Письма в ЭЧАЯ. 2016. Т. 13, № 5(203). С. 1051–1057



# LARGE-SCALE DATA SERVICES FOR SCIENCE: PRESENT AND FUTURE CHALLENGES

*M. Lamanna* [1]

CERN, IT Department — Data Storage Services (IT-DSS), Geneva

Some directions for evolving our data management services in the next years are discussed, using the experience in operating heavy-duty data storage services at CERN, notably for the Worldwide LHC Computing Grid (WLCG). These new developments are potentially useful beyond our community, wherever the success of a project depends on large computing resources and requires the active participation of large and distributed collaborations.

PACS: 07.05.Bx; 07.05.Kf; 07.05.Rm

## INTRODUCTION

The WLCG infrastructure has been one of the success factors of the first phase of LHC (Run 1: 2010–2013) culminating with the Higgs boson discovery. The LHC computing infrastructure has been evolved during the 2013–2015 LHC long shutdown (LS1) in preparation for the second phase (Run 2: 2015–2018) with increased energy and luminosity.

With Run 2 entering a steady operational phase, the computing infrastructure is being analyzed to set up the strategic directions for the next ten years. This is the natural horizon being the continuation of the LHC programme before the upgrade to the high-luminosity mode foreseen for 2025. This talk is a contribution to these discussions from a data-management point-of-view based on recent evolution in the services offered by the CERN IT-DSS group to the CERN laboratory and to the WLCG.

## DATA SERVICES AT CERN FOR LHC

The CERN IT-DSS group is responsible for the main Tier-0 dataflows of the LHC after the event selection. Our primary responsibility is to support the experiments in close connection with the data taking.

All the experiments transfer their data to the computing centre as files (typically in the 1–10 GB file size range) via dedicated links. The receiving end is the CASTOR [1] storage system (for the ALICE and LHCb experiments) or the EOS [1] storage system (for ATLAS and CMS). During data taking, the typical input rates from the experiment data acquisition

[1] E-mail: massimo.lamanna@cern.ch

and filter farms to the CERN computing centre are in the range of several GB/s. The highest sustained rate has been observed during October 2015: about 7 PB/month of data on tape to be compared with 4 PB/month at the end of Run 1. The majority of the data is integrated during the proton–proton programme (several months per year), which is presently dominated by ATLAS and CMS with sustained rates in the 4 GB/s range; similar values, but dominated by ALICE, are observed during the Ion–Ion periods (4 weeks per year).

The CERN disk infrastructure allows the experiments to implement these dataflows:

1) access all data for data quality, calibration, and reconstruction;

2) export all data to the WLCG Tier-1 data centres according to experiments' policies;

3) archival of all data onto tape at CERN.

The main change with respect to Run 1 (ATLAS and CMS only) is that the disk-only system (EOS) receives the data directly from the experiment and is used in the dataflows #2 and #3. ATLAS and CMS raw data are then moved from EOS to CASTOR to provide a tape copy. ALICE and LHCb are still using the same system as in Run 1, where all three

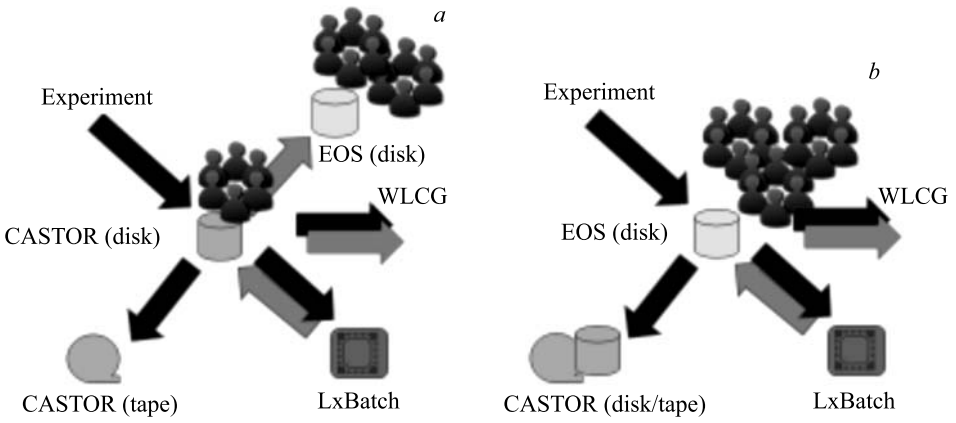


Fig. 1. The simplified schemas for the data handling at the CERN Tier-0. The black arrows represent the raw data and the grey ones are the data generated by the reconstruction. *a*) The schema used by ALICE and LHCb. The CASTOR disk front-end is directly used for reconstruction, while the analysis uses EOS disk. *b*) In the ATLAS and CMS scheme, EOS handles all the user activities, while the CASTOR disk acts as cache for the tape system

