ALICE File Catalogue, SE ops and other topics

costin.grigoras@cern.ch

Current implementation

- MySQL-backed AliEn File Catalogue
  - 7 B logical entries, 8 TB disk footprint
- One DB primary instance
  - 768 GB RAM, 8 SSDs in RAID6
- Two DB replicas for hot standby / backups
  - 6 h to dump, ~5 days to restore
  - Daily backups copied to tape
- Similar, separate DB stacks for the task and transfer queues

Catalogue logical structure

Logical File Name (LFN)

Unix-like paths, eg "`/alice/data/2023/...`" that users and jobs see
Metadata (object type, size, owner, checksum)

Unique identifier (GUID)

New versions of an LFN get a new GUID (UUID v1 – time + MAC)
Any number of physical copies (URLs) associated to GUIDs

Physical files (PFN)

Storage Element ID + URL to the file
GUID-based algorithm to generate paths and file names

Immutable content

If we need to reuse the LFN ("edit" operation) a new GUID+PFNs will be used

One file in a nutshell

LFN Metadata (`stat` command in JAliEn):

```
File: /alice/cern.ch/user/g/grigoras/example
Type: f
Owner: grigoras:grigoras
Permissions: 755
Last change: 2023-09-18 11:40:13.0 (1695030013000)
Size: 72029 (70.34 KB)
MD5: 2e5f0c27e65400ea230c8a94be277b86
GUID: 5d9a8f1e-5607-11ee-90f9-6c02e09897e9
      GUID created on Mon Sep 18 11:40:13 CEST 2023 (1695030013023) by 6c:02:e0:98:97:e9
```

Metadata copied to the GUID record (`stat 5d9a8f1e-5607-11ee-90f9-6c02e09897e9`)

```
GUID: 5d9a8f1e-5607-11ee-90f9-6c02e09897e9
Owner: grigoras:grigoras
Permissions: 755
Size: 72029 (70.34 KB)
MD5: 2e5f0c27e65400ea230c8a94be277b86
Created: Mon Sep 18 11:40:13 CEST 2023 (1695030013023) by 6c:02:e0:98:97:e9
Last change: 2023-09-18 11:42:19.0 (1695030139000)
```

Physical copies (`whereis` on either the LFN or the GUID)

```
SE => ALICE::CERN::EOS          pfn => root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
SE => ALICE::ISS::EOS        pfn => root://mgm.spacescience.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
SE => ALICE::ORNL::EOS       pfn => root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
SE => ALICE::LBL_HPCS::EOS   pfn => root://alicemgm0.lbl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
```

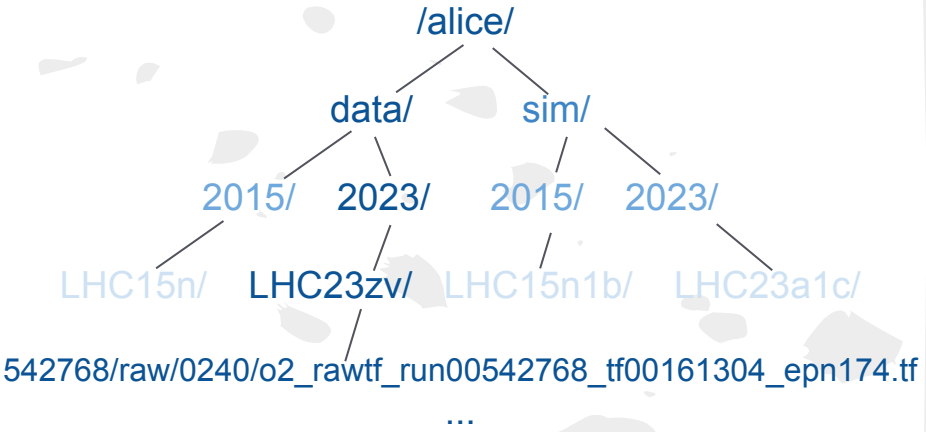
LFN namespace

Namespace is hierarchically split into tables

- 5336 tables (largest 80 M entries)
- 7 B entries in total (including folders)

Tables are split preemptively or as needed to keep them reasonably sized

- per user account
- MC production start
- RAW data new period



LFN tables

Locate the shard for that part of the tree

```
mysql> select tableName, lfn from INDEXTABLE where '/alice/data/2023/LHC23zv/542768/raw/0240/o2_rawtf_run00542768_tf00161304_epn174.tf' like concat(lfn,'%')
       order by length(lfn) desc limit 1;
+------------+--------------------------+
| tableName  | lfn                      |
+------------+--------------------------+
| 1054386710 | /alice/data/2023/LHC23zv/|
+------------+--------------------------+
```

Operation actually done in memory

Get the metadata for that LFN

```
mysql> select lfn, entryId, dir, owner, gowner, perm, type, ctime, jobid, size, md5, binary2string(guid), expiretime from L1054386710L where
       lfn='542768/raw/0240/o2_rawtf_run00542768_tf00161304_epn174.tf' \G
               lfn: 542768/raw/0240/o2_rawtf_run00542768_tf00161304_epn174.tf
           entryId: 156613
               dir: 152205
             owner: alidaq
            gowner: alidaq
              perm: 755
              type: f
             ctime: 2023-09-08 02:50:01
             jobid: NULL
              size: 1865709680
               md5: 9a6bf26f3dd0b8f2fbf61bd66ea426dd
binary2string(guid): A0ED1843-4DE1-11EE-8000-3CECEF03E9E2
        expiretime: NULL
```

