WLCG DATA LAKE PROTOTYPE FOR HL-LHC

I. Kadochnikov [1,3,a], I. Bird [2,b], G. McCance [2,c], J. Schovancova [2,d], M. Girone [2,e], S. Campana [2,f], X. Espinal Curull [2,g]

[1] *Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia*

[2] *European Organization for Nuclear Research, 1 Esplanade des Particules, Geneva, 1211, Switzerland*

[3] *Plekhanov Russian University of Economics, 36 Stremyanny per., Moscow, 117997, Russia*

E-mail: [a] kadivas@jinr.ru, [b] Ian.Bird@cern.ch, [c] gavin.mccance@cern.ch, [d] Jaroslava.Schovancova@cern.ch, [e] Maria.Girone@cern.ch, [f] Simone.Campana@cern.ch, [g] xavier.espinal@cern.ch

A critical challenge of high-luminosity Large Hadron Collider (HL-LHC), the next phase in LHC operation, is the increased computing requirements to process the experiment data. Coping with this demand with today's computing model would exceed a realistic funding level by an order of magnitude. Many architectural, organizational and technical changes are being investigated to address this challenge. This paper describes the prototype of a WLCG data lake, a storage service of geographically distributed data centers connected by a low-latency network. The architecture of an EOS data lake is presented, showing how it leverages economy of scale to decrease cost. The paper discusses first experiences with the prototype and first test computing jobs reading data from the lake.

Keywords: GRID, storage, data lake, EOS, distributed storage, QoS

1. Introduction

High-luminosity Large Hadron Collider (HL-LHC) is the next step in increasing LHC luminosity. With more raw data produced by the ATLAS detector, computing resource requirements are predicted to exceed the limits expected to be available under the current computing model [1]. New and more efficient approaches to organizing distributed computing on this large scale are necessary to meet this demand [2].

Data storage and management is one of the areas where critical improvements to efficiency can be made. With more granular and explicit Quality of Service management for both replication and access latency, redundancy and storage cost may be decreased dramatically while still providing data availability.

Another way to decrease cost of WLCG data storage is to reduce human effort engaged in data center administration. Using large-scale distributed grid sites that consolidate individual institutional resources into a "data lake" is one possible path to that scenario.

CPU and storage resources are no longer always co-located in the GRID. Desktop grids, clouds and HPC are example of CPU-only resources. Caching is being increasingly used instead of pre-staging on individual sites with new software system being developed specifically for this mode of access [3]. In this atmosphere the prospect of dedicated distributed storage-only data lake does not look out of place, as it would in the early WLCG. With improving networks, caching solutions and data management, compute nodes sharing the room with the tape robot might no longer be a necessity.

To study the feasibility of using EOS to create such a highly distributed data lake a prototype with several collaborating sites was created. Support for granular parity-based redundancy of stored data was tested with realistic access patterns.

2. Data lake prototype

2.1. Architecture and participants

The prototype data lake is built as a distributed EOS storage system. In EOS metadata is stored separately from data, with MGM and MQ services responsible for the metadata and multiple FST servers providing data storage [4]. The central EOS metadata services of the prototype are located in CERN datacenters in Switzerland and Wigner Research Centre, Hungary. At the time of testing, in addition to CERN, 8 more research centers provided storage resources for the prototype in the form of local FST servers that joined the data lake.

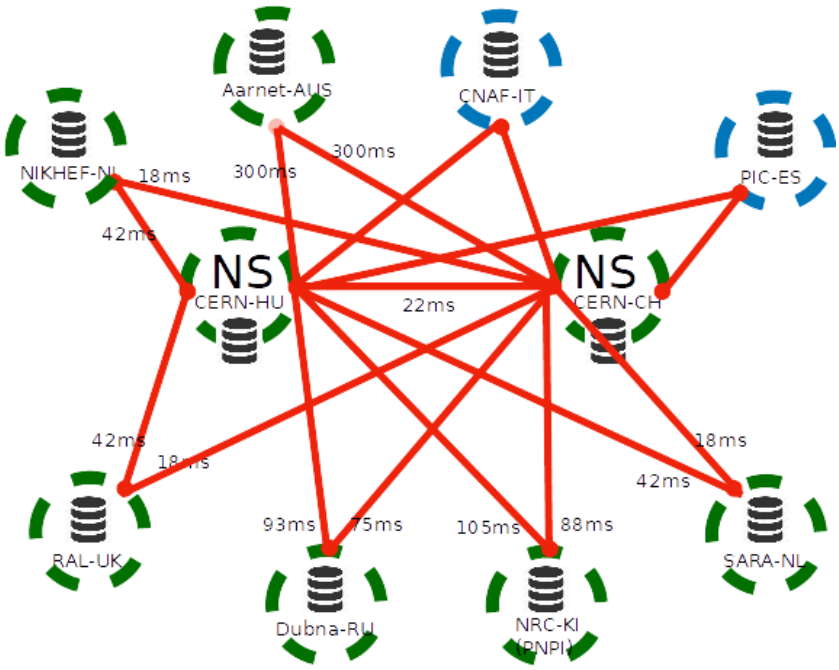

Figure 1. Prototype data lake participating research centers

2.2. Status and testing

The prototype resources are commissioned, the services are deployed, and basic transfer tests had been performed. Prototype transfer monitoring plots are presented in Figure 2.

