Operating dCache at Scale in US ATLAS Tier-1 and Tier-2 Centers

20th International dCache Workshop, NIKHEF, May 2026

Eduardo Bach (UMass Amherst, NET2), on behalf of US ATLAS

With inputs from:
Shawn McKee (U. Michigan, AGLT2)
Wendy Dronen (U. Michigan, AGLT2)
Carlos Gamboa (BNL)
Judith Stephen (U. Chicago, MWT2)
William Axel Leight (UMass Amherst, NET2)
Rafael Coelho Lopes de Sa (UMass Amherst, NET2)

US ATLAS dCache footprint

A federated Tier-1 / Tier-2 storage infrastructure

+ XRootD deployments shown on map

5 dCache deployments

226 servers

98.9 PB disk

190 PB tape

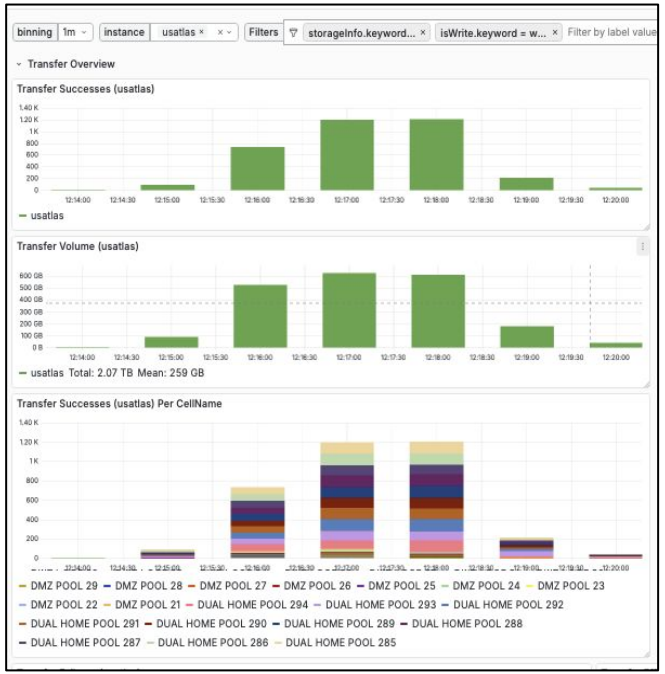
Site	Servers	Disk	Tape
AGLT2 (UM + MSU)	51	16.7 PB	—
NET2 (UMass + Harvard)	35	12.1 PB	50 PB
BNL Tier-1	74	45.0 PB	140 PB
MWT2 (IU)	13	5.1 PB	—
MWT2 (UChicago)	53	20.0 PB	—

US ATLAS dCache operations: in this presentation

- BNL Tier-1 scaling: disk, tape, network, and monitoring
- Tier-2 tape as a opportunistic complement to Tier-1 tape
- SENSE and SC25: testing engineered network paths for dCache traffic
- Redirection vs proxy mode at high throughput
- Toward common US ATLAS dCache telemetry: Kafka, OpenSearch, Firefly
    - For a dedicated discussion on Fireflies, see Shawn's talk.

BNL ATLAS Tier-1 dCache

- dCache v9.2.35+ , OS RH8, and PostgreSQL 16 deployed on redundant NVMe disks
  - Testing 11.2.[1→3] Golden Release on Integration/test instance - production upgrade planned for Summer 2026 (11.2.4 has fixes)
- Pool Infrastructure Expansion:13 pool servers commissioned into production, providing ~13 PB
  - 3 of 13 pool servers repurposed from former Lustre services
- Jumbo frames deployed across all Storage Element (SE) components
- Unify file checksum policy (OnTransfer) accross pools for performance
- Transitioning pools to dual-homed WAN + LAN connectivity
  - Phased deployment synchronized with hardware refresh cycles and capacity expansion
- Applied network kernel parameter tuning across pool and door nodes
  - Post-deployment observations showed WAN read throughput reaching ~98% of the storage network's installed capacity (97.7 GB/s out of 100 GB/s)
- Post-workshop studies on the Evolution of Storage at BNL, focusing on future requirements for HL-LHC and ePIC
  - In deep presentation about dCache services at BNL



Storage Element alerting platform transitioned to VictoriaMetrics, integrated with OpenSearch and Kafka for monitoring

