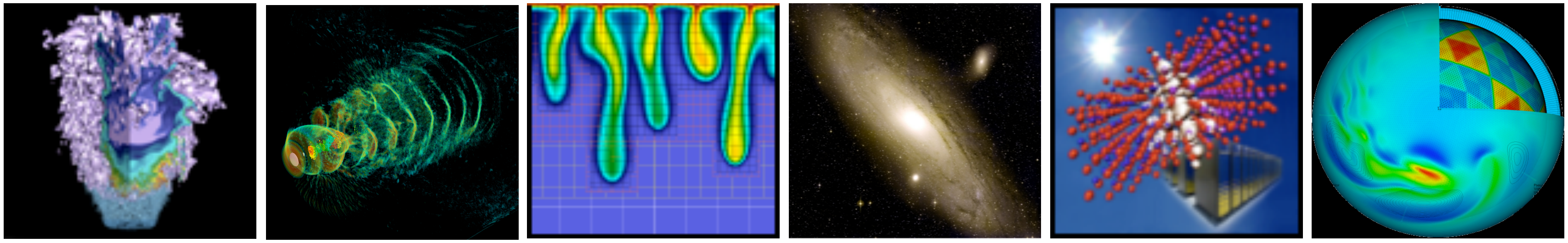


Leveraging the Business Value of Tape

Jason Hick
jhick@lbl.gov
NERSC Storage Systems Group

Fujifilm Global IT Executive Summit
June 8-10, 2011

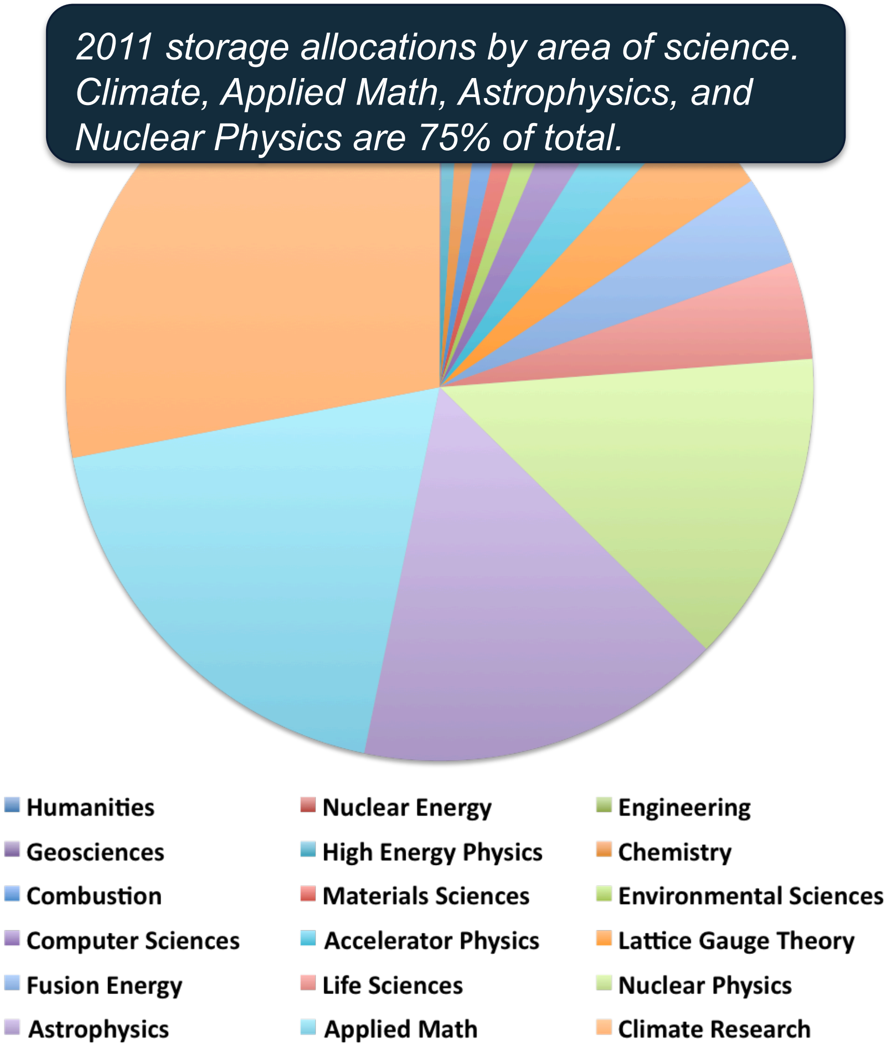
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

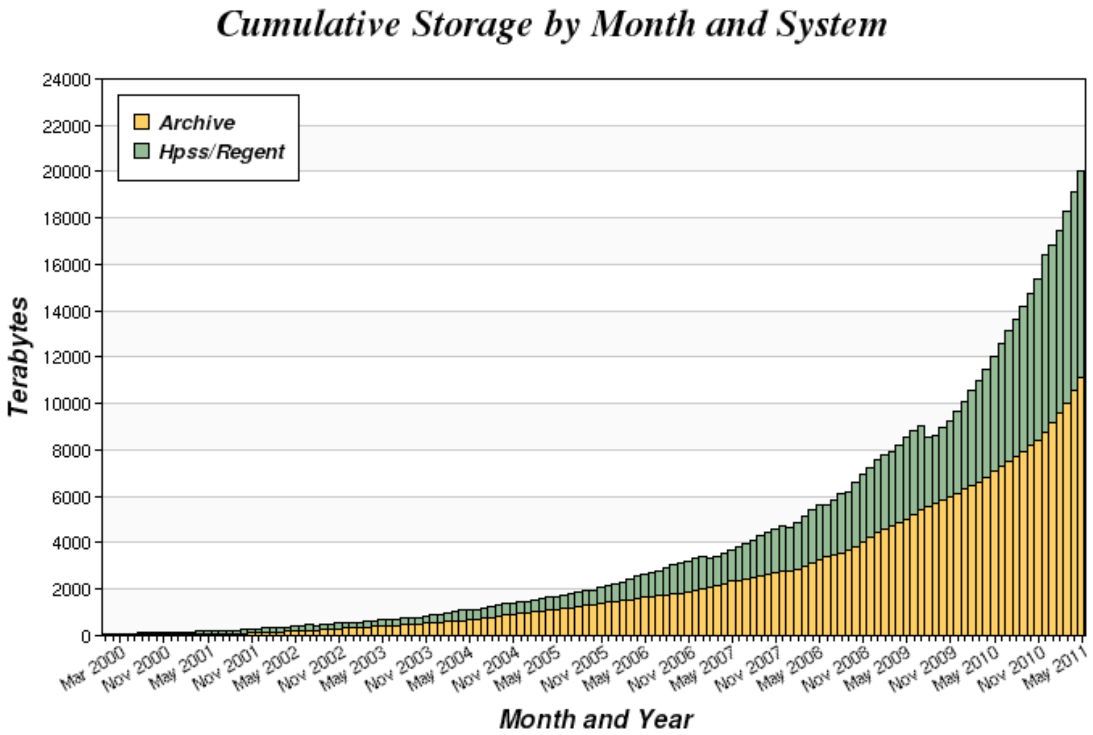
The Production Facility for DOE Office of Science

