Fermilab CTA efforts and migration plan

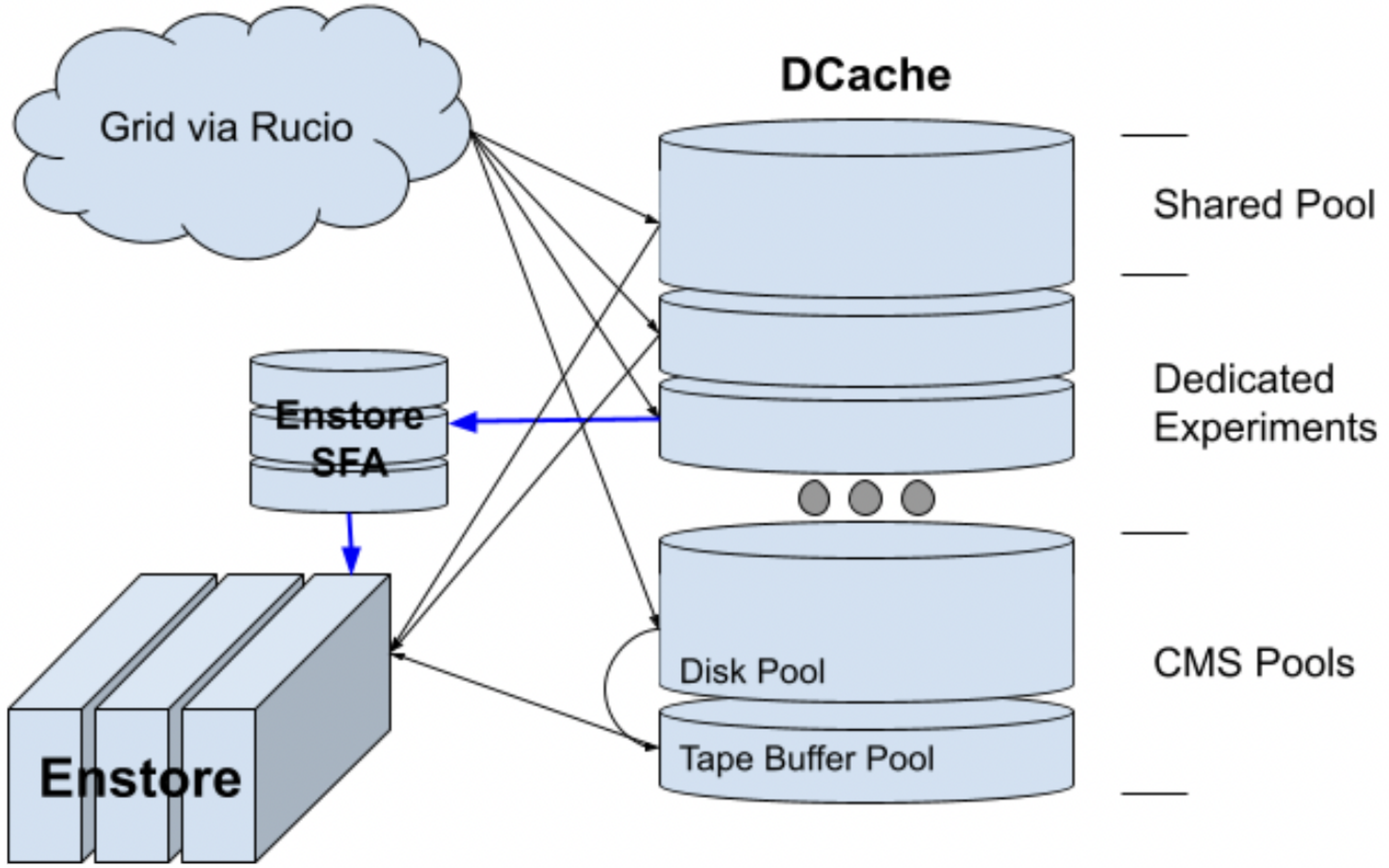
Contributions from many at Fermilab
CTA Workshop
19 March 2024

Outline

- 10% scale testing and results
- Metadata migration from Enstore to CTA
- dCache integration with CTA
- Development needed to read Enstore tapes
- Small File Aggregation options
- Timeframe for migrations

State of Fermilab Tape Storage

- Tape Storage on Enstore
- Two dCache installations
  - CMS: separate disk storage and tape buffer pools, similar to the two disk and buffer EOS instances in CERN's CTA deployment
  - Public: one disk pool backed by tape, auto evicted by LRU
- dCache used for both disk storage and buffer space
- Enstore's Small File Aggregation (SFA) provides capability to stage small files on disk until they can be packaged into a file large enough for tape storage
- All services run on bare metal hardware, no virtualization or release automation