BNL Tape / HPSS

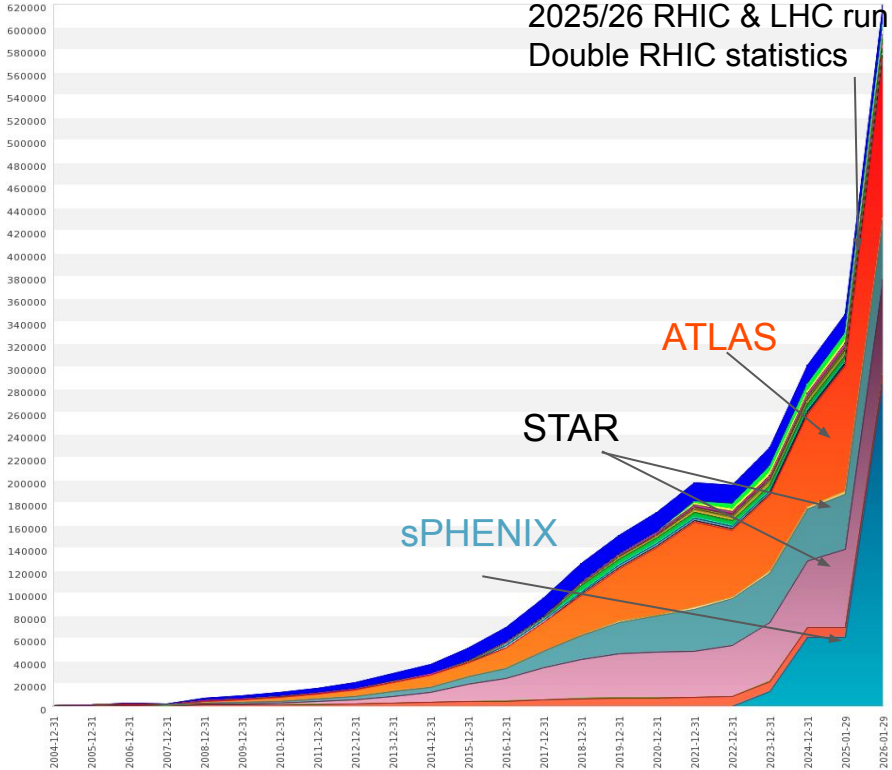
640 PB on tape the largest HEP/NP archive in the US

- 281.7 PB (51.2M files) written & 56.3PB (16.1M files) read in 2025
- 48.9 PB (5.9M files) written & 10.8 PB (4M files) read in first 3 months of 2026
- ATLAS archive data size – 140.0 PB

I/O statistics for ATLAS Data Carousel

- 41.7 PB (9+M files) ATLAS staged in 2025
- 28.9 PB (9+M files) ATLAS injected in 2025
- ATLAS read requests > write
- **Ability to sustain 8 GB/sec (sPHENIX sustained I/O 25 GB/sec)**

ATLAS movers and gateways upgraded to RHEL8



Tape usage, 2004-present

Infrastructure Supporting dCache Services

Core Cells: redundant deployment to ensure high availability.

Core databases: in a primary standby replication mode using postgresql database

Doors : Equipped with 2×25 Gbps internal links and 2×25 Gbps external (WAN) links, providing dual-path connectivity for both LAN and WAN data flows.

DMZ Pools (NVMe) : Handle TPC WRITE operations from external sites. Upon completion, data is automatically flushed into the internal pool infrastructure.

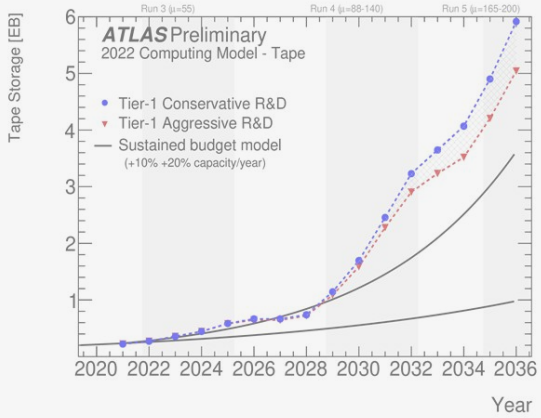
Internal Pools : Connected through 2×25 Gbps internal network links, supporting high-throughput data access and bulk internal workflows.

