Evaluating Tape Storage at MIT

06/07/2025

Motivation for study of a tape robot at MIT

- CMS write O(100 PB) of data per year that need to be stored on tape
- HL-LHC (~2030) will require an order of magnitude increase (Exa Bytes ?)
- Limited number of tape storage sites in CMS, only one in U.S.
- Vulnerability to tape site failures is significant: we had natural catastrophes (like fire, typhoon, and massive rain falls) and other circumstances affect various tape sites
- Opportunity at MIT arose to make use of the Harvard managed tape robot
- Unexplored aspect in CMS: use tape robot that is externally managed **without direct access**

University of Amherst (ATLAS) bought into the same Harvard tape robot.

Tomorrow at the joint ATLAS-CMS section we will present our different approaches with more technical details.

Any Tier2 functionality

- cores to run user jobs
- storage to store data
- data transfer mechanism

Tier1 = Tier2 + Tape Robot

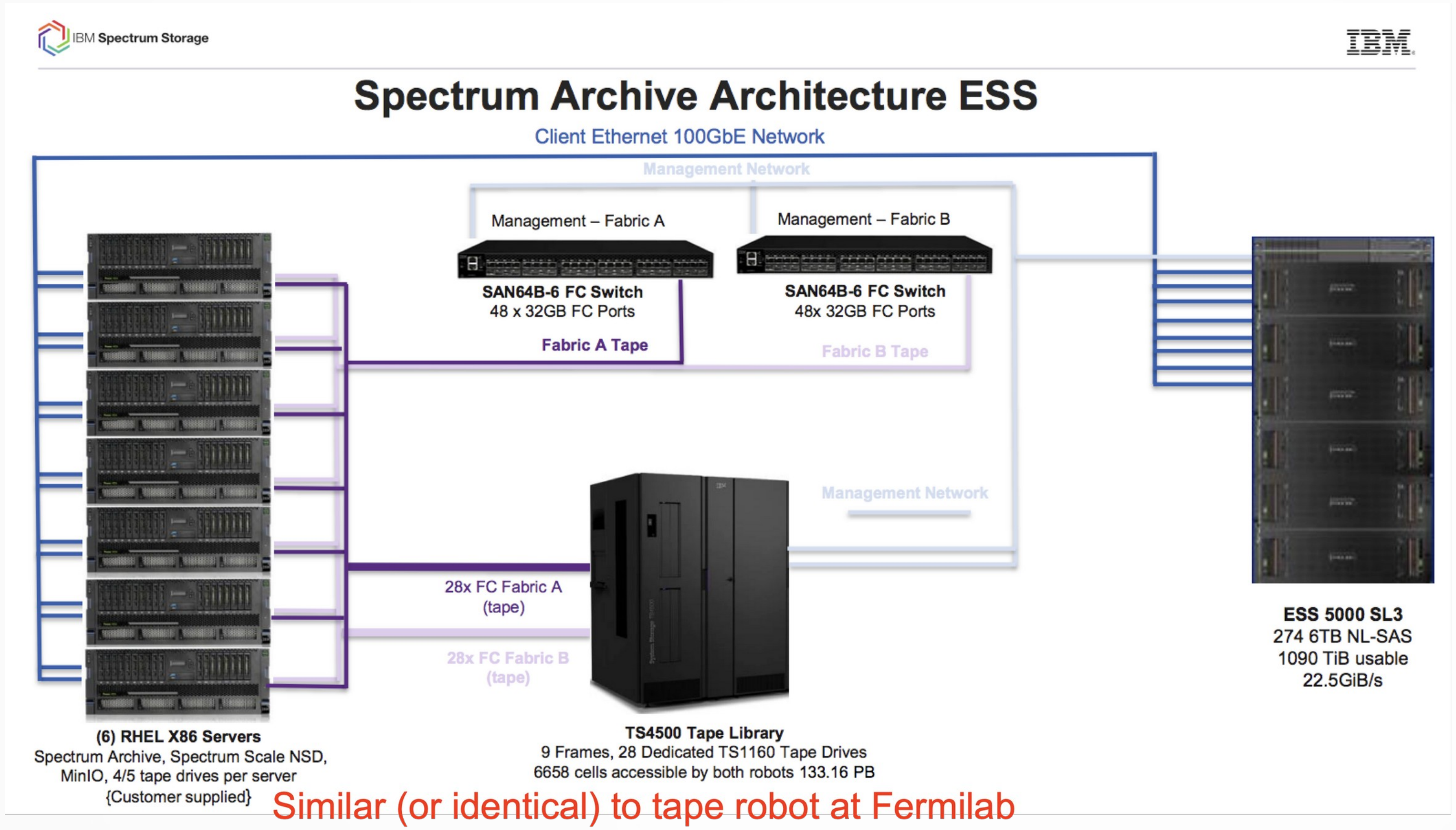
CMS experiment computing structure

- 8 Tier1 sites, one in US at Fermilab
- ~50 Tier2 sites, eight are in US
- our Tier2 site – 25K cores, 23 PB of storage

	Tape capacity
T0_CH_CERN	320 PB
T1_US_FNAL	240 PB
T1_DE_KIT	45 PB
T1_UK_RAL	30 PB
T1_FR_CCIN2P3	47 PB
