Migrations and Recalls in CASTOR

Sébastien Ponce
CERN IT DSS

CASTOR Face to face meeting - 28-30 Nov 2012

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Overall program

- Migration of a CASTOR
  - MigrationJob
  - MigrationRouting, TapePools
  - MigrationMount
  - MigratedSegment
- Recall of a CastorFile
  - RecallJob
  - RecallGroup, RecallUser
  - RecallMount
- Global stager DB schema
- Miscellaneous

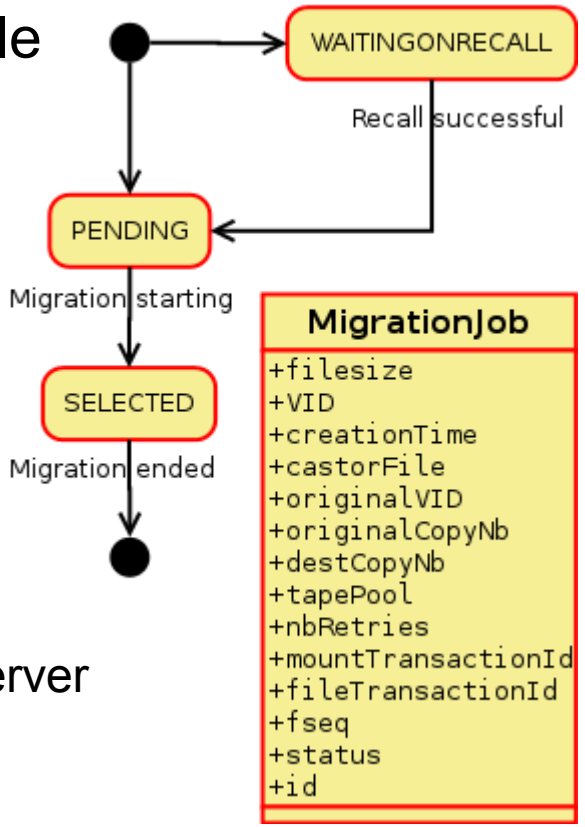
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Migrations

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The migration Job

- Represents a job of migration
  - That is a tape file to be created (from a disk one)
  - e.g. : you have 2 for a dual copy file
- Is created
  - On file closing for writes
    - immediately candidate for migration
  - On RecallJob creation for repack
    - Triggered on recall completion
    - Has data concerning old copy
      - originalVID, originalCopyNb
      - For easy/clean update of nameserver
- Is deleted
  - On successful migrationo or when nbRetries is reached
    - Creation of DiskCopy in CANBEMIGR with no counterpart



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The migrationRouting table

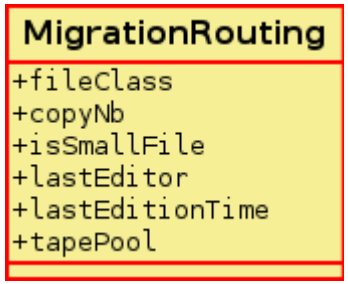
- Allows to decide where to migrate a file
  - As such replaced previous policies
- Content
  - Fileclass, copyNb, isSmallFile
    - input part. Checked by tools for consistency
  - TapePool
    - output part
  - LastEditor, lastEditiongTime
    - For maintainability
- Edition
  - Via dedicated tools
    - enter/delete/modify/printmigrationroute

MigrationRouting
+fileClass +copyNb +isSmallFile +lastEditor +lastEditionTime +tapePool

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The migrationRouting table

- Allows to decide where to migrate a file
  - As such replaced previous policies
- Content
  - Fileclass, copyNb, isSmallFile
    - input part. Checked by tools for consistency
  - TapePool
    - output part
  - LastEditor, lastEditiongTime




```
[root@c2cmssrv401 ~]# printmigrationroute
     FILECLASS COPYNB ISSMALLFILE       TAPEPOOL LASTEDITOR          LASTEDITION
-------------------------------------------------------------------------------
      c3_copy      1            -       cms_user        root 12-Mar-2012 11:08:35
          cms      1            - cmsfamily_new1        root 12-Mar-2012 11:05:10
cms_production     1            -    cms_prod_08        root 12-Mar-2012 11:05:10
       cms_raw     1            -     cms_raw_08        root 12-Mar-2012 11:05:10
  cms_streamer     1            -  cms_stream_08        root 12-Mar-2012 11:05:10
      cms_temp     1            -   cms_testdata        root 12-Mar-2012 11:05:11
      cms_test     1            -     cms_csa_07        root 12-Mar-2012 11:05:11
      cms_user     1            -       cms_user        root 12-Mar-2012 11:08:55
```

