Federated data storage and management infrastructure

A.Zarochentsev[1] , A. Kiryanov[23], A. Klimentov[34] , D. Krasnopevtsev[3,5] and P. Hristov[6]

[1]Saint-Petersburg State University
[2]Petersburg Nuclear Physics Institute
[3]National Research Centre "Kurchatov Institute"
[4]Brookhaven National Laboratory, BNL
[5]National Research Nuclear University MEPhI
[6]European Centre for Nuclear Research, CERN

E-mail : alexei.klimentov@cern.ch

Abstract

The Large Hadron Collider (LHC), operating at the international CERN Laboratory in Geneva, Switzerland, is leading Big Data driven scientific explorations. Experiments at the LHC explore the fundamental nature of matter and the basic forces that shape our universe. Computing models for the High Luminosity LHC era anticipate a growth of storage needs of at least orders of magnitudes, it will require new approaches in data storage organization and data handling. In our project we address the fundamental problem of designing of an architecture to integrate a distributed heterogeneous disk resources for LHC experiments and other data-intensive science applications and to provide access to data from heterogeneous computing facilities. We have prototyped a federated storage for Russian T1 and T2 centers located in Moscow, St.-Petersburg and Gatchina, as well as Russian / CERN federation. We have conducted extensive tests of underlying network infrastructure and storage endpoints with synthetic performance measurement tools as well as with HENP-specific workloads, including the ones running on supercomputing platform, cloud computing and Grid for ALICE and ATLAS experiments. We will present our current accomplishments with running LHC data analysis remotely and locally to demonstrate our ability to efficiently use federated data storage experiment wide within National Academic facilities for High Energy and Nuclear Physics as well as for other data-intensive science applications, such as bio-informatics.

1. Overview and description of the problem

The processing, management and analysis of data in the current Mega-Science-scale projects require integration of computing centers of different sizes, power and architecture into a single computing environment (cyberinfrastructure). When designing such an environment one must consider not only the disk and computing resources, but also the bandwidth of the global computing network and the throughputs and data access time between the computing centers. Aristotle asserted that "the whole is greater than the sum of its parts", so the integration of heterogeneous computing centers into a single federated distributed cyberinfrastructure (FDCI) will allow more efficient utilization of computing and disk resources for a wide range of scientific applications.

Existing problems when creating FDCI are as follows:

1. lack of “building blocks” from which you can build such a federation
2. localized solutions, not going beyond immediate use and a specific task (and/or the center), in other words, it is usually an applied solutions with a low abstraction level of interfaces and software modules
3. the integration in FDCI considered only after the creation of computing infrastructure, and not at the stage of development of the overall architecture of the system
4. using the characteristics of the storage and computing resource as the parameters determining the power of the whole computing system, without taking into account the characteristics of global computer networks, data transmission speeds and the possibility of remote access to them

Thus we need to solve four abovementioned problems and provide access to information about FDCI parameters to dynamic data and computing resource management applications.

During the FDCI architecture plot we have to follow the following fundamental principles:

- uniform method and abstraction level of resource management
- common workload and data management system in heterogeneous computing environment
- integrable and expandable tools to manage FDCI software

Altogether, this comprises the problem of a fundamental nature. The lack of adequate solutions so far leads to economic and functional losses. Requirements for such a system are typical for scientific applications in the areas of science that require storage, analysis, processing and data management in petabyte or exabyte range.

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). At first stage a federated disk infrastructure for data storage was plotted out and deployed.

2. Problem description

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 an end user, while in fact being put together form geographically distributed resources. The system should possess the following properties:

- 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 (optimal) data location
- universality, which implies validity for a wide range of research projects of various sizes, including, but not limited to the LHC experiments [1]

3. 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 [3] 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) [4]
- EOS storage system [5]
- dCache storage system [10]

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

4. Federated data storage prototype

For federated storage prototype we have chosen geographically distributed resource centers located at the vertices of a triangle (fig. 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: in each organization one management server (MGM) and one storage server (FST) are deployed. 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 to have 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.

