CERN

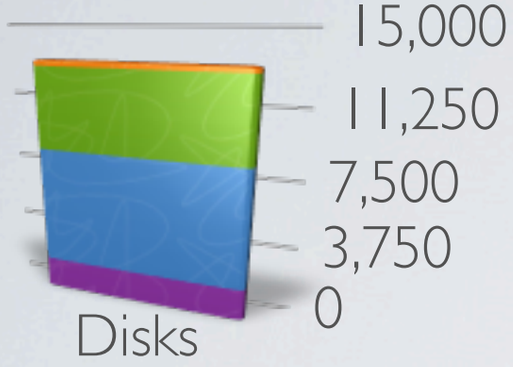
European Organization for Nuclear Research
Organisation Européenne pour la Recherche Nucléaire

IN FEDERATION

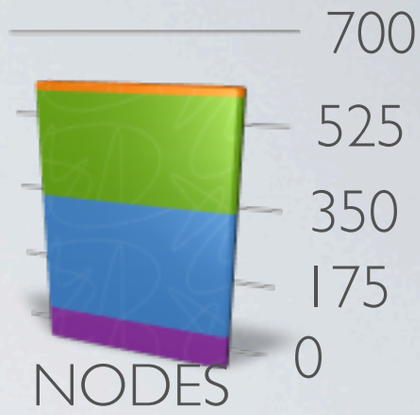
Andreas J. Peters - CERN IT-DSS

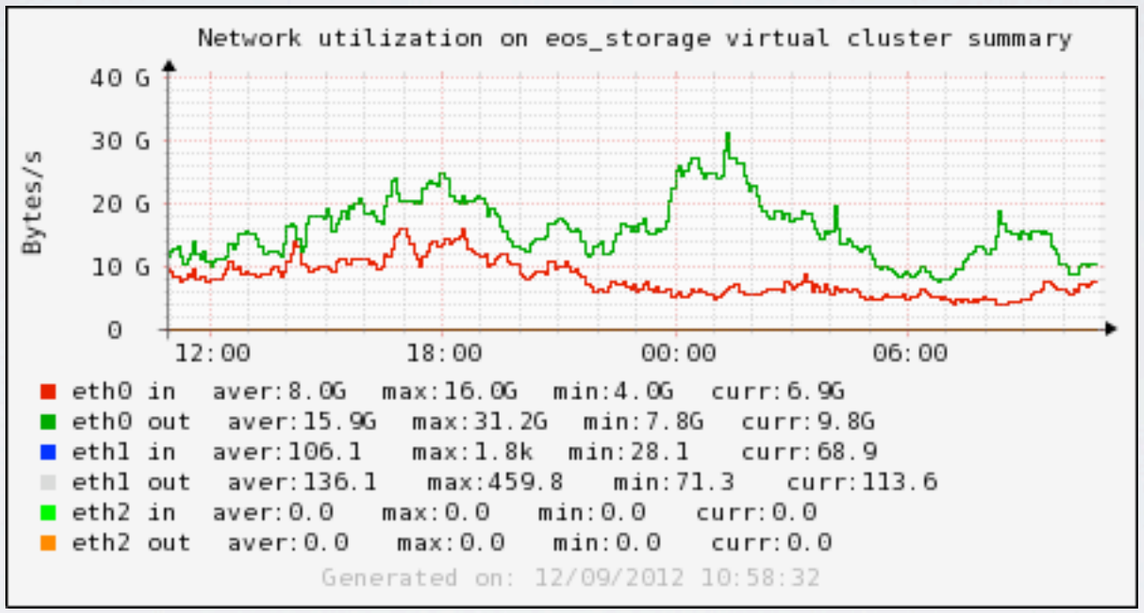
- Why EOS should be part of a federation
  - Usage Today
- Topics for Federation
  - Namespace
    - Name2Name Plugins
    - Prefix Redirection/Virtual storage entry points
  - Monitoring
    - UDP Collector
    - Domain/Application Monitoring
  - Security
  - Performance
  - Federation Model
- Castor in Federation

