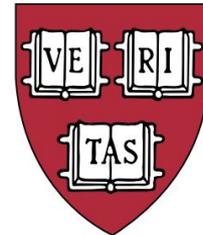


Tape systems for WLCG Tier 2 sites

ATLAS NET2 experience

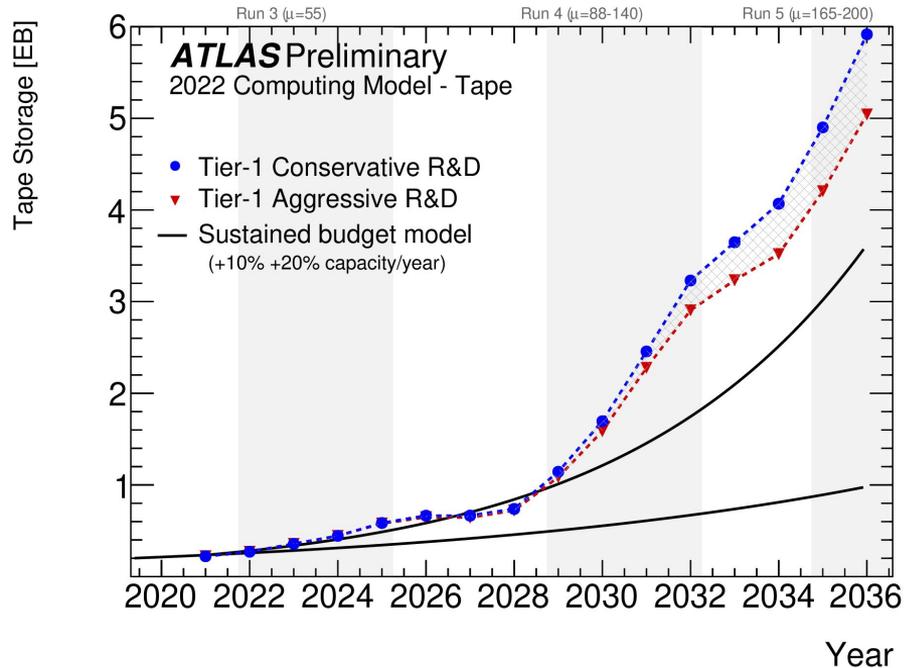
Throughput Computing 2025
Madison, WI
6/5/2025

Eduardo Bach, Rafael Coelho Lopes de Sa, Verena Martinez Outschoorn (UMass Amherst)
Milan Kupcevic (Harvard)



Tape usage at (HL-)LHC

The most recent [ATLAS Software and Computing HL-LHC Roadmap](#) indicates that, even under aggressive R&D, there is a huge deficit in tape resources for HL-LHC.



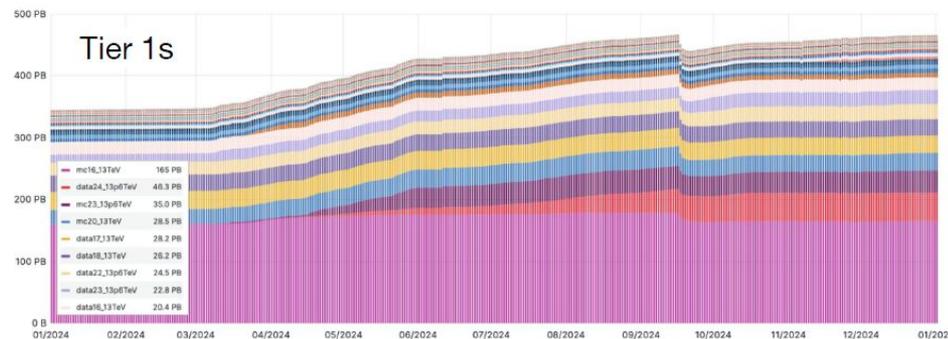
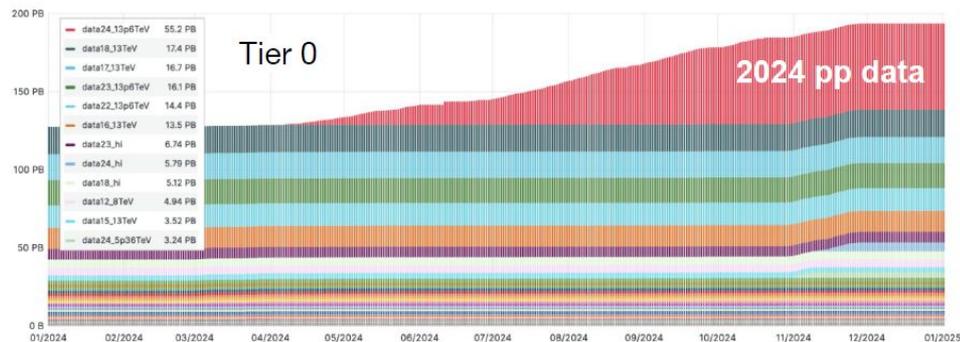
Tape usage at (HL-)LHC

The most recent [ATLAS Software and Computing HL-LHC Roadmap](#) indicates that, even under aggressive R&D, there is a huge deficit in tape resources for HL-LHC.

2024 was a punishing year for tape systems. The exceptional performance of the LHC made it difficult to maintain enough resources at Tier 1s.

2025 is going to be longer than originally planned and we are going to have data in 2026.

