CERN

CERN Tape Archive: production status, migration from CASTOR and new features

Eric Cano[1] Vladímir Bahyl[1] Cédric Caffy[1] Germán Cancio[1] Michael Davis[1] Viktor Kotlyar[2]
Julien Leduc[1] Giuseppe Lo Presti[1] Tao Lin[3] Steven Murray[1]

[1] CERN, Geneva
[2] Institute for High Energy Physics, Protvino, Russia
[3] Institute of High Energy Physics, Chinese Academy of Sciences

What is CTA?

CTA service, integration with EOS and FTS

New features to the software stack

Migration from CASTOR to CTA

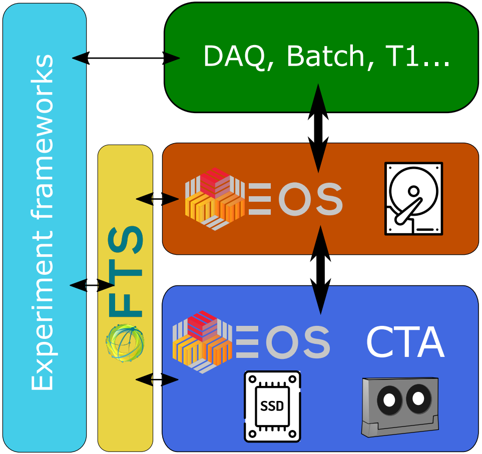
Future functionality

Conclusions

What is CTA?

- The CERN Tape Archive (CTA) is a tape backend for EOS
  - EOS is the CERN developed disk system for physics data
- A central CTA is the tape backend for several EOSCTA instances
- EOS:
  - provides the user interface (XRootd)
  - holds the directory structure, files metadata
  - manages disk (HDD or SSD) buffers
  - handles garbage collection
- CTA:
  - manages transparently file residence on tape
  - transfers tape files to/from disk cache on request from EOS

Current deployment model (I)



HDD icon: https://commons.wikimedia.org/wiki/File:Hard-drive.svg
SSD icon: https://commons.wikimedia.org/wiki/File:Ssd.svg
Tape icon: https://commons.wikimedia.org/wiki/File:Tape_cinta_casette_backup.svg

- Analysis/DAQ goes through the disk based EOS instance
- EOSCTA instances are dedicated to tape archive
- One EOSCTA instance per experiment, with dedicated namespace and storage space
- CTA is a shared tape service providing tape storage to all EOSCTA instances
- The File Transfer Service (FTS) can manage stage-in and transfers between EOS and EOSCTA

Current deployment model (II)

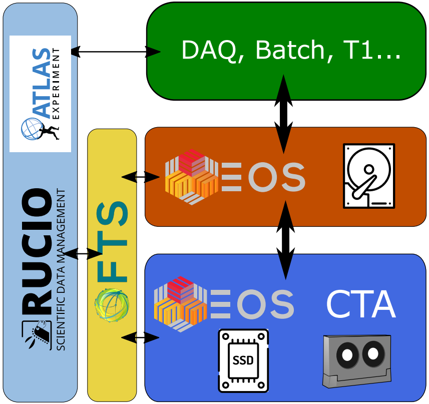
Plans for CERN T0 deployment

- Equipment in procurement
- Design validated in smaller scale pre-production
- 30 disk servers with ≈1 PB of SSD for buffering
- Tape servers transferred from CASTOR
    - Upgrade cycle coming soon

Integration with EOS and FTS

- All protocols and disk storage functionality available via EOS
- EOS has been extended to support tape related operations
  - Lifecycle management
  - Stage in
  - Cache/buffer management
  - Garbage collector
- More details in “EOS architectural evolution and strategic development directions”
- FTS support added for stage in in EOSCTA, transfer to EOS, eviction from buffer
- See also “FTS improvements for LHC Run-3 and beyond”

Integration with experiments: ATLAS



ATLAS EXPERIMENT

RUCIO SCIENTIFIC DATA MANAGEMENT

FTS

DAQ, Batch, T1...