GUID namespace

Version 1 UUIDs (date-time and MAC address)

```
$ uuid -d 7febbfd6-4ecb-11ee-8085-02428de94330
      …
      content: time:  2023-09-09 04:44:02.654818.2 UTC
                    clock: 133 (usually random)
                    node:  02:42:8d:e9:43:30 (local unicast)
```

Sharding on object creation time

- The date-time field of the UUID
- Dynamically, function of current chunk's size
- Switching to a new shard at 50 M entries / table

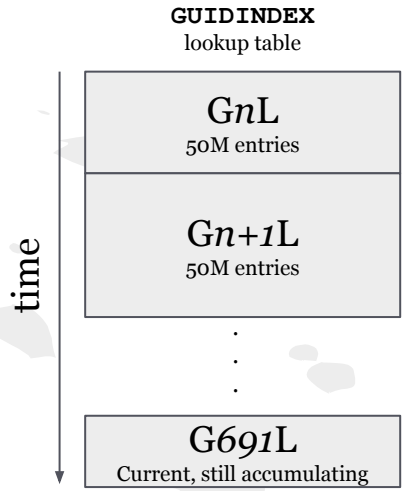
GUID namespace

Version 1 UUIDs (date-time and MAC address)

Sharding on object creation time

- The date-time field of the UUID
- Dynamically, function of current chunk's size
- Switching to a new shard at 50 M entries / table

650 tables, 6.9 B entries in total



GUID tables

The most recently created shard that contains the timestamp

```
mysql> select tableName from GUIDINDEX where guidTime < conv('1EE4C3910000000', 16, 10) order by guidTime desc limit 1;
+-----------+
| tableName |
+-----------+
|       691 |
+-----------+
```

Operation actually done in memory

From that shard extract the metadata for the given object ID

```
mysql> select guidId, ctime, owner, gowner, size, md5, seStringlist from GL691 where guid=string2binary('5d9a8f1e-5607-11ee-90f9-6c02e09897e9');
+----------+---------------------+----------+----------+-------+----------------------------------+---------------------+
| guidId   | ctime               | owner    | gowner   | size  | md5                              | seStringlist        |
+----------+---------------------+----------+----------+-------+----------------------------------+---------------------+
| 32890922 | 2023-09-18 11:42:19 | grigoras | grigoras | 72029 | 2e5f0c27e65400ea230c8a94be277b86 | ',332,382,370,373,' |
+----------+---------------------+----------+----------+-------+----------------------------------+---------------------+
```

Physical file pointers

Record of (Storage Element ID, full URL to the content)

(332, root://eosalice.cern.ch:1094//06/44195/7febbfd6-4ecb-11ee-8085-02428de94330)

Associated to GUIDs (not a separate namespace)

7.7 B entries

Some (most) URLs point to ZIP archive members

(no_se, guid:///7febbfd6-4ecb-11ee-8085-02428de94330?ZIP=AnalysisResults.root)

2.3 B physical files

Distributed in 67 storage elements worldwide

335 PB of data in total (40% on tapes)

PFN tables

Same shard table names as the GUID objects

```
mysql> select seNumber, pfn from G691L_PFN where guidId=32890922;
+----------+-----------------------------------------------------------------------------------------+
| seNumber | pfn                                                                                     |
+----------+-----------------------------------------------------------------------------------------+
|      332 | root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9             |
|      382 | root://mgm.spacescience.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9          |
|      373 | root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9         |
|      370 | root://alicemgm0.lbl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9            |
+----------+-----------------------------------------------------------------------------------------+
```

Additional helper tables, like the SE objects

```
mysql> select seNumber, seName, seQoS, seioDaemons, seStoragePath, seUsedSpace, seNumFiles, sedemotewrite, sedemoteread from SE where seNumber in (332,382,373,370);
+----------+---------------------+-------------------------------------+-------------------------------+---------------+-------------------+------------+---------------+--------------+
| seNumber | seName              | seQoS                               | seioDaemons                   | seStoragePath | seUsedSpace       | seNumFiles | sedemotewrite | sedemoteread |
+----------+---------------------+-------------------------------------+-------------------------------+---------------+-------------------+------------+---------------+--------------+
|      332 | ALICE::CERN::EOS    | ,disk,legooutput,cineca,http,tpcidc,| root://eosalice.cern.ch:1094  | /             | 48040634390738920 |  734753771 |     -0.513661 |            0 |
|      370 | ALICE::LBL_HPCS::EOS| ,disk,legoinput,legooutput,http,ccdb,| root://alicemgm0.lbl.gov:1094 | /             |  2516113854927926 |   55258625 |     -0.512317 |            0 |
|      373 | ALICE::ORNL::EOS    | ,disk,                              | root://ornl-eos-01.ornl.gov:1094 | /          |  1577728286581805 |   35641743 |     -0.545348 |  0.000297619 |
|      382 | ALICE::ISS::EOS     | ,disk,http,                         | root://mgm.spacescience.ro:1094 | /           |  4160034648906618 |   58062164 |     -0.379756 |            0 |
+----------+---------------------+-------------------------------------+-------------------------------+---------------+-------------------+------------+---------------+--------------+
```

PFN URL is automatically generated, as:

`seioDaemon / seStoragePath / hash1(guid) / hash2(guid) / guid`

From LDAP

Job uploading results

```
Output = {
    "log_archive.zip:std*,metrics_summary.json,perf*.json,core*@disk=2",
    "root_archive.zip:bcRanges.root,AnalysisResults.root@disk=2",
    "bcSelection.root@disk=2"
    };
```