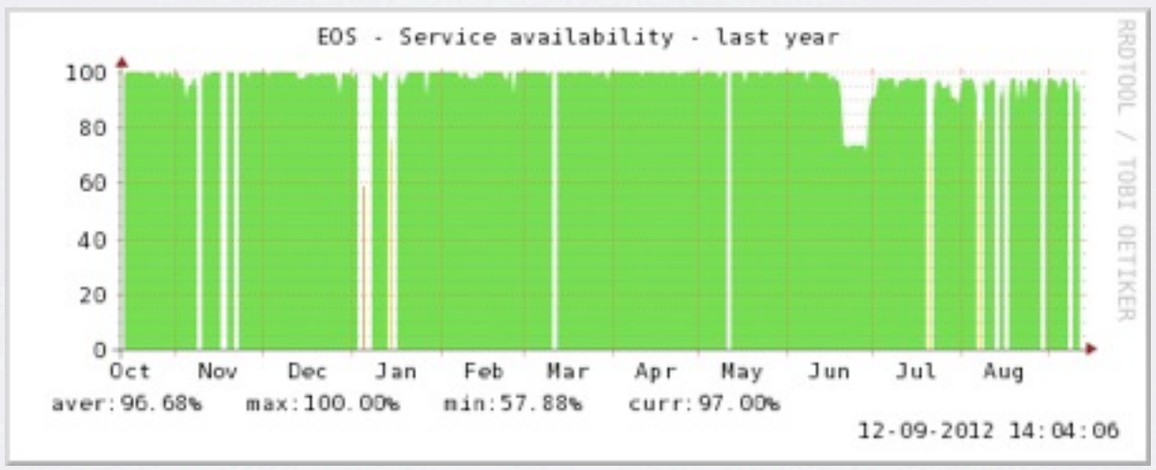
CERN EOS TODAY



ALICE ATLAS CMS LHCB





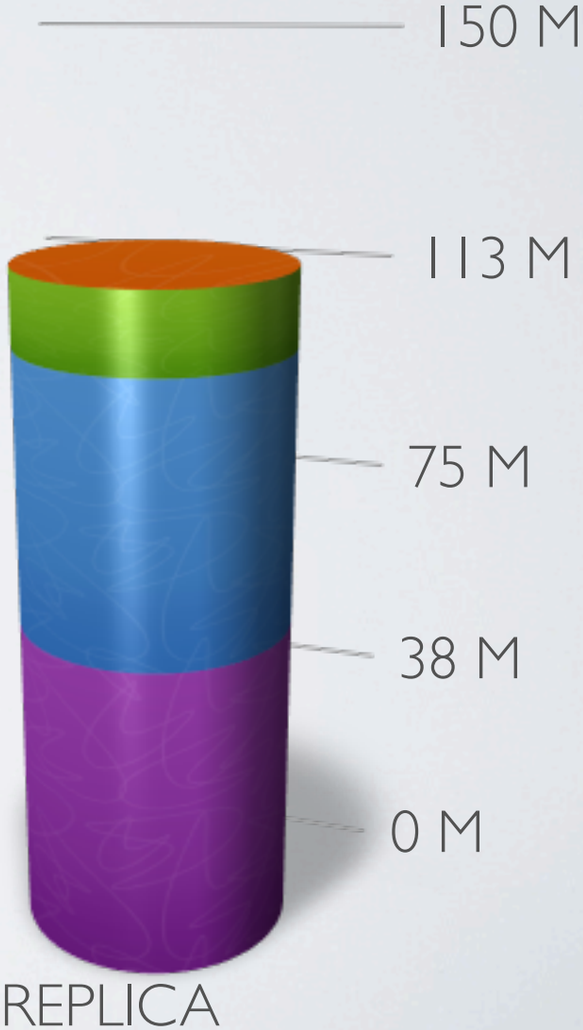


1 Month Averages

Stat/Read-/Create-Open Rates 120 Hz / 70 Hz / 9 Hz

concurrent files Open-Read / # files Open-Write 5.5k / 670

File Checksumming + Data Read 20 GB/s



PROBLEM AREAS FOR FEDERATION OF EOS

Our main headaches ...

