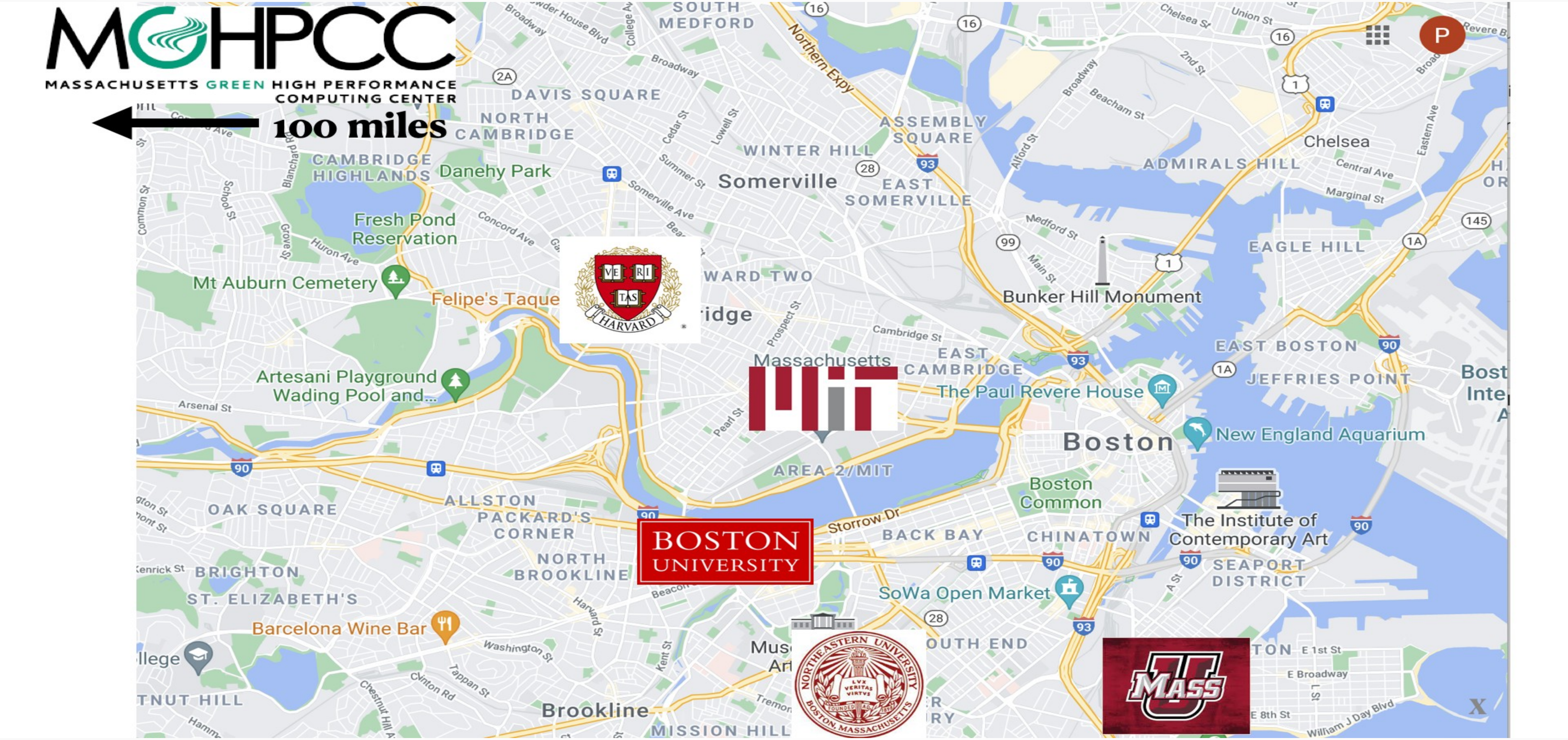
Evaluating Tape Storage at MIT

06/08/2025

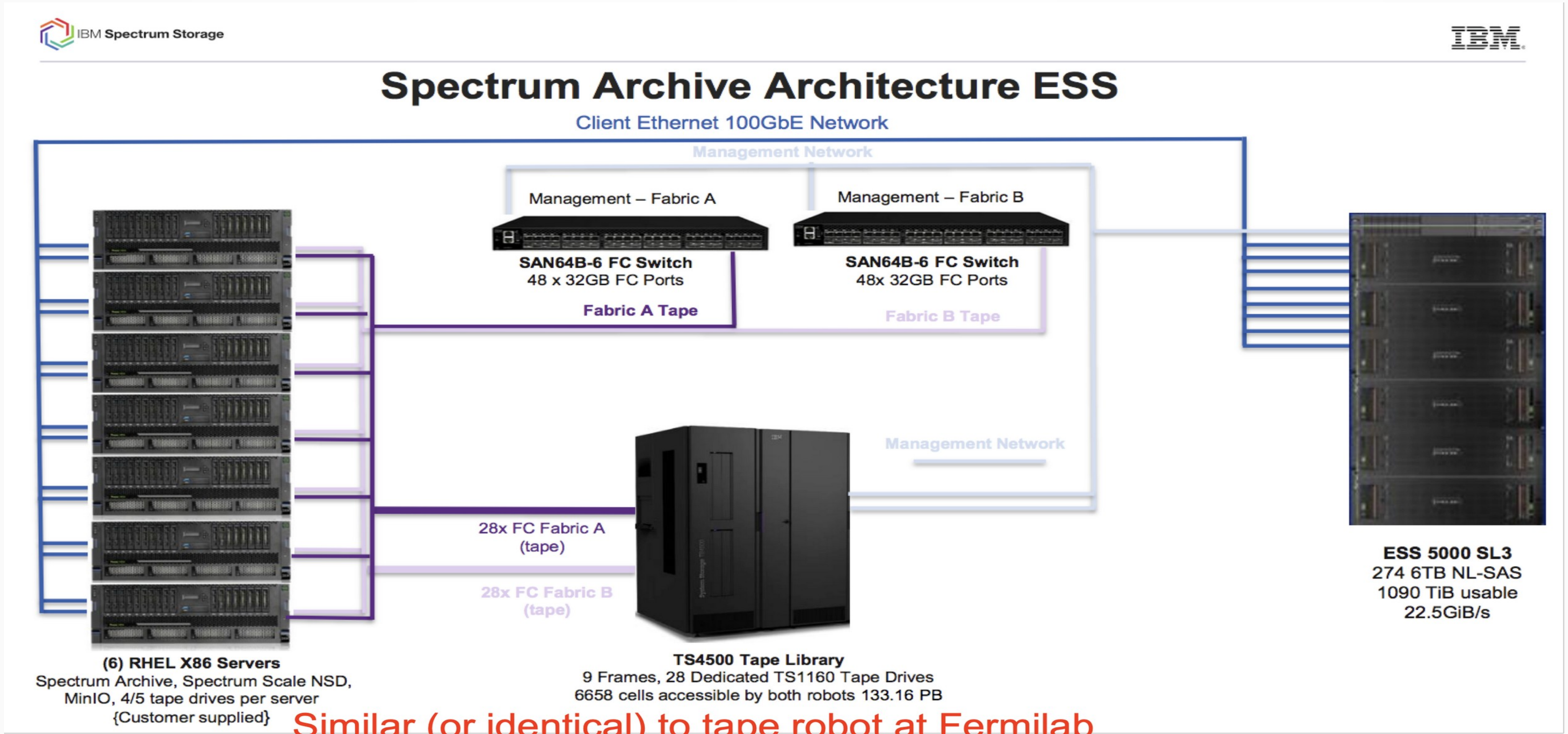
Motivation for study of a tape robot at MIT

- CMS writes 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**


MGHPCC
MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER
100 miles
HARVARD
VE RI TAS
Massachusetts
BOSTON UNIVERSITY
NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS
LUX VERITAS VIRTUS
UMASS
SOUTH MEDFORD
DAVIS SQUARE
NORTH CAMBRIDGE
CAMBRIDGE HIGHLANDS
Danehy Park
Somerville
WINTER HILL
EAST SOMERVILLE
ASSEMBLY SQUARE
Chelsea
ADMIRALS HILL
EAGLE HILL
Fresh Pond Reservation
Mt Auburn Cemetery
Felipe's Taque
WARD TWO
Bunker Hill Monument
EAST CAMBRIDGE
EAST BOSTON
JEFFRIES POINT
The Paul Revere House
Artesani Playground Wading Pool and...
Boston
New England Aquarium
AREA 2/MIT
Boston Common
The Institute of Contemporary Art
ALLSTON
PACKARD'S CORNER
OAK SQUARE
NORTH BROOKLINE
BRIGHTON
ST. ELIZABETH'S
BACK BAY
CHINATOWN
SEAPORT DISTRICT
SoWa Open Market
Barcelona Wine Bar
Brookline
MISSION HILL
Storrow Dr
Northern Expy