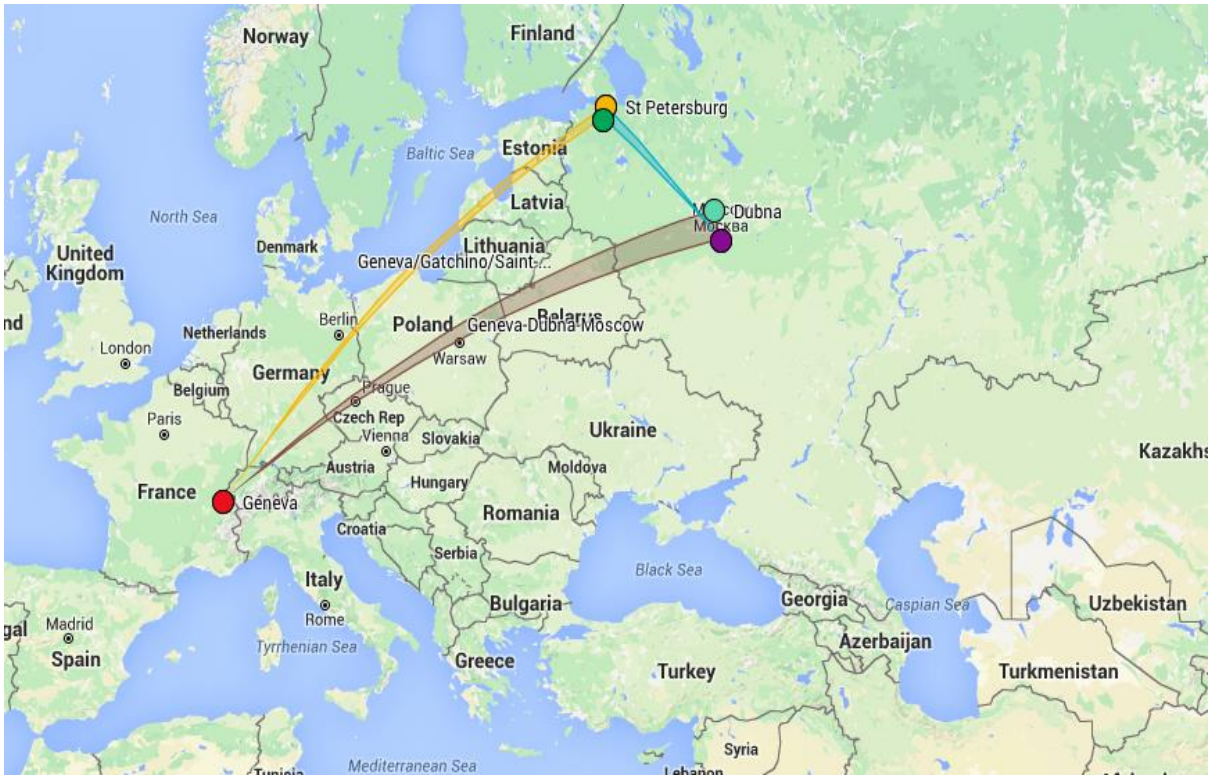


Fig. 1. Participants of the federation prototype.

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.

EOS system includes support for four authentication methods: SSS, Unix, Kerberos and GSI. Unix authentication scheme is not secure enough for WAN use and can only be used within a single data center. SSS scheme is commonly used for authentication between the servers using the shared secret key, while GSI, the most secure one, can be used to authenticate both servers and users via standard X.509 digital certificates. Kerberos scheme is only useful in cases where integration with Kerberos is required.

For the federation prototype the GSI scheme was chosen as the most secure and widely used in the WLCG [6]. GSI authentication implies the need for signed X.509 certificate on each client and server. Our prototype uses certificates signed by Certification Authorities from a standard EUGridPMA list, which is also used in WLCG.

5. 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.

In order to test file I/O performance we used a synthetic Bonnie++ [7] test, which is capable of measuring file and metadata access rate on a filesystem. EOS supports mounting as a virtual filesystem via Linux FUSE [8] mechanism, which makes it possible for Bonnie++ to test both file read-write speeds (FST access) and metadata transaction rates (MGM access) independently.

PerfSONAR [9] suite was used for an 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.

The applicability of such federation for real-life physics experiments was tested with application suites from ATLAS and ALICE experiments.

5.1. 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 xroot protocol.

Upon completion test application produces a comprehensive log file containing information about consumed computing resources on three key stages: environment initialization, event reconstruction and finalization. In addition, during event processing, application records in the log file information about resources consumed during processing of each individual event.

5.2. 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 xroot protocol.

In order to run ATLAS and ALICE test applications we have deployed a standard user interfaces nodes containing all the necessary software stack: user and CA

certificates, EOS client, Bonnie++ for synthetic tests, and LHC experiment software via CVMFS, 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.