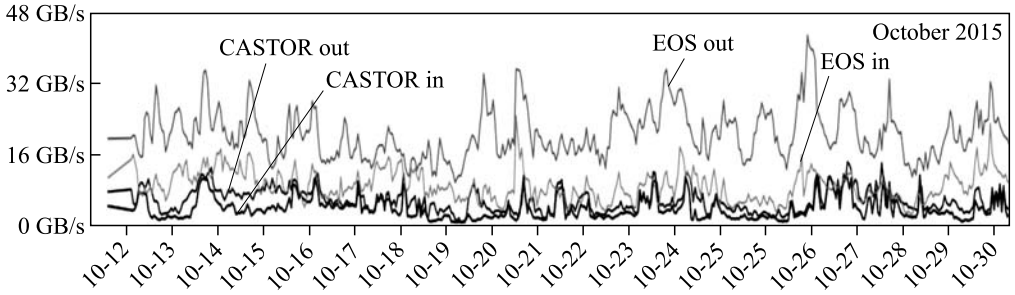


Fig. 2. CASTOR and EOS rates during October 2015

dataflows are implemented by CASTOR. In this case, EOS receives only the products of the reconstructions (and subsequent processing) for user analysis.

In all cases, the data flows are controlled by the experiments. The injection of the data into the computer centre is steered by the data acquisition, while the transfers between EOS and CASTOR and the export (dataflow #3) in general use the File Transfer Service (FTS). The simplified data transfers are illustrated in Fig. 1.

It is interesting to note that EOS, in its infancy at the beginning of Run 1, grew to a mature system offering about 70 PB of usable space for users (for most of the CERN experiments). In parallel, the disk layer of CASTOR went through a substantial simplification (software and deployment) and it is mainly used as a staging area in front of the tape system. At the beginning of Run 2 the CASTOR: EOS disk-resources ratio is about 1:3 ($\sim$ 25 PB vs $\sim$ 70 PB of usable storage). A snapshot of the activity of the CERN farms is shown in Fig. 2.

## SIGNIFICANT CHANGES IN RUN 2

We will discuss two important recent changes in our services, namely, the multisite capabilities of EOS and the introduction of a synchronization and sharing service called CERNBox.

## EOS AS A MULTISITE STORAGE SYSTEM

EOS stores files in multiple replicas or fragments distributed across different disk servers to be resilient to hardware failures and to optimize availability and performance. The disk servers expose all their disks as independent file system (JBOD) and the system chooses the location of file replicas (or fragments).

The CERN Computer Centre is now organized in two locations (the original CERN data centre and a second one at Wigner, near Budapest) connected with a redundant 100 Gb/s connection (two independent 100 Gb/s links with about 22-ms latency). All resources (disk storage and CPU capacity) are evenly distributed across the two locations.

EOS has evolved to support the two-location topology (geolocalization) and uses policies which can improve performance and service continuity. In the two-replica scheme, we use in the majority of cases, new files are created with one replica at each location. The batch nodes access the system preferentially using the local replica without overloading the two 100-Gb/s links and accessing the data with the minimal latency.

The presence of data at different locations opens up the possibility to deploy rather sophisticated resiliency mode. Presently the read activities in one centre can be independent from the performance of the Meyrin–Wigner link due to the deployment of read-only catalogues at each location.

The current system is a concrete example of a 100-PB multisite data centre. This approach suggests that other (groups of) sites in WLCG could simplify their deployment by exposing a set of storage elements at different locations as a single entity (here called “data league”).

WLCG general operations could profit from the reduced complexity if the storage resources were delivered by aggregated data leagues (nowadays around 170 data centres are part of the WLCG infrastructure). The increased resiliency of a “data league” compared to their parts can hide instabilities and downtimes of individual sites in most of the cases.

The deployment model demonstrated by EOS has also important advantages in the initial phase of the deployment (often the most intensive in effort). New sites joining a "data league" can concentrate in deploying only disk capacity, profiting of the central services from the established sites. This means that their resources become available immediately while they are progressively getting the necessary operational experience to run complete autonomous instances. This mechanism allows the sites in a data league to effectively share knowledge and in the second phase to provide redundant cataloging functionality in a reciprocal way (every site maintaining secondary catalogues or backup storage if needed).