- **How-to** support a **global namespace** out of (exotic) naming scheme's in EOS in a multi-spacetoken SE ?

Virtual Storage Elements

  - requires a 1-to-N translation from global to local names

- **How-to** send required **monitoring** information from EOS to a global monitoring system ?

  - who (IP, DN/credential) + what (path)
  - seperate internal from external traffic

Global Monitoring with LAN/WAN separation

  - monitoring interface one layer too high :-(...

EOS INTERNAL N2N MAPPING

- EOS supports dynamic match and replace mapping

CMS

```
[root@eoscmssrv1 ~]# eos -b map ls
/store/                                                    => /eos/cms/store/
```

ALICE

```
EOS Console [root://localhost] |/> map ls
/00/                                                       => /eos/alice/grid/00/
/01/                                                       => /eos/alice/grid/01/
/02/                                                       => /eos/alice/grid/02/
/03/                                                       => /eos/alice/grid/03/
/04/                                                       => /eos/alice/grid/04/
/05/                                                       => /eos/alice/grid/05/
/06/                                                       => /eos/alice/grid/06/
/07/                                                       => /eos/alice/grid/07/
/08/                                                       => /eos/alice/grid/08/
/09/                                                       => /eos/alice/grid/09/
/10/                                                       => /eos/alice/grid/10/
/11/                                                       => /eos/alice/grid/11/
/12/                                                       => /eos/alice/grid/12/
/13/                                                       => /eos/alice/grid/13/
/14/                                                       => /eos/alice/grid/14/
/15/                                                       => /eos/alice/grid/15/
```

ATLAS ?

Mix of many storage area's in one physical storage system ... what to do here? (... indeed now also ALICE ...)

VIRTUAL STORAGE (1)

- EOS can implement several *virtual* SEs to share the same physical backend (SRM spacetoken model)
  - ... deployment model since the beginning
    - one service - each SE uses individual namespace prefixes

      e.g.

      `root://eosatlas//eos/atlas/datadisk`
      `root://eosatlas//eos/atlas/scratchdisk`

- ALICE asked for SEs without prefixes

  `root://eosalice/ /eos/alice/se`
  `root://eosalice-ocdb/ /eos/alice/ocdb`

  => problematic because XRootD protocol allows only to change `host:port` in redirection, not the LFN!

N2N does not work in this case for a global namespace

VIRTUAL STORAGE (2)

- Solution
  - secondary EOS MGM can run as a dummy redirector on different ports and add new prefixes or rewritten LFN as opaque information
  - primary EOS MGM interprets opaque LFN rewrite tags ("`eos.prefix`","`eos.lfn`")
- Caveat
  - works with XrdClient but not with XrdClientAdmin => currently stuck, new Client does support it ...

MONITORING

- EOS difficulty : layer structure of XRootD plugins
  - Strong authentication only on head node - not on disk node (using capabilities) - XRootD protocol layer does not know credentials used on the storage nodes ...
  - XRootD monitoring layer on top of OFS (EOS plugin) => no way to include client identity information into monitoring stream for us ...
- EOS 0.2 allows to send UDP streams from the head node in JSON or Key-Value format compatible with Matevz collector output including authentication information ...

MONITORING (2)
EOS UDP STREAM

- UDP stream is realtime configurable with multiple targets
  - contains full authentication (host, domain, protocol, DN ...) & application information e.g.
    - gridFTP transfer
    - internal drain
    - internal balancing
    - FUSE access
    - applications can use 'eos.app=<app>' as opaque tag to tag IO activities

    => could add filter to restrict to certain application tags only
    (e.g. report only FAX/AAA traffic ...)
    => can adapt to new 'f-stream' proposal

MONITORING (3)
EOS UDP STREAM

- However ... one remaining problem
- OFS layer does not see readV calls - only multiple read calls visible ...
  - EOS can currently not provide this information in the UDP collector output stream ...
    (send UDP from protocol to OFS layer?)