Yields the following in its OutputDirectory

```
> ls -la
-rwxr-xr-x     alidaq    alidaq     18302361 Sep 09 06:44      AnalysisResults.root
-rwxr-xr-x     alidaq    alidaq        14284 Sep 09 06:44      bcRanges.root
-rwxr-xr-x     alidaq    alidaq      6827586 Sep 09 06:44      bcSelection.root
-rwxr-xr-x     alidaq    alidaq        44930 Sep 09 06:44      log_archive.zip
-rwxr-xr-x     alidaq    alidaq         6182 Sep 09 06:44      metrics_summary.json
-rwxr-xr-x     alidaq    alidaq     18316885 Sep 09 06:44      root_archive.zip
-rwxr-xr-x     alidaq    alidaq          539 Sep 09 06:44      stderr.log
-rwxr-xr-x     alidaq    alidaq       463150 Sep 09 06:44      stdout.log
```

And the location can be queried with:

```
> whereis AnalysisResults.root
  ZIP archive member            pfn => guid:///  7febbfd6-4ecb-11ee-8085-02428de94330?ZIP=AnalysisResults.root

> whereis root_archive.zip
  SE => ALICE::CERN::EOS        pfn => root://eosalice.cern.ch:1094//06/44195/   7febbfd6-4ecb-11ee-8085-02428de94330
  SE => ALICE::CNAF::SE         pfn => root://alice-test-xrootdgpfs.cr.cnaf.infn.it:1094//06/44195/    7febbfd6-4ecb-11ee-8085-02428de94330
```

Tag matching, i.e. @disk=2

Storage tags are arbitrary strings, defined in LDAP and copied to the database

JAliEn command `> listSEs -q disk`

		Capacity				Demote			
SE name	ID	Total	Used	Free	File count	Read	Write	QoS	Endpoint URL
ALICE::LBL_HPCS::EOS	370	2.749 PB	2.235 PB	526.6 TB	55259708	0.0000	-0.5402	disk, legoinput, legooutput, http, ccdb	root://alicemgm0.lbl.gov:1094/
ALICE::ORNL::EOS	373	2.725 PB	1.402 PB	1.323 PB	35643399	0.0003	-0.5090	disk	root://ornl-eos-01.ornl.gov:1094/

Demotion value function of recent read and write test results & amount of free space

0 = all's well ; **> 0** when tests are failing / running out of space ; **negative** = artificially biased to attract more data

When matching clients to storages the *demotion* is added to a *distance* metric

Based on network topology and/or geographical location

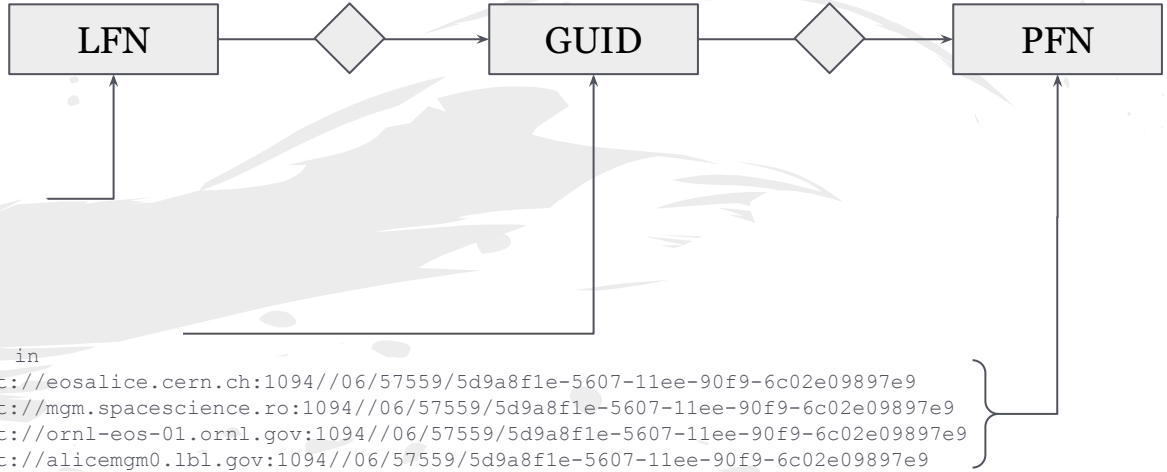
Manually checking what jobs at a site would get when trying to upload something

```
> listSEDistance -qos  disk -site ORNL
       ALICE::ORNL::EOS        (read:  0.000,  write: -0.509,  distance: -0.509)
       ALICE::LBL_HPCS::EOS    (read:  0.000,  write: -0.540,  distance: -0.413)
…
```

Useful JAliEn commands

`whereis`	- PFNs for a LFN or GUID
`stat`	- metadata dump (LFN or GUID)
`xrdstat`	- check the status of the PFNs
`guid2lfn`	- look up the LFN(s) pointing to the GUID
`listSEs`	- details on the SEs
`listSEDistance`	- debugging read and write access of jobs
`mirror`	- schedule a transfer to create more physical copies of a file
`deleteMirror`	- delete one of the replicas of a file
`access`	- get the tokens to read/write/delete files
`testSE`	- perform functional tests and get info from the SE