Dual-Home Pools : Configured with 2×25 Gbps internal links and 2×25 Gbps external (WAN) links, enabling direct WAN and LAN data movement as part of the dual-home deployment model.

dCache instance	Number of VMs+Physical Hardware(PH)	dCache Version	Notes
ATLAS	74(99%PH)	9.2.35	40 PB of data were migrated in 2024 during the transition from MDRAID to ZFS. 1 file replica
Pre-production/Test (AKA dcint)	12(8%PH)	11.2.3	WLCG REST API test endpoint Integrated with ATLAS DDM test infrastructure Dual pool home studies EPiC tests

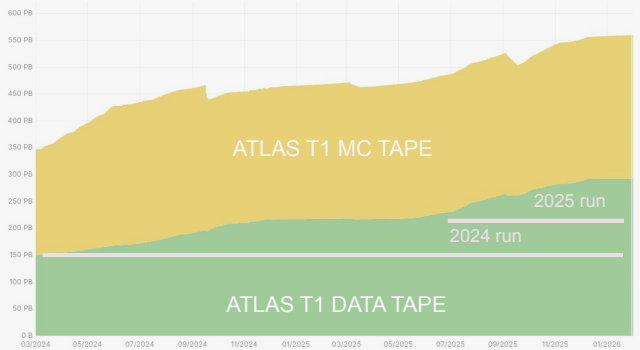
The US ATLAS Northeast Tier 2 Tape

Shared Tape Library - only ~$12/TB tape + $1/TB/year op. Tier 2”-like support: 8x5, no on-call nor alarm tickets



ATLAS Software and Computing HL-LHC Roadmap

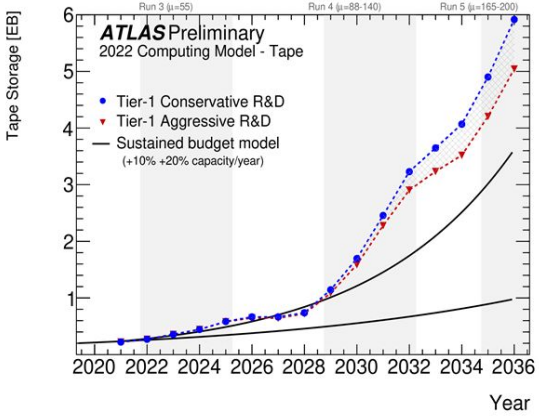
There is a huge deficit in tape resources for HL-LHC



Today, only ~35% of Tier-1 tape is used for RAW; most is AOD/MC-like data.

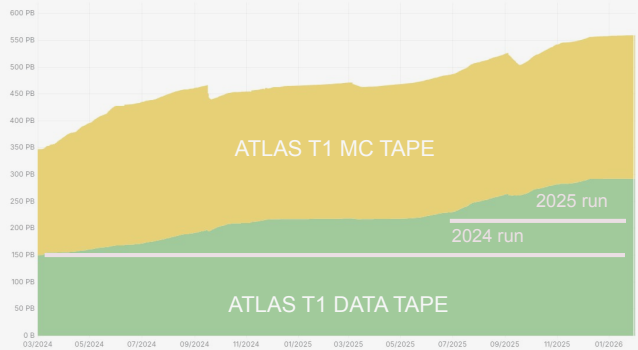
The US ATLAS Northeast Tier 2 Tape

Shared Tape Library - only ~$12/TB tape + $1/TB/year op. Tier 2"-like support: 8x5, no on-call nor alarm tickets



[ATLAS Software and Computing HL-LHC Roadmap]

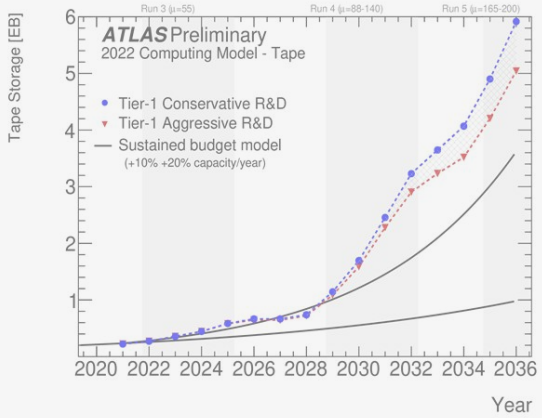
There is a huge deficit in tape resources for HL-LHC



Today, only ~35% of Tier-1 tape is used for RAW; most is AOD/MC-like data.