Figure 2. Data lake prototype I/O and writing by type

QoS management with replicated and striped storage support was tested. Automatic replication of uploaded files based on namespace directory attributes, as well as manual conversion on command was observed to work. Some compatibility issues with several remote FST servers were observed.

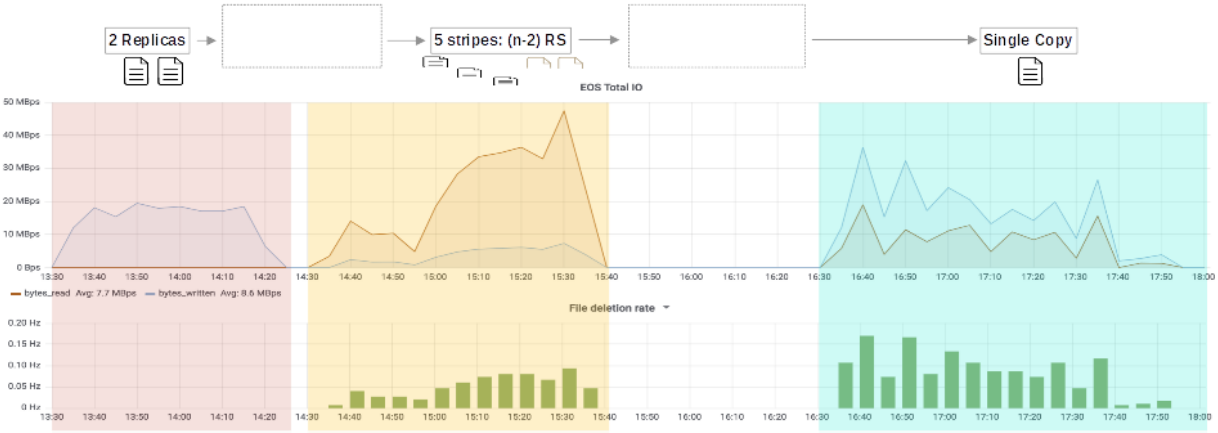

Figure 3. QoS conversions

The prototype was integrated into the ATLAS data management system Rucio[5] as a storage site and 6 input datasets used for Hammercloud[6] tests were copied onto the data lake.

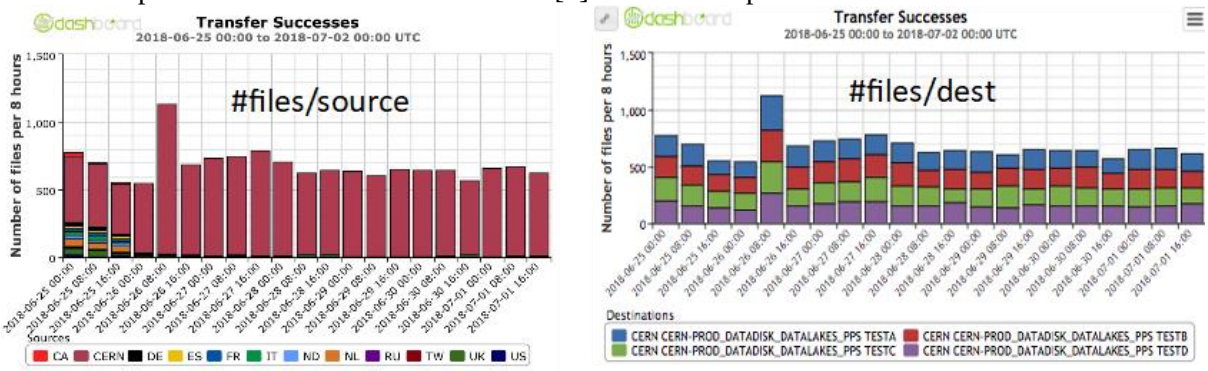

Figure 4. Transfer of data from the grid to the prototype, as seen by the data transfer monitoring

The Hammercloud input datasets were used to compare realistic data access patterns in 4 scenarios: no lake, data local to worker node; replicated data, at CERN; replicated data, not only at CERN; striped data, not only at CERN. In every scenario, the computing worknode was run at CERN.