The TapePool

- A logical group of tape
- With rules for migration triggering
  - On volume (minAmountData)
  - On nb files (minNbFiles)
  - On time (maxFileAge)
- Condition for starting a new mount

  nbDrives not exhausted AND

  (min volume or nbFile still reached with extra mount OR

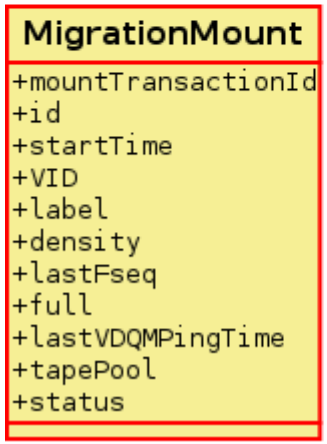
  No migration running and max file age reached for oldest file)
- Edition
  - Via dedicated tools
    - enter/delete/modify/printtapepool
    - tapePool must exist in VMGR

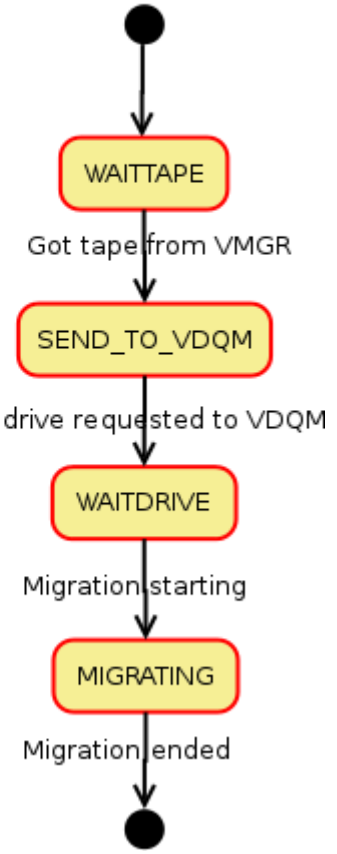
TapePool
+name +nbDrives +minAmountDataForMount +minNbFilesForMount +maxFileAgeBeforeMount +lastEditor +lastEditionTime +id

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The migration Mount

- Represents a mounted tape for migration
- Creation
  - Via DataBase job running every minute
  - Using the TapePool criteria
- Link to MigrationJob
  - During migration, via mountTransactionId
  - So that files are migrated only once (per copy)

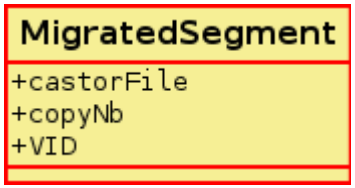




CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The migrated Segment

- Represents existing files on tape
- Created
  - together with migrationJob
    for existing copies of files to migrate
  - at the end of migration for multi copy files
- Allows to not put 2 copies on same tape
  - Within the same mount
  - When repacking or creating extra copy
- Deleted
  - At the end of the last migration of a file
    - When no MigrationJob remains for the file



