EOS deployment at GRIF

Vamvakopoulos Emmanouil

On behalf of Technical Committee at GRIF

EOS Workshop 7-10 Mars 2022

CERN

Outline

- Brief description of GRIF's current storage system
- Motivation for Changes
- Context Diagram of future EOS services at GRIF
- Few details on configuration
- Organization of EOS FS and Scheduler Groups on heterogeneous environment
- Plans and milestones

Storage@GRIF for LHC/EGI VO

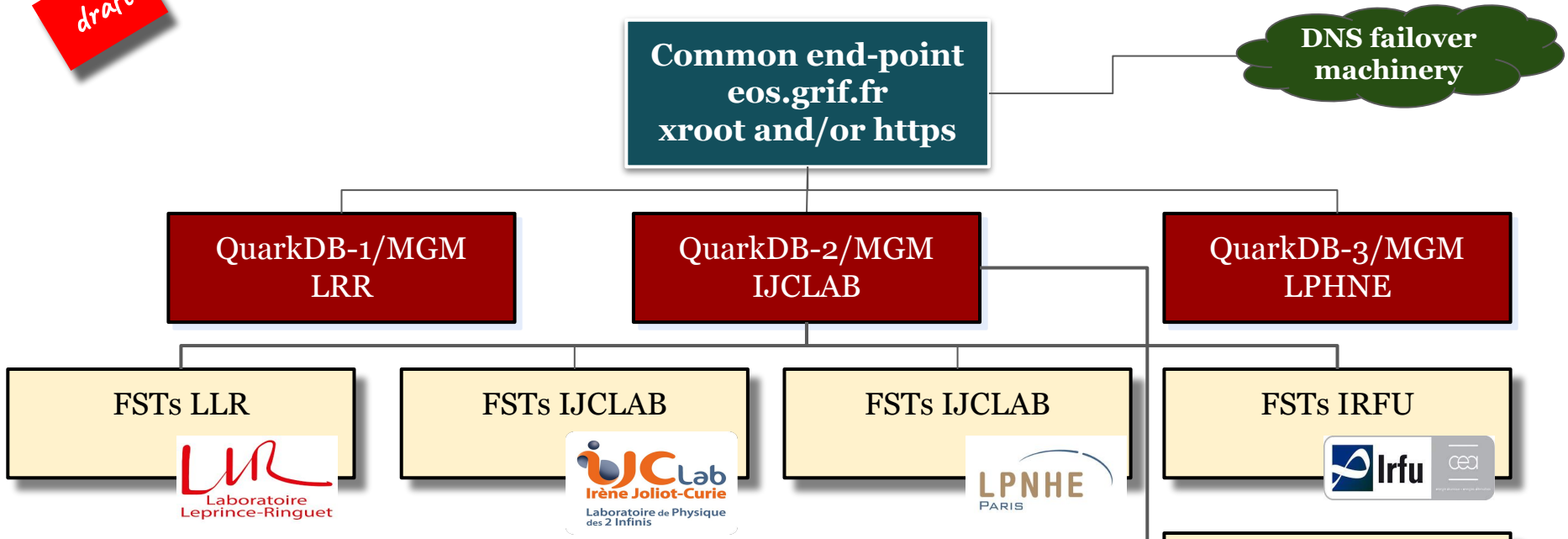
- **GRIF** is a distributed site made of four (4) different subsites, in different locations of the Paris region.
- **IRFU, LLR** and **IJCLAB** are interconnected with 100Gb link.
- The worst network latency between the subsites is within 2-4 msec
- Four (4) independent DPM instances
- Total Pledges Capacity ~10 PBytes
- Supports four (4) WLCG VOs: **Alice, Atlas, CMS and Lhcb** + several EGI VOs
- Hardware configuration is mainly storage servers with 10Gbit nics (or more) with direct attached sata disks
- **Data protection based on RAID-6** done by server's controller
- **Quite heterogeneous hardware layout** and hard drive sizes between the sites and servers' generations