whereis

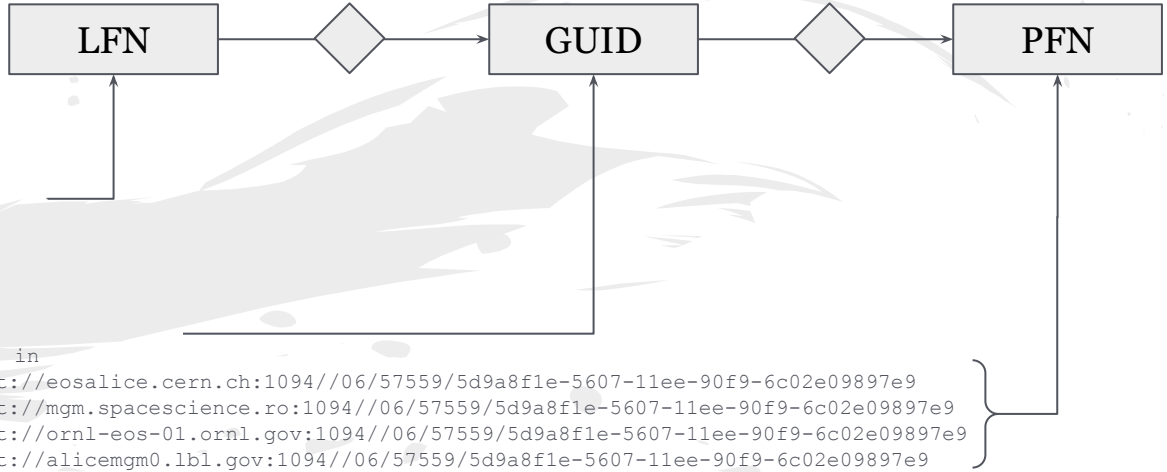



```
> whereis /alice/cern.ch/user/g/grigoras/example
the file example is in
…
```

(or equally well)

```
> whereis 5d9a8f1e-5607-11ee-90f9-6c02e09897e9
the GUID 5d9a8f1e-5607-11ee-90f9-6c02e09897e9 is in
        SE => ALICE::CERN::EOS        pfn => root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
        SE => ALICE::ISS::EOS         pfn => root://mgm.spacescience.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
        SE => ALICE::ORNL::EOS        pfn => root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
        SE => ALICE::LBL_HPCS::EOS    pfn => root://alicemgm0.lbl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
```

whereis




```
> whereis /alice/cern.ch/user/g/grigoras/example
the file example is in
…
```

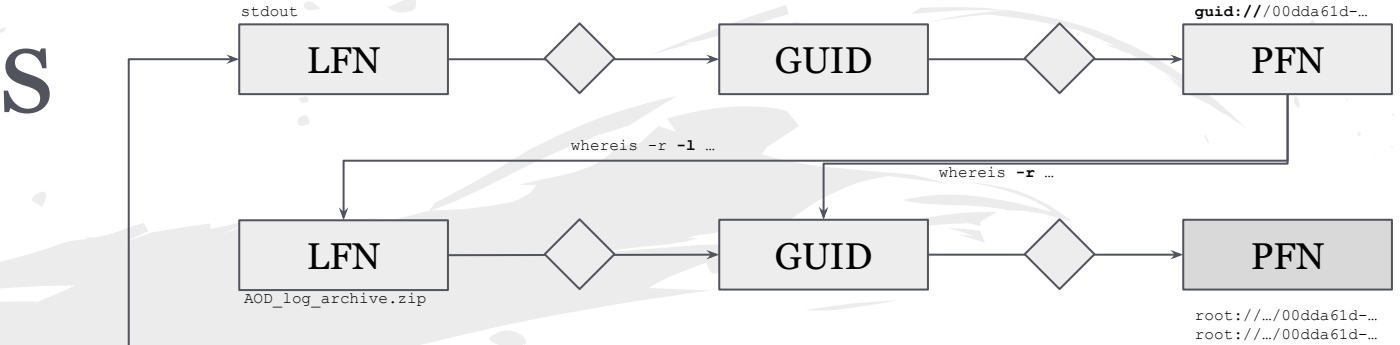
(or equally well)

```
> whereis 5d9a8f1e-5607-11ee-90f9-6c02e09897e9
the GUID 5d9a8f1e-5607-11ee-90f9-6c02e09897e9 is in
        SE => ALICE::CERN::EOS        pfn => root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
        SE => ALICE::ISS::EOS         pfn => root://mgm.spacescience.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
        SE => ALICE::ORNL::EOS        pfn => root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
        SE => ALICE::LBL_HPCS::EOS    pfn => root://alicemgm0.lbl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
```

For members of an archive

```
> whereis stdout
the file stdout is in
        ZIP archive member            pfn => guid:///00dda61d-5c55-11ee-8085-b82a72dd62c0?ZIP=stdout
```

whereis



Only these are visible on the SEs

For members of an archive

```
> whereis stdout
the file stdout is in
       ZIP archive member              pfn => guid:///00dda61d-5c55-11ee-8085-b82a72dd62c0?ZIP=stdout
```

