PoS PROCEEDINGS OF SCIENCE

Russian academic institutes participation in WLCG DataLake project

Andrey Kiryanov[1]
PNPI, NRC "Kurchatov Institute"
Gatchina, Russia
E-mail: Andrey.Kiryanov@cern.ch

Alexei Klimentov
BNL
Berkeley, CA, USA
E-mail: Alexei.Klimentov@cern.ch

Xavier Espinal
CERN
Geneva, Switzerland
E-mail: Xavier.Espinal@cern.ch

Andrey Zarochentsev
SPbSU, NRC "Kurchatov Institute"
Saint Petersburg, Russia
E-mail: andrey.zar@gmail.com

WLCG DataLake R&D project aims at exploring a technology evolution of distributed storage while bearing in mind very high demands of HL-LHC era. The primary objective is to optimize hardware usage and operational costs of a storage system deployed across distributed centres connected by fast networks and operated as a single service. Such storage could host a large fraction of the LHC and other mega-science experiment's data while eliminating inefficiencies due to various levels of fragmentation. In this paper we will describe Russian Institutes' and in particular NRC "Kurchatov Institute"'s role in the DataLake project with highlight on our goals, achievements and future plans.

International Symposium on Grids & Clouds 2019, ISGC2019
31st March – 5th April, 2019
Academia Sinica, Taipei, Taiwan

[1] Speaker

1. Introduction

High Luminosity LHC (HL-LHC) will be a multi-Exabyte challenge where the envisaged Storage and Compute needs are a factor 10 above the expected technology evolution and flat funding [1] (fig. 1).

WLCG community needs to evolve current computing and data organization models in order to introduce changes in the way it uses and manages the infrastructure, focused on optimizations to bring performance and efficiency not forgetting simplification of operations. These are the ingredients that will allow to drive down costs and be able to satisfy HL-LHC requirements.

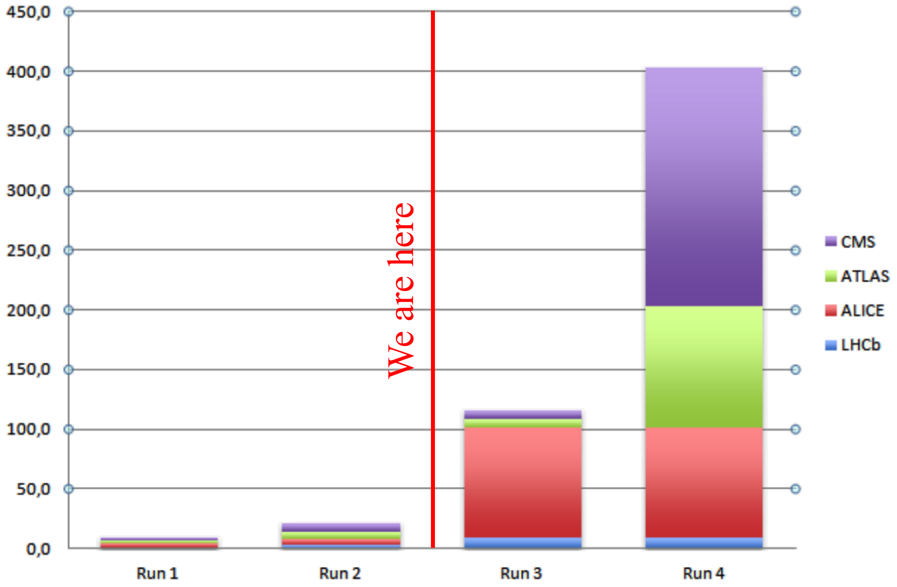

Fig. 1. Rough estimate of raw data volume of major LHC experiments (in PB per year).

Technologies that will address the HL-LHC computing challenges may be applicable for other communities, such as SKA, DUNE, CTA, LSST, BELLE-II, JUNO, etc. to manage large-scale data volumes. One of such technologies that we will discuss in this paper is Data Lake. Generally speaking, Data Lake is a set of sites, associated by proximity, providing together storage services, possibly accompanied by compute ones, to an identified set of user communities, capable to carry out independently well defined tasks. Proximity could be defined by geography, connectivity, funding or a shared user community. This requires that their combined storage capacity and network bandwidth can meet the demands of the designated task and that the usage of different sites is transparent to the users. This implies some form of trust relationship between the sites and a way to locate data, ranging from a simple file catalogue to a full fledged namespace.

While access for users is transparent, the population and management of the storage systems within the Data Lake, including data transition between QoS levels (fig. 2), is a planned and managed activity. These operations are done on the granularity of the Data Lake. Data is moved to or from the Data Lake as a whole, not to or from a specific site, because its internal stucture is not exposed and resource management within the Data Lake remains the responsibility of the Data Lake.