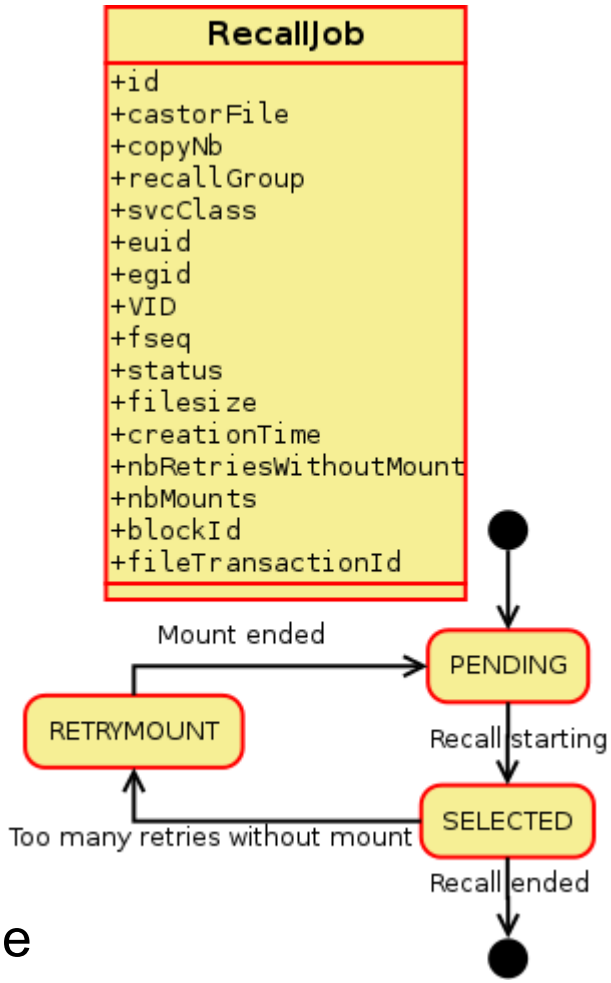
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Recalls

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The recall Job

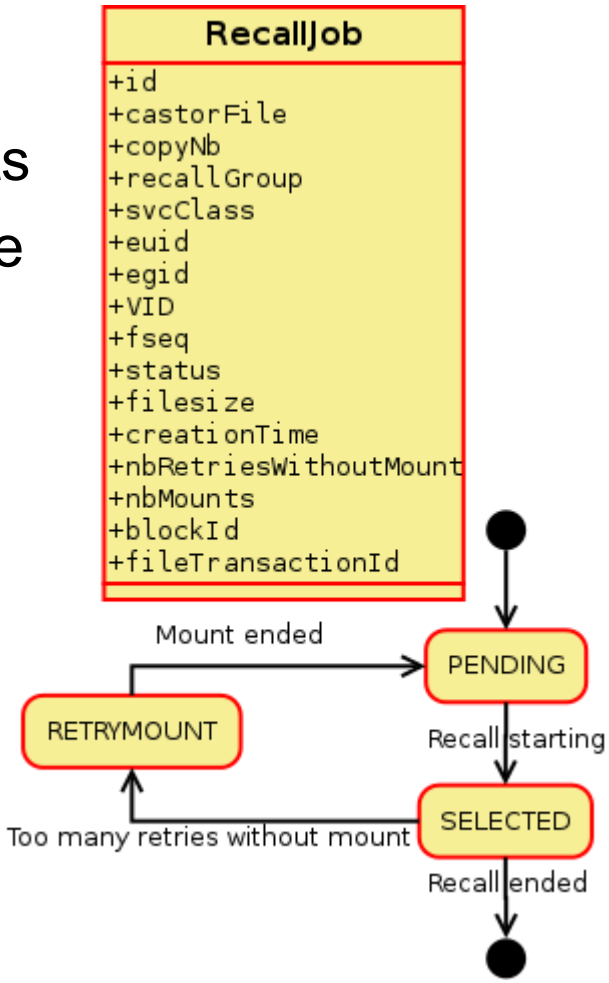
- Represents a potential recall
  - I.e. a tape file to be read potentially
  - e.g. : you have 2 for a dual copy file
    - But only one will be effectively used
- Is created
  - In standar case :
    - On cache miss in the stager
    - For each available tape copy
    - For each recall group
  - In repack case :
    - Only for the repacked copy
- Is deleted
  - On successful recall of the related file
    - all RecallJob for a file go with first successful recall
  - Or when nbRetries is reached
    - Only one RecallJob dropped, others can still go



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

More on recall Jobs

- 2 levels of retry
  - nbRetriesWithoutMount, nbMounts
  - Configurable in CastorConfig table
    - Default is 2 and 2
    - So 4 tries in total
- The svcClass member
  - Is a cache of the svcclass of the request that triggered the recall
  - Tells where to put the file on disk



