EPJ Web of Conferences **295**, 01041 (2024) https://doi.org/10.1051/epjconf/202429501041
CHEP 2023

Operation of the CERN disk storage infrastructure during LHC Run-3

Cedric Caffy[1,*], *Guilherme* Amadio[1], *Maria* Arsuaga Rios[1], *Manuel* Batalha Reis[1], *Niccolo* Capitoni[1], *Cristian* Contescu[1], *Jaroslav* Guenther[1], *Luca* Mascetti[1], *Andreas* Peters[1], *Elvin* Sindrilaru[1], *Ioanna* Vrachnaki[1], and *Abhishek* Lekshmanan[1]

[1]CERN

Abstract.
The CERN IT Storage group operates multiple distributed storage systems to support all CERN data storage requirements. The storage and distribution of physics data generated by LHC and non-LHC experiments is one of the biggest challenges the group has to take on during LHC Run-3.
EOS [1], the CERN distributed disk storage system is playing a key role in LHC data-taking. During the first ten months of 2022, more than 440PB have been written by the experiments and 2.9EB have been read out.
The data storage requirements of LHC Run-3 are higher than what was previously delivered. The storage operations team has started investigating multiple areas to upgrade and optimize the current storage resources. A new, dedicated and redundant EOS infrastructure based on 100Gbit servers was installed, commissioned and deployed for the ALICE Online and Offline (O2) project. This cluster can sustain high-throughput data transfer between the ALICE Event Processing Nodes (EPN) and the CERN's data center.
This paper will present the architecture, techniques and workflows in place allowing EOS to deliver fast, reliable and scalable data storage to meet experiment needs during LHC Run-3 and beyond.

1 Introduction

EOS is the distributed disk storage system developed and used at CERN for storing the large amount of data created by different user communities. It is also the basis of a wider ecosystem allowing to serve different data storage use cases. Among them, CERNBox [2] - that provides cloud storage - and CTA[1] [3] that provides tape storage for long-term physics data archival.

Since the beginning of LHC Run-3 in July 2022 up to end of August 2023, around 750 PB of physics data has been written to EOS by the LHC and non LHC experiments and around 5.3 EB of data has been read out.

In this paper, we will first present the usage of EOS by the experiments and the popularity of the different access interfaces that EOS provides. In a second part, we will present the hardware deployed at CERN allowing EOS to deliver the required storage space and throughput for LHC Run 3. Next, we will show the general LHC experiments data taking workflows and

---

[*] e-mail: cedric.caffy@cern.ch

[1] CERN Tape Archive

present the ALICE setup that is quite different from the other ones. Finally we will describe the deployment in production of Erasure Coding file storage layout allowing to optimize raw storage space capacity and the operation of EOS at CERN during LHC Run 3.

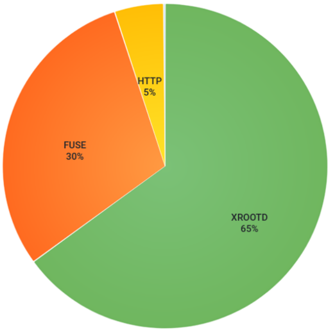
2 The usage of EOS by the experiments

EOS is the main disk storage system used by the CERN experiments to store and access their physics data. Since the beginning of the LHC-Run 3 up to end of August 2023, around 2 billion files for 750 PB of data have been written to EOS and 18.5 billion files for 5.3 EB have been read out.

EOS offers different files access interfaces:

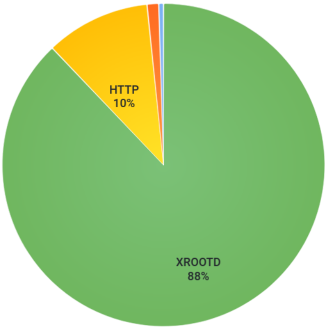
- XRootD - EOS native file transfer protocol
- FUSE - mountpoint on the linux filesystem
- HTTP
- GRIDFTP - extension of the FTP protocol for grid computing

At CERN, the experiments use all of them. The XRootD access interface is the most used for data read and write followed by FUSE access for data read and HTTP access for writes. GRIDFTP is becoming less and less used in favor of HTTP. The figures 1 and 2 show the distribution of the different files access interface used by the experiments.



