Brookhaven™
National Laboratory

U.S. DEPARTMENT OF
ENERGY

HPSS Status and Future Evolution

Tim Chou, Ognian Novakov, Iris Wu

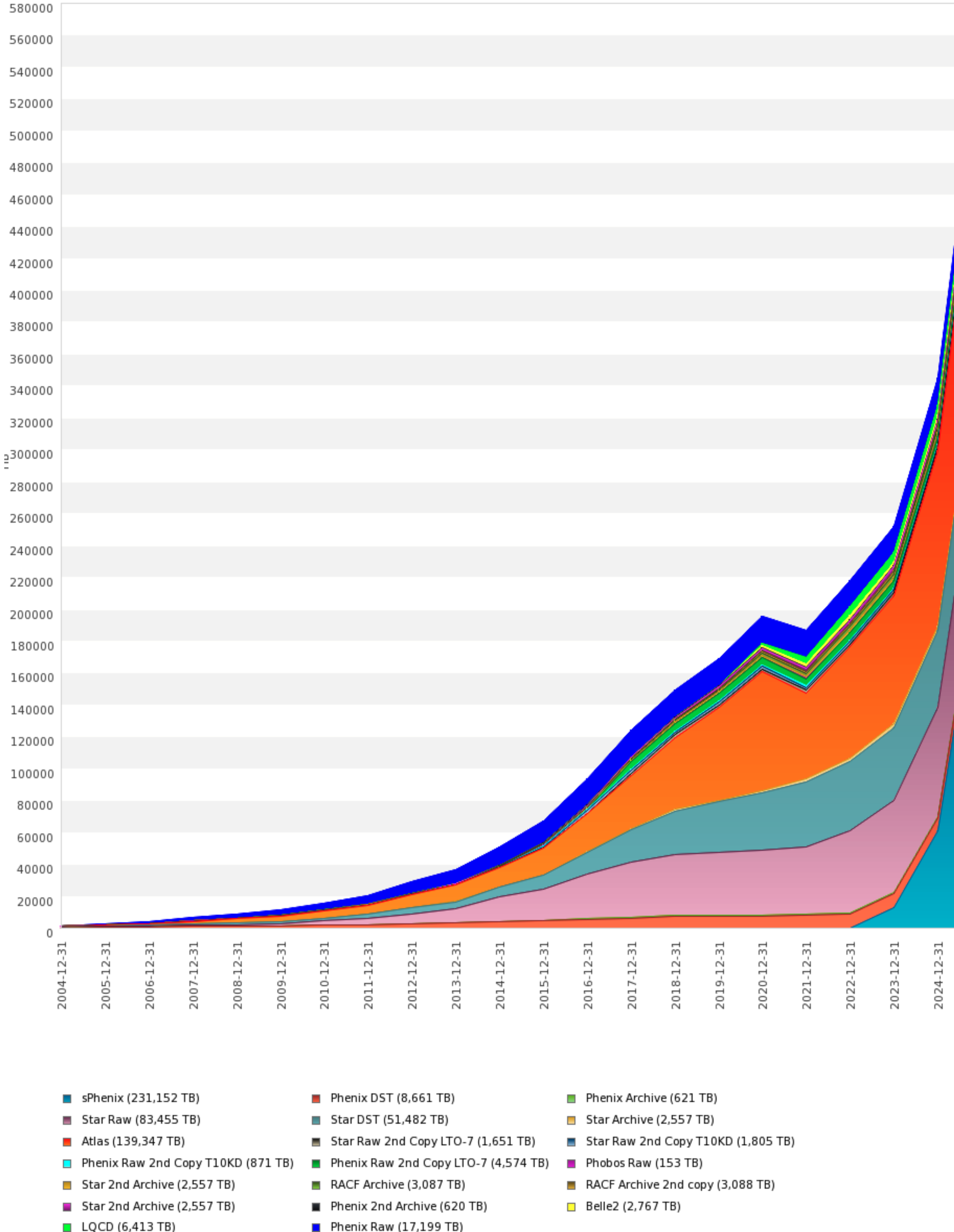
Dec 11th, 2025

SCDF, BNL

@BrookhavenLab

HPSS Statistics 2025

- HPSS 8.3.u20
  - Co-developed by DOE national labs & IBM
  - Supported by IBM
  - Support/license requires annual fee
    - Atlas and RHIC split it 50-50
- Archive data size, 567 PB (Dec, '25)
- Data Movers, 27 servers
  - US Atlas - 8,
  - Star/Phenix - 4,
  - sPhenix - 9,
  - Belle2 - 2,
  - LQCD - 2,
  - QCD/CAD - 1,
  - EIC - 1



Brookhaven National Laboratory

HPSS Statistics 2025

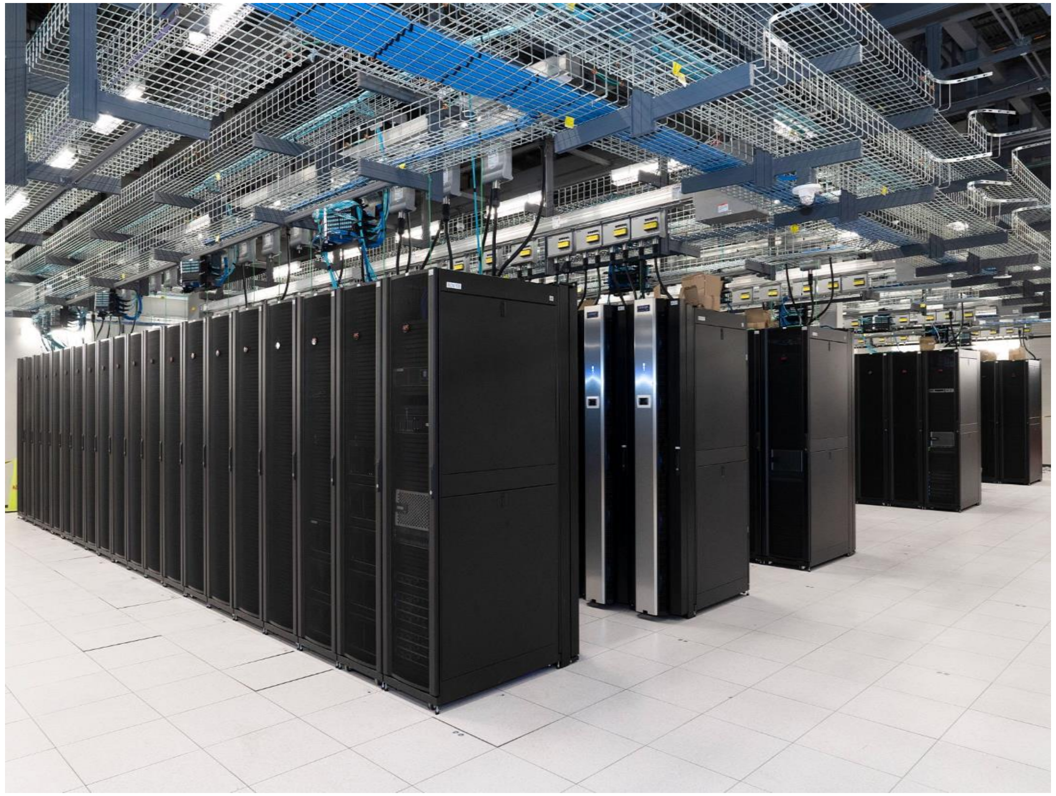
Tape libraries: 16
  - 9 Oracle 8500
  - 7 IBM TS4500

- Tape Drives, 308
  - LTO7 (6 TB) - 49
  - LTO8 (12TB) - 112
  - LTO9 (18TB) - 100
  - Misc. – 47
- Tape slots: 142,464
  - 85,576 on Oracle libraries
  - 56,888 on IBM TS4500

- Active tape volumes: 71,141

HPSS Statistics 2025

- Disk Cache 9.5 PB
  - Atlas – 1.2 PB
  - sPhenix – 5.9 PB
  - Star/Phenix – 1.2PB
  - Belle2 – 537 TB
  - LQCD – 537 TB
  - EIC – 60 TB
  - QCD/CAD - 60 TB

