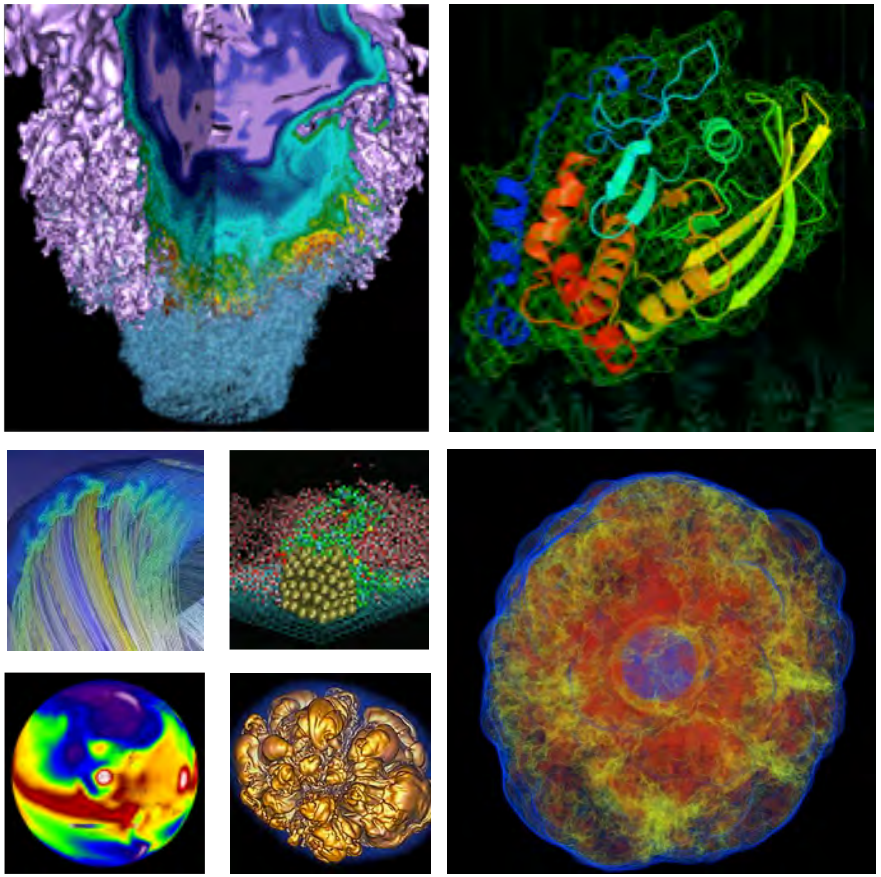
NERSC Tape Tech @MSST



Nicholas Balthaser

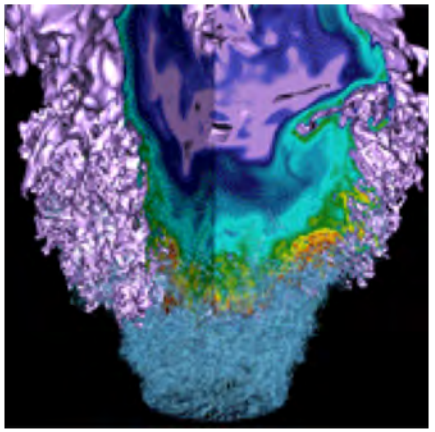
LBNL/NERSC Storage Systems

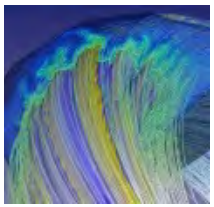
May 16, 2018

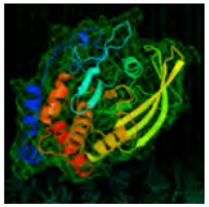
Agenda

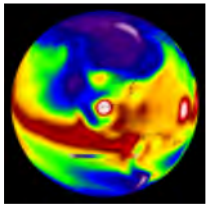
- General Systems and Storage Overview
- (Just one of our) Current Storage Challenges

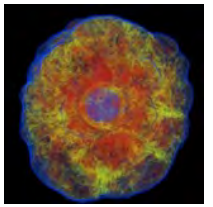
Systems & Storage Overview

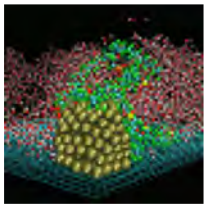






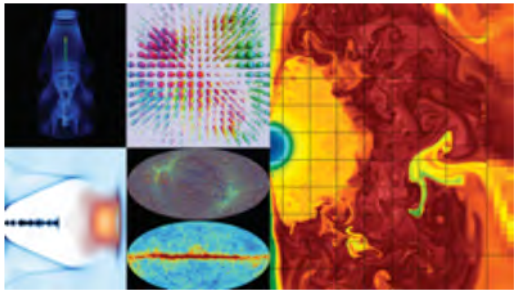


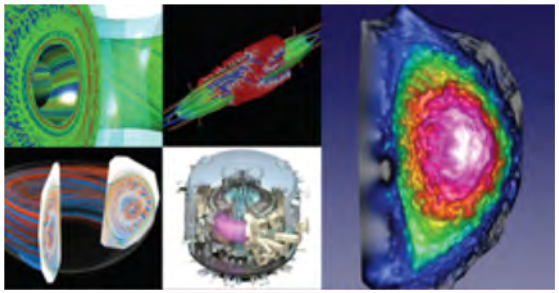




NERSC is the mission HPC computing center for the DOE Office of Science

- **HPC and data systems for the broad Office of Science community**
- **Approximately 7,000 users and 750 projects**
- **Diverse workload type and size**
  - Biology, Environment, Materials, Chemistry, Geophysics, Nuclear Physics, Fusion Energy, Plasma Physics, Computing Research
  - Single-core (not many, but some) to whole-system jobs