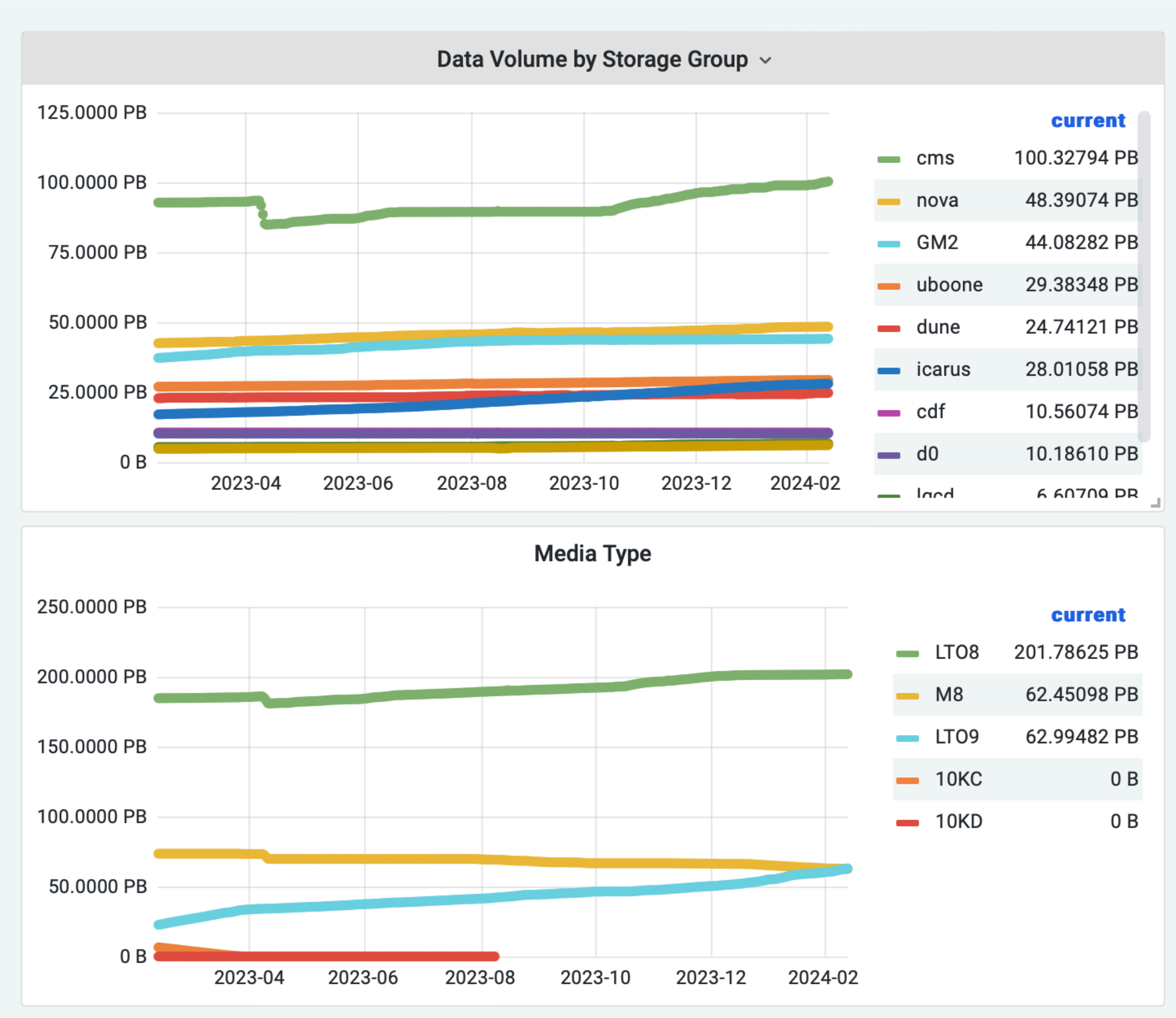


Data ingresses to Fermi from the grid via Rucio, and goes to DCache where, depending on pool, it it can take a variety of paths to Enstore

Fermilab Tape Storage Statistics

Enstore

- Three IBM TS4500 libraries, two Spectra TFinitys. Two libraries dedicated to CMS
- About 350 Petabytes stored data
- 108 LTO8 drives (IBM)
- 80 LTO9 (Spectra Logic)
- 2 drives per server, 32 GB RAM



		current
🟩	cms	100.32794 PB
🟨	nova	48.39074 PB
🟦	GM2	44.08282 PB
🟧	uboone	29.38348 PB
🟥	dune	24.74121 PB
🟦	icarus	28.01058 PB
🟪	cdf	10.56074 PB
🟪	d0	10.18610 PB
🟩	lqcd	6.60709 PB

		current
🟩	LTO8	201.78625 PB
🟨	M8	62.45098 PB
🟦	LTO9	62.99482 PB
🟧	10KC	0 B
🟥	10KD	0 B

