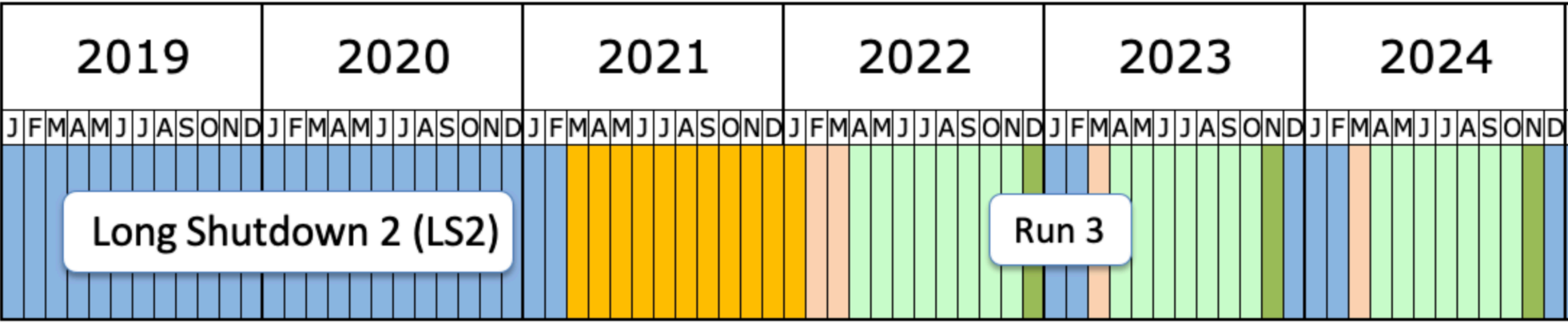


LHC Data Storage: RUN 3 preparation

CERN Tape Archive

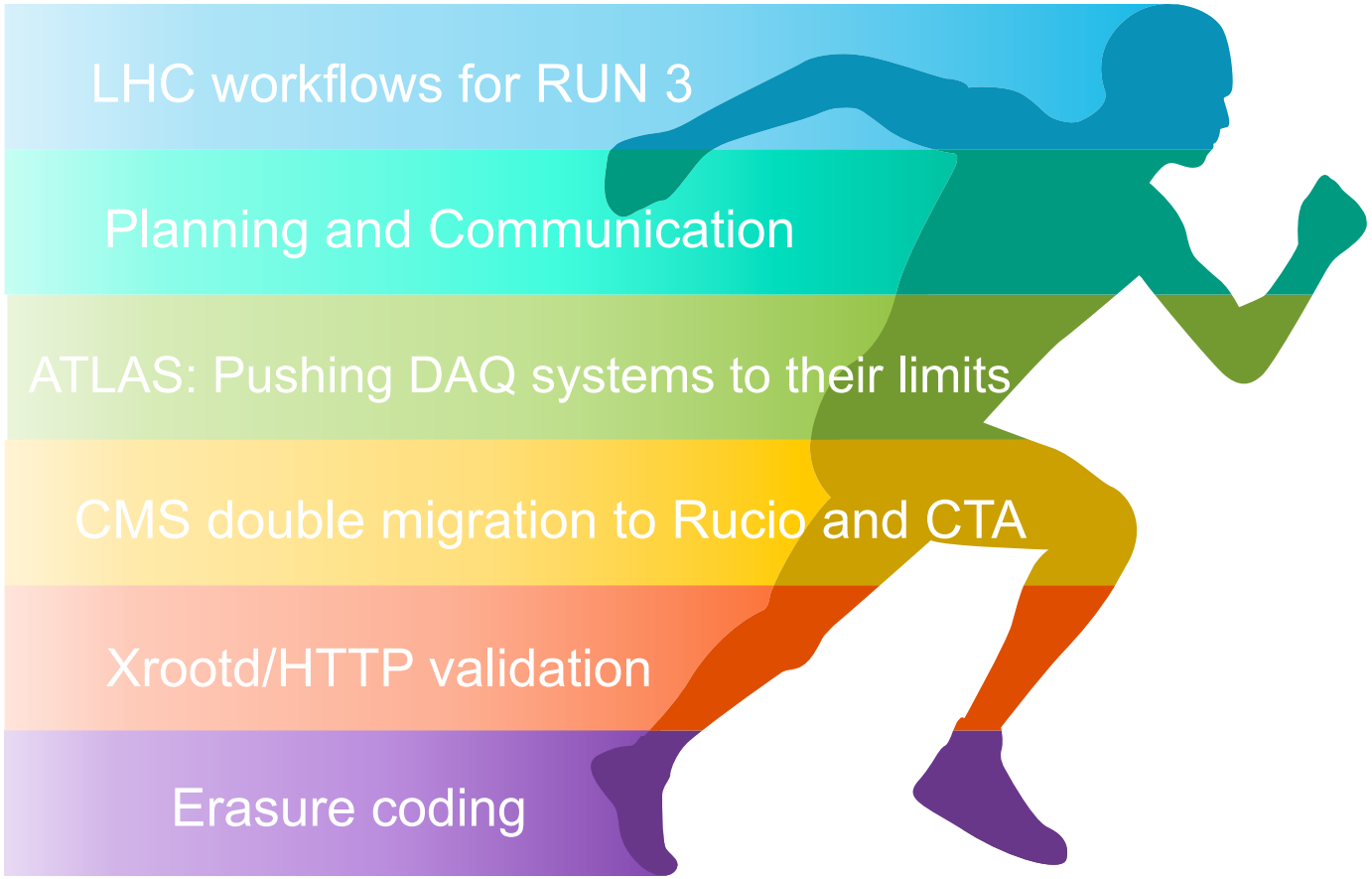
EOS

FTS
File Transfer Service

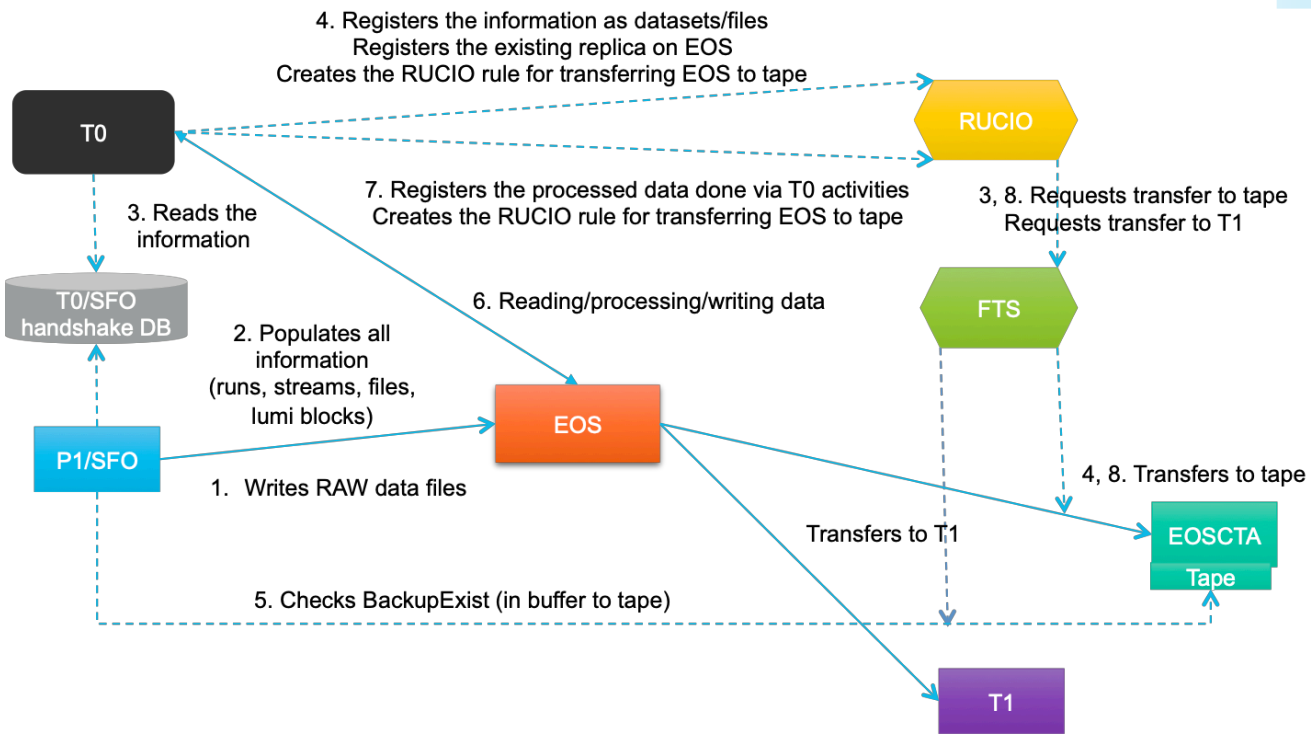
The CERN IT Storage Group ensures the symbiotic development and operations of storage and data transfer services for all CERN physics data, in particular the data generated by the four LHC experiments (ALICE, ATLAS, CMS and LHCb).