Can we use Tier 2 to store some data types and relieve the pressure on Tier 1s?



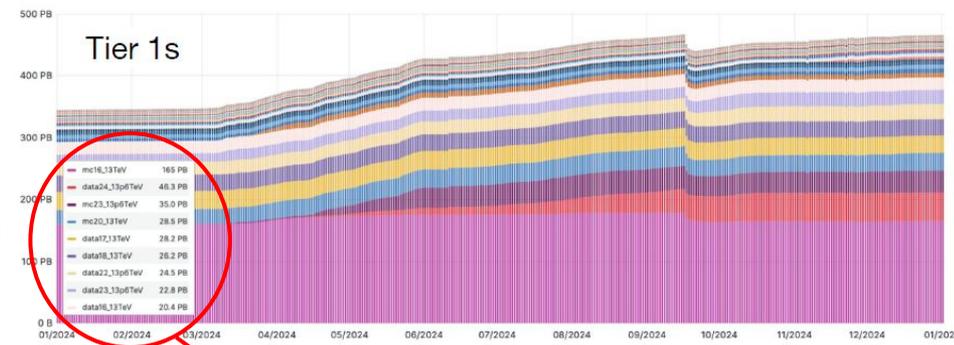
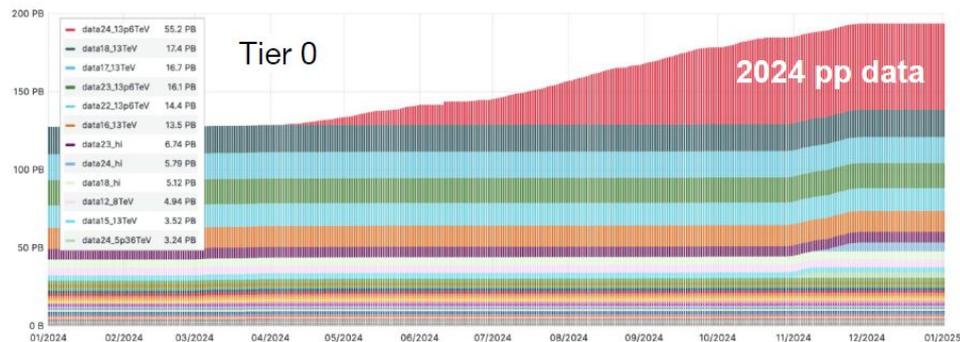
Tape usage at (HL-)LHC

The most recent [ATLAS Software and Computing HL-LHC Roadmap](#) indicates that, even under aggressive R&D, there is a huge deficit in tape resources for HL-LHC.

2024 was a punishing year for tape systems. The exceptional performance of the LHC made it difficult to maintain enough resources at Tier 1s.

2025 is going to be longer than originally planned and we are going to have data in 2026.

Can we use Tier 2 to store some data types and relieve the pressure on Tier 1s?



Tier 1 tape usage: 58% MC, 42% DATA

The US ATLAS NET2 site

The NET2 site is one of the four US ATLAS Tier 2s.

It is operated by UMass Amherst and it is located at the Massachusetts Green High Performance Computing Center (MGHPCC) in Holyoke, MA.



MGHPCC



The data center hosts the [Northeast Storage Exchange \(NESE\)](#) service that provides disk and tape storage for all the universities that use the center.

NET2 uses NESE for all its RSEs

NESE

THE NORTHEAST STORAGE EXCHANGE



HARVARD
UNIVERSITY



Massachusetts
Institute of
Technology



Northeastern
University



Red Hat



UMASS

Yale

Intermezzo 1: Massachusetts geography



M. Goncharov, [Evaluating Tape Storage at MIT](#)

That's not where UMass Amherst is...

Intermezzo 1: Massachusetts geography



M. Goncharov, [Evaluating Tape Storage at MIT](#)



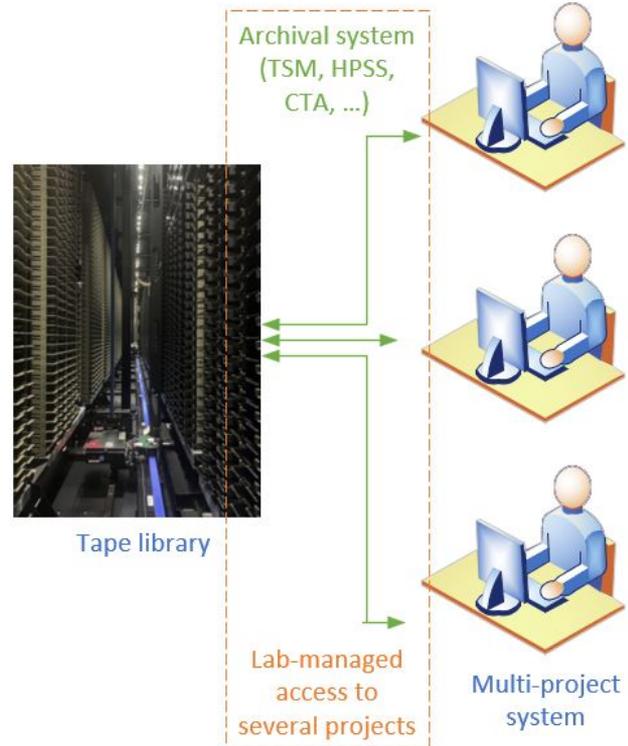
That's not where UMass Amherst is...