6. Test results

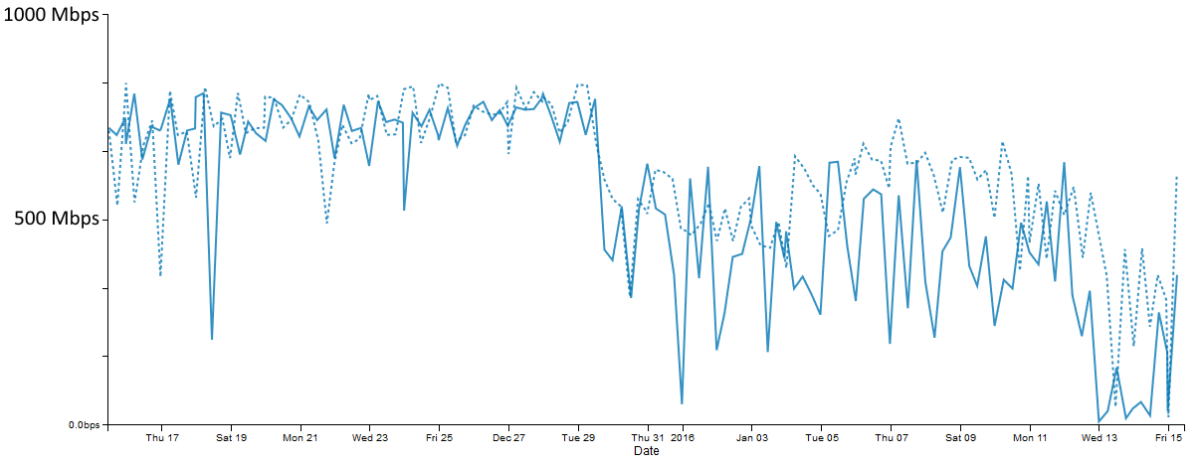
So far we have tested FST and MGM servers, located at CERN, SPbSU and PNPI. We have evaluated the following resource combinations:

- 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)

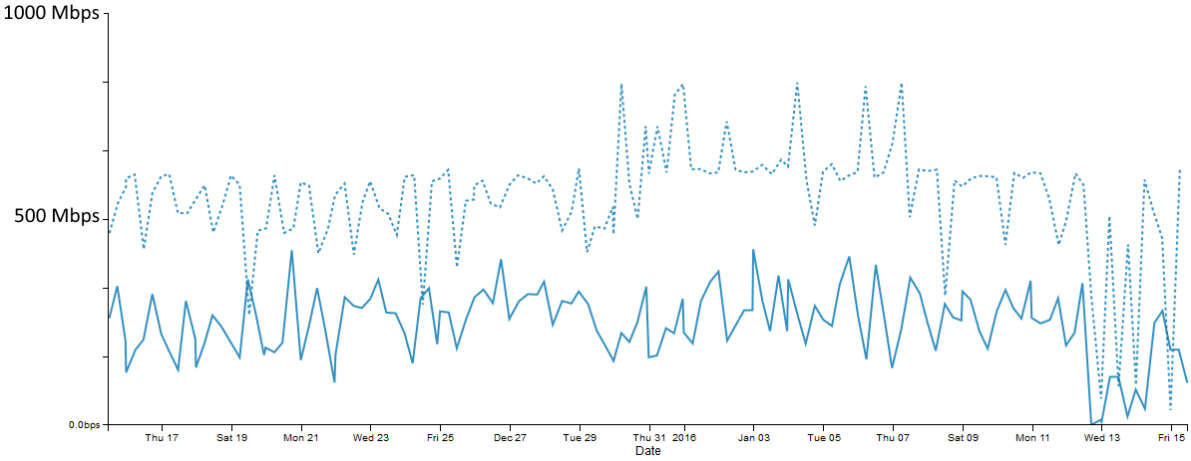
In addition, user interface nodes with at least 4GB of RAM and a gigabit network connection were deployed on all three resource centers composing "the small triangle": CERN, PINP and SPbSU. Tests were performed both before and after consolidation of resources under a single management server. Also, a "hot" role change was tried between MGM without loss of data of accessibility.