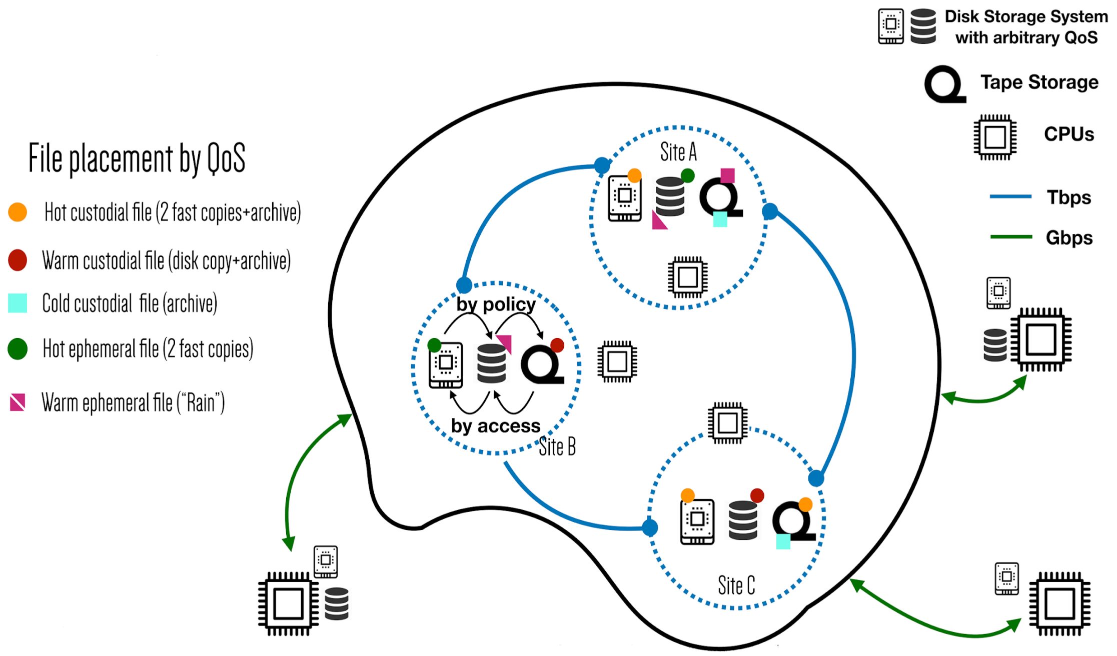
Disk Storage System with arbitrary QoS
Tape Storage
CPUs
Tbps
Gbps
File placement by QoS
● Hot custodial file (2 fast copies+archive)
● Warm custodial file (disk copy+archive)
■ Cold custodial file (archive)
● Hot ephemeral file (2 fast copies)
◩ Warm ephemeral file ("Rain")
Site A
by policy
by access
Site B
Site C

Fig. 2. A Data Lake comprising three sites with different compute capabilities.

Taking aforementioned aspects into account some basic but crucial requirements for the future WLCG data storage infrastructure can be defined:

- Common namespace and interoperability
- Coexistence of different QoS
- Geo-awareness
- File transitioning based on namespace rules
- File layout flexibility
- Distributed redundancy
- Fast access to data, latency compensation via caching
- Built-in fault tolerance

It's worth mentioning that Data Lake is one of the several storage-related R&D projects conducted in parallel. Other R&D projects aimed to address proper handling of storage systems with different QoS include:

- Data Carousel (started by ATLAS)
- Data Ocean (started by ATLAS + Google)
- Data Streaming

All of them are in progress as a part of DOMA [9] or/and IRIS-HEP [10] global R&D for HL-LHC. Authors strongly believe that it is important to develop a coherent solution to address HL-LHC data challenges and to coordinate above and future projects.

2. EULake prototype

In order to evaluate existing storage technologies and their applicability in the Data Lake model, it was decided to build an European Data Lake prototype (EULake) [6] spanning several WLCG sites with headquarters at CERN. Several years ago, in 2015, a similar prototype based on EOS [3] and dCache [4] storage systems was build on Russian sites during the Russian Federated Data Storage project [2]. Existing expertise in building and testing federated multi-site storages allowed us to fruitfully join the EULake with some decent resources and conduct important functional and performance tests.

As of 2019, EULake spans eight sites: CERN, JINR, NIKHEF, PIC, PNPI (part of NRC “KI”), RAL, SARA and UoM. Some of them only provide storage resources; others, including CERN and PNPI, also provide accompanying dedicated compute endpoints that allow to conduct real-life HammerCloud [7] tests on EULake infrastructure. All sites have also deployed perfSONAR [8] servers to automate network monitoring.

