Federated data storage system prototype for LHC experiments and data intensive science

A. Kiryanov[1,2,a], A. Klimentov[1,3,b], D. Krasnopevtsev[1,4,c], E. Ryabinkin[1,d], A. Zarochentsev[1,5,e]

[1] National Research Centre "Kurchatov Institute"

[2] Petersburg Nuclear Physics Institute

[3] Brookhaven National Laboratory, BNL

[4] National Research Nuclear University MEPhI

[5] Saint-Petersburg State University

E-mail: [a]globus@pnpi.nw.ru, [b]alexei.klimentov@cern.ch,

[c]dimitriy.krasnopevtsev@cern.ch, [d]rea@grid.kiae.ru, [e]andrey.zar@gmail.com

Rapid increase of data volume from the experiments running at the Large Hadron Collider (LHC) prompted national physics groups to evaluate new data handling and processing solutions. Russian grid sites and universities' clusters scattered over a large area aim at the task of uniting their resources for future productive work, at the same time giving an opportunity to support large physics collaborations. In our project we address the fundamental problem of designing a computing architecture to integrate distributed storage resources for LHC experiments and other data-intensive science applications and to provide access to data from heterogeneous computing facilities. Studies include development and implementation of federated data storage prototype for Worldwide LHC Computing Grid (WLCG) centers of different levels and University clusters within one National Cloud. The prototype is based on computing resources located in Moscow, Dubna, St.-Petersburg, Gatchina and Geneva. This project intends to implement a federated distributed storage for all kind of operations such as read/write/transfer and access via WAN from Grid centers, university clusters, supercomputers, academic and commercial clouds. The efficiency and performance of the system are demonstrated using synthetic and experiment-specific tests including real data processing and analysis workflows from ATLAS and ALICE experiments, as well as compute-intensive bioinformatics applications (PALEOMIX) running on supercomputer. We present topology and architecture of the designed system, report performance and statistics for different access patterns and show how federated data storage can be used efficiently by physicists and biologists. We also describe how sharing data on a widely distributed storage system can lead to a new computing model and reformations of computing style, for instance how bioinformatics program running on supercomputer can read/write data from federated storage.

Keywords: WLCG, Federated Storage, Big Data, EOS

1. Objective

In 2015 in the framework of the Laboratory “Big Data Technologies for mega-science class projects" in NRC "Kurchatov Institute" a work has begun on the creation of a united disk resource federation for geographically distributed data centers, located in Moscow, St. Petersburg, Dubna, Gatchina (all above centers are part of the Russian Data Intensive Grid of WLCG) and Geneva, its integration with existing computing resources and provision of access to this resources for applications running on both supercomputers and high throughput distributed computing systems (Grid).

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 form geographically distributed resources. The system should possess the following properties:

1. fault tolerance through redundancy of key components;
2. scalability, with the ability to change the topology without stopping the entire system;
3. security with mutual authentication and authorization for data and metadata access;
4. optimal data transfer routing, providing the user direct access to the closest data location;
5. universality, which implies validity for a wide range of research projects of various sizes, including, but not limited to the LHC experiments [LHC].

2. Underlying technology choice

Today, there are particular solutions for abovementioned problems. For example, some of the LHC experiments use storage systems on top of which they build their own data catalogs. ALICE experiment [Aamodt, Abrahantes Quintana,…, 2008] for instance uses a storage system based on xrootd, which keeps track of only physical file names (PFN), and utilizes a separate metadata catalog integrated in ALIEN infrastructure. This approach does not satisfy the requirements of flexibility and scalability. Solutions that satisfy all of the stated requirements are:

- storage based on HTTP Dynamic Federations (DynaFed) [DynaFed];
- EOS storage system [EOS];
- dCache storage system [dCache].

So far as EOS storage system is a complete integrated project, tested and used by all four major LHC experiments, ALICE, ATLAS [The ATLAS Collaboration…, 2008], CMS and LHCb, the authors decided to start with it as a prototype software platform.

