EPJ Web of Conferences **245**, 04038 (2020) https://doi.org/10.1051/epjconf/202024504038
CHEP 2019

CERN Disk Storage Services

Report from last data taking, evolution and future outlook towards Exabyte-scale storage

Luca Mascetti[1*], *Maria* Arsuaga Rios[1], *Enrico* Bocchi[1], *Joao* Calado Vicente[1], *Belinda* Chan Kwok Cheong[1], *Diogo* Castro[1], *Julien* Collet[1], *Cristian* Contescu[1], *Hugo* Gonzalez Labrador[1], *Jan* Iven[1], *Massimo* Lamanna[1], *Giuseppe* Lo Presti[1], *Theofilos* Mouratidis[1], *Jakub T.* Mościcki[1], *Paul* Musset[1], *Remy* Pelletier[1], *Roberto* Valverde Cameselle[1], and *Daniel* Van Der Ster[1]

[1]CERN, Esplanade des Particules 1, 1211 Geneva 23, Switzerland

Abstract. The CERN IT Storage group operates multiple distributed storage systems to support all CERN data storage requirements: the physics data generated by LHC and non-LHC experiments; object and file storage for infrastructure services; block storage for the CERN cloud system; filesystems for general use and specialized HPC clusters; content distribution filesystem for software distribution and condition databases; and sync&share cloud storage for end-user files. The total integrated capacity of these systems exceeds 0.6 Exabyte.

Large-scale experiment data taking has been supported by EOS and CASTOR for the last 10+ years. Particular highlights for 2018 include the special Heavy-Ion run which was the last part of the LHC Run2 Programme: the IT storage systems sustained over 10GB/s to flawlessly collect and archive more than 13 PB of data in a single month. While the tape archival continues to be handled by CASTOR, the effort to migrate the current experiment workflows to the new CERN Tape Archive system (CTA) is underway.

Ceph infrastructure has operated for more than 5 years to provide block storage to CERN IT private OpenStack cloud, a shared filesystem (CephFS) to HPC clusters and NFS storage to replace commercial Filers. S3 service was introduced in 2018, following increased user requirements for S3-compatible object storage from physics experiments and IT use-cases.

Since its introduction in 2014N, CERNBox has become a ubiquitous cloud storage interface for all CERN user groups: physicists, engineers and administration. CERNBox provides easy access to multi-petabyte data stores from a multitude of mobile and desktop devices and all mainstream, modern operating systems (Linux, Windows, macOS, Android, iOS). CERNBox provides synchronized storage for end-user's devices as well as easy sharing for individual users and e-groups. CERNBox has also become a storage platform to host online applications to process the data such as SWAN (Service for Web-based Analysis) as well as file editors such as Collabora Online, Only Office, Draw.IO and more. An increasing number of online applications in the Windows infrastructure uses CIFS/SMB access to CERNBox files.

CVMFS provides software repositories for all experiments across the WLCG infrastructure and has recently been optimized to efficiently handle nightly-builds. While AFS continues to provide general-purpose filesystem for internal

*e-mail: luca.mascetti@cern.ch

CERN users, especially as $HOME login area on central computing infrastructure, the migration of project and web spaces has significantly advanced.
In this paper, we report on the experiences from the last year of LHC RUN2 data taking and evolution of our services in the past year.. We will highlight upcoming changes and future improvements and challenges.

1 Introduction

The CERN IT Storage group at CERN (IT-ST) operates and develop multiple distributed storage systems to support all CERN data storage requirements. Each storage solution is specialised and optimised for a particular use cases, spanning from the storage of physics data generated by LHC and non-LHC experiments, to object and block storage for IT infrastructure services in OpenStack, the CERN cloud system. The total integrated capacity of all the systems exceeds 0.6 Exabytes of storage.

2 AFS migration status

AFS continues to provide a general-purpose filesystem for CERN users, and even today AFS is still the home directory of choice for the central CERN Linux systems. The decades-old architecture has held up surprisingly well, and the code is impressively stable. It is however best suited to modest amounts of data (AFS volume size is limited by the protocols used to transfer these between machines), and modern disk servers are too large for it (space-wise, the entire CERN AFS cell with 640TB would fit into a single host - to keep redundancy, the service is spread over virtual machines and a number of ancient physical machines).