EOS

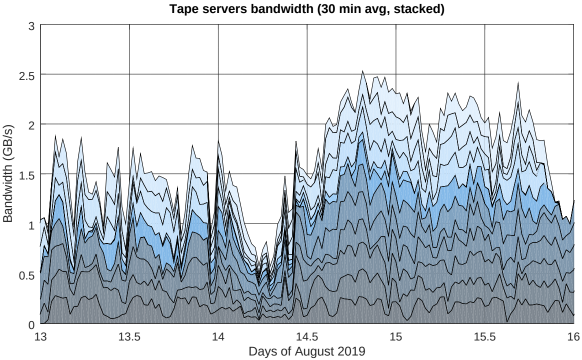
EOS CTA

SSD

HDD icon: https://commons.wikimedia.org/wiki/File:Hard-drive.svg
SSD icon: https://commons.wikimedia.org/wiki/File:Ssd.svg
Tape icon: https://commons.wikimedia.org/wiki/File:Tape_cinta_casette_backup.svg

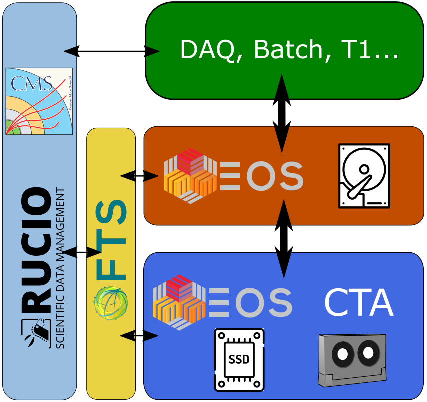
- Most advanced tests with an experiment
- Integration of Rucio
- Data taking: reached hardware saturation (April 2019)
- Data recall: Participated in 2018 reprocessing campaign (August 2019)
- See "The ATLAS Data Carousel Project"

ATLAS 2018 reprocessing campaign (Aug. 2019)



- Opportunity to replace a T1 with small scale EOSCTA instance
- Read-only EOS-CTA instance with imported CASTOR data for ALICE
  - Metadata imported from CASTOR
  - Actual tapes “borrowed” from CASTOR
- Limited number of SSDs (24 × 1 TB)
- 4 disk servers
- ATLAS’ Rucio used FTS to handle EOSCTA to EOS transfers
- Reached network limit for this deployment (2.5 GB/s)

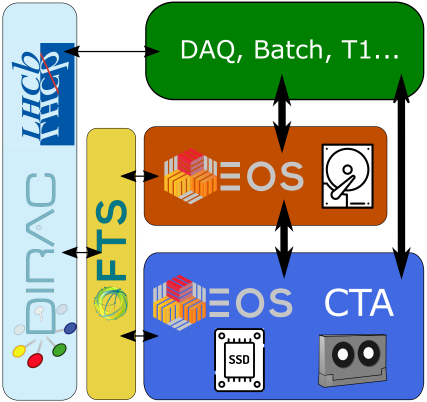
Integration with experiments: CMS



HDD icon: https://commons.wikimedia.org/wiki/File:Hard-drive.svg
SSD icon: https://commons.wikimedia.org/wiki/File:Ssd.svg
Tape icon: https://commons.wikimedia.org/wiki/File:Tape_cinta_casette_backup.svg

- Integration less advanced but expected to be similar to the one of ATLAS with CMS adopting Rucio

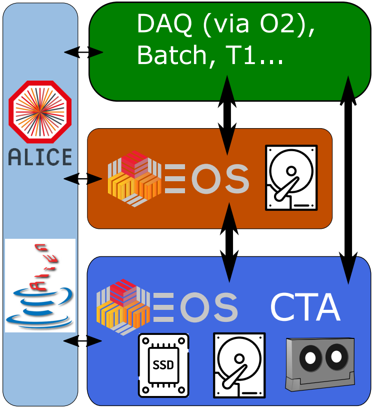
Integration with experiments: LHCb



HDD icon: https://commons.wikimedia.org/wiki/File:Hard-drive.svg
SSD icon: https://commons.wikimedia.org/wiki/File:Ssd.svg
Tape icon: https://commons.wikimedia.org/wiki/File:Tape_cinta_casette_backup.svg