- **Operated by UC for the DOE**
- **NERSC serves a large population**
  - Approximately 4000 users, 400 projects, 500 codes
  - Focus on “unique” resources
    - High-end computing systems
    - High-end storage systems
      - Large shared GPFS (a.k.a. NGF)
      - Large archive (a.k.a. HPSS)
    - Interface to high speed networking
      - Center-wide 10Gb
      - Testing 100Gb (a.k.a. ANI)
- **Our mission is to accelerate the pace of discovery by providing high performance computing, data, and communication services to the DOE Office of Science community.**


2011 storage allocations by area of science. Climate, Applied Math, Astrophysics, and Nuclear Physics are 75% of total.

Tape System at NERSC

- As of Mar 2011, tape holds 18 PB of data with the ability to scale tc over 200 PBs
- Tape provides average compression of 40% for data stored at NERSC.
- Our average annual growth is 50-60%.
- Tape drives doubling capacity every 2 years is essential.
- We manage to a fixed media budget and space footprint.
- We use enterprise tape with a single copy of data.
- Average user file size in HPSS is 65 MB.
- 30-40% of IO to HPSS are reads, so plan/provision for reads off tape.
- Peak day was 170TB with 50% of that reads. Average daily IO is 50TB.



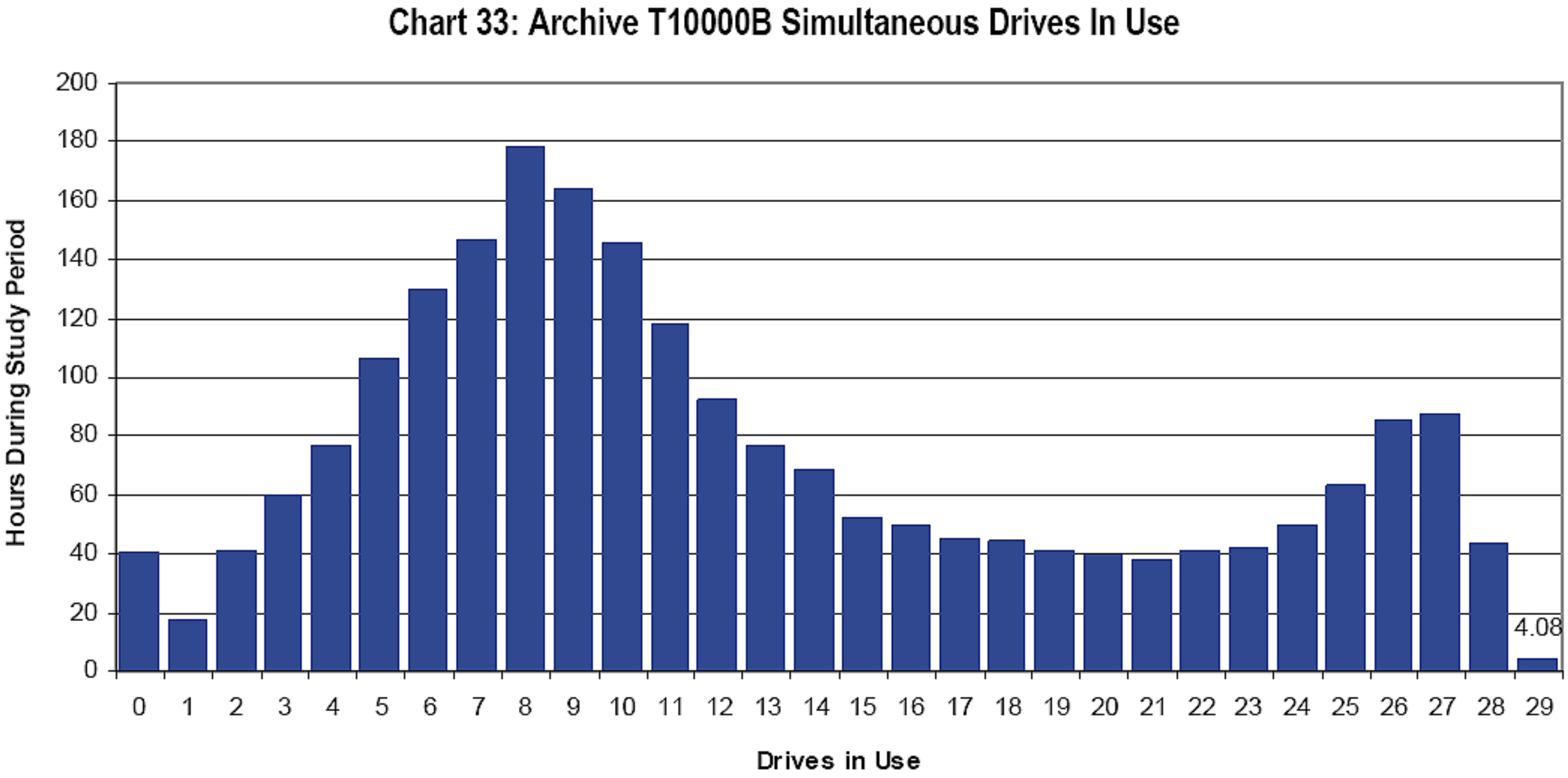
Archival Storage

- **User HPSS**
  - Single transfers 1GB/sec read/write
  - Aggregate bandwidth 4+GB/sec
  - Average daily IO of 30TB, with peak at 150TB
  - 200TB disk cache
  - 24 9840D, 48 T10KB, 16 T10KC tape drives
  - Largest file: 5.5TB
  - Oldest file: Jan 1976