CERN has tried to phase out AFS since 2016[5], both to address concerns about long-term availability and in order to unify the personal storage areas on EOS/CERNBox. This required several changes on the EOS side (including a more scalable namespace and a more robust filesystem), and overall the AFS phaseout has progressed more slowly than hoped for. Nevertheless in the past years, much of the legacy project and experiment content of CERN AFS has been either archived to tape, or migrated to EOS. Also external dependencies into the CERN AFS service have been identified and removed. At the same time, our worries about the upstream OpenAFS project's longevity are reduced, so there is now less pressure to move the core AFS data elsewhere.

However, several challenges lie ahead for the CERN AFS service: it has to adapt to the various CERN IT infrastructure changes (a new Kerberos realm, new resource management, GDPR/OC11 compliance), and most of the remaining legacy hardware will have to be replaced in 2020.

3 Ceph Block Storage, CephFS, and S3

Ceph storage in CERN IT is primarily used for OpenStack. The block storage service allows our cloud users to attach large volumes for bulk, reliable storage; more than 6000 virtual machines are doing this actively, totaling more than 3.5 petabytes of raw data (counting a factor 3 replication).

Since the early days of Ceph, we have built virtual NFS filer appliances using a stack of RADOS block devices, ZFS, and nfsd; this was used extensively as a low-cost NFS solution, and saw major uptake backing our IT infrastructure needs (e.g. configuration management,

documentation, engineering services, etc.) But while our users' needs were evolving to demand higher performance and availability, CephFS matured into a production-quality network filesystem, eventually ticking more boxes than our virtual NFS solution. During 2019, we migrated the majority of our NFS users to CephFS, whilst also taking on board a large number of new POSIX use-cases for container, HPC, and other workloads. We continue to develop the CephFS-based service, having developed a new utility to backup to S3, and now focusing on extensively testing the snapshots feature before enabling it for the general usage.

Through the use of rados gateways, Ceph also provides S3 object storage. The S3 service is configured as a single region cluster, where data is stored with a `(4,2)` erasure encoding schema and bucket indices are replicated 3 times. S3 accounts for 2 petabytes and is integrated with OpenStack Keystone [1] for improved usability for cloud users. In early 2019, the S3 cluster received an hardware upgrade that allowed the storage of bucket indices on SSDs. This increased the metadata performance massively, making the cluster able to sustain rates of 83 kHz, 63 kHz, and 198 kHz for PUT, HEAD, and DELETE requests, respectively.

4 Software Distribution with CVMFS

CVMFS is the storage service for reliable software distribution on a global scale. It provides read-only access to clients by mounting a `POSIX` filesystem at `/cvmfs`. The transport layer makes use of the HTTP protocol, allowing for the use of off-the-shelf web server and web caching software. The primary CVMFS use case is software distribution for the LHC infrastructure, which counts more than 100,000 clients, 400 web caches, and 5 replica servers.

During 2019, CVMFS started to make use of S3 object storage for its repositories. Currently, out of 54 repositories managed at CERN, 11 are S3 native, while other 13 repositories have been migrated from Cinder volumes (hosted on Ceph block devices) to S3. The use of S3 as authoritative storage also allowed the deployment of the CVMFS Gateway. This new component regulates write access to the repository by providing leases that have limited validity in time and restrict the write access to a specific subpath of the repository only. This would allow for parallel publication to the same repository from different release managers (provided they ask for write access to different subpaths), ultimately allowing to reduce the time needed to have a given workload published on CVMFS.

CVMFS is also expanding its field of application to the distribution of Docker container images. In this context, CVMFS allows to store Docker image layers in their uncompressed form and let the Docker daemon fetch the files that are needed at run-time directly from CVMFS. Such a feature allows saving on local disk space, which is typically low on mass processing systems like HTCondor, as it is no longer needed to pull and decompress the full image. This is possible through the development of two specialized components: the *CVMFS GraphDriver*, which installs client-side and allows the Docker daemon to use image layers from CVMFS; and the *CVMFS DUCC*, which converts ordinary Docker images into uncompressed layers and publish them on a CVMFS repository.

