Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

HPSS @ SDCC

Tim Chou, Ognian Novakov,
Iris Wu, Justin Spradley

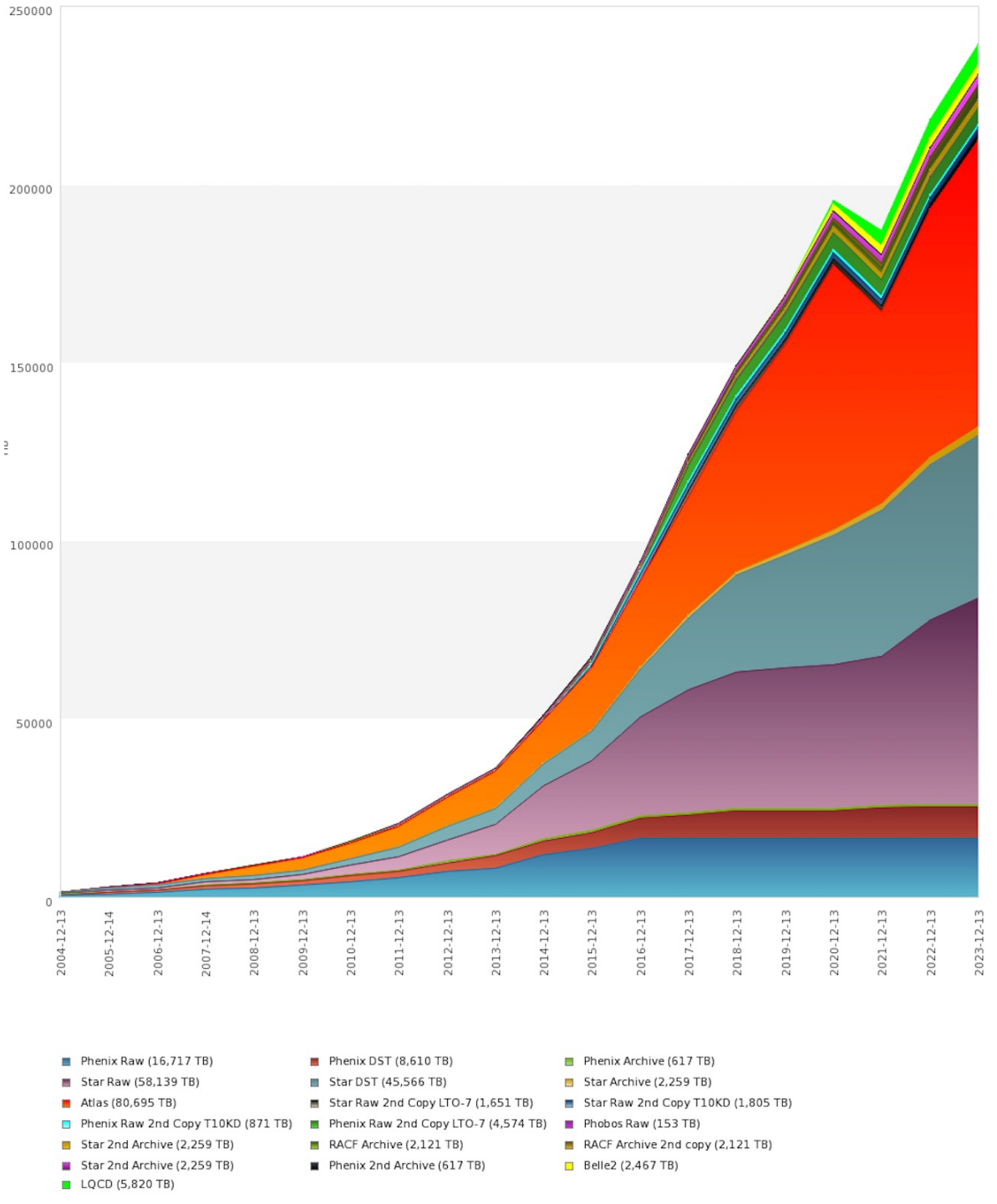
April 22nd, 2024

SDCC, BNL

@BrookhavenLab

HPSS Statistics

- Archive data size: 257.69 PB
  - 239,125,036 files (03/18/2024)
  - Active tape volumes: 75,493
- Data Movers, 25 servers
- Tape libraries: 14
  - Oracle SL8500: 9
    - 83,616 slots
  - IBM TS4500: 5
    - 37,800 slots



