Tamara Belyakhina tamara.belyakhina6@etu.univ-lorraine.fr

File system EOS (CERN)

ITMO University
Saint-Petersburg
31.05.2016

Erasmus Mundus PERCCOM
Course : “Computing clusters, grids, cloud”
Course author: Prof. Andrey Shevel

Contents

- Introduction
- Exabyte Scale Storage (EOS)
- EOS: main features
- Architecture
- Releases
- Deployment
- Recent enhancements and future work

Introduction

- Growing demand for data storage and analysis caused by the data taking of the Large Hadron Collider (LHC)
- 2011: 12 million new files per month in the CASTOR storage system
- 2011: ~ 2.000 storage servers for experiment and user data
- Life cycle management, file system losses and corruptions became major issues for daily operations [1]

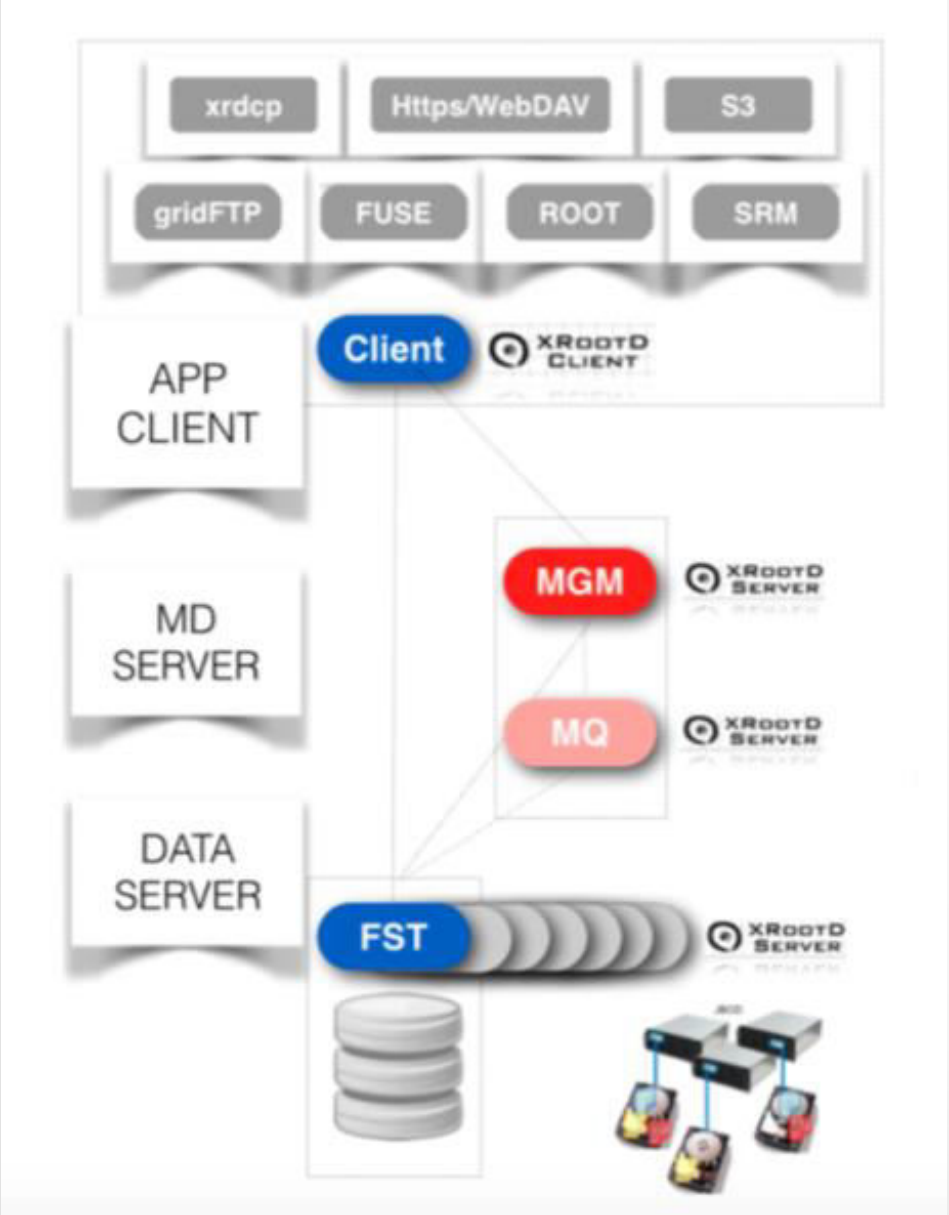
Exabyte Scale Storage (EOS)