Uses of Tape in an Active Archive

- **Migration from disk cache within HSM (HPSS)**
  - Users storing new files
  - Every 30 minutes ~10 drives migrate data from disk to tape (30-150TB per day)
- **Staging from tape to disk within HSM (HPSS)**
  - Users reading/fetching files
  - Random requests for data on tape (above 40MB in file size, disk retains 5 days worth of data)
  - Have to provision for number of concurrent requests
- **Repacks and internal data migration efforts**
  - Provision for some number of tape drives to enable migration to new tape technologies, or data re-mastering (repack)

Planning for Reads

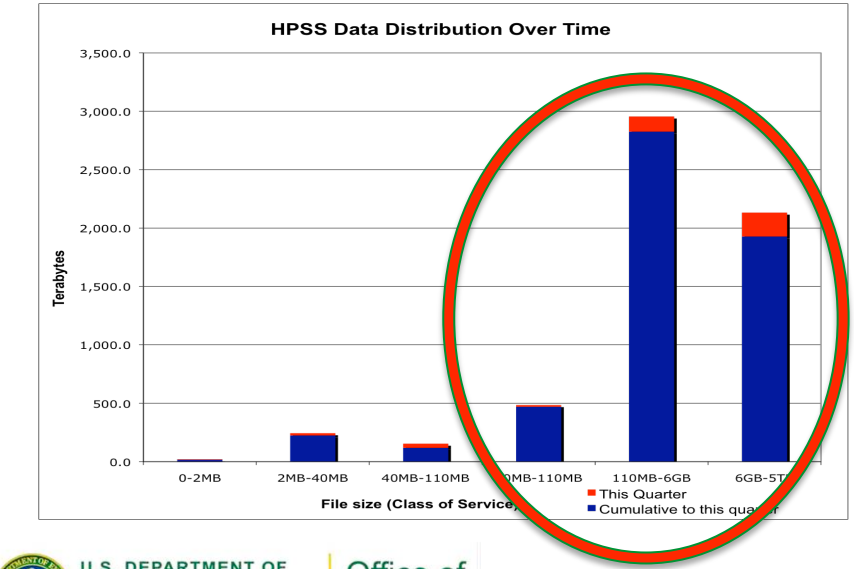


Number of concurrent drives in use is of key interest to us as it gives us a good idea of peak demand, which is what we plan to for the active archive.

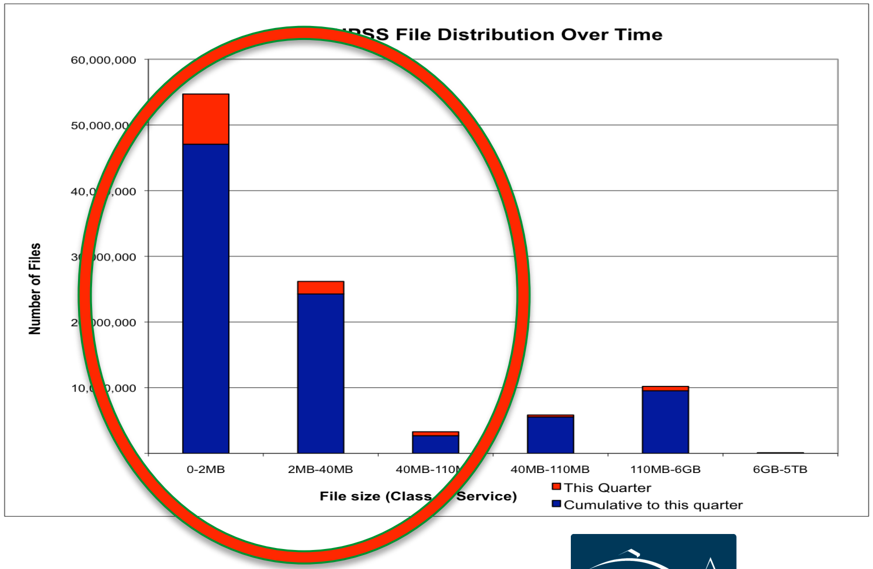
Optimizing Time to First Byte (Part 1)

- Until another strategy proves viable (e.g. aggregation in HPSS v7, partitioning + fast locate on large tapes), NERSC still needs both a fast access and capacity tape drive.
- We also purchase disk and aim to keep all “small” files on disk forever
- 9840D fast access tape 30 seconds to first byte, 75GB native capacity per tape
- T10KB/C capacity tape 1 minute to first byte, 1TB – 5TB native capacity per tape