NERSC - Resources at a Glance 2018

Edison: Cray XC-30

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

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

Data-Intensive Systems
PDSF, JGI, Materials

Data Transfer Nodes
Science Gateways

7.6 PB Local Scratch 163 GB/s

28 PB Local Scratch >700 GB/s

2PB "DataWarp" I/O Accl. Layer >1.5 TB/s

16x FDR IB

32x FDR IB

14x FDR IB

136 GB/s

40 GB/s

5 GB/s

24 GB/s

Ethernet &
IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

/project — **7.1 PB DDN SFA12KE**

Sponsored Storage — **5.1 PB DDN SFA12KE**

/home — ***275 TB NetApp 5460***

HPSS — ***150 PB stored, 240 PB capacity, 42 years of community data***

2 x 10 Gb

2 x 100 Gb

Software Defined Networking

ESnet ENERGY SCIENCES NETWORK

File Systems

- **Lustre scratch for Large Compute systems**
  - Edison: 7.6 PB with 168 GB/s aggregate bandwidth
  - Cori: 28 PB with ~700 GB/s aggregate bandwidth
  - Periodically purged
- **GPFS mounted across all compute platforms**
  - Home Directories: 275 TB, optimized (SSDs) for small files, and high number of I/O operations
  - Project directories: 7.1 PB, 4TB/1M inode quota by default. Intended for larger files (SSD for metadata only), streaming I/O, and data that is actively being used. There is no purge policy, however inactive projects may be moved to the tape archive
  - Sponsored File System Storage: 5.1 PB on a separate file system, but with the same I/O characteristics as the Project file system (minus SSDs). This is a 'buy-in' program for projects that need additional space on disk. Program has a fixed $/TB, and a 5 year service agreement
  - GPFS file systems accessible from Cray compute nodes via DVS I/O forwarding

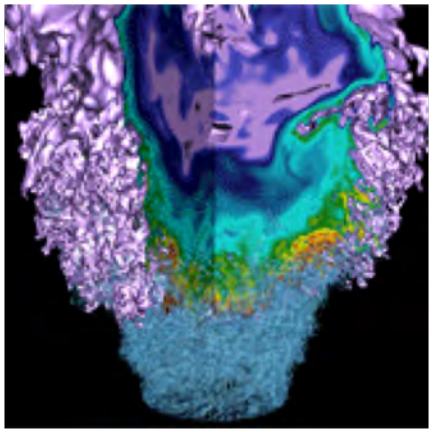
Tape Archive - HPSS

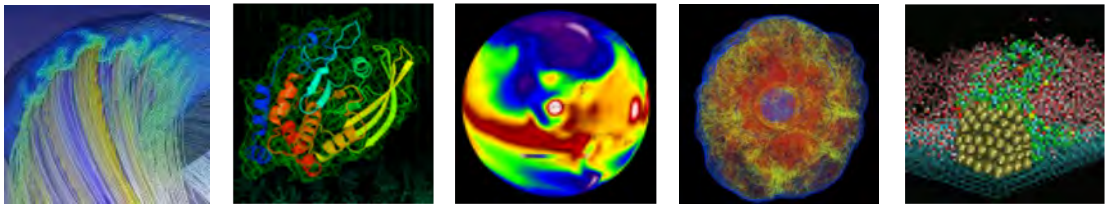
- **High Performance Storage System (HPSS)**
  - Developed over >20 years of collaboration among five Department of Energy laboratories and IBM, with significant contributions by universities and other laboratories worldwide.
  - archival storage system for long term data retention since 1998
  - Tiered storage system with a disk cache in front of a pool of tapes
    - On tape: ~150 PB
    - Disk Cache: 4 PB
  - Contains 42 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]

Top 5 HPSS Sites by PB Stored

Site	PBs Stored	# of files
ECMWF (European Weather)	318.63	330,047,386
UKMO (UK Weather)	239.49	241,498,397
NOAA R&D (US Weather)	133.86	94,260,512
Brookhaven Nat'l Lab	132.75	146,448,877
LBNL-NERSC (Archive)	123.75	224,571,569
LBNL-NERSC (Backup)	22.51	19,674,746
	146.26 PB	244,246,315

