



**Ξ OS deployment at GRIF**

**Vamvakopoulos Emmanouil**

**On behalf of Technical Committee at GRIF**

**GDB Meeting 13 April 2022**

**CERN**

# Outline

- Brief description of GRIF's current storage system
- Motivation for changes
- Context diagram of future EOS services at GRIF
- Comment on installation and version of operating system
- Few notes on configuration
- Organization of EOS filesystems and data protection
- LHC VOs migration plans and Important 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 10+ Gbit NIC and 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

draft

# EOS@GRIF

Common end-point  
eos.grif.fr  
xroot and/or https

DNS failover machinery

QuarkDB-1/MGM  
LRR

QuarkDB-2/MGM  
IJCLAB

QuarkDB-3/MGM  
LPHNE

FSTs LLR  


FSTs IJCLAB  


FSTs IJCLAB  


FSTs IRFU  


Couple of  
PSS components  
(co-exist with a FSTs)

- Quarkdb (and MGMs) cluster: 3 nodes distributed on 3 sites
- FST nodes (disk servers) will be distributed in the 4 GRIF subsites
- Some xrootd PSS gateways for xroot TPC with delegated proxies

# EOS Installation & OS version

- EOS Diopside (5.x.x) installation on Centos Stream 8 and Rocky Linux
  - OS distribution based on GRIF subsite preference
  - Still in testing phase (repositories)
- EOS 5.x.x for Centos Stream 8 is available
  - Always built with the latest CS8 RPMs: not working with RHEL8 and derivatives
- EOS 5.x.x for RHEL 8 will be available soon
  - Not the same repository as CS8
  - build the packages based on the official RHEL8 container image:
    - <https://hub.docker.com/r/redhat/ubi8>
  - May be for production it would be better to have one repo for RHEL8 + CS8

# Installation and Configuration

- Automated deployment and configuration
  - Quattor (3 subsites), Puppet (1)
  - All the EOS/xrootd configuration files managed with the configuration tools
- IPV4, IPV6 public network
- Keytab secrets and macaroon
- MGM endpoint: alias failover through DynDNS managed by a script
  - Update DNS alias based on which instance is the MGM master (quick failover at the MGM level)
  - Used successfully for years for services like BDII
  - Latency: at least a minute (depends on cron job frequency)

# GRIF deployment and erasure coding

Erasure coding considered but abandoned for EOS@GRIF for two major reason:

- Hardware profile at GRIF amongst the four sites is not uniform
  - Different number of servers, different number and size disks due to different local capacity plan
  - Continuous procurement process: no chance to buy the same HW config 2 years, not a volume large enough to build a new homogeneous FS group each year
- Achieving a global resiliency comparable to DPM would require a storage overhead ~30%
  - Failure at one site should not impact more that the data stored uniquely at the site
  - Current RAID6 diskservers have an average overhead of 12%: no budget for increasing it



# Distribution of Used “space” to be migrated

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

FEB 22

- We have servers with total attach capacity (from 100TB, 160TB 240TB up to 760TB)
- Number of servers per subsite: 4 server on LPNHE, 11 on LLR, 14 on IJCLAB, 32 on IRFU

# EOS and Space Organization

- **EOS FS:** no requirement of adopting a unique size as EC is not used
  - EOS provide a balancer to ensure that the usage level of each volume is “the same” with various policies
  - Most FS will be RAID6 volumes, typically in 14+2 configuration, sometimes splitted in several partitions.
  - 500 TB out of Ceph backend, at least for the transition/migration period
  - May want to standardize approximately FS size to limit rebalancing
- **FS groups** with FS from every site, resilience at the FS level as no erasure coding
  - A file is in one FS only: losing a file system will impact neither files not stored in this file system nor ability to write to the FS group
  - Some VOs may be restricted to some subsites using FS geotags
- **One EOS space all the VOs**
  - Uniform utilization of the capacity and the server bandwidth (disks and network) as much we can

# LHC VOs first dialogue and data migration plan

- **Atlas**
  - Atlas DDM group will perform the data migration via rucio and FTS
  - Atlas rucio and panda machinery are having capabilities for seamless transition
  - Thanks to existence of cached/secondaries replicas at GRIF sites the total amount of ATLAS data to migrate is only 3.3PB
- **CMS**
  - Data management group of CMS will perform the data migration via rucio and FTS
  - We need to assess with CMS people the capabilities for seamless transition
  - Not clear yet if we are required to migrate all the data (cache, secondary replicas...)
- **Alice**
  - Data management group will perform the data migration
  - The data migration will be done offline ( maximum 2 steps)
- **LHCb**
  - Still in discussion, LHCb would prefer that GRIF handles the migration but we are not really ready to do it
  - Small volume of data compared to other VOs (~300 TB)

**In all the cases, we need a decent amount (~1PB) of extra capacity for the initial data migration before releasing the first DPMS servers**

# Migration roadmap... ~1 year

- **Remark**: no attempt to balance the data between sites during the migration
- **May 15**: final setup of the EOS instance, configuration of SAM tests for the 4 LHC VOs
- **May 16 - June 19 (5 weeks)**: ATLAS LPNHE (0.7 PB)
  - Hope to migrate 150 TB/week (2 Gb/s); to be validated
  - Need to start draining DPM probably before the end of the migration, after file deletion
- **June 13 - July 31 (7 weeks)**: ALICE IPNO (1 PB)
  - Requires new HW to be delivered and installed
- **June 19 - July 31 (7 weeks)**: ATLAS LAL (0.7 PB)
  - Overlap VO migrations. DPM drain and EOS space addition in parallel with migration
  - Also requires new HW to be delivered and installed
- **August: according to VO and local site availabilities**: not clear if we'll do more than completing the previous steps

