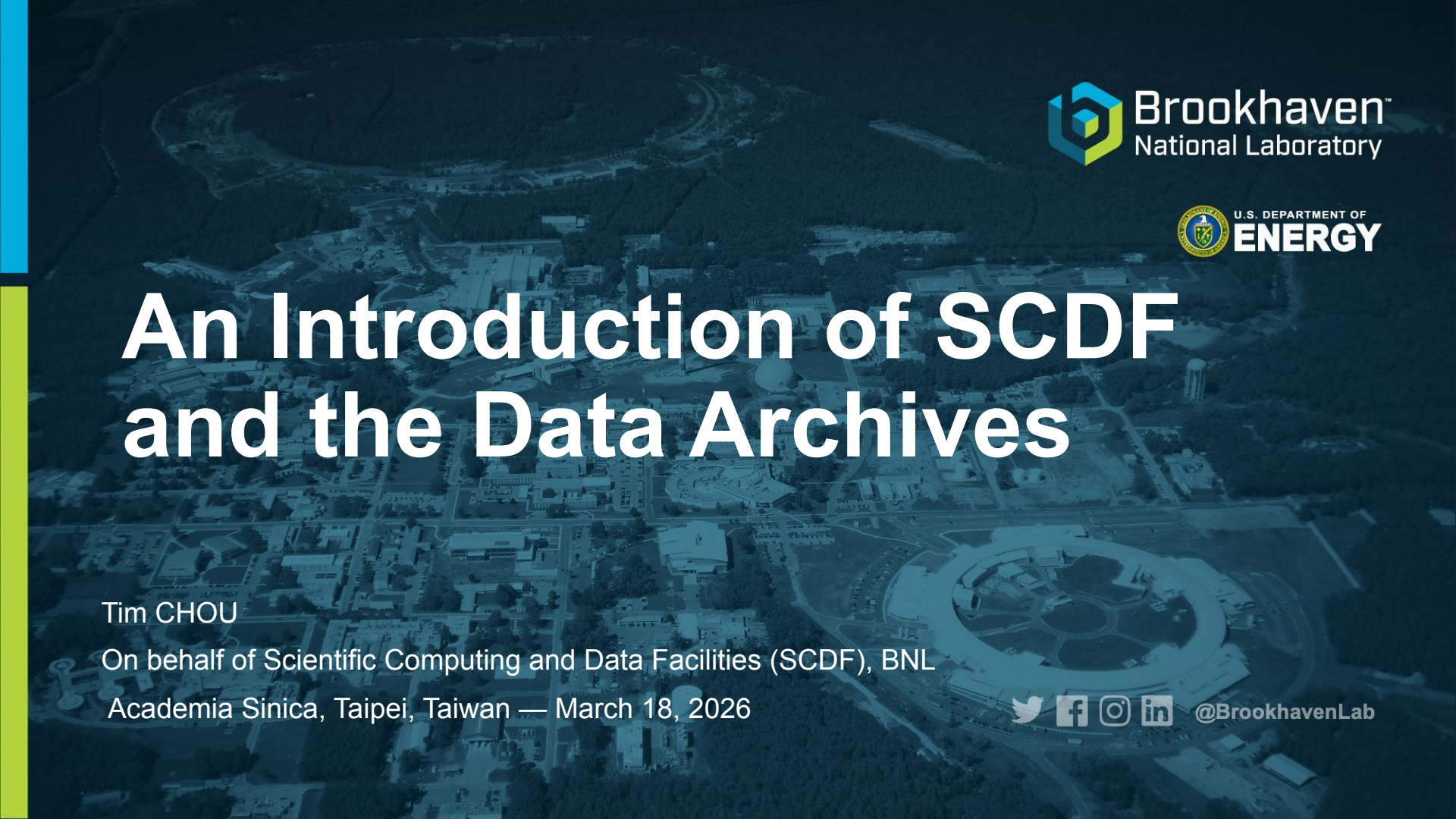


An Introduction of SCDF and the Data Archives

Tim CHOU

On behalf of Scientific Computing and Data Facilities (SCDF), BNL

Academia Sinica, Taipei, Taiwan — March 18, 2026

@BrookhavenLab

Outline

- Scientific Data and Computing Facilities at BNL
- 2025/2026 Operations
- Active on reading
- High throughput
- Fault tolerance
- dCache & HPSS

Scientific Computing and Data Center Overview

- Located at Brookhaven National Laboratory (BNL) on Long Island, New York
- Tier-0 computing center for the RHIC experiments
  - Final RHIC run in 2025 - sPHENIX & STAR
  - BNL is host lab for the future Electron-Ion Collider (EIC)
    - Ongoing discussions about ePIC DAQ, computing and software
- US Tier-1 Computing facility for ATLAS experiment at LHC (*100% R&A Jul-Feb*)
  - Also one of the ATLAS shared analysis (Tier-3) facilities in the US
- Belle II Tier-1 and Data center
- Data center for NSLS-II photon science

