Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Archive Storage at SDCC 2024

Tim Chou, Ognian Novakov,
Justin Spradley, Iris Wu

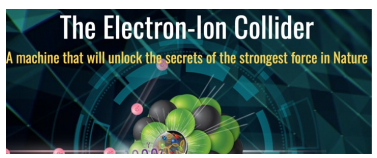
March, 2024

ISGC, Taipei, Taiwan

@BrookhavenLab

SDCC: The Scientific Data and Computing Center

- Located at Brookhaven National Laboratory (BNL) on Long Island, New York
- SDCC was initially formed in the mid-1990s as the RHIC (Relativistic Heavy Ion Collider) Computing Facility



Shared multi-program facility serving ~2,000 users from more than 20 projects

NSLS2

Scientific Data and Computing Center Overview

- Tier-0 computing center for the four RHIC experiments
  - sPHENIX - started taking data in May, 2023
- US Tier-1 Computing facility for the ATLAS experiment
- Computing facility for NSLS-II (National Synchrotron Light Source-II)
- US Tier-1 data center for Belle II experiment
- Providing computing and storage for proto-DUNE/DUNE along w/ FNAL serving data to all DUNE OSG sites
- Also providing computing resources for various smaller / R&D experiments in NP and HEP
- Serving more than 2,000 users from > 20 projects
- Developing and providing administrative/collaborative tools:
  - Invenio, Jupyter, BNL Box, Discourse, Gitea, Mattermost, etc.
- BNL was selected as the site for the major new facility Electron-Ion Collider (EIC/eRHIC)

111 rack frames are already deployed in Main Data Hall

84 RDHx units deployed in B725 MDH, out of which 59 are on racks with equipment while 25 are deployed for the future growth

B725 Central Network Equipment Is Deployed & Active
(10x 400 GbE ready Arista modular chassis with 48x line cards slots in total)

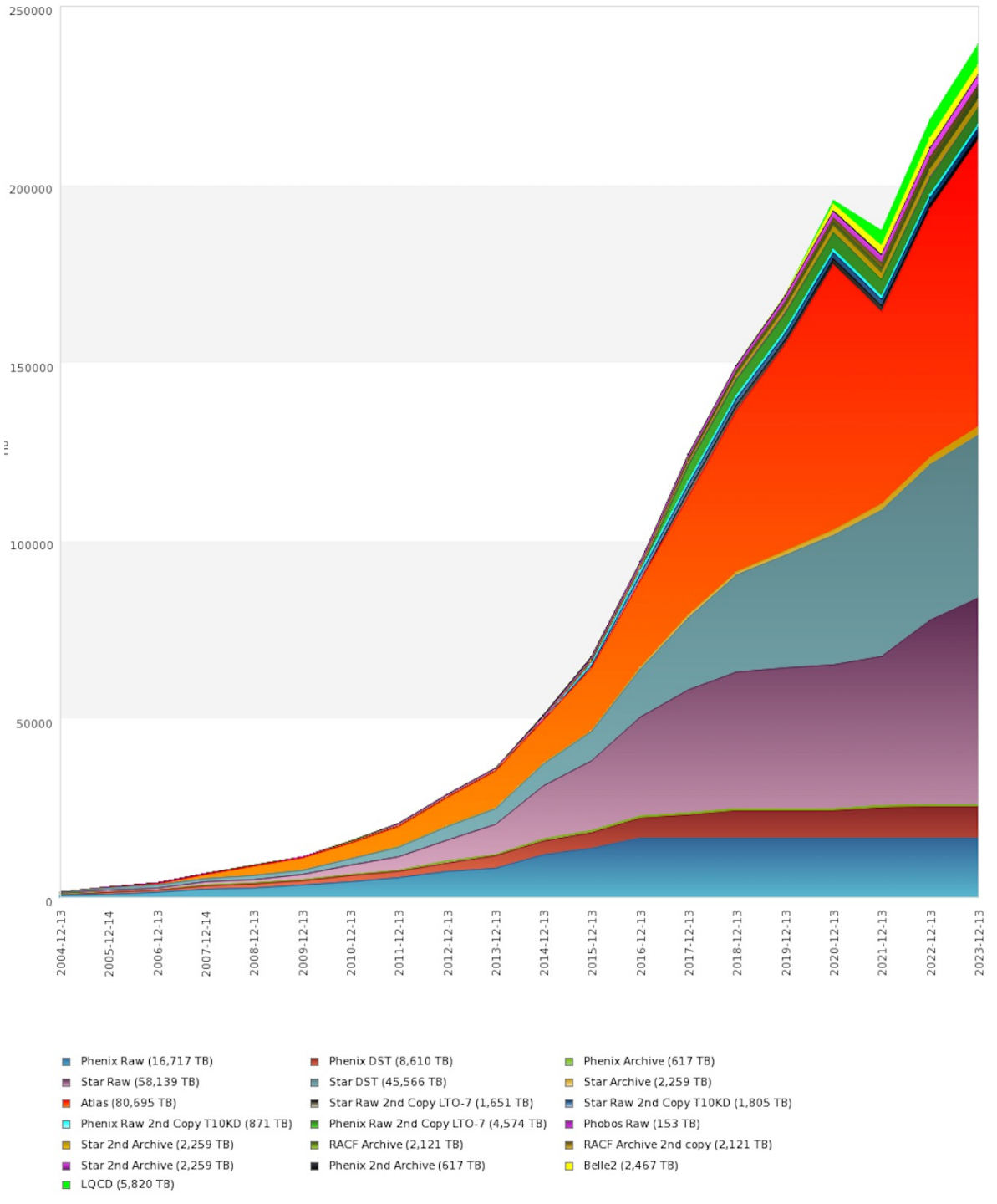
4x 8-frame (352K slots)
IBM TS4500 tape libraries
are installed