Arsenal St

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

Management – Fabric B

SAN64B-6 FC Switch
48 x 32GB FC Ports

SAN64B-6 FC Switch
48x 32GB FC Ports

Fabric A Tape

Fabric B Tape

Management Network

28x FC Fabric A
(tape)

28x FC Fabric B
(tape)

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

(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

Similar (or identical) to tape robot at Fermilab

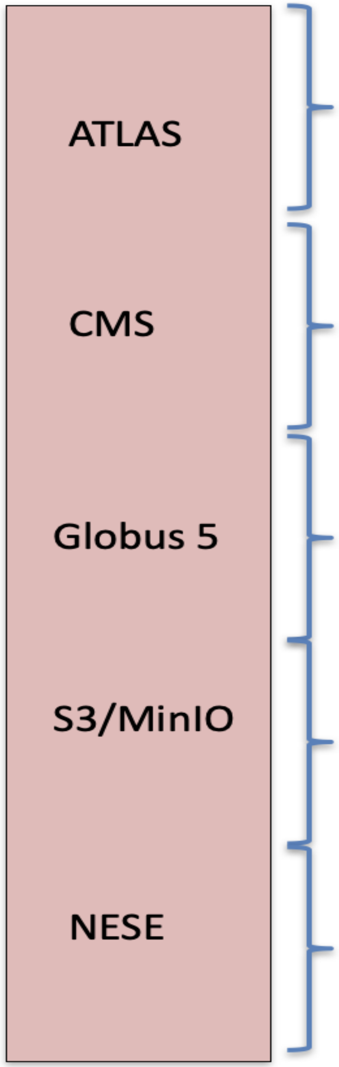
Maximum Capacity: 157 PB
9 Frames, expandable to 18
34 TS1160 Tape Drives, max 11.2 Gb
ESS-5000: 1.1 PB useable
100 Gb network

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

Northeast Storage Echange

Each node has GPFS client and mounts both NESE Tape and NESE CephFS storage

/nese
/nese-ceph



ATLAS — Four node cluster with staging

CMS — Four node cluster with staging

Globus 5 — Four node Globus 5 cluster for general use, with custom staging

S3/MinIO — Four node MinIO cluster for S3 general use, e.g. Harvard librarians CS projects

NESE — Four nodes for testing and spares for the NESE Operations team.

