TOWARDS RUSSIAN NATIONAL DATA LAKE PROTOTYPE

A. Alekseev[1,7,8], S. Campana[2], X. Espinal[2], S. Jezequel[9], A. Kiryanov[1,3,a], A. Klimentov[1,6], V. Mitsyn[4], A. Zarochentsev[1,5]

[1] *Plekhanov Russian University of Economics. 36 Stremyanny lane, Moscow, Russia*

[2] *CERN. 1 Espl. des Particules, Geneva, Switzerland*

[3] *Petersburg Nuclear Physics Institute of NRC "KI". 1 Orlova roshcha, Gatchina, Russia*

[4] *JINR, 6 Joliot-Curie st., Dubna, Russia*

[5] *St. Petersburg State University. 13B Universitetskaya emb., Saint Petersburg, Russia*

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

[7] *Institute of System Programming, Russian Academy of Science, Moscow, Russia*

[8] *Universidad Andrés Bello, Santiago,Chile*

[9] *L.a.p.p. Laboratoire d'Annecy De Physique Des Particules, Annecy-le-Veus, France*

E-mail: [a] globus@pnpi.nw.ru

The evolution of the computing facilities and the way storage will be organized and consolidated will play a key role in how this possible shortage of resources will be addressed by the LHC experiments. The need for an effective distributed data storage has been identified as fundamental from the beginning of LHC, and this topic has become particularly vital in the light of the preparation for the HL-LHC run. WLCG has started an R&D within DOMA project and in this contribution we will report the recent results related to the Russian federated data storage systems configuration and testing. We will describe different system configurations and various approaches to test data storage federation. We are considering EOS and dCache storage systems as a backbone software for data federation and xCache for data caching. We'll also report about synthetic tests and experiments specific tests developed by ATLAS and ALICE for federated storage prototype in Russia. Data Lake project has been launched in Russian Federation in 2019 to set up a National Data Lake prototype for HENP and to consolidate geographically distributed data storage systems connected by fast network with low latency, we will report the project objectives and status.

Keywords: HL-LHC, WLCG, Data Lake, DOMA, Distributed Storage

Aleksandr Alekseev, Simone Campana, Xavier Espinal, Stephane Jezequel, Andrey Kiryanov,
Alexei Klimentov, Valery Mitsyn, Andrey Zarochentsev

Copyright © 2019 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2507/44-50-paper-7.pdf

1. Introduction

The High Luminosity LHC (HL-LHC) will be a multi-Exabyte challenge where the envisaged Storage and Compute needs are a factor 10–100 above the expected technology evolution and flat funding (fig.1).