Data by experiment (top) and Data by tape media (bottom)

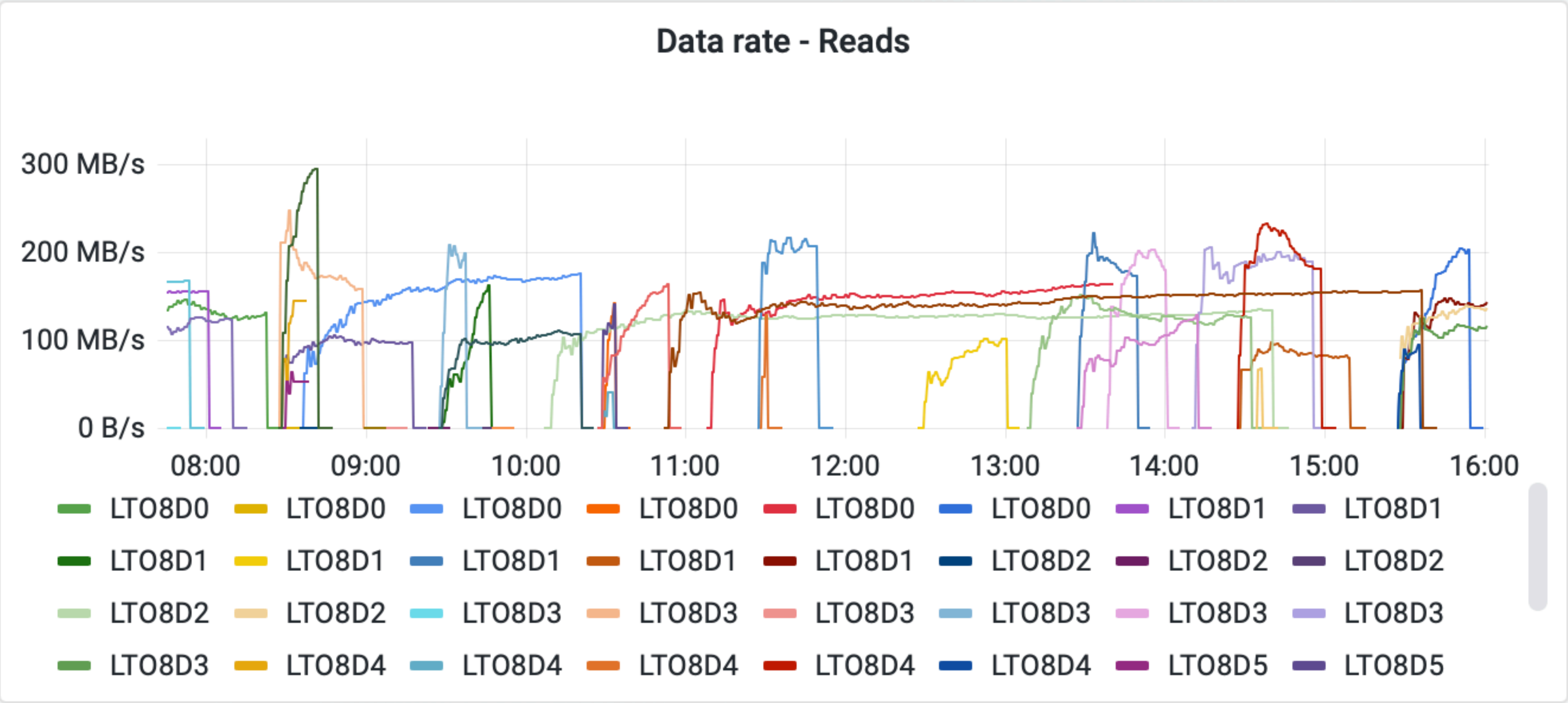
Current status

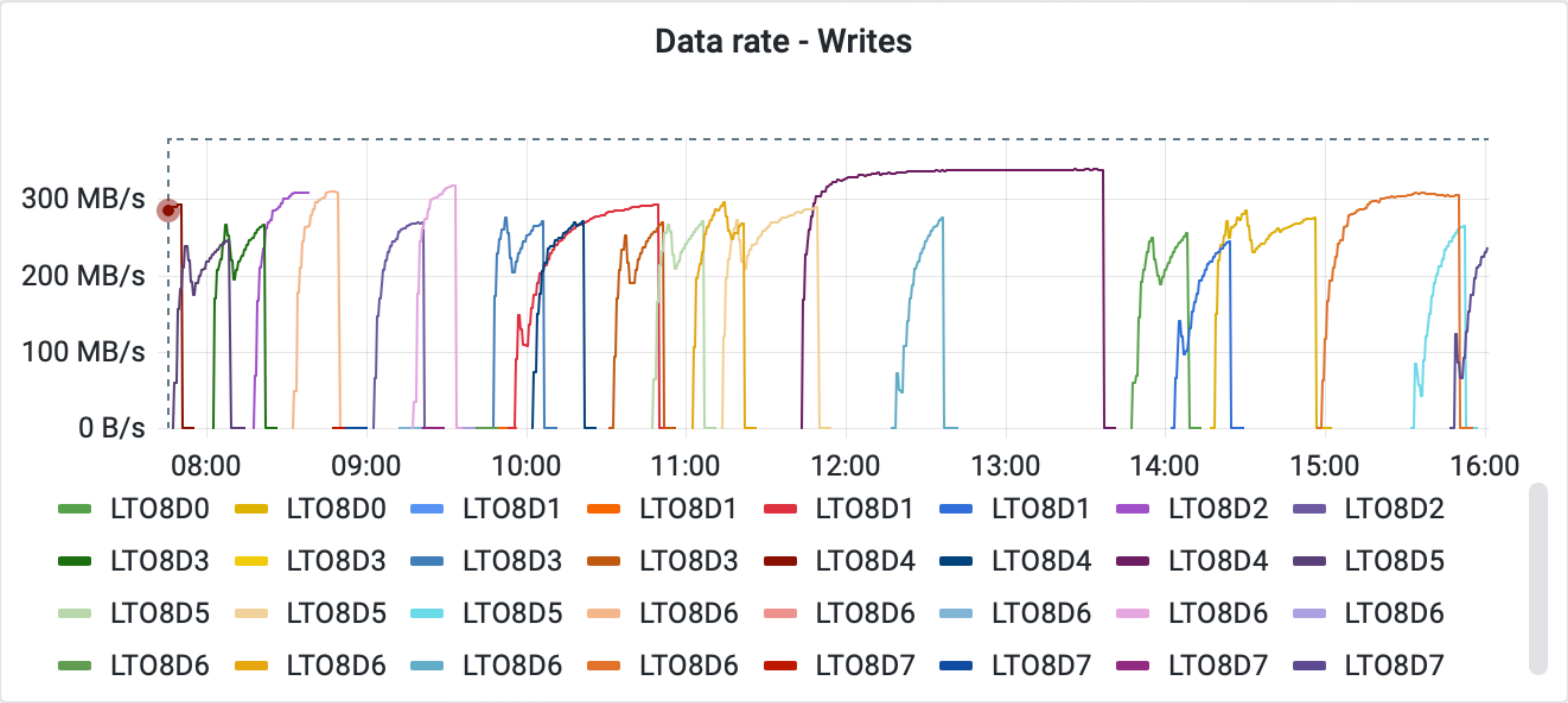
- Fermi decided some time ago to adopt CTA
- Currently in the middle of a 10% scale test: operating 8 drives at production rates
  - CTA 5.10.0
  - First with EOS as a tape buffer, now testing with dCache
  - Will use this experience as a basis to decide EOS or dCache for CMS installation
- Then a period of “operations” on the test system with non-critical data
- Migration of CMS to CTA will happen this year
- Migration of other experiments will happen next year

10% Test methodology

- Define about 20 tape pools with a few tapes per pool
- Try to roughly distribute data over these pools
- Generate incoming traffic with Rucio + FTS. Target is 25 TB/day
  - CMS datasets already on FNAL disk
- Generate recalls of sequential (on disk) files, also 25 TB/day. (About 100 chunks of files during the day)
  - About 10% of a very busy day on the current CMS system (which has 10x the drives)
- Test finished with EOS this weekend.
  - Quite smooth: 50 TB/day bandwidth easily reached. Did 100 TB/day the last day
  - One issue with a tape not being unloaded. cta-smc unload worked.
- dCache testing has started