ALICE ATLAS EXPERIMENT CMS LHCb

RUN 3 preparation: First steps



ATLAS workflow



LHC workflows for RUN 3

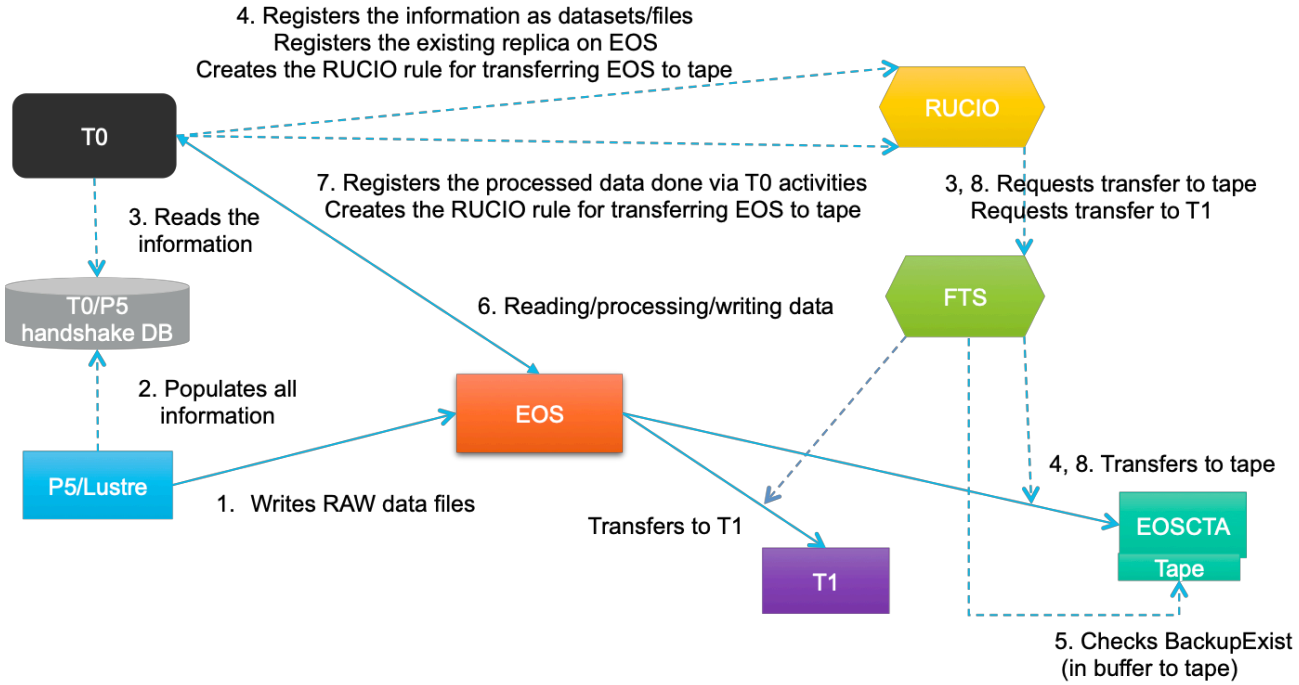
The data logger, or Sub-Farm Output (SFO), writes the RAW data files to EOS. Then it populates the T0/SFO handshake database with all necessary information about runs, “streams“, “lumi blocks“ and files.

Tier-0 (T0) reads all this information, and registers it as datasets/files in Rucio. T0 registers the existing replica on EOS and creates the Rucio rule for transferring files from EOS to CTA.

Finally, SFO checks if a tape copy has been successfully archived before deleting the file from its buffers in LHC Point 1.
During this process, T0 launches offline activities.