6.1. Configuration and testing of the WAN Federation (perfSONAR measurements)

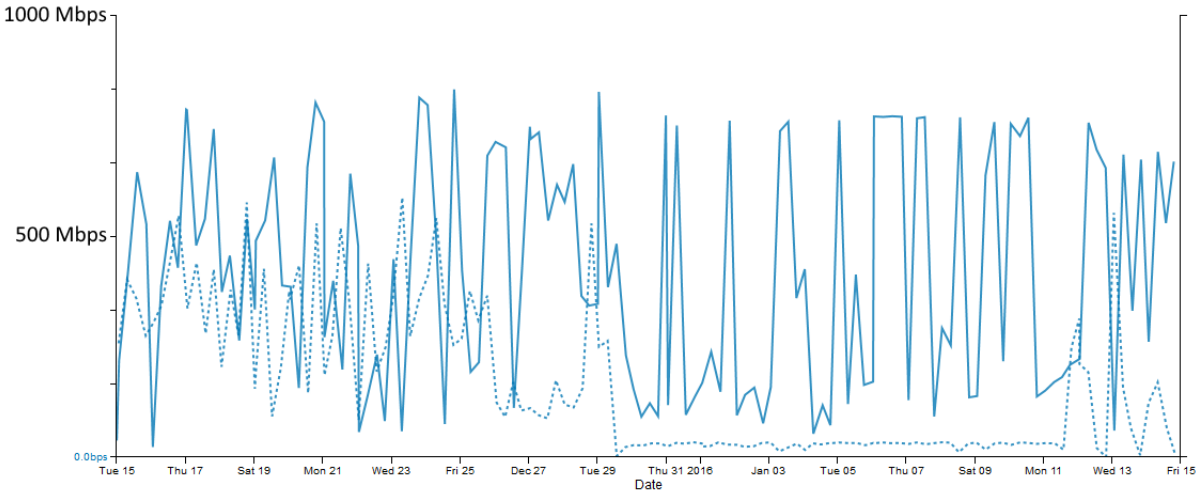
Since the first test results are only available for PNPI, SPbSU and CERN network data channel performance measurements are only of interest between these centers. Network channel conditions were measured with perfSONAR installed on all participating centers of the federation.



PNPI – SPbSU



CERN – SPbSU



PNPI – CERN

Fig.2. Network link bandwidth measured between 15th of December 2015 and 15th of January 2016. Solid and dotted lines represent forward and backward direction respectively.

In fig. 2 we can see that:

- data transfer rates are not symmetric
- communication speed changes significantly with time

This behavior of network channels is due to the fact that 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.

6.2. Bonnie++

For Bonnie++ test the most informative are file (fig. 3) and metadata (fig. 4) 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.

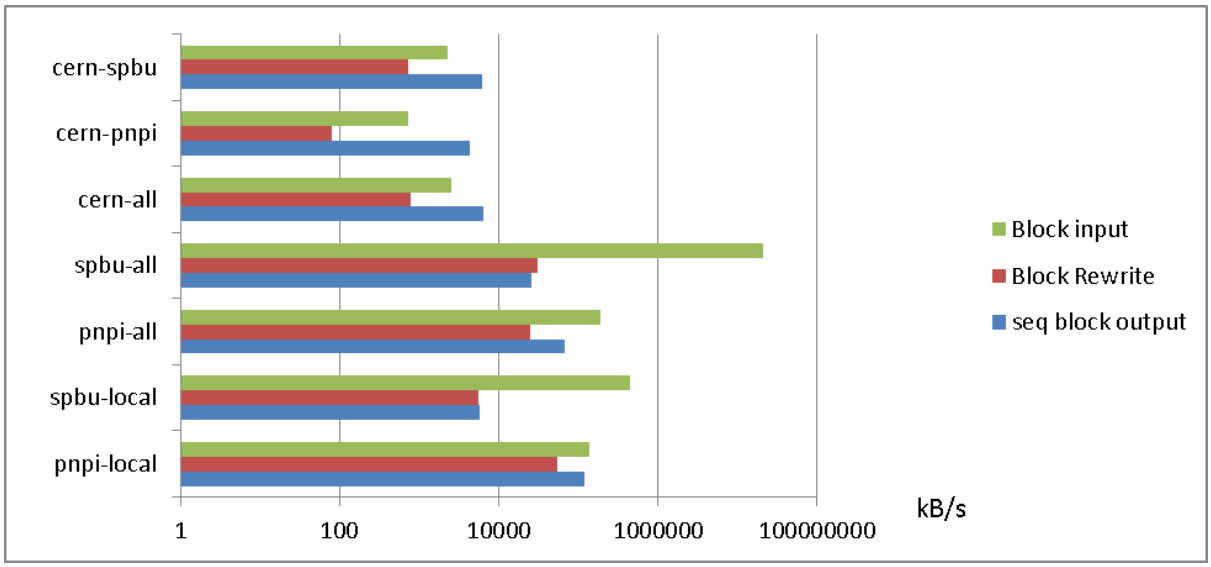


Fig. 3. Bonnie++ test results for block file I/O. On the vertical axis we put a site combination (local means standalone) and test type. On horizontal axis we put a data transfer rate in KBps (logarithmic scale).