Data rates from 10% test


Data rate - Reads

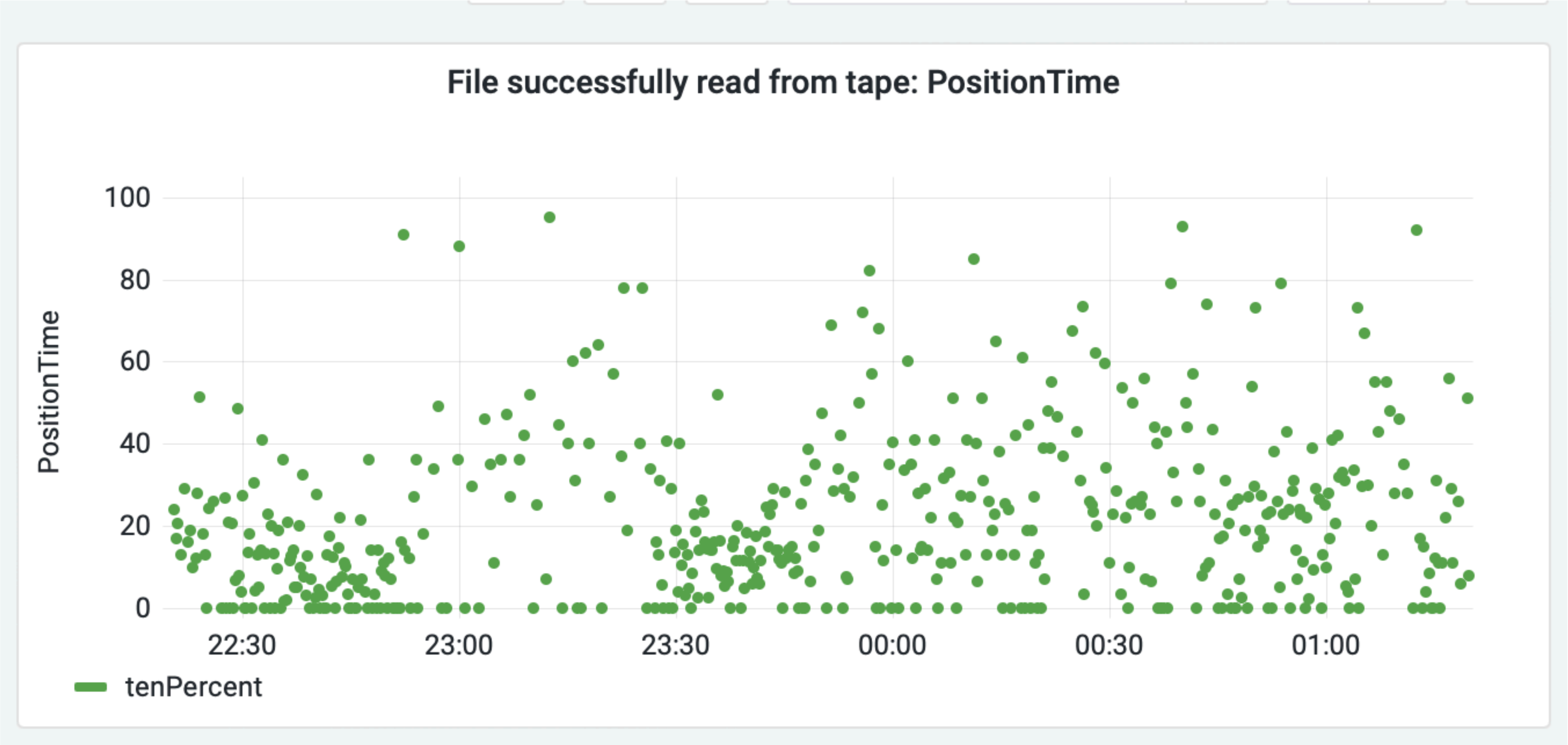

Data rate - Writes

- Long read sessions hitting 150–200 MB/s
  - We have not set the tape values for RAO on LTO-8 tapes
- Writes reach >320 MB/s with sufficiently long write sessions

File positioning times

File positioning times on reads are higher than we would like.