CMS workflow

LHC workflows for RUN 3



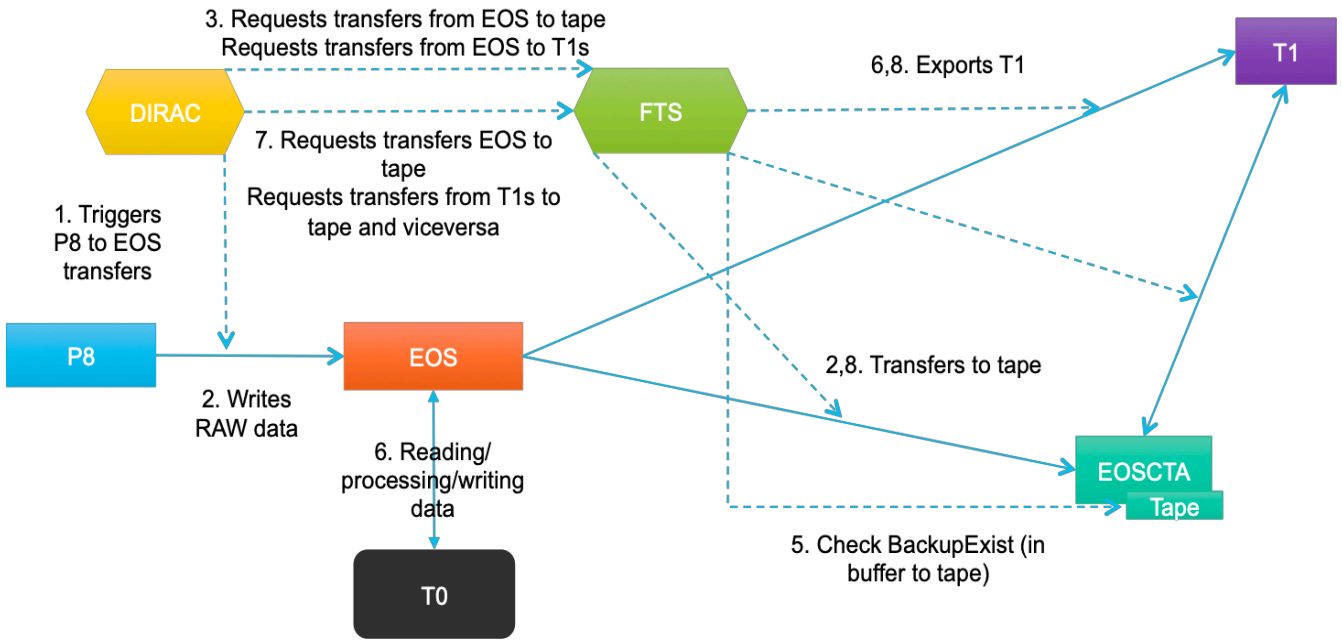
The CMS Data Acquisition system (DAQ), located in LHC Point 5, transfers the data to EOS and notifies Tier-0 (T0) using a database and a handshake protocol.

T0 reads all the information needed and registers it with Rucio, who triggers the transfers to CTA and Tier-1s (T1s), using the File Transfer Service (FTS).

One of CMS' main requirements was that FTS provides a new “Archive Monitoring" feature, which checks if and when the file has been successfully archived to tape.
T0 also performs offline activities