- 157 PB capacity IBM TS4500 tape library, 1 PB GPFS cache
- 110 PB buy-in already
- 50 PB for ATLAS (delivered and installed)
- 15 PB for CMS (delivered and installed)
- LHCb might join with 10 PB

20 endpoints, each with 2 x 25 Gb NICs

4 are in use, remaining 16 waiting for OS

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 data storage; cheap

  - operational costs; data reading is slow

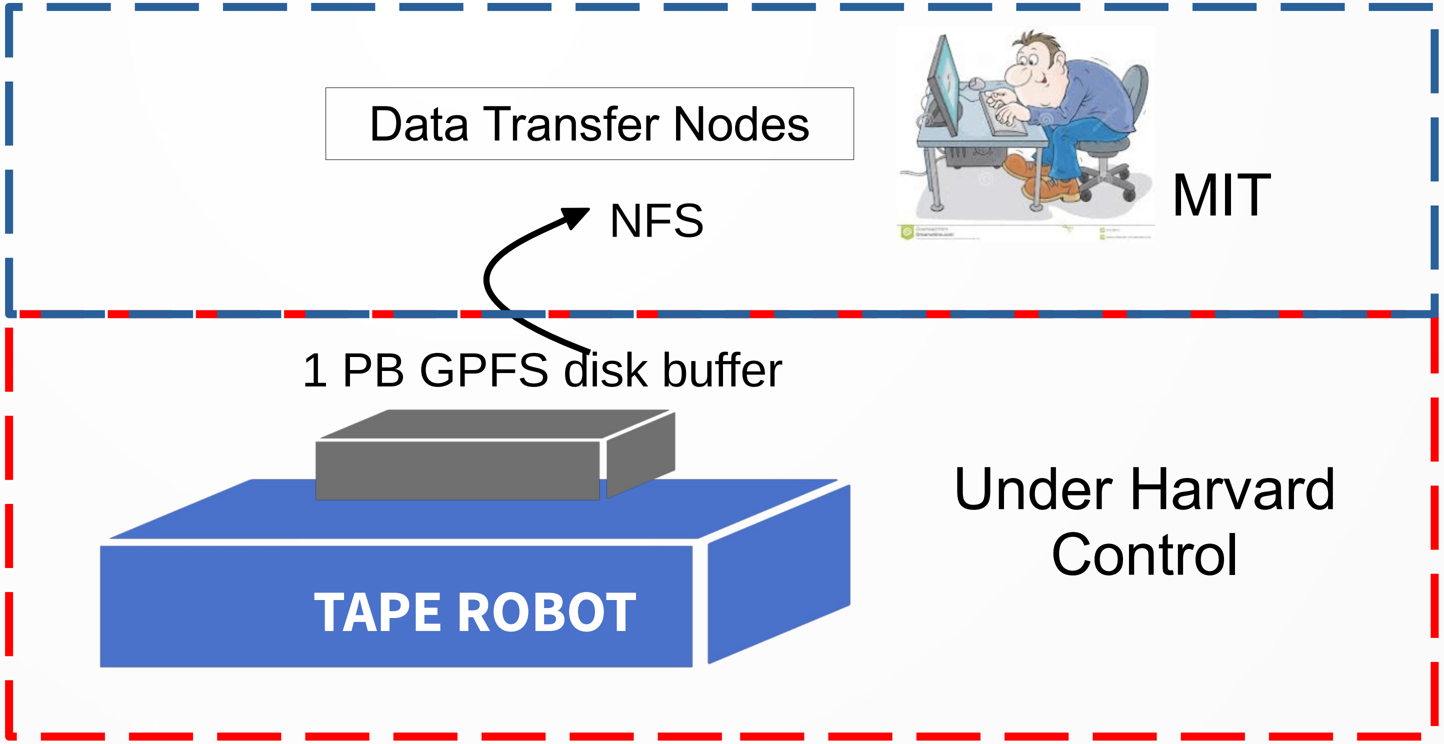
Can we use it for CMS needs?

- as users we do not have access to tape libraries
- restrictive access to file system (security concerns)
- CMS Tier1 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