Will retest with magic numbers for LTO-8 to enable RAO



dCache Integration with CTA

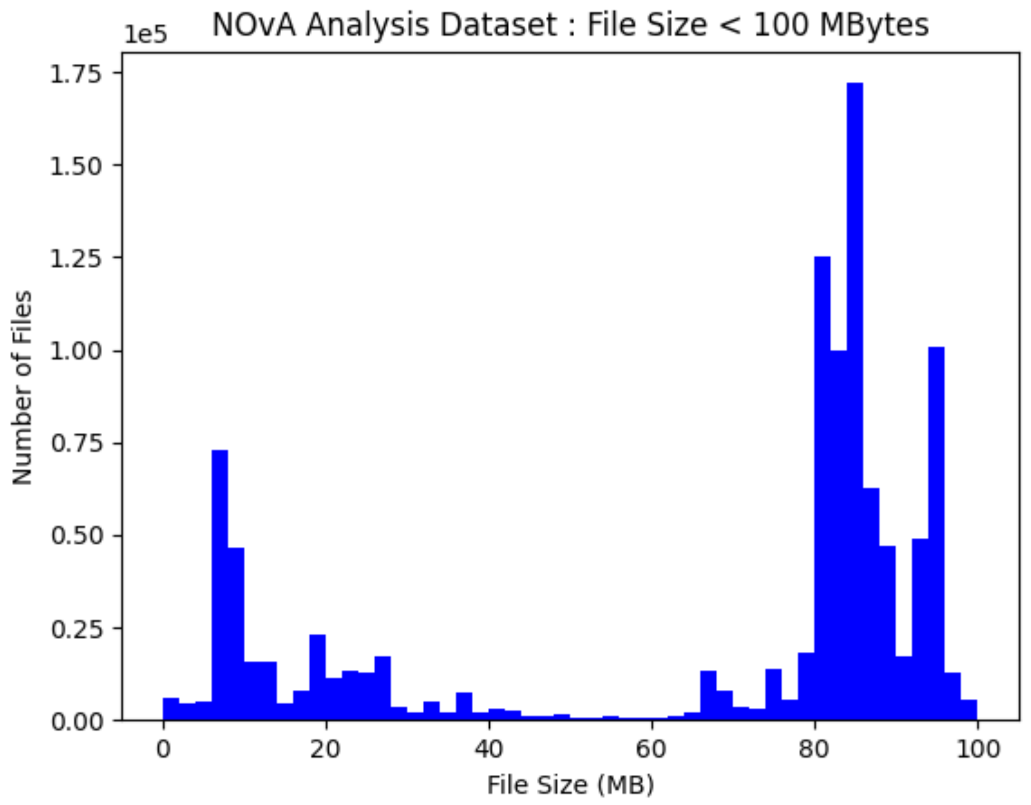
- Retooled CTA system for use as dCache frontend for a 3-day 200 TB test
  - Expected to start soon
  - Same hardware type as EOS test: 2 storage nodes, 72 TB NVMe each. 1 head node
- Earlier connected dCache and CTA as a proof of principle with test hardware
  - Wrote about 12 TB of files (average size 2 GB) to M8 tapes
  - No issues found in a test of this scale
- dCache is needed for one of our two systems because of how we’ve handled small files

CTA and small files

- Small files continue to be a concern for Fermilab
- Enstore had “Small File Aggregation” where files under ~200 MB were tarred and written to tape
  - Recalled by reading and untarring entire archive, effective when multiple files needed
  - This system was well (ab)used by some experiments
- dCache will be slightly modified to easily extract these files from CTA
- We will discourage experiments from writing so many small files
- Currently measuring CTA performance for various file sizes to inform this guidance