Tape system at shared facilities

It is not uncommon for sites to share their tape system among several projects.

In most (all?) Tier 1 sites, the projects have direct access to the tape archival system.

This is because the site can impose strict conditions (protocol, security, ...) on the systems used for communication with the tape system.



Tape system at shared facilities

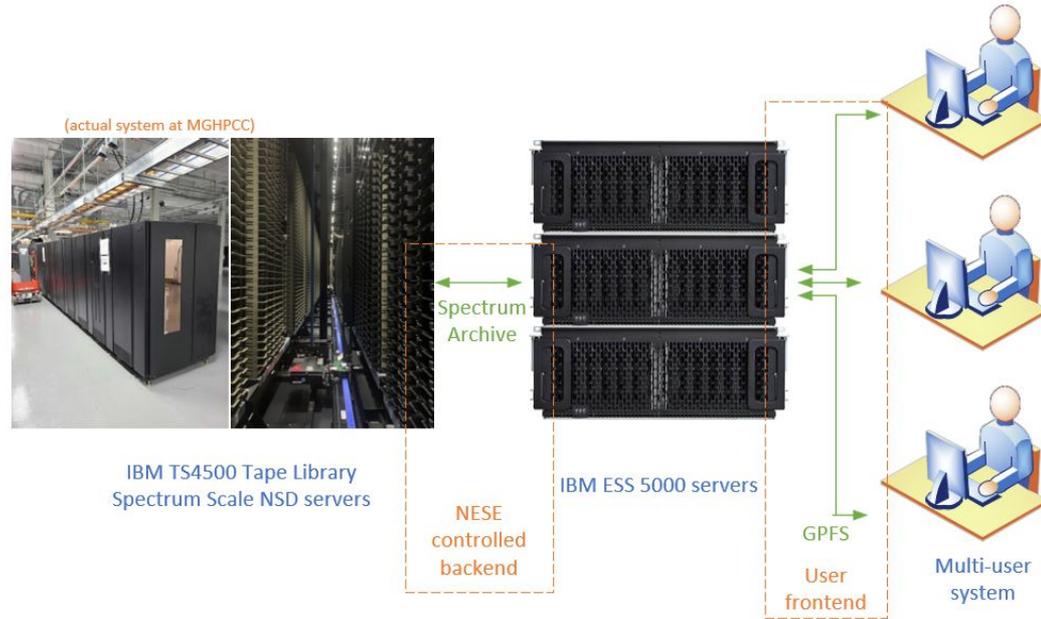
It is not uncommon for sites to share their tape system among several projects.

In most (all?) Tier 1 sites, the projects have direct access to the tape archival system.

This is because the site can impose strict conditions (protocol, security, ...) on the systems used for communication with the tape system.

This is not the case for a shared tape system that needs to serve a large number of heterogeneous systems in a data center.

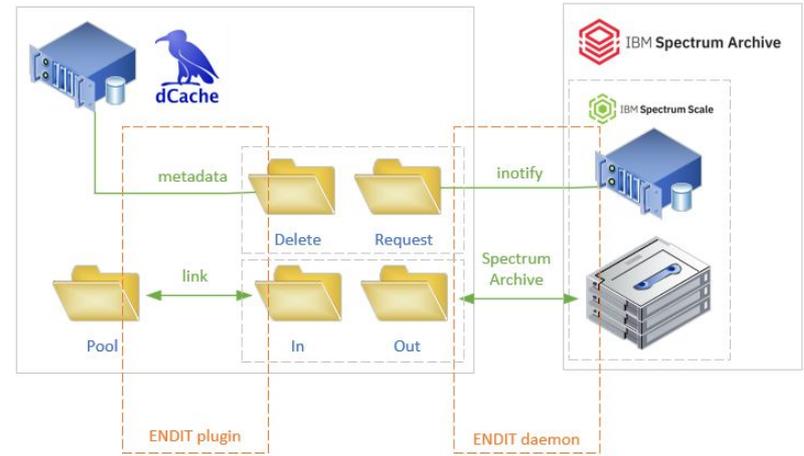
At NESE, we do not have direct access to the tape archival system.



dCache, ENDIT, and NESE

Traditional dCache setups using ENDIT manage tape via a direct tape interface using ENDIT daemons (`dsmc archive` and `dsmc retrieve`, etc.).

In our setup, **we do not control the tape system**, but we do have access to its **GPFS cache layer via NFSv4**.



More details in Mattias Wadenstein [talk at HEPiX Autumn 2023](#)

dCache, ENDIT, and NESE

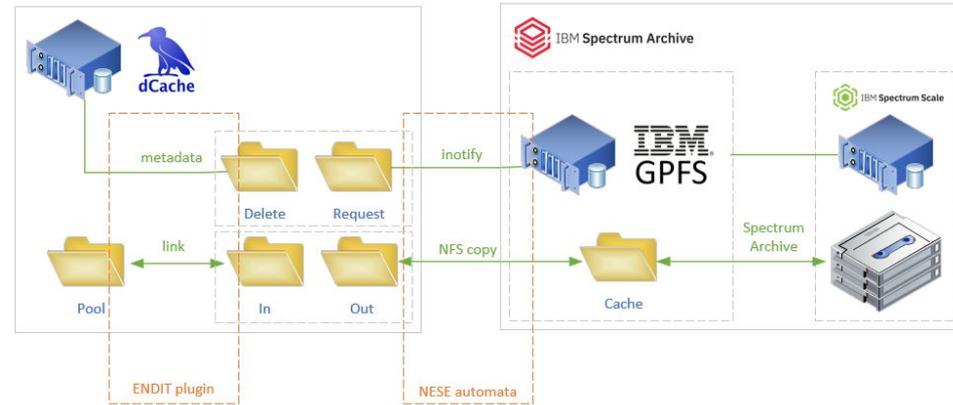
Traditional dCache setups using ENDIT manage tape via a direct tape interface using ENDIT daemons (`dsmc archive` and `dsmc retrieve`, etc.).