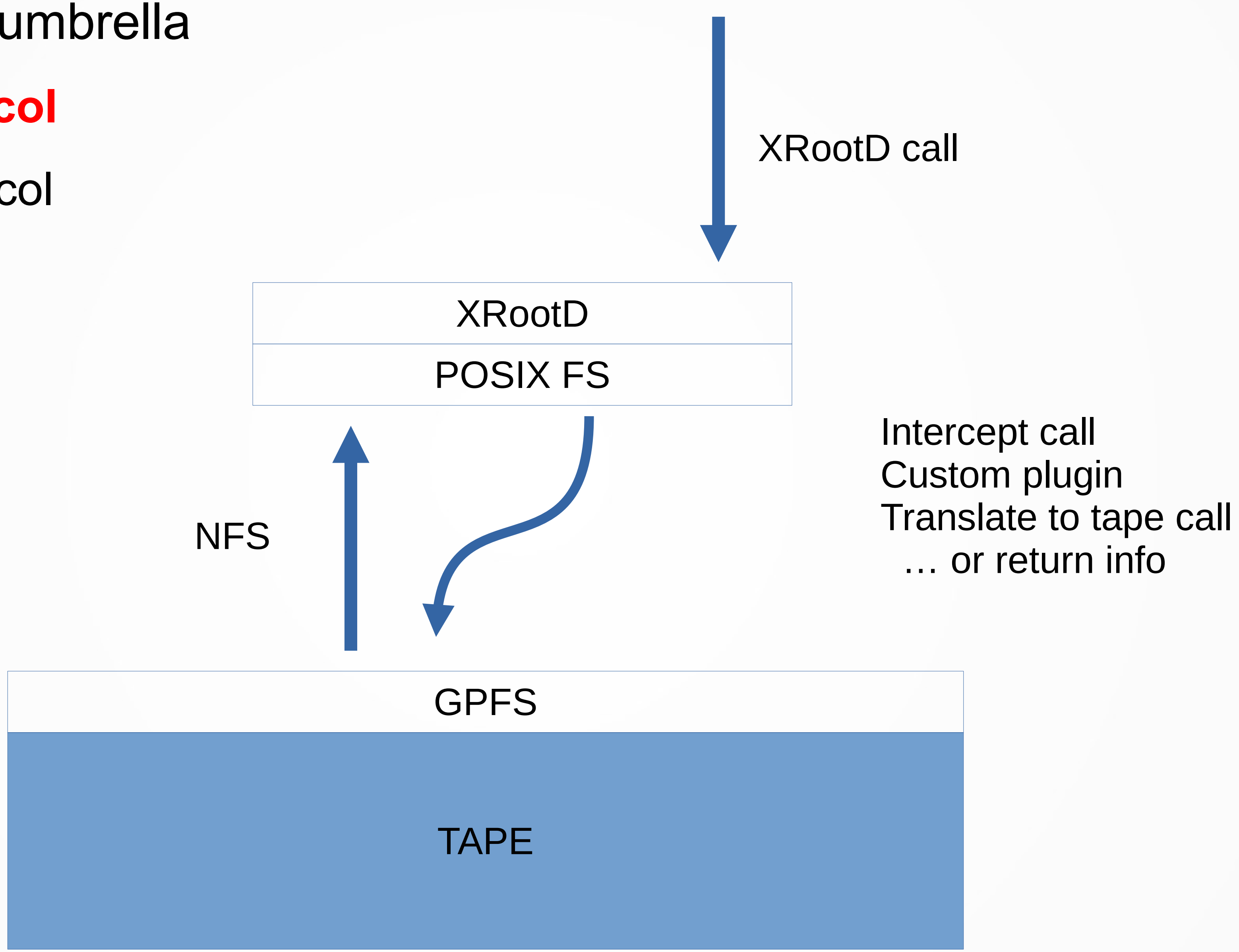
And that means

- 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) ?



Under XRootD umbrella

- **XRootD protocol**
- WebDAV protocol



XRootD call

XRootD

POSIX FS

NFS

Intercept call
Custom plugin
Translate to tape call
… or return info

GPFS

TAPE

XRootD protocol

1. If a file /a/b/c on tape?

```
xrdfs root:/xxx.yyy.edu query prepare /a/b/c
```

=>return Json with “online”=true if file is on disk

2. Request that the file /a/b/c be staged from tape to disk

```
xrdfs root://xxx.yyy.edu prepare -a /a/b/c
```

=> pull file from tape, if required

```
ofs.preplib +noauth libXrdOfsPrepGPI.so -maxresp 14m -maxfiles 256 -maxquery 16 -maxreq 8 -
debug -admit all -run /cms/ops/prepare/prep
```

```
xrootd.chksum max 21 adler32 /mnt/ramdisk/adler/adler.py
```

A callout from XRootD is used to perform the tape staging operations

This is NOT used by FTS any longer

FTS started to use WebDAV protocol (TAPE REST API)