LHCb workflow

LHC workflows for RUN 3

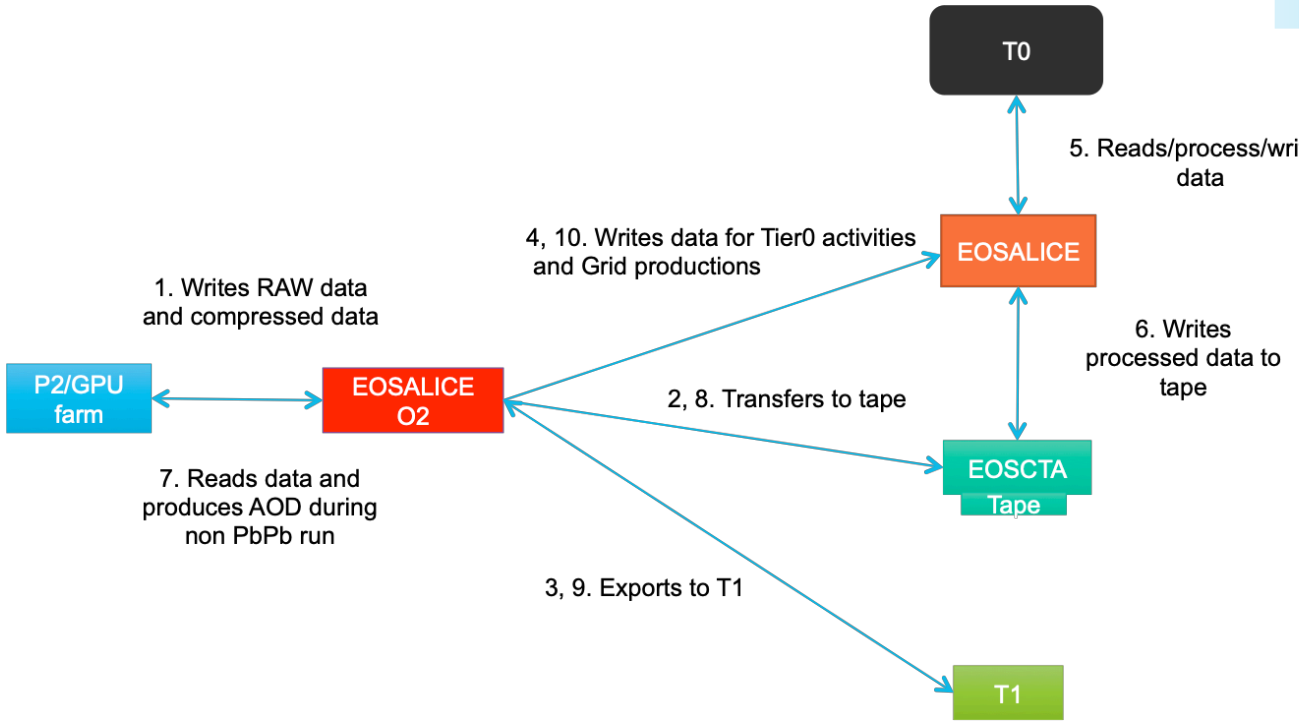


LHCb starts in the same way as the previous experiments, by sending the data from their DAQ system to their EOS instance, triggered by DIRAC, the LHCb data management system.

DIRAC relies on FTS for triggering transfers between EOS to CTA and EOS to Tier-1s.

DIRAC (and FTS) also manages the export and staging in between CTA and Tier-1s for online and offline activities. LHCb, like CMS, will depend on the new FTS Archive Monitoring feature.

ALICE workflow



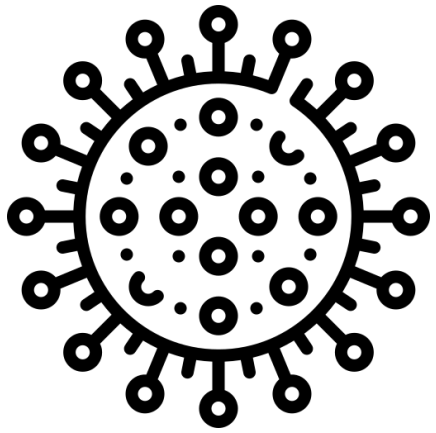
LHC workflows for RUN 3

The DAQ system sends the data to the dedicated EOSALICEO2 instance. EOSALICEO2 is, de facto, an extension of the ALICE Online farm, and acts as a cache to cope with the high data rate and allow high-performance processing.

EOSALICEO2 exports the data to Tier-1s and CTA, and also to other EOS instances such as EOSALICE, for T0 activities and Grid production.

The reprocessing done on EOSALICE by the T0 activities produces data which will be finally stored on tape. Therefore, both EOS instances communicate with CTA.

COVID Impact



- Multiple delays in hardware procurement
- Only virtual communications

Planning and Communication

Standard framework definition

Approve and validate the goals, dates, and blocking factors for individual tests, per component and per experiment

Individual tests per component and experiment

Data challenge tests with more than 3 experiments together

ATLAS : Pushing DAQ systems to their limits

PURPOSE

Handle the accumulated data in SFOs, due to possible issues :

