Distributed Storage System @ CERN

Written by: *Florian Akos Szabo*

Email: florian.akos.szabo@gmail.com

For the Clouds, Clusters and Grids seminar at ITMO University, Saint Petersburg, Russia

7th of June, 2018

Intdoduction

CERN is an acronym that stands for the European Organization for Nuclear Research (CERN) and it is the world laboratory for high energy research located in Geneva, Switzerland. CERN is the birth place of many important discoveries, including several Nobel physics prizes and the World Wide Web. CERN has been one of the world's largest and most respected laboratories for scientific research. There are ~10,000 scientists of ~600 institutions from ~100 countries currently doing their research at CERN.

CERN is home to the multi-billion dollar project called the Large Hadron Collider, which is one of the most ambitious and exciting scientific projects in human history. The goal of LHC is to search for new physics beyond the current known physics frame work. Scientists' expectations for discoveries at the LHC include the Higgs particle which is responsible for the origin of mass, dark matter, extra dimensions, and many other new physics scenarios. The ATLAS (A Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) are the flagship LHC experiments designed to search for such possible new physics. On July 4, 2012, ATLAS and CMS experiments announced the discovery of a new particle which is consistent with the Higgs boson (so-called "God" particle). This important discovery made headline news in all major news media around the world and was named by Science magazine as the 2012 "Breakthrough of the Year" and resulted in the award of the 2013 Nobel Prize for physics to Dr. Peter Higgs and Dr. Francois Englert, the two theoretial physicists who developed the Higgs theory in the 1960's. [1]

EOS is a software solution that aims to provide fast and reliable multi-PB disk-only storage technology for both LHC and non-LHC use-cases at CERN. The core of the implementation is the XrootD framework which provides feature-rich remote access protocol. [2]

EOS is an open source distributed file system developed at CERN. The EOS file system currently houses more than 140 petabytes of data, and the LHC experiment generates up to 10 gigabytes of data per second of run time. As part of CERN’s mission to study the basic constituents of matter, the organization stores and shares the data with partners around the globe using EOS solution.

The EOS project was started in April 2010 in the IT DSS (Date & Storage Services) group. The first release version 0.1 was put into production in June 2010. Since release version 0.1.1 EOS has been available under GPL v3 license. Five EOS instances are currently run at CERN and one at FERMILAB, based on the 0.3 Beryl releases.

Main Features

A fundamental concept of EOS is to use a set of single disks (JBOD) as storage media without the need to build local hardware RAID arrays. All storage nodes are divided into groups and within one group files are placed using RAIN (Redundant Array of Independent Nodes) algorithms.

EOS provides so-called file layouts. These describe how files are structured on the storage system. The most basic type is a plain file, a bit more advanced is a replicated file, which means it is copied to n other instances, and the most advanced is a rain file (Redundant Array of Independent Nodes) which is reed-solomon encoded for the most important files that cannot be corrupted (contains error correcting parity bits).

Why the need for DSS?

The current approach to storage does not work anymore: it is not flexible enough, fast enough or the cost is prohibitively high. By design a distributed storage system solves all of these issues at once.