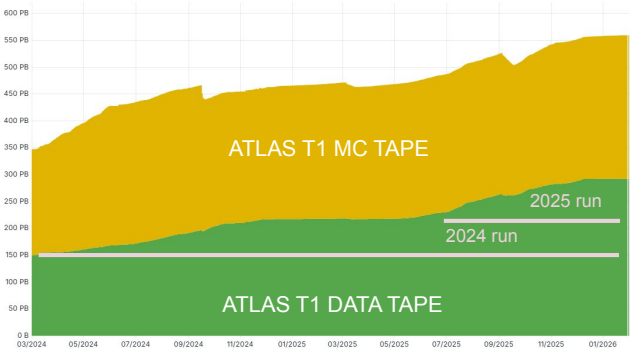
The US ATLAS Northeast Tier 2 Tape

Shared Tape Library - only ~$12/TB tape + $1/TB/year op. Tier 2"-like support: 8x5, no on-call nor alarm tickets



[ATLAS Software and Computing HL-LHC Roadmap]

There is a huge deficit in tape resources for HL-LHC



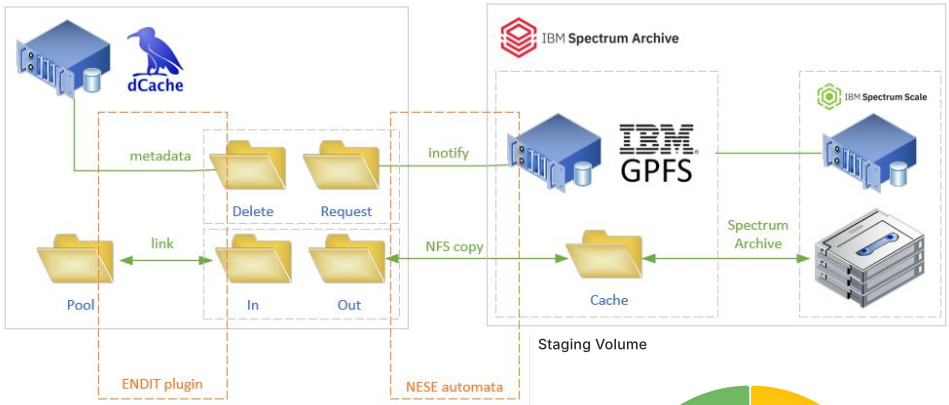
Today, only ~35% of Tier-1 tape is used for RAW; most is AOD/MC-like data.

dCache and the NET2 tape

All communication with the tape backend is **indirect**, via folder-based signaling and file copy operations.

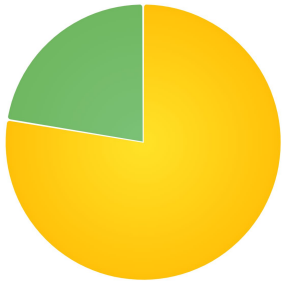
We use the **ENDIT plugin** on the dCache pool to trigger metadata operations (hardlinks, request tracking).

A custom component, called the **NESE automata**, replaces the ENDIT daemons.



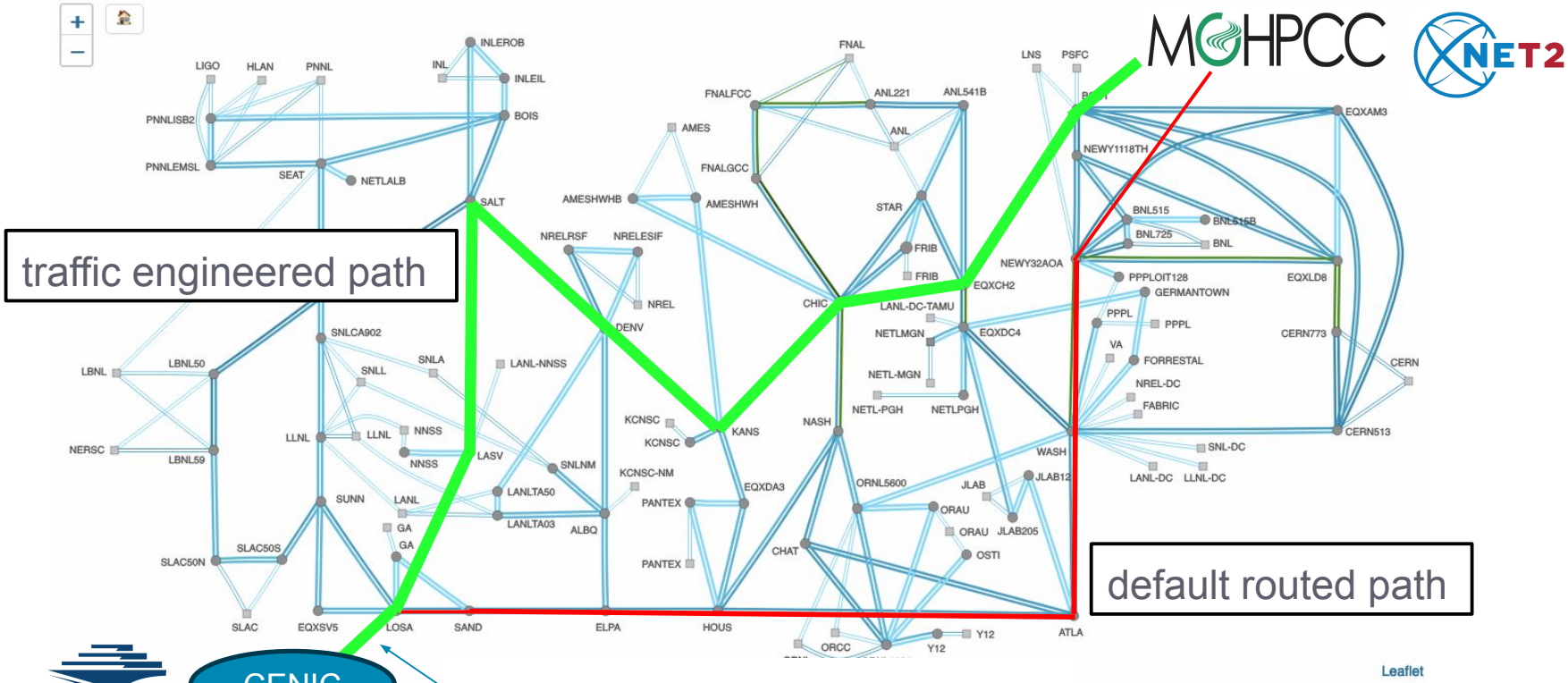
NET2 Delivered 22% of MC like data staged from tape in US ATLAS on the past 6 months.