Brookhaven National Laboratory

HPSS Statistics 2023

- Tape Drives, 272
  - LTO7 (6 TB) - 49
  - LTO8 (12TB) - 112
  - LTO9 (18TB) - 64
  - Misc. – 47
- Tape slots: 122,464
  - 85,576 on Oracle libraries
  - 36,888 on IBM TS4500
- Active tape volumes: 75,588

HPSS Upgrade

- HPSS upgrade to 8.3.20 from 8.3.0
- LTO9 support added
- Core, Movers, Batch servers, Lustre clients, Gateways upgraded
- All HPSS clients upgraded

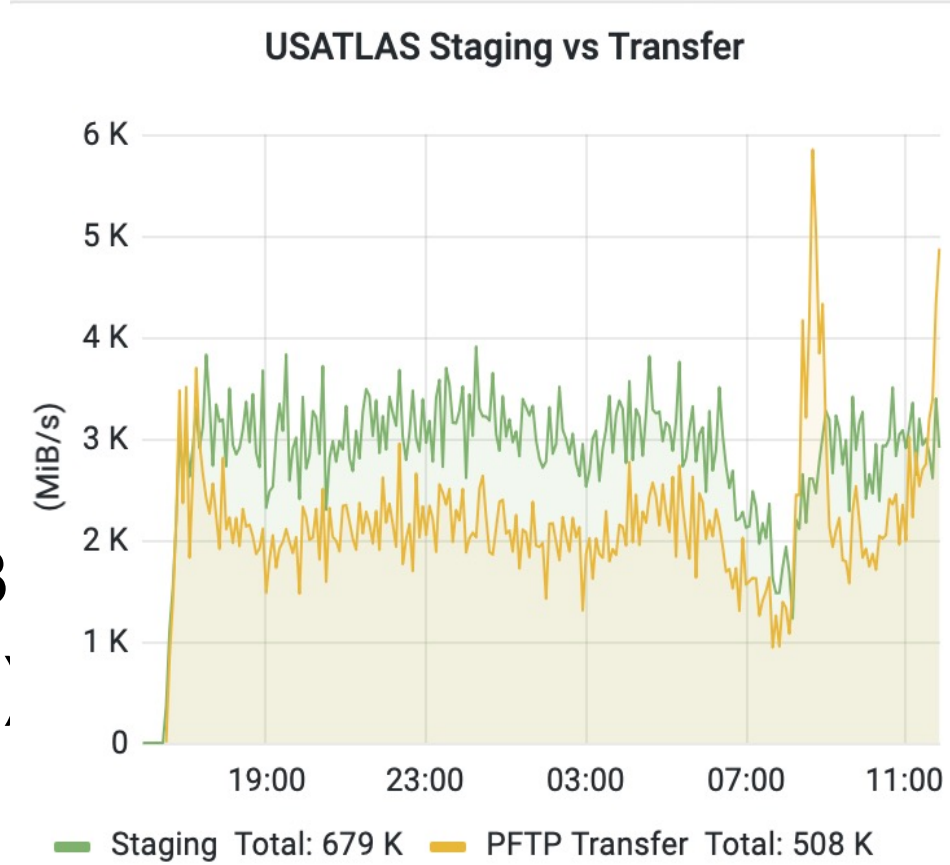
ID ▼	Name	Type	% Used	Total Space
50	Phenix Raw (disk)	Disk	2	121,913,984 MB
52	Star Raw (Disk)	Disk	59	274,306,464 MB
54	Phenix DST (disk)	Disk	81	121,913,984 MB
56	Star DST (disk)	Disk	80	91,435,488 MB
58	Archive (disk)	Disk	30	121,913,984 MB
59	US Atlas (disk)	Disk	33	1,016,069,872 ...
60	Belle2 (disk)	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 ...

