International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.11 (2015), pp.285-292
http://dx.doi.org/10.14257/ijmue.2015.10.11.27

Storage Federations Using Xrootd

Sang Un Ahn[1], Hee Jun Yoon[2] and Sang Oh Park[3,*]

[1,2,3]Korea Institute of Science and Technology Information, South Korea
[1]sahn@kisti.re.kr, [2]k2@kisti.re.kr, [3]sopark@kisti.re.kr

Abstract

In this paper, we introduce a local storage system using Xrootd in which the heterogeneous storage types are combined into a single access point that help clients access to the storage system efficiently without knowing complex architecture underneath. This system is expanding into global storage union federating the storage systems of the WLCG. The WLCG, a collaboration of computing centers to deal with large scale of data produced by the detectors using the LHC at CERN, has distributed storage resources with different protocols and management systems. The heterogeneity of the distributed storage resources prevents efficient access to data and transfer activities between computing centers. The demand on federating storage resources into one single access point has been arising within the WLCG. We introduce the current movement to federating storage resources has been started and expanding into the whole WLCG collaboration upon the successful federation and operation carried by the computing centers supporting ALICE experiment.

Keywords: *Xrootd, WLCG, Storage Federations*

1. Introduction

The WLCG [1] (Worldwide LHC Computing Grid) is an international collaboration of about 150 computing institutes implementing the grid technology to process, store and preserve the data produced by the experiments using the LHC [2] (Large Hadron Collider) at CERN (European Organization for Nuclear Research) in Geneva, Switzerland.

The LHC is the largest particle accelerator in the world; the circumference of the LHC ring is 27km long. It is designed to accelerate proton or lead ions and collide them at the locations where four giant detectors resides. The detectors are: ALICE [3] (A Large Ion Collider Experiment), ATLAS [4] (A Toroidal LHC Apparatus), CMS [5] (Compact Muon Solenoid) and LHCb [6] (LHC beauty experiment). The aim of each detector (in general physicists call it experiment) is different from each other. In principle ATLAS and CMS have the same goal to study the elementary particles forming matters and propagating forces. LHCb is designed to research properties of beauty (or bottom some other physicists call) particle (one of the elementary particle but having special characteristics); ALICE is designed to detect particles from the collisions of the lead ions to research the status of the early Universe just after the Big Bang.

These detectors are equipped with dedicated electronics for fast response and high granularity to catch as many particle tracks as possible. Designed goal of data acquisition (DAQ) for each experiment is different from one another; approximately 1 GB/s on average. The expectation on the amount of annual data taking says 15 PB per year. In fact, the amount of data aggregated on tape media at the CERN Tier-0 center has 100 PB in 2013.

The WLCG is the most successful example of the grid computing and its excellent performance has been proven by the discovery of the new particle called Higgs boson [7-

[*] Corresponding Author

ISSN: 1975-0080 IJMUE

8]. The WLCG has tiered structure: Tier-0, Tier-1s, Tier-2s and so on. There are two Tier-0 centers in the world: one at CERN Computing Center, the other one at Wigner Center in Budapest, Hungary. They are connected with 100 Gb/s link of network circuit. The Tier-0 center plays a role as the host center of the WLCG. It receives and archives the raw data produced from the detectors at the LHC on the fly and then it performs first processing the raw data to convert digital signals from the sensors into a collection of meaningful tracks of particles (Reconstruction) and distributing the raw data to the members of the WLCG. For Tier-1, there are 12 computing centers. They are full members of the WLCG who do their duties: receiving and archiving a portion of the raw data on their tape from Tier-0, performing the reconstruction process and taking part in the mass analysis activities as well. There are many Tier-2 centers, about 140 centers, and Tier-3 or individuals which mainly perform data analysis and simulation data generation.