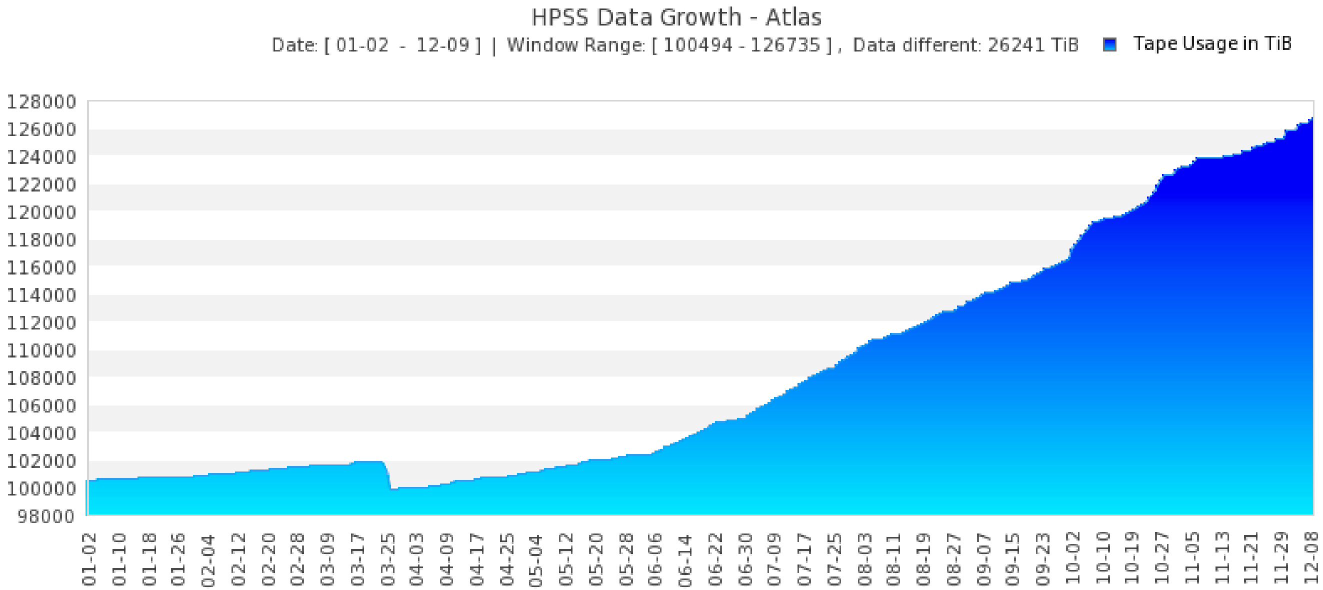


Brookhaven National Laboratory

Atlas Run25

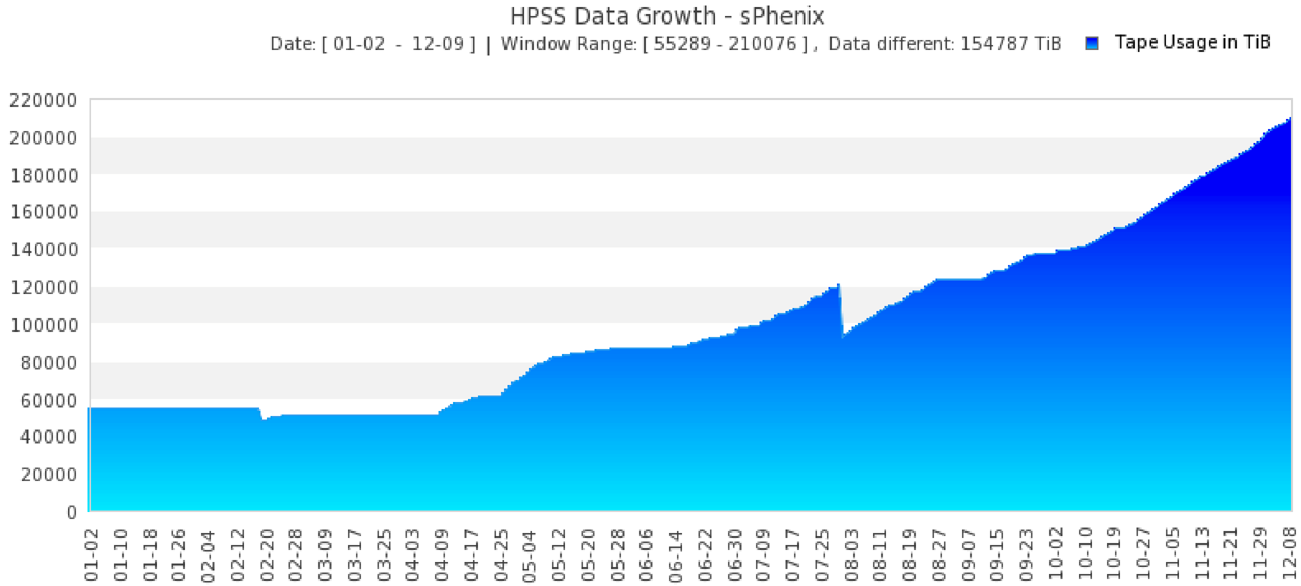
- Archive data size – 139.4 PB
- 39.9 PB (8,868,068 files) staged in 2025 (thru Dec 10th)
- 28.4 PB (9,072,616 files) injected in 2025 (thru Dec 10th)
- Atlas movers and gateways upgraded to RHEL8
- Repacking LTO6 tapes to LTO8 in progress

✓ Sustain 7.5 GB/sec



sPhenix Run25

- Archive data size – 231 PB
- 170.2 PB of data injected to HPSS (Thru Dec 10)
- Two TS4500 libraries added
  - Data written evenly across 4 libraries
- 12,491 total LTO9 tapes used
- Average file size 20GB
- Tools/monitoring plots added
- ✓ Sustain 24 GB/sec