Interface	Volume
XRootD	3.44 EB
FUSE	1.58 EB
HTTP	254 PB
GRIDFTP	3.77 PB

Figure 1: Total data read by CERN experiments per access interface



Interface	Volume
XRootD	660 PB
HTTP	78.1 PB
FUSE	8.15 PB
GRIDFTP	2.96 PB

Figure 2: Total data written by CERN experiments per access interface

The CERN experiments can authenticate to EOS via different ways:

- x509 certificates
- Kerberos 5 tickets
- EOS native tokens (EOS tokens)
- Macaroons tokens
- Scientific Tokens (SciTokens) [4]

The authorization relies on POSIX permission and ACL (Access Control List) that can be set by the experiment at the directory level.

3 The EOS cluster hardware deployed at CERN

The EOS infrastructure is composed of three main components, the Management node (MGM) is the initial point of contact for external clients, responsible for the authentication and authorization of the user, dealing with file metadata operations (registration, modification...) and client-redirection to the physical location of their file. The File Storage node (FST) is a server responsible for managing the physical storage of the files. It is very often attached to several disks. The namespace, provided by QuarkDB [5], is responsible for the persistency of the metadata of every files stored in EOS.

The MGM and the namespace servers are composed of 386 GB of RAM and 64 cores CPU Intel Xeon(R) Silver 4216 CPU @ 2.10GHz.

In terms of storage hardware, 790 PB of raw space is currently deployed. This capacity is provided by around 1300 disk servers for a total of 60000 disks.

The latest storage hardware delivered is composed of 256 GB or RAM, 96 x 18 TB disks and 100Gbps NIC interfaces.

4 General LHC experiment data taking workflow

The LHC is a 27km ring installed 100m underground where particles are accelerated at almost the speed of light and collided at 4 different points. Around these collision points, detectors have been placed in order to detect the particles created by these collisions. Several millions of collision per second occur, however only few of them are of interest for the physicist. The data coming from all these collisions are therefore filtered via different level of High Level Triggers (HLT) to only keep the most relevant ones.

The resulting data is sent by each experiment DAQ (Data AcQuisition) system, located in their respective pit, towards their dedicated disk EOS instance located in the CERN Datacenter - called WLCG [6] Tier-0. The table 1 shows the data transfer throughput towards EOS expected by each LHC experiment during Run-3.

Table 1: Expected data transfer throughput between experiments pits and EOS during Run 3

Experiment	Throughput
ALICE [7]	100 GB/s (up to 150 GB/s)
ATLAS [8]	8-20 GB/s (up to 12 GB/s)
CMS [9]	13-17 GB/s (up to 20 GB/s)
LHCb [10]	10 GB/s (up to 20 GB/s)

Once the data has successfully been written to EOS, it is sent to CTA - CERN Tape Archive - in order to store it on tape medias for long term archival. Finally, the data is also exported to external sites - called WLCG Tier 1 - for data storage and analysis.

The orchestration of the data archival and the data export to Tier 1s is ensured by the FTS [11] software (except for the ALICE experiment).

The figure 3 illustrates the general LHC experiments data taking workflow.

The next section will present the ALICE experiment setup which is quite different from the other LHC experiments one.

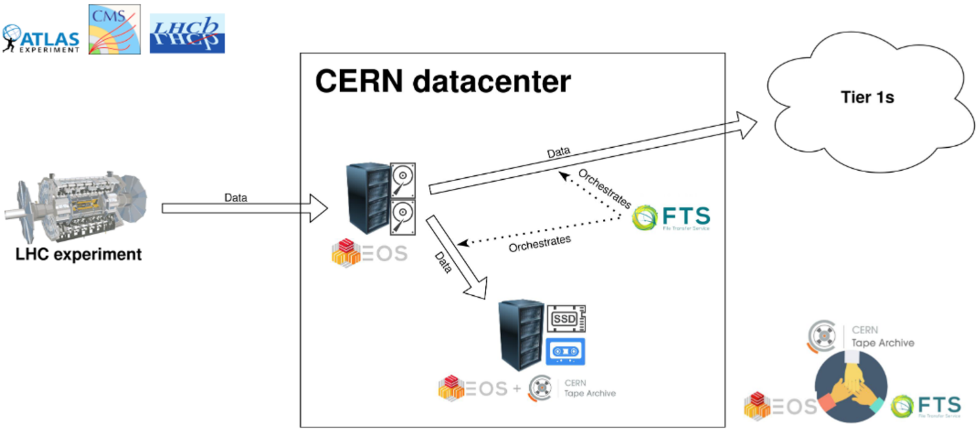


Figure 3: The General LHC experiments data taking workflows using EOS, CTA and FTS