3. Federated data storage prototype

For federated storage prototype we have chosen geographically distributed resource centers located at the vertices of a triangle (figure 1): PNPI and SPbSU in St.-Petersburg district, NRC KI and JINR in Moscow district and CERN in Geneva. CentOS 6 and EOS storage system have been used as a software platform. During the deployment of the federation it was necessary to define the topology of storages and a common authentication scheme. Taking into account that the resources provided by federation participants are roughly equivalent, it was decided to use a simple layout: one management server (MGM) and one storage server (FST) are deployed in each organization. As EOS does not support simultaneous operation of multiple peer MGMs within one segment, one of the MGMs operates in master mode (primary) and others operate in slave mode (secondary) with automatic metadata synchronization. This solution allows having a single access endpoint through the primary MGM, and improves system fault tolerance with the ability to use one of the secondary MGMs as primary in the case of failure.

Normally, the client request first goes to the top-level MGM server on which the authentication and authorization are performed, enabling access to the metadata. Top-level MGM then redirects the client to the appropriate (optimal) storage server (FST), thus ensuring optimal data transfer routing.

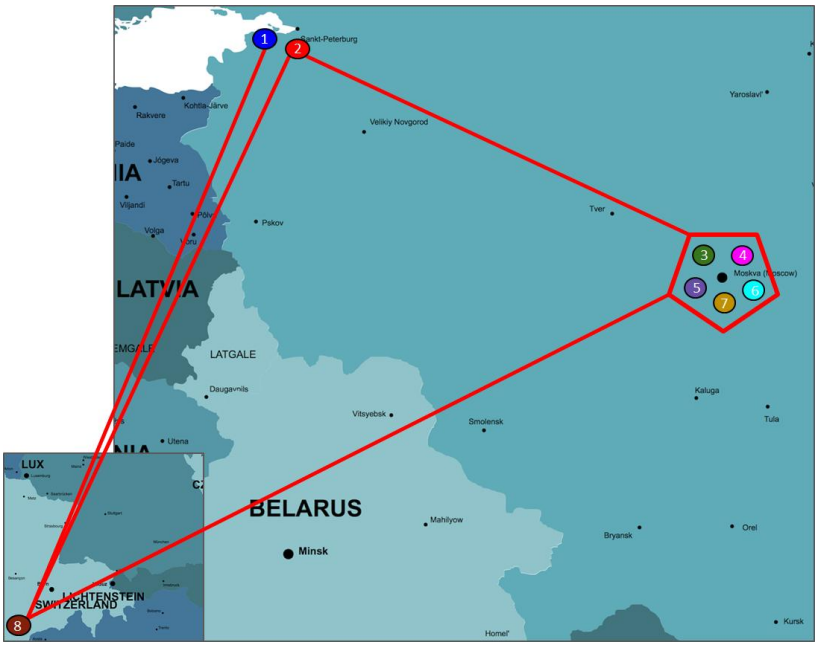

Fig. 1. Participants of the federation prototype. SPbSU(1), PNPI(2), NRC "KI"(3), JINR (4), SINP(5), MEPhI(6), ITEP(7) and CERN(8)

4. Testing procedure

The aim of the test is to verify reliability and obtain quantitative characteristics of performance efficiency of the data analysis jobs for the ATLAS and ALICE experiments. At first testing is carried out with the help of synthetic benchmarks, including remote access over the WAN, followed by efficiency measurements by real-life experiment applications.

As a first synthetic benchmark we used a synthetic Bonnie++ [Bonnie++] test, which is capable of measuring file and metadata access rate on a filesystem. EOS supports mounting as a virtual filesystem via Linux FUSE [FUSE] mechanism, which makes it possible for Bonnie++ to test both file read-write speeds (FST access) and metadata transaction rates (MGM access) independently. In addition, an xrdstress tool bundled with EOS was used for stress-testing. This tool simulates parallel access to EOS from an application by writing and reading back a set of files and is only capable of measuring file I/O performance.

PerfSONAR [perfSONAR] suite was used for independent network performance measurements. Two significant metrics that were measured between each pair of resource centers are bandwidth and latency. These metrics allow to understand the impact of network parameters and topology on performance test results.

Reliability test was performed by simulating a failure of management servers with metadata migration and role switch between master and slave MGMs.