T1_ES_PIC	17 PB
T1_RU_JINR	25 PB
T1_IT_CNAF	63 PB

Harvard University bought and started to operate a tape robot

Harvard offers other groups to buy into tape (purchase tape cartridges)



IBM Spectrum Storage

IBM

Spectrum Archive Architecture ESS

Client Ethernet 100GbE Network

Management Network

Management – Fabric A

SAN64B-6 FC Switch
48 x 32GB FC Ports

Fabric A Tape

Management – Fabric B

SAN64B-6 FC Switch
48x 32GB FC Ports

Fabric B Tape

Management Network

28x FC Fabric A
(tape)

28x FC Fabric B
(tape)

(6) RHEL X86 Servers
Spectrum Archive, Spectrum Scale NSD,
MinIO, 4/5 tape drives per server
{Customer supplied}

TS4500 Tape Library
9 Frames, 28 Dedicated TS1160 Tape Drives
6658 cells accessible by both robots 133.16 PB

ESS 5000 SL3
274 6TB NL-SAS
1090 TiB usable
22.5GiB/s

Similar (or identical) to tape robot at Fermilab

Maximum Capacity: 157 PB
9 Frames, expandable to 18
34 TS1160 Tape Drives, max 11.2 Gb
ESS-5000: 1.1 PB useable
100 Gb network

IBM GPFS POSIX interface
IBM Spectrum Archive Library Software
Xrootd with staging
Globus 5 with staging
S3 via MinIO

As US CMS Tier2 site we have acquired

- ~16 PB of disk storage; accumulated over decade of buying disks
- bought 15 PB of tape storage for 1/10 of disk cost

