Fermilab CTA efforts and migration plan

Eric Vaandering, Ren Bauer, Dmitry Litvintsev, Scarlet Norberg

EOS 2023 Workshop

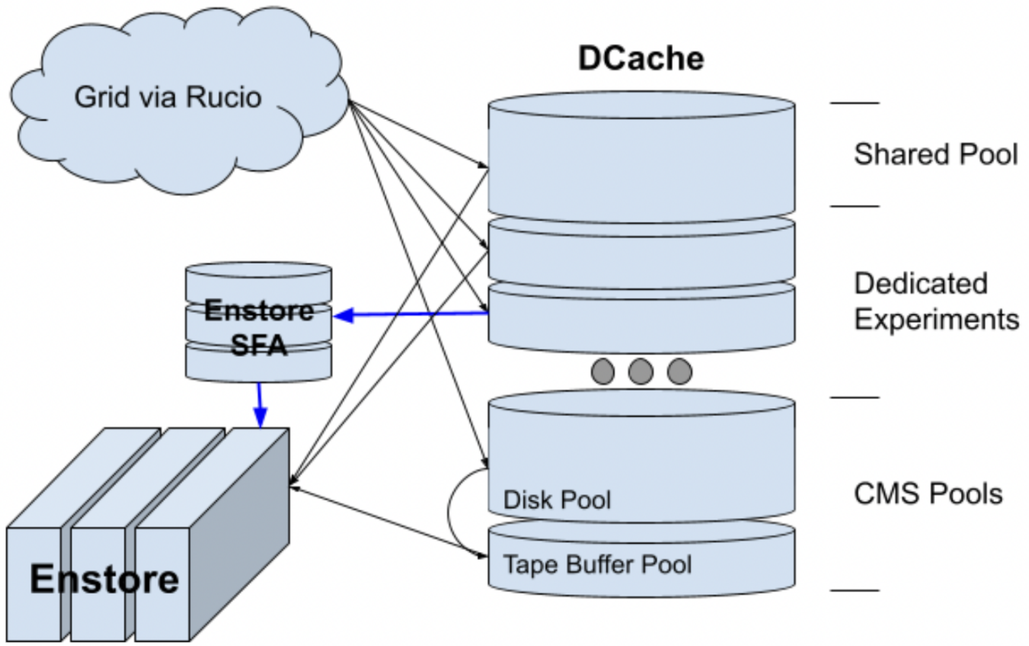
26 April 2023

Outline

- Metadata migration from Enstore to CTA
- Performance testing and monitoring
- Results of dCache integration with CTA at Fermilab
- Small File Aggregation thoughts
- 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, and an Oracle SL8500 (almost deprecated)
- ~300 Petabytes stored data
- 400+ Terabytes R/W per day

