

Disk storage at CERN

L Mascetti, E Cano, B Chan, X Espinal, A Fiorot, H González Labrador, J Iven, M Lamanna, G Lo Presti, JT Mościcki, AJ Peters, S Ponce, H Rousseau and D van der Ster

CERN European Organization for Nuclear Research, 1211 Geneva, CH

E-mail: luca.mascetti@cern.ch

Abstract. CERN IT DSS operates the main storage resources for data taking and physics analysis mainly via three system: AFS, CASTOR and EOS. The total usable space available on disk for users is about 100 PB (with relative ratios 1:20:120). EOS actively uses the two CERN Tier0 centres (Meyrin and Wigner) with 50:50 ratio. IT DSS also provide sizeable on-demand resources for IT services most notably OpenStack and NFS-based clients: this is provided by a Ceph infrastructure (3 PB) and few proprietary servers (NetApp). We will describe our operational experience and recent changes to these systems with special emphasis to the present usages for LHC data taking, the convergence to commodity hardware (nodes with 200-TB each with optional SSD) shared across all services. We also describe our experience in coupling commodity and home-grown solution (e.g. CERNBox integration in EOS, Ceph disk pools for AFS, CASTOR and NFS) and finally the future evolution of these systems for WLCG and beyond.

1. Introduction

CASTOR¹ and EOS² have been developed at CERN and implement a heavy-duty storage for the LHC Computing Grid to receive, store and preserve physics data from experiments and perform full-scale analysis. Our current solution for high-reliability, tape-backed storage is the "CERN Advanced STORage manager" (CASTOR) in production for about 15 years. EOS (in production since 2011) is the CERN disk-only storage, originally designed for large-scale LHC analysis. These two systems hold more than 100PB of physics data on disk and tape. In addition to the two main storage systems for physics data the IT Data & Storage Services (DSS) group operates also other storage systems: AFS, Ceph, NFS and CERNBox.

AFS is a globally-accessible network filesystem offering (Linux) users universal access to their home directories, workspaces and project areas. Close to 3 billion files are stored and accessed at a rate around 75kHz. CERN AFS raw disk installation size is approximatively 3 PB: users get space in RAID-1 volumes. User files add up to about 400 TB with around 30% yearly increase.

Ceph is the storage backend that provides images and volumes for the CERN IT OpenStack cloud (Cinder and Glance components). The service has deployed around 3 PB of raw disks, distributed across different areas of our data centre to overcome single-switch failures. The system currently stores 170 TB of data (77 million objects) with a replication factor of three[1].

¹ <https://castor.cern.ch>

² <https://eos.cern.ch>



CERN-IT has been providing NFS services for some applications via NetApp filers. These services are being migrated to a new NFS service built on top of our Ceph object store (detailed information will be presented in *Section 6.1*). Already 40TB of data and 120 million files have been migrated to the new infrastructure (30% of the NetApp service).

Finally, CERNBox, is the latest DSS storage service being provided to the CERN user community. It is a cloud storage service for file synchronisation based on the ownCloud open-source software stack³ [2].

2. CASTOR

CASTOR is a hierarchical storage management system for handling disk and tape layers. It holds 86 PB of data and 300 M files. Its main role is long-term data archive on the tape system. The current deployment consists of five different instances with dedicated headnodes and disk staging areas, one for each major LHC experiment (ALICE, ATLAS, CMS, LHCb) and one for all other user communities. An additional internal instance (REPACK) is used by the service to perform maintenance of the tape archive (shared across all instances)[3].

During LHC Run1, the CASTOR disk layout configuration was based on hardware RAID-1 controllers. For Run2 our disk servers have no hardware RAID controllers, whose failures were the main source of potential data unavailability and data losses[4]. We have now moved to a software RAID configuration, mainly using RAID-60, to have the best single stream performance required by our tape servers in migrating files to tape.

Before the LHC restart the CASTOR system was upgraded to the latest version (2.1.15) which introduces a new tapeserver daemon and xroot as the main internal protocol[5]. The original RFIO protocol is still available but it will eventually be removed in future major releases.

CASTOR 2.1.15 also introduces the possibility of creating disk pools on top of our Ceph infrastructure (data pools). This is currently under test and will minimise heavy hardware maintenance operations and provide more dependable performances.