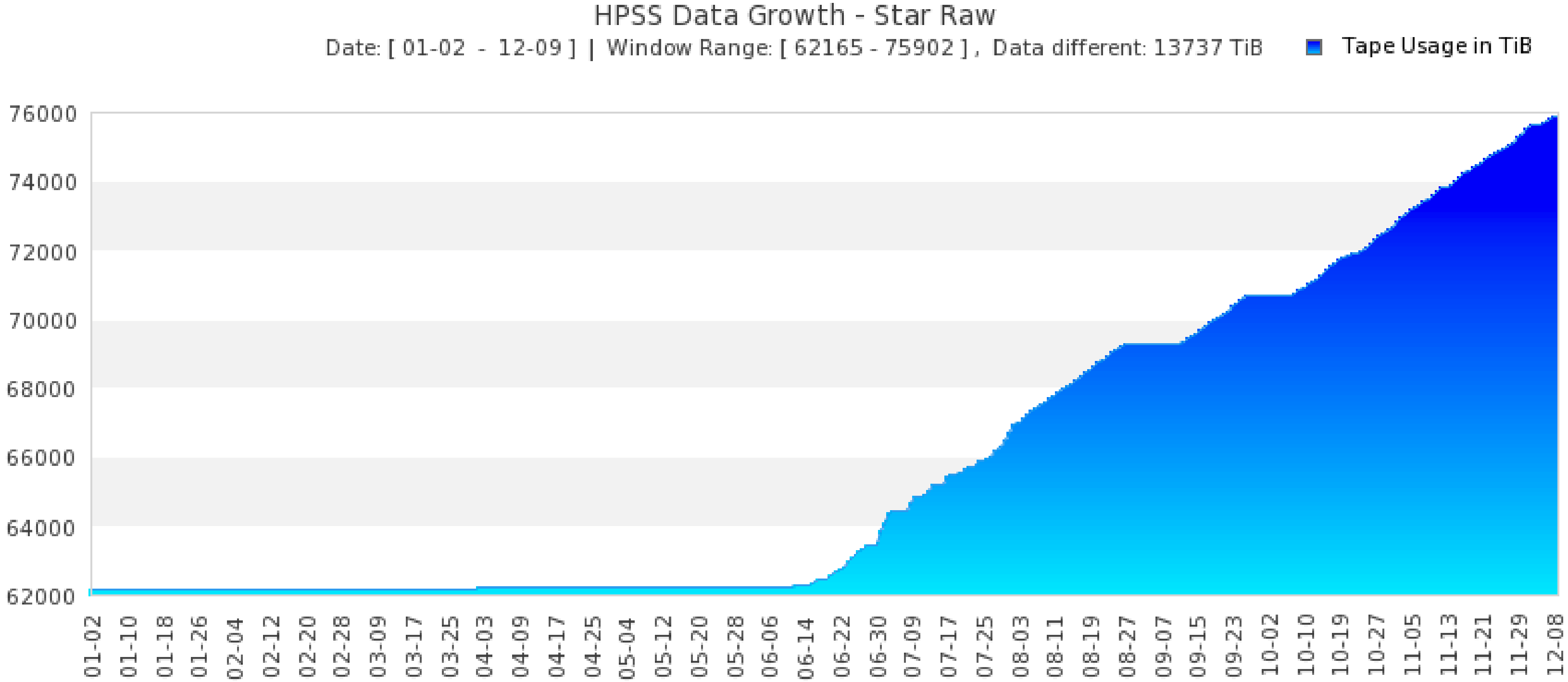


Brookhaven National Laboratory

Star Run25

- Archive data size:
  - Star Raw – 83.5 PB
  - Star DST – 51.5 PB
- Star Raw 15.1 PB injected (1/1/25 thru Dec 10th)
- Star DST 2.7 PB injected (1/1/25 thru Dec 10th)
- Gateways upgrade
- Disk cache upgrade

✓ Sustain 4 GB/sec

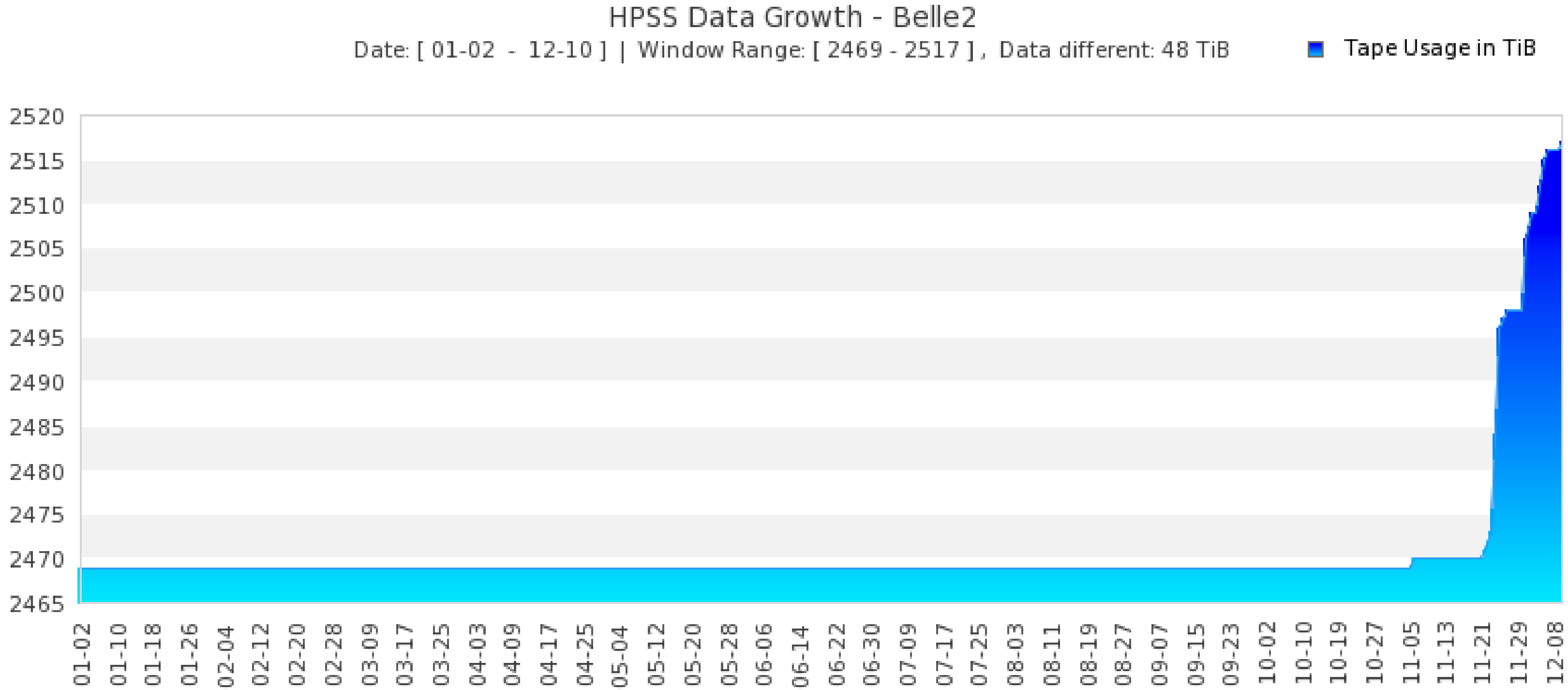


Brookhaven National Laboratory

Belle2 Operations

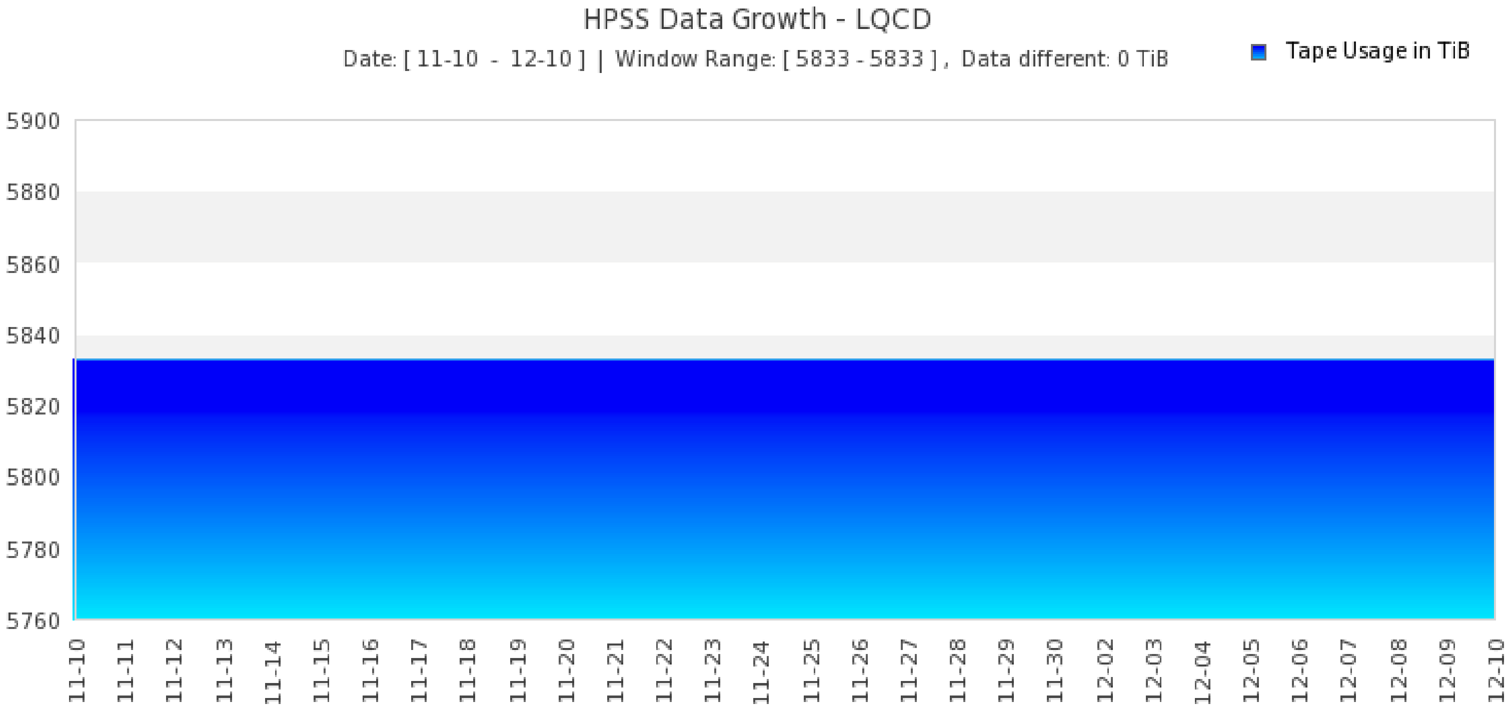
- 2.8 PB archive data
- 10 LTO8 drives
- 537 TB of disk cache
  - disk cache upgraded