  + resilient storage, cheap

  - operational costs, reading is slow

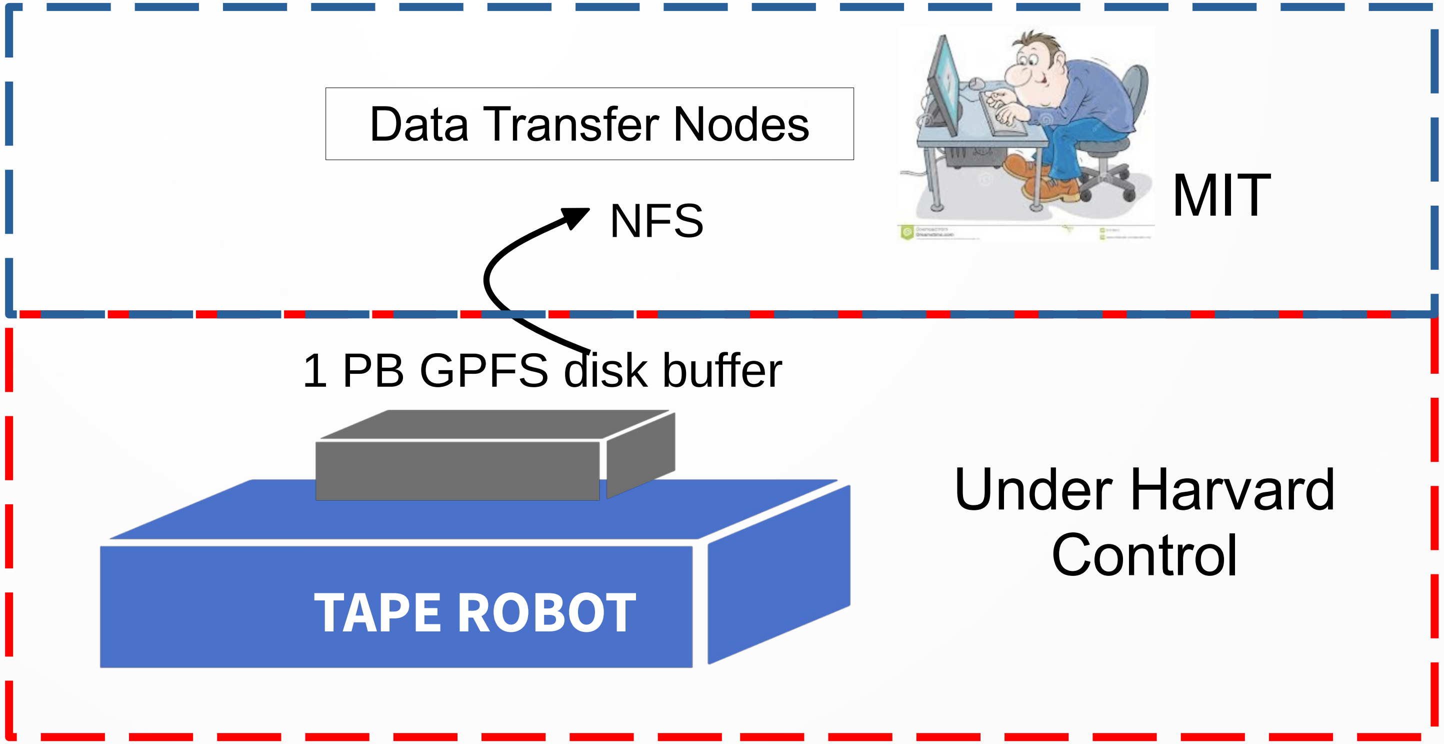
Can we use it for CMS needs?

- as users we do not have access to tape libraries
- restrictive access to file system (security concerns)
- all other CMS tape sites own the robot and do not have any of those restrictions
- Tier1 tape interface with CMS assumes access to tape libraries

GPFS disk buffer is exposed to us though NFS

And that means

- No access to tape libraries
  - ➔ is file available on disk buffer, is it on tape?