In our setup, **we do not control the tape system**, but we do have access to its **GPFS cache layer via NFSv4**.

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.

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



NESE automata developed by M. Kupcevic and E. Bach

dCache, ENDIT, and NESE

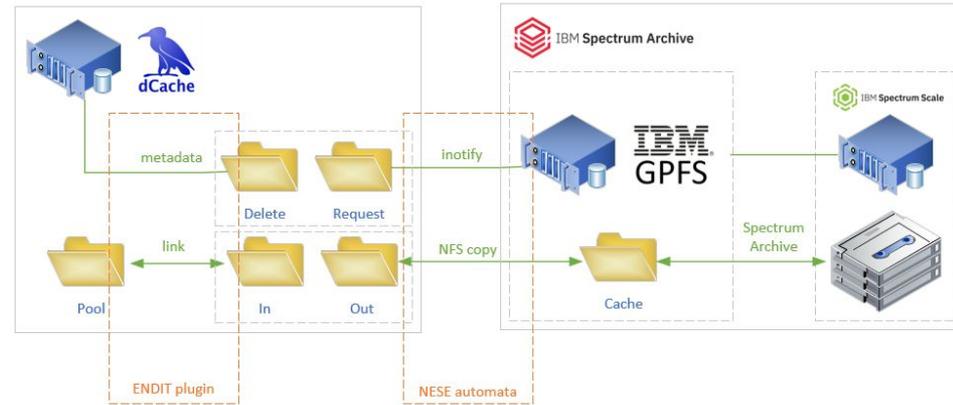
Traditional dCache setups using ENDIT manage tape via a direct tape interface using ENDIT daemons (`dsmc archive` and `dsmc retrieve`, etc.).

In our setup, **we do not control the tape system**, but we do have access to its **GPFS cache layer via NFSv4**.

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.

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

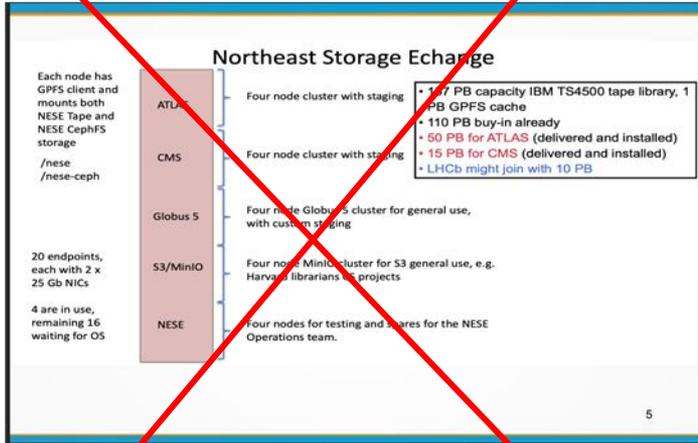


The ENDIT plugin uses **hard links** to expose retrieved files in the `Pool/` directory, enabling efficient use of storage by avoiding data duplication or movement across directories.

We are using the **endit-watching** version of the plugin, which relies on `inotify` to detect new metadata events.

`inotify` offers **very low CPU overhead**, but introduces a limitation: the file restoration rate is approximately **1 file per second** (1 Hz), which can be a bottleneck for high-volume & non-bulk recalls.

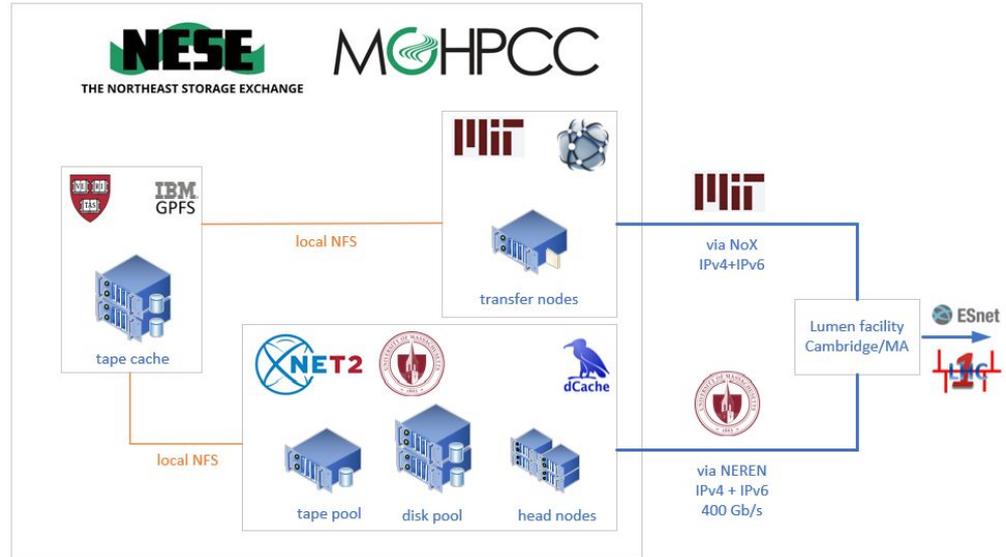
Intermezzo 2: same, but different



M. Goncharov, [Evaluating Tape Storage at MIT](#)

US ATLAS NET2 does not use this scheme.

