NRC "KI" PARTICIPATION IN THE DATALAKE PROJECT

A.K. Kiryanov[1,2], A.A. Klimentov[1,4], A.K. Zarochentsev[1,3]

[1] *NRC "Kurchatov Institute", 1 Akademika Kurchatova sq., Moscow, 123182, Russia*

[2] *Petersburg Nuclear Physics Institute of NRC "Kurchatov Institute", 1 Orlova Rocha, Gatchina, 188300, Russia*

[3] *Saint Petersburg State University, 7-9 Universitetskaya emb., Saint Petersburg, 199034, Russia*

[4] *Brookhaven National Laboratory, Upton, NY, USA*

E-mail: globus@pnpi.nw.ru

WLCG DataLake R&D project aims at exploring an evolution of distributed storage while bearing in mind very high demands of HL-LHC era. Its primary objective is to optimize hardware usage and operational costs of a storage system deployed across distributed centers connected by fat networks and operated as a single service. Such storage would host a large fraction of the WLCG data and optimize the cost, eliminating inefficiencies due to fragmentation. In this article we will emphasize on NRC "Kurchatov Institute" role in the DataLake project with highlight on our goals, achievements and future plans.

Keywords: HL-LHC, WLCG, DataLake, distributed storage

1. Previous work

A concept of a federated data storage system for LHC and non-LHC mega-projects was first presented in the Russian Federated Data Storage research project [1], started in 2015 by Big Data Technologies for Mega-Science Class Projects laboratory of NRC "Kurchatov Institute". This study has demonstrated that such systems could be successfully deployed on existing infrastructure using EOS and dCache storage systems as a basic platform [fig. 1].

