DOE Office of Science

- Diverse workloads
  - Biology & environment, materials & chemistry, nuclear physics, fusion energy, high-energy physics
  - Experimental and AI-driven workloads
- Diverse users (2018)
  - 7,000 active users, 700 projects, 700 apps
  - > 1 exabyte of I/O
  - 2,500 publications
- Operating for 46 years





NERSC's hardware infrastructure for data



- **Scratch (weeks – months)**
  - Mounted on only one HPC system
  - User data purged after 4-12 weeks
  - Discarded when HPC system retired
- **Community (months – years)**
  - Mounted center-wide (HPCs, web, k8s)
  - Quotas
  - User data archived at project end
- **Archive (years – decades)**
  - Not "mounted" anywhere (object-like)
  - No effective quota
  - Infinite capacity, lowest performance

More info: G. K. Lockwood *et al.*, “Storage 2020: A Vision for the Future of HPC Storage,” Berkeley, CA, 2017.

2000	2002	2004	2006	2008	2010	2012	2014	2016	2018

Scratch – data discarded when hardware discarded

Community - data lives longer than hardware

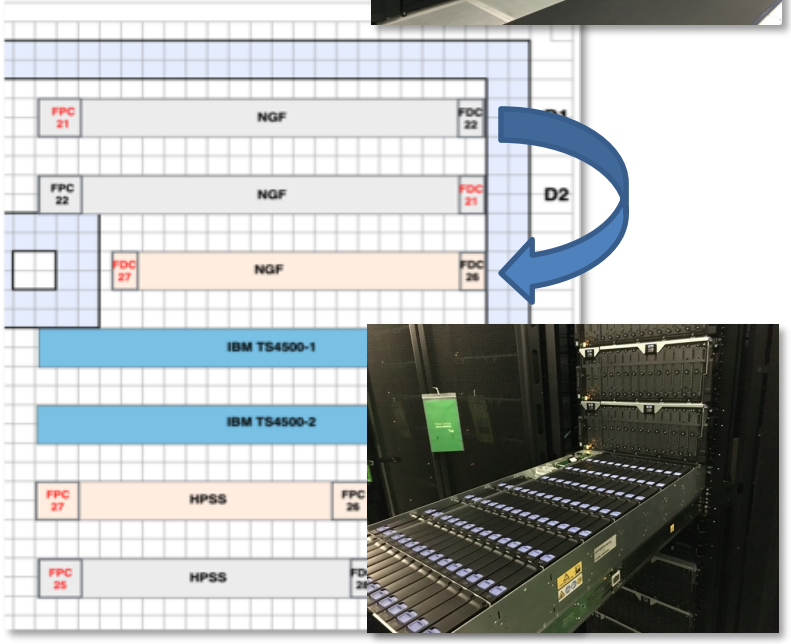
Archive – data lives longer than hardware

Managing generations of storage media:
Long-term data on disk-based file systems

Case Study: File System Expansion

Replaced "project" file system with "community" file system in 1Q2020

Project	Community
DDN SFA12k	**IBM ESS GL8c**
6 PB usable	**64 PB usable**
GPFS	**GPFS**
Supermicro (x86)	IBM (Power8)
4 TB HDs	14 TB HDDs
DDN RAID 8+2	IBM GNR 8+2
4 MiB block	16 MiB block



Standard process for upgrading project

- Use GPFS features:
  1. Add disk array to GPFS
  2. Drain old disk array
  3. Restripe (balance) blocks across remaining arrays
  4. Remove old disk array
- Performed during production
  - 100% online, during business hours
  - Non-disruptive – no user-facing notice announced
- **Not an option for Community!**
  - block layout changes due to scale
  - data must be copied through file interface

Option 1: Users migrate their own data

Pros

1. Staff don't have to manage data
2. Users *might* even clean up their data

Cons

1. Not transparent - significant user support required
2. "Ownership" of project poorly defined
3. Trigger I/O storm the day before the deadline