Initially, EOS storage system developed at CERN was the only software component used to build a working EULake prototype. One of the reasons was a rich feature set of EOS, which maps nicely into the basic requirements defined in section 1:

- Built-in namespace
- Storage groups and catalog attributes
- Geotags and Geo-scheduling
- Layout types (replica, RAIN)
- Support of xrootd [11] protocol and related proxy tools (xCache)
- Support for slave metadata managers (MGMs)

Of course, EOS is not the only software that can be used for such infrastructure. During the Russian Federated Data Storage project is was shown that dCache (version 2 at the time) can also be used in such a distributed installation. Moreover, dCache has significantly improved feature-wise in the last years with the release of version 3.

In order to allow sites and communities to have a freedom of choice of the storage system, and evaluate a slightly more heterogeneous Data Lake, EULake is currently transitioning from a EOS-only system into EOS + Rucio [4].

3. NRC “KI” participation in EULake

In order to take part in EULake a participating site has to provide some resources. Currently in Russia two major scientific centres participate in Data Lake R&D: NRC “Kurchatov Institute” and JINR, both using a virtualized environment.

NRC “KI” resources for the EULake are located at PNPI, Gatchina, in a newly built PIK Computing Centre (fig. 3). Unlike at CERN, EOS is not installed at PNPI on bare hardware, but deployed on top of Ceph [12] storage. The reason for this is added flexibility. EOS at CERN serves as a primary storage solution for LHC data (currently excluding tapes, which are still managed by CASTOR). Effort is being made to integrate virtually all CERN storage into EOS.

On the other hand, at NRC “KI” as a whole and at PNPI in particular LHC data storage and processing is an important one, but just one out of the many workloads. For instance, after a PIK Nuclear Reactor launch PNPI will become an international Nuclear Physics centre with its community requiring to store and process non-LHC data. In such cases Ceph allows for easy re-allocation of storage space between consumers with configurable redundancy for different types of data.

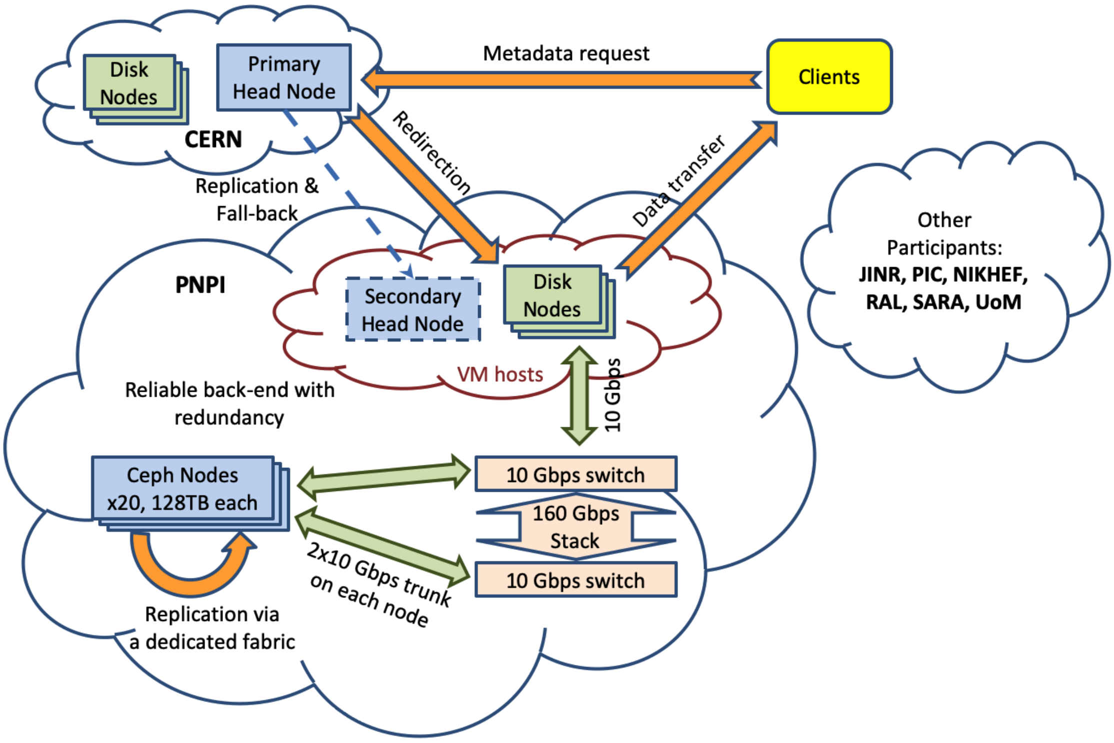

Fig. 3. EULake resources at PNPI.