✓ Sustain 3.5 GB/sec



LQCD Operations 2025

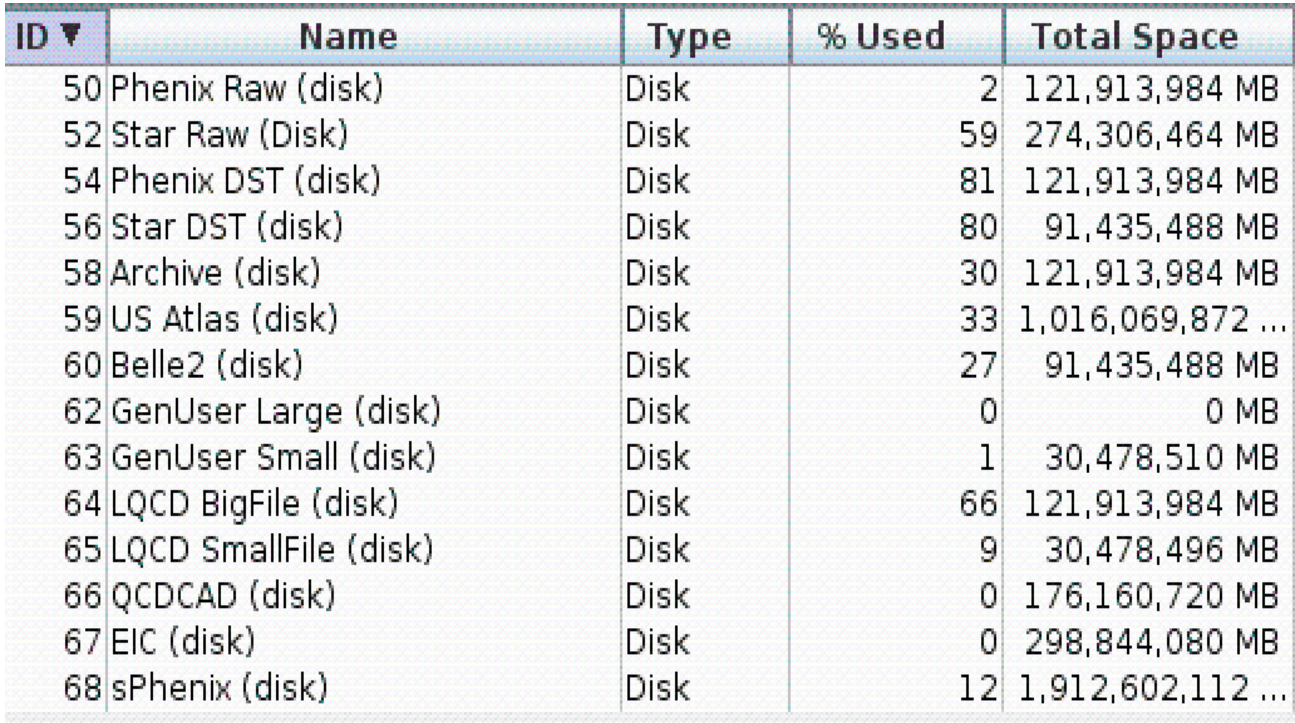
- 6.4 PB archive data
- 8 LTO8 drives
- 320 TB of disk cache
  - disk cache upgraded
- HPSSFS service


HPSS Data Growth - LQCD

Date: [11-10 - 12-10] | Window Range: [5833 - 5833] , Data different: 0 TiB

HPSS Software Upgrades in 2025

- HPSS movers upgraded to RHEL8
- Movers and Gateways upgraded to RHEL8
- HPSS clients upgraded to RHEL8 and Alma9
- HPSSFS service for LQCD



ID ▼	Name	Type	% Used	Total Space
50	Phenix Raw (disk)	Disk	2	121,913,984 MB
52	Star Raw (Disk)	Disk	59	274,306,464 MB
54	Phenix DST (disk)	Disk	81	121,913,984 MB
56	Star DST (disk)	Disk	80	91,435,488 MB
58	Archive (disk)	Disk	30	121,913,984 MB
59	US Atlas (disk)	Disk	33	1,016,069,872 ...
60	Belle2 (disk)	Disk	27	91,435,488 MB
62	GenUser Large (disk)	Disk	0	0 MB
63	GenUser Small (disk)	Disk	1	30,478,510 MB
64	LQCD BigFile (disk)	Disk	66	121,913,984 MB
65	LQCD SmallFile (disk)	Disk	9	30,478,496 MB
66	QCDCAD (disk)	Disk	0	176,160,720 MB
67	EIC (disk)	Disk	0	298,844,080 MB
68	sPhenix (disk)	Disk	12	1,912,602,112 ...

sPhenix Procurements & Installations

- Added Two new 9-frame IBM TS4500 libraries
  - 10,050 slots in each library
  - Total of 4 TS4500 libraries
- 36 LTO9 drives added
  - Total of 100 LTO9 drives
- 9 new data movers
- 9 NetApp disk arrays (5.9 PB)
- Monitoring tools and graphs
- Designed to sustain 20GB/sec
  - 24GB/sec verified

Data Repacks in 2025

- Data repacks of LTO5 and LTO6 to LTO8 media
- US Atlas 3,900 LOT6 tapes
- Phenix 4,700 LTO5 & 2,025 LTO6
- Star 8,850 LTO5 tapes
- Repacking data to new tape technologies allows the retirement of old tape resources and reduce the maintenance costs.
- Retirement of Oracle libraries will save significant service fees annually

Data repacks to CTA from HPSS -- A hypothetical scenario

- No decision has been made at BNL to migrate away from HPSS

- If repacking 600 PB on HPSS to LTO10
  - 300MB/sec/drive on LTO7 drive
  - 63 drive-year to read
  - Assuming 50% overhead for repacking
    - 95 drive-years to repack
  - HPSS and CTA must co-exist till it ends
  - 17K LTO10 tapes needed, cost of $2 million
  - 10 data movers with disk cache needed, cost of $1 million
  - Replacement of LTO7, LTO8 and LTO9, cost of $1 million
  - A new TS4500 library for Belle2, LQCD and other users, cost of $0.5 million

Tasks in the near future

- Preparation for HPSS upgrade to 11.x from 8.3
- HPSS Batch server upgrade
- Data repacks to newer tape technologies
  - LTO6 to LTO8
  - LTO7 to LTO9
- Retirement of Oracle tape libraries and tape drives
- New HPSS test environment
- Explore new technologies
  - Including new features on HPSS and other software
  - New HW technologies

Thank you!