Atlas Procurements, 2022

- Two 8-frame IBM TS4500 libraries
  - 8806 slots in each library
- 64 LTO8 drives
- 4 Movers
- 1.2 PB of disk cache
- PFTP and HSI Clients
- Batch staging service
- Monitoring tools and graphs
- Designed to sustain 8 GB/sec

Atlas Operations

- 64 LTO8 drives
- 30 LTO7 drives (Read-only)
- 1.2 PB of disk cache
- 28.7 PB (7,898,544 files) staged in 2023
- 11.5 PB (5,633,412 files) injected
- Replace gateway load balancer HAProxy with Round-robin DNS
- Provide RPM of HPSS Clients for RHEL 8
- ✓ Sustain 8 GB/sec (limitation of disk cache)

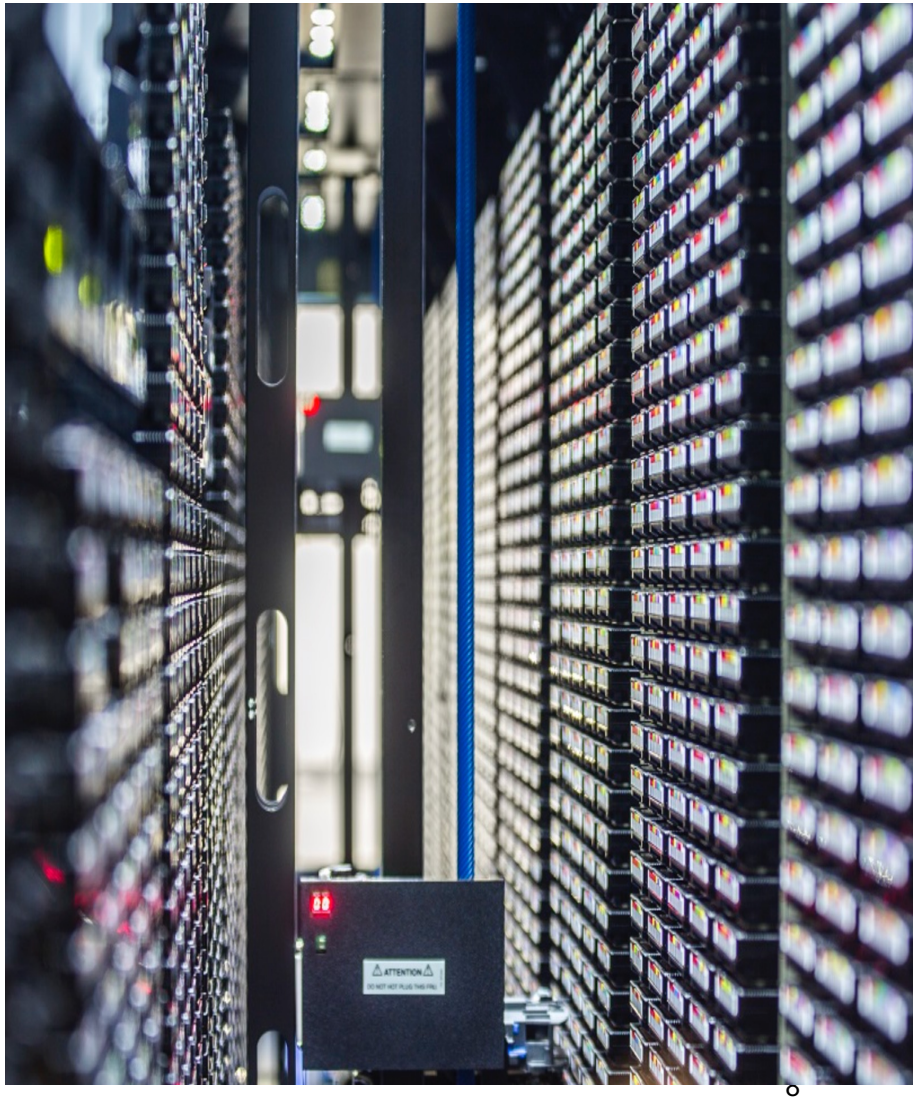


Atlas Operations

- 1,000 LTO8 tapes (12 TB/tape) are purchased in Feb, 2023
- Atlas now has 1,150 blank LTO8 tapes (13.8 PB)
- 8,026 slots available on TS4500 libraries (as of Mar18, 2024)
  - Existing Atlas tape libraries can hold another 110 PB of data
- Evaluation of new LTO10 tape libraries in 2026.
- Data repack of LTO4 to newer LTO technology completed
  - Data repack of LTO6 (2.5TB/tape) to LTO8 (12TB) will start soon