94% of data on capacity tape



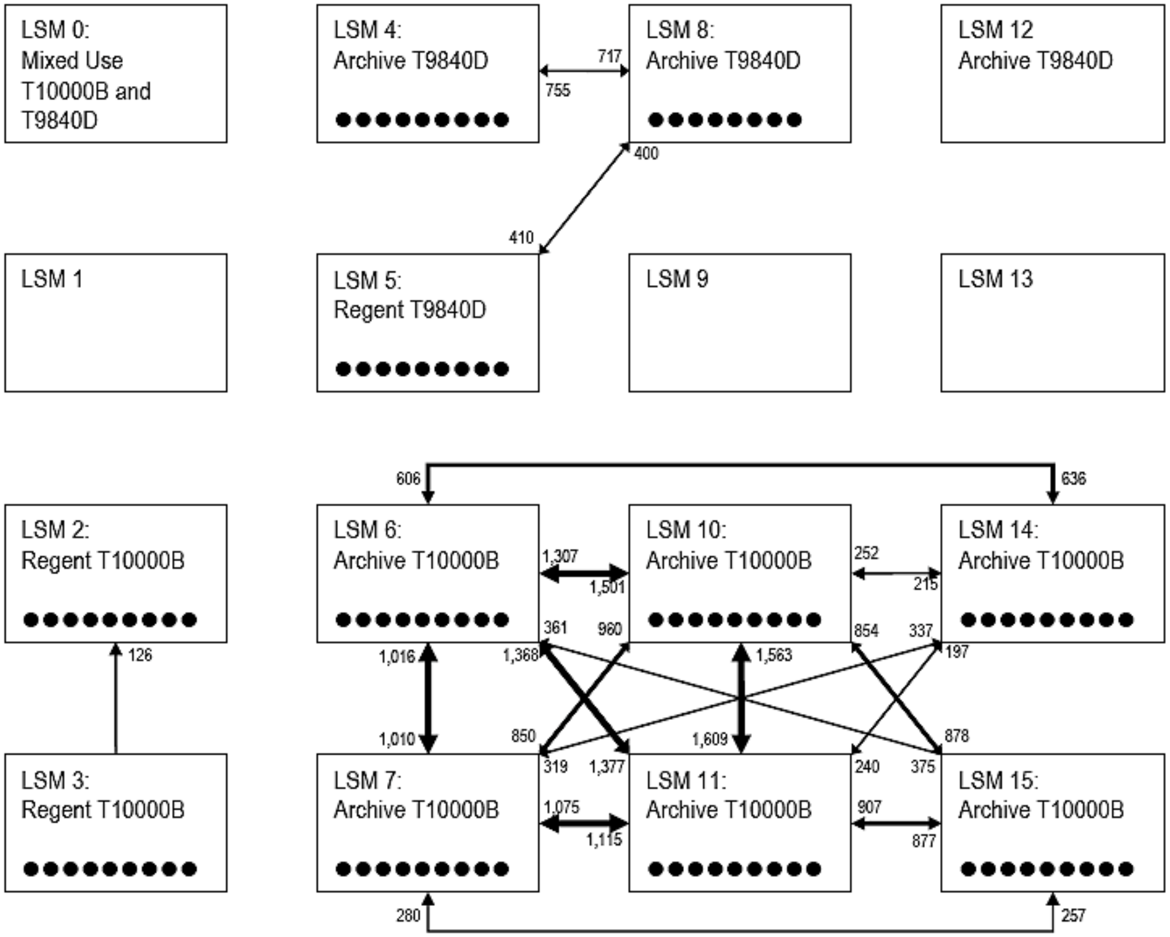
83% of files on fast access tape



Optimizing Time to First Byte (Part 2)

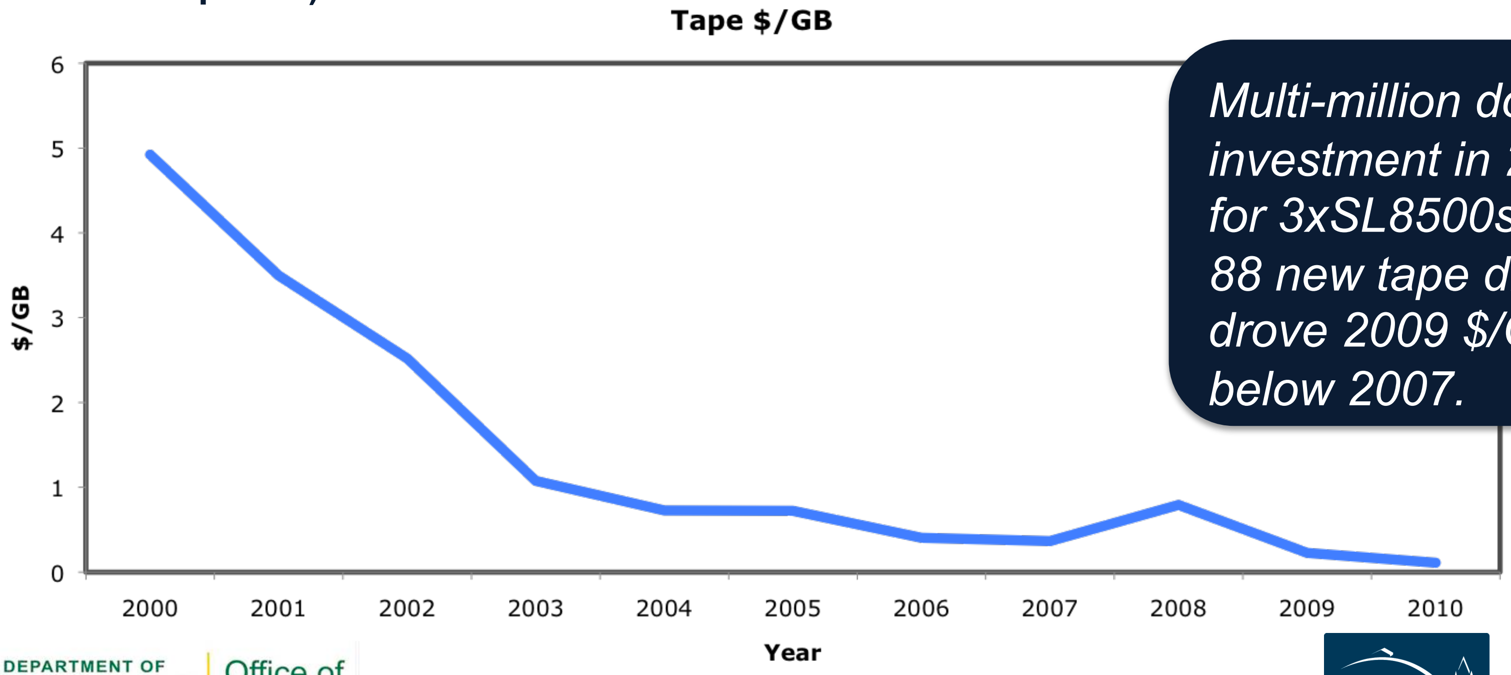
- **Minimizing library hot spots**
  - Easier on hardware
  - Handling peak demand efficiently
- **Optimizing tape to drive locality**
  - Minimize cartridge movement