MONITORING (4)

EOS INTERNAL APPLICATION/DOMAIN MONITORING

```
EOS Console [root://localhost] |/> io stat -d
# ------------------------------------------------------------------------------------
# IO by domain/node name:
# ------------------------------------------------------------------------------------
io         domain                                        1min     5min       1h      24h
# ------------------------------------------------------------------------------------
OUT        .ro                                           0.00  52.65 M   2.11 G  59.66 G
OUT        .cz                                           0.00
OUT        .fr                                         3.74 G
OUT        .uk                                       626.35 M
OUT        .ru                                       664.13 M
OUT        .edu                                          0.00
OUT        .su                                           0.00
OUT        eos                                           0.00
OUT        .dk                                           0.00
OUT        .org                                      64.54 k 4
OUT        lxplus                                        0.00
OUT        .no                                           0.00
OUT        lxb                                       39.02 G 1
OUT        aldaq                                         0.00
OUT        .nl                                        4.29 M
OUT        .it                                       35.13 G
OUT        other                                      1.12 G
OUT        pb-d-128-141                                  0.00
OUT        .se                                        2.44 M
OUT        .ch                                       39.19 G 174.80 G   2.35 T 293.07 T
OUT        .de                                       92.38 M   2.84 G  25.66 G 628.41 G

IN         .ro                                           0.00    0.00     0.00  89.12 G
IN         .cz                                           0.00    0.00     0.00 890.96 M
IN         .fr                                       42.88 M 527.01 M  20.96 G   1.01 T
IN         .uk                                          0.00  25.46 M   1.95 G 121.15 G
IN         .no                                       5.29 k 369.48 k 625.22 k   3.36 M
IN         .edu                                         0.00    0.00     0.00   4.71 G
IN         lxplus                                       0.00    0.00  14.17 k  18.96 M
IN         eos                                          0.00    0.00     0.00  27.04 T
IN         .ru                                          0.00    0.00     0.00 133.03 M
IN         lxb                                       1.32 G   5.38 G 170.96 G   3.77 T
IN         .nl                                          0.00    0.00     0.00  40.92 G
IN         pb-d-128-141                              12.73 k  19.07 M  34.62 M 456.75 M
```

by domain/cluster

```
EOS Console [root://localhost] |/> io stat -x
# ------------------------------------------------------------------------------------
# IO by application name:
# ------------------------------------------------------------------------------------
io         application                                   1min     5min
# ------------------------------------------------------------------------------------
OUT        eos/draining                                  0.00     0.00
OUT        eos/gridftp                                42.27 G  42.27 G
OUT        eos/balancing                              97.30 G  97.30 G
OUT        eos/replication                               0.00     0.00
OUT        other                                     103.14 G 103.14 G

IN         eos/draining                                  0.00     0.00
IN         eos/gridftp                                16.89 G  16.89 G
IN         eos/balancing                              90.01 G  90.01 G
IN         eos/replication                               0.00     0.00
IN         other                                      13.84 G  13.84 G
```

by application

Allows local monitoring & debugging!

MONITORING (5)