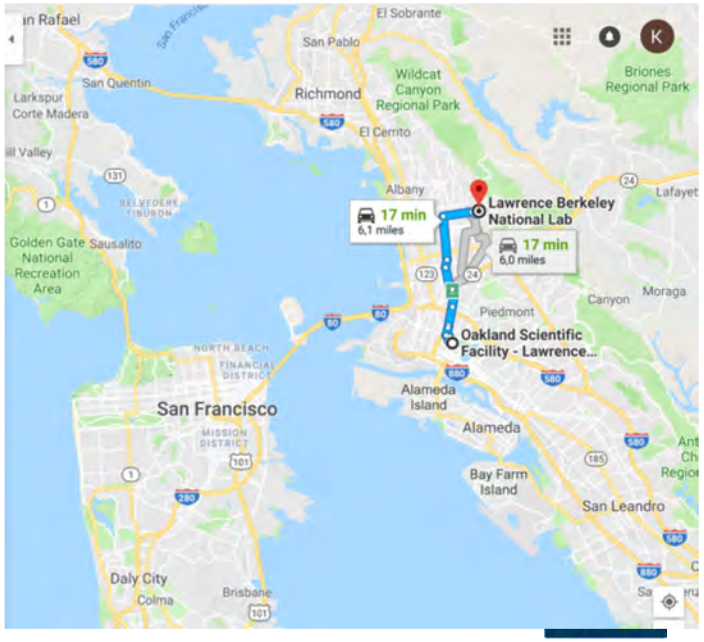
Current Storage Challenge





HPSS Archive – Two significant needs

- **Technology decision needed**
  - 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)

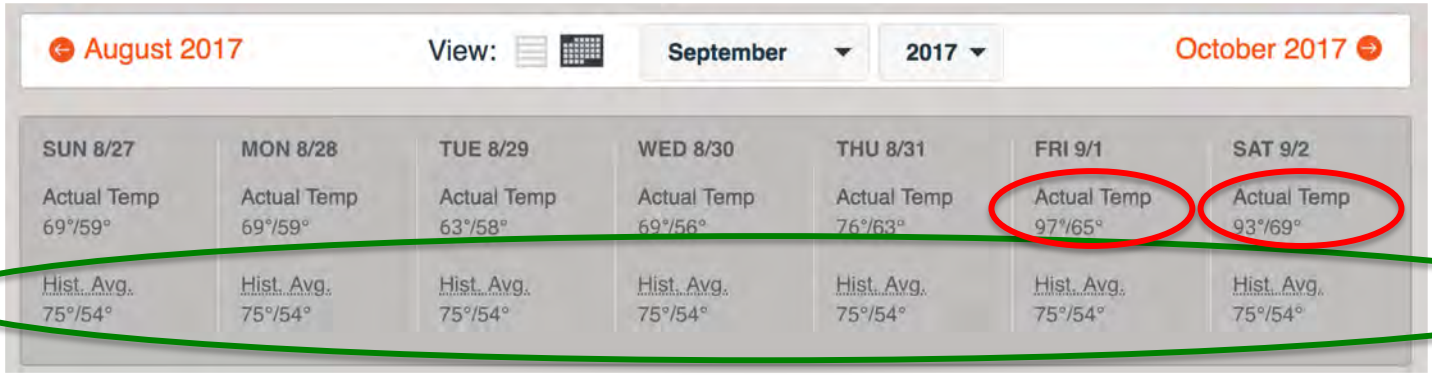


HPSS Archive – Data Center Constraints

- **Berkeley Data Center (BDC) is green (LEED Gold)**
  - Reliance on ambient conditions for year-round "free" air and water cooling
    - Intakes outside air, optionally cooled with tower water or heated with system exhaust
    - No chillers
  - (Ideally) take advantage of:
    - <75°F (23.9C) air year round
    - <70°F (21.1C)for 85% of year
    - RH 10% to 80%, but can change quickly
      - Usually more in 30% - 70% range