Staging Volume



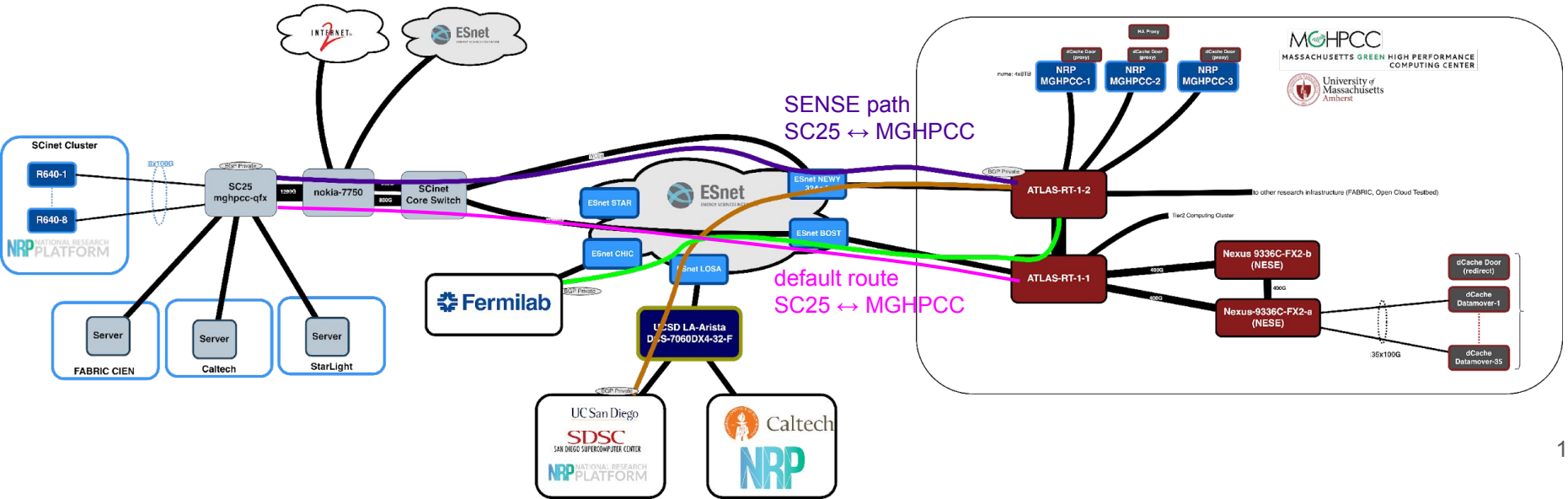
	Value	Percent
BNL-OSG2_MCTAPE	11 PB	78%
NET2_MCTAPE	3 PB	22%