SFA

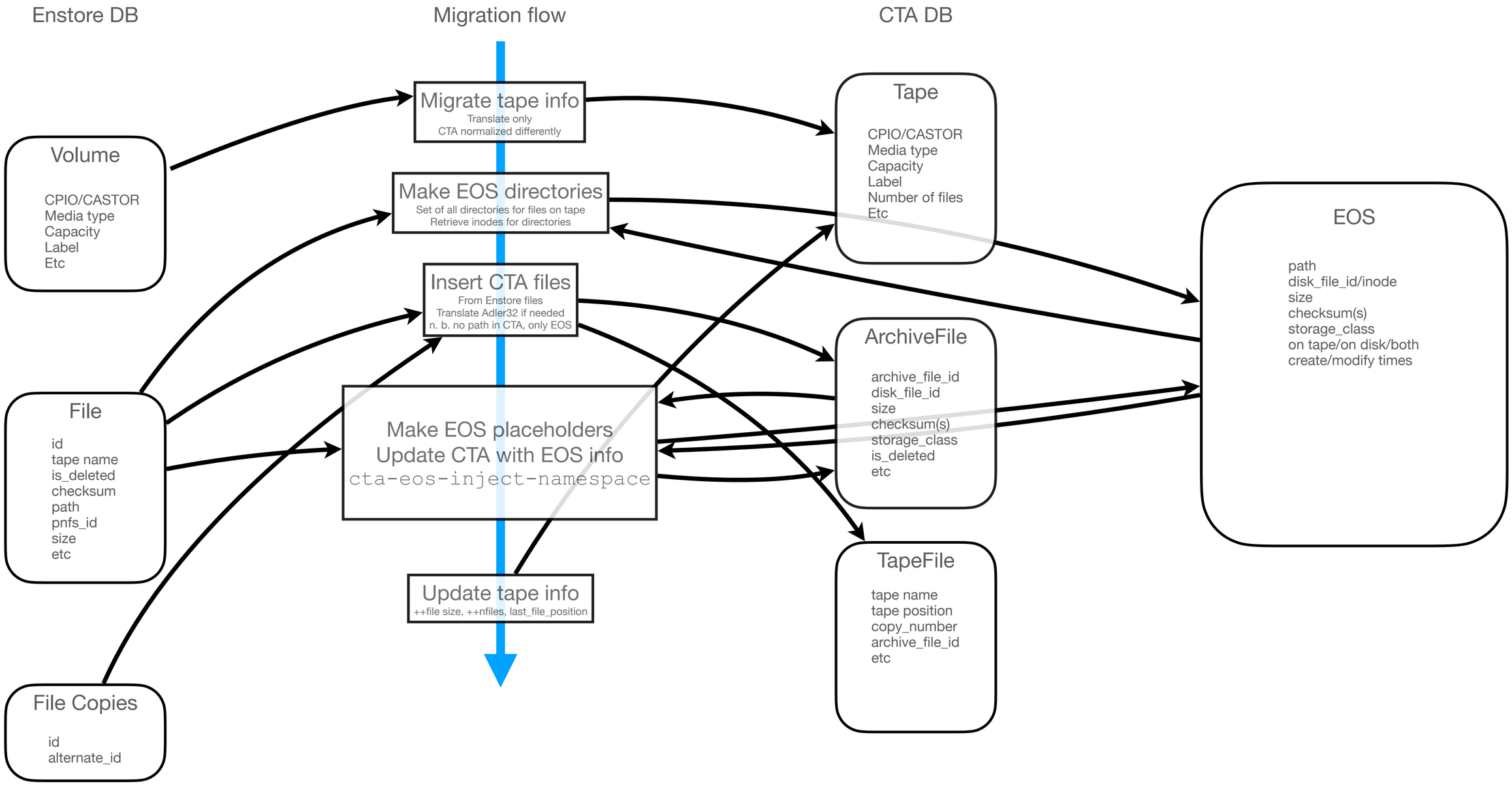
- Small File Aggregation service, utilized by public experiments to package tape data
- Writing about 3 TiB per day

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

Metadata migration from Enstore to CTA

We're using Python scripts to migrate from Enstore to CTA database combined with `cta-eos-inject-namespace`

- Using PostgreSQL for both database
- Similar database structures with differences
  - Enstore keeps the namespace (and so does dCache); delegated to EOS only in CTA
  - Breakdown between ArchiveFile and TapeFile not in Enstore
  - Enstore does have a way to store multiple tape copies
  - Enstore does not store tape block #, becomes important later
- Migration runs (single process) at about 11 Hz (files) or ~1M/day .
  - Borderline of what's acceptable (about 35 days for CMS)
  - 80% of the time taken with EOS operations and checking
  - Migration in dCache should be much faster, metadata only