- Disk-based storage system providing high-capacity and low-latency access for users at CERN
- Online storage system for all LHC and most non-LHC experiments at CERN and life-cycle management
- Six independent failure domains (instances):
  - the four LHC experiments Alice, Atlas, CMS and Lhcb
  - shared experiment instance for smaller experiments PUBLIC
  - generic user instance USER for all CERN users [2]

EOS: main features

- POSIX-like file access (XRootD, gridFTP, FUSE)
- Hierarchical in-memory namespace (10^9 files , $10^6 - 10^7$ directories)
- Strong authentication (Kerberos5, GSI/X509)
- Quota system for users and groups
- File level checksums
- High availability of services
- Tunable reliability of stored data
- High efficiency and low operational costs
- Transparent life cycle management
  - dynamic pool hardware sizing
  - replacement without down-times [1]

Architecture



- Three components (implemented using the XRootD client-server framework): management server, message queue and file storage services
- Separated IO path into meta data access and data access
- Minimal file access latencies:
  - meta data is kept in-memory on meta data server nodes
  - persisted using WAL technology [2]

Architecture

- Set of single disks (JBOD) as storage media without the need to build local RAID arrays
- Storage nodes are divided into groups and within one group files are placed using file-level network RAID algorithms
- The storage cluster is self-healing
- Online migration of file systems between nodes to simplify life-cycle management
- Additional erasure-encoding of files with two or three redundancy stripes (Jerasure library)
- Hierarchical in-memory namespace is a pluggable component which can be exchanged easily (built on the google sparse hash implementation)[2]

Releases

- Amber (2010-2013)
  - First release, not supported or developed anymore
- Beryl (2013-2015)
  - Active-passive meta data server failover, policy engine, recycle bin, erasure encoding
- Aquamarine (2014)
  - Current production release
  - Archiving, backup, extended HTTP support
- Citrine (2014++)
  - Current development branch
  - Infrastructure aware scheduling, placement for multi-site setups, latest XRootD4 version
- Diamond (2015++)
  - Current research branch
  - Namespace, file IO using Rados object store [2]

