EOS deployment at GRIF

Vamvakopoulos Emmanouil

On behalf of Technical Committee at GRIF

EOS Workshop 23-26 Mars 2023

CERN

Grille au service de la Recherche en Ile de France
APC
Irfu
cea
IJCLab
Irène Joliot-Curie
Laboratoire de Physique des 2 Infinis
LLR
LPNHE
PARIS

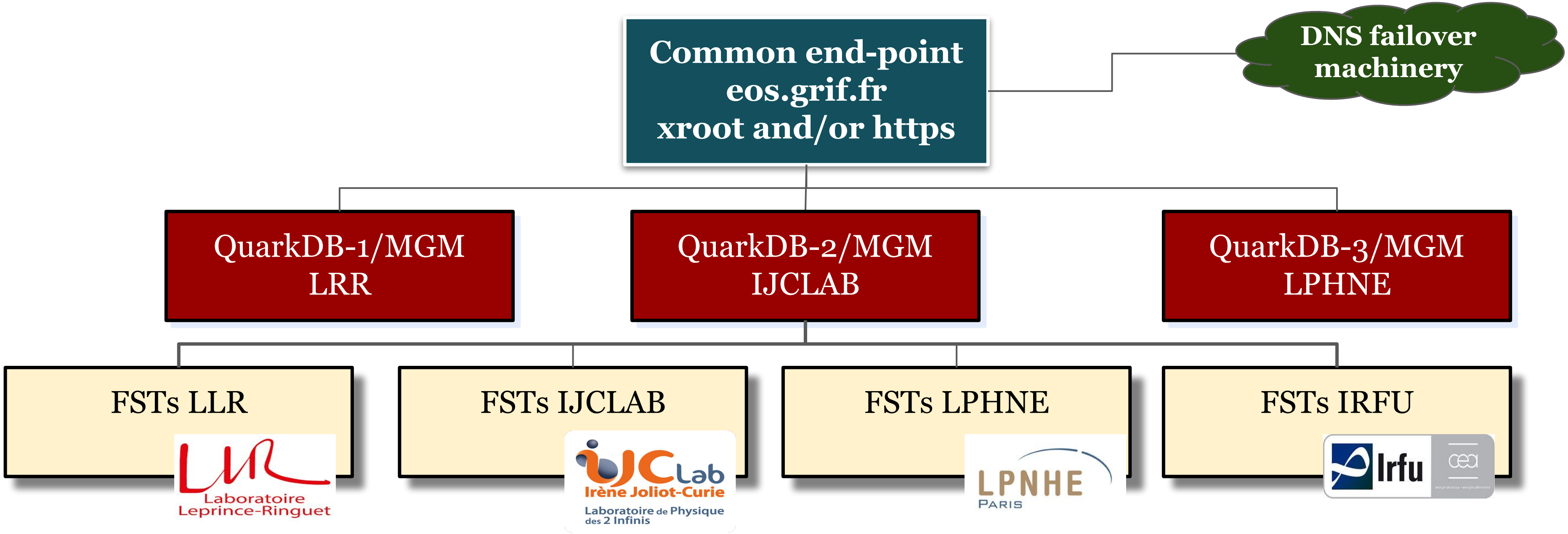
GRIF is a distributed site made of four (4) different subsites, in different locations of the Paris region.

- **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 ~12 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 layou**t** and hard drive sizes between the sites and servers' generations

Plan and milestones

- **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
  - Migrate at least Atlas and Alice LHC VOs
- **Second data Phase Q3 & Q4 -2022**
  - Preparation of data migration of CMS
  - **Third data Phase Q1-2023 (delayed)**
  - Data migration of CMS and LHCb LHC VOs
  - Data migration for non LHC VOs

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 LPHNE

FSTs IRFU

Laboratoire Leprince-Ringuet
IJCLab Irène Joliot-Curie Laboratoire de Physique des 2 Infinis
LPNHE PARIS
Irfu
cea

- Quarkdb (and MGMs) cluster with three (3) nodes
- FST nodes will span over four (4) sites
- Storage accounting

EOS version

- **We are running on 5.1.9**
  - Rocky Linux 8 for IRFU and LPNHE
  - Centos stream 8 for LLR and IJCLAB