During the preparation for the LHC Run2 the system was used mostly by non-LHC experiments, most notably by the Alpha Magnetic Spectrometer (AMS) which were generating substantial activity (data taking, (re)processing, etc...). At the same time the LHC experiments progressively moved their user activities to EOS completing the transition to use CASTOR as an archive system.

3. EOS

EOS is a disk storage system characterised by a low-latency hierarchical in-memory namespace. Its main role is to provide disk-only storage optimised for concurrent access[6]. EOS also offers a complete quota systems for users and groups with secure authentication and authorisation.

The current EOS deployment consists of six different instances. As for CASTOR we have an instance for each of the major LHC experiments plus a public instance for non-LHC experiments. In addition DSS recently deployed a new instance called *EOSUSER*, used as backend to store user data from CERNBox, the CERN cloud storage for file synchronisation and sharing. The total amount of raw disk space installed is around 140 PB, divided between the CERN data centres (Meyrin and the Wigner Data Centres). This translates to about 60 PB of available storage, including back-up files.

EOS is based on an Agile development cycle, new features are rolled out gradually during the service operation, reducing the risk of instability after an upgrade since the amount of changes introduced are relatively small and well tested. One of the latest functionality introduced by EOS are the location awareness and the GEO scheduling, these two features are fundamental

³ <https://owncloud.org>

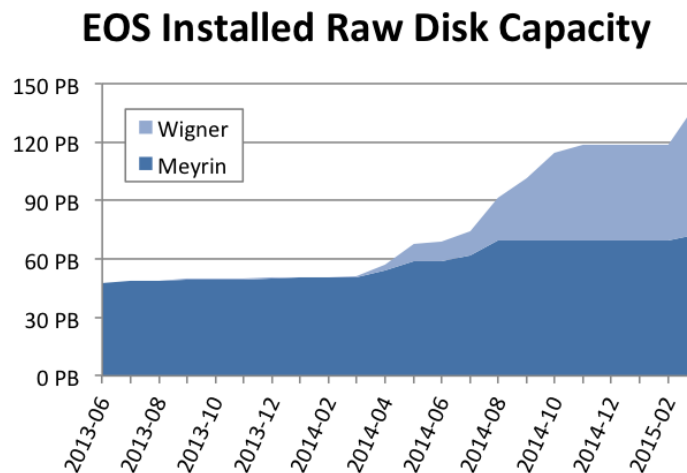


Figure 1. EOS raw capacity deployment for the two data centres.

for operating a system across two different data centres with about 22ms of round trip (Meyrin-Wigner)[7].

Another set of features recently introduced are the archival and backup tools[8]. The archival tool is very useful to delegate file movement between systems, in the CERN case between EOS and CASTOR (developed in close contact with CMS). This tool allows users or group movement of data to be standardised between our two storage systems, throttling requests according to our operational needs, hence avoiding direct access to the tape system for the end-users. The backup tool will be initially used in the context of CERNBox, in order to simplify the current backup procedure of users' files.

4. One year of EOS across two data centres

The CERN-IT datacenter at the Wigner institute in Budapest (Hungary) is officially in production for more than a year. In February 2014 the first storage space was made available inside our EOS production instances. EOS is now optimised for efficiently managing data in the different computer centres and providing a single site view to our user. Each diskserver is tagged with a “geocode” to identify its location. In the next EOS release (codename Citrine) a more accurate tagging mechanism will be available to also identify rooms, aisle, rack and other physical location characteristics inside the computer centre.

The current data placement policy distributes the (two) replicas of each files across the (two) sites (provided available resources are available at both ends). This allows our EOS instances to serve the closest replica to our clients, avoiding unnecessary remote access across the two data centres. Data placement was also enhanced to allow different types of scheduling and data placement as agreed with the users.

With the latest hardware delivery (March 2015) the capacity installed in the two computer centres reached the 50:50 ratio. File replicas are not yet spread equally between the two geolocation (*Figure 2*) due to the original resource asymmetry. Proactive geo-balancing is not yet activated for all EOS instances to avoid saturating the link from Geneva to Budapest.