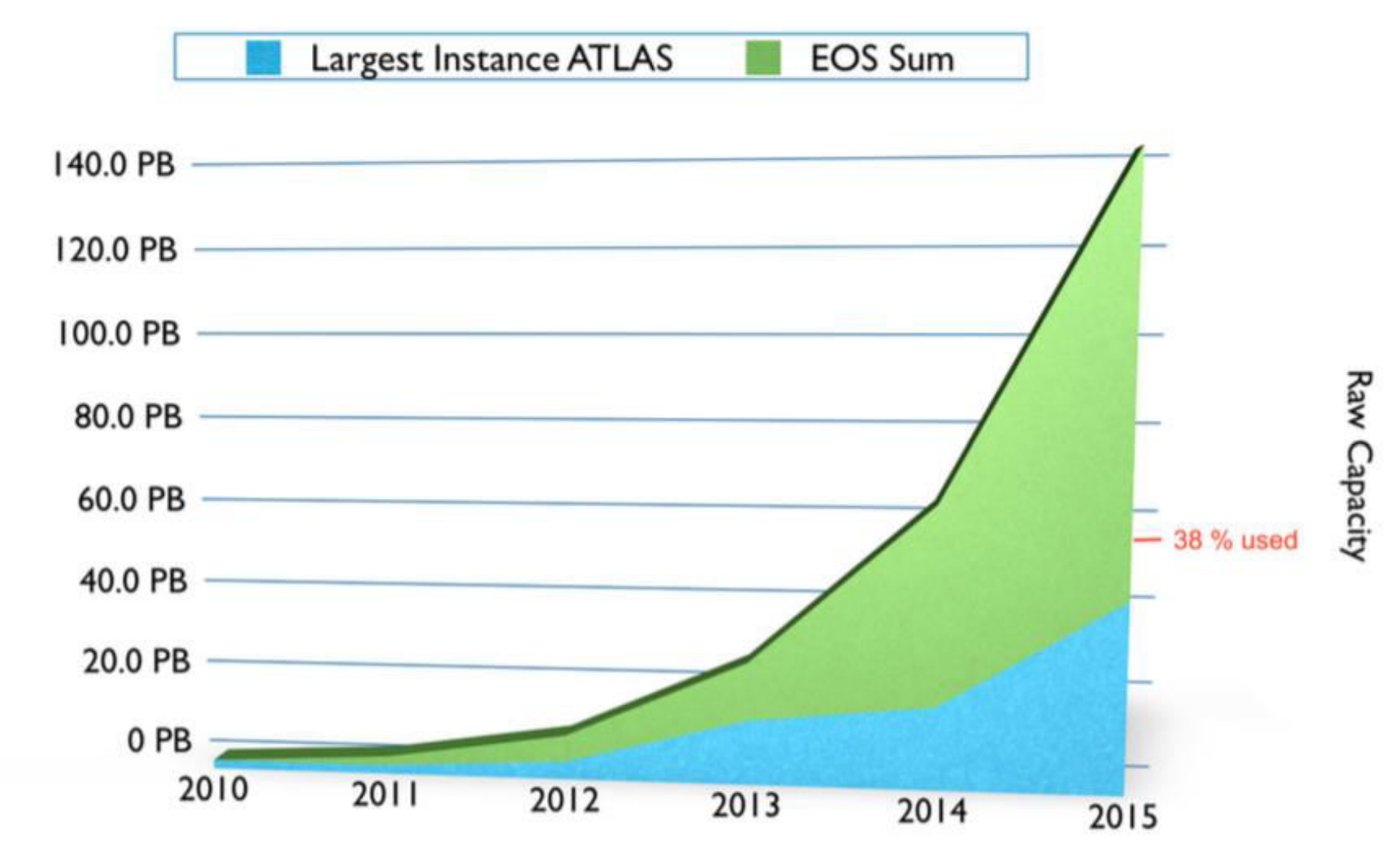
Deployment

- Two computer centers:
  - CERN center in Meyrin(Switzerland) and WIGNER center in Budapest(Hungary)
- Meta data services deployment as six active-passive pairs with real-time failover capabilities on high-memory nodes at the Meyrin center
- ~1.400 server nodes with attached storage (up to 50 disks per node) in both computer centres
- 270 million files (more than half a billion file replicas)
- Three thousand CERN users [2]

Deployment in numbers

Parameter	April 2015
Capacity	140 PB
Server	1.400
Hard Disks	44k
Files	270 M
Directories	26 M
Replicas	0.6 B
theor. Connectivity	13 Tbit
random IOPS	2.2 M
theor. Disk BW	3.3 TB/s
Internal Messaging	150 kHz
State Machine	3 M KV-pairs
Users storing data	3 k
Quota rules	9.600

Deployment


Largest Instance ATLAS
EOS Sum
140.0 PB
120.0 PB
100.0 PB
80.0 PB
60.0 PB
40.0 PB
20.0 PB
0 PB
2010
2011
2012
2013
2014
2015
38 % used
Raw Capacity

Recent enhancements and Future work

- Multi-platform access (Android, IOS, OS X, Windows)
- **FUSE** mount **rm -r** detection and prevention
  - block recursive deletions issued from a shell
- Location defined IO proxy
- Infrastructure Aware Scheduling

- Decrease the time to boot the namespace in memory
- Integration of new cost-efficient IO Plugins [2]

References

- [1] “Exabyte Scale Storage at CERN”, Andreas J Peters and Lukasz Janyst 2011 J. Phys.: Conf. Ser. 331 052015 doi:10.1088/1742-6596/331/5/052015
- [2] "EOS as present and future solution for data storage at CERN", Andreas J. Peters, Elvin A. Sindrilaru, Geoffray Adde 2015

Tamara Belyakhina tamara.belyakhina6@etu.univ-lorraine.fr

Questions?

ITMO University
Saint-Petersburg
31.05.2016

Erasmus Mundus PERCCOM
Course : “Computing clusters, grids, cloud”
Course author: Prof. Andrey Shevel