This scheme was designed for small-scale users. The requirements for large LHC experiments needed a new storage scheme in NESE.



All the NET2 storage, including disk pool, tape pool and head nodes are part of NESE.

A model for other Tier 2 sites

In many institutions, **tape systems exist but are not administratively managed** by the same teams (running dCache).

These tape systems often expose their **cache layers via filesystems** (e.g., **GPFS over NFS**), but lack the typical command-line or API control (e.g., no access to run `dsmc archive/retrieve`).

Our approach **bridges this gap**:

- Using the **ENDIT plugin** (without ENDIT daemons).
- Replacing direct tape commands with a lightweight **automata** that manages metadata and copies files to/from known tape cache locations.
- Leveraging **hard links** for efficient data exposure to dCache.
- Running in **endit-watching mode** via `inotify` for minimal CPU overhead.

The UMass-NESE design

Our design is:

- **Non-invasive** (no changes to the tape infrastructure).
- **Scalable** to other sites with similar access constraints.
- A **template** for how more tape resources can be brought into ATLAS without requiring centralized control or privileged access.

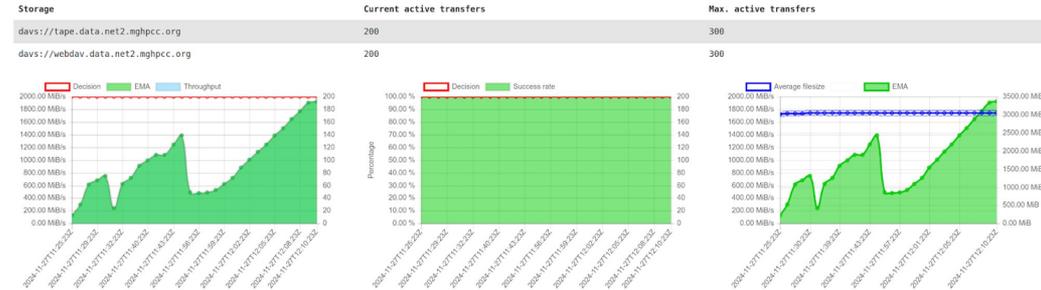
We hope this setup can **inspire other sites** to integrate their locally accessible tape systems into the global ATLAS computing fabric.

Initial tests, commissioning, and operation

The system was made available to ATLAS in the Fall of 2024 and the first tests using FTS were performed in November 2024

The tape system and the new interface worked flawlessly.

Details for [davs://tape.data.net2.mghpcc.org](https://tape.data.net2.mghpcc.org) → [davs://webdav.data.net2.mghpcc.org](https://webdav.data.net2.mghpcc.org)



Overview

Showing 1 to 1 out of 1 from the last 1 hour

First Previous 1 Next Last

Source	Destination	V0	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	Rate (last 1h)	Thr.
davs://mc-tape.data.net2.mghpc	davs://dcintdoor.sdcc atlas		0	2	342	10353	0	1012	0	0	100.00 %	425.04 TiB/s
			0	2	342	10353	0	1012	0	0	100.00 %	-

First Previous 1 Next Last

- Bad shape** There are submitted but no active, less than 3 active with more than 3 submitted, or a failure rate >= 20%
- Undesired** Less than three actives, but no submitted waiting.
- Good shape** Success rate >= 90%, or more than three actives with a failure rate < 20%.
- Nothing special** No active, no submitted, success rate between 80% and 90%.

100% success transfers in testing and commissioning

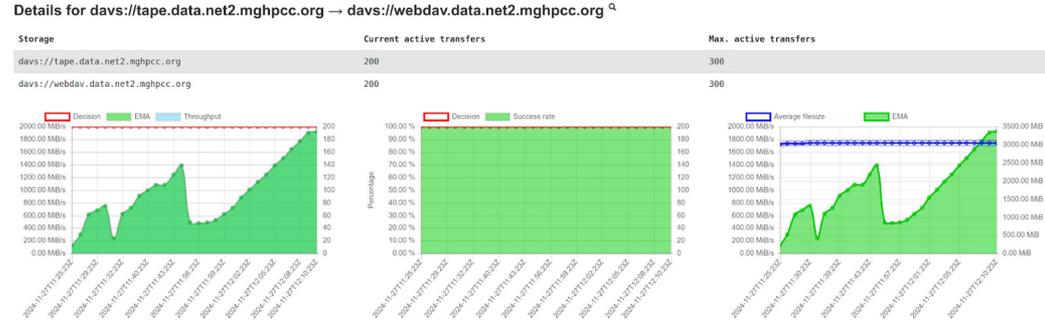
Initial tests, commissioning, and operation

The system was made available to ATLAS in the Fall of 2024 and the first tests using FTS were performed in November 2024

The tape system and the new interface worked flawlessly.