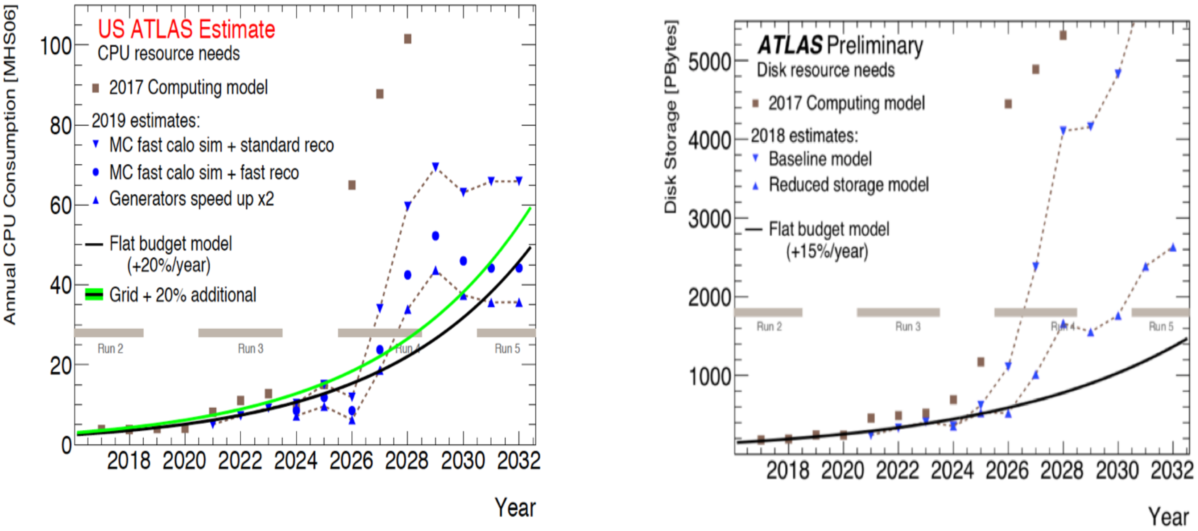


Figure 1. ATLAS projected CPU and storage needs

The 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. Technologies that will address the HL-LHC computing challenges may be applicable for other scientific communities (SKA, DUNE, LSST, BELLE-II, JUNO, NICA, etc.) to manage large-scale data volumes. The evolution of the computing facilities and the way storage will be organized and consolidated will play a key role in how this possible shortage of resources will be addressed by the LHC experiments. The need for an effective distributed data storage has been identified as fundamental from the beginning of LHC, and this topic has become particularly vital in the light of the preparation for the HL-LHC run. WLCG has started several R&Ds within Data Organization and Management Project (DOMA), one of which is a Data Lake project.

Data Lake is a set of sites, associated by proximity, providing together storage services, possibly accompanied by compute nodes 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 usage of the different sites is transparent to the users, which, in turn, 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 storages within the Data Lake is a planned and managed activity. This includes the transitions between Quality-of-Service (QoS) levels. 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. Resource management within the Data Lake is the responsibility of the Data Lake.

Taking the aforementioned into account we can come up with some basic but crucial requirements for the future WLCG sites data storage infrastructure:

- Common namespace and interoperability.
- Coexistence of different QoS.
- Geo-awareness.
- Fault tolerance through redundancy of key components.

- Scalability, with the ability to change the topology without stopping the entire system.
- Security with mutual authentication and authorization for data and metadata access.
- Optimal data transfer routing, providing the user direct access to the closest data location.
- Universality, which implies validity for a wide range of research projects of various sizes, including, but not limited to the LHC experiments.

2. Data Lake. Data Storage and Data Handling R&D Project

In 2015 in the framework of the Laboratory "BigData Technologies for mega-science class projects" at NRC "Kurchatov Institute" a work has begun on the creation of a united disk resource federation for geographically distributed data centres, located in Moscow, Saint Petersburg, Dubna, Gatchina (all above centres are part of the Russian Data Intensive Grid (RDIG) of WLCG) and Geneva, its integration with existing computing resources and provision of access to these resources for applications running on both supercomputers and high throughput distributed computing systems (Grid) [1]. The objective of these studies was to create a federated storage system with a single access endpoint and an integrated internal management system. With such an architecture, the system looks like a single entity for the end user, while in fact being put together from geographically distributed resources. This work was continued as a part of award granted by the Russian Science Foundation to the Laboratory of Cloud Computing of Plekhanov University. The concept of Russian Data Lake for Scientific Data is described in [2]. The resources used for RF data lake prototype are located at PNPI (Gatchina), JINR (Dubna), SPbSU (Saint Petersburg) and MEPhI (Moscow). Project milestones to be addressed in the context of the Data Lake prototype in Russia in the next few years:

- Deploy a working Data Lake prototype
- Develop and deploy a monitoring infrastructure
- Validate data access patterns
- Develop of a testing methodology for sites and data handling, conduct and automate a functional test suite:
  - Synthetic tests including files transfer/replication to have a realistic benchmark and to measure a performance of metadata operations;
  - Experiment-specific tests for I/O-intensive (derivation data production) and CPU-intensive (Monte-Carlo simulation) payloads.
- Create a data distribution model
- Connect Data Lake to the WLCG production infrastructure in Russia
- Use Data Lake for processing of a real experiments' data (LHC experiments + NICA)

Data Lake for Scientific Data project supported by the Russian Science Foundation award has been launched in Russian Federation in 2019 to set up a National Data Lake prototype for HENP and to consolidate geographically distributed data storage systems connected by fast networks with low latency. JINR, SPbSU, PNPI, and MEPhI groups are participating at the first stage of the project and it is anticipated that more centres will be involved during the subsequent stages. The short-term plans for building a distributed Data Lake system in Russian Federation are shown on fig. 2.