Steering dCache traffic onto engineered paths using SENSE

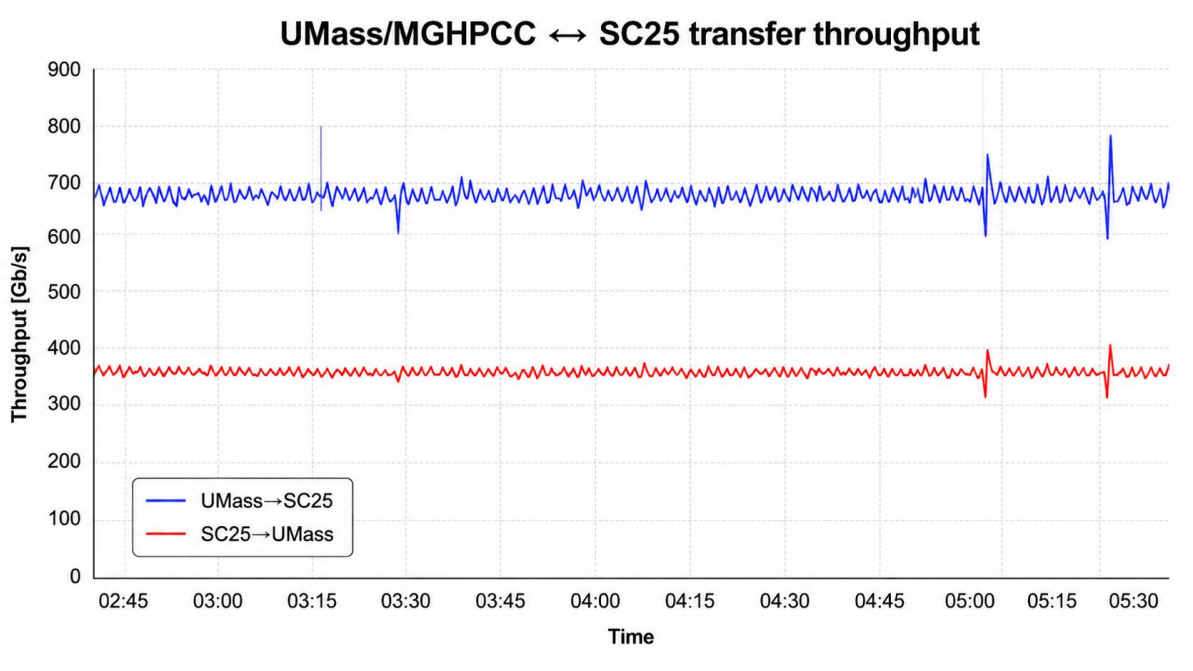


NET2/SENSE SC25 demo setup

- 400Gbps (default route) + 400Gbps (SENSE path)
- **Dedicated docker dCache images created to be used in the NRP k8s**
  - Webdav + Xrootd doors - proxy mode



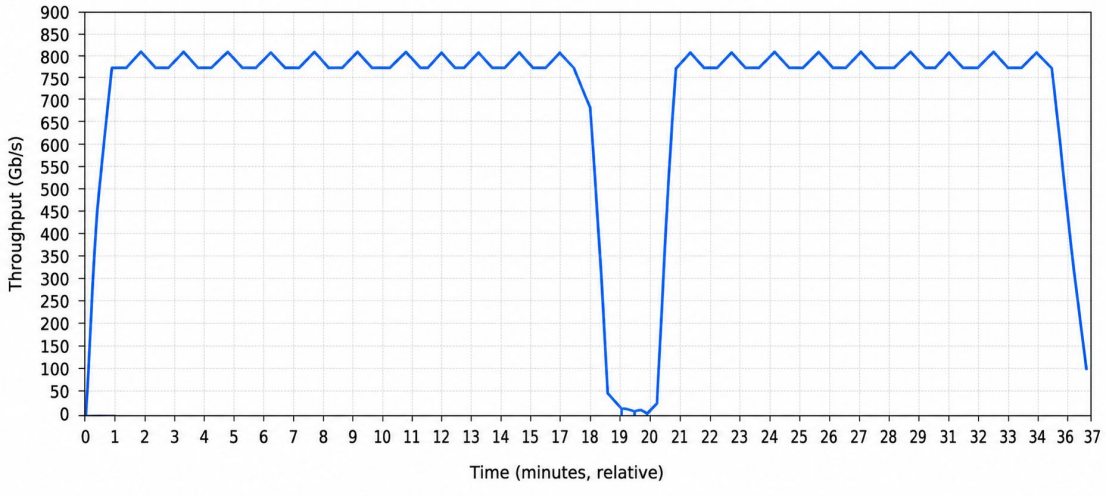
NET2/SENSE SC25 demo result



Data transfers from NESE to the SCinet cluster and to the Caltech booth

- **gfal-copy**: 700 Gb/s (UMass→SC25) + 400 Gb/s (SC25→UMass)
- Redirection mode door used.