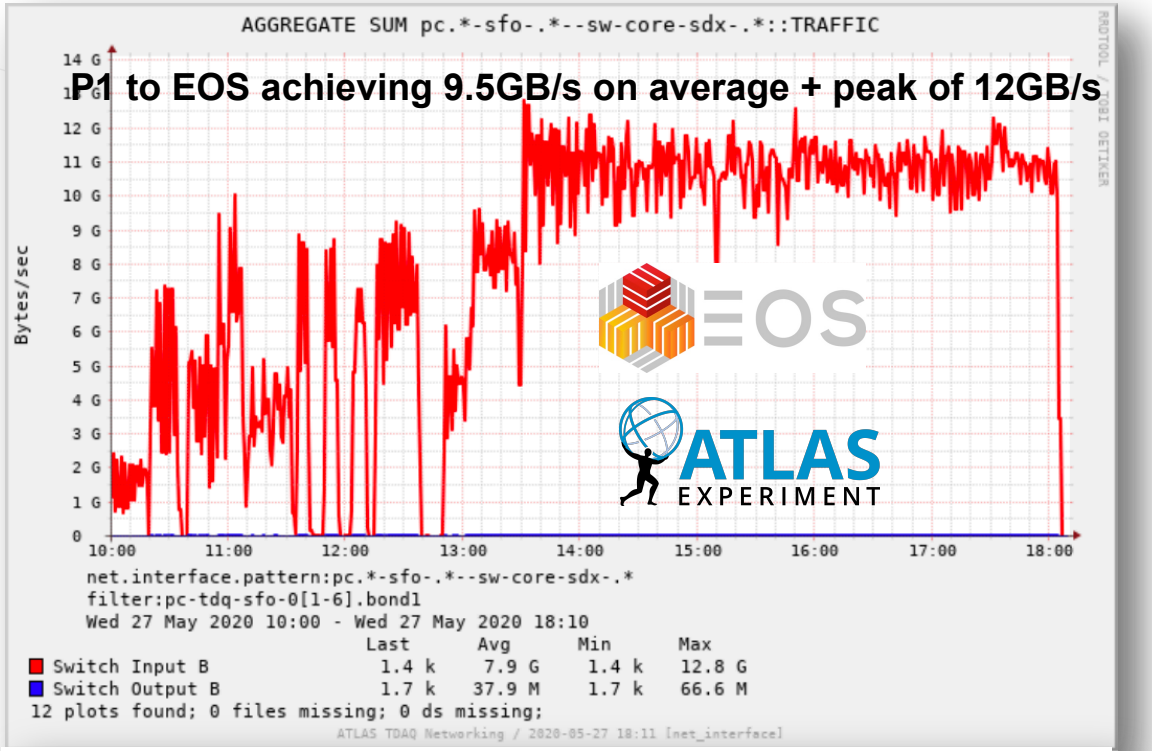
- broken network connection between LHC Point 1 (P1) and the computer center
- short EOS unavailability during a run

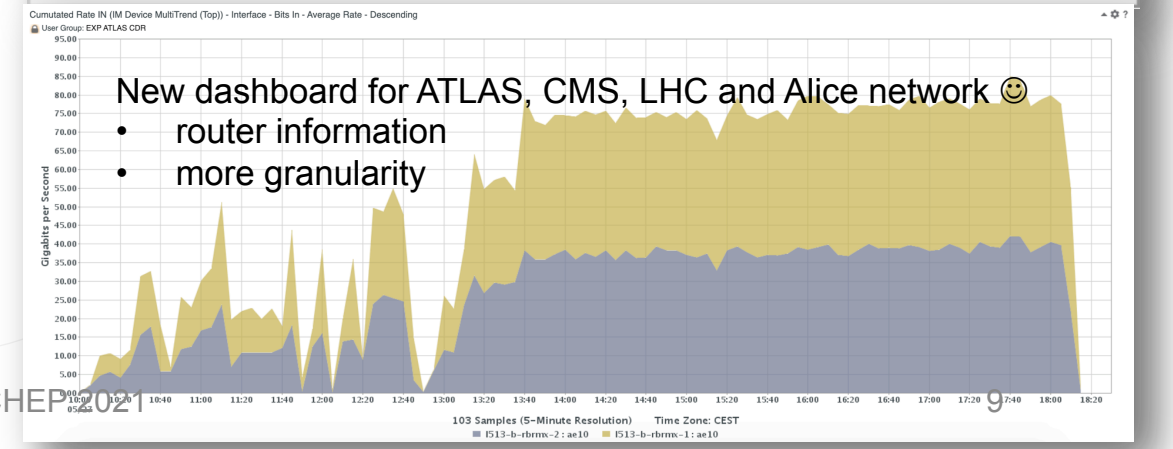
Scenario

P1 generates data without transferring to EOS, thus accumulating it on the SFOs. Once the SFOs are full, P1 stops the data generation and starts moving data to EOS. Data is deleted from the SFO as soon as it reaches EOS.

Numbers

The total amount of data generated is 250 TB with an average file size of 500 MB. SFO counts on six servers achieving 15 GB/s.
Goal: Handle peaks close to 15 GB/s. The maximum achieved was 7 GB/s in 2018.