Extra arguments to follow the GUID pointer to the physical replicas of the ZIP archive

```
> whereis -r stdout
the file stdout is inside a ZIP archive
       pass `-l` to whereis to try to resolve the archive LFN (slow, expensive operation!)
       SE => ALICE::CERN::EOS        pfn => root://eosalice.cern.ch:1094//12/13586/00dda61d-5c55-11ee-8085-b82a72dd62c0   ?ZIP=stdout
       SE => ALICE::BRATISLAVA::SE   pfn =>
root://lcgstorage04.dnp.fmph.uniba.sk:1094//12/13586/00dda61d-5c55-11ee-8085-b82a72dd62c0   ?ZIP=stdout

> whereis -r -l stdout
the file stdout is inside a ZIP archive
       archive LFN: /alice/sim/2023/LHC23f4b2/1/528451/AOD/001/AOD_log_archive.zip
       SE => …
```

stat

```
> stat -v /alice/cern.ch/user/g/grigoras/example
File: /alice/cern.ch/user/g/grigoras/example
Type: f
Owner: grigoras:grigoras
Permissions: 755
Last change: 2023-09-18 11:40:13.0 (1695030013000)
LFN shard: 8 / 10
Size: 72029 (70.34 KB)
MD5: 2e5f0c27e65400ea230c8a94be277b86
GUID detailed information:
       GUID: 5d9a8f1e-5607-11ee-90f9-6c02e09897e9
       Owner: grigoras:grigoras
       Permissions: 755
       Size: 72029 (70.34 KB)
       GUID shard: 8 / 691
       MD5: 2e5f0c27e65400ea230c8a94be277b86
       Created: Mon Sep 18 11:40:13 CEST 2023 (1695030013023) by 6c:02:e0:98:97:e9
       Last change: 2023-09-18 11:42:19.0 (1695030139000)
       Replicas:
              SE => ALICE::CERN::EOS        pfn => root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
              SE => ALICE::ISS::EOS         pfn => root://mgm.spacescience.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
              SE => ALICE::ORNL::EOS        pfn => root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
              SE => ALICE::LBL_HPCS::EOS    pfn => root://alicemgm0.lbl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
```

xrdstat

```
> xrdstat /alice/cern.ch/user/g/grigoras/example
Checking the replicas of /alice/cern.ch/user/g/grigoras/example
        ALICE::CERN::EOS          root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9  OK
        ALICE::ISS::EOS           root://mgm.spacescience.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9  OK
        ALICE::ORNL::EOS          root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9 OK
        ALICE::LBL_HPCS::EOS      root://alicemgm0.lbl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9 OK
```

Checks just the metadata (file exists and has the same size as the catalogue information).
Relies on `` `xrdfs eosalice.cern.ch stat /06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9` ``.

```
 > xrdstat -v -d -c -s ALICE::ORNL::EOS /alice/cern.ch/user/g/grigoras/example
Checking the replicas of /alice/cern.ch/user/g/grigoras/example
        ALICE::ORNL::EOS          root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9 OK
                Downloaded file matches the catalogue details, retrieving took 2s (42.22 KB/s)
export XRD_CONNECTIONWINDOW="3"
export XRD_CONNECTIONRETRY="1"
export XRD_TIMEOUTRESOLUTION="1"
export XRD_PREFERIPV4="1"
export XRD_APPNAME="JBox"
export XRD_REQUESTTIMEOUT="60"
/home/costing/xrootd/bin/xrdcp 'root://ornl-eos-01.ornl.gov:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9?authz=-----BEGIN SEALED CIPHER-----
…
&eos.app=JBox' /tmp/xrdcheck-10893398677280054627-download.tmp
```

The full command above, including the read envelope, can be (re)used for debugging files that are not accessible.
By issuing an `` `xrdcp` `` after the `` `xrdfs stat` ``, we check both the metadata and the consistency on disk.
After downloading, the MD5 checksum of the file is compared with the catalogue information so corrupted content can be identified.

guid2lfn

```
> guid2lfn 5d9a8f1e-5607-11ee-90f9-6c02e09897e9
5d9a8f1e-5607-11ee-90f9-6c02e09897e9    /alice/cern.ch/user/g/grigoras/example
```

Very heavy operation as there is no pointer back from a GUID to the LFN(s) holding it. The command scans all 5000+ LFN tables to find a match.

Useful only for debugging job logs / identifying and cleaning up corrupted files from the catalogue.

Or if a file name shows up frequently in the SE logs…

mirror

```
> mirror /alice/cern.ch/user/g/grigoras/example ALICE::UPB::EOS
ALICE::UPB::EOS: queued transfer ID 865199235

> listTransfer -id 865199235
TransferId          Status       User         Destination              Size          Attempts     File name
865199235           DONE         grigoras     ALICE::UPB::EOS          72029                      3
/alice/cern.ch/user/g/grigoras/example

> whereis example
... (all the previous ones plus)
       SE => ALICE::UPB::EOS            pfn => root://eos-mgm.grid.pub.ro:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9

> deleteMirror example ALICE::UPB::EOS
Mirror scheduled to be deleted from ALICE::UPB::EOS
```

Other forms

```
 > mirror -S ccdb:2 /alice/cern.ch/user/g/grigoras/example
ALICE::KFKI::SE: queued transfer ID 865199519
ALICE::CNAF::CEPH: queued transfer ID 865199518

> mirror -r ALICE::KFKI::SE example ALICE::UPB::EOS
ALICE::UPB::EOS: queued transfer ID 865199541
```

Same *tag:count* syntax as in the JDLs
This is how we decouple calibration or EPN jobs from writing to external SEs

A move operation, after creating the target copy, the indicated replica is deleted

access

```
> access read /alice/cern.ch/user/g/grigoras/example
root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
SE: ALICE::CERN::EOS (needs encrypted envelopes)
Encrypted envelope:
-----BEGIN SEALED CIPHER-----
…
-----END SEALED ENVELOPE-----
```

```
> access -u read /alice/cern.ch/user/g/grigoras/example
https://eosalice.cern.ch:443//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
SE: ALICE::CERN::EOS (needs encrypted envelopes)
Encrypted envelope:
-----BEGIN%20SEALED%20CIPHER-----%0A…0A-----END%20SEALED%20ENVELOPE-----%0A
```