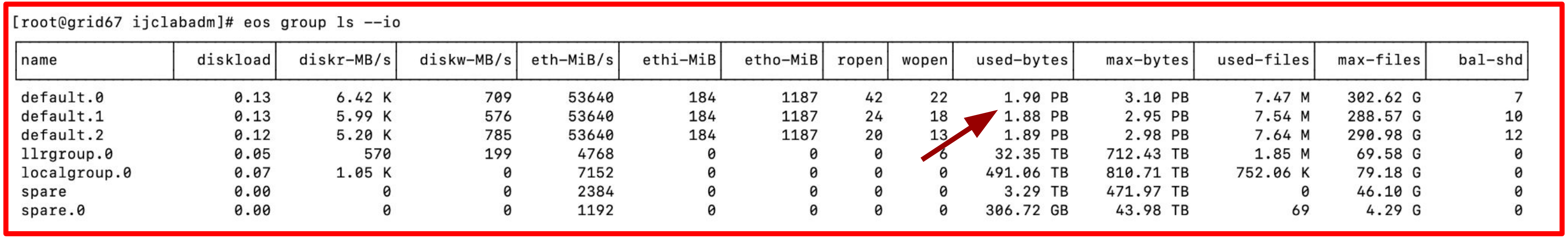
- **Update to Version 5.1.9 (from 5.0.18) solves important bug**
  - The lack of fallocation() usage in https(s) and a block of release of XFS preallocation cause a difference between allocated size and apparent size of the files (this bug mask, up to ~1PB, mask now is 212TB)
- Ambiguity error in https TPC transfers when a mapped DN was not explicitly defined in gridmap-files (e.g. CMS VO, GFAL2 macaroon open/read check error)
- Ambiguities amongst MGMs in failover (?)

Distribution of storage capacity

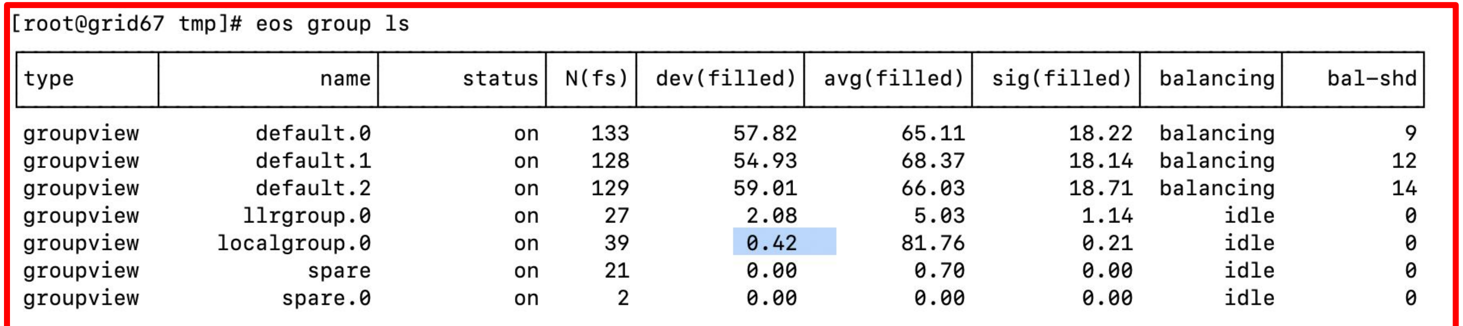
- **We have heterogeneous distribution of storage capacity over the four (4) sites which depends from**
    - Difference of funding streams of each subsite
    - Internal network architecture and cooling capabilities differ at each subsite
    - Different hardware layout due to different purchases campaigns
- Keep the data protection under raid6 and split large (~100-160TB) raid6 volumes on several partitions smaller (FS) partitions
- We have **one (1) default eos "space" for all VOs on production**
    - 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
    - Default Space is made on top of three (3) scheduling group
- We distribute FSs for each site with a round-robin way on each group