Brookhaven National Laboratory

4

Archive Statistics 2023

- Archive data size, 257.69 PB
  - 239,125,036 files (03/18/2024)
- 25 Data Movers
- Tape libraries: 14
  - 9 Oracle SL8500
    - 83,616 slots
  - 5 IBM TS4500
    - 37,800 slots
  - 2nd largest tape archive site in the US.



Brookhaven National Laboratory

Tape Storage Statistics 2023

- Tape Drives, 272
  - LTO6 ~ LTO9
- Tape slots: 122,464
  - 85,576 on Oracle libraries
  - 36,888 on IBM TS4500
- Active tape volumes: 75,588

Disk Storage: Lustre, dCache & XROOTD

Total ~74 PB in dCache

- ATLAS (v8.2.15), Belle II (v7.2.19), PHENIX (v5.2.9), DUNE (v8.2.2)

XROOTD

~11 PB total storage for STAR

- Mix of central and farm node storage

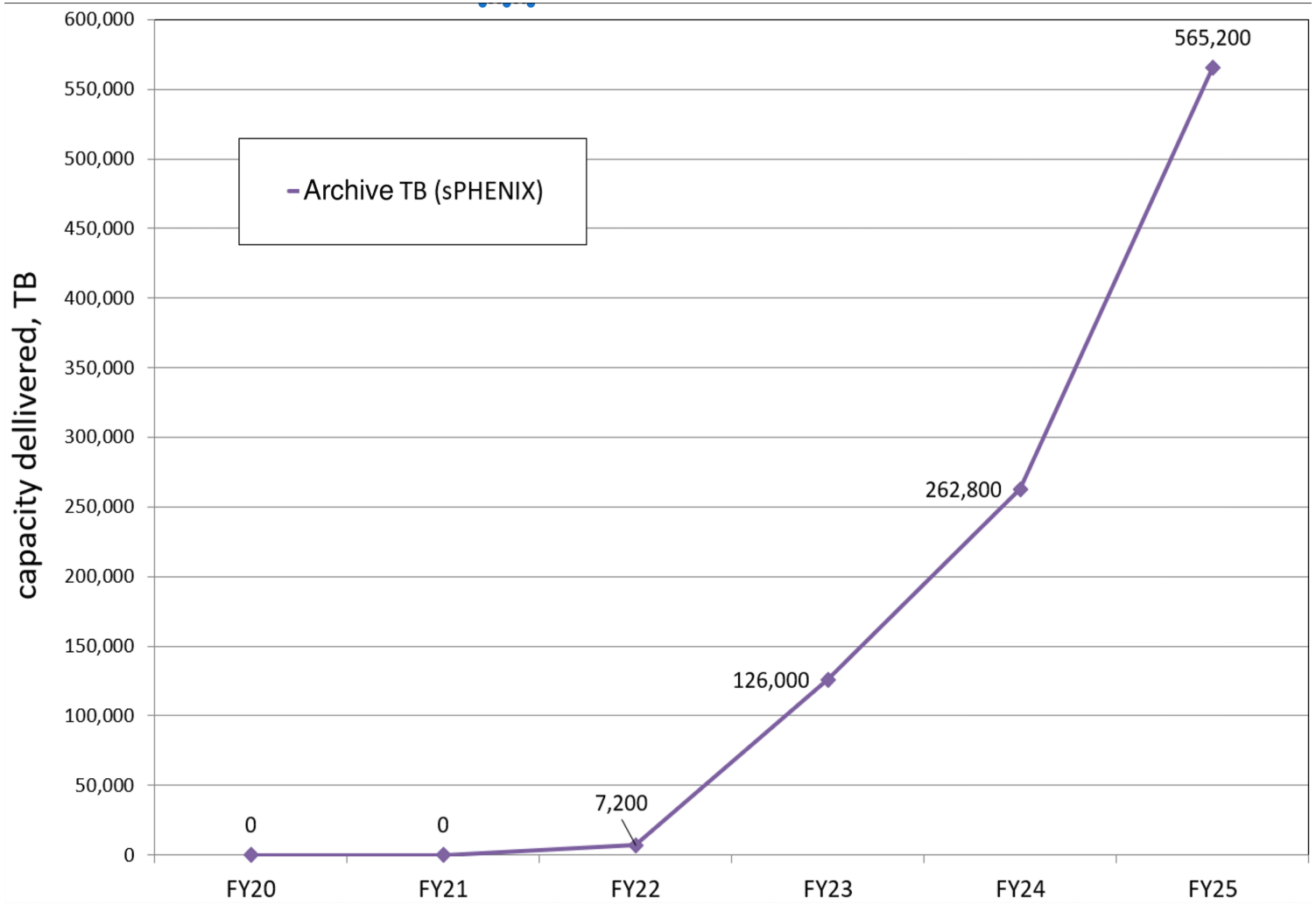
Total ~50 PB in Lustre

- ATLAS, EIC, LQCD, NSLS-II, sPHENIX, STAR
- Growing footprint for Lustre (2.12.8)
- Added 25PB to sPHENIX
- Excellent streaming sequential performance with aggregate throughput of 210 GB/s