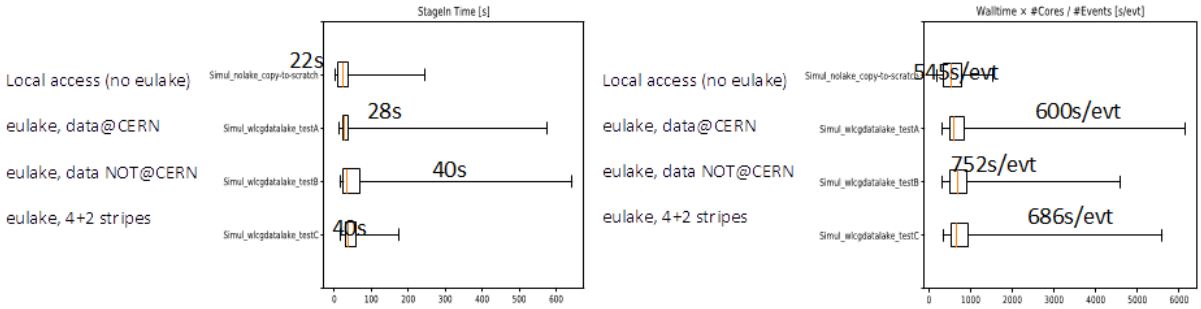

Figure 5. Low I/O test job ~40MB input (1 file), 2 events, ~5 mins/event

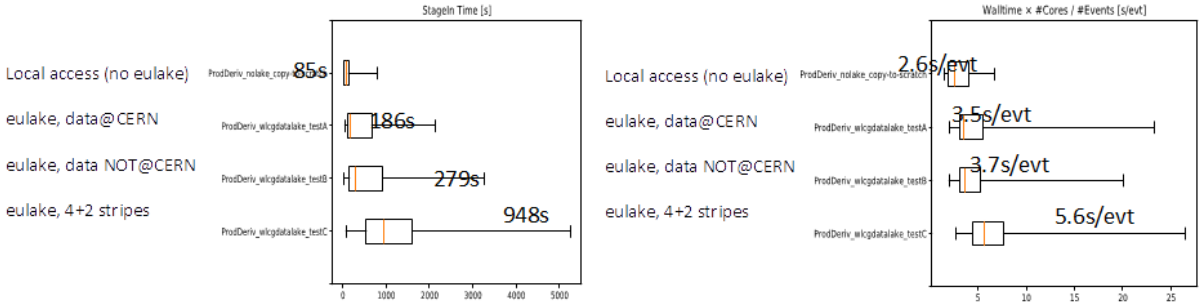

Figure 6. High I/O test job: ~6GB input (1 file), 1000 events, ~2 seconds/event

As can be seen on Figure 5, if a processing job does not require a lot of input data, execution time is not strongly affected by storage, as can be expected. On Figure 6 results for more I/O demanding job are shown. Noticeable degradation of service for striped storage can be attributed to compatibility and configuration problems on some of the storage servers with regards to striped storage support, as well as striped storage requiring data from at least 4 locations, which lowers effective throughput if one of the 4 is bottlenecked.

3. Conclusion

A distributed storage instance based on EOS was set up. This data lake prototype is small in terms of resources, but very geographically distributed. Different EOS deployment options are used by member sites. The lake is integrated with Rucio and Hammercloud. Performance of different data placement scenarios with different access patterns can be measured. CMS integration is the logical next step. The deployment and testing experience on the data lake will help prepare for HL-LHC.

References

[1] H. S. Foundation *et al.*, "A Roadmap for HEP Software and Computing R&D for the 2020s," *ArXiv171206982 Hep-Ex Physicsphysics*, Dec. 2017.

[2] S. Campana, T. Wenaus, and A. collaboration, "An ATLAS distributed computing architecture for HL-LHC," *J. Phys. Conf. Ser.*, vol. 1085, no. 3, p. 032029, 2018.

[3] Y. D. Cheng, Q. Xu, and C. Wang, "LEAF: A data cache and access system across remote sites," *J. Phys. Conf. Ser.*, vol. 1085, no. 3, p. 032008, 2018.

[4] A. J. Peters, E. A. Sindrilaru, and G. Adde, "EOS as the present and future solution for data storage at CERN," *J. Phys. Conf. Ser.*, vol. 664, no. 4, p. 042042, 2015.

[5] M. Barisits *et al.*, "The ATLAS Data Management System Rucio: Supporting LHC Run-2 and beyond," *J. Phys. Conf. Ser.*, vol. 1085, no. 3, p. 032030, 2018.

[6] J. Elmsheuser *et al.*, "Grid site testing for ATLAS with HammerCloud," *J. Phys. Conf. Ser.*, vol. 513, no. 3, p. 032030, 2014.