- 100 transfers per server since 13:00.
- Bottleneck: DB updates executed for registering the transfer status.
- **EOS handled this traffic without any problem.**

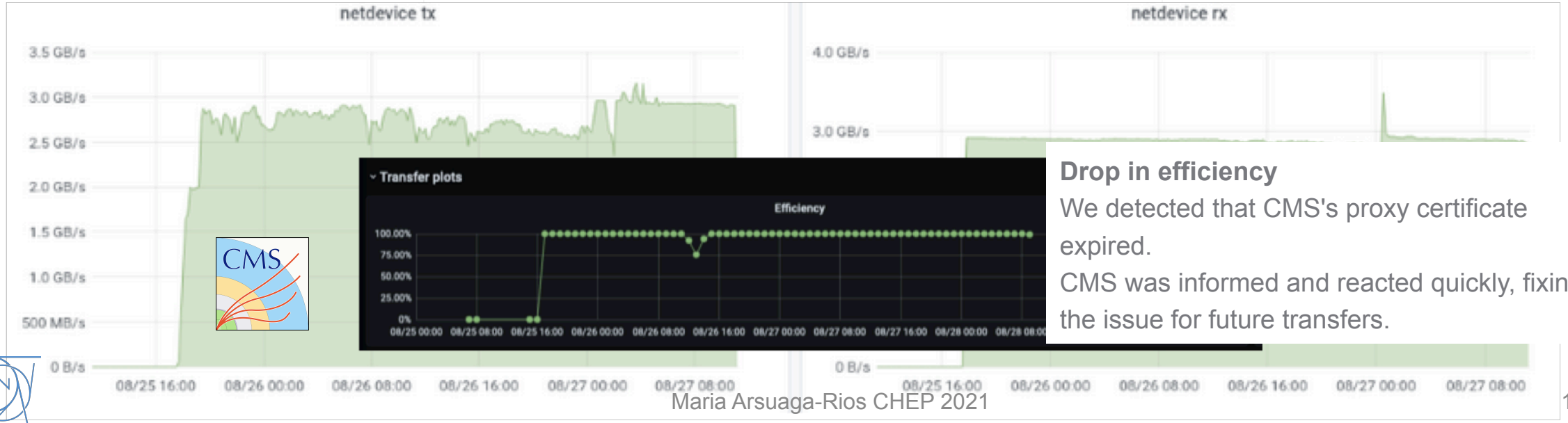
CMS double migration to Rucio and CTA

Main challenge: Migration from PhEDEx to Rucio

Crucial follow up by the storage team as CMS needed to create rule in Rucio to trigger transfers with the help of FTS.
We designed several functional tests to check that their migration uses FTS successfully with full performance transfers from EOS to CTA. Including new specific FTS features as multihop and the new archive monitoring.

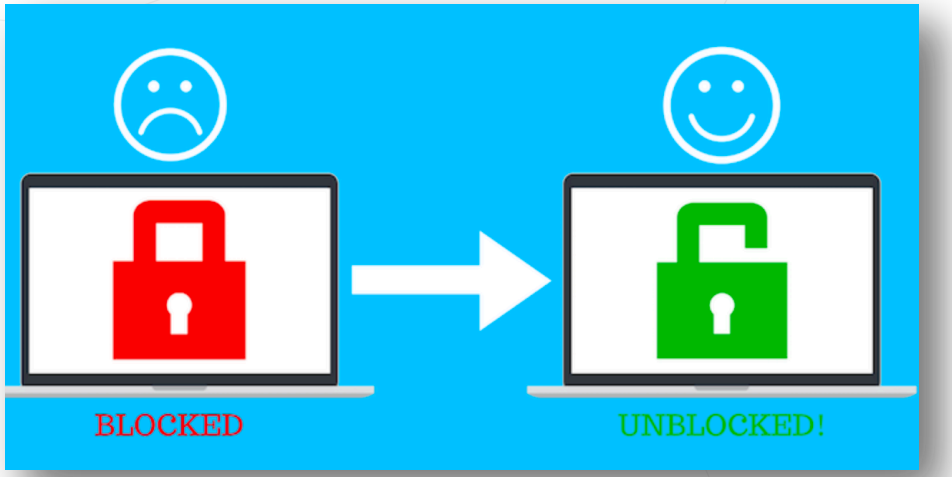
Large test from EOS to CTA (675 TB) with the Rucio production instance

CMS created the corresponding rules in their Rucio instance to send 675 TB from EOS to CTA with 166,510 files from a Neutrino dataset.
According to the CTA setup for this test, the expected throughput was a maximum of 3 GB/s.
This was achieved and sustained.



Drop in efficiency
We detected that CMS's proxy certificate expired.
CMS was informed and reacted quickly, fixing the issue for future transfers.