EOS Architecture

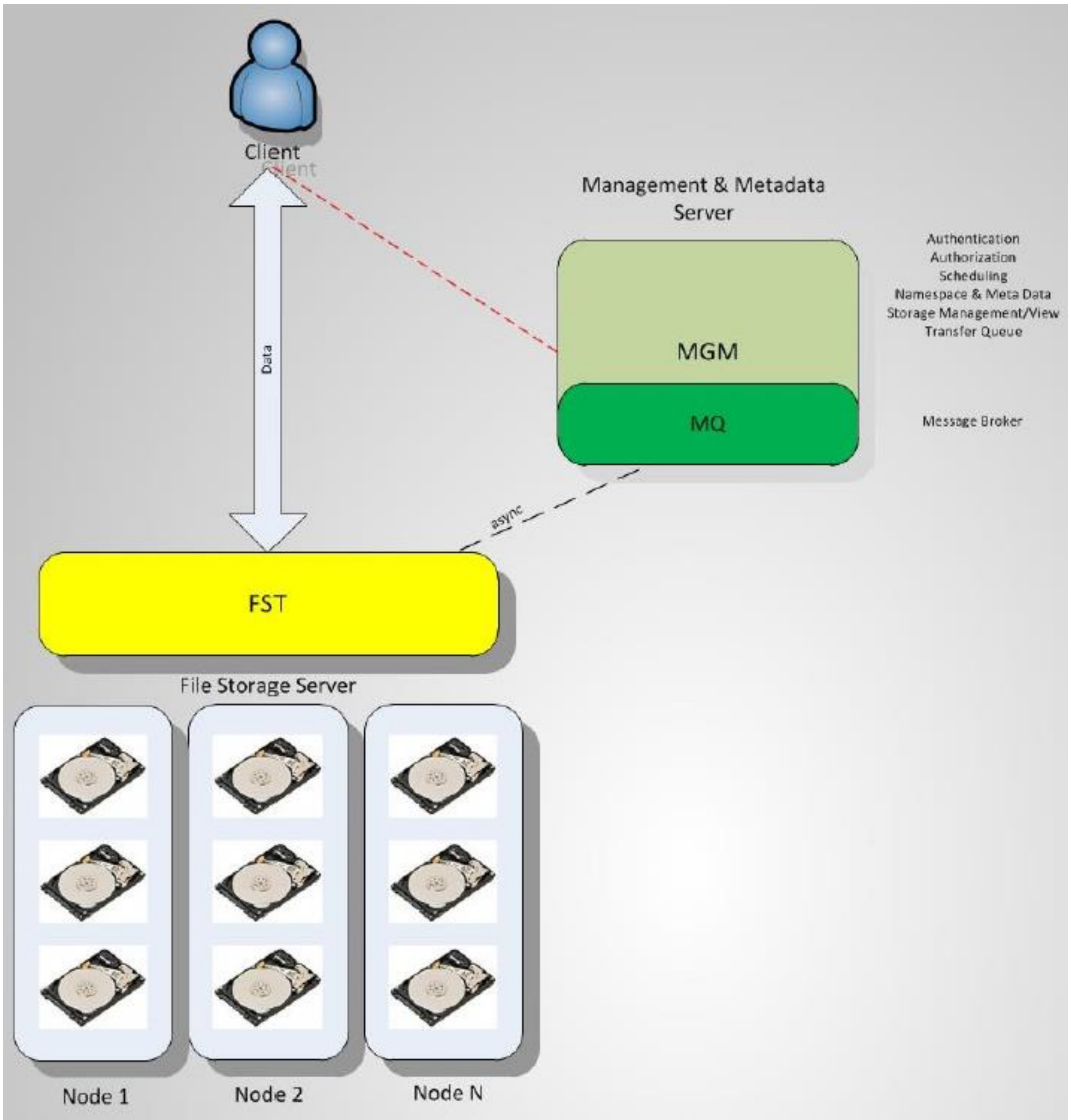
On the below picture, the high-level architecture of the EOS Distributed Storage System can be seen. There are three main components:

- MGM stands for Meta Data Management Server (sometimes called MD)
- MQ stands for Message Parsing / Queueing Server
- FST stands for File Server

EOS's architecture has been designed for simplicity in order to allow easy administration of the instances and to keep the latency at very low levels. An EOS instance is composed of three types of components. It is important to note that EOS separates the IO path into metadata access via a metadata service (MGM) and data access via file storage service (FST). To ensure low latency, all metadata about files is kept in memory on MGM nodes (which require high memory capable nodes).

The MGM node is very important and is responsible for a variety of tasks:

- Authentication, identity management
- file permissions and quota enforcement
- in-memory namespace management
- file placement and access
- instance management
- load/node balancing



The MQ components are responsible for managing the messaging between FST and MGM nodes in an asynchronous manner. MQ is usually run on the same nodes where the MGM is running.