SC25 preparation: dCache in redirection mode + encryption @800Gbps



- 800Gbps connecting the Cluster (150 nodes) and dCache cluster
- One door - redirection mode
  - `webdav.limits.acceptors=10`
  - `webdav.limits.threads.max=2000`
  - `webdav.limits.backlog=20000`
  - `webdav.limits.idle-time=900`
  - `webdav.limits.queue-length=2000`
  - `webdav.limits.threads.min=33`
  - `webdav.mover.timeout=28800000`
  - `Tcp tuning`
- 35 disk servers [27 zfs (tuned) + 8 hardware raid (only read ahead)]
- Transfers between computing cluster and dCache cluster (same datacenter)

Encryption performance comparison (webdav.redirect.allow-https=false/true)
Right - true
Left - false

SC25 preparation: proxy-mode doors and TLS @800Gbps

Proxy mode doors observed throughput on Gen4/5 node with 4x100 Gb/s bonded

- 1 WebDAV proxy door: ~100 Gb/s
- 2 WebDAV proxy doors: ~160–180 Gb/s
- 4 WebDAV proxy doors: ~200 Gb/s

Throughput improves with additional doors, but with clear diminishing returns. At four doors, jetty-* threads consume all the available CPU (100 cores).

TLS JSSE provider comparison

- In the tested proxy-mode configuration, TLS was on the client → door leg. Billing/logging showed that the door → pool leg was unencrypted.
- Tested Conscrypt, a BoringSSL-backed JSSE provider, instead of the default SunJSSE/SunJCE stack.
  - Difference was negligible in the tested configuration.
  - The TLS conclusion is preliminary because only Conscrypt (BoringSSL) was tested.

USATLAS dCache telemetry: logs, billing, and flow visibility

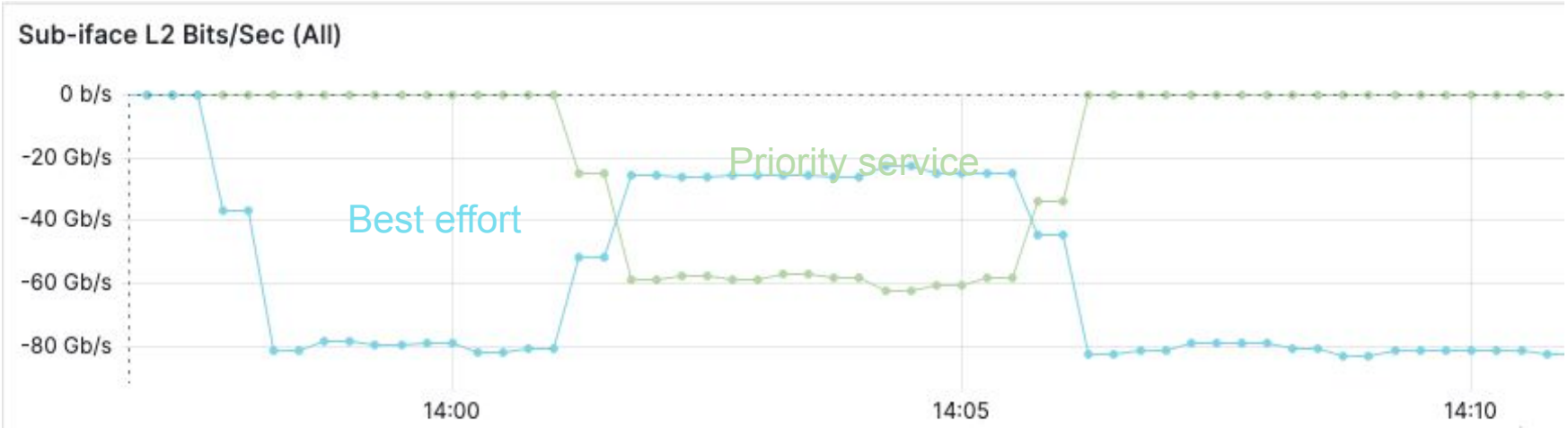
- BNL + AGLT2 + NET2: OpenSearch ingest from Kafka-exported dCache data
  - NET2 : dCache → Kafka → Logstash → OpenSearch → Grafana
  - AGLT2: dCache → independent ZooKeeper/Kafka cluster → OpenSearch → Grafana
  - BNL: VictoriaMetrics alerting integrated with OpenSearch/Kafka monitoring
- Firefly testing: five prereleases tested; most features working; one fix pending
  - Details in Shawn McKee's talk

Thank you.