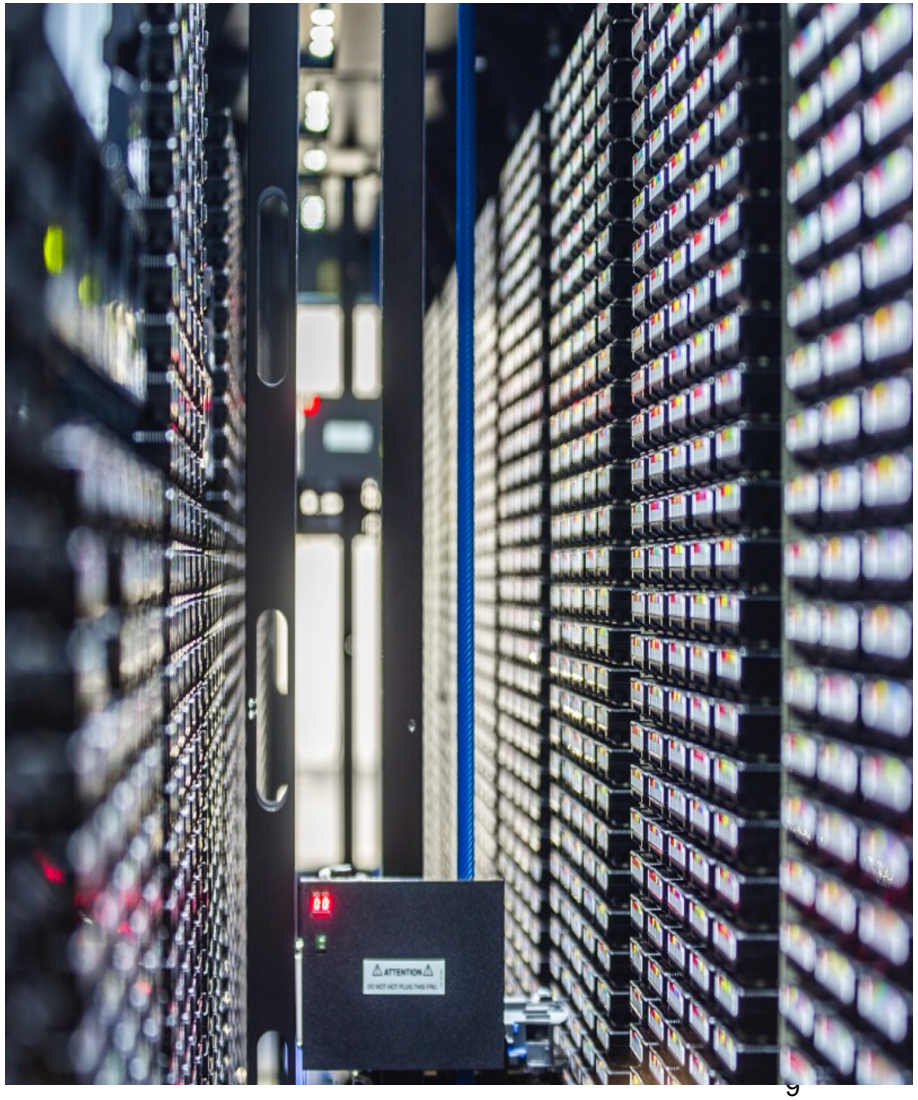
Tape Mount Testing

- Mount 32 drives, 151 sec (4.72 sec/mount)
  - 762 mounts/hour on each library
  - Exclude time for tape loads by the drives.
- Dismount 32 drives, 168 sec (5.25 sec/dismout)
  - 640 dismounts/hour on each library
  - Exclude the time for tape unloads by the drives
  - TS4500 automatically remap the home slot address of a mounted tape to a nearest physical slot. This expedites the subsequent mounts of this loaded tape.
- 361 tapes can be swapped each hour
  - Dismount + Mount = Swap tapes
  - The highest mount rate observed in Atlas is 285/hour
- When tapes go to deeper tiers, it gets slower



Tape Mount Testing - continued

- Each robot has two grippers, fast tape access to the first two tiers
  - 7,044 out of 18,000 slots (84.5PB 12TB/tape) are on the first two tiers in the two libraries
  - With our projected data patterns, the hot tapes are likely all in the fast tiers
  - Tapes with cold data will gradually move to deeper tiers