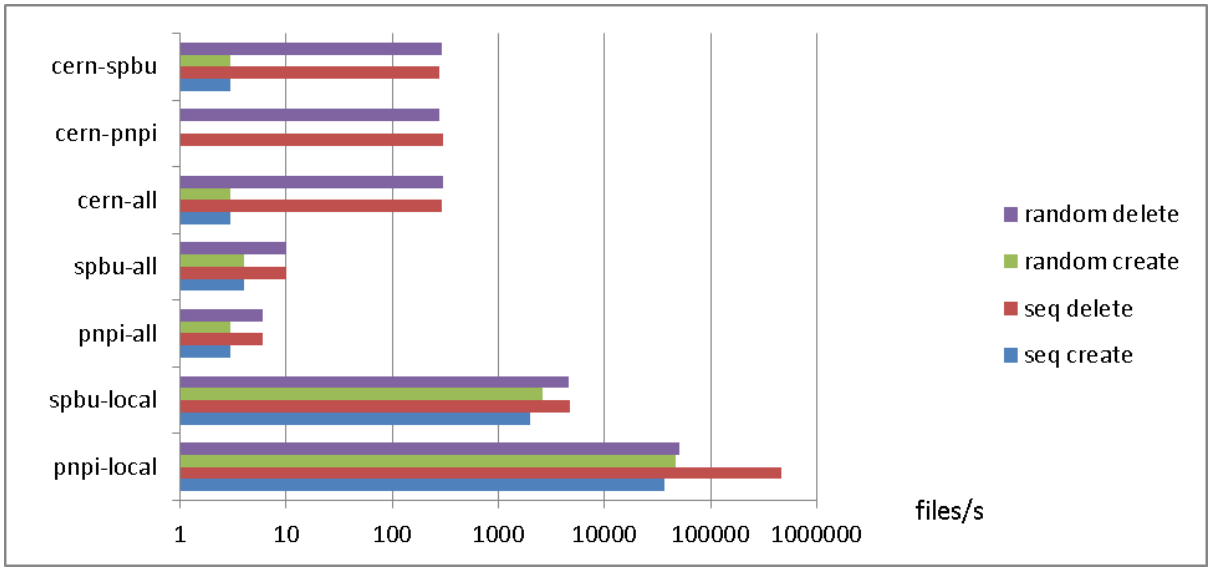


Fig. 4. Bonnie++ test results for metadata transaction rate. On the vertical axis we put a site combination)local means standalone) and test type. On the horizontal axis we put a transaction rate operations per seconds (logarithmic scale).

6.3. ATLAS event reconstruction test

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 xroot protocol. Therefore, in the results of this test we have one additional parameter: access type, which is either FUSE or xrootd. During these tests managers were installed alongside the storage servers, as we were measuring performance of individual storages rather than federation.

The dependence of the test application initialization time on various client-server combinations and data access protocol is shown in fig. 5.