EOS INTERNAL IO NAMESPACE

7 DAYS HISTORY

TOP 10/100/1000/10000

```
EOS Console [root://localhost] |/> io ns
# ------------------------------------------------------------------------------------
rank   by(read count) read bytes     path
# ------------------------------------------------------------------------------------
000001 nread=155666  rb=12.73 TB   /
000002 nread=155666  rb=12.73 TB   /eos/
000003 nread=155666  rb=12.73 TB   /eos/atlas/
000004 nread=78260   rb=8.07 TB    /eos/atlas/atlasdatadisk/
000005 nread=46107   rb=7.29 TB    /eos/atlas/atlasdatadisk/data12_8TeV/
000006 nread=44379   rb=1.80 TB    /eos/atlas/atlasgroupdisk/
000007 nread=39793   rb=801.37 GB  /eos/atlas/atlasgroupdisk/perf-idtracking/
000008 nread=39793   rb=801.37 GB  /eos/atlas/atlasgroupdisk/perf-idtracking/dq2/
000009 nread=39793   rb=801.37 GB  /eos/atlas/atlasgroupdisk/perf-idtracking/dq2/mc12_8TeV/
000010 nread=39793   rb=801.37 GB  /eos/atlas/atlasgroupdisk/perf-idtracking/dq2/mc12_8TeV/NTUP_MINBIAS/
# ------------------------------------------------------------------------------------
rank   by(read bytes) read count     path
# ------------------------------------------------------------------------------------
000001 rb=12.73 TB   nread=155666  /
000002 rb=12.73 TB   nread=155666  /eos/
000003 rb=12.73 TB   nread=155666  /eos/atlas/
000004 rb=8.07 TB    nread=78260   /eos/atlas/atlasdatadisk/
000005 rb=7.29 TB    nread=46107   /eos/atlas/atlasdatadisk/data12_8TeV/
000006 rb=4.82 TB    nread=3385    /eos/atlas/atlasdatadisk/data12_8TeV/AOD/
000007 rb=2.94 TB    nread=1147    /eos/atlas/atlasdatadisk/data12_8TeV/AOD/f475_m1223/
000008 rb=1.83 TB    nread=661     /eos/atlas/atlasdatadisk/data12_8TeV/AOD/f475_m1223/data12_8TeV.00209864.physics_JetTauEtmiss.merge.AOD.f475_m1223/
000009 rb=1.81 TB    nread=1737    /eos/atlas/atlasdatadisk/data12_8TeV/AOD/f475_m1218/
000010 rb=1.80 TB    nread=44379   /eos/atlas/atlasgroupdisk/
```

SECURITY (1)

- storage internal traffic is authenticated with 'sss' protocol
- client internal & external traffic requires kerberos or X509 credentials
- CMS: DN mapped to CERN account=kerberos principal
- ATLAS: DN mapped to pool account
  (non-symmetric access in- and outside grid jobs => difficult to access user private directories in this way)
- ALICE: asymmetric+symmetric encrypted authorization token issued by ALICE - no authentication

SECURITY (2)

- ACLs/Unix permissions on parent define protection of children
  - std. unix permission schema user/group/other
  - ACLs for listing, chmod, delete, quota-set and write-once
  - add virtual group permissions via EGROUPs
  - bind secondary owner (user) via authentication principle (not only mapped uid/gid pair) e.g. owner of a directory can be attached as 'krb5:cernprod' or 'gsi:DN=/...../'

SECURITY (3)

- Impact on Federation
  - EOS 'lookup' restricted to 'sss' authenticated connection from local federation xrootd/cmsd pair
- File access possible only with strong authentication and suitable permissions
  (no anonymous access)