Chart 23: LSM to LSM Movements With More Than 100 Tapes Moved


LSM 0: Mixed Use T10000B and T9840D
LSM 4: Archive T9840D ●●●●●●●●●
LSM 8: Archive T9840D ●●●●●●●●
LSM 12 Archive T9840D
717
755
400
410
LSM 1
LSM 5: Regent T9840D ●●●●●●●●●
LSM 9
LSM 13
606
636
LSM 2: Regent T10000B ●●●●●●●●●
LSM 6: Archive T10000B ●●●●●●●●●
LSM 10: Archive T10000B ●●●●●●●●●
LSM 14: Archive T10000B ●●●●●●●●●
1,307
1,501
252
215
361
960
854
337
126
1,016
1,368
1,563
197
1,010
850
1,609
878
319
1,377
240
375
LSM 3: Regent T10000B ●●●●●●●●●
LSM 7: Archive T10000B ●●●●●●●●●
LSM 11: Archive T10000B ●●●●●●●●●
LSM 15: Archive T10000B ●●●●●●●●●
1,075
1,115
907
877
280
257

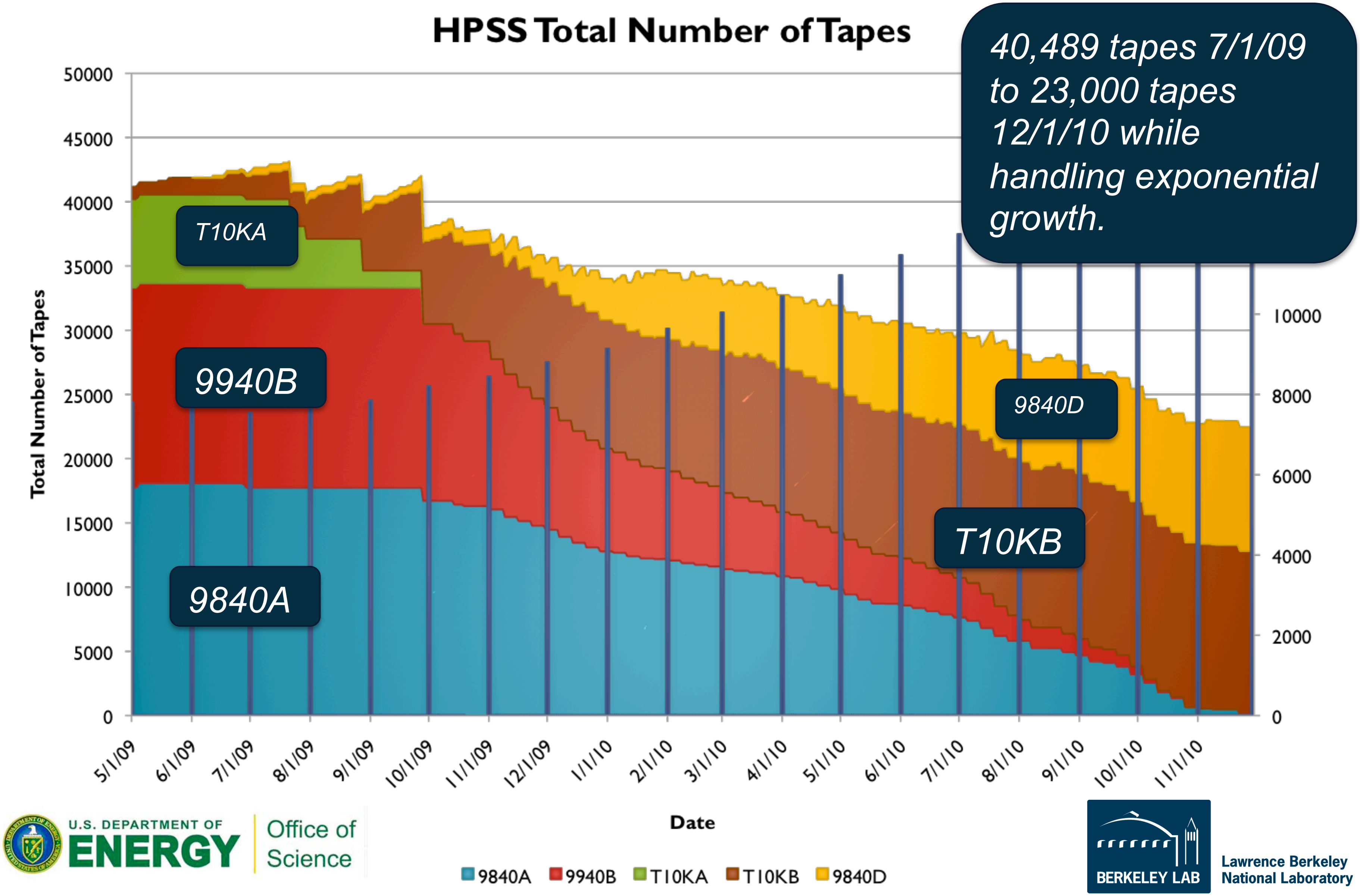
Early Adoption of New Capacity

- **Early adoption of new tape capacity provides operational savings immediately upon use because more data fits on similar cost cartridge.**
- **Small capacity (9840) media reuse**
  - **Bought only 1,500 9840 tapes since 2005**
  - **Migration to higher capacity media continues freeing up tapes that are rewritten at 3.5X previous capacity**
  - **As of 2009, $/GB is about 20x less than disk solutions at our site (includes all costs except staff and power).**



Multi-million dollar investment in 2008 for 3xSL8500s and 88 new tape drives drove 2009 $/GB below 2007.