Q & A…

HPSS interfaces for flexibility

HPSS Storage Broker for datasets container

HPSS GHI for IBM Storage Scale automation

- Policy driven space management, and
- Scale-Out Backup-Restore (SOBAR) support

HPSS S3 Proxy for on-prem cloud workloads

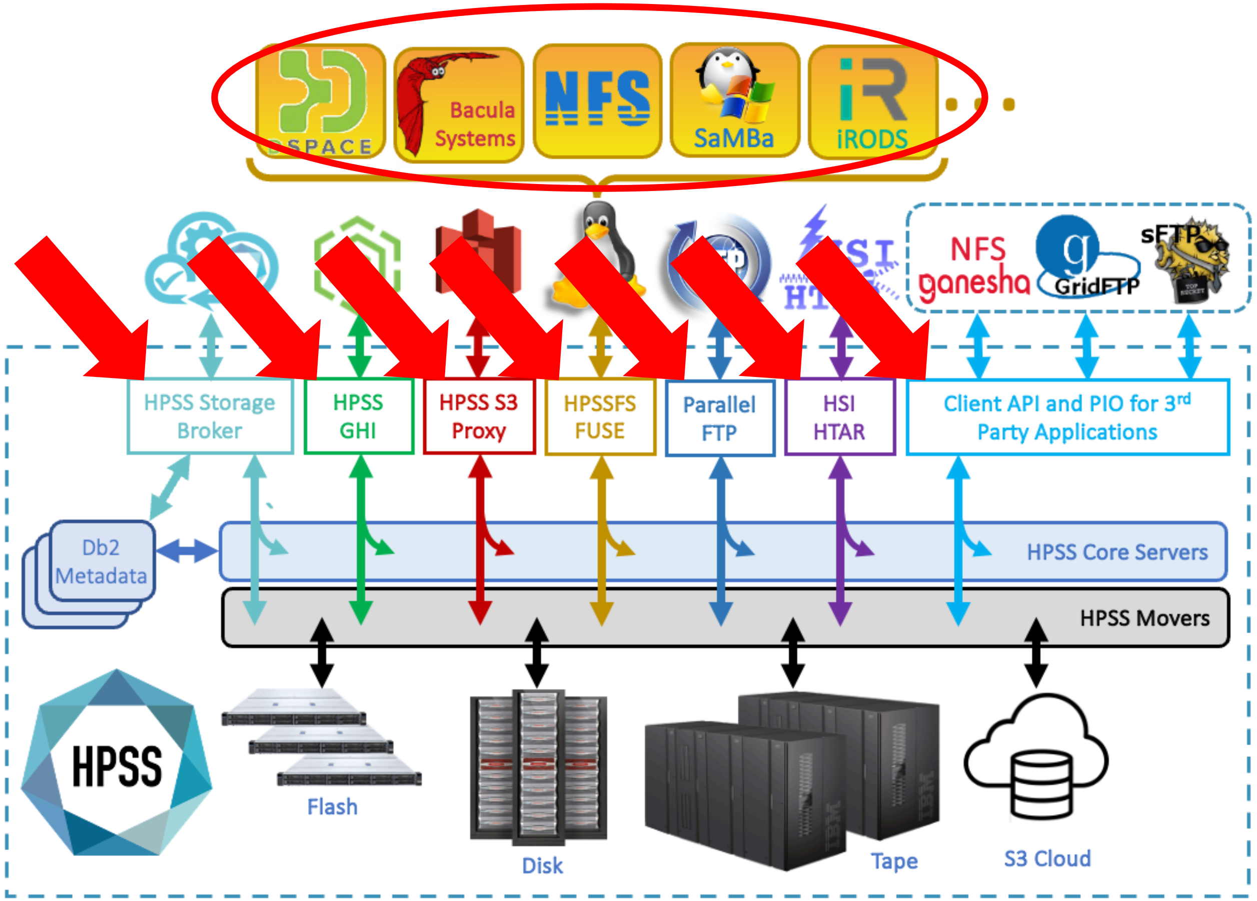
HPSSFS FUSE for tape-friendly mount points

- Export Linux mount point for third party tools:
  - D-Space and Globus for sharing digital repositories
  - Bacula Enterprise for site backups
  - NFSv4 export for non-Linux users
  - Open Samba providing HPSS on Windows

Standard and Parallel FTP

HSI and HTAR are HPSS high performance tools

C and Python client API – for programmers



Brookhaven National Laboratory

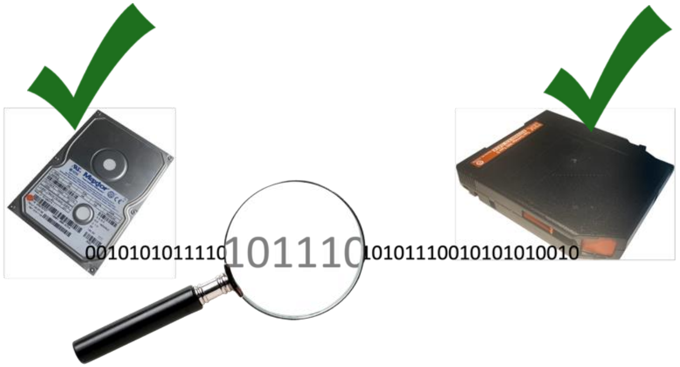
Best of breed features

Best of breed for tape data integrity

- HPSS uses file and block-level checksums
- Read-after-write verifies data on disk and tape
- Corrupted data never makes it to tape
- High-performance and efficient data-re-validation

Battle Against

SILENT DATA CORRUPTION



Export a copy of HPSS data to LTFS tape for ‘portability on-demand’

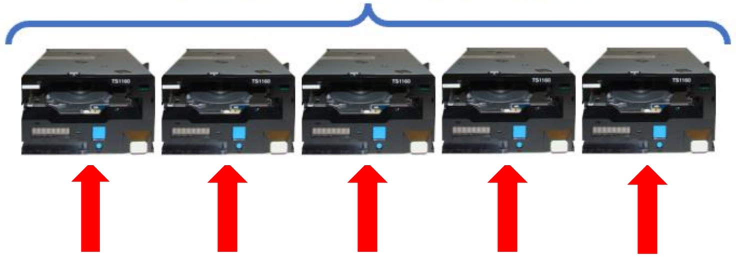
HPSS
LTFS
Linear Tape File System

Tape stripes with rotating parity to cut redundant tape costs, and to improve data durability and tape library durability with RAIT and RAIL

- 4+P HPSS RAIT cuts redundant tape costs by 75% over dual-copy
- 4+P HPSS RAIT stripe @ 1.6 GB/s to write a 1 TB file in 11 minutes

Up to 16-wide RAIT stripes

Five drives = 4+P RAIT or 3+PQ RAIT



Best of breed performance and efficient…

90-10 Rule – our experience

- 90% of files by count take 10% of the capacity
- 10% of files by count take 90% of the capacity

HPSS achieves near-native tape performance with small and large files

- HPSS uses fewer tape drives to meet requirements
- Small file aggregation and buffered tape marks to stream small file tape writes

HPSS enables high-performance tape recalls

- HPSS automatically groups files on tape by policy
  - Reduces tape mounts and tape seeks on future recalls
- Full aggregate recall enable streaming tape reads for small files
- Fully integrated with modern tape drive feature called “RAO” (recommended access ordered)
  - Reduces tape seeks, thereby improving recall efficiency and tape cartridge longevity