To be used respectively in :

```
xrdcp root://eosalice.cern.ch:1094//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9?authz=<multi line envelope> /tmp/local.copy
curl -kL https://eosalice.cern.ch:443//06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9?authz=<urlencoded envelope> -o /tmp/local.copy
```

I could have used LBL_HPCS::EOS as an example but it doesn't work atm

```
$ curl -kL https://alicemgm0.lbl.gov:8443//06/…
curl: (7) Failed to connect to alicemgm0.lbl.gov port 8443 after 303 ms: Connection refused
```

Envelope testing

http://alimonitor.cern.ch/work/decryptToken.jsp

Enter token here:

```
-----BEGIN SEALED CIPHER-----
YyBiAC2Wev8eim+SqaPimZ-ZtgutkwMFb624wZ3-S9VdwYZpuFrDOg5skRBvMiq
pvxmyP-kjjcQfjfY55WEiNiLIDaJPu4guyGV+0wHfgL44jLOkqnuoJUBfraL9hU
Ru5biKmjTORRmzGB+L4=
-----END SEALED CIPHER-----
-----BEGIN SEALED ENVELOPE-----
AAAAgDDRjFLPjUqEbkQpcPOVbHlPcW9sE+SnTDjPOkZU-DIaRmfkeYZuTU1Gl+9
9MmQ7OoyeVRiQlDvZ7wqMiOI+zhOMLbh1VGZbPsC-Z1z8bkjCtQhl7vSHNIGoVO
OnDBPdmNZwCbr0KidKI4GGb8CQImhpY80vmqsi29s1v8tB9wSzoz4MopxVTg5iP
SdsIB9NO3oa5oku5v2U8k-XwQkrwHyDA1zUTKdkiwYycPaB0el3K36PRb3Yhdq-
0hunqW-QqdmXcPpyCJ-oNQdJuKI1Vn+hJGvSr7g02xNWCnfHUMerrK681NqCVw6
I7Nrm2C5LCWfHLfwIvWLCaQna7wu7OhnKK2beR0PY2w9oOxg3J9agI-QoqlDvf-
pVxHO3iPuVg-9v2lkWRp6QNq9Bs48axQa9HJ4Spm4pmDWw-a2G8hL8mugtHWGvm
aEzxGxlgQ1pV5OixsKIxd7GXihm5fzffqhCQI0E5qgQCRBLIMBR4QwYDRJdoeXM
I90RpHG6Cv9HGvAb9U5KLxn6xhAHqUQfu8hqJW5hX35mcy3xUSl53xWiHExbt49
c98u1Vevag2nO1g+x1zRHpxOWGLMYkDNNvBu01w5Ff+FeNA-AAJ7LAAKu9+KKsa
EsAoFYHPFlz2O6Tnhcm5QxwloMwguWZIIo8IrG+YWja3xUCkzXR6AxVxq7i+ztv
QmaHiOIaFWeB9PHtbqaIaQ4CPJve1ZeBOze-p8mEI7BgaCK+KnOq8vXWC0doLF7
0zaa+qSmN+e4Sb-jVWviQFx1dgJ7oj5q-m9LwjbboQTrYZMH0b5eJsVz0ipR71o
uldA5XO60mZx3EHprkh5H6oPvlbBqpSPMcAE4vvbf7qpHI8xkIKumI+PVNvNTG3
-----END SEALED ENVELOPE-----
```

Submit Query

```
-----BEGIN ENVELOPE-----
CREATOR:     AuthenX
UNIXTIME:    1696436689
DATE:        Wed Oct  4 18:24:49 2023
EXPIRES:     0
EXPDATE:     never
CERTIFICATE: none
-----BEGIN ENVELOPE BODY-----
<authz>
  <file>
    <access>write-once</access>
    <turl>root://ornl-eosprf-01.ornl.gov:1094//12/46771/2b19222f-468e-11ed-8f89-3cecef04a72e</turl>
    <lfn>/NOLFN</lfn>
    <size>2113029846</size>
    <guid>2B19222F-468E-11ED-8F89-3CECEF04A72E</guid>
    <md5>bea633347e69791dd36b1a2b263ef1c6</md5>
    <pfn>/12/46771/2b19222f-468e-11ed-8f89-3cecef04a72e</pfn>
    <se>ALICE::ORNL::PRF_EOS</se>
  </file>
</authz>

-----END ENVELOPE BODY-----
-----END ENVELOPE-----
```

testSE

```
> testSE ALICE::LBL_HPCS::EOS
 Open write test: cannot write (expected)
 Authenticated write test: could write (expected)
 Open read test: read back failed (expected)
 Open HTTP read test: read back failed  (expected)
 Authenticated read: file read back ok (expected)
 Authenticated HTTP read access: read back failed NOT OK)
 Open delete test: delete failed (expected)
 Authenticated delete: delete worked ok (expected)
 Space information:
Path:  /
Total: 3.48 PB (LDAP setting: 3825766236160)
Free:  1.288 PB
Used:  2.191 PB
Chunk: 64 GB
Version: Unknown 5.5.10
 LDAP information:
SE: seName: ALICE::LBL_HPCS::EOS
seNumber        : 370
seVersion       : 0
qos     : [disk, legoinput, legooutput, http, ccdb]
seioDaemons     : root://alicemgm0.lbl.gov:1094
seStoragePath   : /
seSize: : 3825766236160
seUsedSpace     : 2521997480791012
seNumFiles      : 55551585
seMinSize       : 0
seType  : File
exclusiveUsers  : []
seExclusiveRead : []
seExclusiveWrite        : []
options:        {https_port=8443}
```