## SYNCHRONIZATION AND SHARING SERVICES FOR HEP

For about two years, CERN IT-DSS is running a synchronization and sharing service called CERNBox [2]. It is based on the ownCloud open-software product [1] and initially delivered a cloud synchronization service similar to Dropbox™ or equivalent systems. In order to integrate it in our service offer and benefit of our large-scale disk services, we evolved EOS to efficiently inter-operate with the ownCloud system. The result is a successful service (over 3,500 users as in October 2015 with a three-fold increase in the last 12 months).

CERNBox extends the initial service provided by ownCloud in a significant way:

1) the CERNBox backend (EOS) exposes a HTTP/WebDAV interface such that the synchronization capacity will scale as the system capacity grows;

2) EOS as a file system maintains the information needed by the ownCloud clients for performing synchronization as extended attributes. This is a scalable solution to substantially offload the internal ownCloud data base;

3) EOS as a file system allows users direct access to their data. Users access all their data directly (e.g., for heavy-duty batch and Grid processing), while subset of data (e.g., selected input files or compact outputs) are *also* available to be synchronized and shared across local resources.

The first CERNBox extension is essential to scale up the number of users syncing their laptops. The load on the synchronization service is directly proportional to the number of users which have to regularly verify the coherence between their local copy and the central repository. An important byproduct of offering HTTP/WebDav access is that users (most notably Mac and Windows) can "mount" the EOS repository and allow direct access all the data (in addition to the FUSE mount available for Linux platforms).

The second extension removes a redundant cataloging level that would preclude the scale up of the system, for example, to match the complexity of the CERN home directory system (presently over 3 billion files).

The third extension has the potential to revolutionize the way the experiments interact with the data storage by federating all the data for users and groups (sharing). Probably for the first time this can be done in a uniform way across personal data (home directories) and experiment data (EOS experiment repositories). The goal is to provide a *coherent* system view for all the data, with high-performance access for the entire data repositories alongside with the option to snapshot data on local resources for developments and sharing.

[1] More information under http://owncloud.com.

## EVOLUTION OF EXISTING SERVICES

A recent WLCG survey [3] pointed out that the data-management operations absorb significant resources. At CERN we operate the vast majority of the storage for physics and for the IT infrastructure with a team of 10 engineers responsible for several major services with about 200 PB disk (2,000 servers) [1]. One should note that several important functions are performed by other CERN IT groups or via support contracts, notably to cope with the hardware replacement and interventions (in our case dominated by disk failures).

In order to improve our efficiency and minimize the hardware costs, all services use the same building blocks. The latest hardware acquisition corresponds to about 50 PB (about 250 data servers). The most recent data servers are built by a 10 Gbit/s node with 64 GB RAM. The storage attached to each server is about 200 TB (two 24-disk crates with 4-TB HDDs). While this hardware solution corresponds to the lowest price per disk space, it can be inefficient in providing storage to services and users. For example, operations like retiring a server are slow since they require large quantities of data to be moved out of big servers. In addition, single-disk performance cannot be scaled up easily. This problem is solved for EOS but remains for most of the other services notably small infrastructure services.

The problem has been solved by virtualization techniques, in analogy with the thin provisioning approach used for computing resources. CERN has virtualized its infrastructure and manages about 150,000 virtual machines CERN OpenStack cloud. For about three years, we have been operating a Ceph installation to support the CERN cloud and now we are using it to provide back-end storage to other mid-size services [4]. The original deployment (3 PB) was for long the largest Ceph instance worldwide. We have explored the possibility to scale up by a factor of 10 with dedicated tests which generated a lot of improvements to the code base to become part of future Ceph releases.

For example, we have replaced proprietary solutions to support legacy and internal services requiring the NFS protocol. The solution is a new infrastructure of independent NFS servers hosted in OpenStack virtual machines using Ceph block devices capabilities (Ceph RBD) [5].

In the near future, the CASTOR disk layer will be implemented with Ceph resources instead of physical disk servers. This will require a Ceph installation of 10/20 PB to buffer CASTOR data from/to tape using Ceph object store capabilities (RADOS). This will drastically simplify the most labour-intensive tasks in the CASTOR service (disk–server retirement and disk–usage rebalancing) using the same amount of raw disk space (using Ceph erasure coding capabilities). We have developed (and contributed to the upstream Ceph software stack) a library to be used by CASTOR (RADOS striper) to handle files as streams of objects from/to the Ceph infrastructure. The RADOS striper library allows one to reach high performance for single streams, which is a requirement to use efficiently the CASTOR tape infrastructure. During stress testing, we confirmed stable streaming performance above 300 MB/s (which matches the tape-driver requirements).