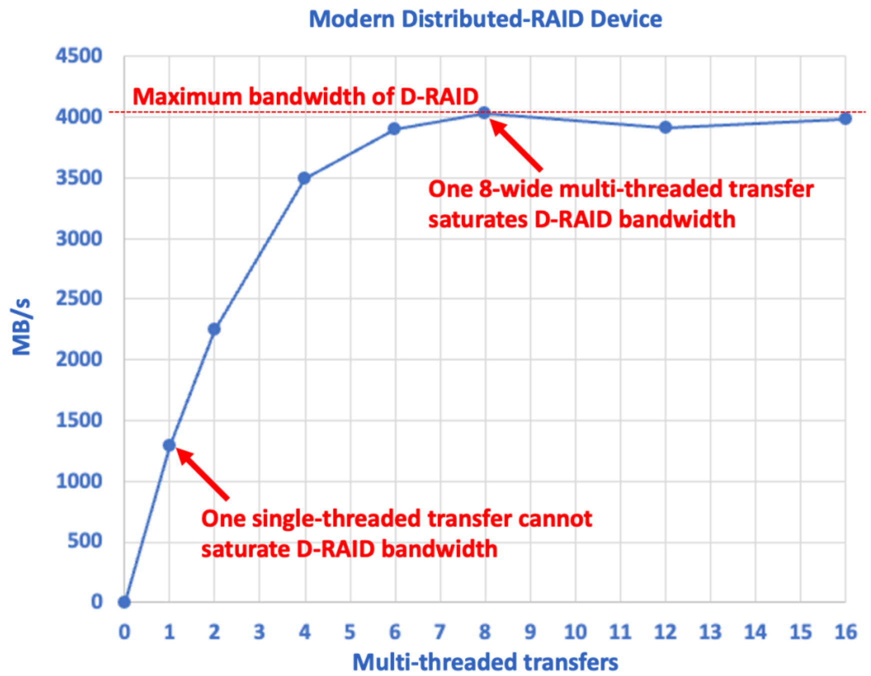
90-10 Rule

More Data

Less Hardware

Best of breed performance and efficient…

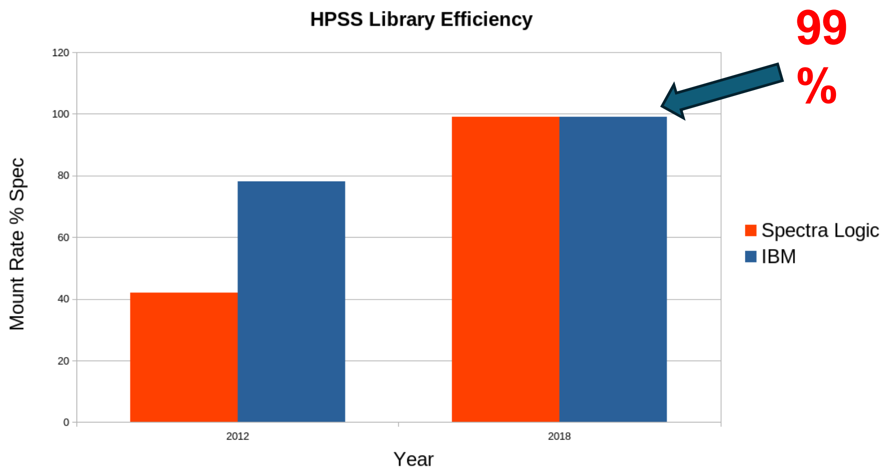
Multi-socket disk transfers



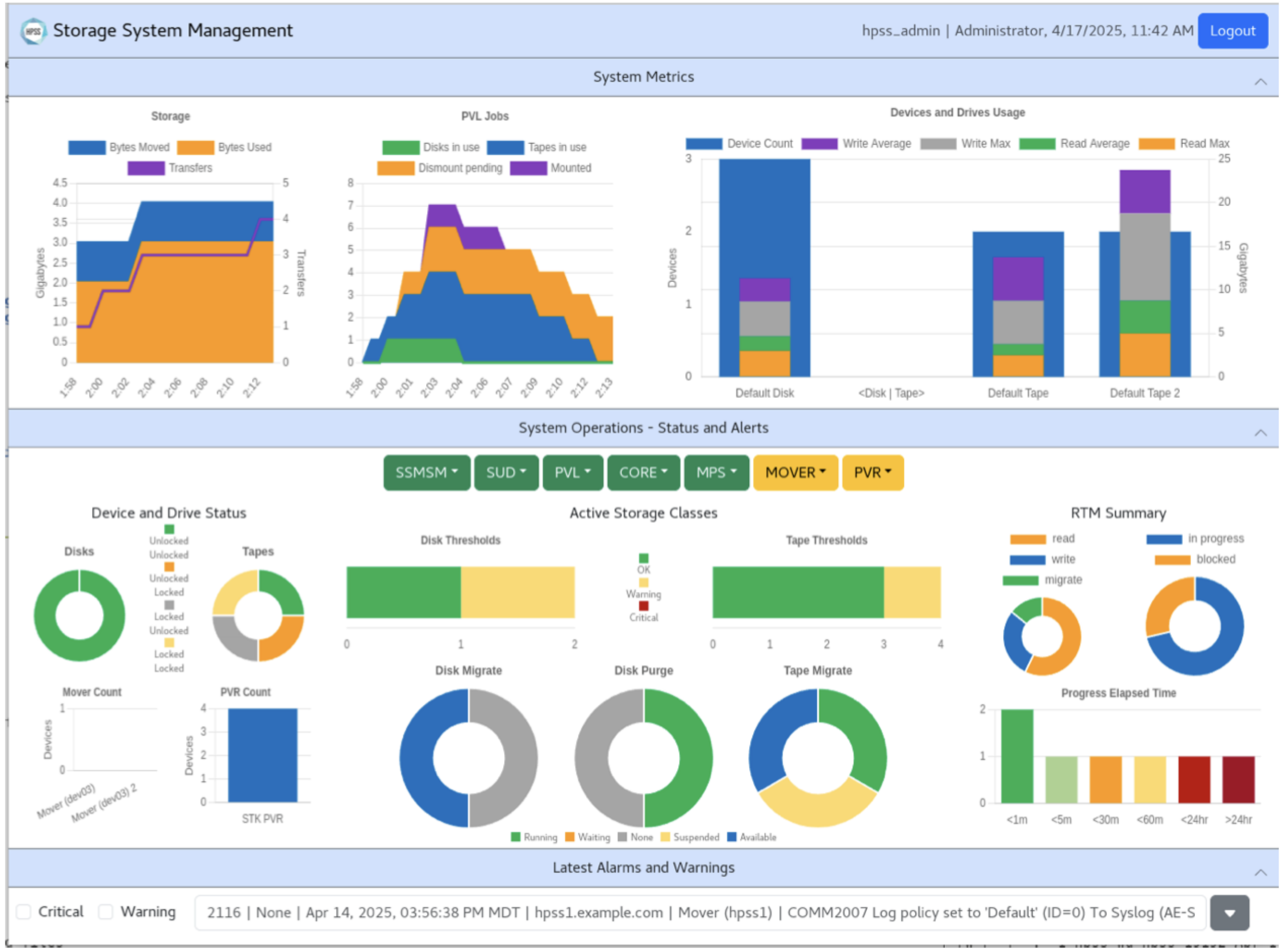
Optimized tape movement within the library

Library striping, tape striping, and RAIT

Achieve native tape read and write performance



Best of breed performance and efficient…



Storage System Management

hpss_admin | Administrator, 4/17/2025, 11:42 AM

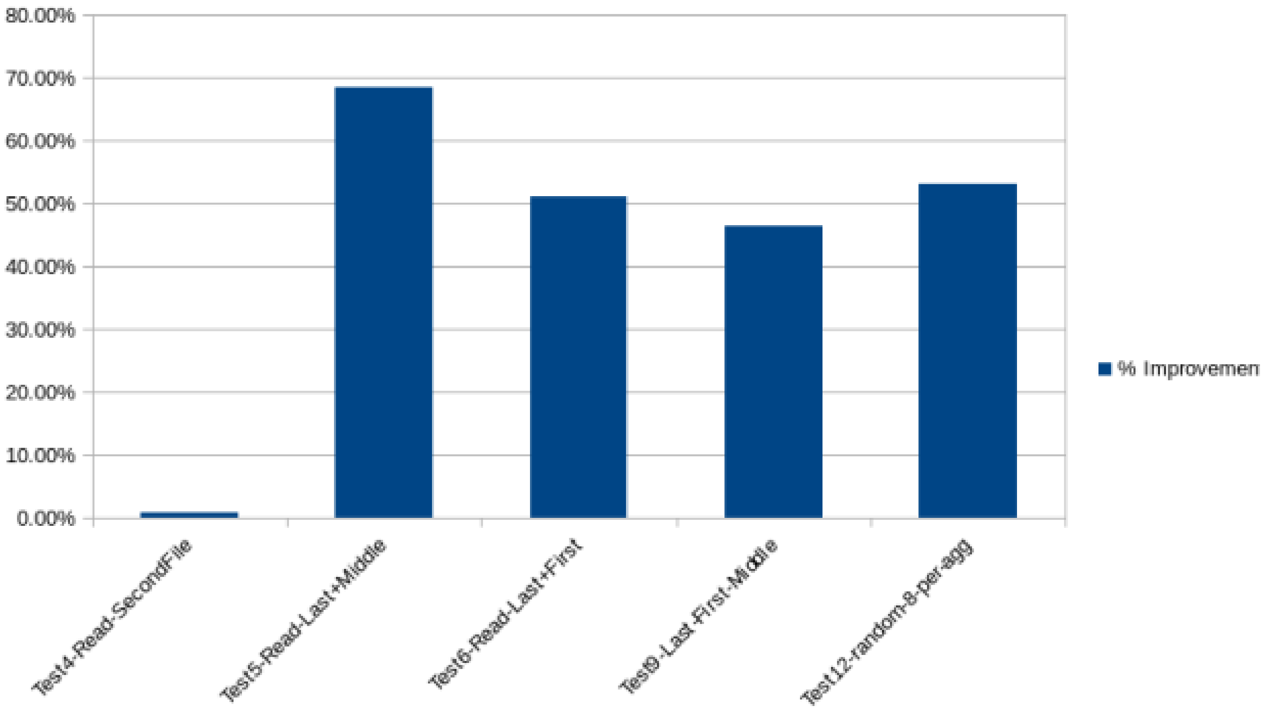
2116 | None | Apr 14, 2025, 03:56:38 PM MDT | hpss1.example.com | Mover (hpss1) | COMM2007 Log policy set to 'Default' (ID=0) To Syslog (AE-S