The most important thing of the data processing capability of the WLCG is to enable efficient data transfer among computing centers and efficient data access by user analysis task jobs or data management. However, heterogeneous storage systems deployed at the computing centers of the WLCG could make the activities concerning data by users or administrators inefficient. This implies that different protocol or interface tool is required each time users want to access the data at the remote location where various storage systems are installed. The possible solution for this problem is Xrootd9 which is able to enable a single point access to the data upon the heterogeneous storage system because it has a simple pluggable and scalable architecture.

In this paper, we present a local storage system using Xrootd in details in Section 2 and we introduce the current trend on the activities for federating storage recently arising in the WLCG in Section 3. Finally, we conclude, in Section 4, this paper mentioning about the limitation on the functionality of Xrootd and the future study to improve its performance.

2. Local Xrootd Storage System

In this section, we present a simple Xrootd storage system and discuss its architecture as well as some strong points of using Xrootd as the storage system. And setting up hierarchical storage system using Xrootd is described and expanding this to local application for storage federation is discussed.

2.1. Architecture of A Simple Xrootd Storage System

A storage system using Xrootd can consist of Xrootd redirector (or header) and at least one Xrootd server. Xrootd servers subscribe to the redirector. There is no limitation on the number of nodes to be subscribed. The header redirects all requests concerning data access or data management to the corresponding servers; the actual services (data access or data management) are carried out by the server itself who received the request. Therefore Xrootd storage system is able to process the requests in parallel without much of loads on the redirector. In principle, the more Xrootd servers are introduced, the better performance of the storage system could be achieved.

Figure 1 shows a schematic view of a simple local Xrootd storage system. As stated above, when a request is submitted to the Xrootd redirector by a user or by an analysis script for data access or by an administrator for data management, the Xrootd redirector scans a namespace of the system. The namespace has cached information about the location of data where the data actually is stored on which Xrootd servers. At first, the Xrootd redirector scans the namespace to find the location of the data and returns the result to the client. Thus, the namespace is used as a reference when a user request is placed to query the location of the data file. The physical location of the data is unknown to the client; the client is using the logical file name (provided as a catalogue accessible via the grid) which is human-readable is used for the request.

Once the client requests Xrootd redirector to access the data, the redirector locates where the data is stored by using the cached information and redirects the request to the relevant server. If the Xrootd redirector fails to find any information concerning the request, it will send queries to the servers via CMSd (Cluster Management Service daemon). The Xrootd server who receives the request then serves the request directly to the client. Any activities concerning data access or data management are provided via Xrootd protocol.

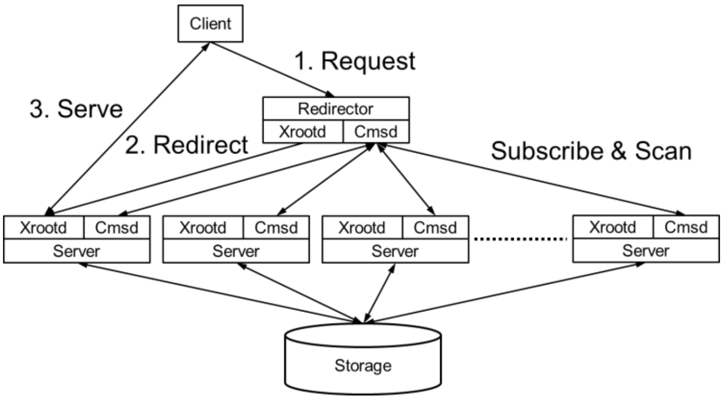

Figure 1. A Schematic View of a Simple Local Xrootd Storage System

2.2. Advantages of Xrootd as a Storage System

In the following, we list three strong points of using Xrootd as a storage system:

1) Common tool used for data analysis in high energy physics, which is called ROOT10, provides an interface to Xrootd so that users or administrators can easily access data on the Xrootd storage system without changing tools or learning additional tools.

2) RDBMS (Relational Database Management System) is not needed for data management. Typically, one can easily imagine that using RDBMS requires a high performance machine in order to deal with huge number of queries for data access or data management. As Xrootd use a single namespace instead of RDBMS for data management, one can reduce the cost for setting up storage system and assure the performance of the storage system with a mid-range server.