Motivation for changes

- DPM is reaching its end of life soon as a WLCG/EGI service
- GRIF represents a total of ~10 PB but is seen as 4 medium-size sites
  - Avoid duplication of data amongst the subsites (depending on the VO's DDM workflow)
  - Optimum usage of storage resources in a common pool
- Datalakes perspective makes GRIF configuration inappropriate
  - Has the potential to be a major player in a French datalake if it can expose one GRIF endpoint for each VO
- Management not optimal: we can share experience/tools but each subsite has to be managed independently
- Manpower/expertise is not increasing, we need to consolidate our efforts amongst the four subsites
- In addition, work started on a distributed Ceph instance could open the way for more things in common

EOS@GRIF

draft



- Quarkdb (and MGMs) cluster with three (3) nodes
- FST nodes will span over four (4) sites
- Representative number of xrootd PSS gateways for xroot TPC with delegated proxies
- Usage proxy and firewall nodes under considerations
- Storage accounting and BDII publication

Installation and Configuration

- Usage of Quattor and Puppet configuration tools for deployment
- IPV4, IPv6 public network
- Firewall Rules
- Grid Certificates key/pair
- Grid General configuration (Pool account, CAs, vomses, edg-gridmapfile)
- EOS rpms repositories (exclude xroot and microhttpd from epel and umd)
- Install EOS and quarkdb rpms
- Keytab secrets and macaroons
- Sysconfing environmental file → /etc/sysconfig/eos_env
- Base EOS configuration files: xrd.cfg.xxx files for fst, mgm and mq
- ssh from MGM to FST without password would be convenient
- Setup of the DNS failover mechanism
- Setup of EOS internals
- And last but not least: monitoring

Current capacity plan

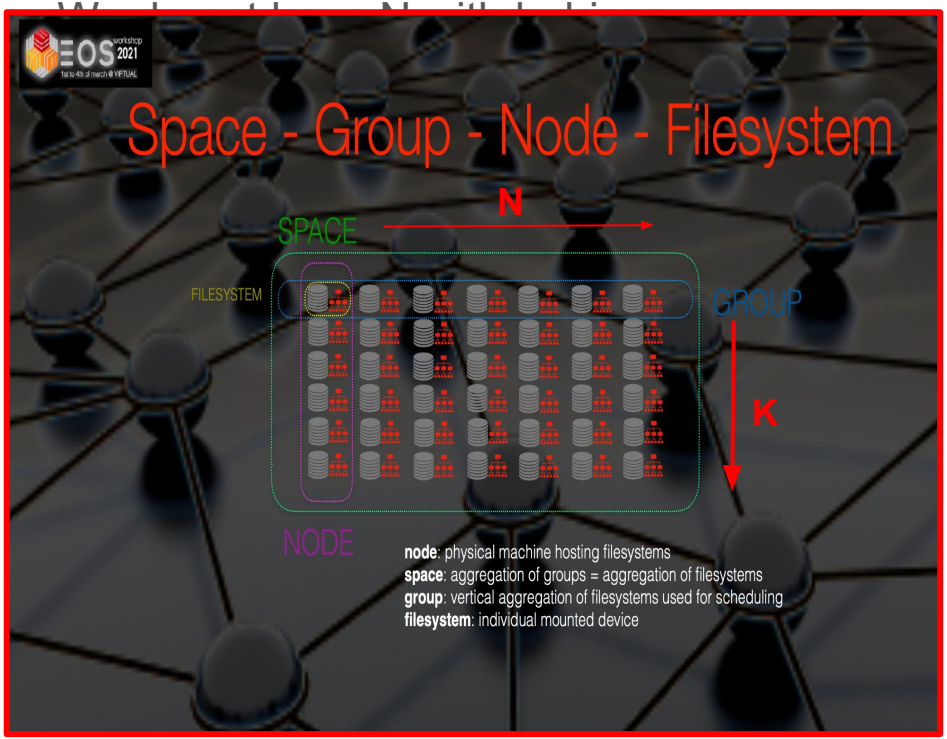
- We have heterogeneous distribution of storage capacity over the four (4) sites which depends from
  - Difference of funding streams of each subsite
  - Internal network architecture and cooling capabilities differ at each subsite
  - Different hardware layout due to different purchases campaigns
  - Different # of servers because of the Internal distribution of the WLCG Pledges on top of each site
- With have servers with total attach capacity (from 100TB, 160TB 240TB up to 760TB)
- Indicative number of servers per subsite: 4 server on LPNHE, 11 on LLR, 14 on IJCLAB, 32 on IRFU