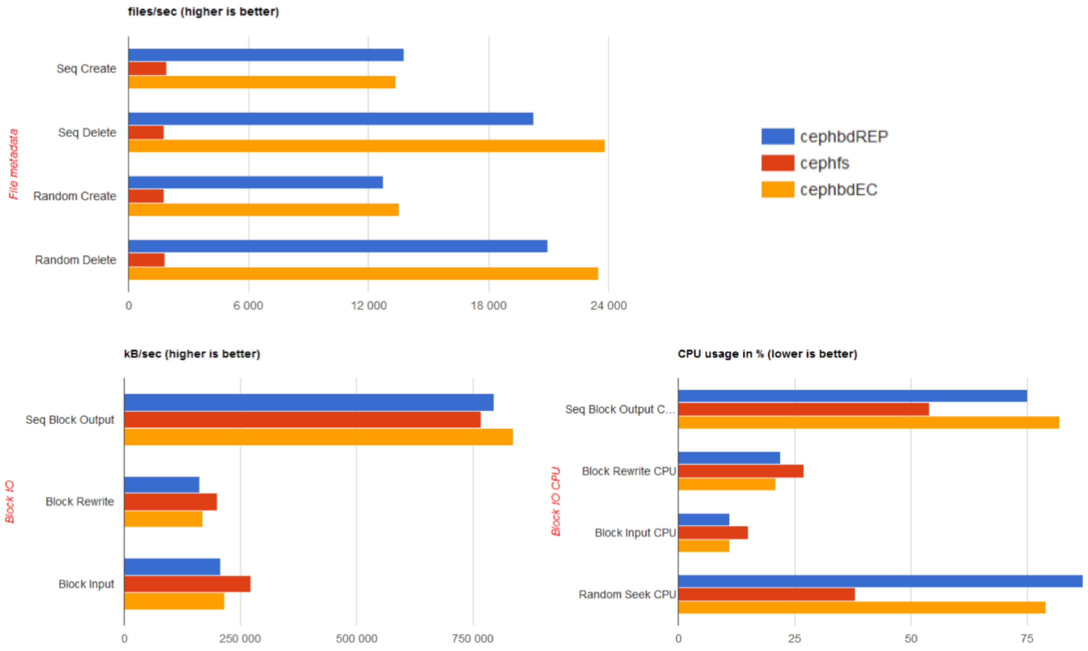

Fig. 4. EOS on top of Ceph performance measurements.

During the initial allocation of EULake resources at PNPI, an interoperation between EOS and Ceph had to be verified for any possible incompatibilities and performance bottlenecks. We have conducted EOS performance tests in three possible configurations: Ceph block device with replication, Ceph block device with Erasure Coding and Ceph filesystem (fig. 4). In our tests we were using the latest version of Ceph available at the moment: Mimic 13.2.1.

As it can be seen from the results, block I/O performance of Ceph replicated (cephrbdREP) and Erasure Coded (cephrbdEC) block devices as well as Ceph filesystem (cephfs) with EOS was on par, including the CPU utilization, while performance of metadata operations was significantly slower with Ceph filesystem. This was expected as Ceph filesystem maintains coherent metadata across all clients which adds latency overhead. As a conclusion we have decided not to deploy EOS on top of Ceph filesystem, but keep the deployment on top of Replicated and Erasure Coded block devices, which can be seen as different QoS types in the Data Lake terms.

4. NRC “KI” and JINR backbone networks

One of the crucial components of the Data Lake is a high performance network that is necessary for transferring data in and out of the Data Lake, as well as for efficient data access and reshuffling inside the Data Lake.

In 2018 NRC “KI” and JINR joined forces in an attempt to dramatically improve throughput and reliability of the primary backbone that connects Russian academic institutes to the international scientific networks like GEANT, NorduNet, etc. As a result of this effort new 100 Gbps network ring interconnecting hubs at Moscow, Amsterdam and CERN is being put in operation (fig. 5) and is expected to become fully operational in 2019.