5 The ALICE setup

During the LHC Run, the ALICE detector creates a continuous flow of 3.5TB/s of raw data that is sent to a filter, the First Level Processor (FLP). The filtered data is then sent to around 500 Event Processing Nodes (EPN) at 635GB/s. These EPNs are located in containers outside of the ALICE experiment main building and are used to filter again the data before sending it to the CERN datacenter for storage. This workflow is part of the ALICE Online/Offline [12] (ALICE O2) computing system.

The network between the ALICE detector and the EPNs is using InfiniBand [14]. Therefore, 4 redundant InfiniBand-to-Ethernet gateways have been installed in order to convert the traffic from InfiniBand to Ethernet before sending it to the data center. In total, the network infrastructure between the ALICE site and the CERN data center allows to reach up to 3.2Tbps of bandwidth.

Within one of the EPN containers, an EOS instance has been created to act as a backup in case of a network link failure happening between the ALICE experiment point and the CERN datacenter. This instance is called *EOSP2*. It can sustain 100GB/s of throughput and has 13.5 PB of raw storage allowing to have around 1.5 days of buffer in case of a major system or network failure.

In the CERN datacenter, the *EOSALICEO2* instance sustains between 100GB/s and 150GB/s and is used by ALICE to store the data (called Time Frames) coming from the EPNs. From this instance, the data is archived on tapes and exported to Tier 1 sites. Some fraction of this data is also copied to the *EOSALICE* instance that is used by the physicists to perform some data analysis.

The figure 4 shows the ALICE data workflow from the experiment detector to the different areas of storage.

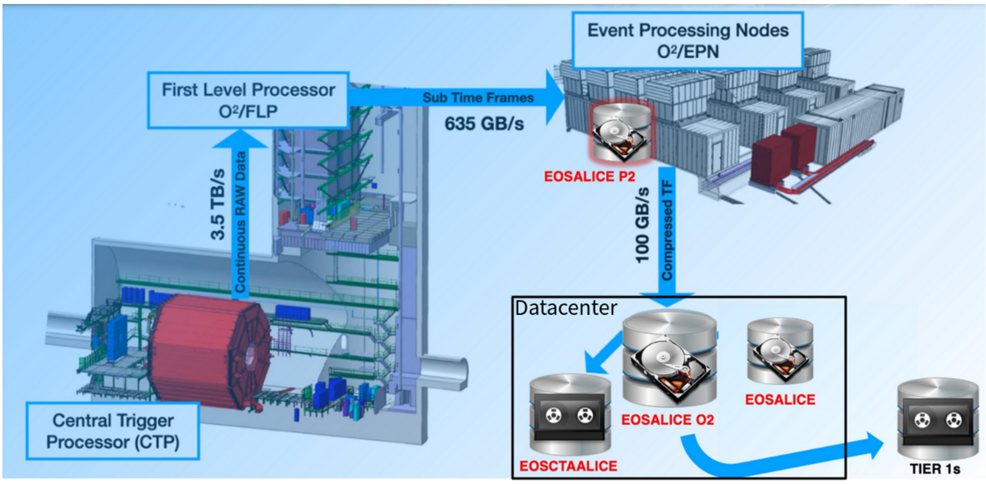


Figure 4: The ALICE experiment data taking workflow

As opposed to the other LHC experiments, ALICE does not use FTS to manage their data transfers. They instead use their own software called JAlien [13].

Despite the need of the ALICE experiment to export their data to the CERN datacenter at a throughput between 100GB/s and 150GB/s, benchmarks showed that the EOSALICEO2 instance can actually reach up to 350GB/s of data write throughput and 720GB/s of data read throughput. These performance are reached thanks to the high parallelism offered by the use of erasure coded storage layout that we present in the next section.