testSE

```
> testSE -v -c ALICE::ORNL::EOS
 Open write test: could write,  ( NOT OK), please check authorization configuration
…xrdcp --nopbar --verbose --force --posc --cksum md5:source /etc/hostname root://ornl-eos-01.ornl.gov:1094//04/59628/1df2df96-592e-11ee-989a-6c02e09897e9
 Open read test: reading worked ( NOT OK) please check authorization configuration
…xrdcp root://ornl-eos-01.ornl.gov:1094//04/59628/1df2df96-592e-11ee-989a-6c02e09897e9?eos.app=JBox /tmp/xrootd-get12072414171333287299.tmp
 Open HTTP read test: read back failed  ( expected)
 Open delete test: delete worked ( NOT OK)
…xrdfs ornl-eos-01.ornl.gov:1094 rm /04/59628/1df2df96-592e-11ee-989a-6c02e09897e9
 Authenticated write test: could write ( expected)
 Authenticated read: file read back ok ( expected)
 Authenticated HTTP read access: read back failed  ( NOT OK)
No route to host
 Authenticated delete: delete worked ok ( expected)
```

Storage operations

Catalogue pointers to SE files

`SEUtils.masterSE(true, "ALICE::ORNL::EOS");`

Produces a "*ALICE::ORNL::EOS.file_list*" with the dump of all PFNs that we expect to be there

```
#PFN,size,MD5,ctime,guid
root://ornl-eos-01.ornl.gov:1094//02/09446/14ed401c-548d-11dc-99d5-000423b5ab42,100184,ac083478db8c77cd2e651d51646d6da9,1188212583248,14ED401C-548D-11DC-99D5-000423B5AB42
root://ornl-eos-01.ornl.gov:1094//11/38803/996bcdfa-f122-11e0-91fd-6709dc177d98,128537807,a9801309a6c800708bccd9db166e8aba,1318019172627,996BCDFA-F122-11E0-91FD-6709DC177D98
```

Remote listing of SE content

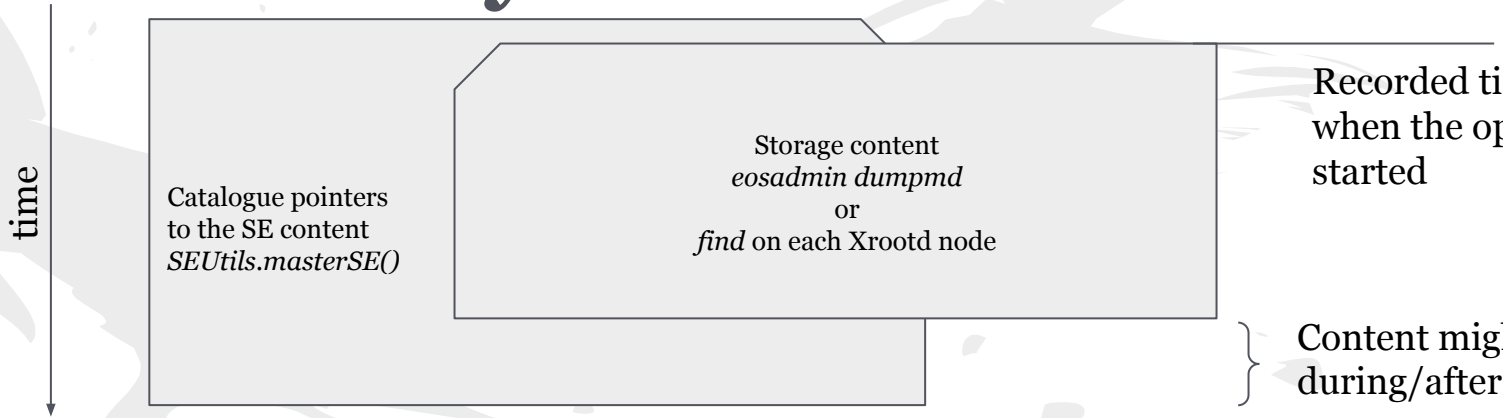
```
$ xrdfs eosalice.cern.ch ls -l /06
…
drwxr-xr-x aliprod        z2 2642650605680 2023-09-21 14:00:16 /06/57543
drwxr-xr-x aliprod        z2 2241740340036 2023-09-21 14:12:48 /06/57559
drwxr-xr-x aliprod        z2 1454379360762 2023-09-21 14:06:10 /06/57574
…

$ xrdfs eosalice.cern.ch ls -l /06/57559
…
-rw-rw-r-- aliprod        z2       527170 2021-10-07 08:05:42 /06/57559/5d97cf18-2745-11ec-bf72-0baed1c3d396
-rw-r--r-- aliprod        z2        72029 2023-09-18 09:40:13 /06/57559/5d9a8f1e-5607-11ee-90f9-6c02e09897e9
-rw-rw-r-- aliprod        z2        29686 2020-01-02 11:45:25 /06/57559/5d9afe7c-2d55-11ea-aa51-275c57918c1f
…
```

Recursive listing and matching the catalogue content: `XrootdCleanupSingle`

Slow operation; only removes files on the SE that are not in the catalogue (dark data)

In the general case we cannot rely on the output of `` `xrdfs ls` `` to remove pointers from the catalogue