HPSS Archive – Data Center Realities

– And, sometimes it's too sunny in California

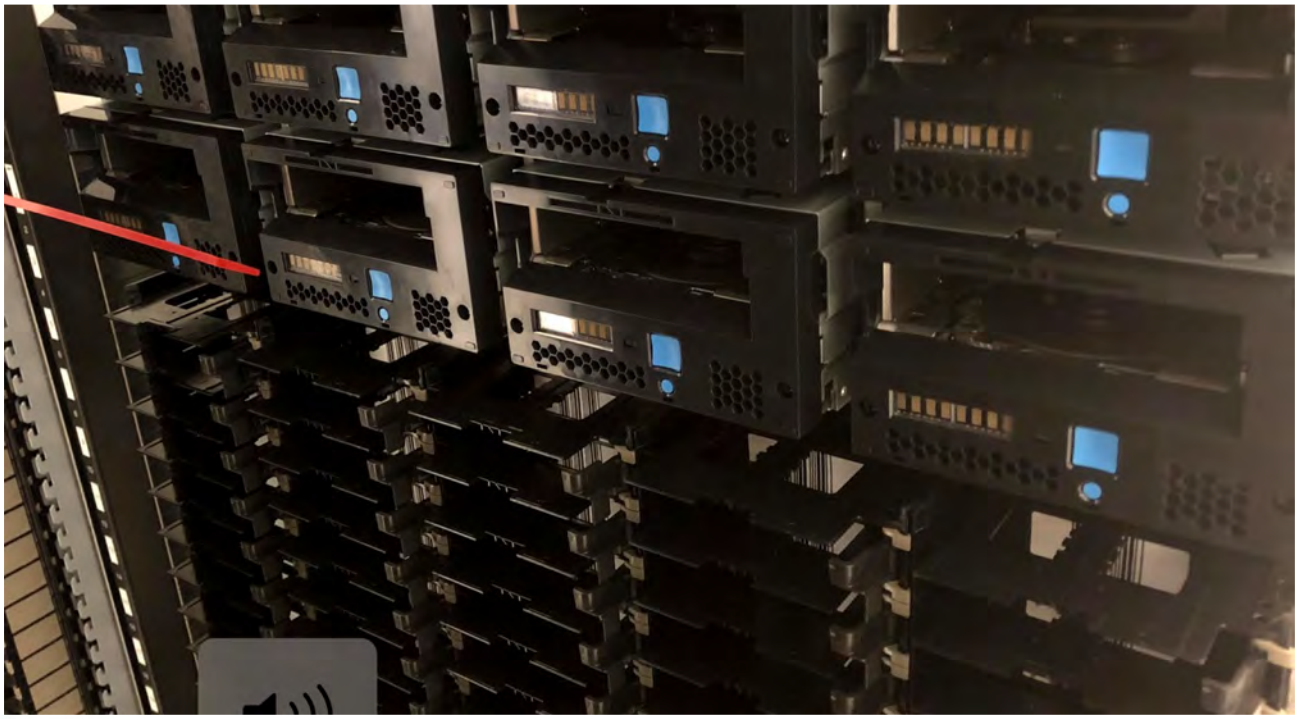


Alternatives to Consider

- **Co-lo to a data center elsewhere**
  - Costly
  - System management concerns
- **Cloud**
  - not performant enough
    - queue issues – need my data now
  - too costly
    - General expectation that data retrieval is the cloud killer,
      - but just letting 150PBs sit is more expensive
- **Room-within-a-room**
  - Space utilization goals and budget constraints
- **Disk-only**
  - Too costly
  - Increased power consumption

HPSS Archive – Green Data Center Solution

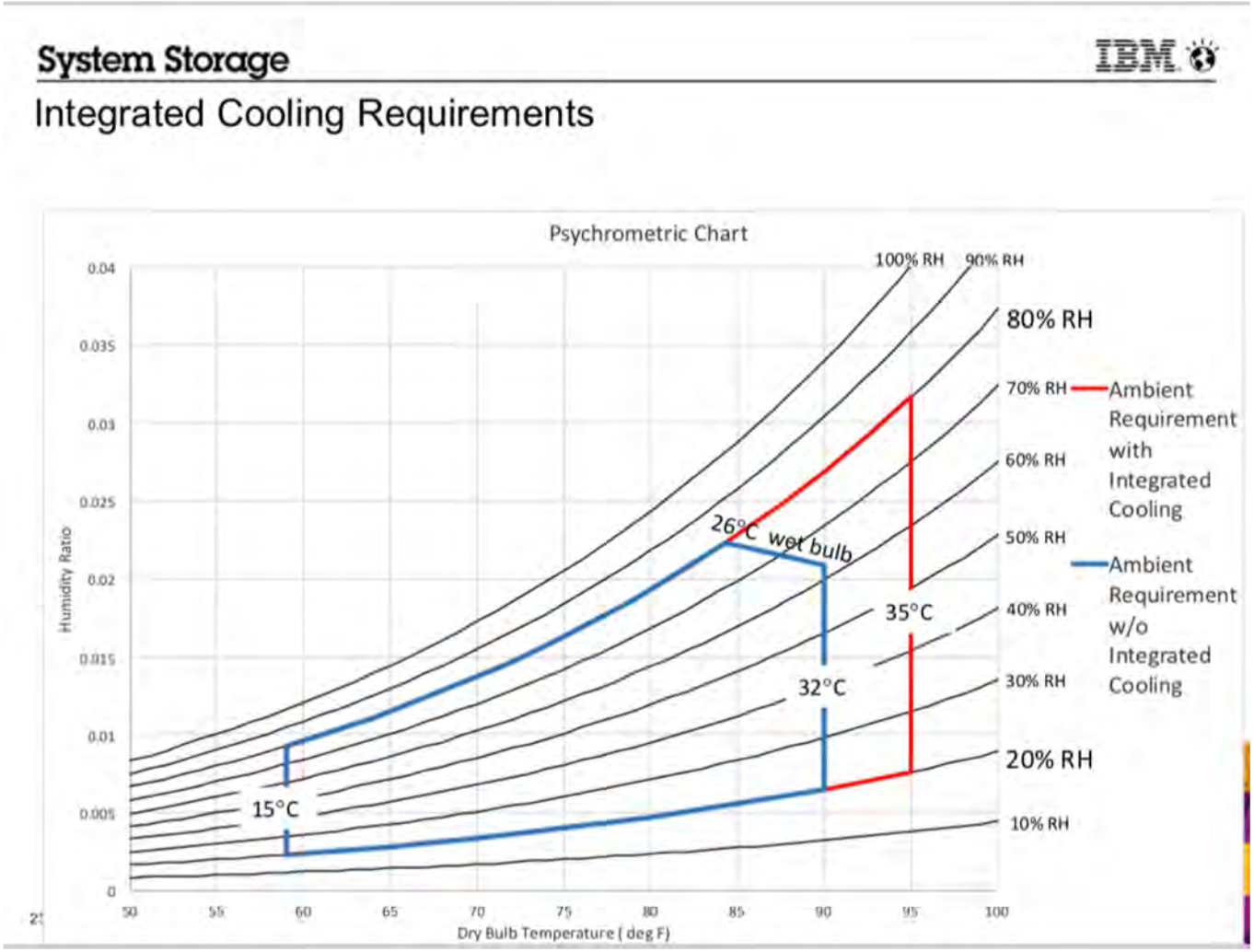
- **IBM TS4500 Tape Library *with Integrated Cooling***
  - seals off the library from ambient temperature and humidity
  - built-in AC units (atop library) keeps tapes and drives within operating spec