TAPE REST API Calls: there are two types of them

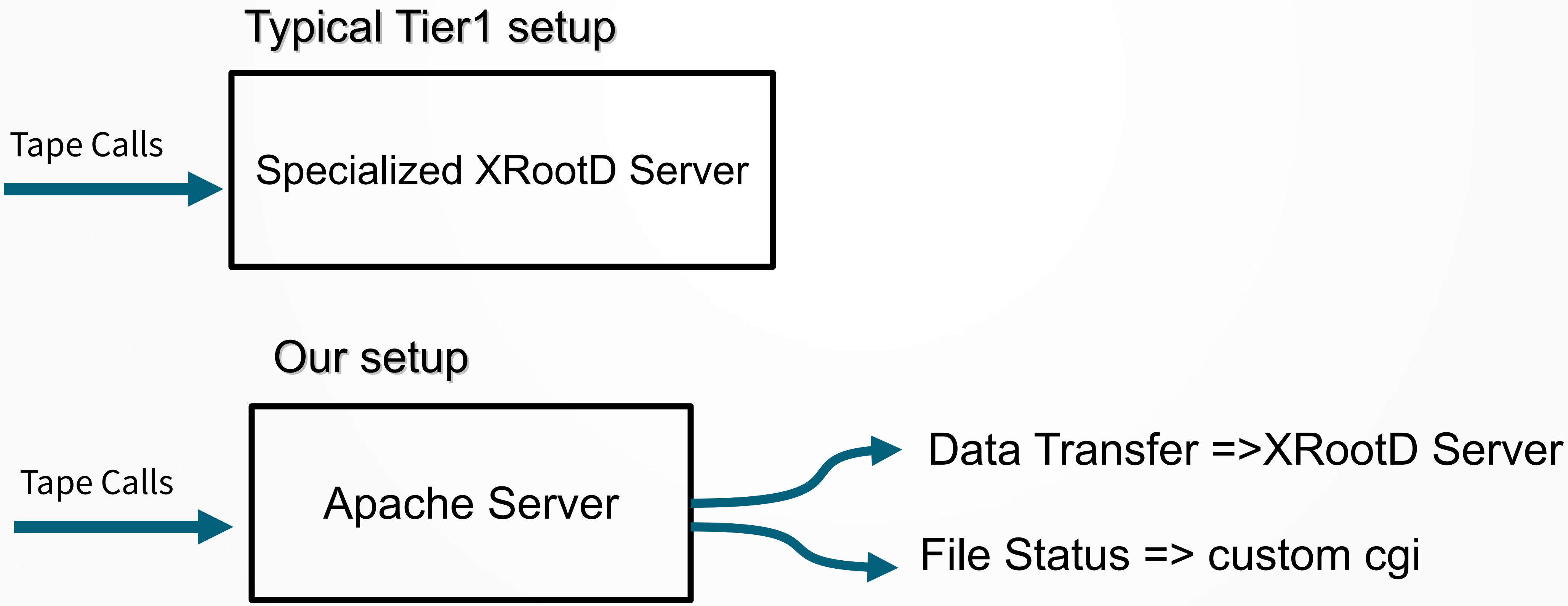
- File status calls
  - is file available for immediate read (is it on disk)?
  - stage it from tape onto disk
  - has this file been written to tape?
- Data transfer calls (XRootD or GridFTP)
  - do you have this file?
  - what is checksum ?
  - read a file
  - write a file

In CMS all above calls will be handled by Tier1 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

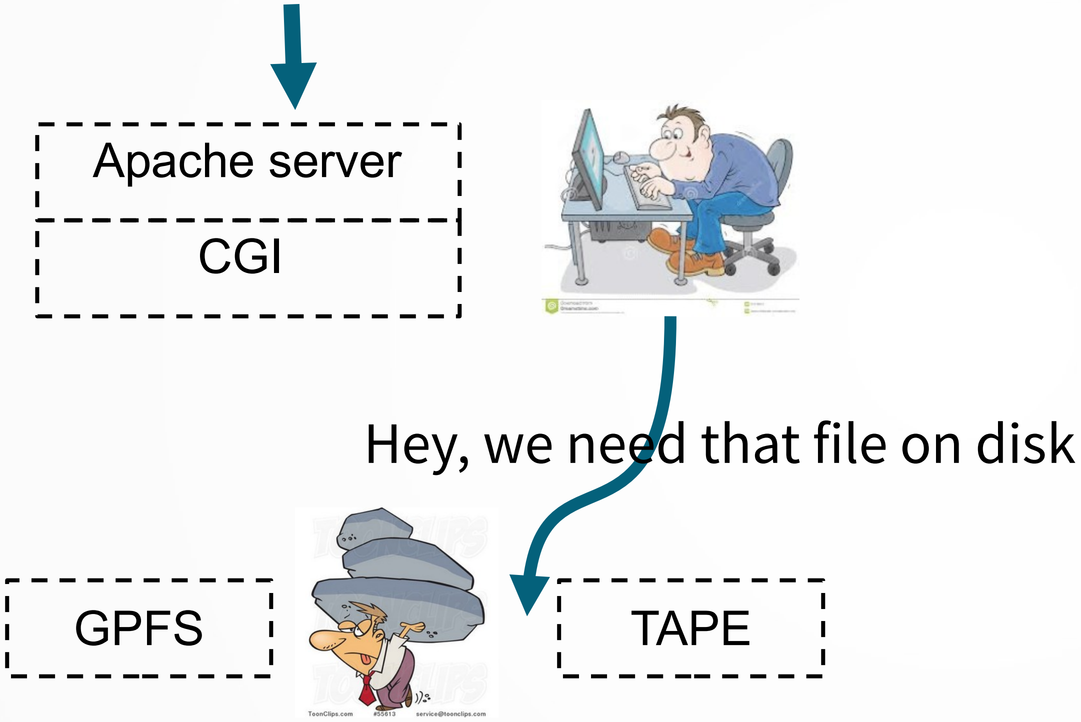
- data transfer calls are forwarded to XRootD servers
- file status calls are handled by custom cgi (python) scripts