3) Pluggable architecture of Xrootd makes setting up a single access point on top of the heterogeneous storage system possible. There are plugins for almost all of currently being used for the data management deployed in the WLCG. For example, a proxy, which is called dCap, is developed for dCache11 storage system which is widely used in the big computing centers. This is why Xrootd is currently being highlighted as the key for federating heterogeneous storages over the world. This is discussed in detail in Section 4.

In addition, as mentioned already some in Section 2.1, Xrootd has other strong points: scalable (and low-cost) for expanding the system, lightweight (low-load on redirector as well as on servers), easy installation and configuration for administration.

2.3. Hierarchical Storage System Using Xrootd

Archiving data and preserving them for decades are one of the most important duties of the WLCG. Traditionally physicists have used tape media for data preservation rather than hard disks. Yet still disk storage is required as cache before the data to be written on the tape because reading or writing on tape is not much slower than disk (or rather than faster) but it is sequential, and also the architecture of tape libraries is not suitable for runtime operation (tape medium should be searched and loaded on tape drive before reading or writing which might take minutes in any case all tape drives are in operation). In this case, sending data from disk cache to the backend storage (tape) and retrieving the data written on tape by any request should be done dynamically upon the management policy.

In order to implement the hierarchical storage system, Xrootd provides FRM (File Redundancy Management) as an interface to the backend storage. Figure 2 shows a schematic view of a hierarchical storage system using Xrootd. Once any data is migrated to the backend storage by FRM, Xrootd does not keep the information about the data in the namespace any more. When any requests on the data migrated is made, the Xrootd redirector, which does not have any information on that, sends a query to any available Xrootd server for the data then FRM reacts upon the request. In fact, since the data does not exist on the server, FRM holds the request and makes the data to be staged (copied) from the backend storage (call-back process). After the call-back process is finished, the complete file is available for data access by the client.

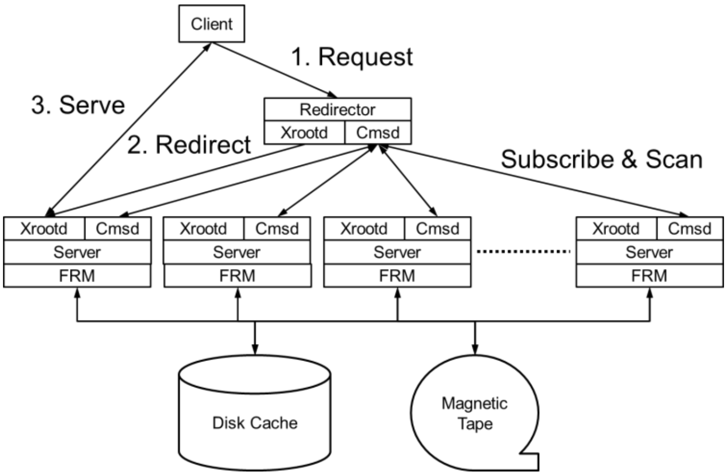

Figure 2. A Schematic View of a Hierarchical Storage System Using Xrootd

FRM also makes storage-less system possible for those who do not have enough storage resources such as Tier-3 or individuals. This can be done by having FRM point to the Xrootd redirector at larger computing center.

2.4. Local Storage Federation Using Xrootd

As an outcome of the discussion in the previous sections, in principle, setting up a single service point on top of the heterogeneous storage system (or file system) is possible locally via Xrootd.

In Figure 3, we present a schematic view of a storage system consisting of different type of data management pools together with a hierarchical storage system using magnetic tape library. Note that we simplify the complexity of the Xrootd architecture for the example. In this case, two Xrootd servers are interfaced with the other data management pools with plugins. These data management pools can have any kind of storages managed by their own: disk storage or tape. The other two Xrootd is interfaced with tape management server and the other Xrootd storage system out of the local domain via FRM. To make a single access point for this storage system, we present a global redirector here.