# Migration roadmap

- **September - November (~12 weeks):** ATLAS Irfu (1.9 PB)
- **September - November (12 weeks):** CMS LLR (1.8 PB)
- ~~**December (3 weeks):** ALICE Irfu (0.45 TB)~~
- **December - February (12 weeks):** CMS Irfu (1.9 PB)
- **March - June:** LHCb: other VOs

# Risks and mitigations

- **Currently on track for the setup of the EOS instance:** should be able to start migration mid-May
  - Important to avoid delays: summer period will be less favorable to start the the initial work with VOs
  - Subject to new HW delivery delays leading to insufficient temporary space in EOS
- **DPM drain longer than expected,** introducing delays in the roadmap before we can start a new migration phase
  - Normally, should not be too long as we move forward as migrated files will be deleted from DPM
- **Underestimation of migration time:** clearly difficult to assess before we started
  - We think our current numbers are very conservative: 2x less that what we observe in real production... but production will continue in
  - Non LHC VOs migration may take time as not much contact
  - They also have no real tools to do it

# Acknowledgements

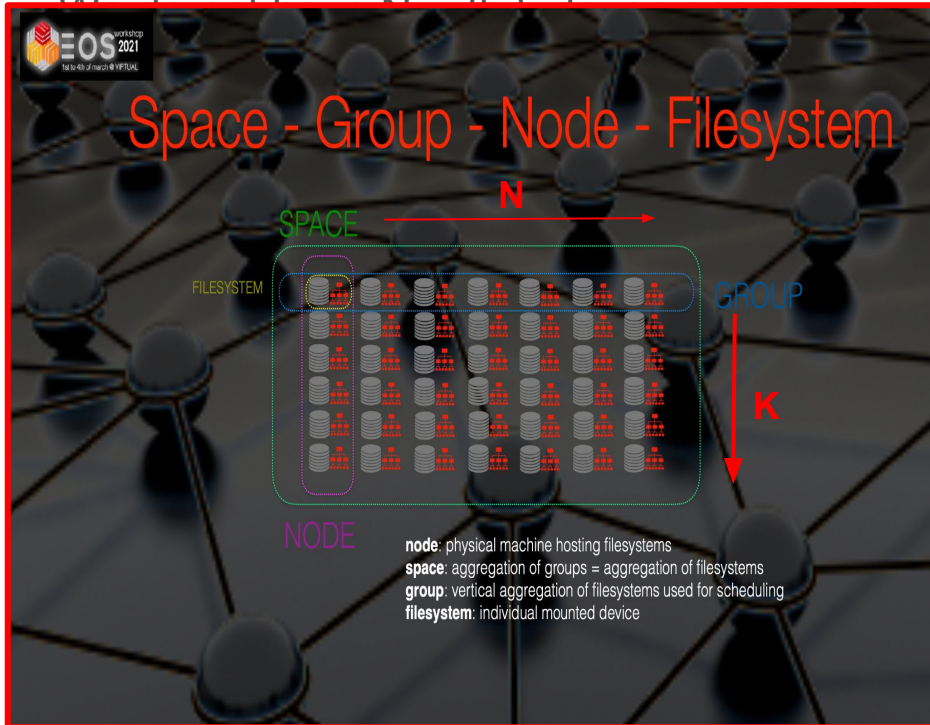
*Many thanks to EOS developers team and LHC VOs technical representatives for the discussions  
and for the recommendations*

*Many thanks for yours attention  
Questions and Comments ?*

# **BACKUP slides**

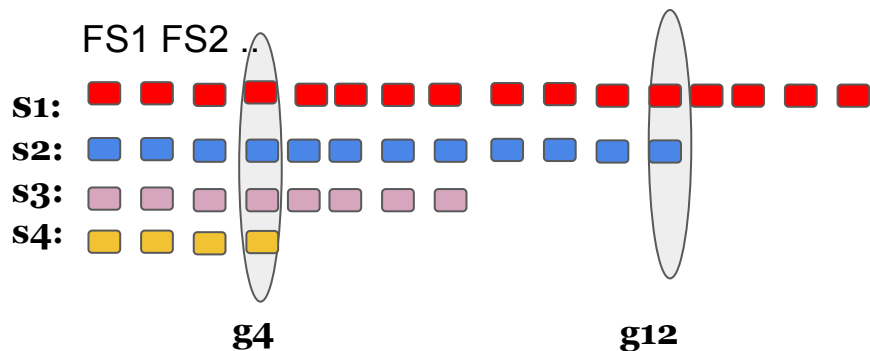


# An Ideal Matrix: N server by K Filesystem (of same size)

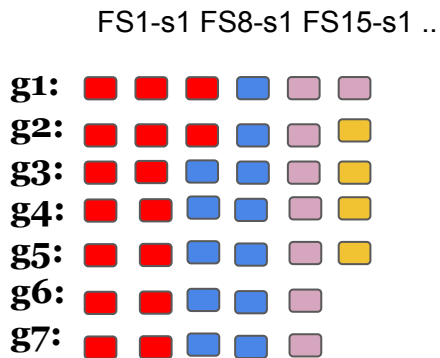


- On Ideal case we have:
- N servers with **K** individual FS on each server (of the same size)
- Thus we have **K** groups with N filesystem on each group (from N different servers)
- Easy to add a new server of same size (of K individual FS )

# A non-uniform example of EOS FS 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
- $k = \text{int}(\text{sqrt}(40)) + 1 = 7$  ( 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



# 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.secextractor /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.protparam 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.secextractor /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`