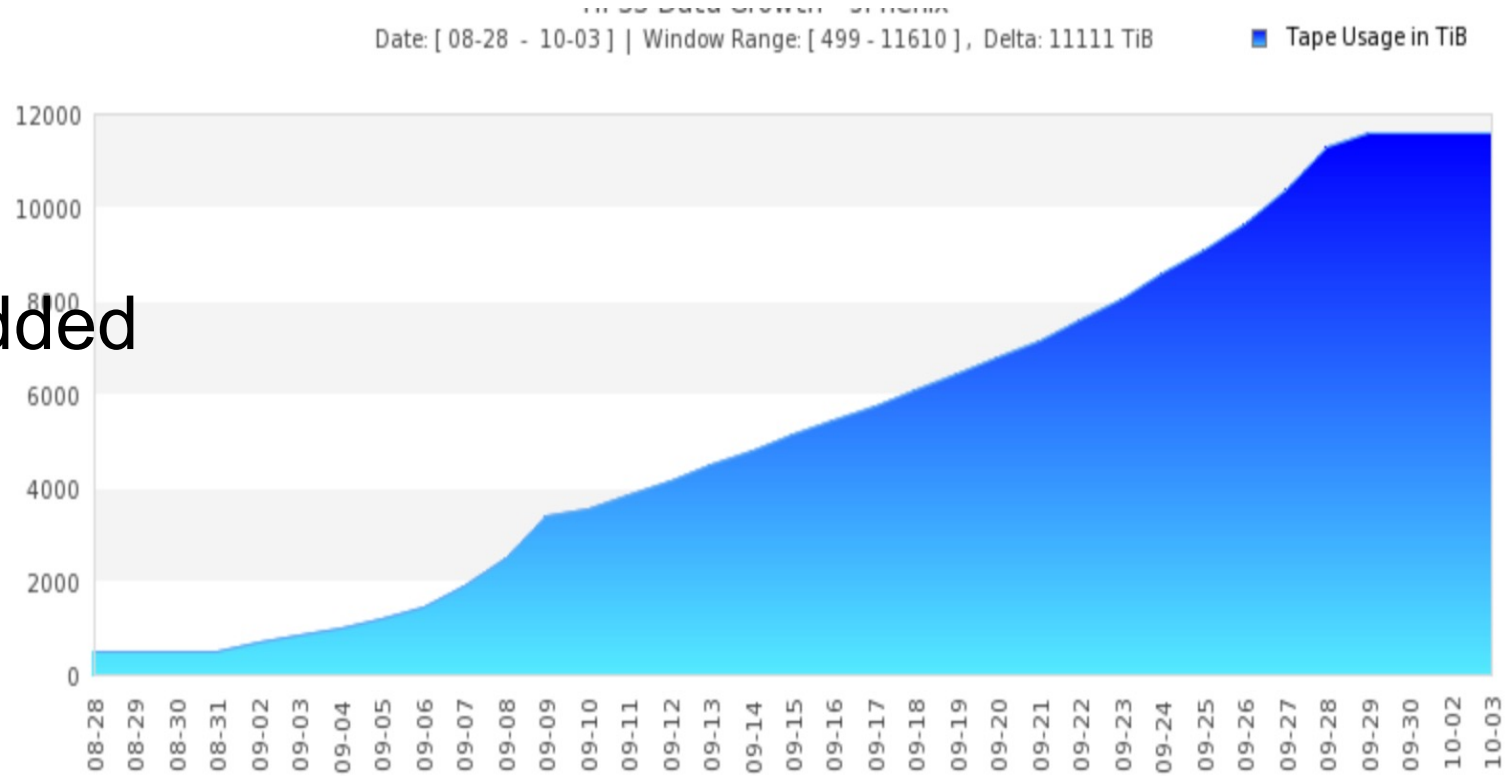
sPhenix Procurements, 2023

- Two 8-frame IBM TS4500 libraries
  - 8806 slots in each library
- 64 LTO9 drives
- 4 Movers
- 1.8 PB of disk cache
- PFTP and HSI Clients
- Batch staging service
- Monitoring tools and graphs
- Designed to sustain 10 GB/sec