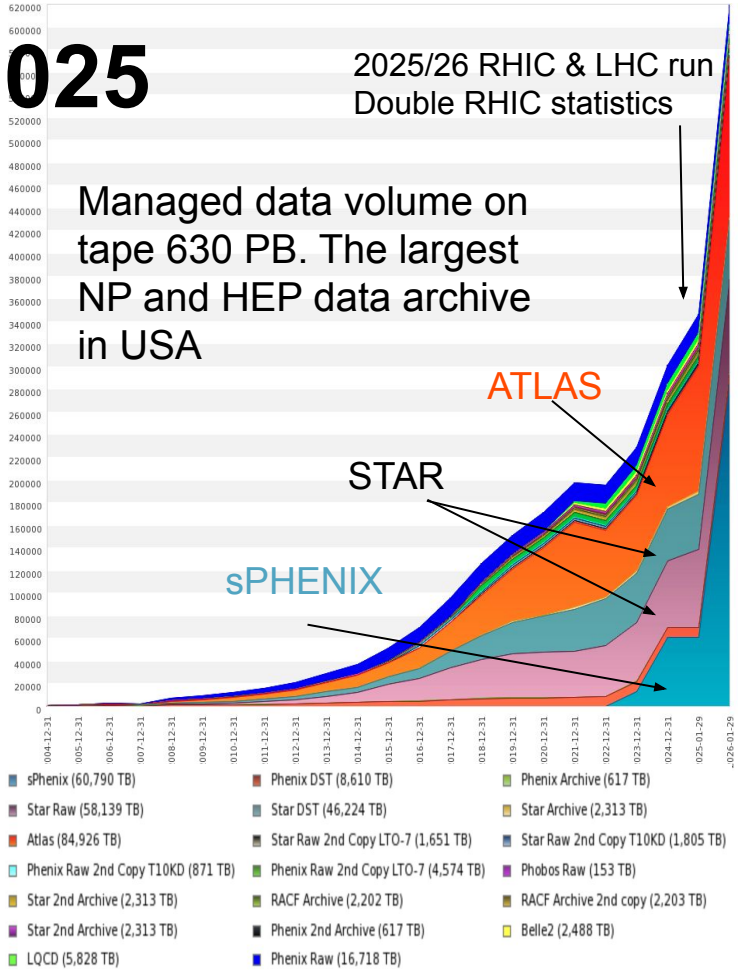
- Providing computing and storage for proto-DUNE/DUNE along w/ FNAL serving data to all DUNE OSG sites
- Providing computing resources for a number of smaller experiments and R&D projects in NP and HEP
- Serving more than 2,000 users from >20 projects

Brookhaven National Laboratory

Data Archive Operations 2025

- 630 PB of scientific data on tape
  - The largest NP and HEP data archive in the USA

- New installations and upgrades
  - Two new sPHENIX libraries
  - New sPHENIX movers and disk cache
  - Replace STAR movers and disk arrays
  - Replace Belle2 disk array
  - Replace LQCD mover and disk array
  - New monitoring charts