6 The deployment of erasure coding (EC)

Currently, the availability of the data stored in EOS is ensured by creating a second replica of each file written. Each of these replicas is located on different storage nodes ensuring that the data is still available even if one of the node is unavailable (due to maintenance or hardware issue). This storage layout is called RAIN (Redundant Array of Inexpensive Nodes).

Ensuring data availability with RAIN works very well, but is very costly in terms of storage capacity. Indeed, one needs twice the amount of raw storage that it is actually needed to store the data.

The deployment of erasure coding [15] layout allows to optimize the raw capacity needed to store the files and at the same time, provides high availability of the data stored on EOS. As opposed to replicating each file written to EOS, the data is split into chunks, each chunk is encoded and parity stripes are created. Finally, each piece is sent to different storage nodes (figure 5).

https://doi.org/10.1051/epjconf/202429501041

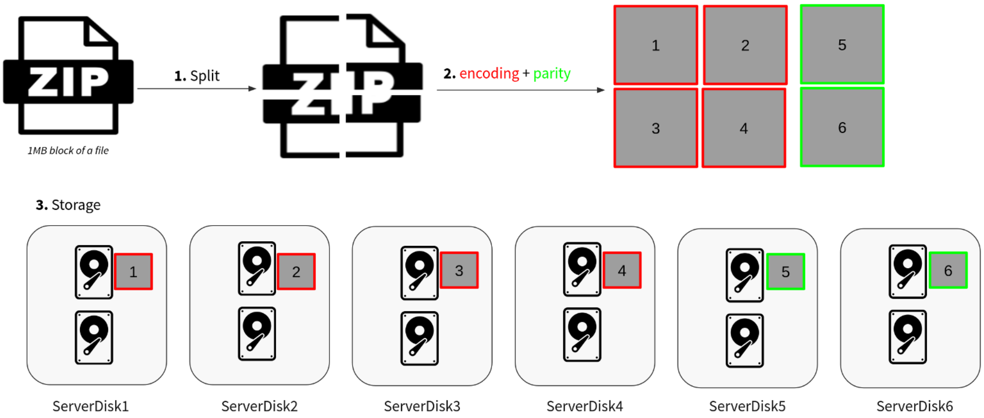


Figure 5: Erasure coding EC(4+2) 4 stripes + 2 parities allowing to have two disk servers unavailable and still have the data accessible

This way, the raw capacity needed to store the data can be optimized while preserving the file availability.

As an example, the table 2 shows the storage capacity needed to store a 1GB file depending on the storage layout chosen.

Table 2: The storage capacity needed to store a 1GB file depending on the layout

Storage layout	Capacity needed
RAIN(2)	2 GB
EC(4+2)	1.5 GB
EC(8+2)	1.25 GB
EC(10+2)	1.2 GB

At CERN, erasure coding is deployed on 3 instances. Among them, the EOSALICEO2 instance is fully erasure coded with EC(10+2) allowing to optimize 55.3 PB of storage space.

EC provides higher performance because of the parallelism provided by the several servers and disks involved. However, it is intensive in terms of CPU usage for striping and encoding, as well as network usage, as the various stripes must be sent to multiple storage nodes.

7 Operating EOS

Operating the EOS infrastructure at CERN encompasses various aspects and facets that require attention from the operations team.

First of all, every EOS server needs to be configured in order to deliver the service they are meant for. The entire configuration of the EOS clusters relies on CERN's puppet infrastructure [16]. All the needed configuration files are stored in a gitlab repository, which the operators can modify and easily deploy thanks to Puppet.

Another important aspect of the operation of EOS is to ensure the software is up to date with the latest version containing bug fixes, stability and performance improvements. Therefore, software upgrades are performed on a regular basis. The MGM and namespace nodes are the most visible upgrades as a service restart is required for the upgrade to be applied. As a consequence, the upgrade of the MGM is always planned and announced at least one week in advance in collaboration with the computing experts of each experiment. The disk servers upgrades, on the contrary, do not need any communication with the experiment and are performed one by one via an automated job running in Rundeck. As the experiments data is replicated twice or erasure coded, the upgrade of the storage nodes does not impact the files availability at all.