sPhenix Archive Storage Projections

- FY22 7.2 PB
- FY23 126.0 PB
- FY24 262.8 PB
- FY25 565.2 PB

- Data archived to tape will not be purged
- Requires 10GB/sec



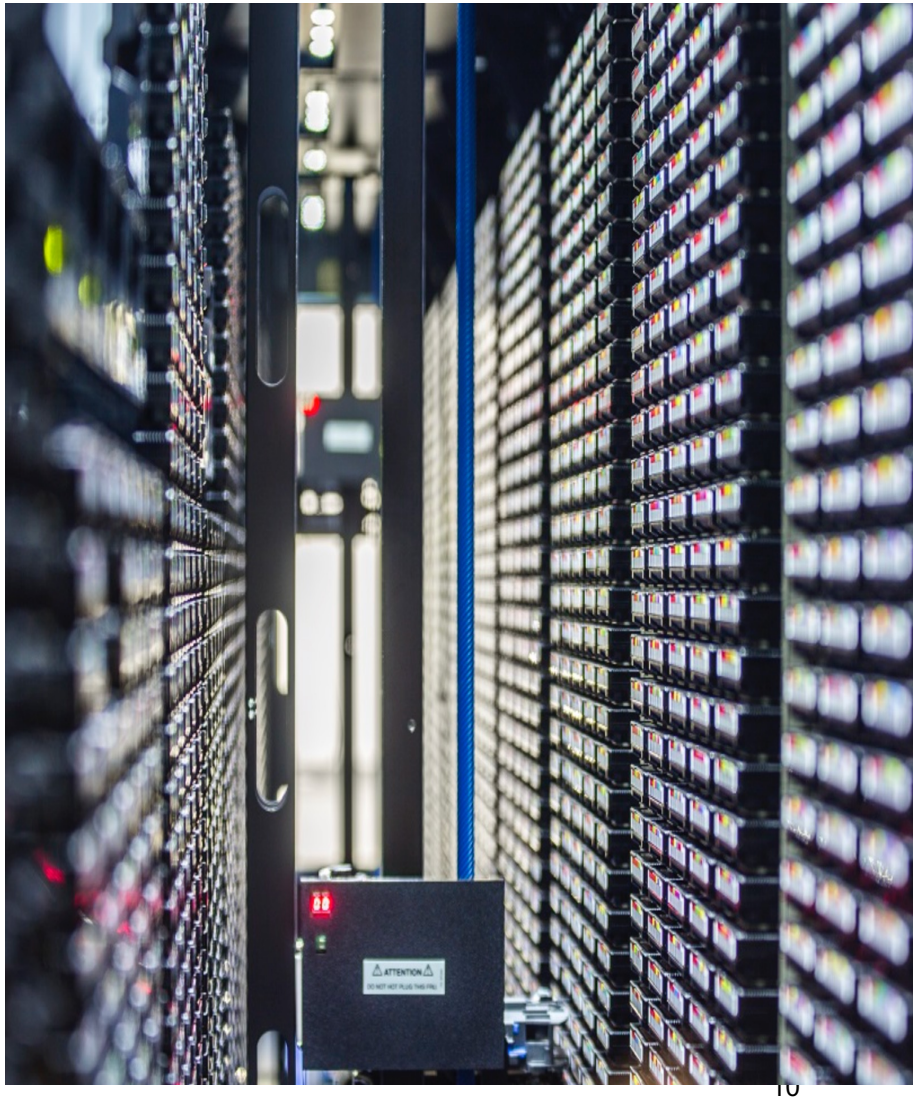
Brookhaven National Laboratory

sPhenix Procurements & Installations