PERFORMANCE IN FEDERATION

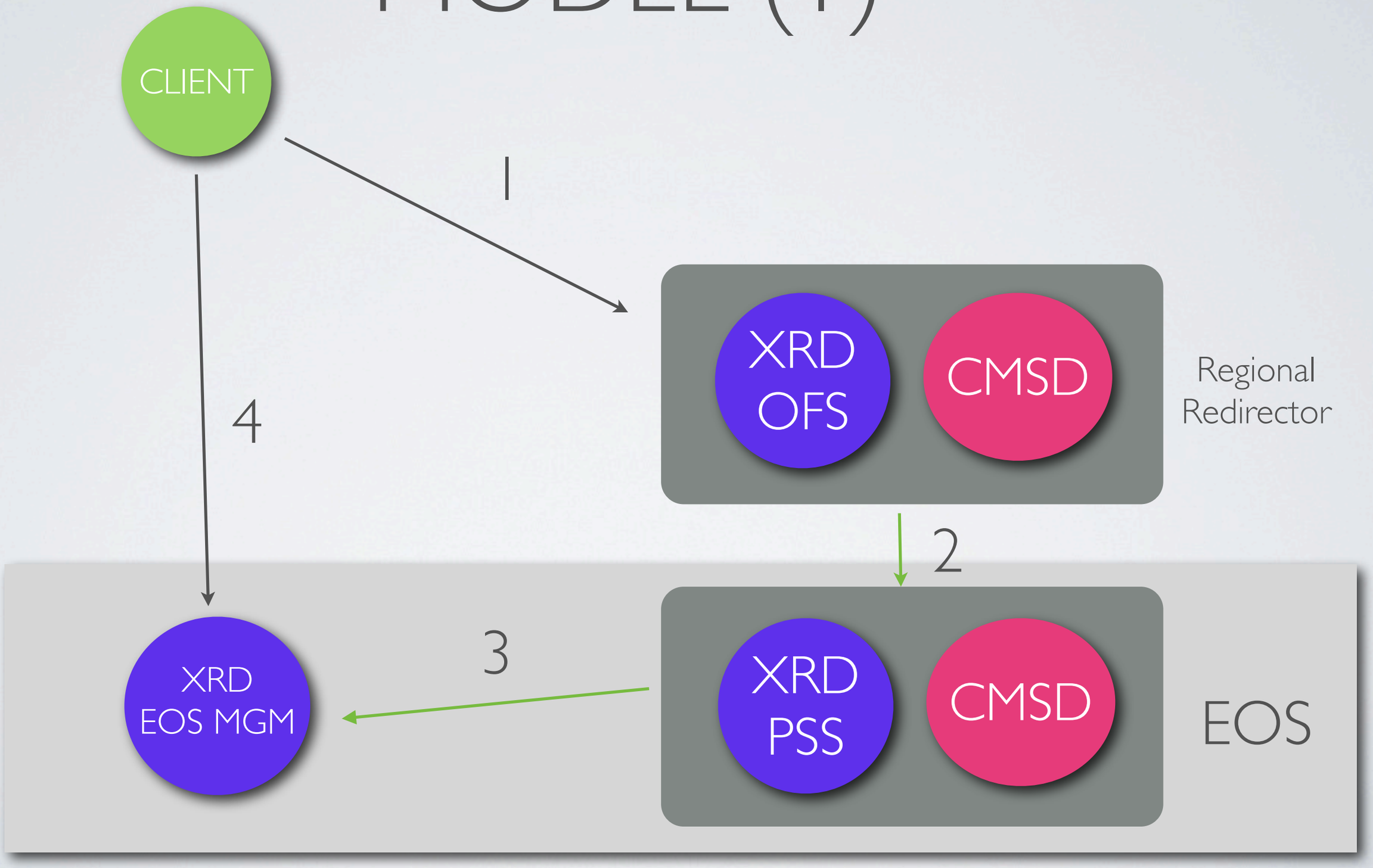
- locate requires stat performance (measured 24 kHz)
  all files are 'located' in a federation setup on the MGM node which then redirects to disk server based on real-time load measurements (weighted disk + network IO + random)
  (ok!)

- requires global access - WAN access to all EOS instances configured - throughput limit given by used network (OPN or world) (ok!)

- Federation access could saturate CERN networks and degrade T0-T1 replication etc. (alert!)

- requires additional storage performance
  - storage performance >>> WAN limits (ok!)

FEDERATION DEPLOYMENT MODEL (I)



FEDERATION DEPLOYMENT MODEL (2)

- Works for CMS
- ALICE requires several virtual SE entry points
  (once new client is in production ...)
- ATLAS model will require a modification of previous picture
  => see Elvin's talk!

CASTOR IN FEDERATION

- Primary use-case T1 (T0 not excluded)
- CASTOR federation should locate files which are currently staged => stager DB lookup
  - rate must be controlled
    - otherwise could also gain from an additional cache mechanism
  - the current stat implementation does a *stager query* by default and set's **`st_mode=-1`** and **`std_dev=0`**
  - test setup possible without development for trivial name translations or standard ATLAS N2N for 1:1 translation

SUMMARY

- Useful to subscribe EOS to federations <= Performance, Size
- *Trivial* Federation has been tried already with ATLAS,CMS
  - however experiment setups and naming conventions require extensions for ALICE & ATLAS
    - there is no complete solution for unified monitoring yet ...
      - architecture of OFS vs readV+monitoring
      - probably no show-stopper (f-stream sounds good)
- CASTOR might take part in federations in T1(0)