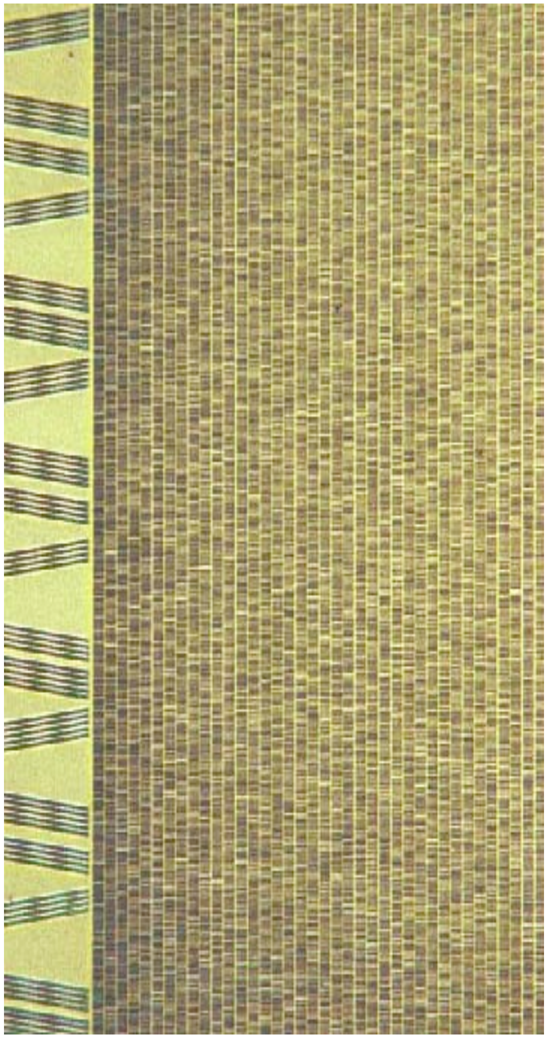
Migrating Data from Old to New



40,489 tapes 7/1/09 to 23,000 tapes 12/1/10 while handling exponential growth.

Actual Reliability of Enterprise Tape

- **We read all data on 40,489 tapes**
  - 6,859 T10KA (up to 2yrs old)
  - 15,572 9940B (up to 8yrs old)
  - 18,058 9840A (up to 12yrs old)
- **We found 36 tapes that had some data that couldn't be read.**
  - 24 9840A, 8 9940B, 4 T10KA
  - One of those had 558 files and couldn't be mounted.
  - Two others had 136 and 43 files that couldn't read, the remainder had less than 6 with most having only 1.
- **0.0009% error rate for tape cartridges or 99.9991% with 100% readable data.**
- **But wait! It's not the whole cartridge, the unreadable data was contained in 850 files (84.6 M total) representing 3.0 TB of data (8,056 TB total).**
- **0.00001% error rate for files or 99.99999% of files with 100% readable data.**
- **Unreadable data is normally in one or two blocks of data (250-500MB of data) with remainder of file readable, but we don't recover partial files unless user requests.**



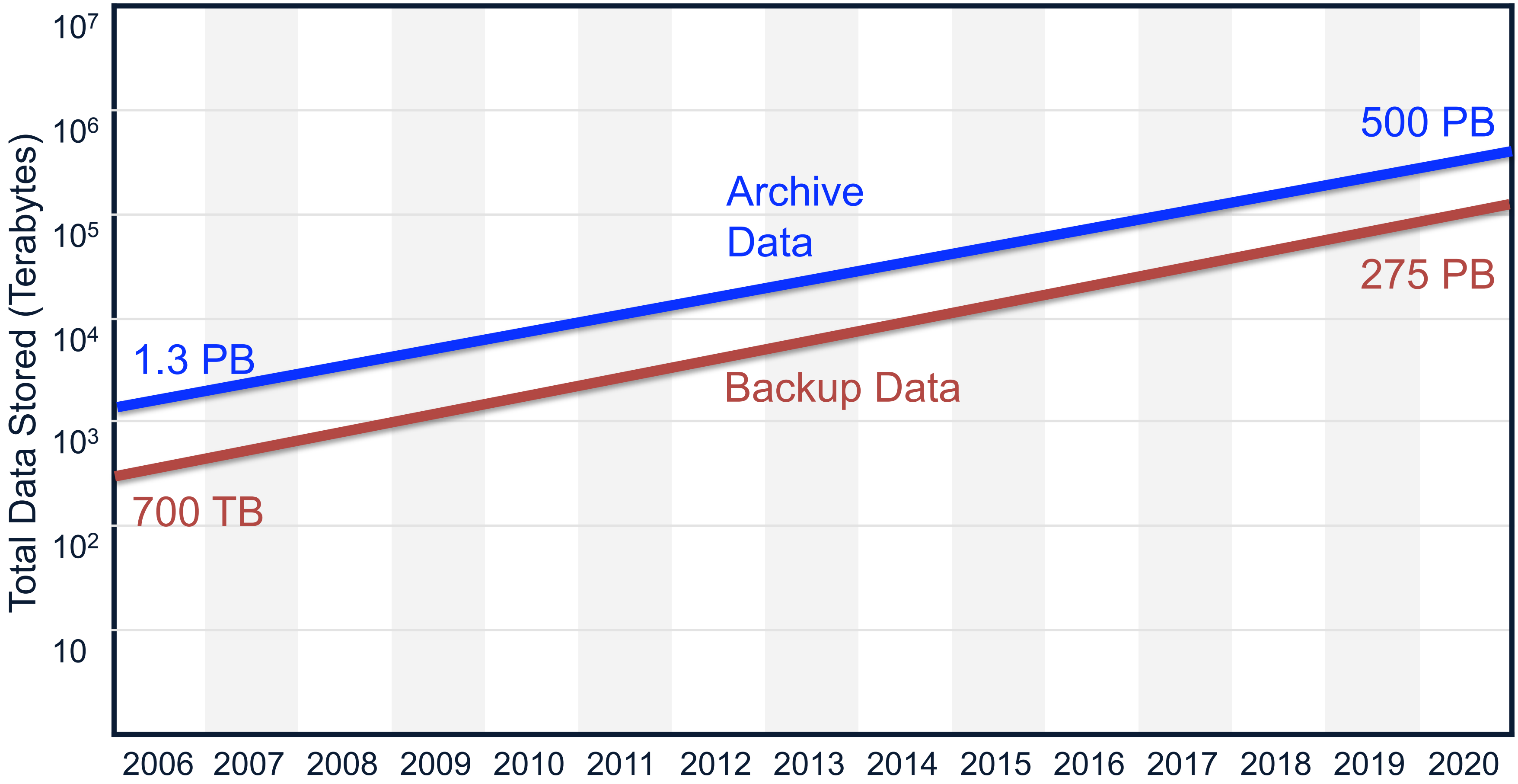
Our Experience Shows Enterprise Tape is Reliable

The data migration (7/09 – 12/10) involved reading 22,065,763m of tape, about the distance of flying San Francisco to Tokyo to Paris to Nova Scotia.