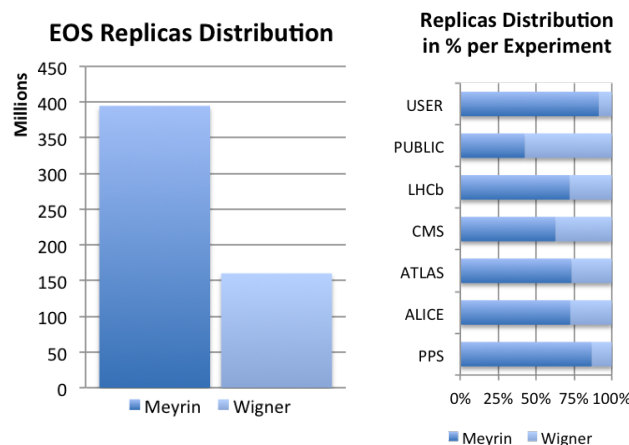


Figure 2. EOS file replicas distribution per experiment across the two sites.

5. Storage usage for LHC data taking

In the last two years the usage of CASTOR and EOS services was modified to better control the data flow and access. Our storage systems evolved to separate tape and disk endpoints. EOS is the disk endpoint for high concurrency and low latency access while providing the required data reliability. CASTOR is the tape endpoint for long-term data storage oriented in particular to archive data. Data stored in CASTOR is kept for preservation purpose and to offload the disk endpoint.

As foreseen in our previous CHEP paper[4] two main scenarios have been put in place for LHC Run2:

Hot and Cold Storage ATLAS and CMS apply this scenario for their data flow. The data generated from their DAQ is sent directly to EOS, where the processing takes place. The RAW data to be archived are transferred from EOS to CASTOR and to Tier-1s. The data movement from one storage system to the other is controlled by the experiments' data management tools. Data processing and user analysis is only scheduled on EOS. In this scenario CASTOR behaves like a "Cold Storage": after the transfer the data is only accessed in a controlled way by the experiments' data manager experts (prestaged from CASTOR to EOS).

Disk/Tape Separation ALICE and LHCb preferred to use a different scenario. The data recording flow for Run2 will send files directly to CASTOR for long-term archiving and Tier-1 export. The (re)processing will require a copy from CASTOR to EOS, performed by the experiments' data managers.

6. A common hardware foundation for all data services

Nowadays most of the disk capacity is absorbed by the EOS service: during the last 24 months (Long Shutdown 1) its usable capacity went up by about 2 PB per month. This has also been made possible by the availability of efficient hardware configurations prepared by the CERN procurement team (IT-CF). The most recent batches of disk capacity are composed by disk servers of about 200 TB each, with 2x24x4TB disks layout and 64 GB RAM with a 10-Gbps NIC. In the next 12 months we expect to see these systems evolving to support 6TB disks and possibly more disks (from 2 to 3 24-bay disk trays) and reach about 400 TB per server.

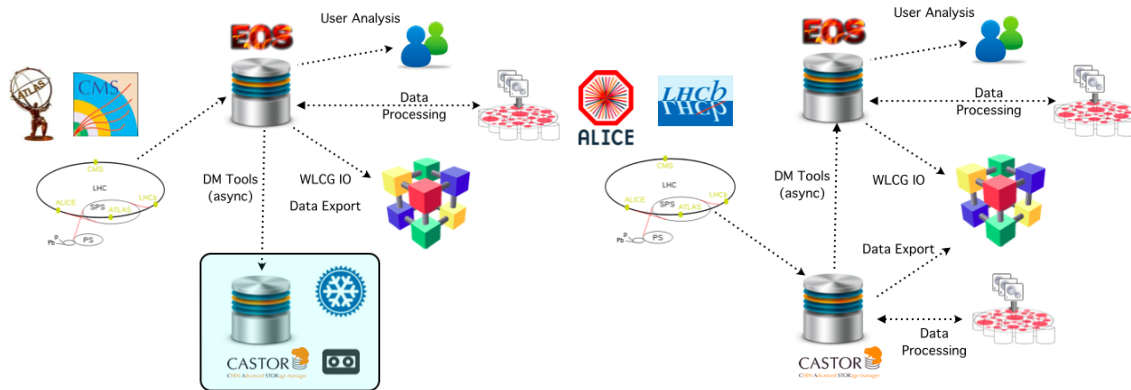


Figure 3. Data taking flow for LHC Run2.

This diskserver configuration is optimised from a cost-of-ownership point of view for EOS and CASTOR, while it can be suboptimal for other services (most notable AFS and NFS) because of a too-large ratio of the disk capacity over network capacity.

The Ceph infrastructure uses similar configuration and the needed optimisation, for the block storage uses cases, requires increasing the performance of the servers with a small number of SSDs to host the journaling files as demonstrated by our team[1]. On the other hand, for supporting other use cases (like streaming files/objects for CASTOR) the CERN standard hardware can be used.