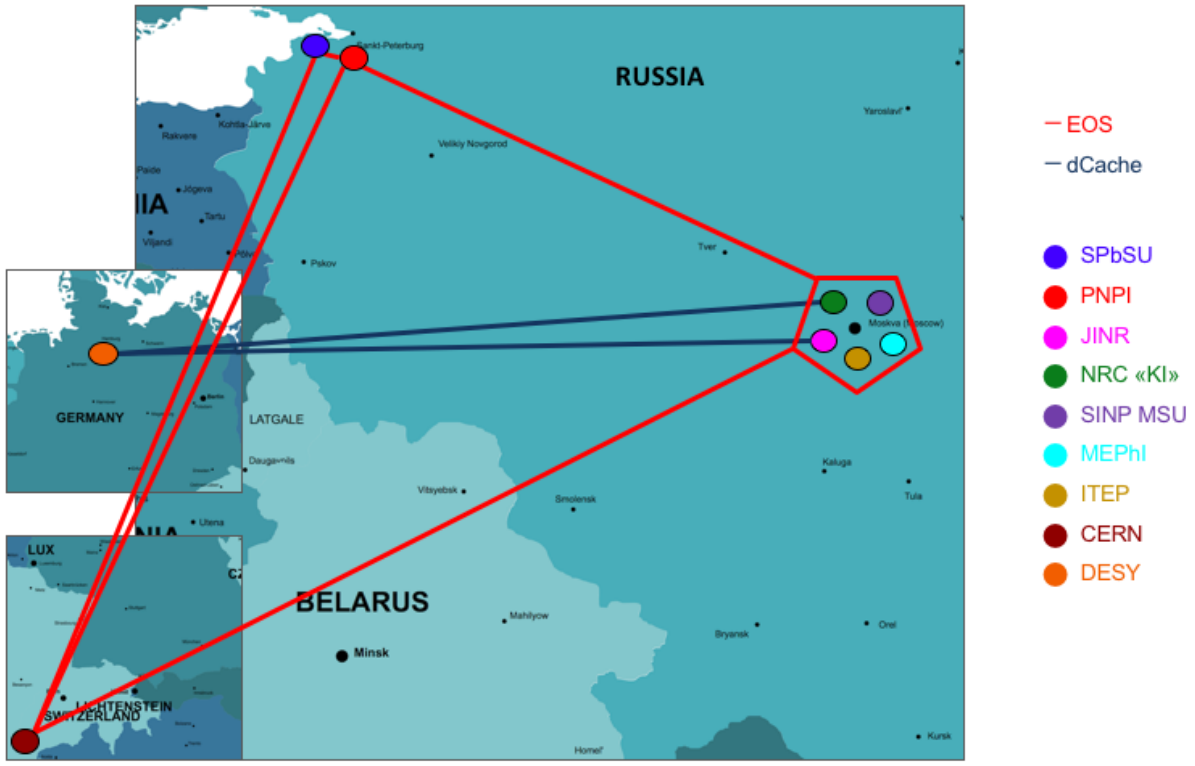

Figure 1. Russian Federated Data Storage project participants

Results of this project were discussed with WLCG collaboration and, along with other projects like Google Data Ocean, led to the emergence of an EU DataLake project with similar goals: providing a usable and homogeneous storage service with low requirements for manpower and resource level at sites.

2. WLCG DataLake project prototype (EUlake)

This project was first presented at the Joint WLCG and HSF workshop in Napoli in March 2018. One of the main specific reasons was the fact that HL-LHC storage needs to go far beyond standard expected storage technology evolution and funding [fig. 2].

WLCG DataLake project's main goal is to explore distributed storage evolution to improve overall costs (storage and ops) while maintaining the following set of features:

- Common namespace and interoperability
- Co-existence of different QoS (storage media)
- Geo-awareness
- File transitioning based on namespace rules
- File layout flexibility
- Distributed redundancy

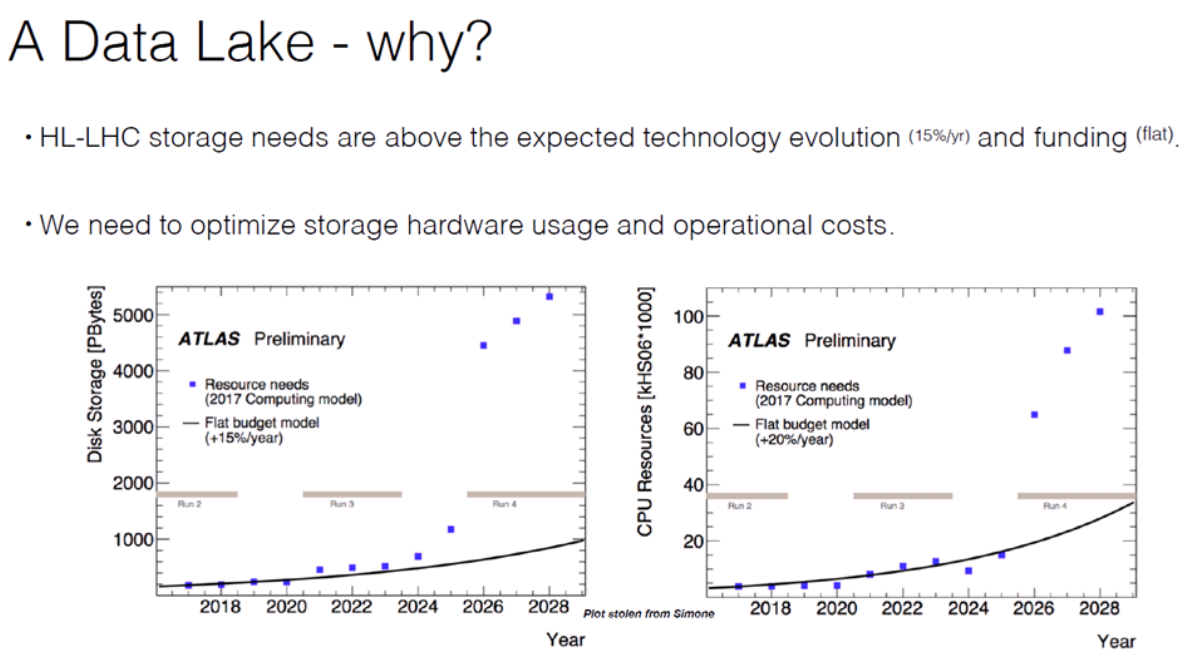

Figure 2. Reasons for the DataLake project

Xavier Espinal's slide at Joint WLCG and HSF workshop, Napoli, 26-29 March 2018

This R&D project aims to demonstrate that a dynamically distributed storage system with a common namespace:

- Has the potential to lower the cost of stored data
- Has the potential to ease local administration and world-wide operations
- Has the acceptable efficiencies in performance, reliability and resilience
- Is compatible with HL-LHC computing models

Currently the prototype is based on EOS storage system, but other storage technologies (dCache, etc) and their possible interoperability are also considered. The primary EOS namespace server (MGR) is deployed at CERN, it is agreed that a secondary namespace server will be deployed at NRC "KI" (PNPI) in Gatchina. Storage endpoints deployed at SARA, NIKHEF, RAL, JINR, NRC "KI", PIC, CNAF and Aarnet. perfSONAR endpoints are deployed at participating sites for network performance monitoring. EOS monitoring for all sites is hooked up to Grafana and performance tests are ready to be run in continuous mode.

3. NRC "KI" resources and role in the project

Kurchatov Institute decided to join this project because of an existing extensive expertise in deployment and testing of distributed storages gained during deployment of a similar prototype on Russian sites during Russian Federated Storage Project (which is still alive and available for testing). Furthermore, such an appealing universal storage technology may be useful not only for HL-LHC and HEP experiments, but also for other applications and fields of science like NICA, PIK and XFEL [2].

Our equipment for EUlake prototype is located at PIK Data Centre in PNPI, Gatchina [fig. 3] and has the following features:

- 10 Gbps connection to international scientific networks (GEANT, NorduNet) with both IPv4 and IPv6 protocols
- 300 TB of Ceph storage as a backend
- Extensive use of modern virtualization technologies like SR-IOV made it possible to deploy not only EOS but also perfSONAR endpoints on VMs