North America
Europe
Asia
North America
Atlantic Ocean
Africa
South America
Indian Ocean
Australia
Pacific Ocean

Unreadable data resided in at least one block of 850 files. These files represent 178m of tape, approximately the length of two Boeing 777 jets (70m) or half the length of most cruise ships (350m).

Our Tape Roadmap to Exascale Archival Storage



- **Annual growth projected at 50% (historical)**
- **Need 2X capacity tape drives every two years to handle**

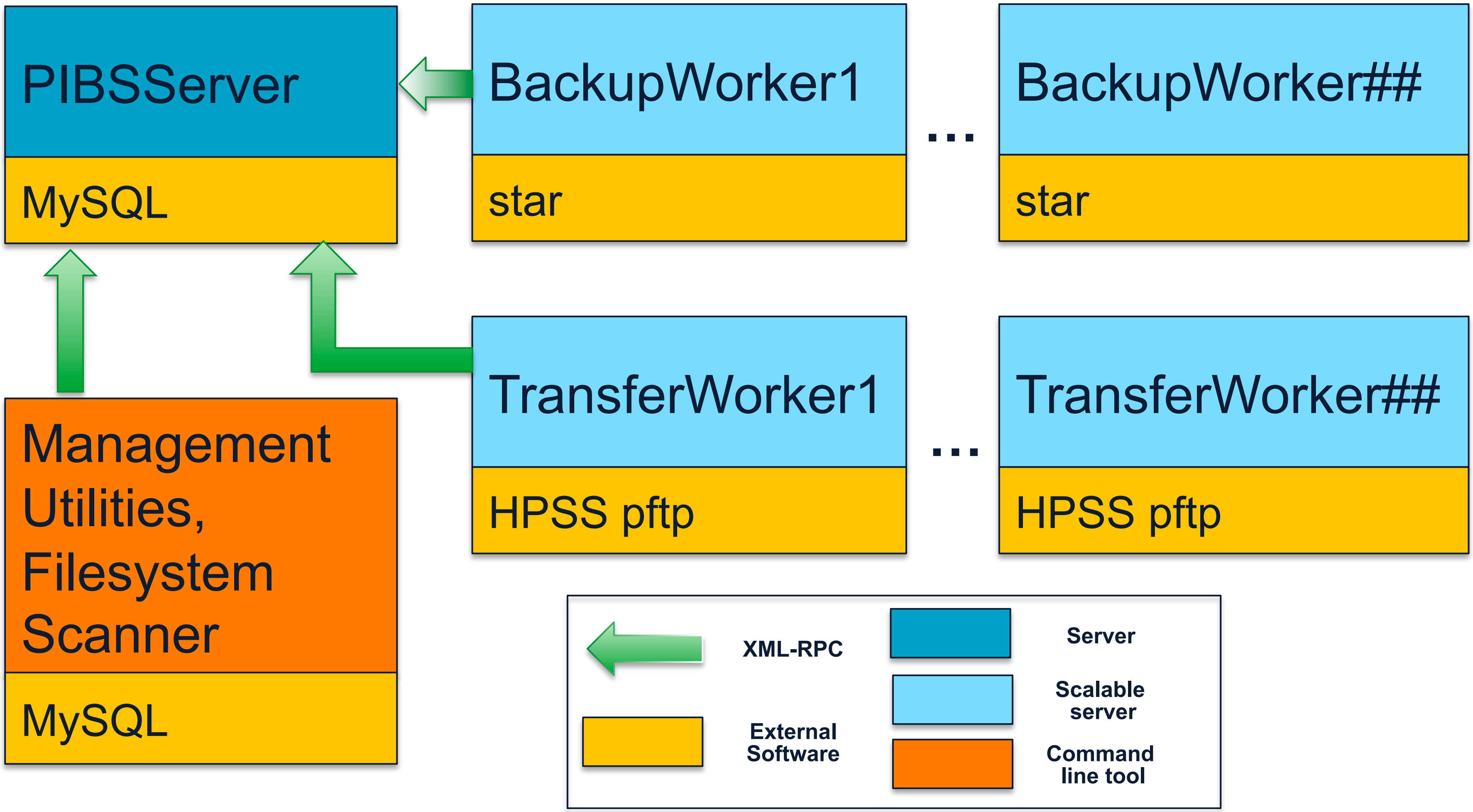
A Practical Exascale Storage Story Using Tape

- **The year is 2020 and the archive has 775 PB of data**
- **Assuming tape drive capacity doubling every 2 years and bandwidth doubling every 4, we end up with:**
  - BaFe media can theoretically hold 64TB of data, so assume no media formulation change (unrealistic)
  - 80 TB tape cartridge, for NERSC at 40% compression = 112 TB/tape
  - Just under 7,000 tapes holding this 775 PB, but would likely have significant amount of data on previous cartridge type (maybe 14,000 cartridges)
  - Annual growth in 2020 would require another 3,500 80TB tapes per year which at today’s cost is about $1M in media budget
  - Single drive bandwidth is near 1GB/sec
    - Will take 31 hours to read/write a 80TB tape at maximum bandwidth (1GB/s)
    - Assuming we have the same tape library and drive footprint as today, we will occupy about 6 tape libraries, and have approximately 200 tape drives
    - We typically can afford to dedicate about 25% of our drives to migrating data from old technology to new, which means about 50 drives.
    - Migrating 14,000 tapes with 25 drives reading, 25 drives writing at the max speed of the previous generation of drive (480MB/s) would take about 2 yrs.