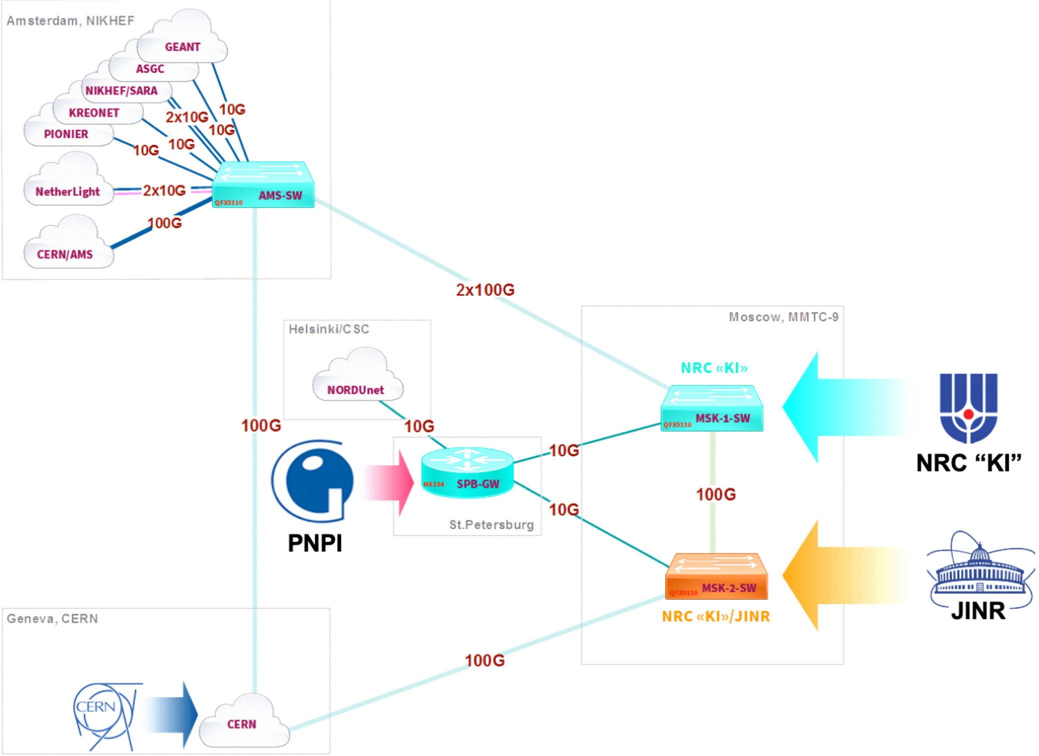

Fig. 5. NRC “KI” and JINR 100 Gbps backbone networks.

At Moscow Internet Exchange (also known as M9 Data Centre) two new high-performance network switches working in failover mode were installed, first of which is operated by NRC “KI” and the second one by JINR.

In addition to the 100 Gbps network links to Moscow, NRC “KI” also operates two 10 Gbps network links between Moscow and Saint-Petersburg, where PNPI’s PIK Data Centre is connected, enabling high-throughput connection between PNPI and international scientific networks.

5. Conclusions

Data Lake is currently operational as a Proof-of-Concept EULake prototype with real-life and synthetic tests running continuously on the infrastructure. Still, there’s much more to be done. In particular, we have to conduct an extensive testing of different types of QoS (possibly simulated) with different storage groups, test automated data migration, exploit different caching schemes and their impact on performance, figure out and implement proper fault-tolerance of the core components and finally evolve the infrastructure from a simple Proof-of-Concept to an infrastructure capable of measuring performance of future possible distributed storage models.

From NRC “KI” part we are particularly interested in deploying and testing a heterogeneous Data Lake in Russia that will allow diverse user communities of our major scientific organizations to effectively store and process big data from the future experiments.

6. Acknowledgements

This work was supported by the NRC “Kurchatov Institute” (№ 1608).

The authors express appreciation to the computing centres of NRC “Kurchatov Institute”, JINR and other institutes for provided resources.

References

[1] D. Adamova, M. Litmaath, *New strategies of the LHC experiments to meet the computing requirements of the HL-LHC era*, in proceedings of *55th International Winter Meeting on Nuclear Physics*, PoS (BORMIO2017) 053, 2017.

[2] A. Kiryanov, A. Klimentov, D. Krasnopevtsev, E. Ryabinkin, A. Zarochentsev, *Federated data storage system prototype for LHC experiments and data intensive science*, J. Phys.: Conf. Ser. 898 062016, 2016.

[3] https://eos.web.cern.ch/

[4] https://www.dcache.org/

[5] https://rucio.cern.ch/

[6] X. Espinal, *Data Lake R&D: high level goals*, Joint WLCG and HSF workshop, 2018.

[7] http://hammercloud.cern.ch/hc/

[8] https://www.perfsonar.net/

[9] https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities

[10] http://iris-hep.org/

[11] http://xrootd.org/

[12] https://ceph.com/