5 CERNBox: CERN Cloud Collaboration Hub

CERNBox[10–12] is the ubiquitous CERN cloud storage hub and collaboration platform. The service allows the whole CERN community to synchronize and share files on all major desktop and mobile platforms (Linux, Windows, MacOSX, Android, iOS) providing universal access and offline availability to any data stored in the CERN EOS infrastructure (currently for users and projects, 7.5 petabytes and more than 1 billion files). With more than 25,000 users registered in the system, CERNBox has responded to the high demand in our diverse

community to an easily and accessible cloud storage solution that also provides integration with other CERN services for big science: visualization tools, interactive data analysis and real-time collaborative editing. The service is used daily by more than 6,000 unique users, provides more than 500 project spaces for group collaboration and has more than 140,000 shares spanning all CERN departments and groups, an important highlight of the collaborative aspect of the platform. In addition, CERNBox has also been declared apt to deal with personal data handling and confidential data (thanks to the anonymous upload-only public link feature[9]), depicting the added value the service brings to the organization for such heterogeneous use-cases.

During 2019 the CERNBox infrastructure was upgraded (SSDs caches, multiple 10gbps network cards, automatic load balancing and better KPIs/SLIs) resulting in a factor two increase in the availability of the service. A feature flag mechanism was also introduced (enable Canary Mode), to allow end-users to test latest features before they land in production and provide valuable feedback. For example, the OnlyOffice document suite was successfully introduced with this mechanism to explore the feasibility of replacing use-cases from the commercial Microsoft Office365 platform in the context of the MALT project[17] to reduce the dependency in commercial software and move to open source solutions. Secondly, a new backup system for the service was prototyped based on tool called Restic[8], bringing to the service long sought improvements: backing up the data into a different technology (from EOS to a S3 cluster based on Ceph), data de-duplication to minimise costs and the ability to have point-in-time data restores (individual files and full snap-shoots). Finally, a new back-end micro-services architecture[13, 14] was introduced based on Reva[7, 16] to manage the evolution of the service and support the its various use-cases.

In the context of MALT, CERNBox plays a significant role to expose alternative open source solutions to the diverse and ever growing community of CERNBox users. These are some examples of integrations available in CERNBox: the DrawIO platform is used for diagrams and organigrams sketching as alternative to Microsoft Visio, OnlyOffice and Collabora Online for documents editing as an alternative for Microsoft Office365 and DXHTML Gantt for project management as an alternative to Microsoft Project. Furthermore, CERNBox was also the chosen solution to migrate the home folder of more than 37,000 user accounts from the Windows-centric DFS file system. As for March 2020, 30% of the accounts have been successfully migrated[15].

In 2020 the aim is at drastically improve the usability of the service for the end-users in major areas like sharing and ACLs handling. In addition to that, more applications integration will be introduced into the platform alongside the opening of development of applications to other teams thanks to the recently introduced architecture. There is also the plan to fully phase-in the new backup system and deprecate the previous one.

6 Tape Storage

Tape storage for Physics data has traditionally been provided by CASTOR [2], which in 2018 saw a significant load as the LHC Run 2 ended with the Heavy Ion programme covered below. Currently, the system stores 340 PB of total data, with minimal increase in 2019 as the major LHC experiments did not take any raw data.

The new CERN Tape Archive software, CTA, is expected to replace CASTOR during 2020 and 2021 for the LHC experiments and take over the whole Physics tape storage at CERN [3].

7 EOS large-scale disk storage system

At CERN we currently operate several EOS instances amounting for approximately 340 PB of raw disk capacity and storing almost 6 billion files. 2019 was a year of many changes in the EOS service at CERN. This included, but was not limited to, deploying the new namespace technology, QuarkDB[20][21], across all EOS instances, evaluating a new replica layout system and, finally, converting most of the ALICE experiment's data acquisition files from a 2-replica layout to an erasure encoded layout across storage nodes with N+2 redundancy (RAIN 8+2).

The deployment of the new namespace backend system, together with other software improvements has particularly improved the overall service availability across all EOS instances, reducing to a minimum the restart time of the namespace, which was the main cause of service unavailability.