In such a way, it is obvious that a client can access the data without knowing about the internal architecture or the file system of the storage and the backend storage. Providing a single access point to clients or administrators make the data access and the data management efficient. Because of the scalable architecture of the Xrootd storage system, the load on the redirector is balanced.

3. Storage Federations in the WLCG

In this section, we present an example of the federated storage in the WLCG: the ALICE uses a case and the current setup of the storage system that we provide for ALICE. And then, we briefly introduce the current trend of the storage federations in the WLCG by exploiting Xrootd.

3.1. Federated Storage of ALICE

The WLCG is the collaboration of the computing center that provides their computing resources to the experiments using the LHC. The provisioning and the installation of the storage system of each computing centers is different from one another. And also it depends on the computing model of the experiment which the computing center is associated with. The storage federation has been already done and has been well performed by the ALICE collaboration8.

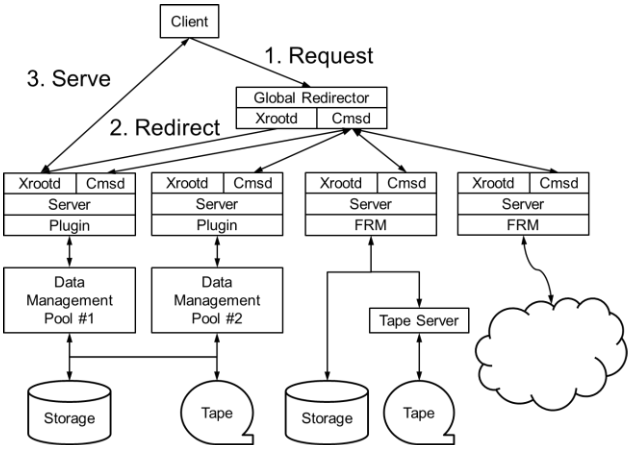

Figure 3. A Schematic View of the Local Application for Storage Federations Using Xrootd

We provide our computing resources to ALICE experiment as a Tier-1 computing center of the WLCG and the computing model of ALICE experiment is to have distributed storages federated as a single storage to be shown to users (mostly physicists). ALICE experiment introduces Xrootd as a standard protocol for the storage. They also provide a global redirector where all of the Xrootd storages distributed over the global can subscribe to it.

We have two different storage domains: disk domain and tape domain. Disk domain is a typical storage system using Xrootd; tape domain is a hierarchical storage system interfaced with tape management server via FRM. For disk domain, we have one Xrootd redirector and 9 Xrootd servers are subscribed to the redirector. The storage we use for the disk domain is a SAN type of storage. We deploy 10 Xrootd servers and one Xrootd redirector for tape domain. FRM is enabled on all of the Xrootd servers to perform the migration of data to be archived in the backend storage: in our case, tape media. Between FRM and tape, there is tape management server that carries out scheduling of data I/O from tape upon the operation policy. And also in order to achieve better performance on data migration from FRM to tape, we introduce a parallel file system on tape management server. For the recall, tape management server keeps all information about the state of data file. A schematic view of our setup for the storage system using Xrootd is shown in Figure 4.

The tape domain is responsible for receiving the raw data flow on the fly: data flow from the detector produced by the collision of particles. The network bandwidth is important factor in order to achieve this. Since the data generation rate is 1 GB/s as mentions, the connectivity between CERN and our site has to be sufficient bandwidth: 10 Gb/s for minimum. In fact, to have a dedicated link to CERN and to be included within a private circuit for the WLCG Tier-1s is one of the most crucial requirements. And also all the hardware used for Xrootd nodes have to be equipped with 10 GbE card.