Option 2: MPI fileutils, fpsync, Globus, etc

Pros

1. Don't reinvent the wheel

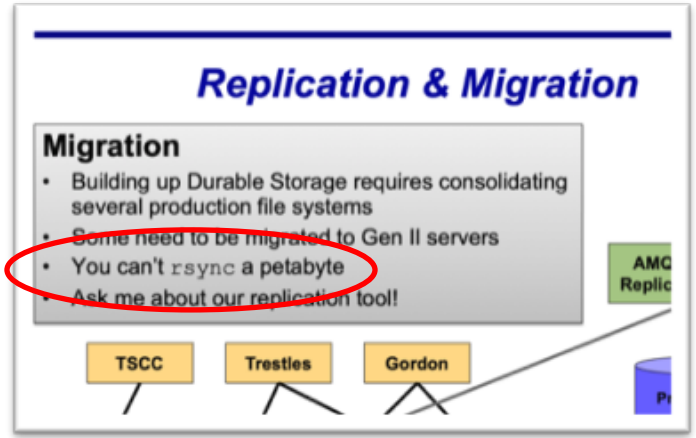
Cons

1. One month deadline, limited ability to test at scale
2. Edge cases may result in undefined behavior
   - Sparse files, gargantuan (500 TiB) files
   - “Creative” filenames (spaces, pipes in names)
   - ACLs, xattrs, hard links, ...
3. Not confident that user data will be transferred perfectly

Option 3: rsync, fpart, parallel, cp, tar

Pro: Won't mess up user data

Con: Engineering effort



From R. P. Wagner, "SDSC's Data Oasis Gen II: ZFS, 40GbE, and Replication." 2015 Lustre User Group.
http://cdn.opensfs.org/wp-content/uploads/2015/04/SDSC-Data-Oasis-GEn-II_Wagner.pdf

1. **Initial asynchronous copy (14 days)**
   - GPFS ILM scan to build work list
   - fpart + GNU parallel + 16 mover nodes
   - cp/tar and rsync
2. **Daily snapshot sync (12 – 48 hours)**
   - GPFS snapshots
   - Per-project rsync + checksum
3. **Final cut-over (12 hours)**
   - Old FS goes read-only
   - Final rsync of entire file system
   - Remount

For details, see Kallback-Rose (2020). https://storagetechshow.com/wp-content/uploads/16-NERSC-Kristy-Kallback-Rose.pdf

Big picture: using file systems for long-lived data

1. Avoid fork-lift upgrades (lots of work)
   - Plan for months of migration testing
   - Plan for outage for final cut-over
   - Plan to avoid block layout changes!
2. Drain/rebalance essential for long-term expansion, maintenance
3. Consider drain/rebalance granularity
   - Upgrade granularity is usually 1 disk array min
   - Due to assumption of reliable block LUNs
   - Fine-grained add/remove is preferable
     - Disaggregated block + network erasure probably better
     - Enables fail-in-place, dynamism is first-class feature

Managing generations of storage media:
Long-term data on tape-based archival storage

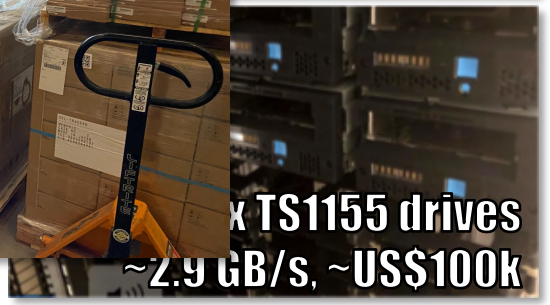
Case Study: Tape Archive Expansion

Replaced Oracle SL8500 libraries with IBM TS4500 libraries starting in 2018

Oracle SL8500	IBM TS4500
4 libraries	3 libraries
60 T10KC drives 68 T10KD drives	128 TS1155 drives 36 TS1150 drives
40,000 slots	39,000 slots
5 TB T10KC cartridges 8.5 TB T10KD cartridges	15 TB 3592-JD cartridges
31.5 GB/s peak	59.0 GB/s peak

Case Study: Tape Archive Expansion

- Usual refresh process relies on re
  1. Load new tape cartridges into library
  2. Rewrite [sparse] old tapes to new ta
  3. Remove old tape cartridges (if need
- Expansion cadence
  - Buy new cartridges every 3 – 4 mon
  - Buy new drives every 24 – 48 months
  - Buy new libraries every 5 – 10 years
- Enterprise tape: *everything* backwards compatible by ≥ 1 gen



Case Study: Tape Archive Expansion

- Oracle cartridges incompatible with IBM drives/libraries
  - Oracle drives, libraries also out of support (so this is urgent!)
  - Rely on archive software reading Oracle, writing IBM
- Repacking 150 PB of data takes months to years
  - Handling 29,000 cartridges takes a long time, period
  - Data migration must be done online
- Strategy
  1. Freeze Oracle library state – redirect incoming data to IBM
  2. Repack Oracle to IBM over the wire asynchronously