The CERNBox service, which uses EOS as the storage backend, is also built on the CERN standard hardware.

6.1. NFS service and AFS prototype server

CERN IT manages a series of NetApp appliances for NFS applications. The IT DSS group is consolidating the NFS services with the goal of avoiding proprietary solutions. Given the size of our storage servers, to efficiently use the disk capacity and have effective operations, we developed a different approach.

The strategy for decommissioning the current NFS storage is to provide an NFS-based service able to show the same performance as the previous one with similar or better availability and reliability. A solution based on ZFS and Ceph was prototyped and presented to our users. This

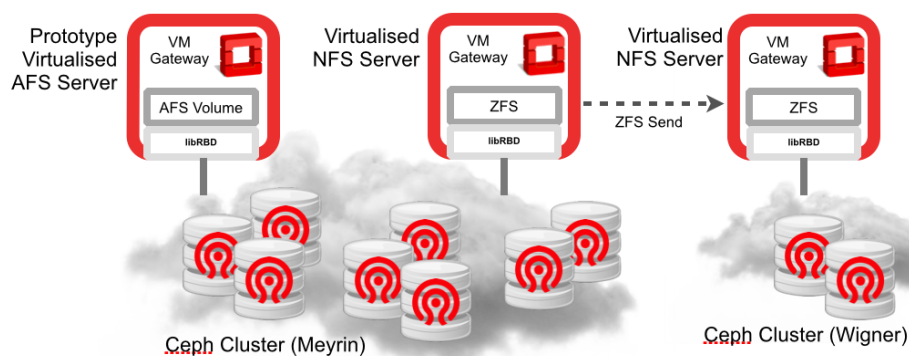


Figure 4. NFS service architecture and AFS virtualised server.

new NFS service is based on a virtual machine that exports a Ceph block storage (using RBD block devices) formatted with ZFS. The virtual machine needs to be correctly sized with the amount of RAM in order to have a good cache-hit ratio to sustain heavy read operations. On the other hand we can effectively provision each NFS server with the appropriate-size ZFS volumes from the Ceph pool, hence efficiently using the physical disk space.

One of the strong points for preferring ZFS is the possibility to easily provide snapshots and backups. With ZFS these operations are very advanced and give users the possibility to easily roll back to previous snapshots. Snapshots are also essential for providing disk and tape backups and these options will be offered to users in the near future. The disk backup will be done off-site (using the Wigner Ceph installation - all primary disks are located in the Meyrin installation).

A similar approach was tested for AFS servers. The idea behind is the virtualisation of the AFS servers, having a virtual machine that uses a Ceph RBD volume instead of local physical storage. The advantage of this approach is the reliability of the Ceph volume compared to the local storage (three replicas on different nodes vs. a RAID-1 on two local disks) and the possibility to have more, smaller AFS servers. As in the NFS case, smaller servers are less prone to be overloaded by concurrent activities and in general more memory can be guaranteed in an exclusive way to each applications.

6.2. CASTOR data pool

CASTOR 2.1.15 introduces the possibility to create data pools on top of our Ceph infrastructure. This is currently under test and will drastically improve the operations of the CASTOR disk layer.

A data pool is a storage pool built on top of the Ceph common disk infrastructure: it gives us full flexibility in adjusting the resources across different activities to best use resources to improve the total-cost-of-ownership. In addition, operations-intensive activities (e.g. balancing data across physical services when hardware is replaced) will be radically simplified since these will be delegated to the Ceph storage backend.

This innovative approach is based on the CERN contribution to the Ceph software (libradosstriper⁴, now part of the standard Ceph distribution starting with the Giant release).