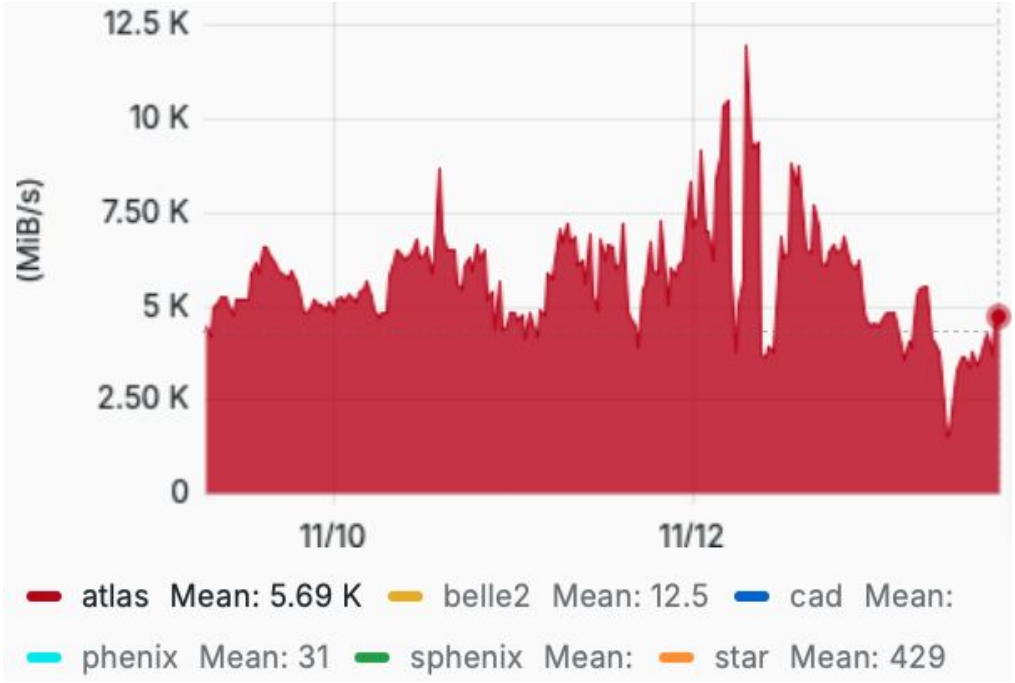
Tape Storage Infrastructure

- Data Movers - 27
- Tape libraries - 16
  - 9 Oracle 8500
  - 7 IBM TS4500
- Tape Drives, 308
  - LTO7 (6 TB) - 49
  - LTO8 (12TB) - 112
  - LTO9 (18TB) - 100
- Tape slots: 142,464
  - 85K+ on Oracle libraries
  - 57K+ on IBM TS4500
- Disk Cache 9.5 PB
- Active tape volumes: 81K+

ATLAS Run 2025

- Archive data size – 142.0 PB
  - The largest Atlas Tier-1 data site
- 41.7 PB (9+M files) staged in 2025
- 28.9 PB (9+M files) injected in 2025
- ATLAS read requests > write
- Atlas movers and gateways upgraded to RHEL8

✔ Sustain 8 GB/sec



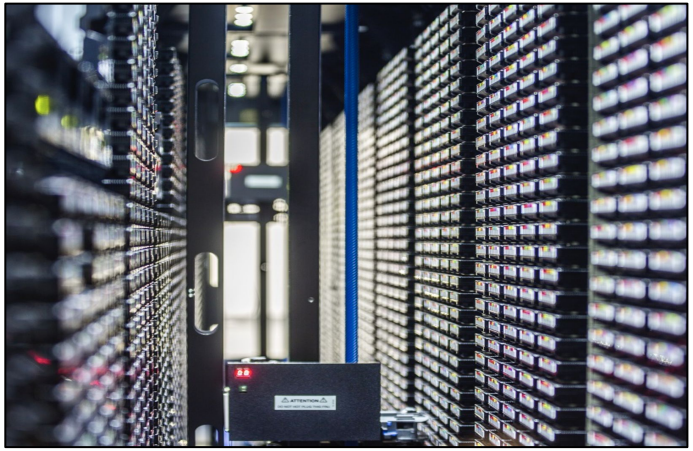
Atlas 11/09 - 11/13/2025 stages average at 6.11GB/sec

Active Reading: ATLAS

- Active reading, More read traffic than write
  - Require optimization to access tape media
  - Require optimization for positioning to files on tape
- File sizes 2 to 5 GB
  - Small file Aggregation (< 1GB/file)
- Throughput configured for 7 GB/sec (8 GB/sec achieved)
  - Data injection with concurrent tape migration

ATLAS requires fast robotics

- Mount 32 drives, 151 sec (4.72 sec/mount)
  - 762 mounts/hour (without tape positioning)
- Dismount 32 drives, 168 sec (5.25 sec/dismount)
  - 640 dismounts/hour
- **360/2=180** tapes can be swapped each hour
  - Dismount + Mount = Swap tapes
  - ATLAS requires **285**/hour (180 < 285)
  -
- **Conclusion**
  - **We need to double the mounts/hr by deploying two libraries.**
    We'll write ATLAS data sets evenly across two tape.



ATLAS File Staging Optimization

- File co-locations on tape
  - HPSS writes files to tape in the order of directory, not time stamps
- dCache/ENDIT issues staging requests to HPSS
  - 200K requests limit
- BNL batch queues group staging requests, when in bulk
  - Group requests by tapes
    - to minimize tape mounts
  - Sort requests in the order of the file positions on tape
    - To minimize tape repositionings
  - Allow long-waited files to be staged first

Tape Info	Lock	Tape ID	# of Files	LSM	LSM Status	Lock	Status	Waited Time
Phenix Raw LTO-7		AE7887	15	4,16	ONLINE		4,16,1,12	00:48:37
Phenix Raw LTO-7		AE7907	14	4,13	ONLINE		4,13,1,3	00:28:37
Phenix Raw LTO-7		S77123	10	4,12	ONLINE		4,12,1,12	00:07:37
Phenix Raw LTO-7		S77122	8	4,12	ONLINE		4,12,1,0	00:07:37
Phenix Raw LTO-7		S77127	7	4,14	ONLINE		4,14,1,12	00:06:37
Phenix Raw LTO-7		S77125	7	4,18	ONLINE		4,18,1,13	00:17:37
Phenix Raw LTO-7		S77128	6	4,15	ONLINE		4,15,1,12	00:23:37
Phenix Raw LTO-7		AE7925	6	4,19	ONLINE			00:06:37