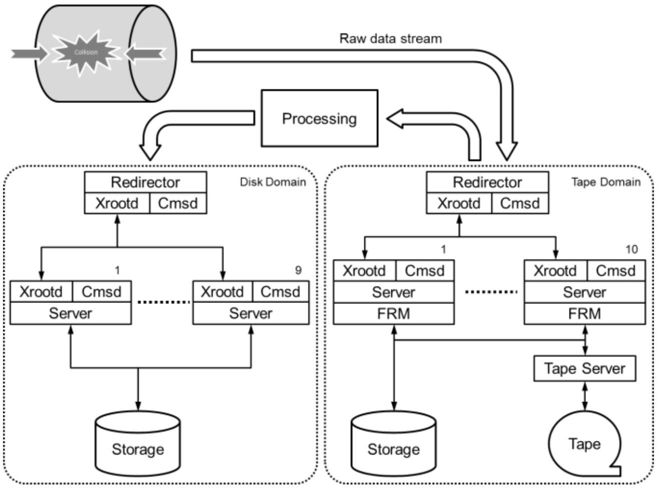

Figure 4. An Example of the Storage System Using Xrootd for ALICE Experiment

Once raw data files are archived in tape, heavy processing activities to use the raw data are foreseen. In order to prevent frequent recall or reduce delay time due to recall, the most frequently accessed data files are cached in the disk storage of the tape domain. After the processing of the raw data, the output of the processing is stored on the disk storage. Depending on the priority of the data produced after the processing, some of data can be archived back to the tape, however, mostly resides in the disk storage.

Centrally, the ALICE point of view, there are two different storage points as explained above: disk and tape. The disk storage is open to public within the collaboration. The user produced data is automatically stored on the nearest storage where the user is seated and the data is accessible wherever the user now is. However, the tape storage is used for archiving only. Because of the explicitly different purpose of disk and tape storage, in general, they are not federated into one single access point. However, within the same storage domain, ALICE establishes a federated storage system by using Xrootd.

3.2. Current Trend of the Storage Federation

Contrary to ALICE case, computing centers of the WLCG supporting other LHC experiments such as ATLAS and CMS have been in different situations. Although there are definitely active data access activities and data transfer rates among the computing centers that support ATLAS or CMS, data access and data transfer are not centrally organized. Since there is no single access point to the storages, a list of the storage location and its type has to be kept and well maintained.

Recently Xrootd storage system is considered as a solution to solve the inefficient data access problem caused by the heterogeneous storage systems deployed within the WLCG. Member centers of the WLCG have different implementations of the storage system depending on their technical preferences or by the requirement of the experiment they support: for example, dCache, CASTOR, SRM, and Xrootd. In order to access the data storage on such various storage systems, the clients are asked to implement different protocols corresponding to the relevant storage systems into their applications. The direct access to the data via dedicated data transfer tools usually requires the physical location of the data. For client side, unnecessary effort to research the physical location is foreseen.

The idea to resolve such an inefficient situation is to federate the heterogeneous storage systems by deploying Xrootd with the appropriate plugins upon the existing storage system and to subscribe them to a virtual global Xrootd redirector with the global namespace in order to provide a single access point to the clients. This is called storage federations. As the local Xrootd storage system, the clients over the world can ignore various site-specific implementations for storage system and can access to the data with the logical file name to be matched at the global namespace to locate the physical data.

Recently ATLAS and CMS collaboration started federating their grid storage using Xrootd; they are called FAX (Federated ATLAS Xrootd) system and AAA (Any Data, Any Time, Anywhere) system, respectively12. The current FAX system consists of the virtual global redirector hosted by BNL (Brookhaven National Laboratory, U.S.) where Tier-1 center for ATLAS is placed and several Tier-2 centers in the U.S. The federated storage system for CMS is organized by the Tier-1 center at FNAL (Fermi National Accelerator Laboratory, U.S.) and the Tier-1 center at RAL (Rutherford Appleton Laboratory, U.K.). The WLCG is currently encouraging the member centers for both collaborations to deploy Xrootd for expanding the federations.

4. Conclusions

Xrootd is a simple but powerful and scalable data management system that provides various plugins to support heterogeneous storage system. In this paper, we presented a local Xrootd storage system and also a hierarchical storage system by using FRM. Especially we provide a federation of distributed storage or different type of storages by exploiting the functionality of FRM. The computing centers of WLCG supporting ALICE experiment have been successfully federating their distributed storages. As an example, we presented our current storage setup using Xrootd.