We are considering EOS and dCache storage systems as a backbone software for data federation and xCache for data caching. Synthetic tests and experiments specific tests have been developed by ATLAS and ALICE for federated storage prototype in Russian Federation.

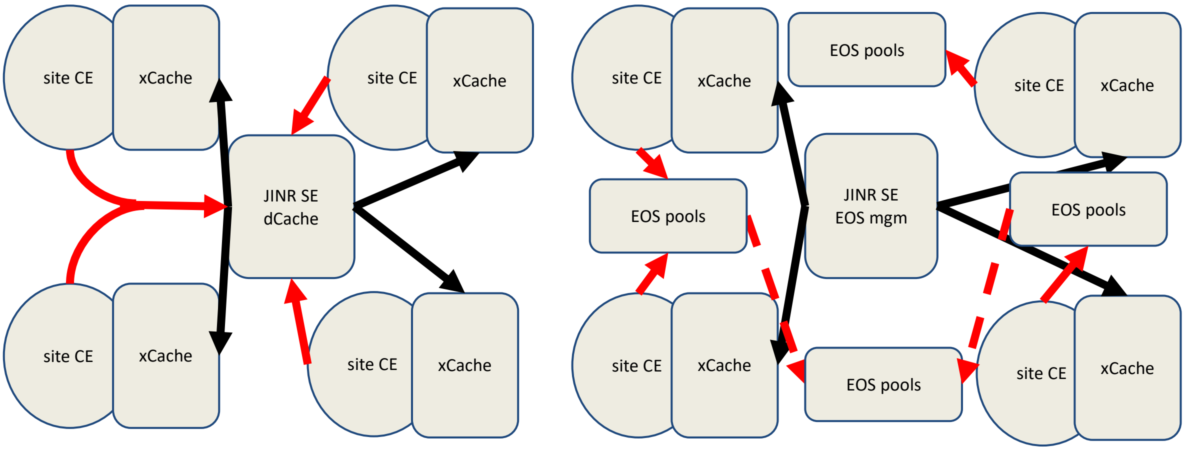


Figure 2. The short-term plans for creating a distributed Data Lake system in Russia for 2019 (left side) and 2020 (right side). Black arrows – data reading, solid red arrows – data writing, dashed red arrows – data replication

The Computing Element (CE) + xCache computing infrastructure has been set up at PNPI and access to it has been configured from JINR. Sites' technical characteristics are shown below:

- Worker Node @ JINR: 8 cores, Xeon E5420, 16GB RAM, 8.74 HEP-SPEC06 per Core
- Worker Node @ PNPI: 8 cores, Xeon E5-2680, 32GB RAM (VM), ~11 HEP-SPEC06 per Core
- Local network @ JINR (SE ⇄ CE) 1 Gbps
- Local network @ PNPI (SE ⇄ CE) 10 Gbps
- Network IPv4,6 JINR → PNPI: Latency ~5 ms
- Network IPv4,6 PNPI → JINR: Latency ~10 ms
- Network IPv4,6 JINR → PNPI: Throughput ~1 Gbps
- Network IPv4,6 PNPI → JINR: Throughput ~1,5 Gbps

The following authorization parameters were configured and tested:

- PNPI xCache → JINR SE: GSI authorization by local gridmapfile on JINR SE
- PNPI WN → PNPI xCache: GSI authorization by VOMS (ALICE & ATLAS)
- PNPI UI → JINR CE, PNPI CE (for local tests): GSI authorization by VOMS (ALICE & ATLAS)
- Hammer Cloud → ALL: GSI authorization by VOMS (ATLAS)
- An external library for VOMS authorization in xCache [3]

The following tests were conducted during the infrastructure set-up phase:

1. Synthetic tests from Worker Nodes and through Cream-CE
2. ATLAS HammerCloud tests:
    a. Copy2scratch: data copy from WN to scratch area
    b. Directaccess: remote data access

Tests were conducted on 3 configurations:

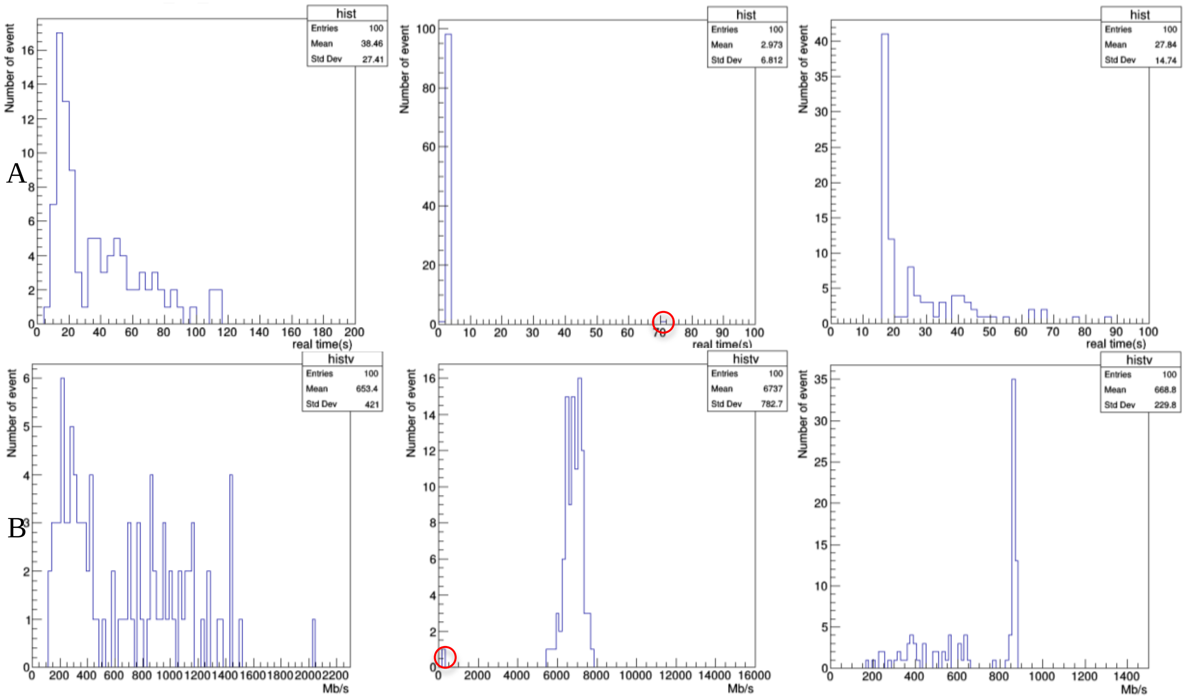
1. Direct access from PNPI WN to JINR SE: "PNPI direct"
2. Access through xCache from PNPI WN to JINR SE: "PNPI xCache"
3. Local access from JINR WN to JINR SE: "JINR local"

3. Synthetic tests

Synthetic tests are composed of sequential transfers of identical files accessing the storage in three different ways: directly from the remote site, through xCache at the remote site (PNPI) and directly from the site local to the storage (JINR). Synthetic test results are as follows:

- For "PNPI direct" the speed was 650±40 Mbps (fig. 3-1).

- For "PNPI xCache" the speed was 6700±700 Mbps, but if we remove the first hit (cache warm-up) we will get the same speed with a much lower deviation. As a result, with 100 hits we have a 95% transfer speed gain and 92% transfer time gain (fig. 3-2). For xCache transfers the first warm-up access to a new file is highlighted with a red circle.
- For "JINR local" with a 1 Gbps internal network tests show an expected transfer speed of 670±220 Mbps with a pretty high deviation (fig. 3-3).