4.1. Processing and analysis of the ATLAS experiment data

ATLAS test application performs a proton-proton event reconstruction using so-called "raw" data as an input. Specific of this task is the need for reconstruction of all particle jets in the Transition Radiation Tracker (TRT) detector. Reconstruction of a signal from each of the proportional drift tubes in

TRT is one of the most challenging and highly CPU-bound task, especially in high-luminosity conditions.

Reconstruction if performed in several stages, each of which requires intensive read and write of data stored in the federation. In addition, test application also analyzes the kinematic distribution in TRT, which allows to compare results obtained with federated and traditional storages.

Input and output data, be it a single file or a dataset, may be accessed both locally, on the filesystem, and remotely via xrootd protocol.

4.2. Processing and analysis of the ALICE experiment data

ALICE test application sequentially reads events from a file, analyzes them and produces information about the most "interesting" event according to the specified selection parameters. This application was specifically invented to evaluate the performance of the storage system. Just like before, input dataset can be accessed both locally and remotely via xrootd protocol.

5. Test results

In order to run ATLAS and ALICE test applications we have deployed a standard user interfaces nodes containing all the necessary software stack which fully corresponds to the classical data processing environments used in both experiments. Following the developers' recommendations, perfSONAR nodes were deployed on dedicated physical machines with at least 4GB of RAM.

So far, we have tested FST and MGM servers, located at CERN, SPbSU and PNPI. We have evaluated the following resource combinations:

SPbSU (FST and MGM on the same server);

PNPI (FST and MGM on the same server);

PNPI (MGM master and FST on different servers) + SPbSU (MGM slave and FST on different servers);

CERN (MGM master) + PNPI (MGM slave and FST on different servers) + SPbSU (FST).

5.1. Configuration and testing of the WAN Federation

Since the first test results are only available for PNPI, SPbSU and CERN network data channel performance measurements are only of interest between these centers. Network channel conditions were measured with perfSONAR installed on all participating centers of the federation.

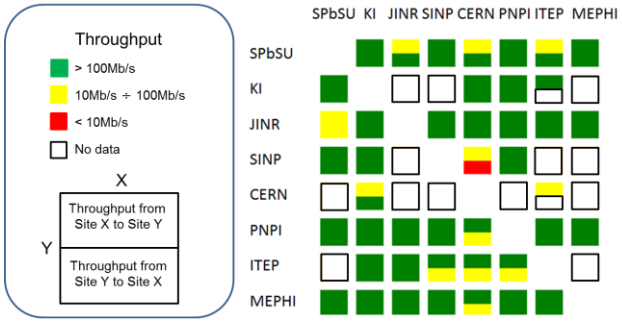

Fig. 2. Throughput between participants

In figure 2 we can see that data transfer rates are not symmetric. This behavior is due to the fact that network links of participating resource centers are not dedicated and also handle production workload. It makes our test results harder to interpret but at the same time gives us the opportunity to test federation in the real-world conditions with saturated network channels.

5.2. Bonnie++

For Bonnie++ test the most informative are file (figure 3 left) and metadata (figure 3 right) read-write performance diagrams, since in our scenario files and metadata are stored at different locations. During these tests CERN was used as a master MGM. We can see that metadata I/O performance depends solely on a link between client and manager while data I/O performance does not depend on a link between client and manager.