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


Typical Tier1 setup
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

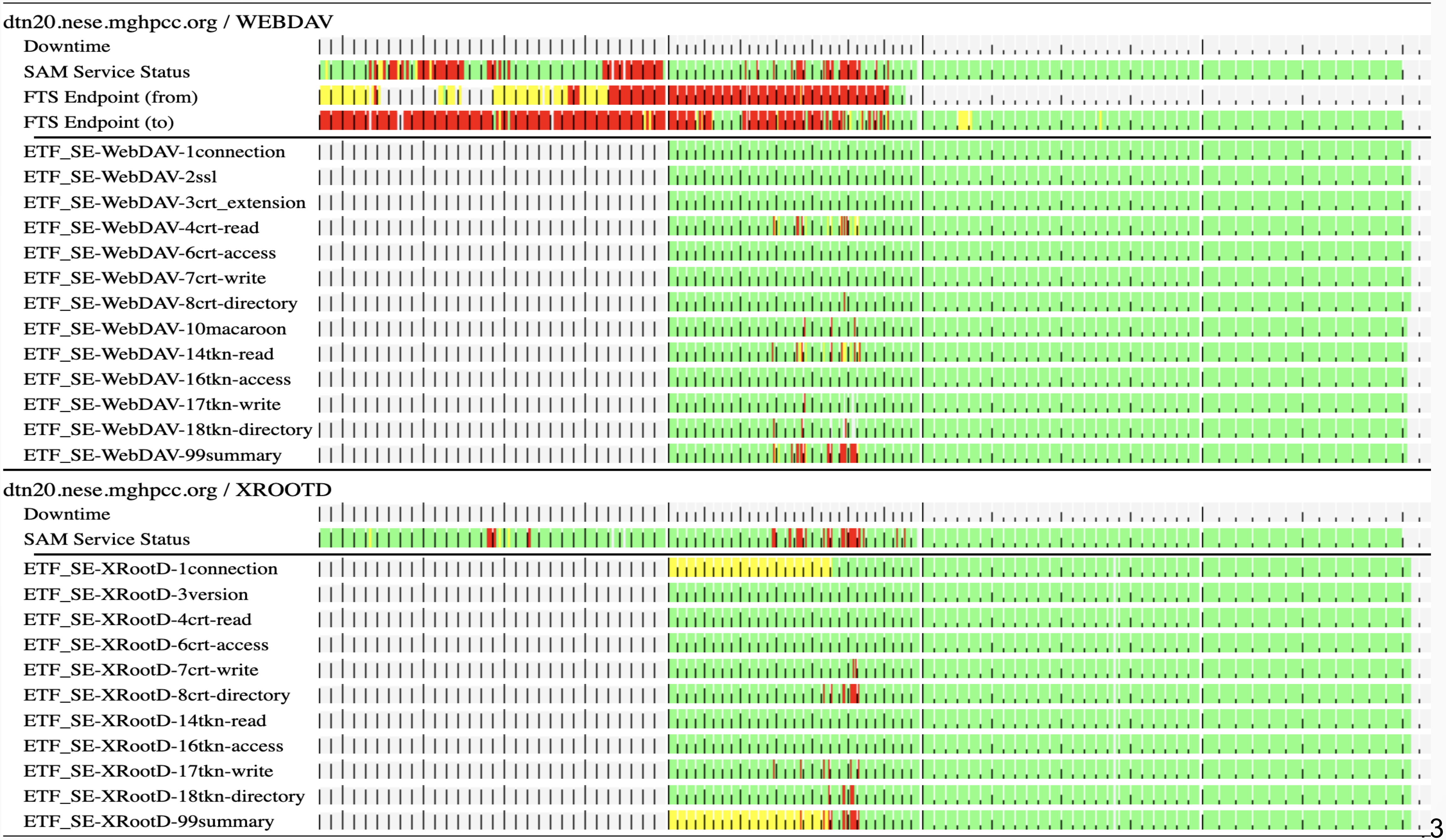
Hey, we need that file on disk

GPFS

TAPE

Harvard team does heavy lifting

The file becomes available for reading

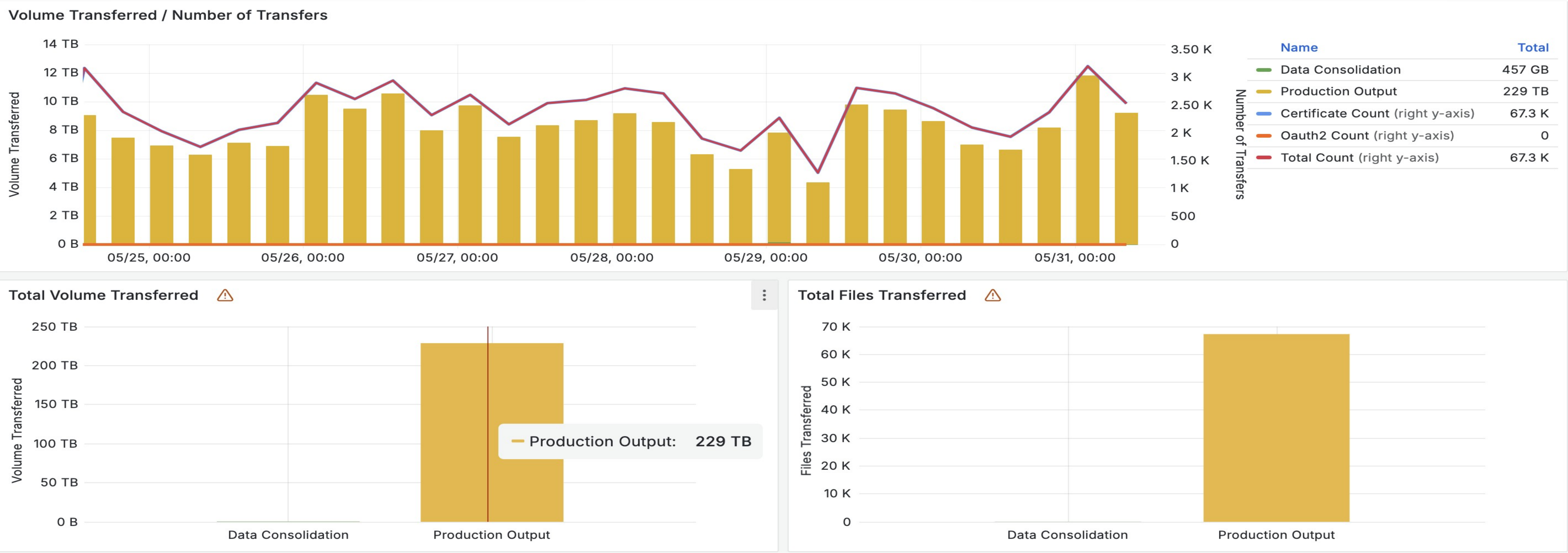

dtn20.nese.mghpcc.org / WEBDAV
Downtime
SAM Service Status
FTS Endpoint (from)
FTS Endpoint (to)
ETF_SE-WebDAV-1connection
ETF_SE-WebDAV-2ssl
ETF_SE-WebDAV-3crt_extension
ETF_SE-WebDAV-4crt-read
ETF_SE-WebDAV-6crt-access
ETF_SE-WebDAV-7crt-write
ETF_SE-WebDAV-8crt-directory
ETF_SE-WebDAV-10macaroon
ETF_SE-WebDAV-14tkn-read