1) 2) 3)

Figure 3. Synthetic test results: copy of a 1.9 GB root file from JINR-SE (100 iter.). Top plots (A) show time in seconds, bottom plots (B) - transfer speed in Mbps

4. ATLAS HammerCloud tests

Table 1 – Hammercloud tests results

Data access method	**Metric (avg. time in sec.)**	**PNPI direct**	**PNPI xCache**	**JINR local**
copy2scratch	Wallclock	2698 ± 577	1934 ± 139	1518±150
	Download input files	811 ± 574	53 ± 137	117 ± 17
directaccess	Wallclock	2150 ± 70	1906 ± 30	1200 ± 9
	Athena Run Time	2111 ± 46	1856 ± 22	1096 ± 0

HammerCloud is a software tool developed for ATLAS, CMS and LHCb experiments, it provides rich opportunities for automatic computing resource testing [4, 5]. Using Hammercloud framework we have created two templates namely directaccess and copy2scratch to test data access methods. The following parameters were defined to test computing infrastructure:

- Test type: stress
- Jobs template for derivation jobs with high IO access

The tests have been conducted for two days and results are presented in table 1. We can estimate a file transfer speed for Copy2Sctartch tests by "Download input files" metric, and for directaccess tests by "Athena Run Time" metric. Results of HammerClouds tests demonstrate 30% gain in time for copy2scratch and 12% gain in time for directaccess.

5. Monitoring

All of the components in the aforementioned tests were monitored by various monitoring systems. Synthetic tests were run through WLCG middleware (CREAM-CE) and tasks execution progress was monitored directly by CREAM-CE logs and monitoring tools. Tasks that were launched using HammerCloud were tracked using HammerCloud and BigPanda monitoring [6]. The state of network links was monitored using the mesh of perfSONAR [7] systems deployed at all participating sites sites. Separately, the work of the xCache service was monitored by Kibana monitoring [8] (fig. 4). All virtual nodes at PNPI were monitored by the Zabbix [9] suite (fig. 5).