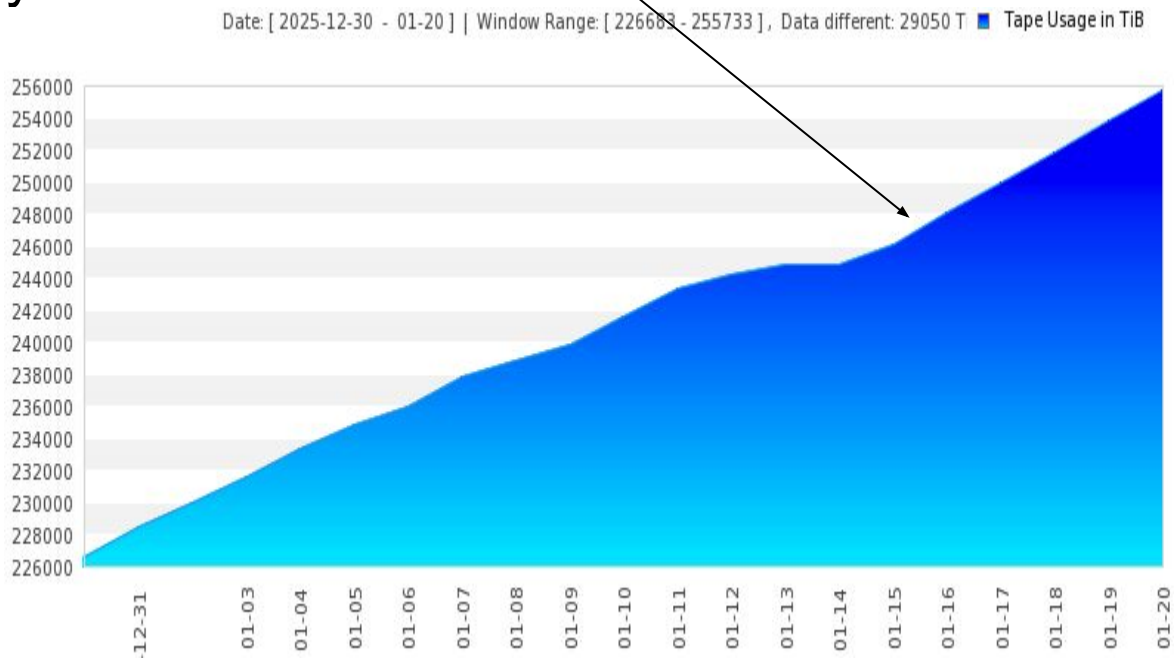
Staging requests grouped into each tape in tape positioning order

sPHENIX Run2025

- Archive data size – 300 PB
- 170.2 PB of data injected in 2025
- Data sets written evenly across 4 libraries
  - Expedite tape mounts
- New Monitoring plots

✔ Sustain 25 GB/sec, 50 GB/sec burst mode

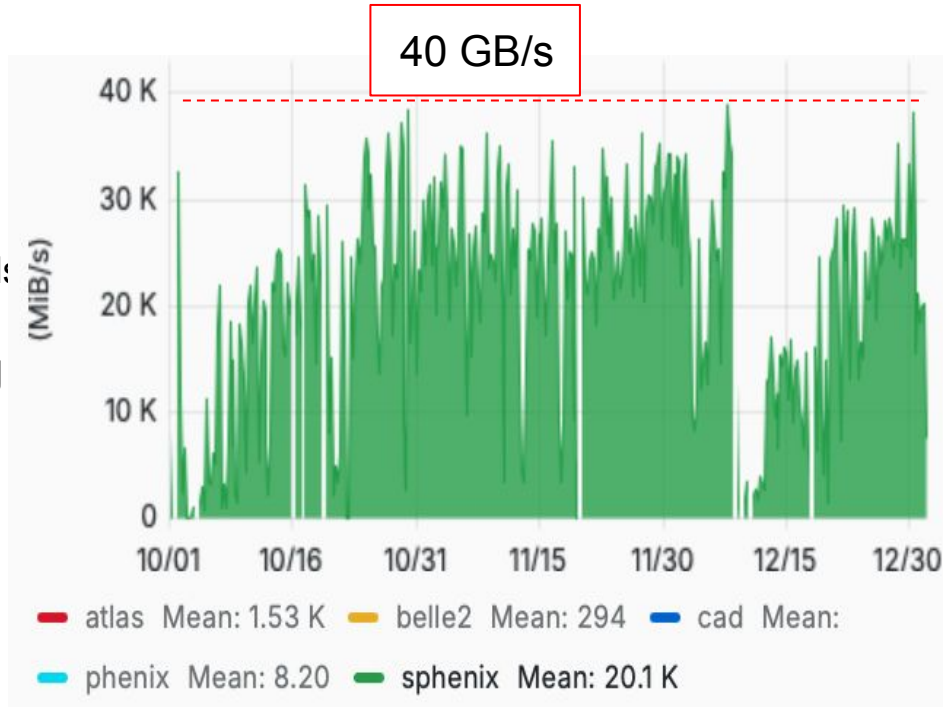
sPhenix injected 32PB in 21 days (12/31/25 - 1/20/26)
29,050TiB = 32 PB = 1.5PB/day