ETF_SE-WebDAV-16tkn-access
ETF_SE-WebDAV-17tkn-write
ETF_SE-WebDAV-18tkn-directory
ETF_SE-WebDAV-99summary
dtn20.nese.mghpcc.org / XROOTD
Downtime
SAM Service Status
ETF_SE-XRootD-1connection
ETF_SE-XRootD-3version
ETF_SE-XRootD-4crt-read
ETF_SE-XRootD-6crt-access
ETF_SE-XRootD-7crt-write
ETF_SE-XRootD-8crt-directory
ETF_SE-XRootD-14tkn-read
ETF_SE-XRootD-16tkn-access
ETF_SE-XRootD-17tkn-write
ETF_SE-XRootD-18tkn-directory
ETF_SE-XRootD-99summary

Data flow never stops. This is a typical week.

The amount of data on tape right now – 6.1 PB

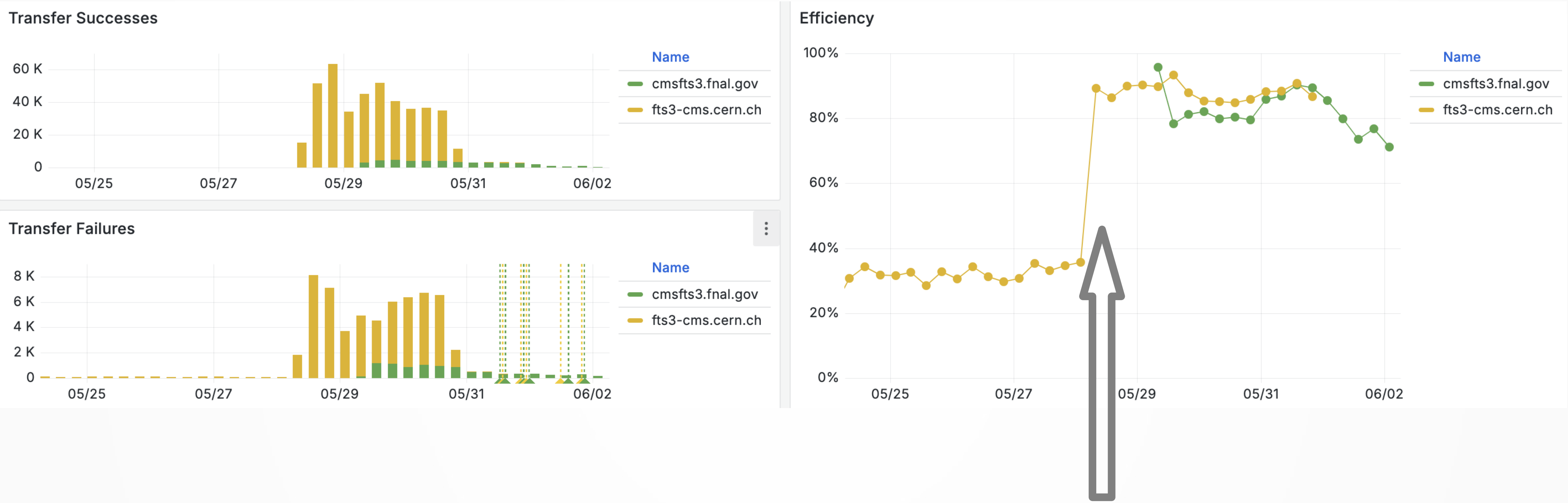
Data writing is limited by network available: ~7 Gb