sPhenix Run23

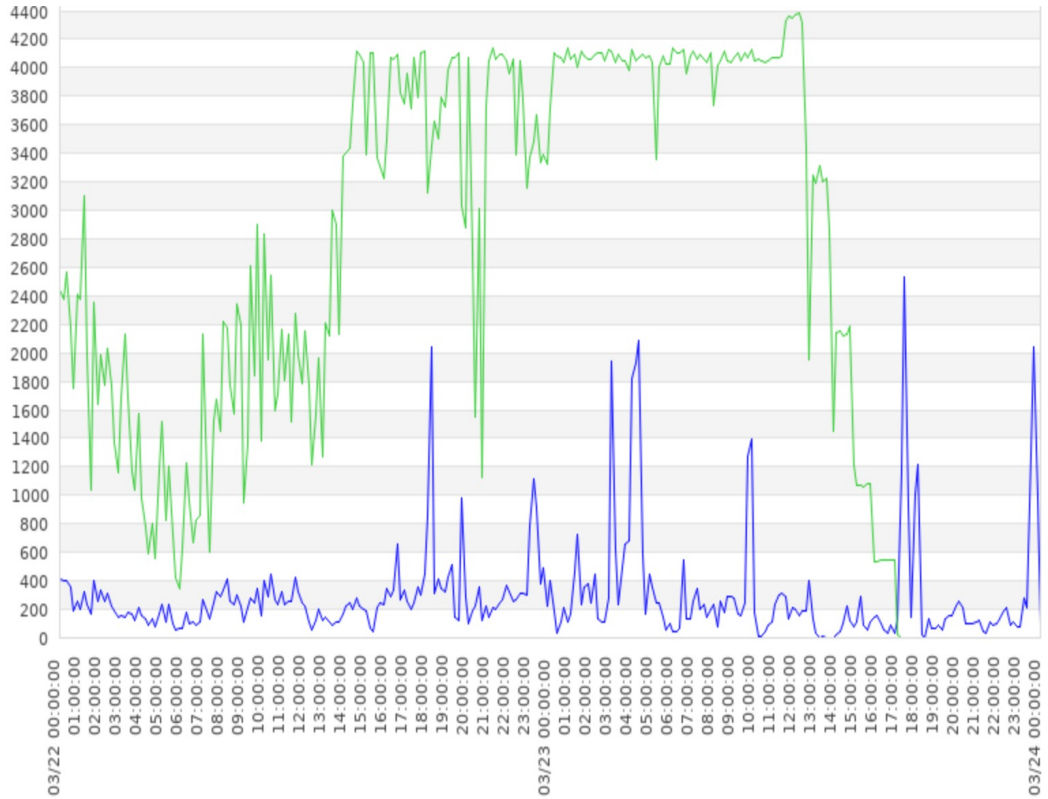
- 11.6 PB of data injected to HPSS
- 2,400 LTO9 prepared
- 321 LTO9 tapes used
- Average file size 20GB
- Tools/monitoring plots added

✓ Sustain 10 GB/sec



Star MDC 2023 (03/22 ~ 03/24)