Brookhaven National Laboratory

High Throughput: sPHENIX

- 20 GB/sec injection required
- 6 client buffer boxes with NVME
  - Running concurrent pftp processes
  - No dCache/Endit for data injections
- 9 HDD disk arrays
  - 2+ PB of cache space needed
    - Cache size for 24 Hrs, in case tape fails
    - SSD over budget limit
- 72 tape drives concurrent streaming
- 25 GB/sec accomplished
  - Data injection with concurrent tape migration
- 50 GB/sec in burst mode
  - Data injection without tape migration
- Switch to READ mode in 02/2026



20.1GiB = 21.5GB, average injection over 3 months

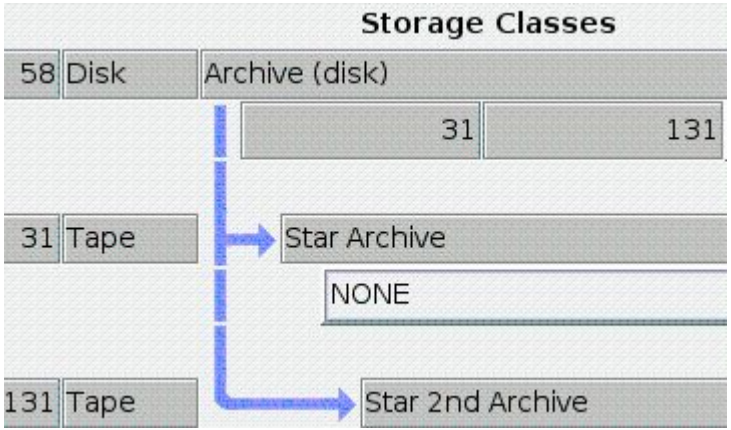
Brookhaven National Laboratory

Tunings for sPHENIX Throughput

- Movers upgraded to PCI-5 from PCI-3
- Data mover I/O buffers increased to 32MB from 4MB
- Disk multipath increased to quad paths from dual paths
  - Each fiber connection is 32Gb/sec
- Nine NetApp disk arrays
  - Seven arrays required to sustain 25GB/sec of concurrent migration and injection traffic
- Move 100Gb NIC's to 16-lan PCI slots from 8-lan slots
- Increase Kernel memory setting

Fault Tolerance: ARCHIVE classes

- ARCHIVE classes require tape fault tolerance
- Solution candidates
  - RAIT, 3+1 fault tolerance
  - Dual tape copies
- Dual tape copies was selected
  - File sizes of 1 - 5 GB not suitable for RAIT
- When a tape read fails, the other copy will be used



Data Repacks

Actively repacking data to new tape technologies

- Maintain the integrity of the scientific data
- Lower the costs by retiring old hardware
- Lower the library slot counts
-

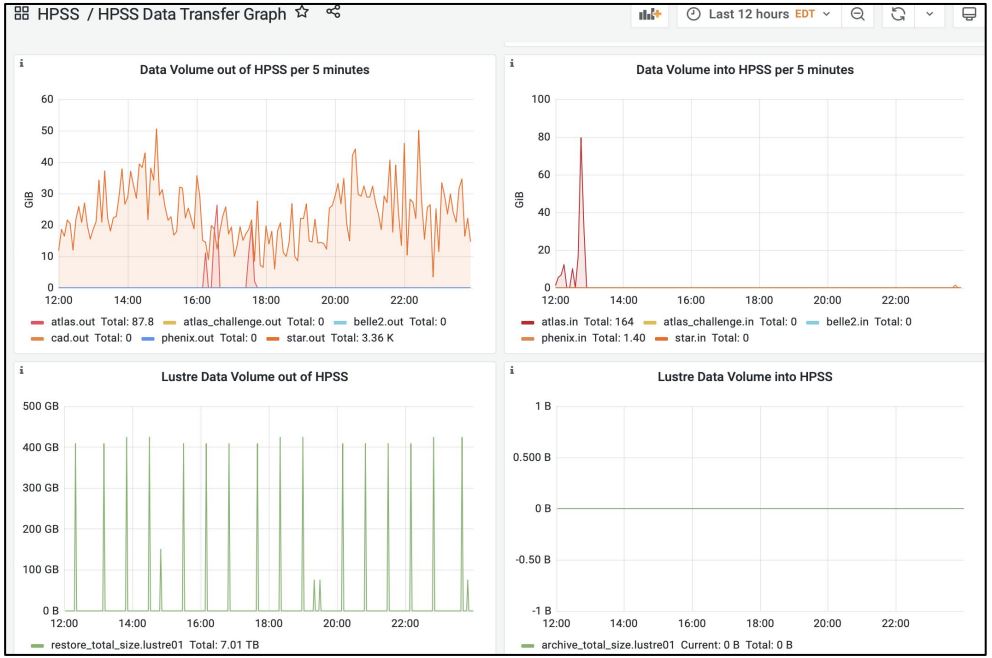
Repack LTO6 (2.5TB) to LTO8 (12TB) operations…