The main reason behind the activities regarding layout change is the possibility to lead to a reduction of the storage cost, while allowing a substantial capacity increase, moreover the disk capacity will be more efficiently used and spread across storage servers.

Strictly on the operational side, one of the big achievement in 2019 was the retirement of approximately 90 PB of storage and the repatriation of approximately 50 PB of disks from the now decommissioned data center in Wigner, Hungary. This activity was scheduled in five stages across the entire year and it was successfully concluded beginning of December.

In addition the operations team completed the migration of all end-users account in the EOS instance used as a backend for CERNBox (named EOSUSER) to a new re-designed EOS system (EOSHOMEs). The new EOS backend for this activity is now composed of 8 separate instances (or shard), where account are redistributed. This new architecture allows a better handling of the metadata load coming from the users and a partition and reduction of the failures domains.

2019 saw as well the phase-in in production of the latest generation EOS fuse client, `FUSEx`[22]. The stability, memory footprint and performance of this client has been substantially improved by our developers. This mount interface is heavily used by many CERN users from desktops, lxplus and lxbatch nodes. For example the IO traffic on the EOSHOMEs instances has surpassed the traffic seen on the complete AFS service during 2019, reaching daily read peaks of half a petabyte.

On the development side there have been several improvements as well, many of them taking more and more advantage of the new namespace. This also lead to the creation of an external tool to detect files possibly at risk and fix them before they get into an unrecoverable state. This system has started to be ported to EOS natively (EOS FSCK), hoping to have a working EOS-native component ready by Q2 2020.

8 Heavy Ion Run 2018

Throughout the LHC Run2, the data flow of the ATLAS and CMS experiments has seen their EOS instances as the receiving hubs for the Data Acquisition (DAQ) systems. Data are then distributed to batch for the reconstruction, via FTS to Tier1s for replication, and to CASTOR for tape archival. These accesses are essentially controlled by the experiments themselves. In 2018, ALICE modified their data flows along those lines as well.

For the Heavy-Ion run, the LHC was expected to operate with a duty cycle of 50% to 60%, this value is directly proportional to the overall data volume to store in our systems. ALICE announced a target throughput of 8-9 GB/s, whereas ATLAS and CMS targeted up to 4 GB/s. LHCb did not change their target at about 1 GB/s in anti-coincidence with the other experiments, as data are preferentially transferred off-collision. The announced throughput

Figure 1. Throughput from the Experiments' DAQ systems to EOS during one of the LHC Heavy-Ion fills.

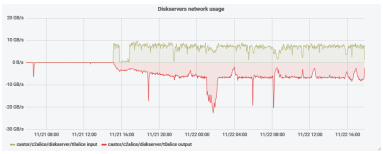

Figure 2. ALICE throughput for CASTOR during one fill.

was quickly met as apparent from fig.1, where the alternating on-off pattern in the network bandwidth corresponds to the behavior of the Experiments' DAQ systems. When particles are colliding in the experiments' Pit, data is generated, processed and transferred from the acquisition systems to the EOS storage infrastructure.

After ALICE started to copy data to tape, the system has reached a steady state whereby the observed influx rates have been 2.5-3.5 GB/s for ATLAS, 4-6 GB/s for CMS, and 8-9 GB/s for ALICE. The LHC duty cycle has been lower than expected at about 40%, which allowed to migrate to tape all data without any significant backlog. In particular, a typical ALICE session in CASTOR is shown in fig. 2, where the input rate is dominated by the influx traffic from the EOS buffer (EOSALICEDAQ) and the output rate is the superposition of tape migrations at 6.6 GB/s (where 20 tape drives write data at their full speed) and occasional user accesses that create peaks of over 20 GB/s.

Following the last LHC beam dump, all experiments kept pushing their data out for several hours. During this period, CASTOR accumulated the largest backlog of data to be migrated to tape (about 100 TB for ALICE alone), and ALICE kept recording data to tape for two more weeks, as most of the data was still only stored on their EOSALICEDAQ disk buffer. Overall, the entire pipeline including the Experiments' DAQ systems, EOS, and CASTOR, has performed remarkably smoothly and has allowed to achieve the largest data rates to date, with 13.8 PB recorded to tape in one month. A full detailed report of the exercise is available at [4].