Unplugging the fire hoses

4Q17	1Q18	2Q18	3Q18	4Q18	1Q19	2Q19	3Q19	4Q19	1Q20

Data Repacking

Over-the-wire + Sneakernet

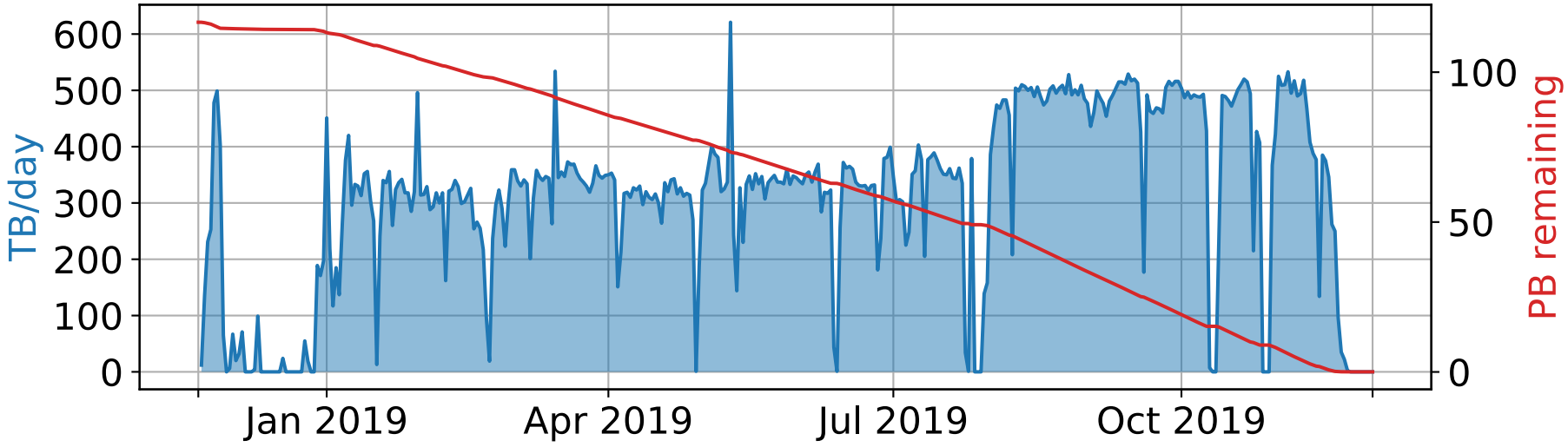
Repacking 150 PB of user data

Sneakernet

- 3,000 IBM cartridges
- 30 PB of user data
- 15 days (23 GB/s)

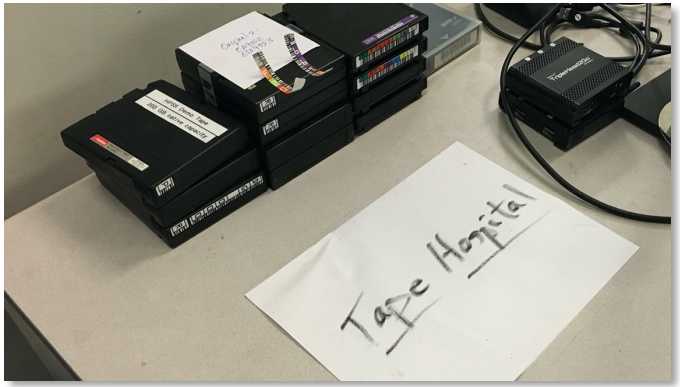
Network

- 23,910 cartridges
- 121 PB of user data
- 426 days (3.2 GB/s)



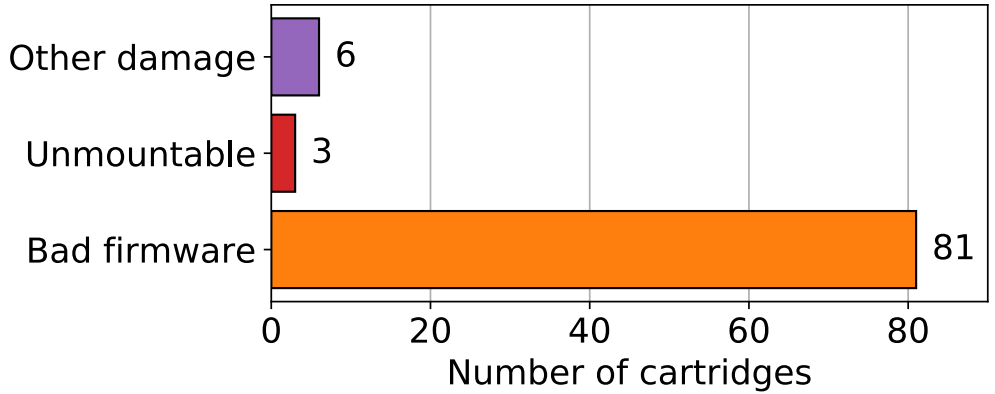
You don't know what you've lost until you need it