  - ➔ how do you stage out from tape onto disk?
- No capability for file extended attributes
  - ➔ was transfer ok (checksum) ?



TAPE REST API Calls: there are two types of them

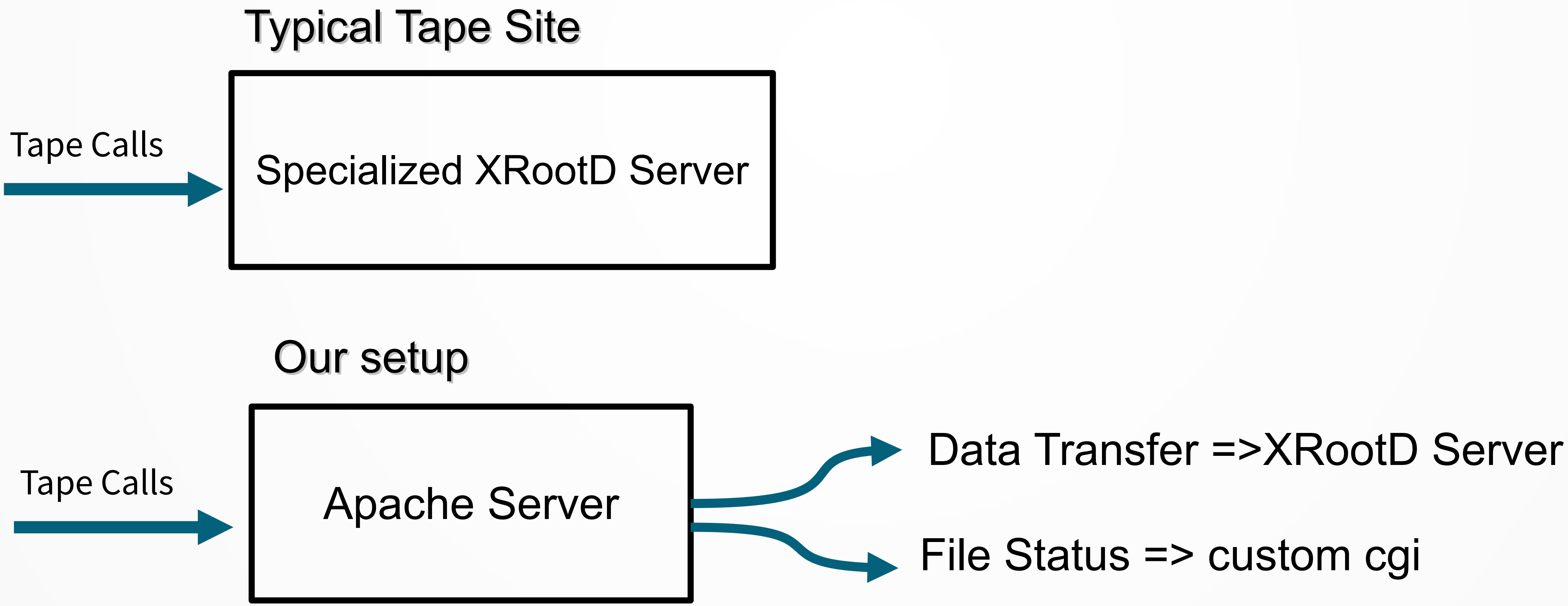
- File status calls
  - ➔ is file available for immediate read (is it on disk)?
  - ➔ stage it from tape onto disk
  - ➔ has this file been written to tape?
- Data transfer calls (XRootD or GridFTP)
  - ➔ do you have this file?
  - ➔ what is checksum ?
  - ➔ read a file
  - ➔ write a file