9 Outlook

During 2019 all our storage systems were fully operational and several enhancements were developed and rolled-out in production. The Heavy-Ion Run was a success from the operations point of view of the two major storage system for the physics use-case (CASTOR and EOS). In 2020 we are duly committed to continue evolving the CERN storage infrastructure to support WLCG data services, experiments and the laboratory needs for the next LHC Run, which is expected to scale up to 10 times the data rates of Run2.

We highlight the EOS unprecedented storage capabilities and scalability in addition to the ability to seamlessly discontinue a data centre while running normal operations. EOS is at the heart of the CERN IT strategy to further innovate the end-user services and as well the future tape access with the CTA project. The synergy between EOS and other projects will shape the data access and analysis for the physics community.

On the infrastructure side, Ceph continues to successfully provide the common backbone for a variety of services for the CERN Computer centre data infrastructure.

We are convinced that innovative developments will allow us to cope with the challenges of LHC Run3 and provide interesting opportunities for other communities dealing with data-intensive applications.

References

[1] J. Castro Leon, *RadosGW Keystone Sync*, CERN TechBlog, https://techblog.web.cern.ch/techblog/post/radosgw_sync_ec2_keys/

[2] CERN Advanced STORage manager, http://cern.ch/castor

[3] E. Cano et al., *CERN Tape Archive: production status, migration from CASTOR and new features*, CHEP 2019 Proceedings.

[4] C. Contescu, G. Lo Presti, H. Rousseau, *Data Taking for the Heavy-Ion Run 2018*, CERN IT Note 2019-006

[5] J Iven, M Lamanna, A Pace; *CERN's AFS replacement project*; J. Phys.: Conf. Ser. **898** 062040 (2017)

[6] ownCloud, https://owncloud.com (access time: 10/03/2020)

[7] Reva, https://reva.link (access time: 10/03/2020)

[8] Reva, https://restic.net/ (access time: 10/03/2020)

[9] https://home.cern/news/news/computing/computer-security-file-drop-confidential-data-0 (access time: 10/03/2019)

[10] Mascetti, Luca and Labrador, H Gonzalez and Lamanna, M and Mościcki, JT and Peters, AJ, Journal of Physics: Conference Series **664**, 062037, *CERNBox + EOS: end-user storage for science* (2015)

[11] H. G. Labrador, *CERNBox: Petabyte-Scale Cloud Synchronisation and Sharing Platform* (University of Vigo, Ourense, 2015) EI15/16-02

[12] H. G. Labrador et all, EPJ Web of Conferences **214**, 04038, *CERNBox: the CERN storage hub* (2019)

[13] Thönes, Johannes, IEEE software 32.1 *Microservices*, (2015)

[14] Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L, Springer, Cham, *Microservices: yesterday, today, and tomorrow. In Present and ulterior software engineering* (2017)

[15] Hugo G. Labrador et al., *Evolution of the CERNBox platform to support the MALT project*, Proceedings of 24th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2019)

[16] Hugo G. Labrador et al., *Increasing interoperability: CS3APIS*, Proceedings of 24th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2019)

[17] Maria Alandes Pradillo et al., *Experience Finding MS Project Alternatives at CERN*, Proceedings of 24th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2019)

[18] Peters, AJ and Sindrilaru, EA and Adde, G, Journal of Physics: Conference Series **664**, 062037 (2015)

[19] Peters, Andreas J and Janyst, Lukasz Journal of Physics: Conference Series **331**, 052015 (2011).

[20] Peters, Andreas J., Elvin A. Sindrilaru, and Georgios Bitzes. "Scaling the EOS namespace." International Conference on High Performance Computing. Springer, Cham, 2017.

[21] Bitzes, Georgios, Elvin Alin Sindrilaru, and Andreas Joachim Peters. "Scaling the EOS namespace–new developments, and performance optimizations." EPJ Web of Conferences. Vol. 214. EDP Sciences, 2019.

[22] Peters, Andreas J et al. "Evolution of the filesystem interface of the EOS Open Storage system", CHEP 2019 Proceedings.