Volumetrics

- **Total pledge install capacity ~9.5 PB (max 12.5PB)**
- **Total unpledge capacity for local usage ~1.5TB**
- **486 filesystems over 55 fst nodes over 4 sites :**
  - **(23 IJCLAB, 6 LLR, 9 LPNHE, 12 IRFU)**



```
[root@grid67 ijclabadm]# eos group ls --io
┌────────────────┬────────┬──────────┬──────────┬────────┬────────┬────────┬─────┬─────┬──────────┬──────────┬──────────┬─────────┬───────┐
│name            │diskload│diskr-MB/s│diskw-MB/s│eth-MiB/s│ ethi-MiB│ etho-MiB│ropen│wopen│used-bytes│ max-bytes│used-files│max-files│bal-shd│
└────────────────┴────────┴──────────┴──────────┴────────┴────────┴────────┴─────┴─────┴──────────┴──────────┴──────────┴─────────┴───────┘
 default.0            0.13     6.42 K        709    53640      184     1187    42    22    1.90 PB    3.10 PB     7.47 M  302.62 G        7
 default.1            0.13     5.99 K        576    53640      184     1187    24    18    1.88 PB    2.95 PB     7.54 M  288.57 G       10
 default.2            0.12     5.20 K        785    53640      184     1187    20    13    1.89 PB    2.98 PB     7.64 M  290.98 G       12
 llrgroup.0           0.05        570        199     4768        0        0     0     6   32.35 TB  712.43 TB     1.85 M   69.58 G        0
 localgroup.0         0.07     1.05 K          0     7152        0        0     0     0  491.06 TB  810.71 TB   752.06 K   79.18 G        0
 spare                0.00          0          0     2384        0        0     0     0    3.29 TB  471.97 TB          0   46.10 G        0
 spare.0              0.00          0          0     1192        0        0     0     0  306.72 GB   43.98 TB         69    4.29 G        0
```



```
[root@grid67 tmp]# eos group ls
┌─────────┬─────────────┬────────┬─────┬───────────┬───────────┬───────────┬─────────┬───────┐
│type     │         name│  status│N(fs)│dev(filled)│avg(filled)│sig(filled)│balancing│bal-shd│
└─────────┴─────────────┴────────┴─────┴───────────┴───────────┴───────────┴─────────┴───────┘
 groupview     default.0       on   133       57.82       65.11       18.22 balancing       9
 groupview     default.1       on   128       54.93       68.37       18.14 balancing      12
 groupview     default.2       on   129       59.01       66.03       18.71 balancing      14
 groupview    llrgroup.0       on    27        2.08        5.03        1.14      idle       0
 groupview  localgroup.0       on    39        0.42       81.76        0.21      idle       0
 groupview         spare       on    21        0.00        0.70        0.00      idle       0
 groupview       spare.0       on     2        0.00        0.00        0.00      idle       0
```

Virtual Organizations (VOs)

- **WLCG VO**
  - **alice**
  - **atlas**
  - **ops**
  - **dteam**
  - **cms (under progress)**
  - **lhcb (under progress)**

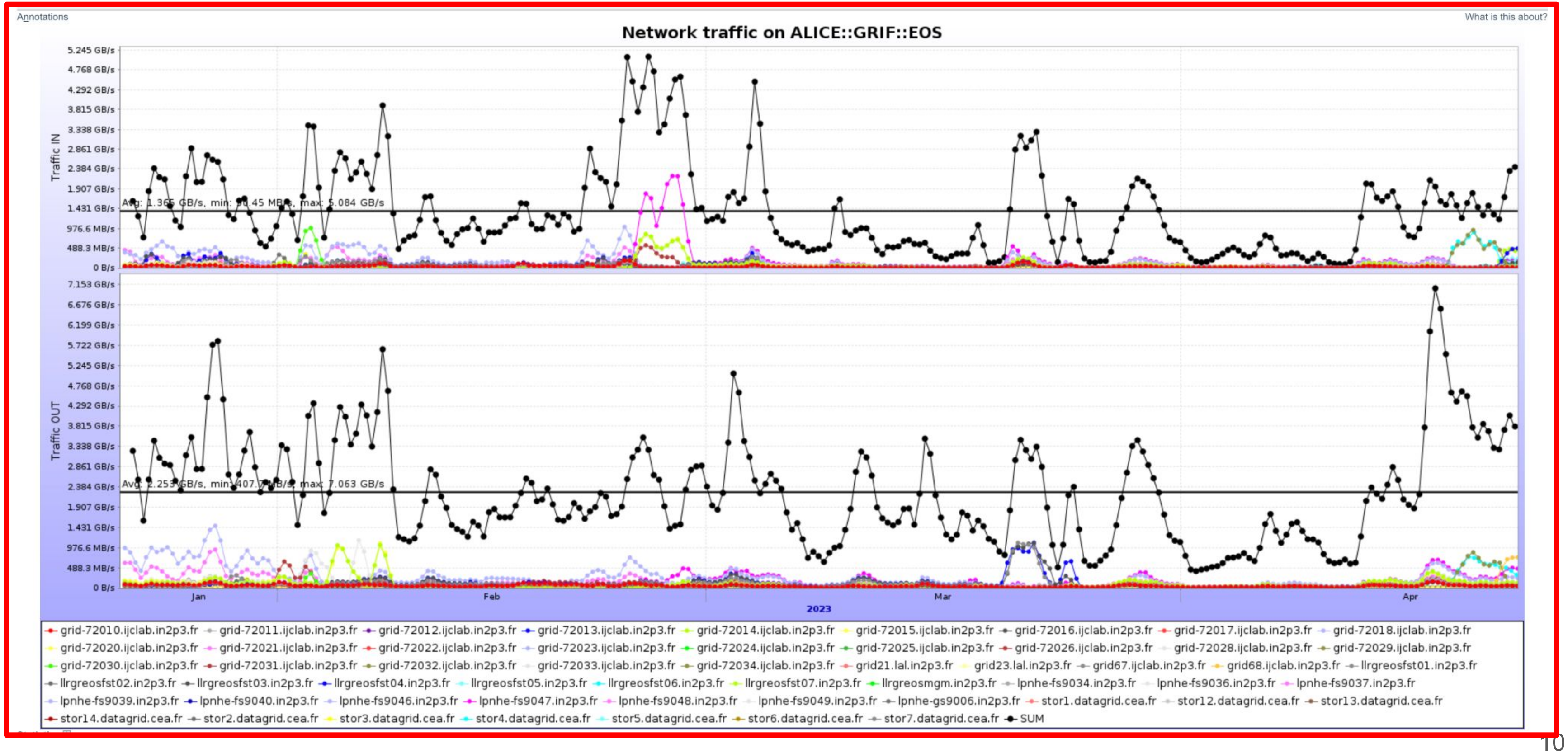
- **EGI VOs**
  - **complex**
  - belle II
  - VO based on Dirac WMS
    - cta
    - hess
  - Other EGI VOs