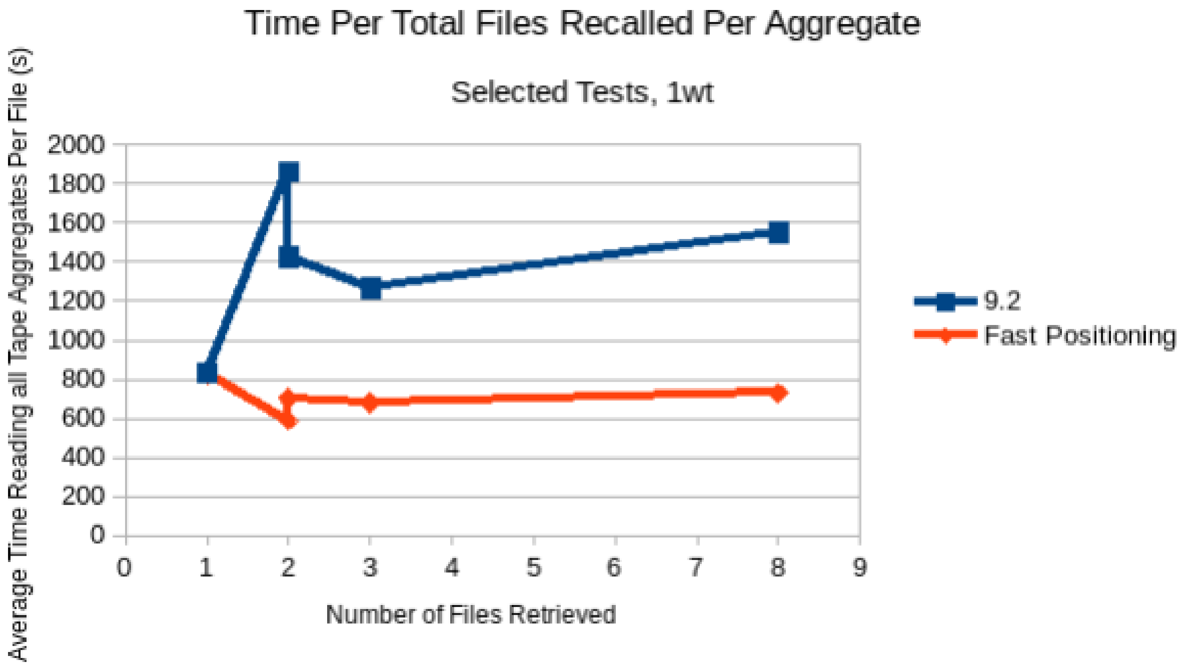
Brookhaven National Laboratory

Mover Improvements

FAST Locate

- HPSS will position directly to the data, bypassing header reads





Brookhaven National Laboratory

HPSS scales incrementally

Add HPSS Core Server and Db2 Off-Host Nodes to scale file transaction performance

Add HPSS Metadata Storage to scale file count capacity and Db2 performance

Add Disk Cache Storage Units to scale disk cache bandwidth and capacity

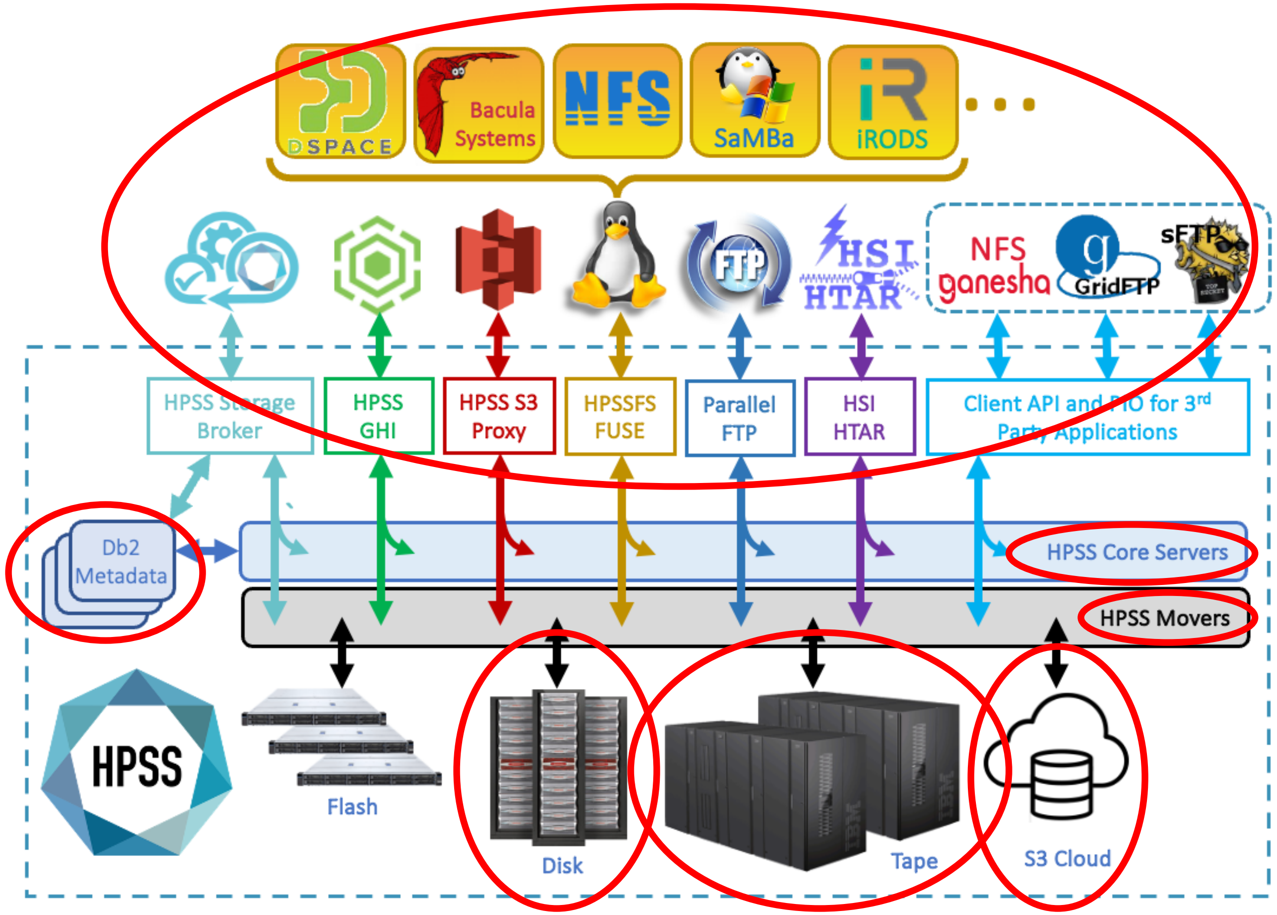
Add Tape Libraries to scale tape capacity, tape drive count and tape mount rate