- 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 10GB/sec

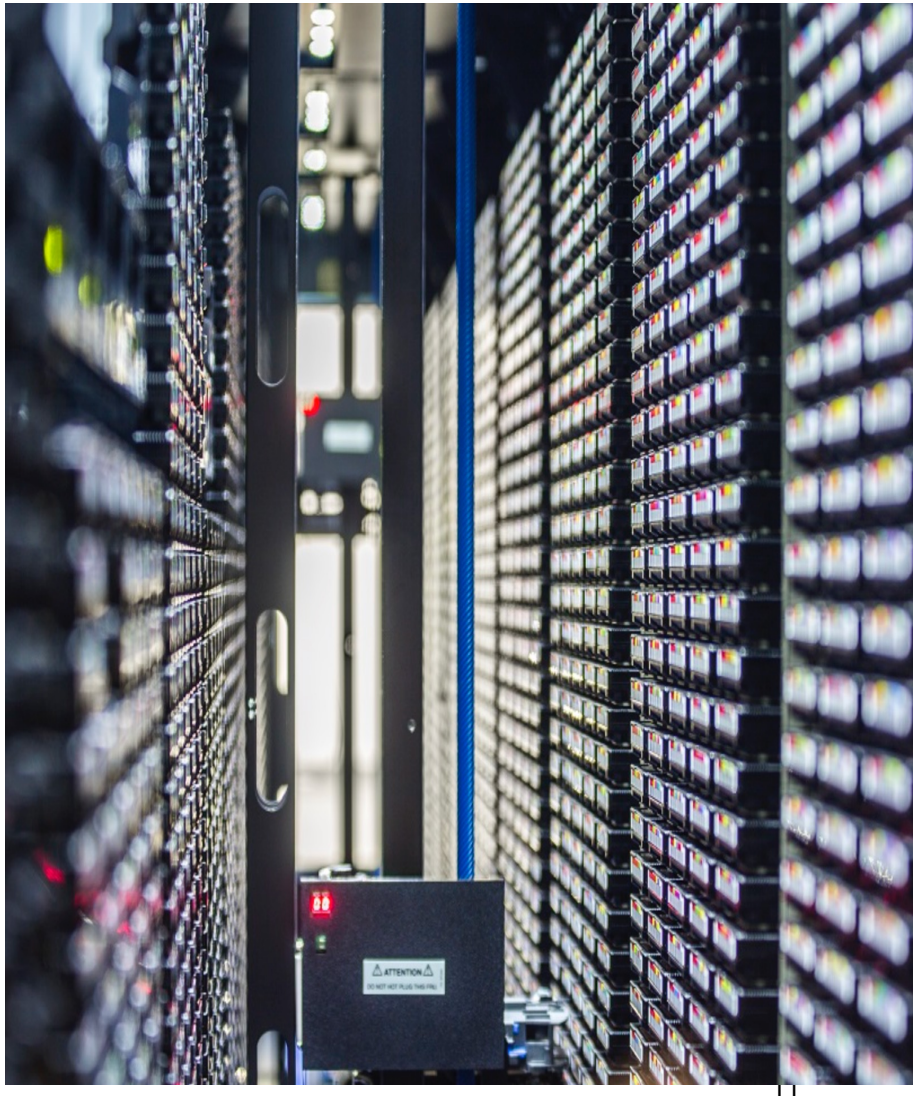
Tape Mount Testing

- 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/dismount)
  - 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