Operating 1300 disk servers for a total of 60000 disks means that hardware faults can occur at any time. With Network Interface Cards (NIC) failures, broken DIMM (Dual Inline Memory Module), CPU malfunctions and out-of-order disks, it happens very frequently that hardware intervention are needed. These hardware replacements are handled by a dedicated repair team dealing with CERN's datacenter hardware. Before doing any intervention, the repair team contacts the EOS service managers in order to ensure that the server having hardware failures can be taken out of production.

In order to ensure the EOS disk infrastructure is running properly and performs as expected by the experiment, the service managers need to monitor different metrics. For instance, the overall EOS cluster network ingestion and export rates, the data throughput between the experiments DAQ system and their associated EOS instance, the metadata operation rates, the memory consumption of each MGM and namespace servers are crucial data for the operators.

Therefore, different monitoring technologies have been put in place. The data is acquired thanks to many tools that EOS provides via its command line interface (CLI). The output is parsed (very often by using fluent-bit) and sent to monitoring back-ends like Graphite, InfluxDB, Prometheus or ElasticSearch. The monitoring visualisation is done using Grafana and Kibana.

Monitoring plays a crucial role in the day-to-day operation of EOS. It involves many different technologies that the operation team needs to take care of. As a consequence, the storage group of CERN is actively working on harmonizing the usage of these technologies in order to lower the burden of their maintenance.

8 Conclusion

The CERN disk storage infrastructure delivers the capacity and the throughput expected by the LHC and non-LHC experiments during Run 3. XRootD and FUSE are still the predominant file access interfaces, but the popularity of the HTTP protocol is increasing and will be more and more used by the physics community in a close future.

The LHC experiments data taking workflows involves the use of several softwares that are interacting with each other. EOS provides the disk storage, CTA the tape storage and FTS orchestrates the data transfers between EOS and CTA but also between EOS and external sites.

The ALICE experiment setup is quite different from the other LHC experiments. It has a dedicated EOS instance on their site that acts as a backup in the case there is a link failure between the ALICE site and the CERN data center. ALICE main EOS cluster (EOSALICEO2) is fully erasure coded which allows to optimize the space needed to store the data while ensuring its high availability and delivers very good read and write performance.

Operating EOS during LHC Run-3 involves many different aspects that the operators need to deal with. Different operational procedures have been put in place in order to ease the disposal of production servers for hardware repair. The EOS software evolves a lot over time and software upgrades need to be performed on a regular basis in order to get the latest improvements and bug fixes. Different monitoring technologies are used allowing operators to watch the EOS clusters to make sure the system is stable and that data is transferred at the expected rates.

In the future, it is expected that the data rates produced by the High-Luminosity LHC (LHC Run-4) will increase by an order of magnitude. The EOS developers and operators team are constantly improving the software and the operational procedures in order to face the new upcoming challenges and ensure that EOS will still continue to deliver fast and reliable storage for the physics communities.

References

[1] EOS - https://eos-web.web.cern.ch/eos-web/
[2] CERNBox - https://cernbox.cern.ch/
[3] CTA - https://cta.web.cern.ch/cta/
[4] scitokens - https://scitokens.org/
[5] QuarkDB - https://quarkdb.web.cern.ch/quarkdb/docs/master/
[6] WLCG - https://wlcg.web.cern.ch/
[7] ALICE - https://alice-collaboration.web.cern.ch/
[8] ATLAS - https://atlas.cern/
[9] CMS - https://cms.cern/
[10] LHCb - https://lhcb.web.cern.ch/
[11] FTS - https://fts.web.cern.ch/fts/
[12] Buncic, M. Krzewicki, P. Vande Vyvre, Tech. Rep. CERN-LHCC-2015-006. ALICE-TDR-019 (2015), https://cds.cern.ch/record/2011297
[13] JAliEN - https://alien.web.cern.ch/
[14] Daniele Pomponi, Transferring data from ALICE to the Computer Center - 2.4Tbps capacity with DWDM lines - HEPiX 2022 - https://indico.cern.ch/event/1200682/contributions/5087580/
[15] Andreas-Joachim Peters, Michal Kamil Simon, Elvin Alin Sindrilaru, Erasure Coding for production in the EOS Open Storage system - CHEP 2019 - https://doi.org/10.1051/epjconf/202024504008
[16] Steve Taylen, A Puppet Infrastructure at CERN - Puppet Camp - https://cds.cern.ch/record/2258306