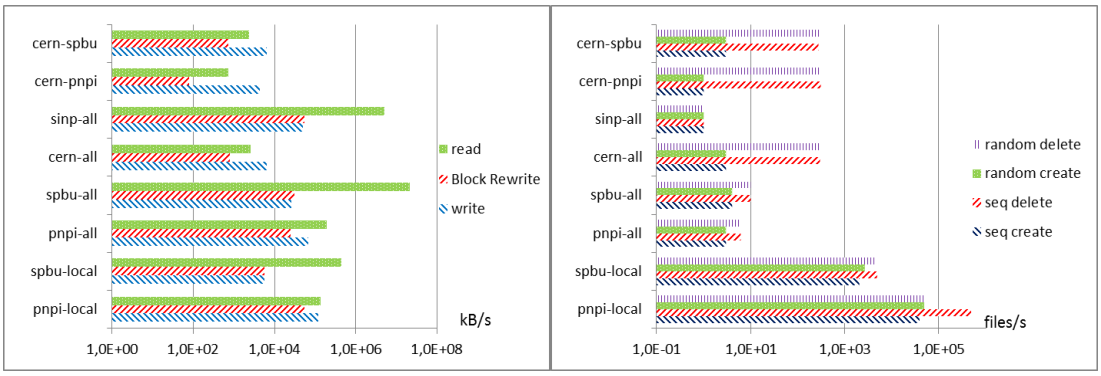

Fig. 3. Bonnie++ test results for block file I/O (left) and for metadata transaction rate (right). On the vertical axis we put a site combination (local means standalone) and test type. On the left plot on horizontal axis we put a data transfer rate in KBps in logarithmic scale; On the right plot on the horizontal axis we put a transaction rate in operations per seconds in logarithmic scale.

5.3. Experiment-specific tests

During ATLAS event reconstruction test we had to reasonably choose application output parameters that appeared the most suitable for our needs. With a single input file, the most reasonable performance estimate was an application initialization time, i.e. the duration of the stage where input data are preloaded from the storage. Another particularity in comparison with Bonnie++ test is that the input files do not necessarily have to be available in the locally mounted filesystem and can be read remotely via xrootd protocol. Therefore, in the results of this test we have one additional parameter: access type, which is either FUSE or xrootd. The dependence of the test application initialization time on various client-server combinations and data access protocols is shown in figure 4 right.

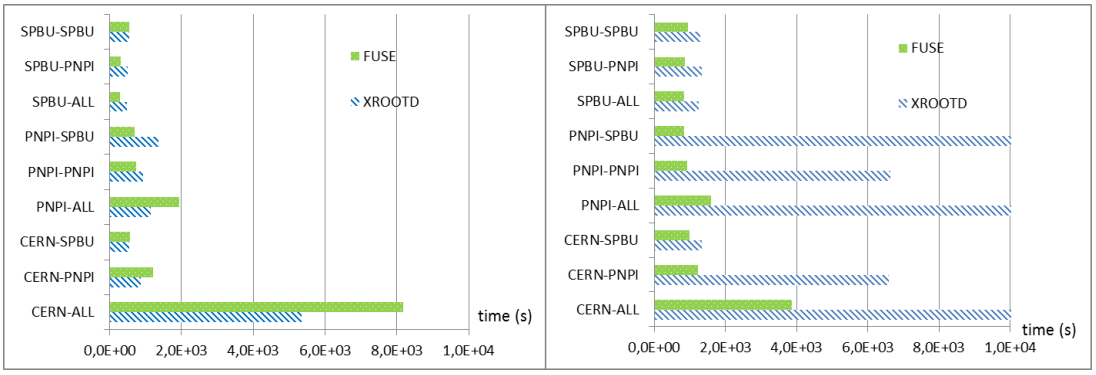

Fig. 4. The left plot is the ALICE test application runtime. The right plot is the ATLAS test application initialization time. On the both plots on the vertical axis we put a UI -FST combination and data access protocol and on the horizontal axis we put time in seconds.

ALICE tests were specifically written to test the performance of the federated storage and do not have performance information in the output as ATLAS tests do. Therefore, a complete test runtime

measured with a system ‘time’ utility was taken as a primary parameter. The dependence of the ALICE test application run time on various client-server combinations and data access protocols is shown in figure 4 left.

As we can see from the plots above, experiment-specific tests for different data access patterns have contradictory preferences with respect to data access protocol (pure xrootd vs. FUSE-mounted filesystem).

6. Data placement policy

Number of data replicas depends from data type (replication policy has to be defined by experiments / user community);

Federated storage may include reliable sites(“T1s”) and less reliable sites (“Tns”);

Taking aforementioned into account we have three data placement scenarios which can be individually configured per dataset:

Scenario 0: Dataset is randomly distributed among several sites;

Scenario 1: Dataset is located as close as possible to the client. If there’s no close storage, the default reliable one is used (NRC “KI” in our tests);

Scenario 2: Dataset is located as in scenario 1 with secondary copies as in scenario 0.