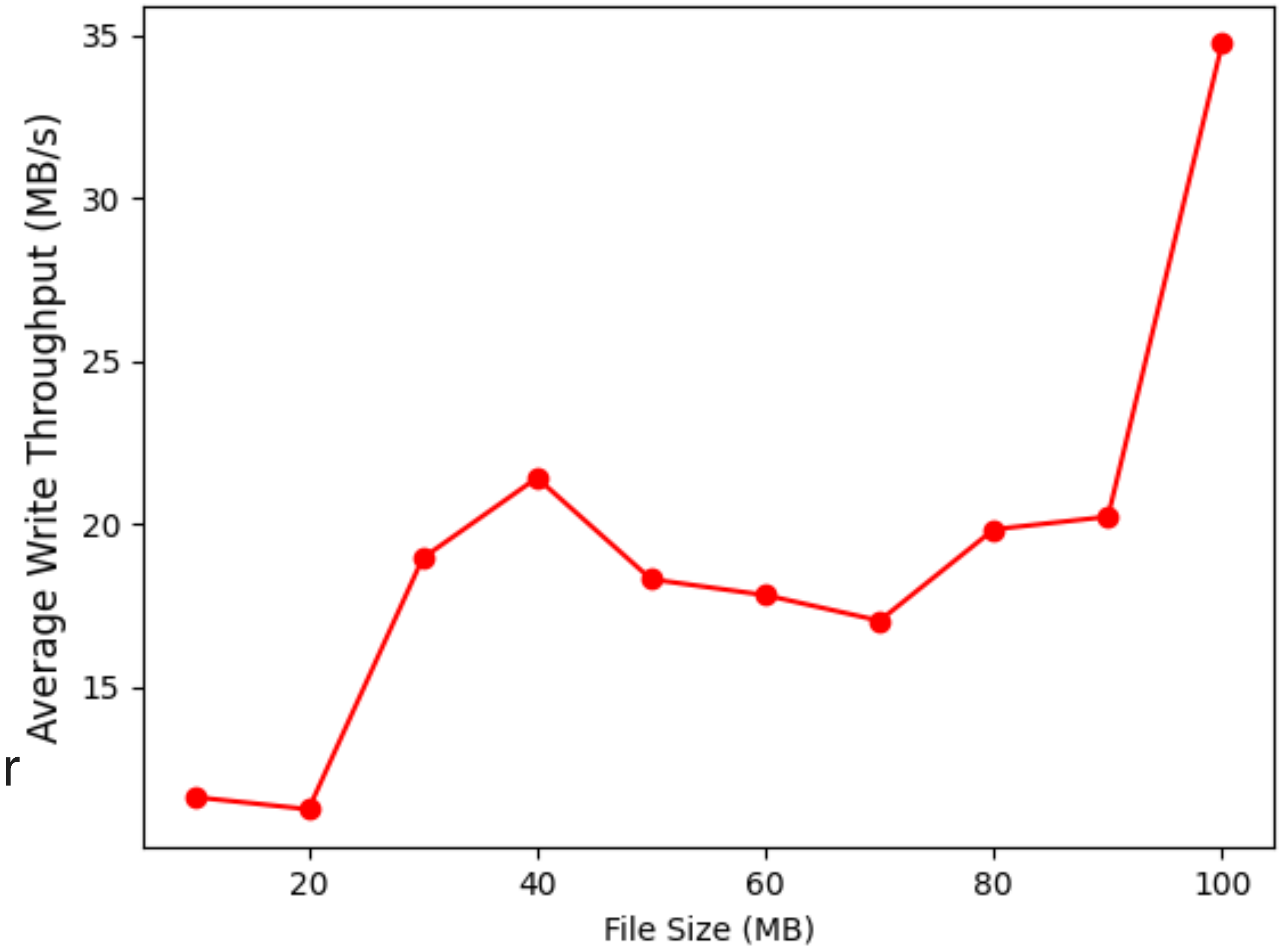
Testing Small Files

- Many intensity frontier experiments (such as NOνA) create files with size less than 100 MB



Standard analysis dataset

- $N(\text{files})_{\text{Total}} = 1{,}088{,}578$
- $D(\text{size})_{\text{Total}} = 74.4$ TB
- File size ranges from 127 KB to 5 GB
- 97% of the file sizes are < 100 MB
- The peaks correspond to different types of data streams



- Testing Setup (using our dev setup):
  - Data range from 100 MB – 10 MB in increments of 10MB
  - Original data are stored on Fermilab dCache, copied to EOS with docker container

Changes needed to read Enstore tapes

We modified CTA to be able to read normal Enstore tapes

- This is tape_label_format=2 in CTA

Enstore also writes into what we call the “CERN” format if files exceed 8 GB

- This is similar to, but not quite the CASTOR format
  - Volume label and headers can have multiple 80-byte blocks
  - Tape mark after volume label
- As with Enstore format, we don’t know the block IDs of files, have to position by file sequence #

To make matters even worse, when tapes are reused they are not relabeled

- We likely have tapes with “CERN” format and Enstore label
- And tapes in the Enstore format with the “CERN” label

PIC has used a dCache migration procedure and encountered these problems. We will develop a tape_label_format #3 and perhaps modify ‘format 2’ to account for these possibilities

- We hope to have this ready to merge sometime next month. Will update migration procedures

Metadata migration from Enstore to CTA

CTA metadata is in two places: CTA (tape file metadata) and EOS or dCache (namespace and file metadata).

Enstore metadata is sufficient to populate both

Similar database structures

We have two different metadata migrators. For dCache this can be completely in database(s).

- For EOS we need a part that interacts with the filesystem. Done in python
- PIC has successfully used the dCache migrator (except for “CERN” format)

Suggestions

A prerelease public RPM repository would be welcome, especially for Alma 9

The practice of announcing CTA dev meetings (or not) on the same day in CERN’s morning is not very friendly for people in the Americas.

Is there ops functionality which should be moved into the core project?