Alice VO and TkAuthz.Authorization

- cat /etc/grid-security/xrootd/TkAuthz.Authorization
  - EXPORT PATH:/   VO:* ACCESS:ALLOW CERT:*
  - RULE   **PATH:/eos/grif/alice/** AUTHZ:delete|read|write|write-once| NOAUTHZ:| VO:*| CERT:IGNORE
  - KEY   VO:* PRIVKEY:/etc/grid-security/xrootd/privkey.pem PUBKEY:/etc/grid-security/xrootd/pubkey.pem

- sec.protbind * only gsi sss unix
- (**a client with GSI has to authenticate to the MGM with GSI and requires UNIX on the FST**)

Apmon for Alice



Further Steps

- Conclude with VOs migration
- Intention to remove of static DNs and use only Vid (for role based acls)
- Increase capacity , add more FSTs
- Incorporate wlcg tokens (e.g. for CMS)
- Make some tests with “Jambo Frames”
- Test LRU and deletion for temporary areas in namespace (e.g. cms temp dir)
- Understand better the namespace structure, fsck and durability process
- MGM and failover verification

Acknowledgements

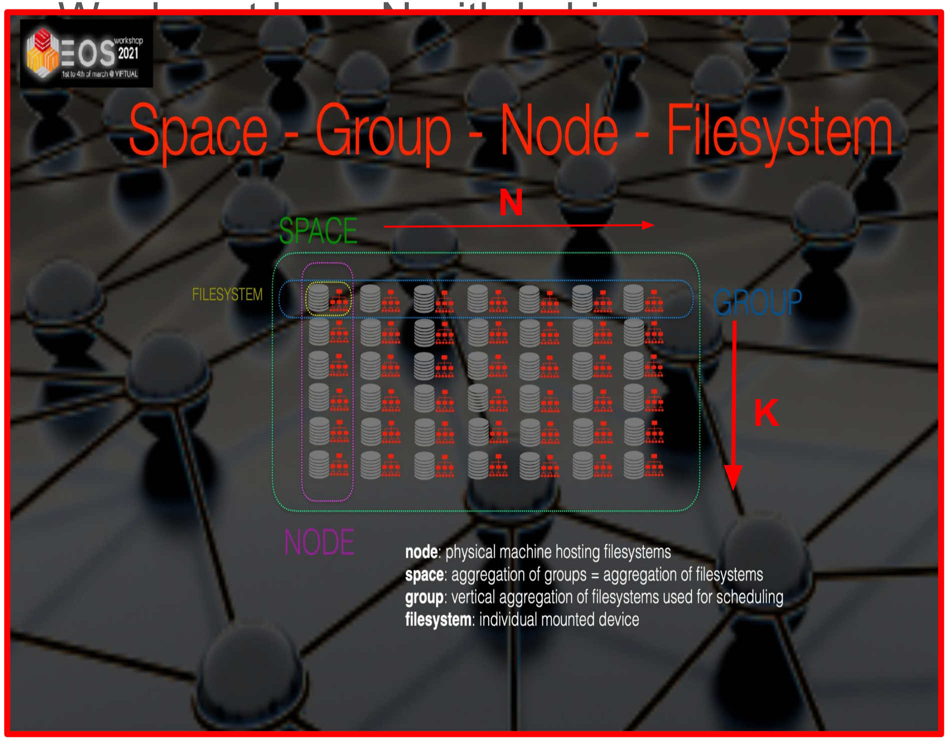
Many thanks to EOS developers team for the discussions and the recommendations

Many thanks for yours attention

Questions and Comments ?

BACKUP slides

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



Space - Group - Node - Filesystem

node: physical machine hosting filesystems
space: aggregation of groups = aggregation of filesystems
group: vertical aggregation of filesystems used for scheduling
filesystem: individual mounted device

- 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

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