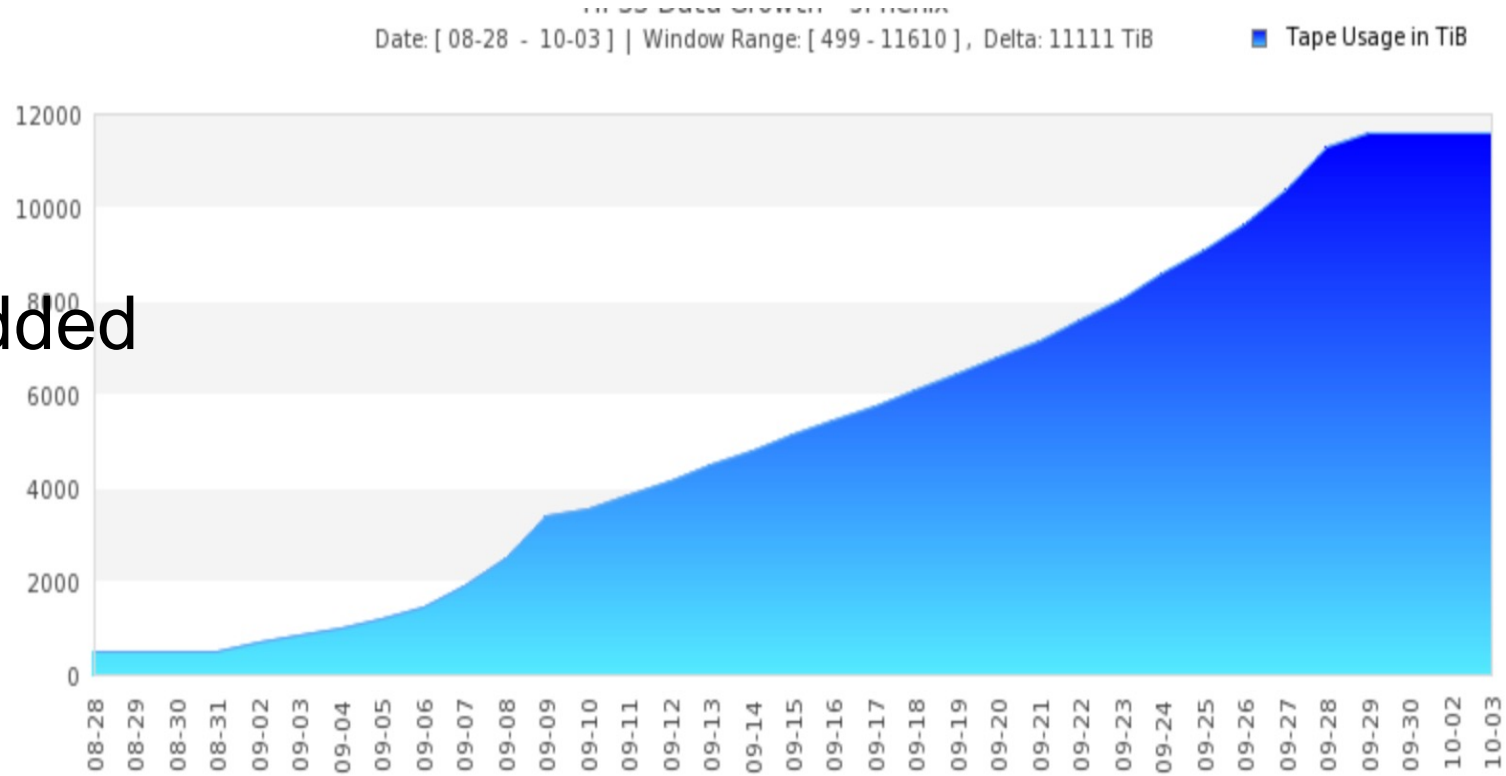
- Each robot has two grippers, fast tape access to the first two tiers
  - 7,044 out of 18,000 slots 126.8 PB 18TB/tape) are on the first two tiers in the two libraries
  - With our projected data patterns, the hot tapes are likely all in the fast tiers
  - Tapes with cold data will gradually move to deeper tiers