In CMS all above calls are handled by tape sites utilizing specialized XRootD protocol (dCache). Underneath it assumes full access to the tape robot.

Our Solution: all calls are handled by an apache server

- data transfer calls are forwarded to XRootD servers
- all file status calls are handled by custom cgi (python) scripts

Same setup would apply to Globus GridFTP as a transfer protocol.



Typical Tape Site

Tape Calls

Specialized XRootD Server

Our setup

Tape Calls

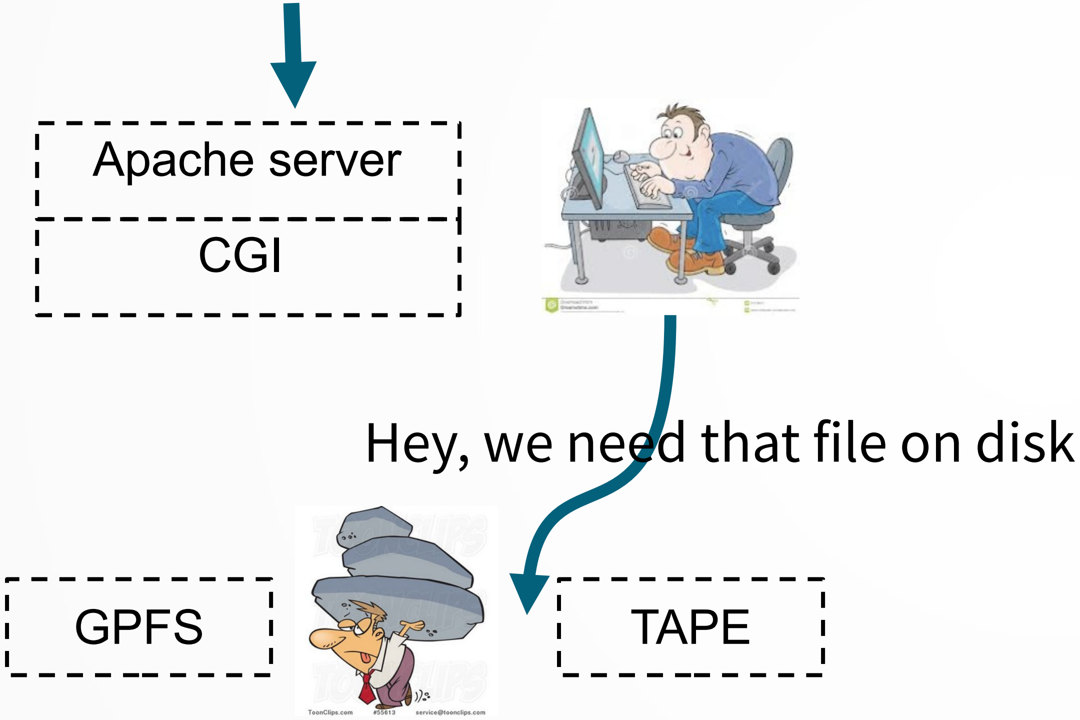
Apache Server

Data Transfer =>XRootD Server

File Status => custom cgi

For example: stage out from tape call

Bring me from tape onto disk that file



Apache server

CGI

Hey, we need that file on disk

GPFS

TAPE

Harvard team does heavy lifting

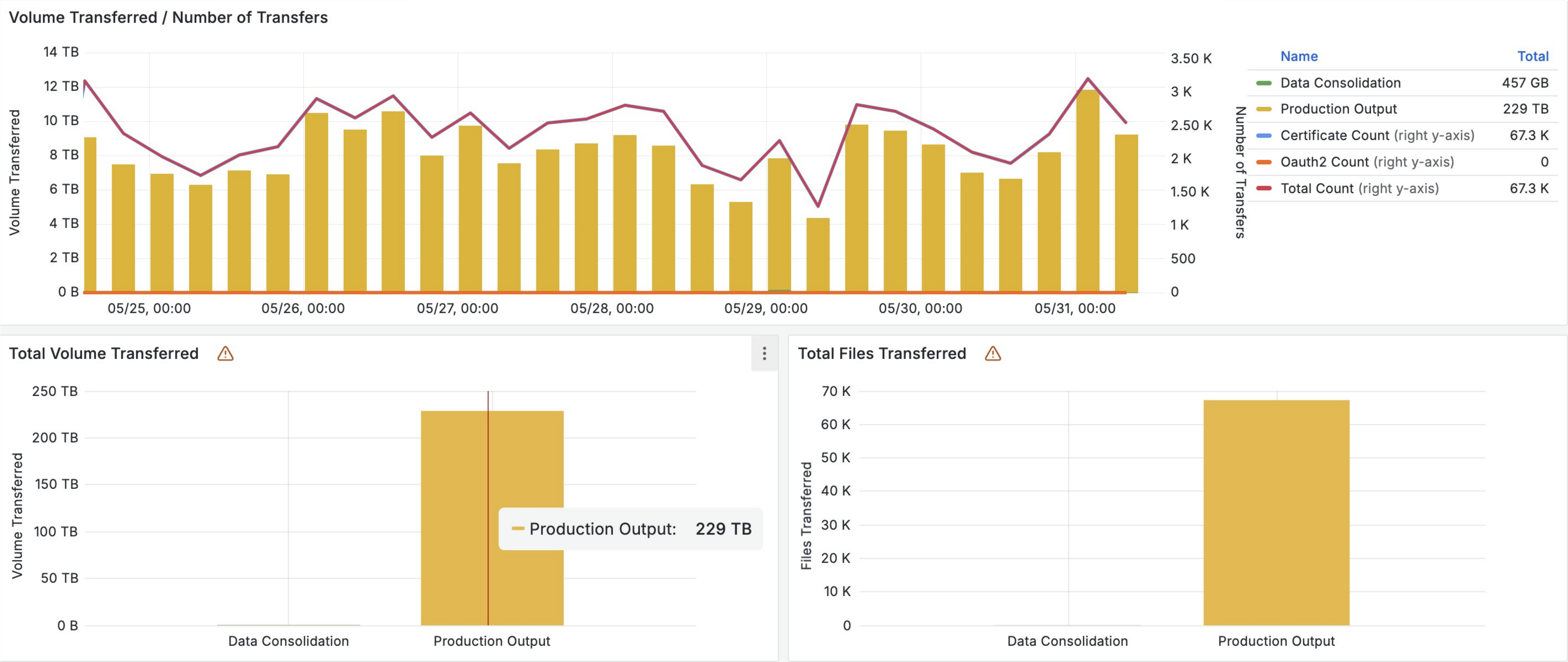
File becomes available for reading

Data flow never stops. This is a typical week.