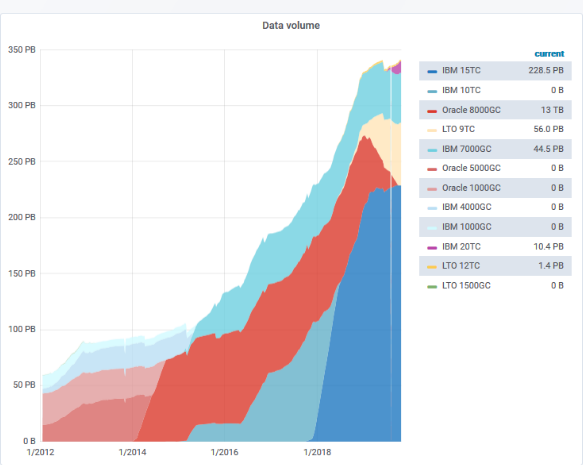
- Integration less advanced but expected to be similar to the one of ATLAS, with Dirac
- Direct T1 export for a small fraction of cases

Integration with experiments: ALICE



HDD icon: https://commons.wikimedia.org/wiki/File:Hard-drive.svg
SSD icon: https://commons.wikimedia.org/wiki/File:Ssd.svg
Tape icon: https://commons.wikimedia.org/wiki/File:Tape_cinta_casette_backup.svg

- Different use case
- SSD buffer for data intake
- Garbage collected ≈5 PB HDD cache for retrieves
- Occasional T1 access to EOSCTA



		current
🟦	IBM 15TC	228.5 PB
🟦	IBM 10TC	0 B
🟥	Oracle 8000GC	13 TB
🟧	LTO 9TC	56.0 PB
🟦	IBM 7000GC	44.5 PB
🟥	Oracle 5000GC	0 B
🟥	Oracle 1000GC	0 B
🟦	IBM 4000GC	0 B
🟦	IBM 1000GC	0 B
🟪	IBM 20TC	10.4 PB
🟨	LTO 12TC	1.4 PB
🟩	LTO 1500GC	0 B

Latest development (I)

Repack

- Behind the scene operations for tape repair, free space reclaim, media migration...
- Completion of repack enables production operations

Requests queueing

- Queueing was further optimized to reach 250 Hz

Latest development (II)

Retrieve scheduling: Activities scheduling

- Following discussions with ATLAS
- Intra-VO scheduling (other experiments unaffected)
- Prevents high latency for small request during big retrieve campaign

Retrieve scheduling: FIFO scheduling

- Avoids starving small mounts forever in high activity periods
- Optional switch between per-size or pre-age scheduling for retrieve (in development)

Tape file catalogue

- 3 backends for production: Oracle, PostgreSQL, MySQL
  - Oracle used for migration and current production
  - Targeting future production on PostgreSQL (no schedule yet)
  - MySQL support contributed by IHEP (Chinese Academy of Science). Many thanks to them!
- All flavors validated (like all of CTA) in continuous integration tests
  - See “System testing CERN physics archival software using Docker and Kubernetes”

Migration from CASTOR to CTA

- Metadata-only migration: tape format unchanged
- Migration from a single CASTOR namespace to 5 EOSCTA instances
- Oracle to Oracle for CASTOR → CTA
- Injection tool for CASTOR → EOS
- Migration experiment by experiment
- Requires preparation and cleanup on CASTOR side
- Can be done as a read-only import (validated during ATLAS test)
- Typical time few hours

Future functionality

- Provide fast positioning on LTO
  - Software solution addressing lack of drive-provided recommended access order (RAO)
- Pre-emptive scheduling
- Study of dataset collocation (PhD student)
- Various operations related features (drive dedication...)

Conclusions

- EOSCTA core ready for production usage
- Promising tests validated the deployment model
- Delivered full tape performance on limited hardware
- Migrations of experiments at various stages, some requiring adaptation in the pipeline
- Focus in 2020 will be migration out of CASTOR for LHC experiments

CERN