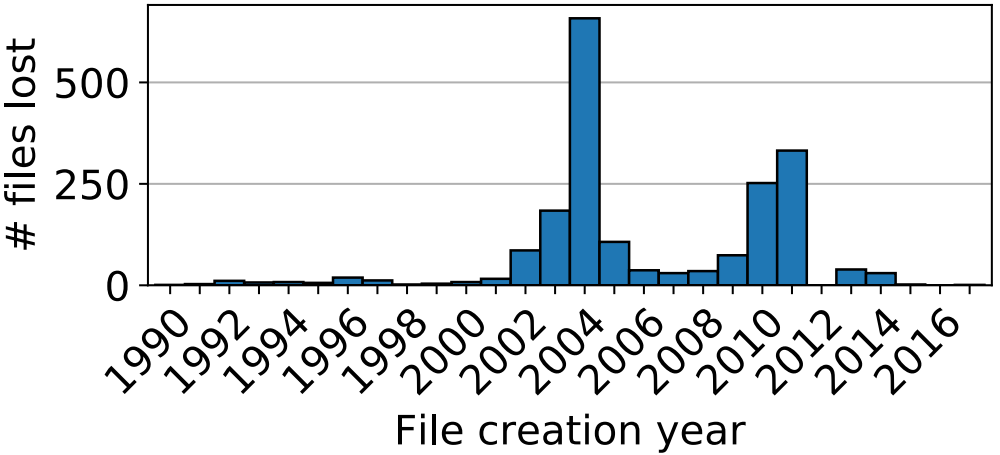
- **Data loss expected**
  - We only replicate small files
  - No routine scrubbing of data on tape
  - Rely on robustness of enterprise cartridges
  - Rely on built-in parity (UBER = 10^{-19})
- **Data loss uncovered**
  - 22 TB over 1,964 files unreadable
  - cf. 151,000 TB and 230,000,000 files
  - 148 users affected
- **Other issues – stickers, RFID, etc**



Data loss in practice

- Most data loss caused by bad drive firmware
  - 2011 incident caused drives to damage tapes
  - 3,000 tapes affected
  - 500 suffered loss in 2011
  - 81 deemed lost in 2019
- Unknown root cause for nine damaged tapes





Managing Generations of Data Centers

Data centers are not static

1974 – NERSC founded at LLNL (Livermore, CA)

1976 – Move to new data center (Livermore, CA)

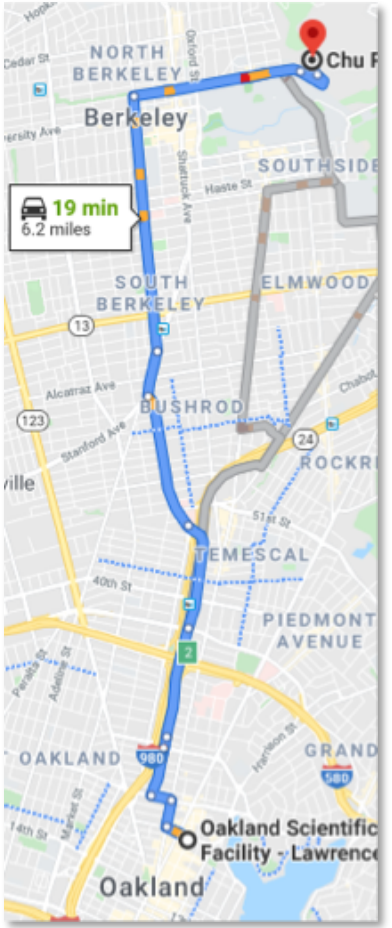
1996 – Move from LLNL to LBNL (Berkeley, CA)

2001 – Move to new data center (Oakland, CA)

2016 – Move to new data center (Berkeley, CA)