All described tests have been performed on extended testbed.

6.1. Data population performance test from CERN

Data population procedure is as follows:

Scenario 0: Files are copied to several file servers;

Scenario 1: All files are copied to the default file server at NRC “KI”, because there’s no storage close to CERN in our testbed;

Scenario 2: All files are copied to the default file server at NRC “KI” with secondary replicas on several servers.

As we can see in figure 5, there’s a slight increase in transfer speed with distributed write. Replication costs are less than 20%

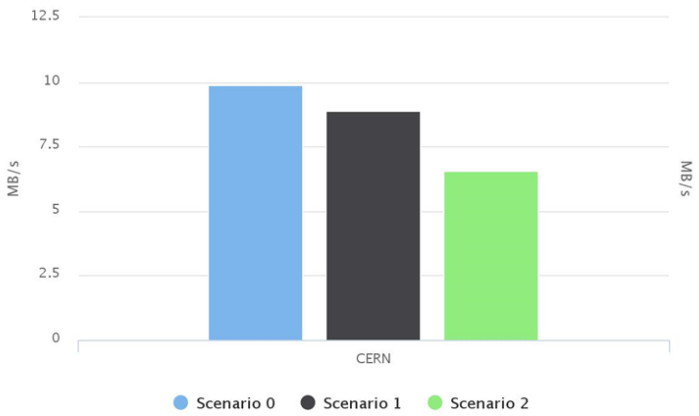

Fig. 5. Populate dataset of 21 ROOT files (~30 GB) from CERN with three scenarios. Plot above shows mean write speed per dataset per scenario.

6.2. ALICE read test

Read procedure is as follows:

Scenario 0: Files are scattered among several file servers;

Scenario 1: All files are on the default file server at NRC "KI";

Scenario 2: All files are on the default file server at NRC "KI" with replicas that may end up on a closest file server.

Test results are shown in figure 6.

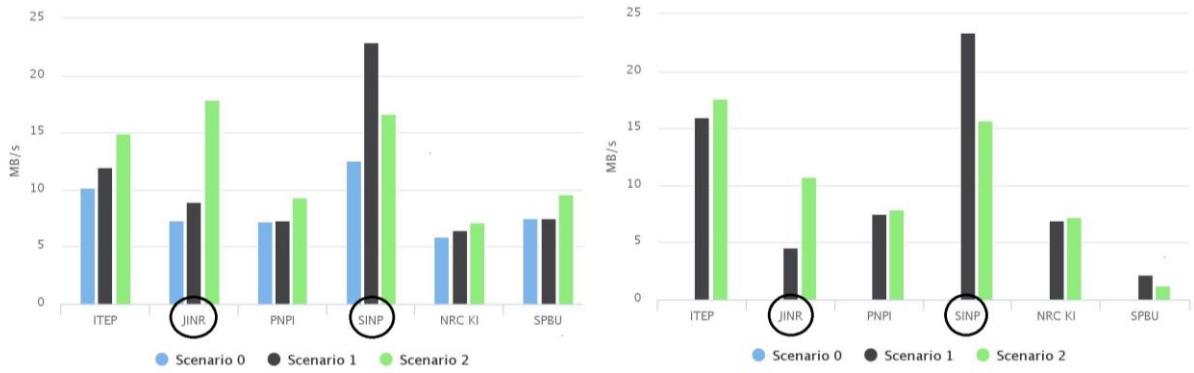

Fig. 6. In the left plot all clients read simultaneously. In the right plot only one client reads data at any given time. Clients are shown on X axis. Marked clients may have files on the closest file server.

Impact of a system load is negligible at this scale.

6.3. Synthetic data placement stress test

Stress test procedure is as follows:

Scenario 0: Files are written to and read from random file servers;

Scenario 1: Files are written to and read from a closest file server if there is one or the default file server at NRC "KI";

Scenario 2: Primary replicas are written as in Scenario 1, secondary replicas as in Scenario 0. Reads are redirected to a closest file server if there is one or to the default file server at NRC "KI".

Test results are shown in figure 7. In contrast with the data population test, here distributed write is a bit slower than write to the default storage. At low transfer speeds replication costs are almost negligible. With many small files there's almost no difference in transfer speed between close and remote dataset, network fluctuations have more impact.