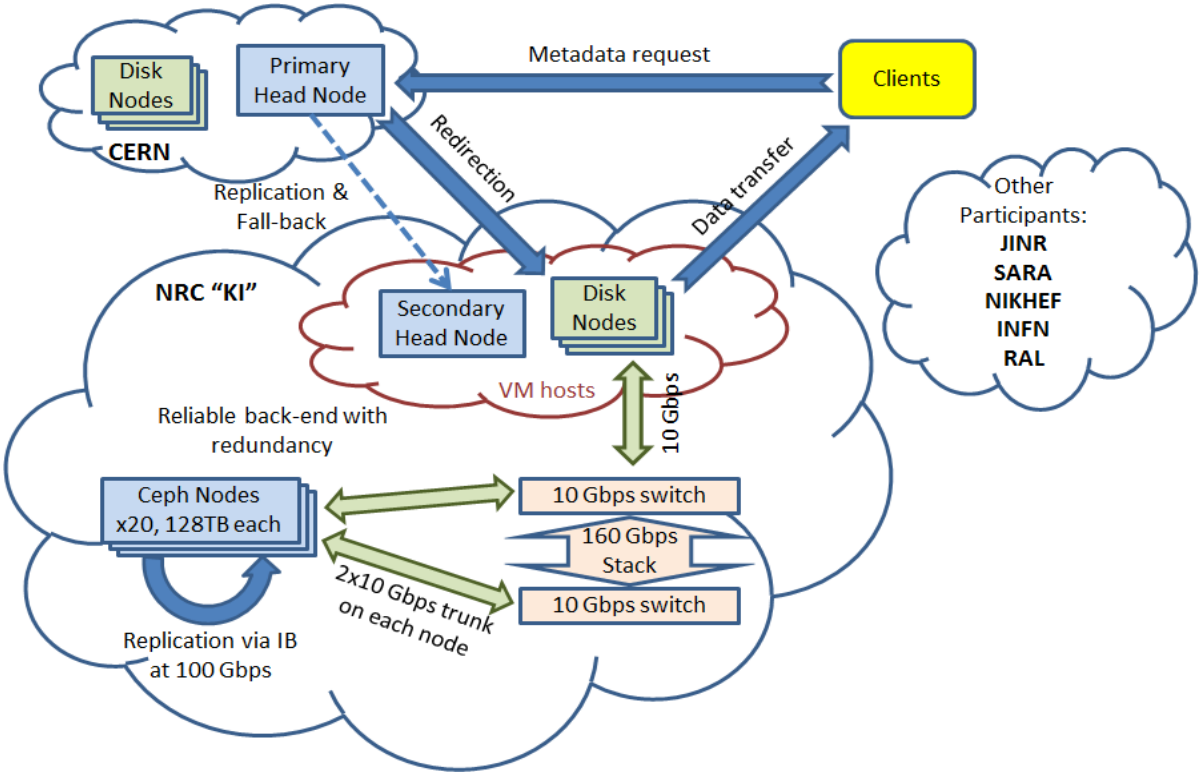

Figure 3. NRC "KI" resources for the DataLake project

We have decided to deploy EOS on Ceph because of a flexibility of such solution. While offering decent performance, it allows to reallocate storage resources without the need to re-install software and re-partition disk shelves, which is important when you're building a prototype where resource allocation may need to change. Out of 2.5 PB of Ceph storage available we currently provide 300 TB for EUlake prototype.

Ceph offers different types of storages for the clients: a filesystem and a block device both of which can be allocated on replicated or erasure coded pools. First of all we had to figure out which one is best for EOS. Our very first tests started with Ceph Luminous, but we quickly moved on to Mimic release because of its new monitoring features and improved filesystem performance.

All types of EOS storages were initially tested with Bonnie++ [fig. 4]. As we can see, block I/O performance was on par so we decided to expose and test all possible types at DataLake testbed, but in order to efficiently test our EOS storage within DataLake it was necessary to deploy a set of placement rules [3] which tied some directories in the virtual namespace to FSTs at NRC "KI". After that, the tests were conducted using xrdstress tool from EOS software bundle [fig. 5].

As expected, visible I/O performance of a client located close to the file servers (FSTs) increases for larger files and block sizes. The reason for this is less calls to the distant metadata server (MGM) which at the time of tests was located only at CERN. Maximum write speed for a single file was around 50 MB/s which corresponds to Ceph write speed.

After that, it was decided to augment the results of synthetic tests with some real-world data processing. The Russian part of EUlake is being intensively tested with ATLAS HammerCloud tests [4], for which some additional infrastructure nodes were deployed at NRC "KI".

4. Future plans

Our immediate plans are to identify a specific preferred Ceph storage configuration for EOS, polish geotags topology and geoscheduling rules, and to deploy a secondary metadata server (MGM) at NRC "KI".