This library allows files to be disassembled in parallel stripes of objects which are then written in the Ceph infrastructure. Conversely, objects are read by the client which reassembles

⁴ <https://github.com/ceph/ceph/tree/giant/src/libradosstriper>

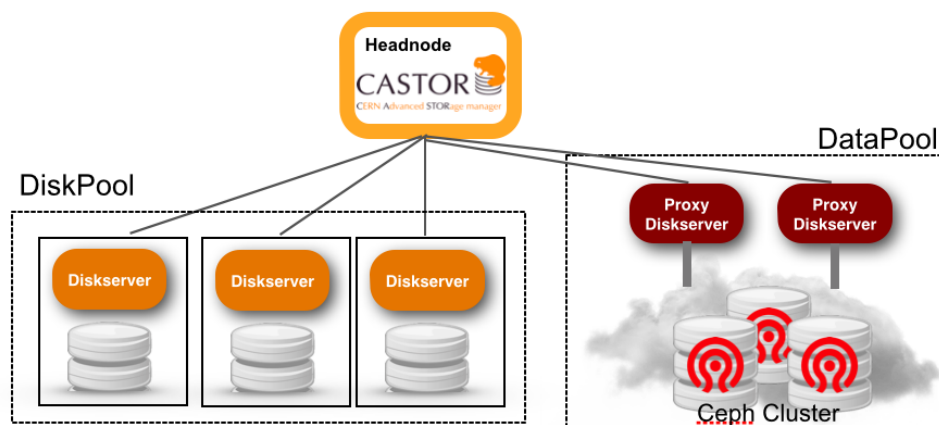


Figure 5. CASTOR architecture for DiskPool and DataPool.

the complete file (self-described by the objects themselves). A single file transfer can benefit from the throughput of several Ceph disks as objects sent/received in parallel go through several network interfaces on the server side. During the CASTOR release process (stress testing) we observed single-stream performances of the order of 350 MB/s.

In this new schema the disk servers are mere proxy servers providing the translation between the client code and the Ceph-aware striping process: the performance of the whole system can be adjusted by varying the number of proxy servers. On the other hand, the CASTOR tape servers access the data natively (via Ceph libradosstriper), hence receiving the fragments of the required file from several Ceph disk servers at the same time. This removes the single disk limitation which can prevent efficient usage of the tape system. Despite the streaming performances of spinning disks (~ 100 MB/s) not increasing in the last years, tape servers can already accept (and read) data at about 300 MB/s and are expected to increase substantially in the next few years. The fact that each file is split into small objects greatly helps in avoiding hot-spots (concurrent clients accessing the same disk at the same time for long periods of time) making the overall system more resilient to overload conditions under heavy usage.

7. Summary and outlook

During the Long Shutdown 1 our storage systems were all fully operational and several enhancements were developed and rolled-out in production in preparation for the LHC Run2. The hardware uniformisation has been pushed further with operational improvements at scale. Globally we are committed to continue evolving the CERN and WLCG data services while supporting experiments and laboratory needs during LHC Run2 and beyond.

We highlight that EOS now provides storage capabilities across distributed computer centres. EOS is a key service in the IT DSS strategy to further innovate the end-user services. The synergy between EOS and the cloud synchronisation and sharing service of CERNBox will shape data access and data analysis in the next years.

On the infrastructure side, Ceph provides the common backbone for a variety of services for the CERN Computer centre data infrastructure. We have built a convincing “virtualisation” layer to efficiently provide a broad spectrum of data services. This ranges from the support of OpenStack to virtualised AFS and NFS services to a flexible, robust and performing disk layer for CASTOR.

We are convinced that innovative developments will allow us to cope with the challenges of LHC Run2 and provide interesting opportunities for other colleagues dealing with data-intensive applications.

References

- [1] D van der Ster et al. “Ceph-based storage services for Run2 and beyond.” CHEP 2015 - these proceedings.
- [2] L Mascetti et al. “CERNBox + EOS: end-user storage for science.” CHEP 2015 - these proceedings.
- [3] DF Kruse “The Repack Challenge.” *Journal of Physics: Conference Series*. Vol. **513**. No. 4. IOP Publishing, 2014.
- [4] X Espinal et al. “Disk storage at CERN: Handling LHC data and beyond.” *Journal of Physics: Conference Series*. Vol. **513**. No. 4. IOP Publishing, 2014.
- [5] E Cano et al. “Experiences and challenges running CERN’s high-capacity tape archive.” CHEP 2015 - these proceedings.
- [6] AJ Peters and J Lukasz “Exabyte Scale Storage at CERN.” *Journal of Physics: Conference Series*. Vol. **331**. No. 5. IOP Publishing, 2011.
- [7] J Iven et al. “di-EOS - “distributed EOS”: Initial experience with split-site persistency in a production service.” *Journal of Physics: Conference Series*. Vol. **513**. No. 4. IOP Publishing, 2014.
- [8] EA Sindrilaru et al. “Archiving tools for EOS.” CHEP 2015 - these proceedings.