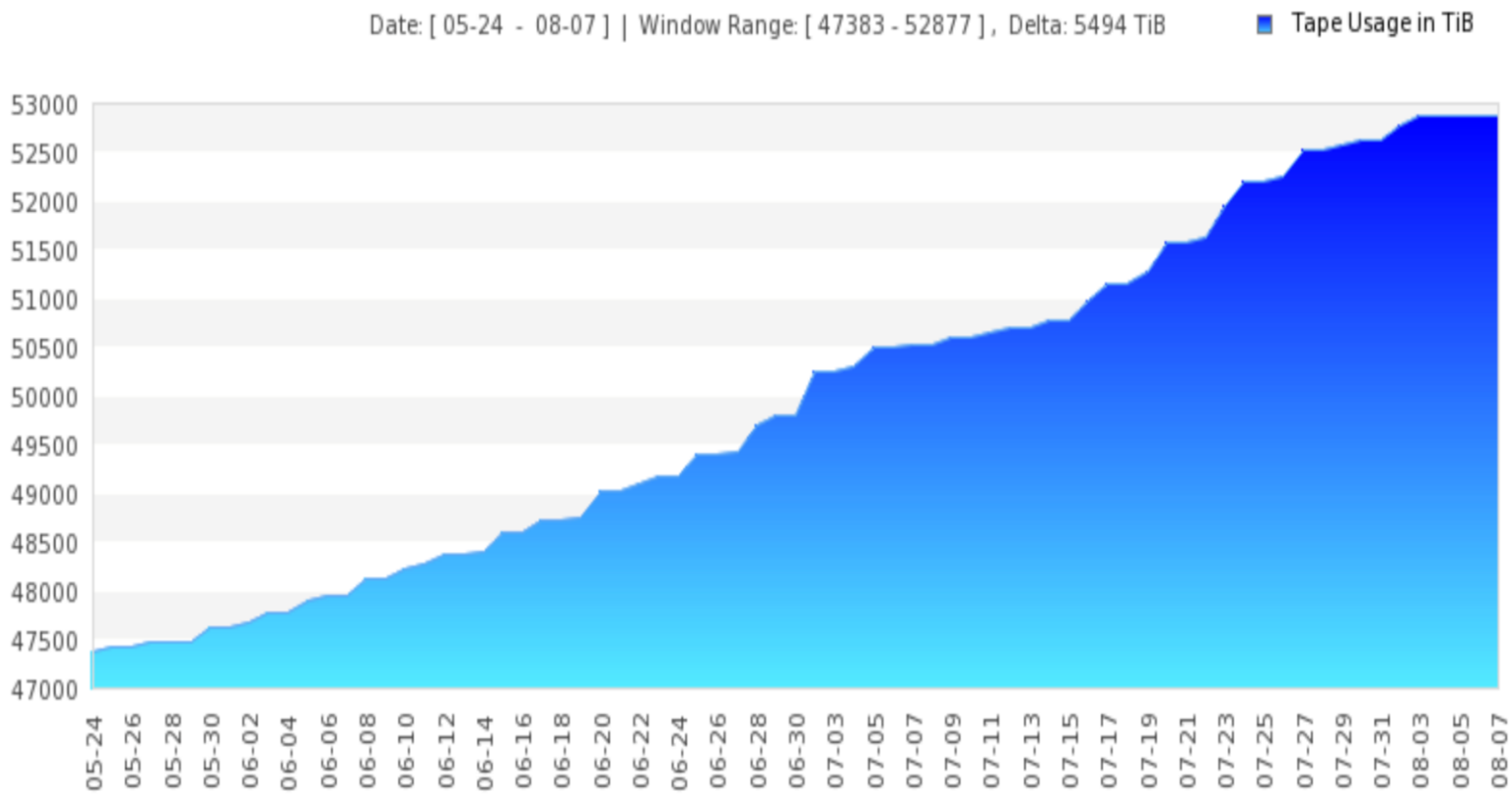
- 48 hours of continuous data injection
- Injection rate stabilized at 4.1 GB/sec
- 557.5TB injected
- 14 LTO8 drives migrating concurrently
- All used LTO8 tapes reclaimed after MDC
- MDC has met the 4.0GB/sec requirement