HPSS Archive – Tape Tech Decision

- **IBM had the library tech ready**
- **We have experience with TS3500/3592 drives for our backup system**

HPSS Archive – Tech Change

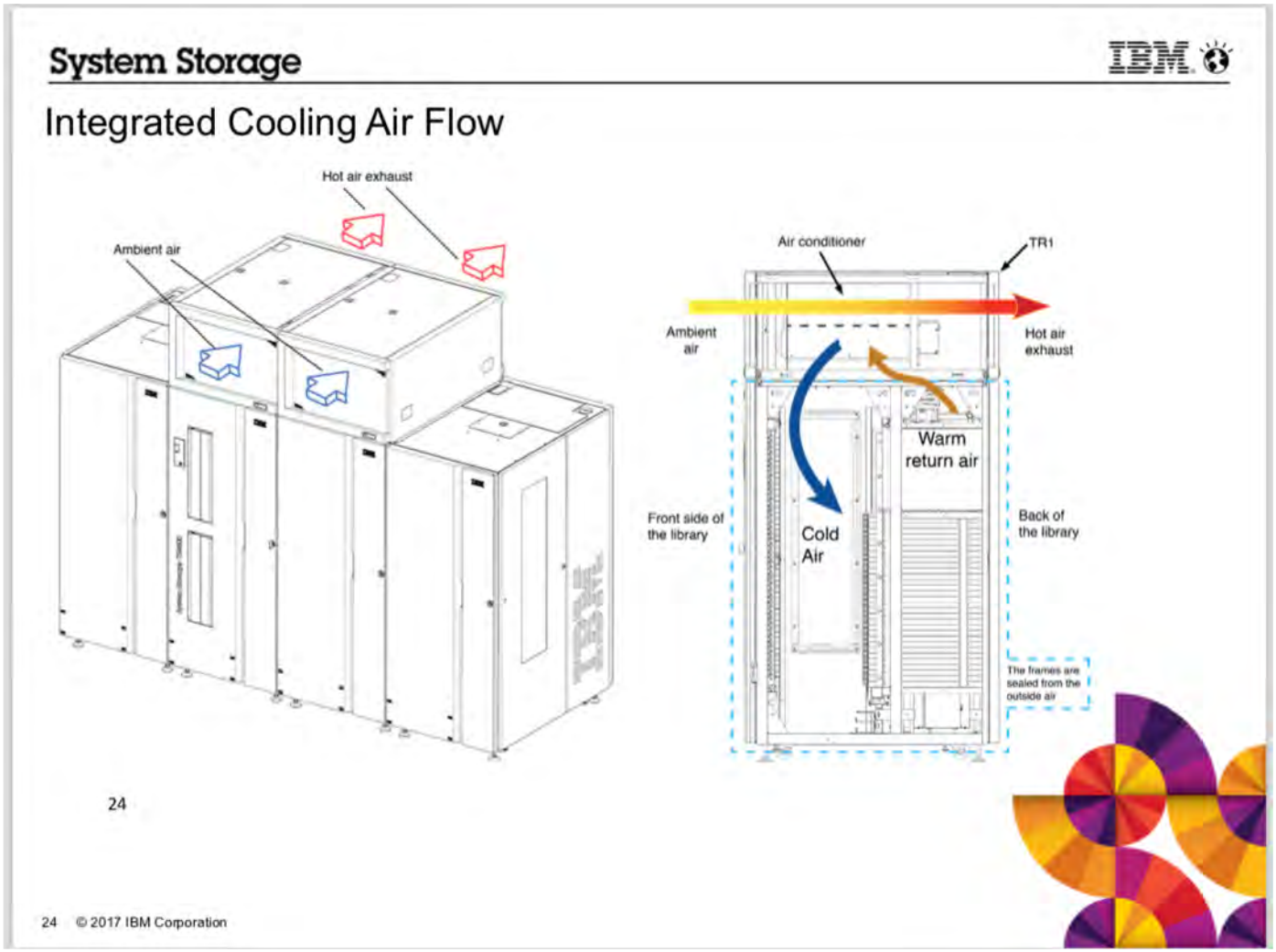


Slide from Lee Jesionowski: https://conference-indico.kek.jp/indico/event/28/session/12/contribution/30/material/slides/0.pdf

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 zones no L25, D25 instead)
  - 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



Slide from Lee Jesionowski: https://conference-indico.kek.jp/indico/event/28/session/12/contribution/30/material/slides/0.pdf

HPSS Archive – Tech Change (CRT)

- **Each library will be 4 "cooling zones"**
  - Cooling zone –logical separation
  - 16 frames
  - 64 TS1155/3592-55F(FC)/Jag(uar)6 tape drives
  - ~13,000 tape slots
    - JD media @15TB/cartridge
- **We'll install 2 of the above**

HPSS Archive – Tech Change (CRT)

- **Some things to get used to/improve**
  - No ACSLS
  - Can get some data out of IBM library, but
    - Need to re-do/build some tools
    - AFAIK, no pre-aggregated data like ACSLS
  - No Pass Through Port (PTP)
    - One Physical Volume Repository (PVR) per library
      - Since no shuttle option with IBM libraries
      - Could need more drives
    - But time to first byte should improve
      - PTP can fail, be slow

Value of Enterprise over LTO

NERSC Growth Rates

2016-2017: 2PB/mo.