After the EoY break, NET2_MCTAPE was added to Rucio.

Since the beginning, our dCache instance has exposed a WLCG Tape REST-compliant interface, supporting standard operations — **status**, **stage**, **release**, **cancel**, — through our RSE endpoint via the well-known discovery mechanism.



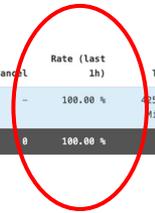
Overview

Showing 1 to 1 out of 1 from the last 1 hour

Source	Destination	V0	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	Rate (last 1h)	Thr.
davs://mc-tape.data.net2.mghpc	davs://dcindoor.sdcc atlas		0	2	342	10353	0	1012	0	0	100.00 %	425.04 TiB/s

- Bad shape** There are submitted but no active, less than 3 active with more than 3 submitted, or a failure rate >= 20%
- Undesired** Less than three actives, but no submitted waiting.
- Good shape** Success rate >= 90%, or more than three actives with a failure rate < 20%.
- Nothing special** No active, no submitted, success rate between 80% and 90%.

100% success transfers in testing and commissioning



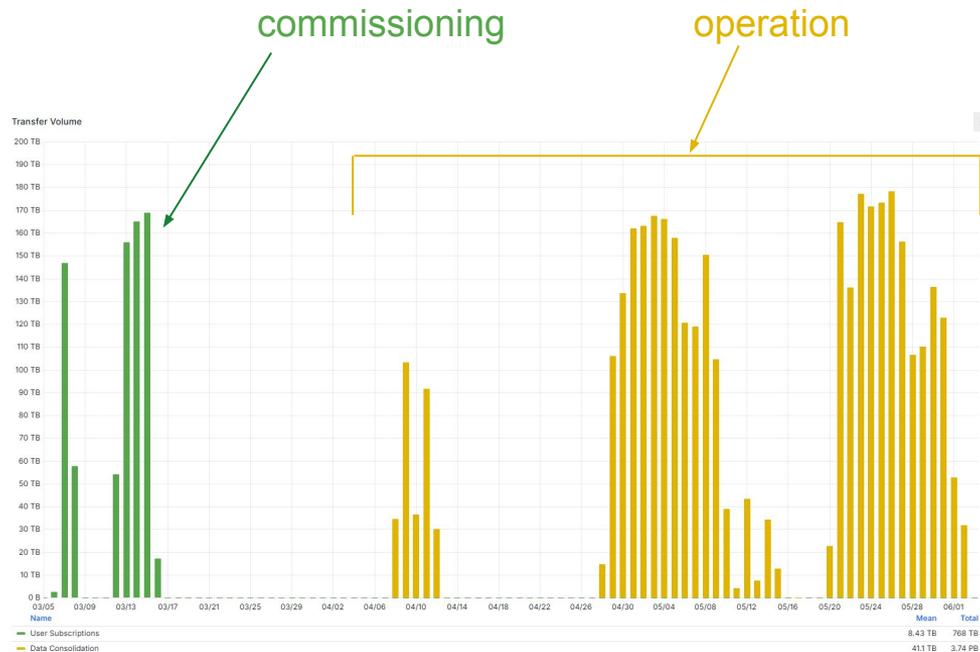
Initial tests, commissioning, and operation

The system was made available to ATLAS in the Fall of 2024 and the first tests using FTS were performed in November 2024

The tape system and the new interface worked flawlessly.

After the EoY break, NET2_MCTAPE was added to Rucio.

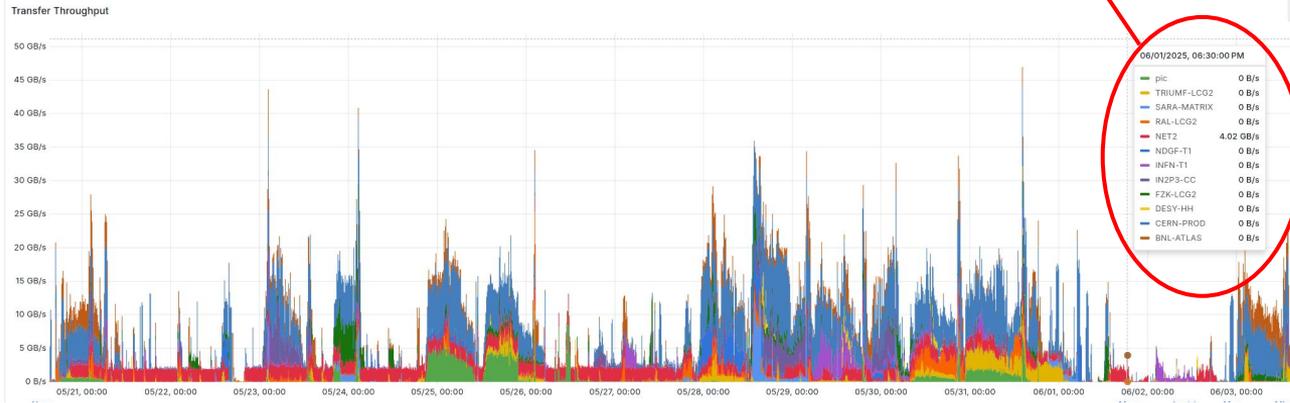
A period of commissioning followed in March 2025. Operations started in April 2025.