Upon the successful federation and operation done by ALICE, the other computing centers in the WLCG supporting other experiments such as ATLAS and CMS are now federating their storage resources using Xrootd. Active movements are now performed within the U.S.

Acknowledgments

This work was supported by the program of the Global hub for Experiment Data of Basic Science, 2015 funded by the NRF (N-15-NM-CR01-S01).

References

[1] Bird, "Computing for the Large Hadron Collider", Annual Review of Nuclear and Particle Science, vol. 61, **(2011)**, pp. 99-118.

[2] L. Evans and P. Bryant, "LHC Machine", Journal of Instrumentation, S08001, vol. 3, **(2008)**.

[3] The ALICE Collaboration, "The ALICE experiment at the CERN LHC", Journal of Instrumentation, S08002, vol. 3, **(2008)**.

[4] The ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", Journal of Instrumentation, S08003, vol. 3, **(2008)**.

[5] The CMS Collaboration, "The CMS experiment at the CERN LHC", Journal of Instrumentation, S08004, vol. 3, **(2008)**.

[6] The LHCb Collaboration, "The LHCb Detector at the LHC", Journal of Instrumentation, S08005, vol. 3, **(2008)**.

[7] The ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", Physics Letters B, pp. 1-29, vol. 716, no. 1, **(2012)**.

[8] The CMS Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”, Physics Letters B, vol. 716, no. 1, **(2012)**, pp. 30-61.

[9] A. Dorigo, P. Elmer, F. Furano and A. Hanushevsky, “XROOTD-A Highly scalable architecture for data access”, WSEAS Transactions on Computers, vol. 4, **(2005)**, pp. 348-353.

[10] R. Brun and F. Rademakers, “ROOT – An object oriented data analysis framework”, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 389, no. 1-2, **(1997)**, pp. 81-86.

[11] A. P. Millar, T. Baranova, G. Behrmann, C. Bernardt, P. Fuhrmann, D. O. Litvintsev, T. Mkrtchyan, A. Petersen and A. Rossi, “dCache, agile adoption of storage technology”, Journal of Physics: Conference Series, 032007, vol. 396, **(2012)**.

[12] L. Bauerdick, D. Benjamin, K. Bloom, B. Bockelman, D. Bradley, S. Dasu, M. Ernst, R. Gardner, A. Hanushevsky, H. Ito, D. Lesny, P. McGuigan, S. McKee, O. Rind, H. Severini, I. Sfiligoi, M. Tadel, I. Vukotic, S. Williams, F. Wurthwein, A. Yagil and W. Yang, “Using Xrootd to Federated Regional Storage”, Journal of Physics: Conference Series, 042009, vol. 396, **(2012)**.

Authors

Sang Un Ahn, received his B.S., M.S., and Ph.D. degrees from the Department of Physics at Konkuk University, in 2007, 2009, and 2012, respectively and Ph.D. from the Department of Subatomic Physics at University Blaise Pascal, in 2011. He has been serving as a Senior Researcher of Global Science experimental Data hub Center at Korea Institute of Science and Technology Information since 2012. His research interests are Linux system management, batch system, and grid system.

Hee Jun Yoon, received his M.S from the Computer Engineering Department at Chungnam National University, KR in 1997. He has been serving as a Senior Researcher at Korea Institute of Science and Technology Information since 2000. His research interests include Large data processing, Parallel Computing, Cyber-Infra system, and Computer Education.

Sang Oh Park, received his B.S., M.S., and Ph.D. degrees from the School of Computer Science and Engineering at Chung-Ang University, in 2005, 2007 and 2010, respectively. He has been serving as a Senior Researcher of Global Science experimental Data hub Center at Korea Institute of Science and Technology Information since 2012. He served as a Research Professor at Chung-Ang University. His research interests include big data system, tape storage system, embedded system, cyber physical system, home network, and Linux system.