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



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

Incoming transfers

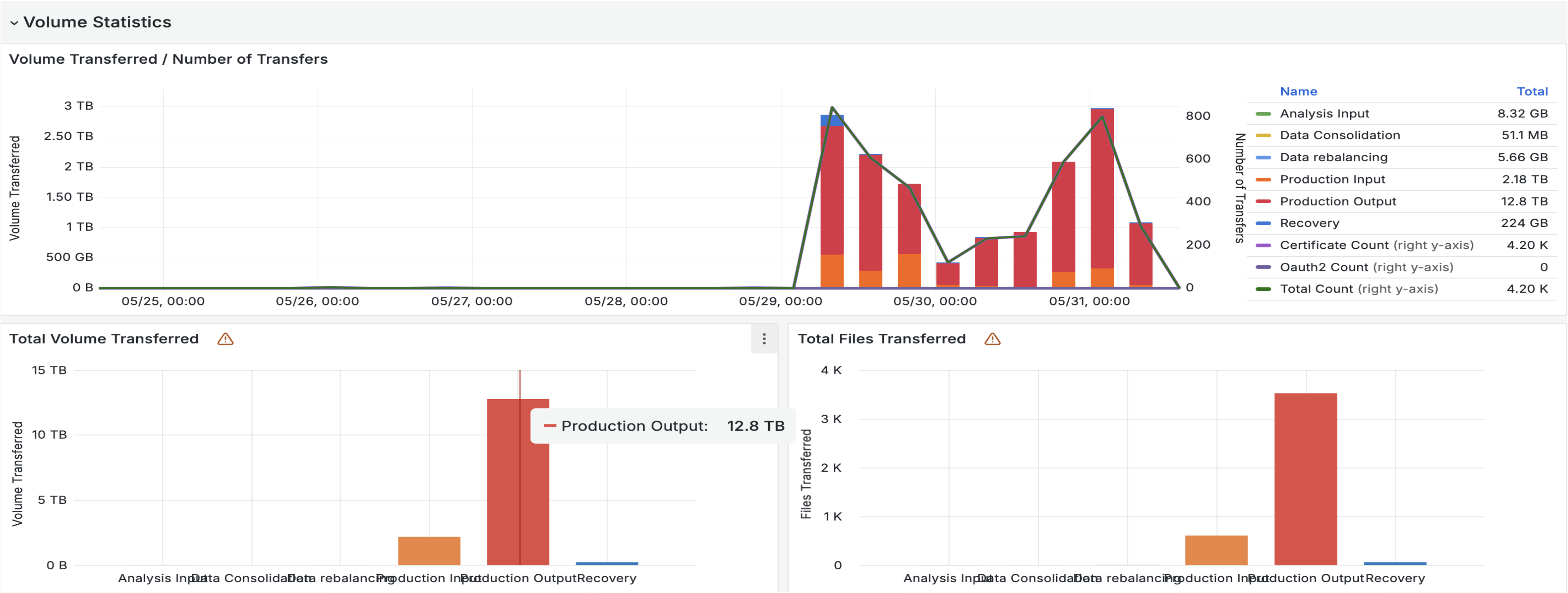


Setting network routing through ESNet was challenging (Harvard is not a member)

Mistake in network routing was blocking incoming ipv6 traffic from Europe

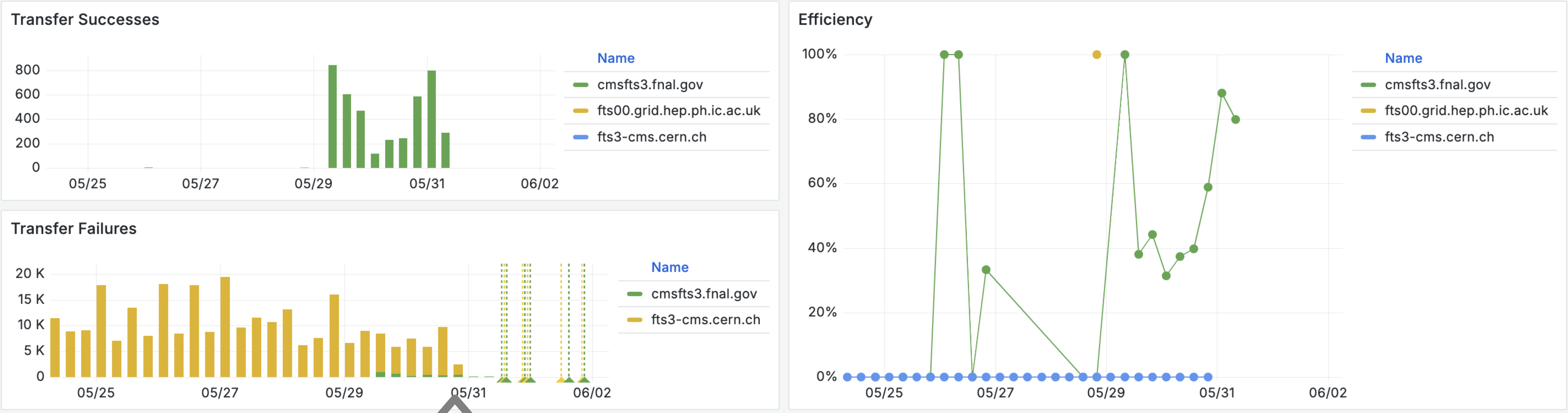
Tape reading is not as common at the moment.

Will scale tape staging out in the future.



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

Outgoing transfers



⇧

Mistake in network routing was blocking incoming ipv6 traffic from Europe