2018: > 3PB/mo.

LTO Generations	LTO-6	LTO-7	LTO-8	LTO-9
Max Format Capacity (Native)	2.5 TB (L6)	6 TB (L7)	12 TB (L8)	Up to 25 TB (L9)
Other Format Capacities (Native)	1.5 TB (L5) (800 GB L4 R/O)	2.5 TB (L6) (1.5 TB L5 R/O)	6 TB (L7)	Up to 12 TB (L8) (6 TB L7 R/O)
Native Data Rate	160 MB/s	300 MB/s	360 MB/s	Up to 450 MB/s

2012 2015 2017

2011 2014 2017

TS1100 Generations	TS1140	TS1150	TS1155	Gen-6	Gen-7
Max Format Capacity (Native)	4 TB (JC) 1.6 TB (JB)	10 TB (JD) 7 TB (JC)	15 TB (JD) 7TB (JC)	Up to 20 TB (JE) 15 TB (JD) 10 TB (JC)	Up to 50 TB (JF) Up to 30 TB (JE) 15 TB (JD)
Other Format Capacities (Native)	1 TB (JB) 700 GB (JB) (All JA R/O)	4 TB (JC)	10 TB (JD) 4 TB (JC, R/O)	10 TB (JD) 7 TB (JC) 4 TB (JC, R/O)	10 TB (JD)
Native Data Rate	250 MB/s	360 MB/s	360 MB/s	Up to 420 MB/s	Up to 1000 MB/s
Attachment	FC-8	FC-8	FC-8, 10 GigE (RoCEv2)	FC-16, 25 GigE (RoCEv2)	TBD

Ability to up-format at higher density; gain capacity with same media, same DC footprint

Value of Enterprise over LTO

- **Time to first byte is important**
  - Recommended Access Order (RAO)
    - NERSC read-back rate is high ~40%
    - Supported under HPSS as Tape Ordered Recall
  - High Resolution Tape Directory
    - 1/128th of a tape wrap resolution per block versus 1⁄2 wrap for LTO
      - reach the target data block more efficiently
  - Higher search and rewind velocity about 30% faster then LTO

Value of Enterprise over LTO

- **42 years of data, reliability & stewardship are important**
  - Media Physical durability
    - Improved materials/construction over LTO
  - Loader supports 3x more load/unloads than LTO
  - Pro-active drive and media alerts

That being said…

- **We will have some LTO drives and media for second copies**
  - Gain experience
  - Diversity of technology

HPSS Archive – Loose timeline

April:

- **TS4500s w/integrated cooling were delivered at BDC**

May - July:

- **Fire suppression, system integration**
- **Oakland facility becomes read-only**
- **Data migrated to BDC via HPSS "repack" functionality**
  - >= 20 Oracle T10KC source drives at Oakland
  - >= 20 TS1155 destination drives at BDC
  - 400Gbps Oakland <-> BDC link

2020 (or earlier?) data migration complete

Future

GPFS HPSS Integration (GHI)– DMAPI-based integration of GPFS filesystem and HPSS tape archive (DR, Space Management)

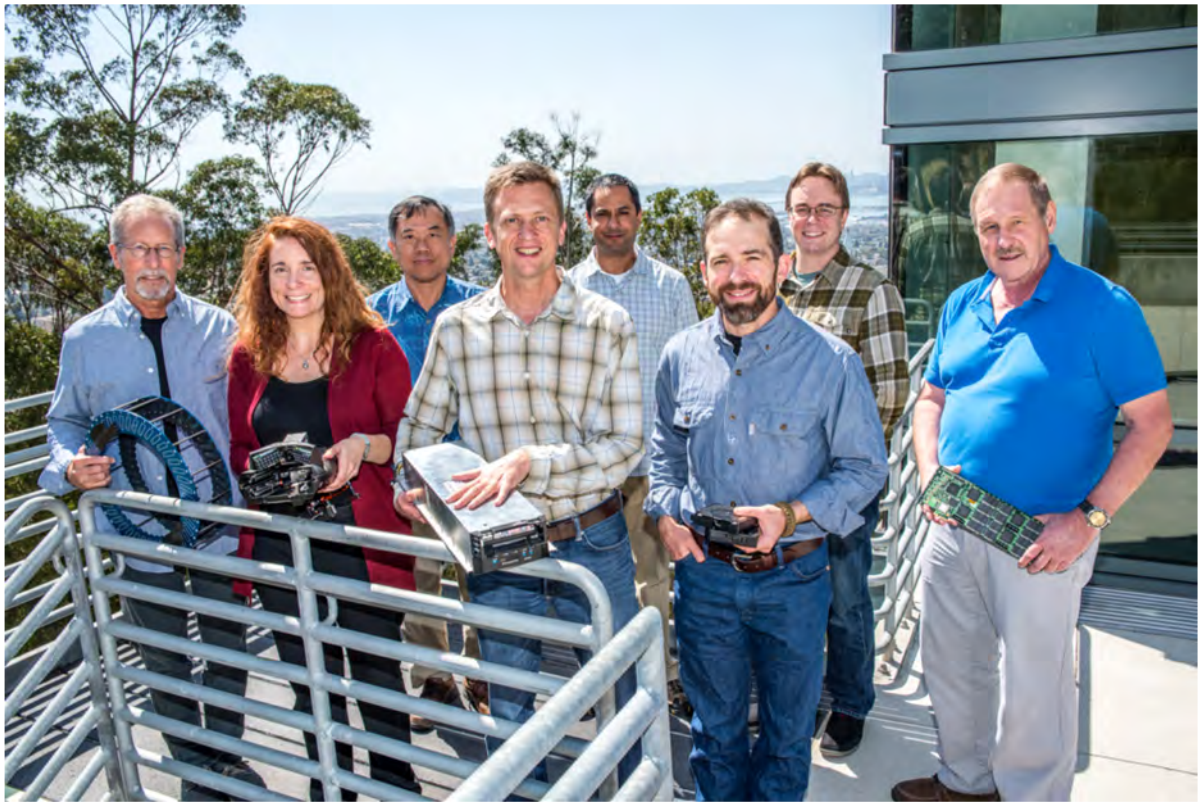
Upgrade HPSS – Leverage recent HPSS Features:

- RAIT
- RAO
- E2EDI
- Globus DSI support

Storage 2020 – Discussion of storage strategy at NERSC looking forward to 2020 and 2025

https://escholarship.org/uc/item/744479dp

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

Thank you. Questions?

National Energy Research Scientific Computing Center

Backup Slides

Library Eval Notes…

- **Mount rates**
  - SL8500 complex avg. 1600 mounts/day
  - TS3500 @OSF can sustain 600 mounts/hr in current config
  - TS4500s in Berkeley should be able to sustain similar rate
- **2011 Spectra TFinity Eval**
- **Capacity Projection through 2021**
  - Expect 300PB by end 2019, 450PB by 2021
  - 2xTS4500 max capacity 400PB
  - 3rd library will increase max capacity to 585PB
    - Drive/media capacity bump to 20TB in 2019 – 2020
- **Drive quantity TS1155s**
  - Handle projected load + data migration from OSF
  - No PTP required more drives to reduce cartridge mount wait

Glacier Eval Notes…

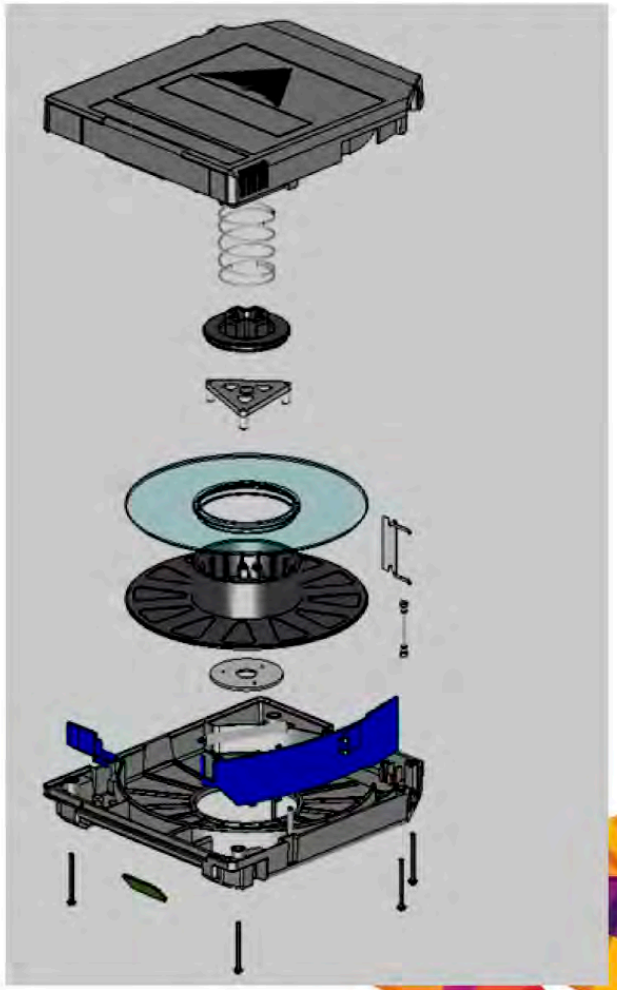
- **2 – 5x more expensive incl. staffing costs**
  - Plus limitations:
    - 4 hour retrieval
    - 4GB file limit
- **In house/DIY spec**
  - TS3500 w/2000 10TB cartridges – 20PB
  - expandable to 100PB with additional $2M media
  - $5.5M for 100PB for 5 years
- **Glacier**
  - Retention cost for 77PB is $10M/yr
  - Allows 5% retrieval for free – retention in this case is more expensive

3592 Cartridge Details

System Storage

3592 Robust Cartridge Design

- Robust cartridge design
  - Thicker plastics (vs. LTO)
  - Ribs to hold a reel
  - Five screws to tighten shells
  - Spec'd to withstand 1-m drops from all 6 axes without dataloss (not recommended!)
- Dust-proof curved door design
  - Effective dust-proof design for higher recording density
  - Passed an open/close test more than 50K cycles