The amount of data on tape right now – 6.1 PB

Data writing is limited by network available: ~7 Gb

University promised to have 100 Gb available by the end of the summer



Volume Transferred / Number of Transfers

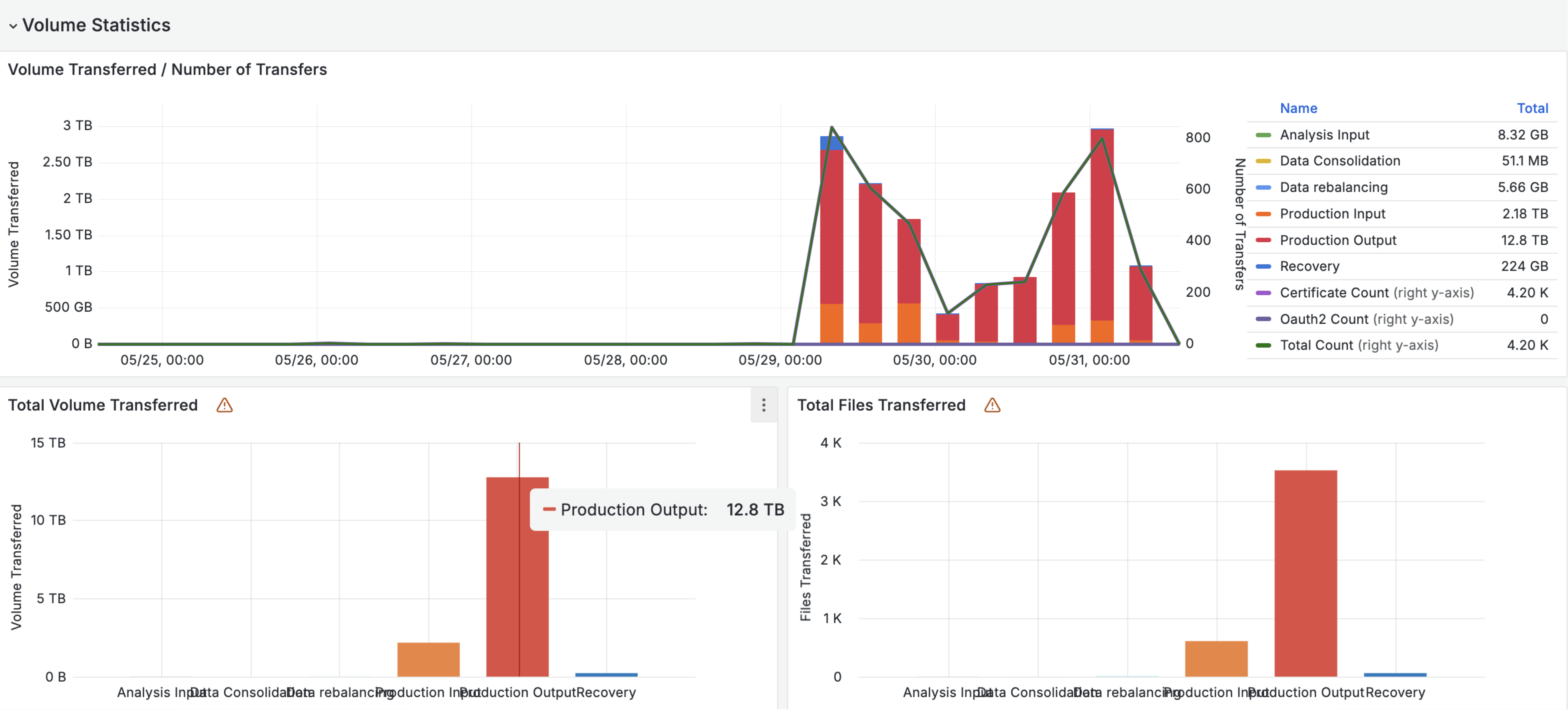
Name	Total
🟩 Data Consolidation	457 GB
🟨 Production Output	229 TB
🟦 Certificate Count (right y-axis)	67.3 K
🟧 Oauth2 Count (right y-axis)	0
🟥 Total Count (right y-axis)	67.3 K

Total Volume Transferred ⚠

Total Files Transferred ⚠

Tape reading is not as common at the moment.

Will scale tape staging out in the future.



⌄ Volume Statistics

Volume Transferred / Number of Transfers

Name	Total
Analysis Input	8.32 GB
Data Consolidation	51.1 MB
Data rebalancing	5.66 GB
Production Input	2.18 TB
Production Output	12.8 TB
Recovery	224 GB
Certificate Count (right y-axis)	4.20 K
Oauth2 Count (right y-axis)	0
Total Count (right y-axis)	4.20 K

Total Volume Transferred

Production Output: 12.8 TB

Total Files Transferred