The FST the the place where disks are attached for storing the files and replicas. It also handles requests to access files and for storing new ones from clients. FST nodes are also in charge of calculating file checksums to ensure integrity and mitigate errors. Typically, the FST component is running on many instances to accommodate the varying number of dks storage components.

EOS at CERN

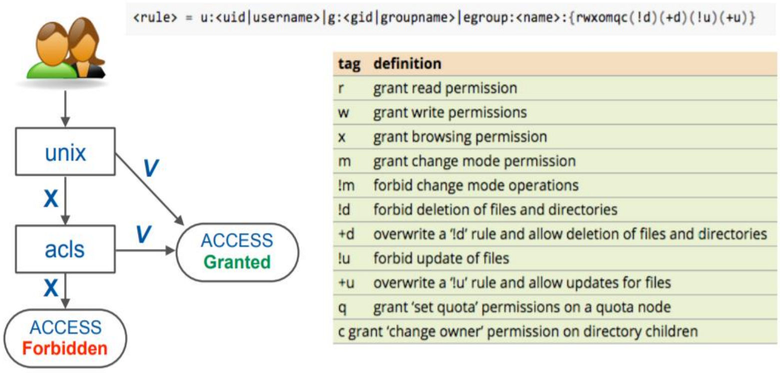
Since 2011, when it was initially deployed, it has grown considerably at CERN. It is the default destination for much of the data that is generated from various physics experiments and measurements. By 2015 it has grown to the below proportions:

- 140 Petabytes of Data
- 1400 servers
- 44.000 hard disks
- 271 million files in 26 million directories
- 3000 users

Protocols and Client implementations

The primary native protocol for communications in EOS is XrootD [3]. It is used for file IO, internal communication, replication, CLI. Secondary protocols are gridFTP, HTTP, HTTPS, WebDaV, S3.

Strong authentication is handled by Kerberos for local area and X.509 for remote access. An optional Simple Shared Secret scheme is also avialable.



```
<rule> = u:<uid|username>|g:<gid|groupname>|egroup:<name>:{rwxomqc(!d)(+d)(!u)(+u)}
```

tag	definition
r	grant read permission
w	grant write permissions
x	grant browsing permission
m	grant change mode permission
!m	forbid change mode operations
!d	forbid deletion of files and directories
+d	overwrite a '!d' rule and allow deletion of files and directories
!u	forbid update of files
+u	overwrite a '!u' rule and allow updates for files
q	grant 'set quota' permissions on a quota node
c grant 'change owner' permission on directory children	

Extensible ACLs can be used in combination with UNIX file permissions to enable and enforce user/group/project level quotas across the whole EOS domain.

Scalability and Data Integrity

The described namespace implementation can scale to approx. 100 million files using 80 - 100 GB of memory. The boot time for such a namespace is around 10 minutes. Assuming an average file size of 200 MB/File and 2 replicas the total space of a storage pool would be 40 PB.

To ensure integrity for the storage system, it's important to add a checksum verification field to every chunk of data so that its integrity can be verified, independently from the error correction. Checksumming is an essential requirement of any storage service. It measures the ability to ensure that for some reasons the data stored has not involuntarily changed due to hardware failure, software bugs, human error or during an I/O operation like reading, writing, moving, copying. Loss of data integrity can be the worst nightmare for a storage systems engineer / administrator.

Recommendation

EOS is recommended for any organization or company that…:

- Needs large scale distributed storage system
- Is on a low budget (no commercial license fees as EOS is under GPL V3)
- Has large amounts of data to store (several petabytes)

The only draw-back of EOS (that I'm aware of) is that the in-memory namespace consumes a lot of memory resources and takes relatively long time to boot up. Otherwise it can be a really good fit for big companies and organizations who need to handle lots of data in a cost-effective and relatively easy manner.

Glossary

[1] - http://zimmer.csufresno.edu/~yogao/ATLAS/CERN-LHC-ATLAS.html

[2] – EOS Introduction and Architecture, COMTRADE, https://cernbox.cern.ch/cernbox/desktop/index.php/s/Nbpi3hGZYYqHN93

[3] – http://xrootd.org