## NEW DIRECTIONS FOR THE FUTURE

In the last few years we have witnessed a spectacular rise in analysis activities. Simply by observing the EOS input and output rates we see about a threefold increase between the end of Run 1 and the end of LS1 (about 24 months). This is clearly going to continue due to more

demanding analyses. On the other hand, the analysis model has not changed substantially in the last several years, notably since the beginning of LHC.

The existing technology is mature for the development of integrated *data analysis* services. The integration of the JavaScript ROOT [6] viewer allows one already to share and manipulate results of analysis (histograms) within CERNBox. This is a concrete example of a service capable to support data access across a number of different client types which could be part of a novel analysis environment.

The keystone of such services will be the coherent cooperation of storage and compute resources. Users will transparently choose the best environment for the given tasks (e.g., the laptop for development and final steps of the analysis while using batch and grid resources for heavy-duty processing *without* the distraction of changing tool/environment at every step and automatically getting coherent configuration sets).

User analysis is usually done using the ROOT framework [6] relying on a complex environment of experiment-specific libraries, calibrations, and other read-only data (typically experiment specific and with different versions) and the actual files containing the analysis objects (reconstructed physics events). ROOT is already exploring Jupiter Notebooks technologies to prototype the idea of "ROOT as a Service" along recent development in the virtualization and cloud computing [7].

For the first time we have all the main ingredients and a clear structure to build a solid and attractive set of services to enhance analysis activities. Data services like CERNBox/EOS and CVMS [8] can cover all the data access part for analysis facilities based on ROOT notebooks, now at the level of prototype [7]. The construction of an analysis layer using the power of ROOT (and its evolution to support Python and R as supported languages) and the data technologies beyond CERNBox/EOS provide a general environment with a potential outside high-energy physics [9].

The CERN and the WLCG computing infrastructures will be an excellent showcase to demonstrate the value of our developments to other scientific or engineering communities which are confronted to problems requiring access to large datasets, the availability of heavy computational resources, and the sharing across broad collaborations of experts. The LHC experience in analysis tools and in handling large data sets can be of interest for other communities interested in collaborating in these areas.

**Acknowledgements.** The author is grateful to the entire CERN IT-DSS group. In addition, he acknowledges fruitful discussions with P. Mato, D. Piparo, and E. Tejedor (CERN PH-SFT) and J. Moscicki (CERN IT-DSS) in preparing the first white paper describing innovative data analysis services for HEP and other communities.

## REFERENCES

1. *Mascetti L.* CASTOR/EOS. Presentation at the CHEP Conf., Okinawa, Japan, April 2015;
   *Peters A. J.* EOS as the Present and Future Solution for Data Storage at CERN. Presentation at the CHEP Conf., Okinawa, Japan, April 2015.
2. *Mascetti L.* CERNBox. Presentation at the CHEP Conf., Okinawa, Japan, April 2015.
3. *WLCG Ops Team.* 2015 Site Survey. https://twiki.cern.ch/twiki/bin/view/LCG/WLCGSiteSurvey.
4. *van der Ster D.* Ceph at CERN: A Year in the Life of a Petabyte-Scale Block Storage Service // Openstack Summit, Vancouver, May 12–18, 2015.
   http://openstacksummitmay2015vancouver.sched.org/event/cd13cc9ba60d66d5d10c97be448975db#.

5. *Cano E.* Evolution of the NFS File Service. CERN ITTF Presentation, Oct. 2015. https://indico.cern.ch/event/438636.
6. ROOT web site. https://root.cern.ch/.
7. *Piparo D.* RaaS: ROOT as a Service and *Bellenot B.* Javascript ROOT. Presentations at the ROOT Workshop, Saas-Fee, Switzerland, Sept. 15–18, 2015; http://indico.cern.ch/event/349459/timetable/#20150916.detailed.
8. *Blomer J.* The Evolution of Global Scale Filesystem for Scientific Software Distribution // Comp. Sci. Engin. V. 17, No. 6. P. 61–71.
9. *Lamanna M.* Scalable Sync and Share Services as a Platform for Novel Applications: the CERN Experience. Presentation at the "ownCloud Connects Science, Research and Education Market" Workshop, Vienna, Oct. 2015.