Distribution of Used “space” to be migrated

FEB 22

	IRFU	IJCLAB	LLR	LPNHE	Total
ALICE	450TB	966TB	0	0	1,4PB
ATLAS	1.9PB	1.3PB	0	1.3PB	4,5PB
CMS	1.5PB	0	1.8PB	0	3,3PB
LHCB	0	156TB	0	113TB	289TB

An Ideal Matrix: N server by K Filesystem (of same size)



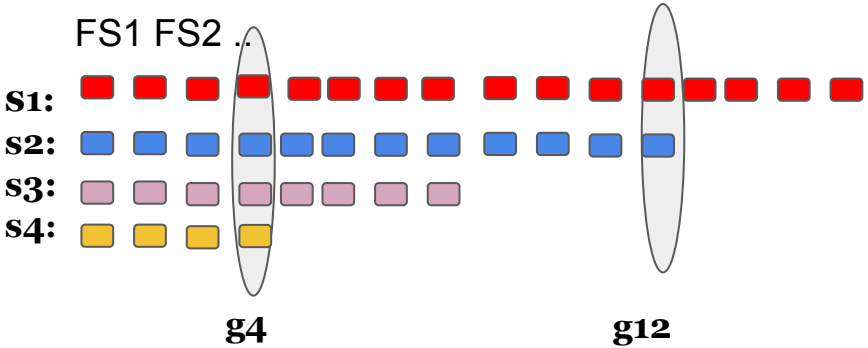
- On Ideal case we have:
- N servers with **K** individual FS on each server (of the same size)
- Thus we have **K** groups with N filesystem on each group (from N different servers)
- Easy to add a new server of same size (of K individual FS)

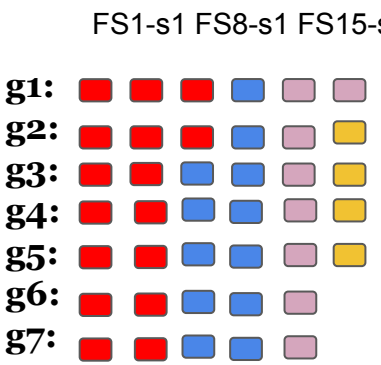
Andreas Joachim Peters et al. EOS Basic Concepts and Design, EOS Workshop 2021

EOS and Space Organization

- **One eos “space” for the four (4) LHC VO**
  - All FSTs will support all the VOs
  - All subsites will support “Filesystems” for all VOs
  - Uniform utilization of the capacity and the server bandwidth (disk and network) as much we can
- **Hardware available is not able** to support Erasure Code (e.g. physical memory , size and # of the disks
  - Keep the data protection under raid6 and split large (~100-160TB) raid6 volumes on several partitions smaller (FS) partitions
- **Try to establish a procedure** and organize the FSs in Scheduler Groups according the following requirements :
  - Each FS file system should have the same size of equal size ~20TBytes (easy to manage, respect the limits and marks per FS in Scheduler Logic, load of fsck ?)
  - Each scheduler group should have as much as minimum number of FS’s per same server (for fixed group’s size this maximize the network and disk throughput)
  - Each scheduler group should have as much the same total capacity in order have an uniform usage of the groups via a round-robin selection.