(daily volume for the last 3 months)

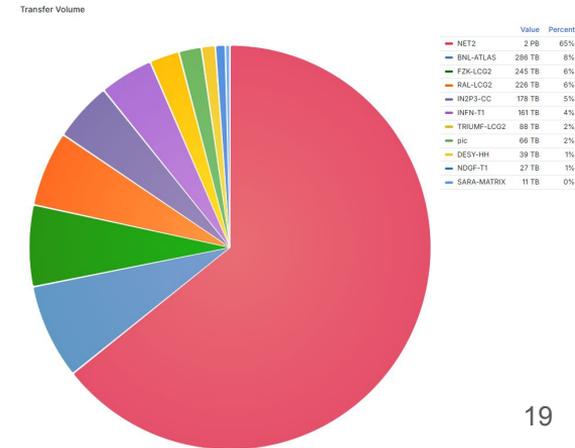
Current performance

The system can currently reach up to 4 GB/s throughput (simultaneous reading + writing). But further development is underway and we are limiting to 2 GB/s in the meantime. We are working to improve the the throughput as much as possible.



(past 14 days)

NET2_MCTAPE continues to be filled at full speed to increase usage.



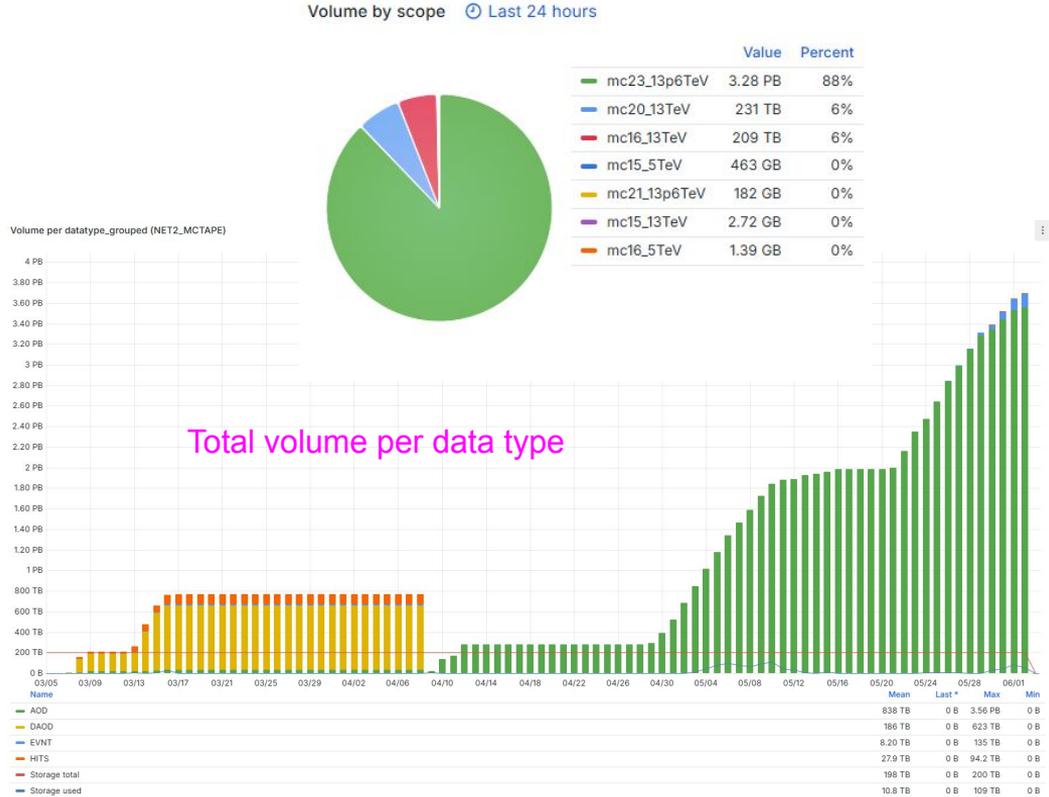
Workflows

A dedicated campaign to fill the tape with recent AOD (mostly mc23) has started during the DAOD rederivation in April 2025

We were able to participate in the very end of the campaign.

We are now receiving copies of EVNT, HITS, and AODs. The tape system is being used normally like the T1s by Data Carousel to serve as input for production

Tier 0 exports still to pledged Tier 1s only.