Brookhaven National Laboratory

Star Run23

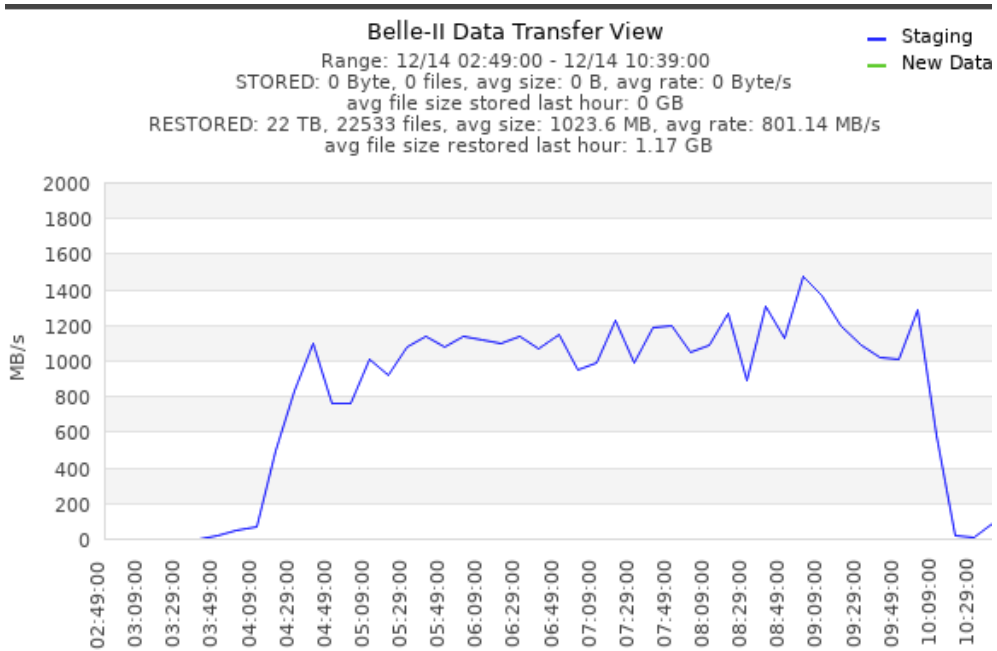
- 5 additional LTO8 drives installed
  - 18 LTO8 drives total
- 275 TB of Disk cache
- 4 data movers
- 5.5 PB injected
- ✓ Sustain 4 GB/sec



Belle2 Operations

- 10 LTO7/8 drives
- 92 TB of disk cache
- 708.2 TB (740,826 files) staged in 2023
- 0.3 TB (500 files) injected

✓ Sustain 2 GB/sec



Phenix Operations

- Four LTO8 drives are acquired for media repack

- Four concurrent repack streams are constantly running to migrate Phenix data on LTO5 to LOT8

Data Repacks

- Manage the data migration of LTO5 to new LTO8 media

- Approx. 8,000 LTO5 tapes repacked
- Approx. 7,500 library slots freed up

- Data repacks keep data in new tape technologies and allow the retirement of old tape resources to reduce the maintenance costs.

Smart Writing, colocations

- Files injected to HPSS are usually grouped into directories
- The files on disk cache are sorted in the order of directories
- The files on the same directories are collocated on tapes
- Optimal number of tape drives are used for multi-stream concurrent injections

Smart Writing, file sizes

- Files smaller than 1GB in size are aggregated into larger files on tape
- File sizes matter, Atlas files are about 4GB in size (75% of tape throughput)
- File sizes larger than 10GB are recommended for better tape performance (85% of throughput)

File Data Set	LTO7 write (300MB/sec Max)	LTO8 Write (360MB/sec max)
16 MB x100	9.0 MB/sec	9.1 MB/sec
32 MB x 100	16.5 MB/sec	17.1 MB/sec
64 MB x 100	28.0 MB/sec	29.4 MB/sec
128 MB x 100	43.6 MB/sec	46.2 MB/sec
256 MB x 100	66.8 MB/sec	67.5 MB/sec
512 MB x 100	101.6 MB/sec	105.1 MB/sec
1 GB x100	156.3 MB/sec	164.6 MB/sec
2 GB x 50	202.5 MB/sec	221.5 MB/sec
4 GB x 50	238.0 MB/sec	268.1 MB/sec
8 GB x 50	259.2 MB/sec	300.1 MB/sec
16 GB x 50	272.9 MB/sec	318.8 MB/sec
32 GB x 50	279.4 MB/sec	328.2 MB/sec

Batch staging,

- At staging, requests are submitted in bulks to Batch queues.
- To minimize tape mounts and repositioning, Batch will group staging requests by tapes and order them by its logical positions on tape.
- For better staging performance, submit staging requests in the same directories in bulks of high numbers
- RAO on LTO9 and enterprise tape drives requires developments on Batch

Development of new Batch

- Evaluation of HPSS staging tool: QUAID
- Development of New HPSS Batch System with QUAID
- Function & Performance test of New HPSS Batch System

Tasks in the near future

- Preparations for Run24 for all experiments
- Continue Smart writing optimizations to improve performance
- TSM tape subsystem installation and configurations.
- Batch (data staging) development with HPSS LORI
- Continue data repacks to newer medium technologies
- New test environment
- HPSS upgrade to 10.x
- Prepare new tape libraries for sPhenix after Run24
- Explore new technologies

Thank you!

Q & A…