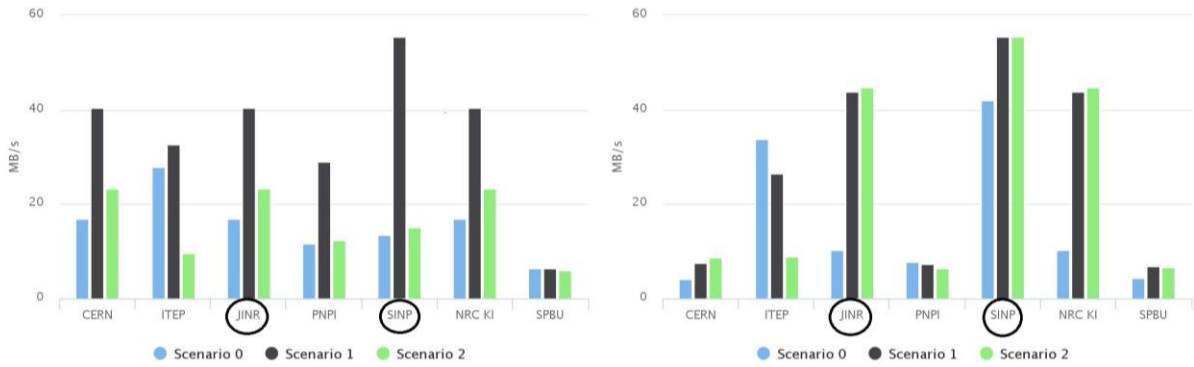

Fig. 7. The left plot is xrootd write stress test, the right plot is xrootd read stress test. Clients are shown on X axis. Marked clients may have files on the closest file server.

7. Conclusion

The first prototype of geographically distributed federated data storage comprising of CERN and RDIG centers has been set up. It was demonstrated that such storage system can be efficiently used for data processing and analysis by the LHC scientific applications.

Performed synthetic and real-life application tests from ATLAS and ALICE have shown a reasonably high performance of the federation mostly limited by the local disk speeds and the bandwidth of network connections.

Bonnie++ tests have shown the expected dependency of metadata access speed on the speed of access to the management server that already speaks in favor of the proposed system, since metadata access times are not bound by the location and availability of the storage servers. Also, file I/O speeds correlate reasonably with the throughput of the network channels. On the other hand, is not yet entirely clear why there's so big difference in the performance of the remote and local clients. In this case, additional testing is considered with further study of configuration parameters.

Taking into account all already obtained test results authors conclude on the applicability of the federated storage for the considered usage scenario.

8. Acknowledgments

This work was funded in part in part by the Russian Fund of Fundamental Research under contract "15-29-07942 офи_м" and U. S. DOE, Office of Science, High Energy Physics and ASCR under Contract No. DE-AC02-98CH10886 and SPbSU under research grant 11.38.242.2015.

References

LHC – The Large Hadron Collider. [Electronic resource]: http://lhc.web.cern.ch/lhc/

The ATLAS Collaboration (G Aad et al.). The ATLAS Experiment at the CERN Large Hadron Collider.// *J. Inst.* **3** S08003 — 2008.

K Aamodt, A Abrahantes Quintana, R Achenbach, S Acounis, D Adamová, C Adler, M Aggarwal, F Agnese, G Aglieri Rinella, Z Ahammed, et al. The ALICE experiment at the CERN LHC. // *J. Inst.* **3** S08002 — 2008

DynaFed. [Electronic resource]: https://svnweb.cern.ch/trac/lcgdm/wiki/Dynafeds

EOS. [Electronic resource]: https://eos.web.cern.ch

Jamie Shiers C. The worldwide LHC computing grid (worldwide LCG). // Phys. Comm. 177:219–223 — 2007

Bonnie++. [Electronic resource]: http://www.coker.com.au/bonnie++/

FUSE. [Electronic resource]: http://fuse.sourceforge.net/

perfSONAR. [Electronic resource]: http://www.perfsonar.net/

dCache. [Electronic resource]: https://www.dcache.org/