Add HPSS Disk Movers to Scale Disk Cache Performance

Add HPSS Tape Movers to scale tape drive count and tape bandwidth

Add Client Nodes to scale HPSS client connections and client performance

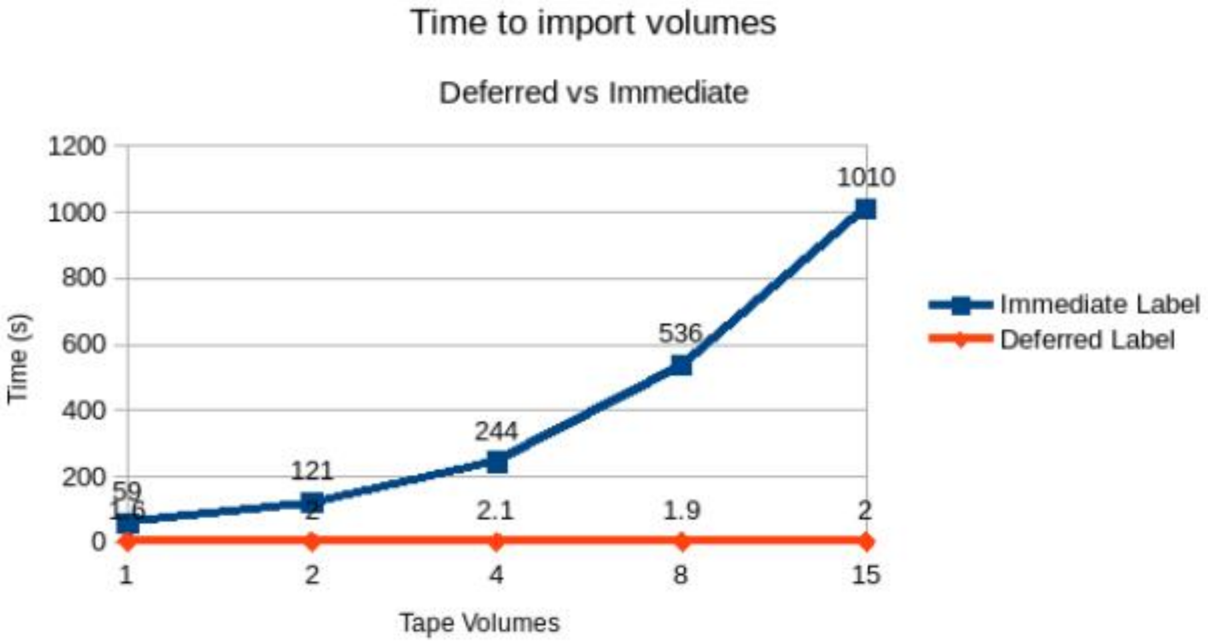
HPSS cloud-tiering to send HPSS files to clouds (with AWS Glacier support)



Brookhaven National Laboratory

Defer Tape Labeling Test Results

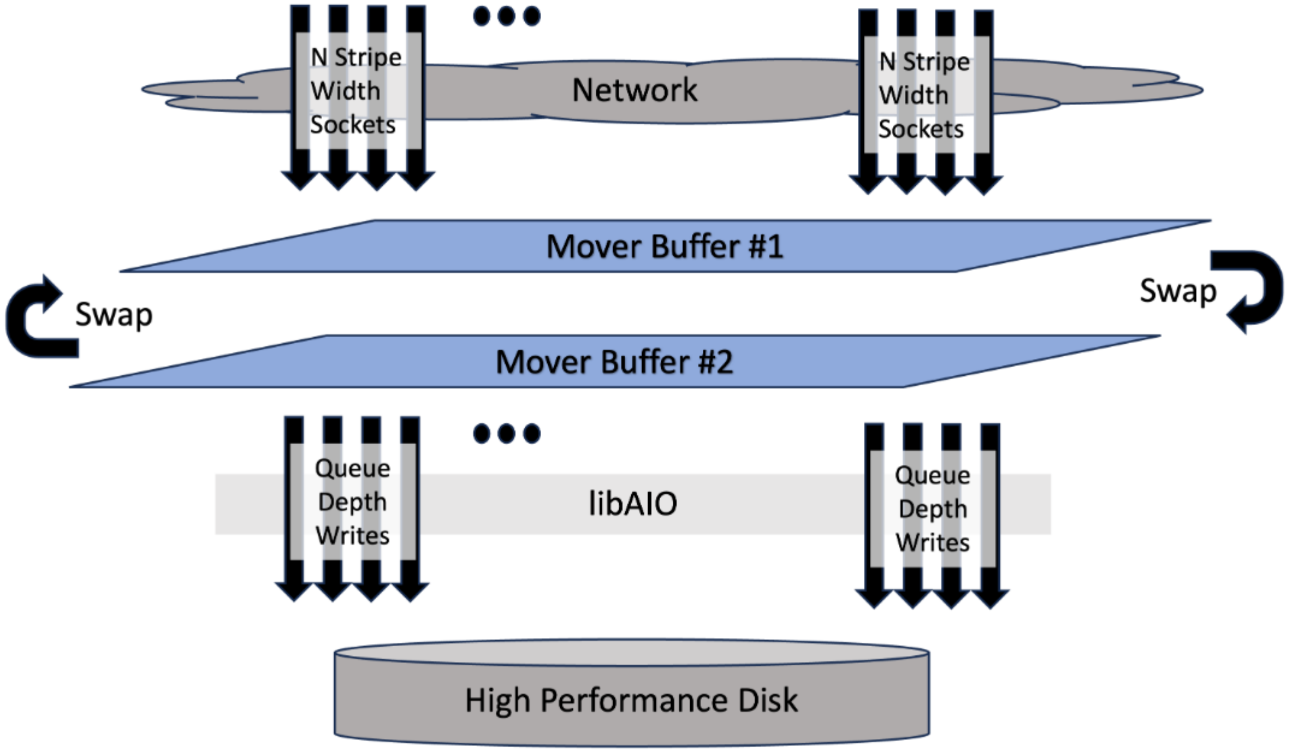
- This was done with JC media and TS1140 drives.
- Deferred tape labeling added an average of 10 seconds to the load time on the first write.
- Deferred tape imports are a metadata only operation that takes very little time.
- Overall savings of **50 seconds** of drive time **per cartridge**.
- HPSS no longer requires babysitting long running import operations.



Mover Improvements

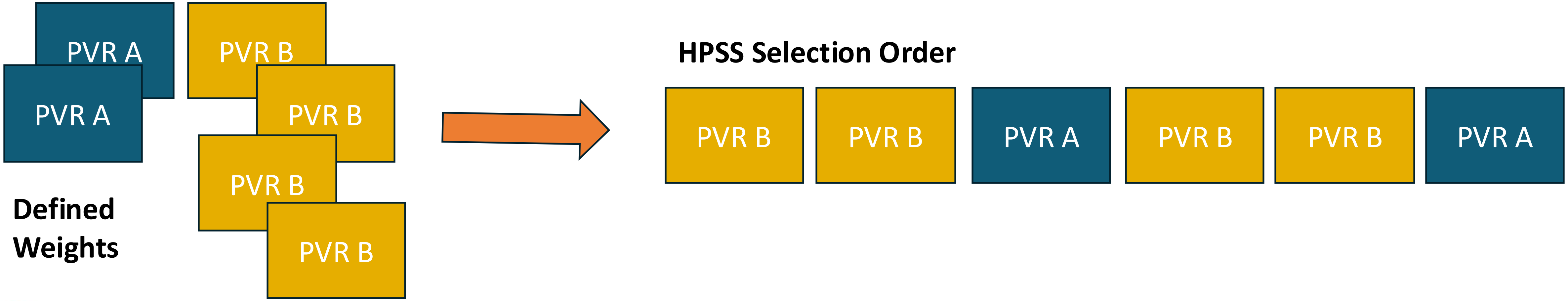
Multi-socket Transfers

- HPSS clients can request multiple sockets for a data transfer, improving bandwidth availability to disk



Balance New Tape VV Selection

- More on policy-based selection
  - Weighting scheme
  - For a given tape SC, each PVR has a weight
    - Higher weight = chosen more often
    - Lower weight = chosen less often



AI	Performanc	Security
Ease of Use	Best of Breed	Hybrid Cloud

HPSS roadmap – release cadence

Release	Date	Features and Goals
10.2	January 2023	Purge on migrate flag; Low overhead read interface (LORI); Optional ASAN-enabled versions of HPSS Servers; Disable Direct IO option; Simplify AWS tape storage gateway support.
10.3	September 2023	dumpv_pvl can display the PVR name; Migration and Purge State Changes; New HPSS S3 Proxy; Read Queue APIs; Persisted Read Queues; Updates to avoid server restarts; Added purge filters for file size and age.
11.1	August 9, 2024	Native Cloud Tiering replacing AWS tape gateway; S3 Proxy Glacier Support for optimized recall workloads.
11.2	February 2025	Updated HTAR standard (lixed limitations); Stronger User Authentication; Faster Per File Performance.
11.3	September 2025	Restful SSM, Native In-Flight Data Encryption; Containerization; ARM support for clients.
12.1	May 2026	Replication and Multi-Region Support.
12.2	January 2027	Security, Performance, Ease of Use, Best of Breed for Tape; Hybrid Cloud; AI