- Concurrent repack streams may be lower, while operations are busy
- Significant savings by retiring old LTO6 drives

System Monitoring

Grafana and MySQL DB

- Operational numbers such as network traffic, tape mounts, disk and tape usage … etc are monitored and recorded,
- Recorded numbers are displayed on Grafana

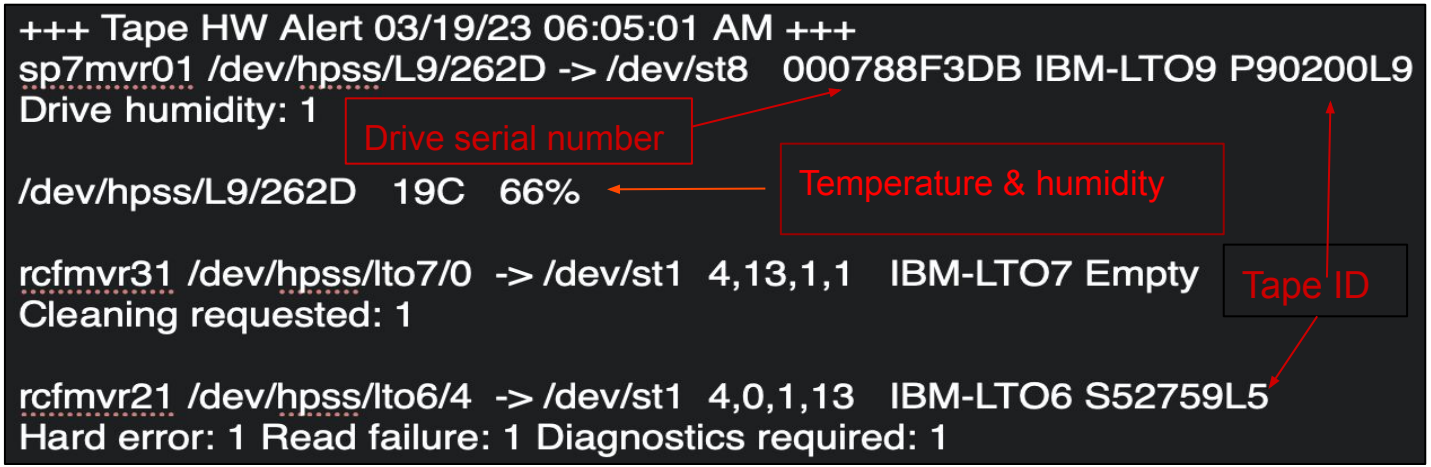


Brookhaven National Laboratory

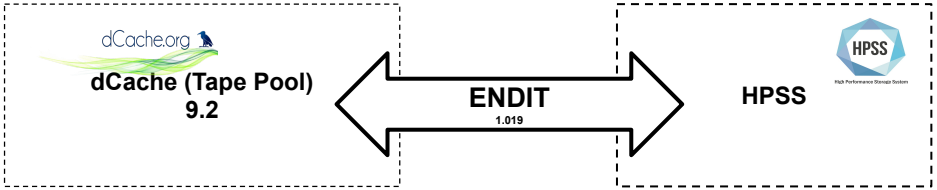
System Alerts

Alerts on software and hardware errors

- Email alerts are sent to related staffers on system errors, include software and hardware errors
  - A tape media error may trigger data repack to new tapes.
  - A humidity alert will inform Ops team to adjust Air Control systems.
  - A robotics or drive error will inform admins for investigations… etc.