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

dCache Integration with CTA

- FNAL's public dCache is HSM, assumption is this instance, at least, will be dCache+CTA
- Configured dCache / CTA interface on dCache pool
- Wrote abut 12 TiB of files (average size 2 GB) to M8 tapes
  - `cta-admin dr ls` shows something like 234 MB/s rate
    - Not sure that this is particularly accurate
- Observed excessive mounts
  - corrected by: `cta-admin mp ch --name ctasystest --minarchiverequestage 600 --minretrieverequestage 600`
  - After that no issues - one mount and files written until all done
- Randomly read back written files

Changes needed to read Enstore tapes

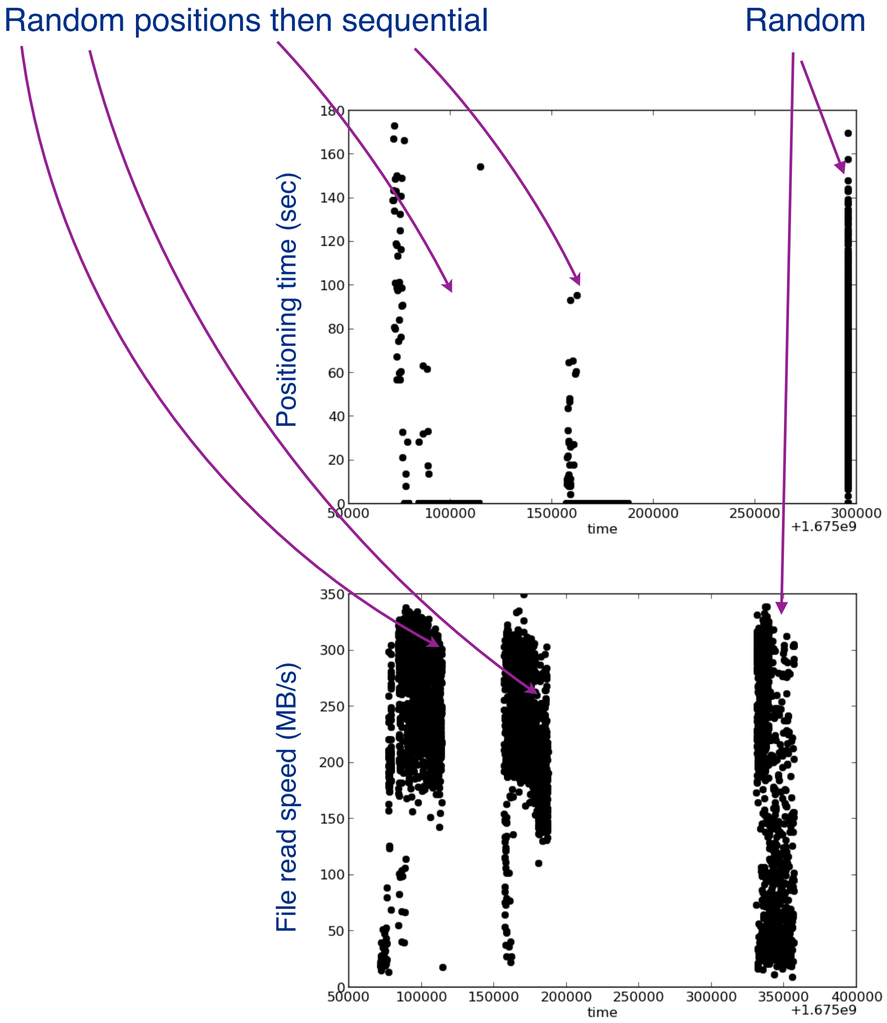
Enstore tapes have similarities to CTA tapes and OSM tapes

- Tape label is *almost* CTA. Same format, slightly different fields
  - CTA has been modified to understand this (tape_label_format=2)
- Like OSM, no headers/footers for files. CPIO-based file wrapper
  - Unlike OSM we only use this for files under 8 GB
- For files larger than 8 GB we have some tapes written with the CTA/CASTOR format
  - Have not tried reading these back yet (unclear if the label is CTA or Enstore)
- Because we don’t know the block IDs of Enstore files, have to position by file sequence number
  - This imposes quite a performance penalty

Performance monitoring

Some early monitoring based on log parsing (Thanks RAL!)