Testing between US ATLAS sites

First successful test between two ATLAS sites during the latest capability mini-challenge! (see other capability tests in Shawn's talk)

FDT: 80 Gb/s circuit, 60 Gb/s requested for high-priority transfer



Infrastructure Supporting dCache Services

Core Cells: redundant deployment to ensure high availability.

Core databases: in a primary standby replication mode using postgresql database

Doors : Equipped with 2×25 Gbps internal links and 2×25 Gbps external (WAN) links, providing dual-path connectivity for both LAN and WAN data flows.

DMZ Pools (NVMe) : Handle TPC WRITE operations from external sites. Upon completion, data is automatically flushed into the internal pool infrastructure.

Internal Pools : Connected through 2×25 Gbps internal network links, supporting high-throughput data access and bulk internal workflows.

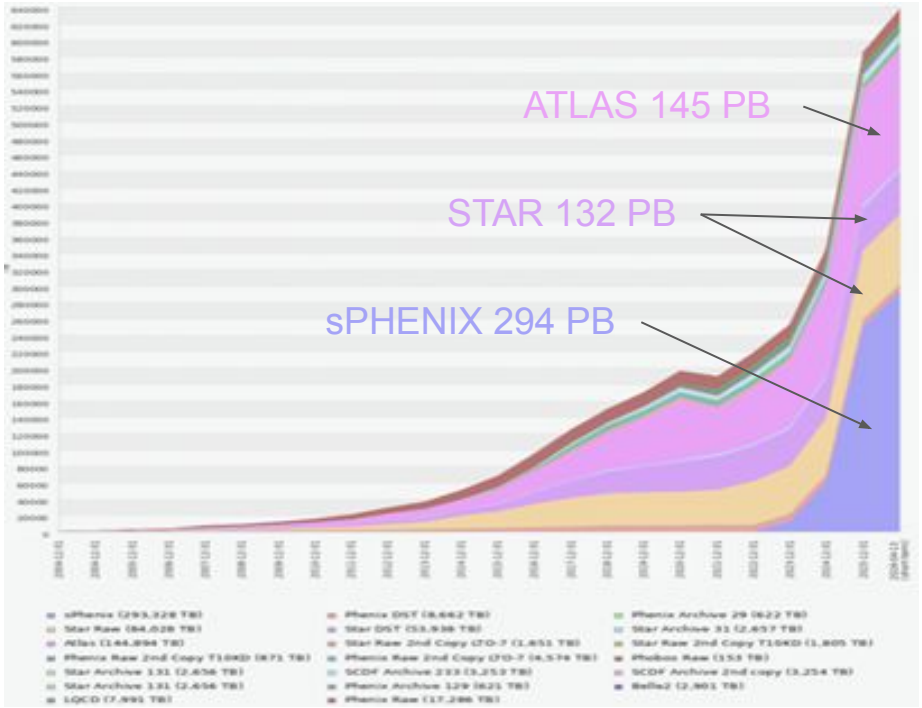
Dual-Home Pools : Configured with 2×25 Gbps internal links and 2×25 Gbps external (WAN) links, enabling direct WAN and LAN data movement as part of the dual-home deployment model.

dCache instance	Number of VMs+Physical Hardware(PH)	dCache Version	Notes
ATLAS	74(99%PH)	9.2.35	40 PB of data were migrated in 2024 during the transition from MDRAID to ZFS. 1 file replica
BELLE2	13(92%PH)	9.2.35	Upgrades are subject to a yearly schedule, primarily taking advantage of detector downtime. 1 file replica
DUNE	12(42%PH)	9.2.35	Legacy hardware in a resilient configuration 2 copy/file
Pre-production/Test (AKA dcint)	12(8%PH)	11.2.3	WLCG REST API test endpoint Integrated with ATLAS DDM test infrastructure Dual pool home studies EPiC tests

Tape Storage

Approximately 640 PB of scientific data

- The largest HEP/NP data archive in the USA
- 281.7 PB (51.2M files) written & 56.3PB (16.1M files) read in 2025
- 48.9 PB (5.9M files) written & 10.8 PB (4M files) read in first 3 months of 2026



US ATLAS and dCache usage