sPhenix Run23

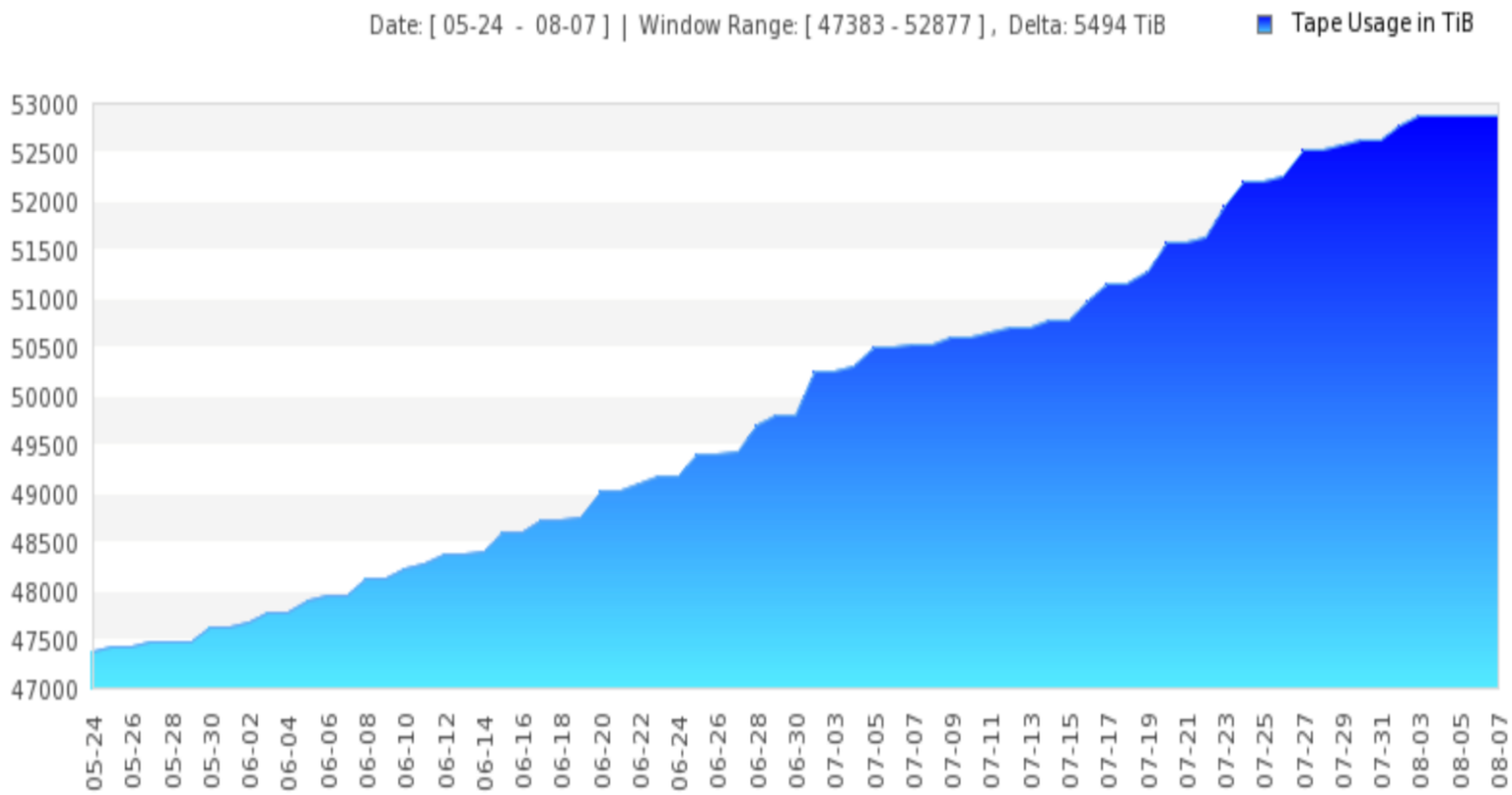
- 11.6 PB of data injected to HPSS
- 2,400 LTO9 prepared
- 321 LTO9 tapes used
- Average file size 20GB
- Tools/monitoring plots added

✓ Sustain 10 GB/sec