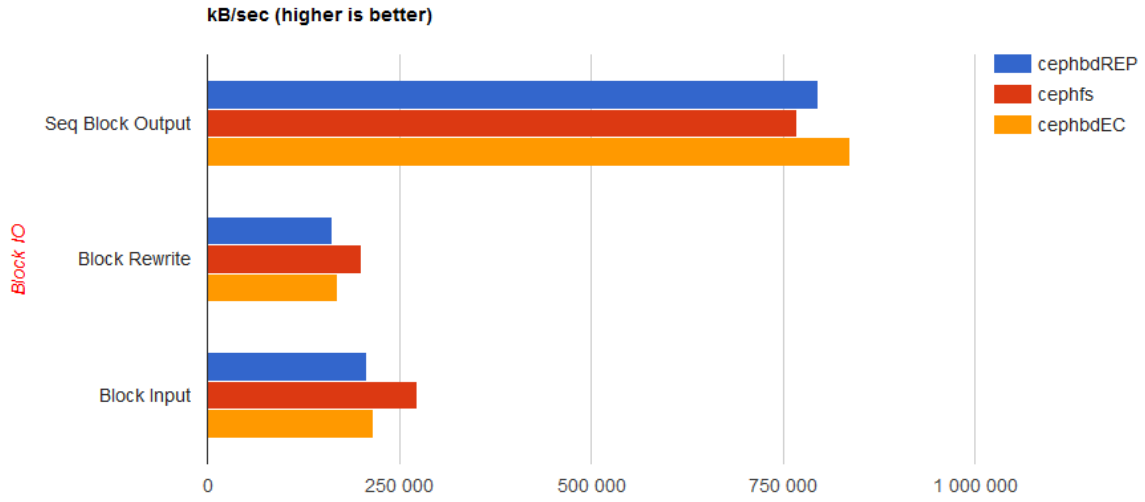

Figure 4. Bonnie++ test results for Ceph filesystem, Ceph block device on replicated pool and Ceph block device on erasure coded pool

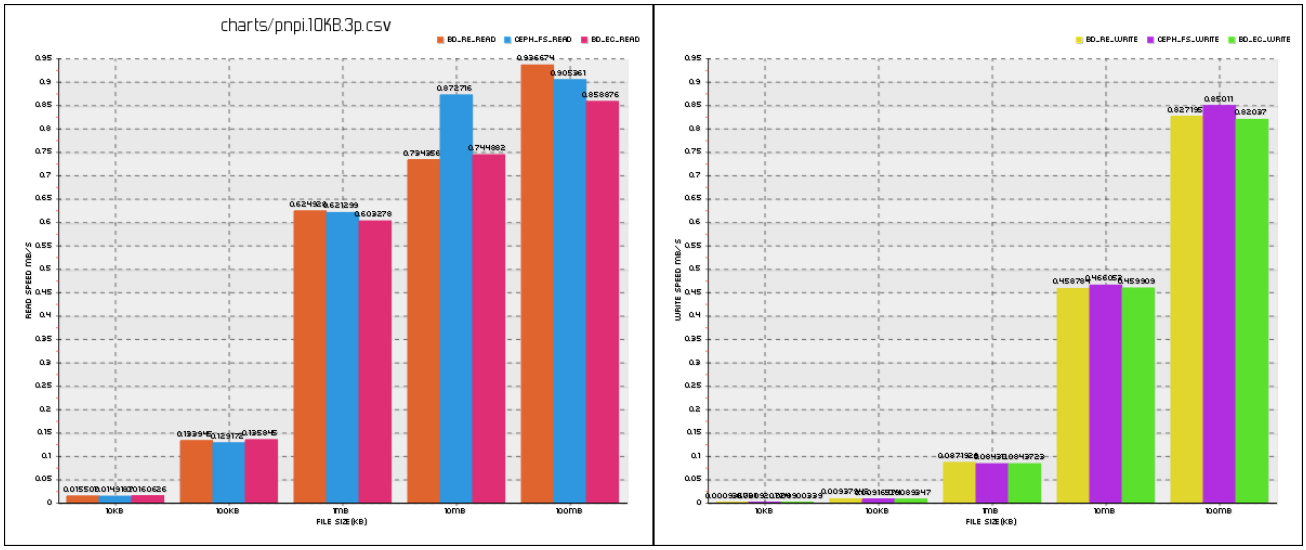

Figure 5. Xrdstress test from a client at NRC "KI" (block size is 100KB)

References

[1] A. Kiryanov, A. Klimentov, D. Krasnopevtsev, E. Ryabinkin, A. Zarochentsev. Federated data storage system prototype for LHC experiments and data intensive science // Journal of Physics: Conference Series, vol. 898, issue 6, 2017

[2] A. Kiryanov, A. Klimentov, A. Zarochentsev. Russian scientific data lake // Open Systems Journal, issue 4, 2018, https://www.osp.ru/os/2018/04/13054563/

[3] https://twiki.cern.ch/twiki/bin/view/WLCGDatalakes/FileLocality

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