dCache and HPSS tape interaction via ENDIT



ENDIT Provides

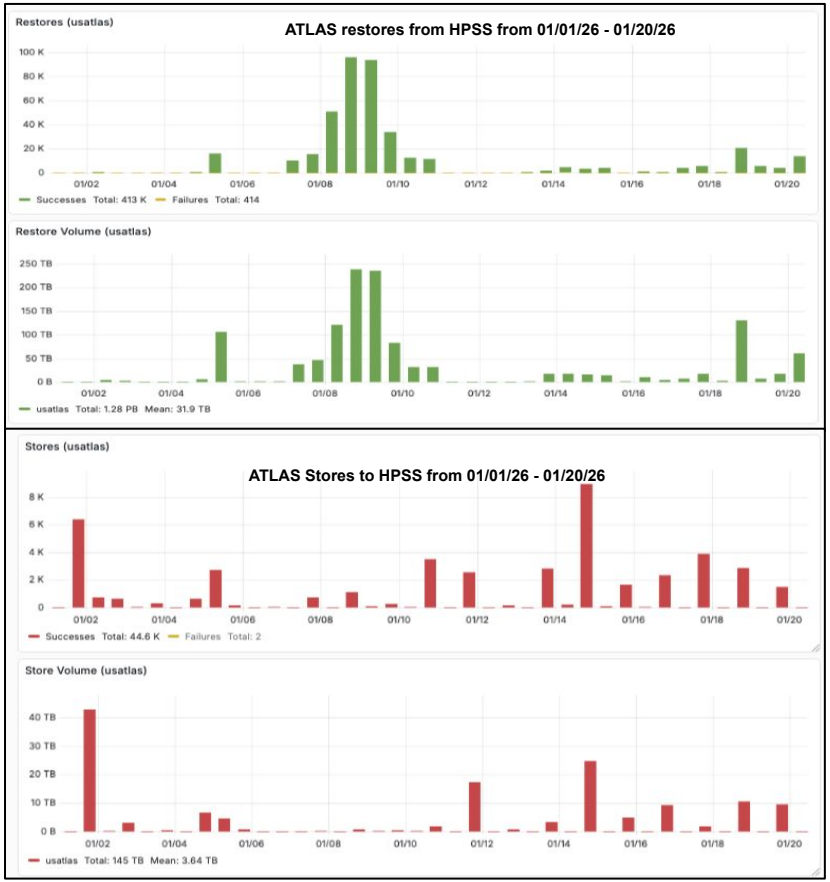
- Tape request orchestration and state tracking.

Operational Impact

- Successful adoption of the ENDIT retriever enabled extended support for write interactions to HPSS.
- Enabled consolidation of legacy software and custom code previously used for HPSS write workflows.

ENDIT Future Plans

- Evaluation and validation with the dCache 11.2 “golden” release, starting in 2026/02.

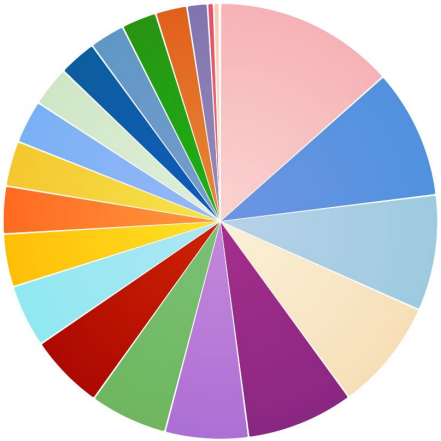


ATLAS dCache and HPSS systems configured to support 200k+ simultaneous staging requests

Brookhaven National Laboratory

Data Carousel at BNL

- RHIC experiments have been running in Data Carousel mode for over 20 years
- BNL HPSS heavily used by ATLAS Data Carousel
  a. >20% of data staged by ATLAS during Run3 from Tier-1/2 sites are from BNL Tier-1
- Besides the current tape optimization, we are also exploring tape optimization strategies based on archival metadata
  a. One particular solution we are looking into is the KIT one, unfortunately the KIT developer left ATLAS.



	Value	Percent
BNL-OSG2_MCTAPE	5 PB	14%
RAL-LCG2_MCTAPE	4 PB	10%
BNL-OSG2_DATATAPE	3 PB	9%
FZK-LCG2_MCTAPE	3 PB	8%
IN2P3-CC_MCTAPE	3 PB	8%
RAL-LCG2_DATATAPE	2 PB	6%
TRIUMF-LCG2_MCTAPE	2 PB	6%
INFN-T1_MCTAPE	2 PB	6%
FZK-LCG2_DATATAPE	2 PB	5%
TRIUMF-LCG2_DATATAPE	2 PB	4%
SARA-MATRIX_DATATAPE	1 PB	4%
NET2_MCTAPE	1 PB	3%
SARA-MATRIX_MCTAPE	1 PB	3%
IN2P3-CC_DATATAPE	1 PB	3%
INFN-T1_DATATAPE	1 PB	3%
NDGF-T1_MCTAPE	1 PB	3%
PIC_DATATAPE	1 PB	3%
NDGF-T1_DATATAPE	919 TB	2%
PIC_MCTAPE	583 TB	1%
RRC-KI-T1_MCTAPE	150 TB	0%
DESY-HH_MCTAPE	138 TB	0%
IN2P3-CC_GROUPTAPE	17 GB	0%

link on ATLAS DDM dashboard

Summary and future plans

- sPHENIX accomplished 25GB/sec injection rate to HPSS
  - Experience learned can be applied to ATLAS HL-LHC
  - SCDF is well-positioned to support DAQ streaming with ePIC at the EIC
- HPSS upgrade planned for 2027
- ATLAS disk cache and mover refresh planned for 2027
- Two new tape libraries with LTO11 drives planned for 2029
- Evaluation of file co-locationing on data sets
- dCache successfully adopted ENDIT to interact to tape
  - ATLAS dCache and HPSS systems configured to support 200k+ simultaneous staging requests
  - Evaluation for ENDIT prodiver for dCache 11.2 release starting 02/26

Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Tape Service at SDCF

Acknowledgements:

C. Gamboa, Q. Huang, J. Liu, A. Klimentov, J. Smith, T. Wong, Y. Wu, O. Novakov, X. Zhao

WLCG OTF #8, CERN, Geneva, Switzerland — February 3, 2026

@BrookhavenLab

Thank you

dCache Storage and HPSS Tape Services for ATLAS and Belle II

dCache Storage and Tape (HPSS)

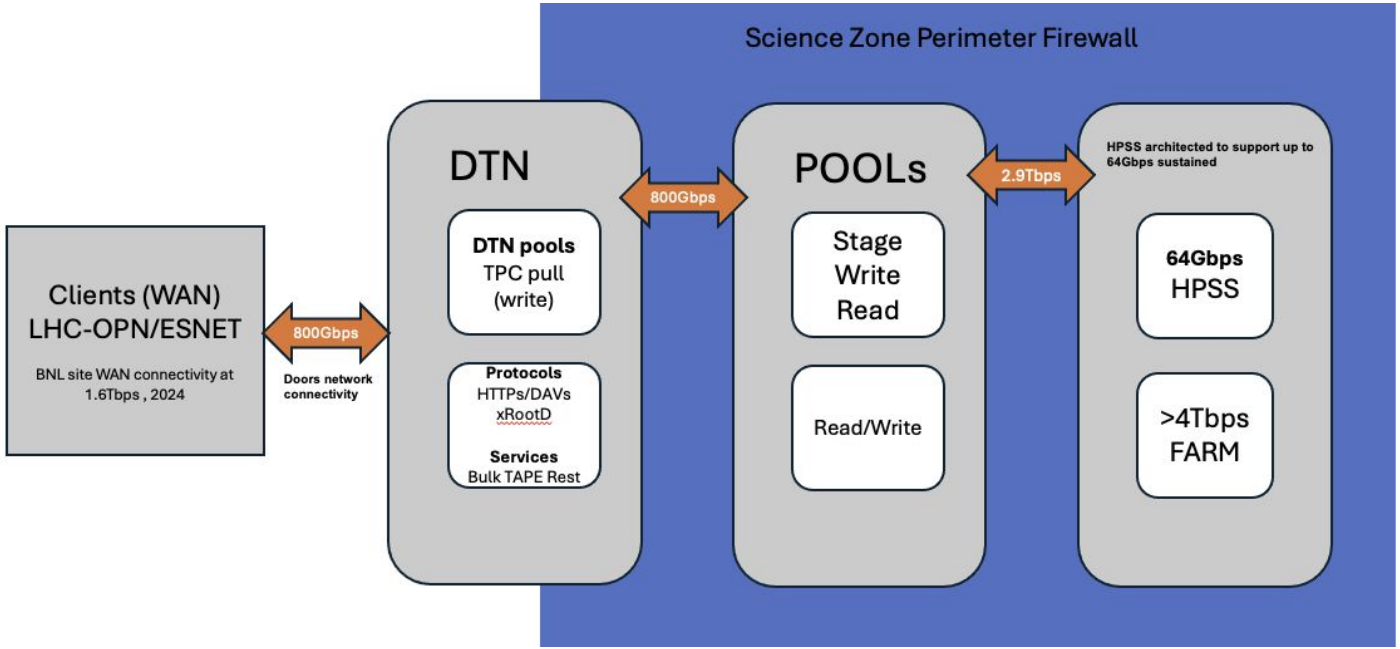
Experiment Integration

- Collaborations access HPSS tape services through the dCache Storage Element (Bulk TAPE service).
- BNL Storage Elements supports collaborations as a Tier 1 site
- dCache provides a unified interface between disk-based workflows and tape-backed storage.

Tape Usage

- >140.45 PB of data stored on HPSS tape
  - 97% ATLAS
  - 0.3% Belle II
- dCache–HPSS connectivity is tailored to experiment requirements.
- Belle II: smaller dedicated dCache instance, sustained throughput up to 2 GB/s to HPSS.
- ATLAS: sustained throughput up to 8 GB/s to HPSS

dCache general layout (ATLAS)



Staging Workflow with ENDIT HPSSRetriever