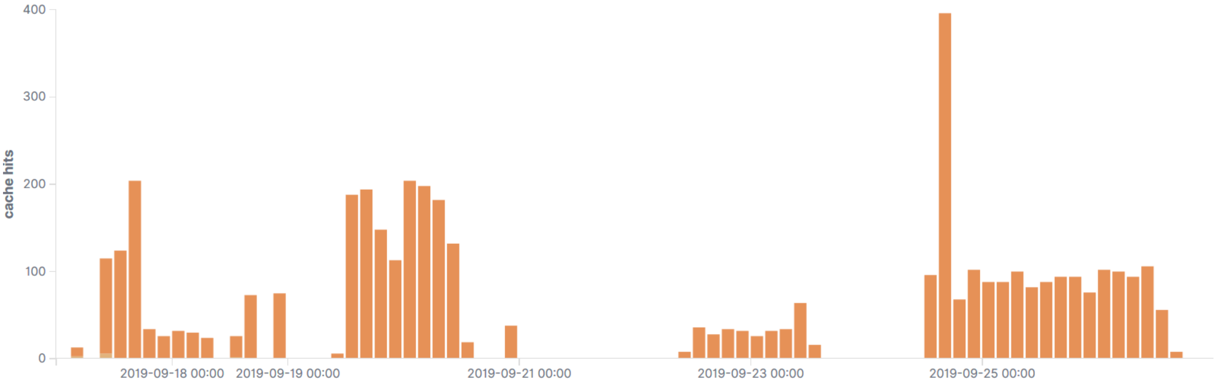


Figure 4. xCache Monitoring in Kibana: cache hits

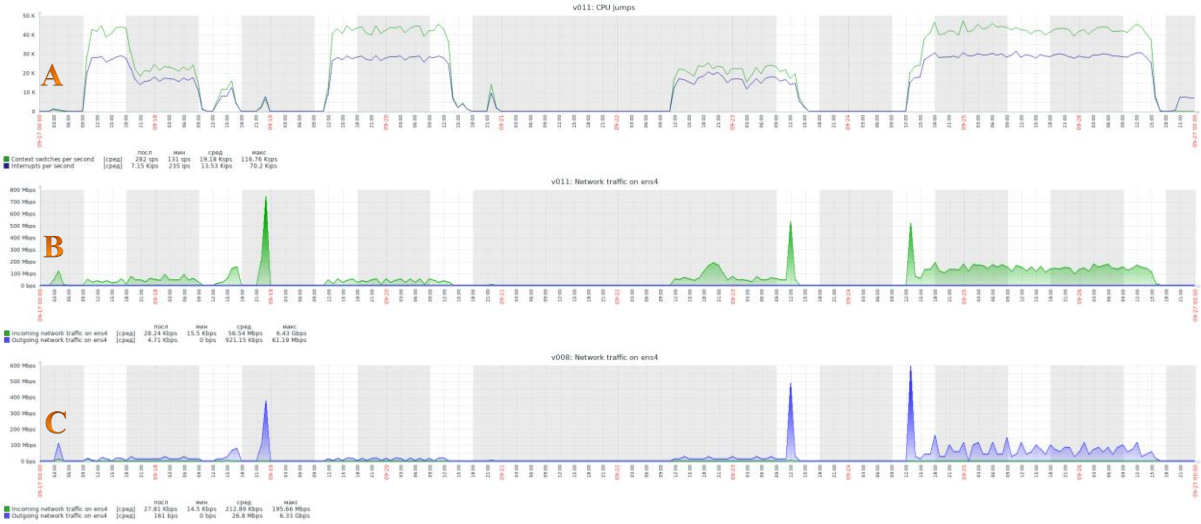


Figure 5. Zabbix monitoring at PNPI. A) WN CPU jumps, B) WN traffic, C) xCache traffic

Figures 4 and 5 show the monitoring data for the same period of time: 17.09-27.09. We can see a clear correlation of the data between Kibana and Zabbix which both show load peaks for the dates of HC tests - 19.09-21.09 and 24.09-26.09. This fact provides us with the opportunity to use these monitoring systems for further work on debugging and tuning the system as a whole as it becomes more complex at scale.

6. Summary

Data Lake for Scientific Data is a three years R&D project started in Russian Federation in 2019 as a continuation of the successful Federated Data Storage project. Production-grade computing resources and 10–100 Gbps network connectivity with low latency is used to prototype a production Data Lake. The project will have two phases:

1. Four Russian WLCG centres (JINR, PNPI, SPbSU and MEPhI) will participate and they will be configured to test an initial prototype. The test methodology and consolidated monitoring tools should be developed, as well as we need a better understanding of xCache data access control scenarios.
2. More RDIG centres will participate to have a realistic set of production resources and scattered storages. This prototype will be tested for real LHC use-cases, monitoring tools will be provided to sites and experiments.

7. Acknowledgements

The Russian Data Lake R&D project is funded by the Russian Science Foundation under contract No.19-71-30008 (research is conducted in Plekhanov Russian University of Economics). The work at PNPI is supported by the NRC "Kurchatov Institute" order No. 1571.

References

[1] A. Kiryanov, A. Klimentov, D. Krasnopevtsev, E. Ryabinkin, A. Zarochentsev, Federated data storage system prototype for LHC experiments and data intensive science // 2017 J. Phys.: Conf. Ser. 898 062016

[2] Kiryanov, A. Klimentov, A. Zarochentsev. Russian scientific data lake // Open Systems Journal, issue 4, 2018. Available at: https://www.osp.ru/os/2018/04/13054563/ (accessed 13.11.2019)

[3] XRootD repository. https://github.com/opensciencegrid/xrootd-lcmaps (accessed 13.11.2019)

[4] HammerCloud Distributed Analysis testing system. Available at: http://hammercloud.cern.ch/hc/ (accessed 13.11.2019)

[5] Johannes Elmsheuser et al, Improving ATLAS grid site reliability with functional tests using HammerCloud// 2012 J. Phys.: Conf. Ser. 396 032066

[6] A. Alekseev, A. Klimentov, T. Korchuganova, S. Padolski, T. Wenaus, ATLAS BigPanDA monitoring// 2018 J. Phys: Conf. Ser. 1085. 032043

[7] PerfSONAR official site. Available at: https://www.perfsonar.net/ (accessed 13.11.2019)

[8] ATLAS Kibana in Chicago. Available at:

https://atlas-kibana.mwt2.org:5601/s/xcache/app/kibana (accessed 13.11.2019)

[9] Zabbix official site. Available at: https://www.zabbix.com/ (accessed 13.11.2019)