Workflows

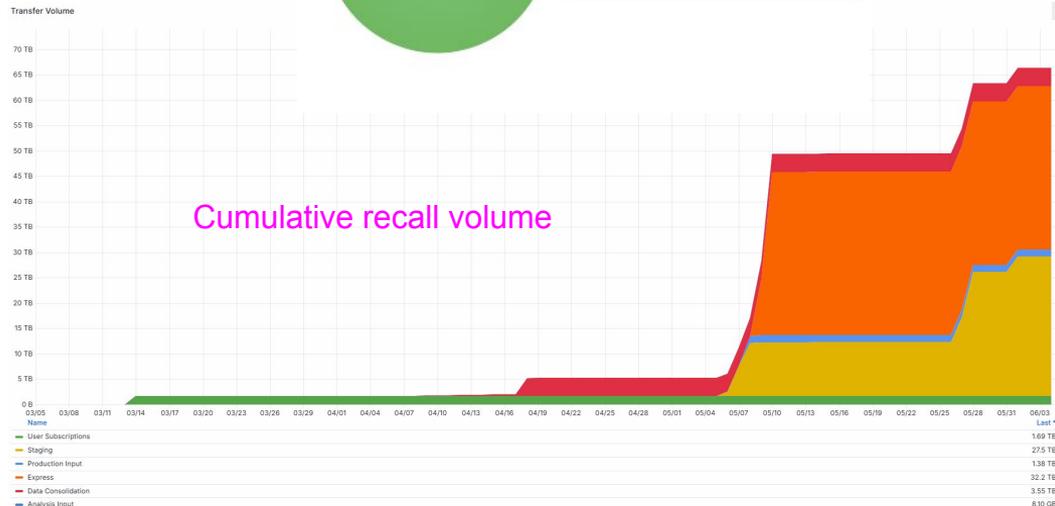
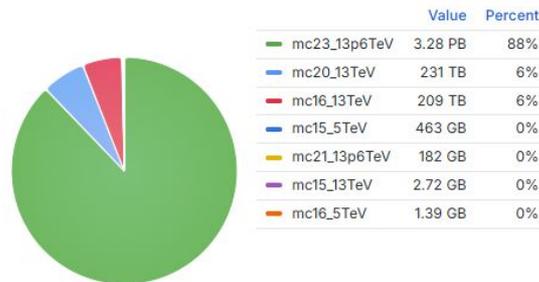
A dedicated campaign to fill the tape with recent AOD (mostly mc23) has started during the DAOD rederivation in April 2025

We were able to participate in the very end of the campaign.

We are now receiving copies of EVNT, HITS, and AODs. The tape system is being used normally like the T1s by Data Carousel to serve as input for production

Tier 0 exports still to pledged Tier 1s only.

Volume by scope [Last 24 hours](#)



Workflows

A dedicated campaign to fill the tape with recent AOD (mostly mc23) has started during the DAOD rederivation in April 2025

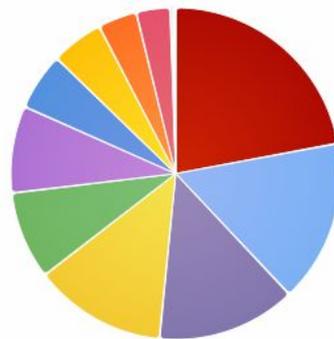
We were able to participate in the very end of the campaign.

We are now receiving copies of EVNT, HITS, and AODs. The tape system is being used normally like the T1s by Data Carousel to serve as input for production

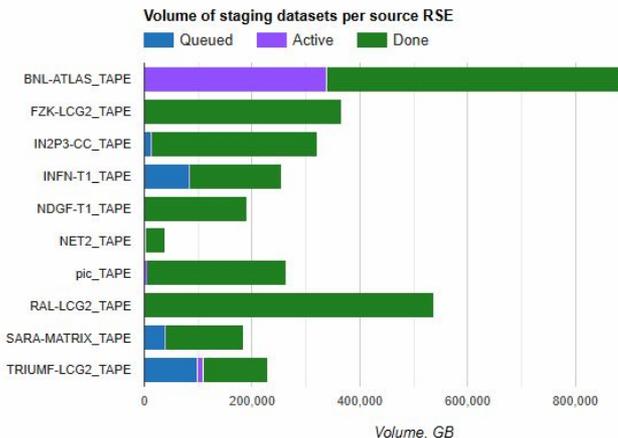
Tier 0 exports still to pledged Tier 1s only.

Analysis and reconstruction workflows have been staging files from NET2_MCTAPE successfully.

Volume by rse [🕒 Last 24 hours](#)



	Value	Percent
BNL-OSG2_MCTAPE	24.3 PB	22%
RAL-LCG2_MCTAPE	17.4 PB	16%
IN2P3-CC_MCTAPE	15.0 PB	14%
FZK-LCG2_MCTAPE	14.2 PB	13%
TRIUMF-LCG2_MCTAPE	9.36 PB	9%
INFN-T1_MCTAPE	9.25 PB	8%
NDGF-T1_MCTAPE	6.08 PB	6%
SARA-MATRIX_MCTAPE	5.57 PB	5%
PIC_MCTAPE	4.13 PB	4%
NET2_MCTAPE	3.58 PB	3%
IN2P3-CC_GROUPTAPE	112 TB	0%



Only two weeks in the Data Carousel already showing on the monitoring!

Looking forward

Operational enhancements

- Improve availability and scaling

Leverage access logs to inform storage strategy

- Analyze file access patterns
- Classify datasets into access tiers
- Correlate performance with size and usage

Optimize tape usage and resource allocation

- Identify candidates for multi-copy tape storage
- Use predictive recall based on historical trends
- Derive storage tiering heuristics

Discussion points

- Benefit/cost for experiments
- Quality of service
- Sustainability

Acknowledgements

We would like to express our thanks for the support provided by Mattias Wadenstein (NeIC) with the ENDIT code, Fabio Luchetti (U. Washington) with the tests and commissioning of the new tape system, and the ATLAS S&C team including Mario Lassnig (CERN) and David South (DESY) for the support.

The support from US ATLAS S&C and ATLAS S&C allowed us not only to offer the tape RSE, but also to have it fully incorporated in the ATLAS Data Carousel workflows.

We are constantly looking for other sites and institutions that would like to collaborate in the development of the tape interface for Tier 2s. Don't hesitate to contact us at net2-technical@groups.umass.edu

Thank you