Content synchronization



Potential data loss

Potential dark data

AliEn SE		Catalogue statistics						Storage-provided information						Functional tests			
AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version	add	get	rm	3rd
ALICE::LBL_HPCS::AF_EOS	2	1.101 PB	161 GB	1.101 PB	0.014%	742	FILE	1.101 PB	8.024 TB	1.093 PB	0.712%						
ALICE::LBL_HPCS::EOS	2	3.48 PB	2.24 PB	1.24 PB	64.38%	55,549,440	FILE	3.48 PB	2.191 PB	1.288 PB	62.97%	Xrootd 5.5.10	5.1.27				
ALICE::ORNL::EOS	2	4.943 PB	1.737 PB	3.206 PB	35.14%	36,395,889	FILE	4.943 PB	1.793 PB	3.151 PB	36.26%	Xrootd v5.5.5	5.1.27				
ALICE::ORNL::PRF_EOS	2	807.7 TB	234.7 TB	572.9 TB	29.06%	115,878	FILE	807.7 TB	240.4 TB	567.3 TB	29.76%		5.1.14				

LDAP static value, manually updated

`sum(size)` from all G*L_PFN tables

`count(*)` from all G*L_PFN tables

`xrdfs <endpoint> spaceinfo /` or `xrdfs <endpoint> query space /` (updated hourly)

`rpm -qa {xrootd,eos-server,eos-xrootd}` from *MLSensor* or *eosapmond*, or `xrdfs <endpoint> query config version`

Content synchronization

Two input lists

- Storage-provided list of files (+timestamp)
- Catalogue content (up to that timestamp)

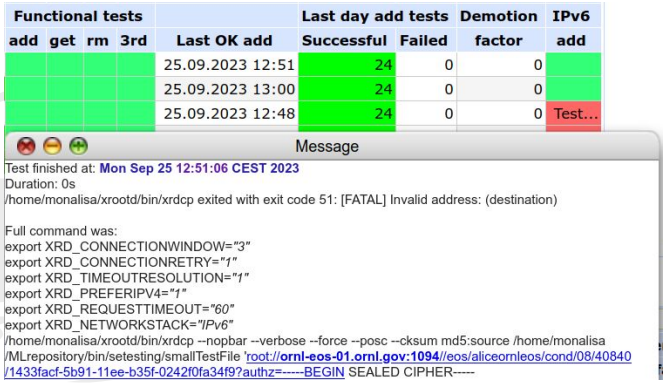
Two output lists

- Missing from catalogue: dark data / can be deleted
- Missing from SE: to recover / notify the catalogue
  - Mirror back if it has other copies
  - Delete the pointers if this was the only replica of that file
    - Delete the GUID + all ZIP archive GUIDs
    - Find and delete all LFNs pointing to the physical file or the ZIP entries

Intersecting the two sets and performing the necessary actions are implemented by `Recover`

- Operating on GUID and file size alone
- Metadata dumps are not reliable without background consistency checks

Functional tests



Functional tests					Last day add tests		Demotion	IPv6
add	get	rm	3rd	Last OK add	Successful	Failed	factor	add
				25.09.2023 12:51	24	0	0	
				25.09.2023 13:00	24	0	0	
				25.09.2023 12:48	24	0	0	Test...

Similar to the `testSE` command

Once per hour, silently retried 3x if `add` fails

1. Add a fixed file (10MB) (includes `stat`) - new GUID every time
2. Read it back (+compare checksum) (or a different existing file if `add` fails)
3. Delete it
4. Use Third Party Copy to replicate a fixed file from CERN::EOS (`05422528-a162-11e8-b5f9-a310691b2def`)
5. Repeat 1-3 with IPv6 only, on a small file - only `add` is shown

Full command that failed can be copy-pasted to be retried

`add` and `get` history contributes to the Demotion factor

for write and respectively read operations

This page deliberately left blank

CCDB usecase

- CCDB files so far: 7.2 M / 1.6 TB
  - Average size: 243 KB
  - Max size: 100 MB (57 objects under `TPC/Calib/CorrectionMap*` - 5 paths)
  - Distinct paths access during processing: 240
  - Total size of a set of CCDB objects: 640 MB => 2.6 MB/file
- CCDB objects must be found close to the processing node
  - ORNL and LBL copies are both needed
- The few large objects can create hotspots - internal replication solves them
  - Same for Run1/2 `OCDB{sim,rec}.root` snapshots - ORNL::CCDB also receives a copy
  - LBL - if hotspots are detected we can do the same

Infrastructure points

OS - CentOS 7 → AlmaLinux, or another variant of 9, on the host

Cgroups v2

- Depends on the OS to implement them
- And on the BQ to delegate the slices. Max is waiting for both Slurm and HTCondor to implement the patches.

CVMFS module - upgrade to `2.11.0` (*LBL_HPCS* is at `2.10.{0,1}`, *ORNL* at `2.9.3`), set and then reload `autofs`:

- `CVMFS_CACHE_REFCOUNT=1`
- `CVMFS_NFILES=1048576`

} `/etc/cvmfs/default.conf`

IPv6

	SE	VoBox	WNs
LBL	✅	❌	❌
ORNL	❌	❌	❌

HTTP(S) SE endpoint

- Command options are available to return it; not used by default
- Any stats on the service side on (attempts of) accesses on this port?

JAliEn changes since May

8 releases (1.7.3 to 1.8.0)

- 64b JDK used by default, bumped to Java 17
- Nested containers for JW and payloads
- Automatic resubmission of jobs in case of env setting up
- CE refresh of LDAP configuration and apply it to the submitted scripts without restart
  - And in case of multiple queues on a site, fix loading the correct one for a host
- CPU pinning - fix same core assignment to multiple payloads
- Job accounting data saved to *QUEUEPROC* (basis for job quotas)
- Fix limiting which accounts can run jobs on which queue
- Experimental *aarch64* support