Star Run23

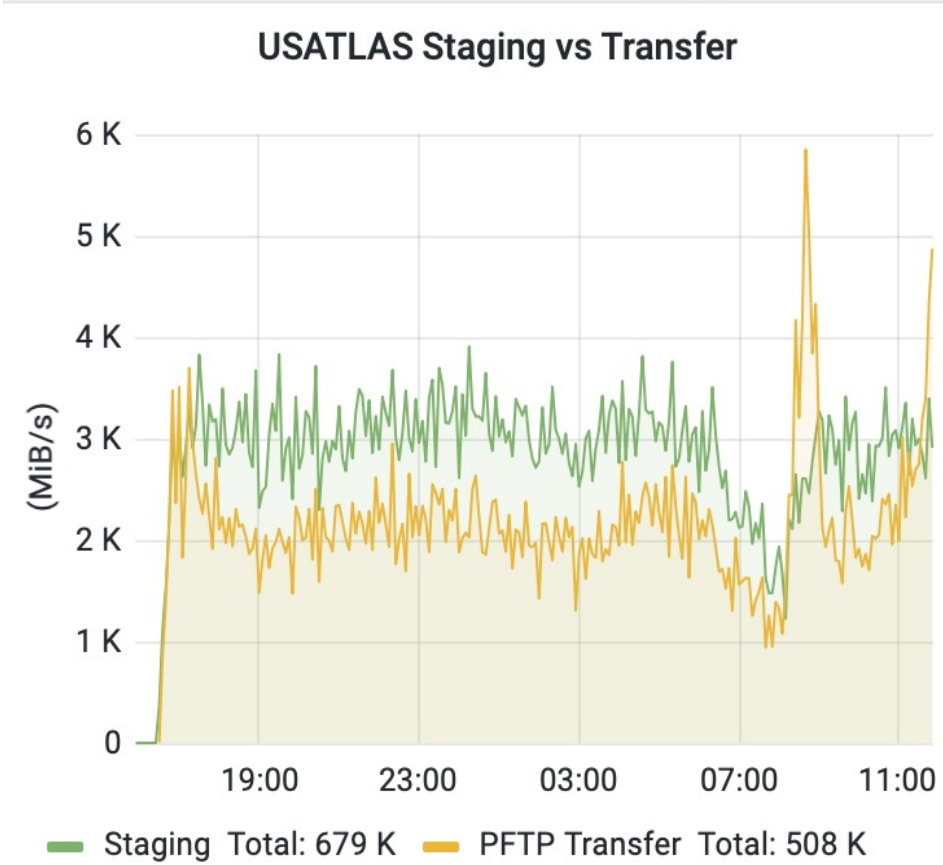
- 5 additional LTO8 drives installed
  - 18 LTO8 drives total
- 275 TB of Disk cache
- 4 data movers
- 5.5 PB injected
- ✓ Sustain 4 GB/sec