Parallel Incremental Backup System (PIBS)

- **Developer: Matthew Andrews (mnandrews@lbl.gov)**
- **Backup HPSS (Tape component)**
  - Single transfers 1GB/sec read/write
  - Aggregrate bandwidth 3+GB/sec
  - Average daily IO of 20TB, with peak at 130TB
  - 40TB disk cache
  - 8 9840D and 18 T10KB tape drives
  - Largest file: 3.5TB
  - Oldest file: May 1995
  - Single biggest user is our group backing up GPFS file systems (daily)
    - Full backups involve 600+TB of data as of 6/1/2011
    - Incrementals are taken daily and typically involve 5-10TB of data

PIBS Software Components



PIBSServer

MySQL

BackupWorker1

star

…

BackupWorker##

star

Management Utilities, Filesystem Scanner

MySQL

TransferWorker1

HPSS pftp

…

TransferWorker##

HPSS pftp

XML-RPC

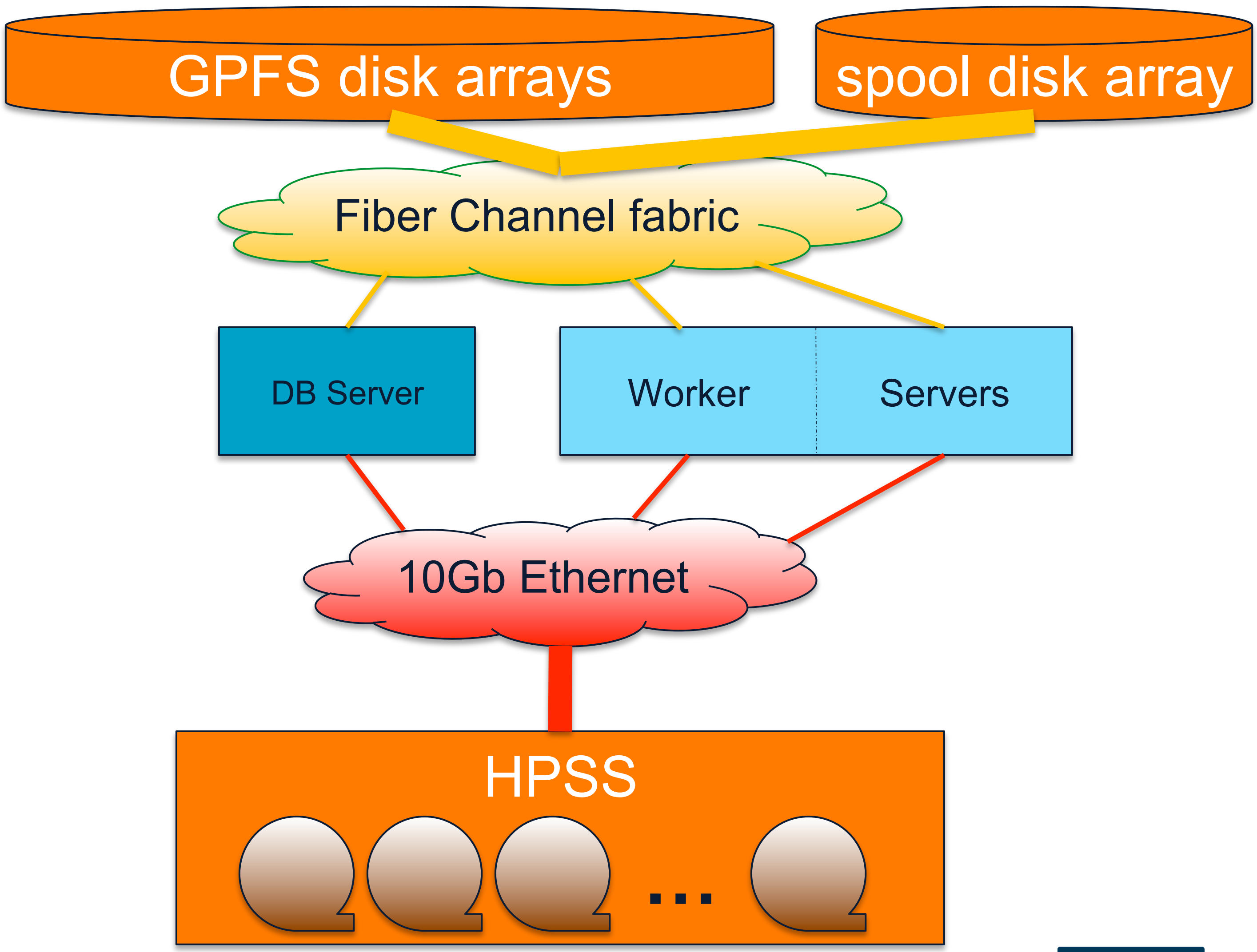
Server

Scalable server

External Software

Command line tool

PIBS Hardware Layout



PIBS Performance

- **Uses multi-threaded namespace walk and efficient GPFS inode scan interface.**
  - Project: 108 Million files
    - Name scan: 3 hours, 14 minutes (opportunity for future work)
    - Inode scan: 3 minutes
    - Sort: 15 minutes
    - Merge names and inodes and compute lists: 42 minutes(PERL – could be made faster if re-written in C)
- **Worker performance scales well with available hardware**
  - Full restore of /project
    - Moved over 500TB in 7 days
    - Peak performance with current hardware of over 100TB in a single day.

PIBS Scalable Restores

- Stood up new file system copy from restore off tape
- 500TB of data moved in 7 days from tape using 10 tape drives
- No impact to current production file system, other than requiring regular backups
- Full backup of our file system costs about $50,000 in reusable media. A second system on disk would cost at least $1 million if not several.

Summary

- Tape is more of an investment than a cost
  - Useable lifetime
  - Most competitive $/GB
- Early adoption of new tape capacity provides immediate operational savings
- With proper design, active tape system involves no shame
- Tape requires very little power & cooling
- Thus far, we have found that 99.9991% of our enterprise tape cartridges are 100% readable
- It provides us a scalable backup and restore solution for a multi-PB parallel file system
- The technology will enable Exascale archival storage