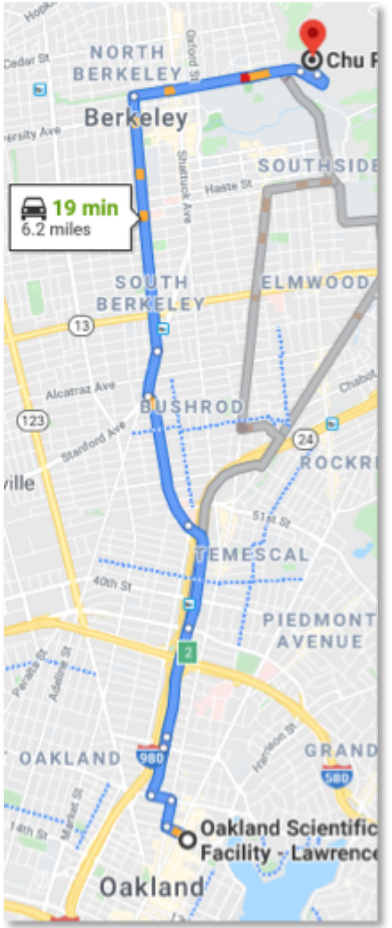
Migrating 150 PB archive in 2019-2020

- Actually occurred over 6 miles between Oakland and Berkeley
  - Sneakernet = trucks (3 PB/truck/day, ~100 GB/s)
  - Network = 400 GbE "superchannel"
- Relied on archival software features
  - repack over Ethernet
  - users requesting data from tapes that are on a truck



Migrating file systems over the wire

- Live migration of 4.8 PB of user data from Oakland to Berkeley in 2015
  - 400 GbE "superchannel"
  - 14x parallel routers
  - 20 GB/s transfer rate
- Relied on software support for
  - Live restripe of file system data from LUNs in Oakland to LUNs in Berkeley
  - VyOS to bridge Ethernet and InfiniBand



Lessons learned & best practices

What makes a good archival storage system?

Long-term storage = transparent data management

Data lives longer than hardware

∴

Long-term storage must be upgradeable

Must be able to change hardware without altering metadata

∴

Must migrate data transparently – avoid forklifts

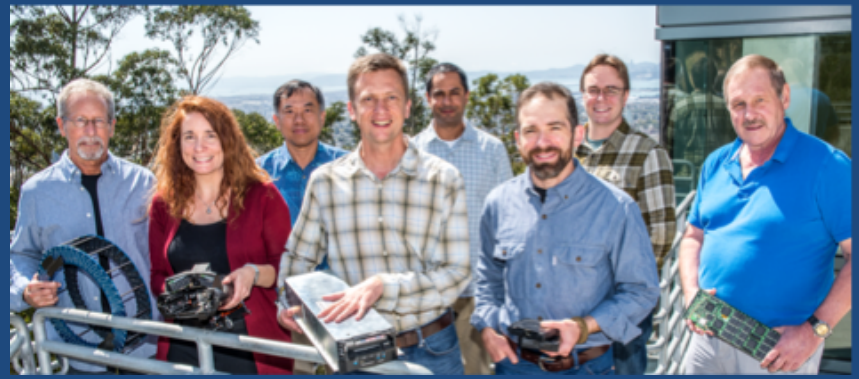
Long-term data management requirements

- Opposite of a forklift upgrade? Fine-grained, piecewise upgrades
- More granularity = more freedom in upgrade options
  - Good: RAIDed LUNs, controllers, enclosures, servers
  - Better: upgrade individual drives instead of whole RAID LUNs
  - Best: Tape cartridges, tape drives, tape enclosures, servers
  - Bestest: + data centers

Long-term storage software enables all of this

- Good archival storage systems are aware of full granularity
  - Strong networking enables disaggregation of devices
  - Disaggregation enables network erasure, fail-in-place, geo-distributed data/parity
- Geo-distribution simplifies data center migration
- Manageability is a first-class feature alongside performance
  - Live repack/restripe and online maintenance for hardware break/fix
  - Data migration over Ethernet, not just SAS/FC

Thank you!



NERSC Storage Systems Group (R-L):

- Wayne Hurlbert
- Kristy Kallback-Rose
- Rei Lee
- Damian Hazen (now net/security)
- Ravi Cheema
- Nick Balthaser
- Kirill Lozinskiy
- Greg Butler
- Melinda Jacobsen (not pictured)

We're hiring!

https://jobs.lbl.gov/jobs/hpc-storage-infrastructure-engineer-2697

Big-picture philosophies around long-term data

- Hardware/software diversity is a strength
  - Bad firmware is leading cause of device failure – showed tape, but also true for network, HDD, NVMe
  - Small data – replicate on different media (disk + tape, tape + tape)
  - Large data – spread over multiple media, firmware levels
- Preventing data loss requires active effort
  - Reading + checking is only way to verify data
  - Costs time, bandwidth, people, and hardware wear and tear