LTO v. Enterprise Media

Table 1. *Tape Format Specifications*

Technical Feature	LTO-8 M Format	LTO-8	IBM TS1150	IBM TS1155	Oracle T10000D
Native Capacity	9 TB	Up to 12 TB	10 TB	15 TB	8.5 TB
Compressed Capacity‡	22.5 TB @ 2.5:1	Up to 30 TB @ 2.5:1	30 TB @ 3:1	45 TB @ 3:1	21.25 TB @ 2.5:1
Native Performance	300 MB/sec	Up to 360 MB/sec	360 MB/sec	360 MB/sec	252 MB/sec
Compressed Performance‡	750 MB/sec	Up to 900 MB/sec	900 MB/sec	900 MB/sec	800 MB/sec
Speed Matching	Yes	Yes	Yes	Yes	Not Specified
Speed Matching Range	112–300 MB/sec	112–360 MB/sec	112–360 MB/sec	112–360 MB/sec	Not Specified
Unrecoverable Bit Error Rate	1 in 10^{19}	1 in 10^{20}	1 in 10^{19}	1 in 10^{19}	1 in 10^{19}
Media Coating	BaFe	BaFe	BaFe	BaFe	BaFe
Reliability (Load/Unload Cycles)	80,000	300,000	Not Specified	Not Specified	150,000
Reliability (Head Life)	250,000 hr	250,000 hr	Not Specified	Not Specified	Not Specified
Load Time	22 sec	22 sec	16 sec	16 sec	13 sec
Average File Access	50 sec	50 sec	40 sec	40 sec	50 sec
Maximum Rewind	98 sec	98 sec	76 sec	76 sec	97 sec
Native Drive Encryption	Yes	Yes	Yes	Yes	Yes
Encryption Key Standard	KMIP	KMIP	Proprietary	Proprietary	Proprietary
WORM	No	Yes	Yes	Yes	Yes
LTFS Support	Yes	Yes	Yes	Yes	Yes
Power Consumption	32 W max, 11 W idle	32 W max, 11 W idle	46 W max, 24 W idle	46 W max, 24 W idle	90 W max, 36 W idle
Interface Ports	8 Gb FC	8 Gb FC	8 Gb FC, 10 GbE	8 Gb FC, 10 GbE	16 Gb FC, FICON
Warranty	3 years	3 years	1 year	1 year	1 year
Drive Price	$30,000	$30,000	$42,000	$42,000	$40,000
Media Price	$80	$150	$175	$175	$250

‡ Compressed Capacity and Performance: Specifications for compressed data are estimated based on normal data types; no vendor can guarantee these results.

LTO vs. 3592 Media

System Storage

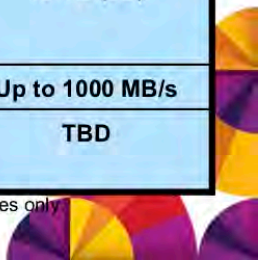
IBM Tape Drive History and Roadmaps

LTO Generations	LTO-6	LTO-7	LTO-8	LTO-9
Max Format Capacity (Native)	2.5 TB (L6)	6 TB (L7)	12 TB (L8)	Up to 25 TB (L9)
Other Format Capacities (Native)	1.5 TB (L5) (800 GB L4 R/O)	2.5 TB (L6) (1.5 TB L5 R/O)	6 TB (L7)	Up to 12 TB (L8) (6 TB L7 R/O)
Native Data Rate	160 MB/s	300 MB/s	360 MB/s	Up to 450 MB/s

2011 2012 2014 2015 2017 2017

TS1100 Generations	TS1140	TS1150	TS1155	Gen-6	Gen-7
Max Format Capacity (Native)	4 TB (JC) 1.6 TB (JB)	10 TB (JD) 7 TB (JC)	15 TB (JD) 7TB (JC)	Up to 20 TB (JE) 15 TB (JD) 10 TB (JC)	Up to 50 TB (JF) Up to 30 TB (JE) 15 TB (JD)
Other Format Capacities (Native)	1 TB (JB) 700 GB (JB) (All JA R/O)	4 TB (JC)	10 TB (JD) 4 TB (JC, R/O)	10 TB (JD) 7 TB (JC) 4 TB (JC, R/O)	10 TB (JD)
Native Data Rate	250 MB/s	360 MB/s	360 MB/s	Up to 420 MB/s	Up to 1000 MB/s
Attachment	FC-8	FC-8	FC-8, 10 GigE (RoCEv2)	FC-16, 25 GigE (RoCEv2)	TBD

Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.



HPSS Archive – Loose timeline

April:

- **TS4500s w/integrated cooling were delivered at BDC**

May:

- **½ Disk Cache + ½ of disk movers moved to Berkeley DC (BDC)**
  - Not yet enabled, just in preparation
- **½ Disk Cache + ½ of disk movers remain at Oakland**

June:

- **TS4500s w/integrated cooling installed at BDC**

July:

- **Move HPSS core server to BDC**
- **Enable ½ Disk Cache, ½ disk movers at BDC**
- **Enable Writes to new library at BDC**
- **Oakland tapes now read-only [user access, repacks]**
- **Remaining ½ disk cache made read-only, drained**
- **Move remaining ½ disk cache, movers to BDC**
- **Data migrated to BDC via HPSS "repack" functionality**
  - >= 20 Oracle T10KC source drives at Oakland
  - >= 20 TS1155 destination drives at BDC
  - 400Gbps Oakland <-> BDC link

2020 (or earlier?) data migration complete

- **We're not in Oakland any more, Dorothy.**