- Again with M8 tapes. When reading sequential files from a tape, performance approaches 400 MB/s theoretical max
- Any tape repositioning is very slow (average a minute or more)
  - With our file sizes, more time spent positioning than reading
  - Will try to compare with native Enstore and CTA
  - Frequently read tapes may be repacked to CTA before EOL

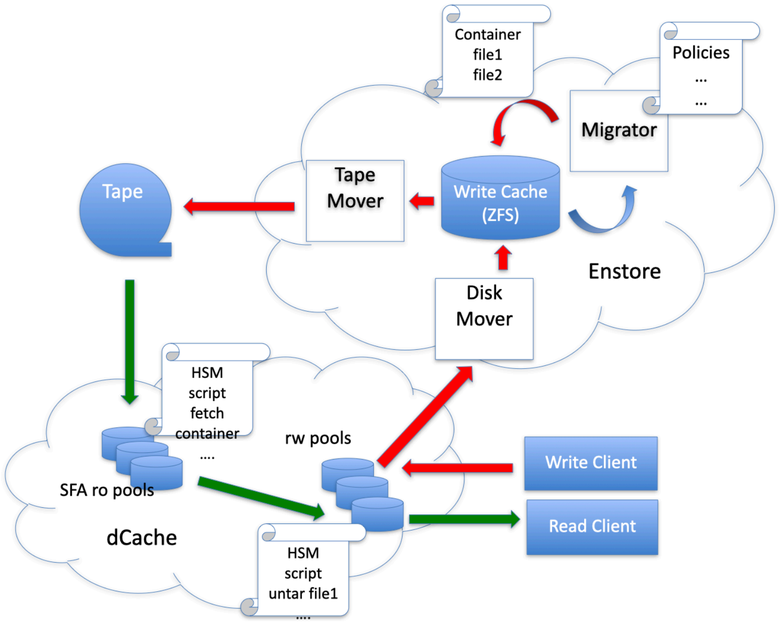


Three tests over the course of a few days
(x-axes don't quite line up)

Small File Archive (SFA)

Unique to Fermilab

- To minimize filemark writing, long seeks, and tape wear
- Small files are cached until, by policy, a group in a storage group are ready to write
  - Written to a single tar file, tracked in Enstore DB
- dCache is responsible for retrieving the small files from Enstore
  - Read-only SFA from CTA can be done similarly
    - A small of Enstore metadata to be migrated to dCache schema
- For writing small files we have options:
  - Is CTA fast enough with buffered writes and positioning to read and write?
  - If not, adopt “Sapphire” from DESY giving SFA in dCache



Fermilab

Migration Plan and Timeline

We have made the decision some months ago to move Fermilab to CTA

Two distinct migrations, CMS and Public (all other experiments)

CMS

- Decide if disk pool in front of tape will be dCache or EOS
  - What are the pros and cons of each approach?
- Begin with an $\mathcal{O}(10\%)$ setup including a Rucio RSE to test the whole stack
- Aim to have the full migration started by end of 2023
  - IBM allows tapes to be moved between virtual libraries and resizing libraries. Spectralogic appears not to

Public

- Will follow on a somewhat longer timescale

What Will Our CTA Setup Look Like?

Currently using Ceph for the object store, looking forward to PostgreSQL support in the future

We will run two drives per taped node. Probably implies containerized tapes

PostgreSQL for CTA database supplied by FNAL database group

Would like to virtualize as much as possible (e.g. CTA frontend) in our OKD cluster

Still need some advice here about hardware setups, etc. Looking forward to learning from what everyone else is doing.

Summary

Will be migrating to CTA

We're still working out what the components of our CTA system will look like

    We hope to leave this week with some ideas

10% test later this year should inform the migration process

Examples of monitoring, operations scripts/jobs would be helpful