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

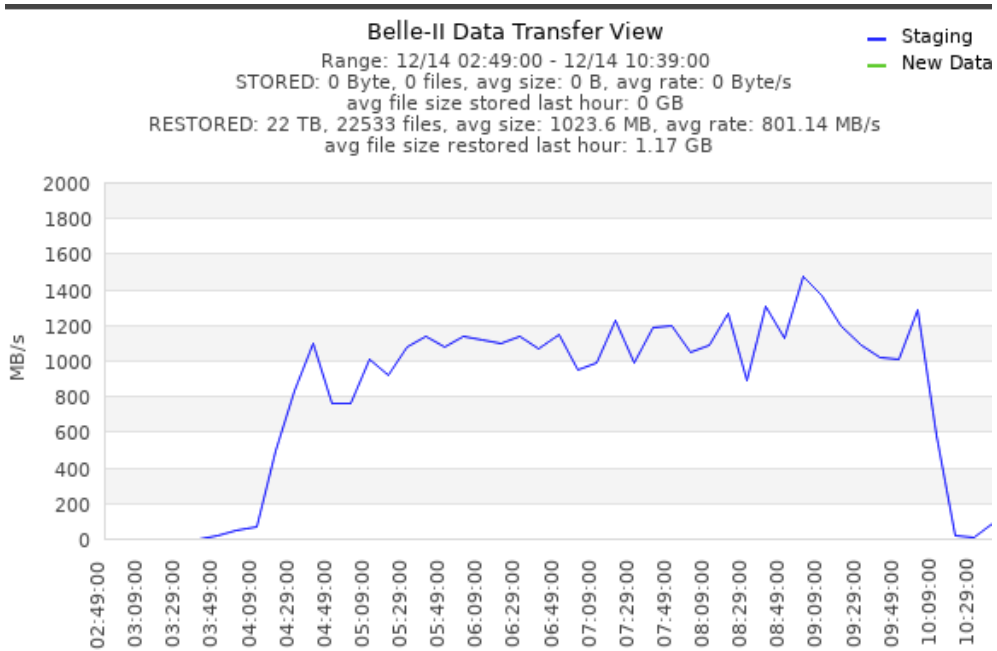
✓ Sustain 8 GB/sec



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



Brookhaven National Laboratory

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 on latest 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)
- Buffered tapemarks

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

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 technologies
- New test environment
- HPSS upgrade to 10.x
- Prepare new tape libraries for sPhenix after Run24
  - Projected data size for Run24 is 565.2 PB
- Explore new technologies

Thank you!

Q & A…