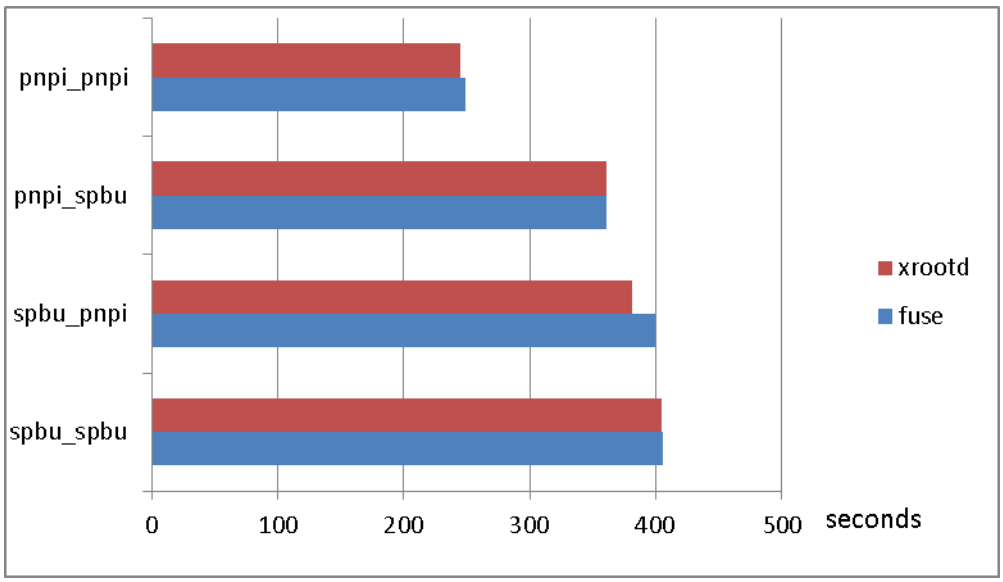


Fig. 5. ATLAS test application initialization time. On the vertical axis we put a client-server combination and data access protocol. On the horizontal axis we put time in seconds.

6.4. ALICE event analysis and retrieval test

These tests were specifically written to test the performance of the federated storage and do not have such a complex output structure as in ATLAS tests. Therefore a complete test runtime measured with a system ‘time’ utility was taken as a primary parameter.

Fig. 6 shows the results of reading from a single FST at PNPI but from different user interfaces and through different MGMs. Tests were performed using the dataset of 25 files with a gross size of ~ 34 GB.

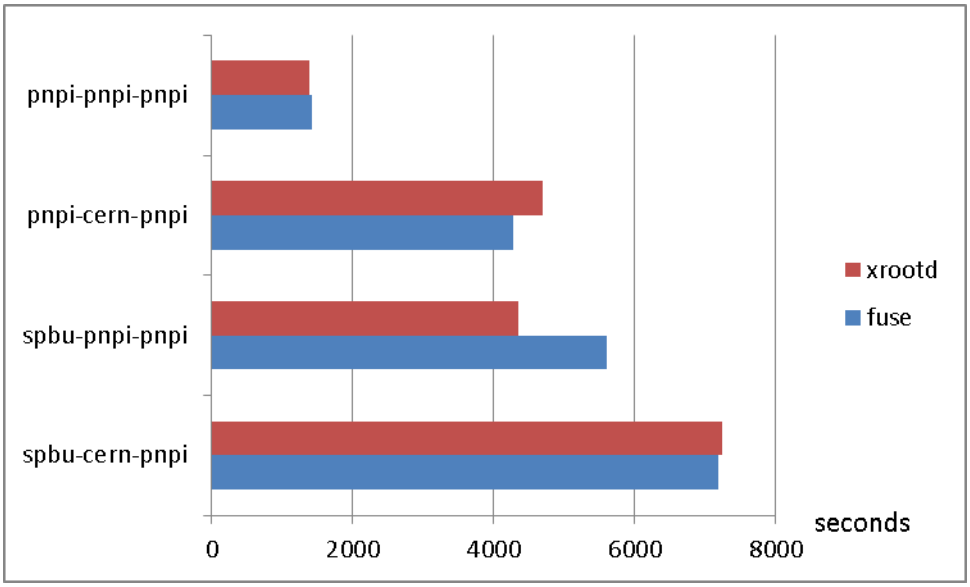


Fig. 4. Runtime of the ALICE test. On the vertical axis we put a UI-MGM-FST combination and data access protocol. On the horizontal axis we put time in seconds.

Noteworthy is the fact that with other things being equal relocation of the MGM server entails a significant (more than twice) difference in runtime. Thus, there is a clear correlation between the ALICE test performance and the geographical location of MGM server.

7. Conclusions

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.

Acknowledgements:

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.

The authors express appreciation to SPbSU Computing Center and PNPI ITAD Computing Center for provided resources.

References:

[1] LHC - The Large Hadron Collider. http://lhc.web.cern.ch/lhc/

[2] The ATLAS Experiment at the CERN Large Hadron Collider. The ATLAS Collaboration (G. Aad et al.). JINST 3 S08003 doi: 10.1088/1748-0221/3/08/S08003 (2008)

[3] Kenneth Aamodt, A Abrahantes Quintana, R Achenbach, S Acounis, D Adamová, C Adler, M Aggar- wal, F Agnese, G Aglieri Rinella, Z Ahammed, et al. The ALICE experiment at the CERN LHC. *Journal of Instrumentation,* 3(08):S08002, 2008.

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

[5] EOS https://eos.web.cern.ch

[6] Jamie Shiers. The worldwide LHC computing grid (worldwide LCG). *Computer physics communica- tions,* 177:219–223, 2007.

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

[8] FUSE http://fuse.sourceforge.net/

[9] perfSONAR http://www.perfsonar.net/

[10] dCache https://www.dcache.org/