Xrootd/HTTP Third Party Copy (TPC) validation between Tier-0 and Tier-1s in LHCb



01

SRM-less process in dCache namespace migration

Deprecate space tokens in dCache T1s
Gridka, PIC, RAL, IN2P3, and SARA

02

Deploy and configure XRootD or HTTP Third Party Copy (TPC)

Our EOS instances at CERN were the first to support both TPC protocols
Validate all the T1s and debug any problems with their configuration

03

Validation test of 200 TB from EOSLHCb to CTA

04

Multiprotocol submission model:
XRootD stage-only and HTTP-TPC transfers from T0 to T1s

Optimized multihop feature in FTS to accomplish LHCb requirements
Tested and integrated with DIRAC to effect T0 to T1 transfers via HTTP-TPC

Erasure coding performance in ALICE

High performance streams

ALICE O2 use case requires high-performance streams due to limited transfer time budget and minimal storage costs with a nominal performance of up to 100 GB/s.

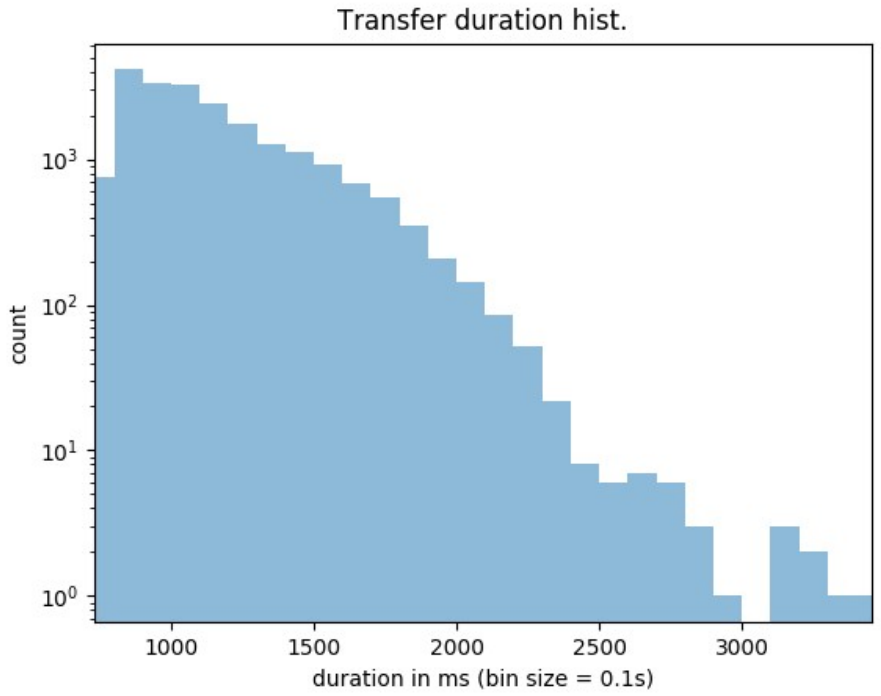
Erasure coding deployment for ALICE02

Scaling single stream performance with the number of data disks.
The number of parity disks and the ratio of data to parity disks are selected to yield the desired redundancy level and storage volume overhead.

Tails in the upload distribution

A possible problem for the ALICE O2 use case is tails in the upload time distribution.
If we run 1000 streams as fast as possible, we observe a wide distribution of upload times with a long tail, where some of them exceed the available upload time window.

Erasure coding performance in ALICE



ALICE Erasure Coding (EC): Upload time distribution for RS(12,10) using3×100GE disk servers with 96 hard disks and max. 300 streams

Reduce tail effect by limiting the upload bandwidth

Testing a client-driven EC plug-in, which creates data and parity blocks client-side and uploads them to data and parity server locations without going through a storage gateway.

Bisection in network traffic for writing and reading

Testing the client-driving EC plug-in

With only 3 server nodes achieves an average file transfer speeds of 1.8 GB/s, 15 GB/s in total and not a single transfer exceeds the defined upload window.
The estimated write bandwidth for the complete ALICE O2 pool would be 375 GB/s based on the performed measurements.

Conclusions and future work

Validate the migration to a new data management system (Rucio in case of CMS)

Verify the new Archive Monitoring feature of FTS

Planning multiple collective data challenges with more than two experiments simultaneously

Validate the CTA migrations with extensive tests

Obtain the throughput objectives of the DAQ systems

Confirm correct setups for Third Party Copy protocols