A non uniform example of EOS File systems Organization





- Let's imagine 4 servers with 16,12, 8, and 4 FS of the same size
- The original organization of FS can not be deployed as we are going to have a group with a non-uniform number of FS
- in total, We have 40 groups
- *k=int(sqrt(40))+1 = 7 (a rule of thumb)*
- Sort the server by the # of filesystems
- Take the server with the largest number of FS and fill cyclically the group table
- And continue to the next one
- At the end, we have a matrix **of k group x k fs** which looks more uniform than the initial one
- We have as much as the minimum # of FS from the same server for each group
- We expect that with a larger number of server/fs this will converge better (more uniform groups)
- This procedure is easy to deploy when we add a new FST
- This procedure is not unique

Plan and milestone

- **Preparation Phase Q1-2022**
  - Functional Quattor and Puppet modules
  - Have a running EOS instance under pre production some SAM test for the four (4) LHC VO + dteam
  - Have a working FTS TPC with https/xrootd for each LHC VO
  - First contact with the four (4) LHC VOs and discuss about the data migration plan
- **First data Phase and Preparation Q2-2022**
  - Have the final workflow and plan for data migration
  - Start to Migrate at least one (1) LHC VO
  - First version of a local operational guide for EOS - documentation
- **Second data Phase Q3 & Q4 -2022**
  - Data migration of LHC VOs
- **Third data Phase Q1-2023**
  - Data migration for non LHC VOs

Potential risks and mitigations

- Phase Q1 delay by ½ a month: not a big impact
- Data migration for large VOs (CMS and ATLAS): may need to do it by subset, not completely clear what the real impact is
    - Spare space for the migration: ~1.5 PB
- Underestimation of the migration time: delayed completion, need of maintain 2 storage services for a longer period
- Small, not really managed, VOs: how to coordinate with them ?

Acknowledgements

Many thanks to EOS developers team for the discussions and the recommendations

Many thanks for yours attention

Questions and Comments ?

BACKUP slides

Configuration details

- EOS 5.0.x
  - Mixing nodes with Centos 7 and Centos 8 flavors
- Identical gridmap file along the sites
- Identical pool unix accounts for the VOs
  - Logically we need 2-3 accounts (depending on VO internal DN/proxies usage)
  - VOs, which give access to each user can drive to a large gridmapfile
  - We are not sure if we need the VOMS extension matching or not (?)
  - **e.g. http.secxtractor /opt/eos/xrootd/lib64/libXrdVoms.so -vomsfunparms:certfmt=pem\|vos=atlas,dteam\|grps=/atlas,/dteam,/dteam/france\|grpopt=10\|dbg**
  - **Plus the vid mapping: DN/voms role→User**
- Usage of native http(s) xrootd interface only on specific ports
  - Do not use microhttpd interface - under decommission
  - EOS_MGM_HTTP_PORT=9000 and EOS_FST_HTTP_PORT=9001
- Looking forward for the redirection from Slave to Master MGM (for xroot and http(s))

EOS@MGM

- sec.protparm gsi -vomsfun:/opt/eos/xrootd/lib64/libXrdSecgsiVOMS.so -vomsfunparms:**certfmt=pem|vos=atlas,dteam|grps=/atlas,/dteam,/dteam/france|grpopt=10|dbg**
- sec.protocol gsi -crl:3 -cert:/etc/grid-security/daemon/hostcert.pem -key:/etc/grid-security/daemon/hostkey.pem -gridmap:/etc/grid-security/grid-mapfile -d:4 **-gmapopt:11** -vomsat:1 -moninfo:1 -gmapto:1

...

- http.cadir /etc/grid-security/certificates/
- http.cert /etc/grid-security/daemon/hostcert.pem
- http.key /etc/grid-security/daemon/hostkey.pem
- http.gridmap /etc/grid-security/grid-mapfile
- http.secxtractor /opt/eos/xrootd/lib64/libXrdVoms.so -vomsfunparms:certfmt=pem|**vos=atlas,dteam|grps=/atlas,/dteam,/dteam/france**|grpopt=10|dbg
- http.trace all
- http.exthandler xrdtpc /opt/eos/xrootd/lib64/libXrdHttpTPC.so
- http.exthandler EosMgmHttp /usr/lib64/libEosMgmHttp.so **eos::mgm::http::redirect-to-https=1**

…

- mgmofs.cfgtype quarkdb
- mgmofs.nslib /usr/lib64/libEosNsQuarkdb.so
- Mgmofs.qdbpassword mystrongsecret