- Moving tapes from supply to pools — first class concept in Enstore

Continued interest in sharing ops information and monitoring tools

— cta-ops effort is very welcome

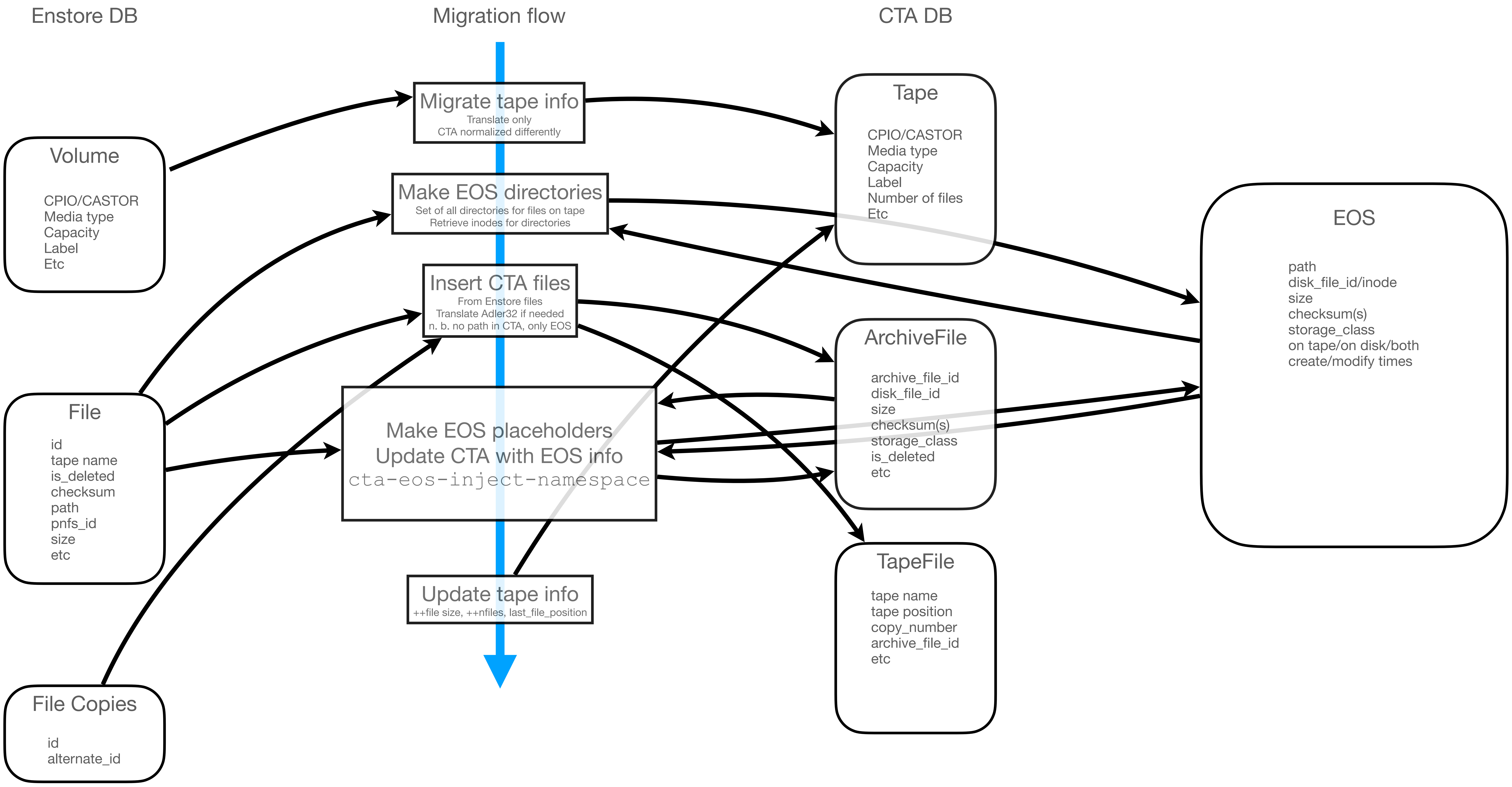
— Is there room for a regular forum (meeting) for this? Suggested by Julien – thank you!

Summary

Migration of our systems will being this year, finish next year

Looking forward to Alma9 and ability to move away from Ceph (although Ceph has been working OK for us)

More and more of our team is developing operational experience



Enstore DB

Migration flow

CTA DB

Volume

CPIO/CASTOR
Media type
Capacity
Label
Etc

File

id
tape name
is_deleted
checksum
path
pnfs_id
size
etc

File Copies

id
alternate_id

Migrate tape info

Translate only
CTA normalized differently

Make EOS directories

Set of all directories for files on tape
Retrieve inodes for directories

Insert CTA files

From Enstore files
Translate Adler32 if needed
n. b. no path in CTA, only EOS

Make EOS placeholders
Update CTA with EOS info
`cta-eos-inject-namespace`

Update tape info

++file size, ++nfiles, last_file_position

Tape

CPIO/CASTOR
Media type
Capacity
Label
Number of files
Etc

ArchiveFile

archive_file_id
disk_file_id
size
checksum(s)
storage_class
is_deleted
etc

TapeFile

tape name
tape position
copy_number
archive_file_id
etc

EOS

path
disk_file_id/inode
size
checksum(s)
storage_class
on tape/on disk/both
create/modify times

Not shown: tables for libraries, media types, storage classes