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

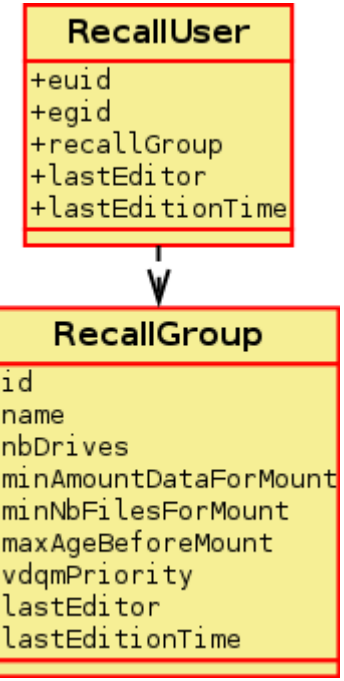
The recallGroup

- A logical group of RecallUser
  - Users can be defined by euid/egid
  - Or by group (null euid, egid given)
- With rules for recall triggering
  - Identical to TapePool for migration
  - Plus vdqmPriority
- Condition for a new mount

  nbDrives not exhausted AND

  (min volume or nbFile still reached with extra mount OR

  No recall running and max file age reached for oldest file)
- Edition via dedicated tools
  - enter/delete/modify/printrecallgroup
  - enter/delete/printrecalluser



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

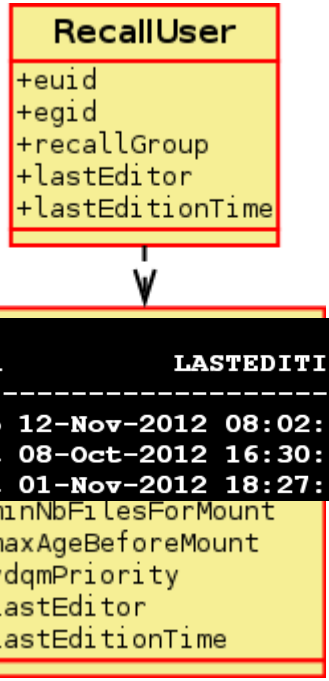
The recallGroup

- A logical group of RecallUser
  - Users can be defined by euid/egid
  - Or by group (null euid, egid given)
  - Plus vdqmPriority
- Condition for a new mount

  nbDrives not exhausted AND

  (min volume or nbFile still reached with extra mount OR

  No recall running and max file age reached for oldest file)
- Edition via dedicated tools
  - enter/delete/modify/printrecallgroup
  - enter/delete/printrecalluser

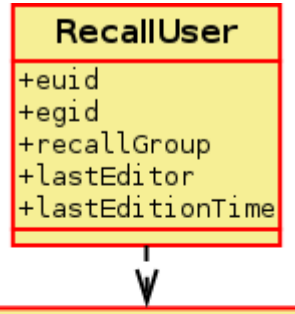



```
[root@c2cmssrv401 ~]# printrecallgroup
     NAME NBDRIVES MINAMOUNTDATA MINNBFILES MAXFILEAGE VDQMPRIORITY         ID LASTEDITOR          LASTEDITION
-------------------------------------------------------------------------------------------------------------
  default       20         10GiB         10        4h            0 23575325766 gcancio 12-Nov-2012 08:02:55
      vip       40        100GiB       1000      10mn          100 23575331556    root 08-Oct-2012 16:30:57
immediate        5           1B          1        5s         1000 23983267548    root 01-Nov-2012 18:27:07
```

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The recallGroup

- A logical group of RecallUser
  - Users can be defined by euid/egid
  - Or by group (null euid, egid given)
  - Plus vdqmPriority




```
[root@c2cmssrv401 ~]# printrecallgroup
     NAME NBDRIVES MINAMOUNTDATA MINNBFILES MAXFILEAGE VDQMPRIORITY         ID LASTEDITOR          LASTEDITION
-------------------------------------------------------------------------------------------------------------
  default       20         10GiB         10        4h            0 23575325766 gcancio 12-Nov-2012 08:02:55
      vip       40        100GiB       1000      10mn          100 23575331556    root 08-Oct-2012 16:30:57
immediate        5            1B          1        5s         1000 23983267548    root 01-Nov-2012 18:27:07
```

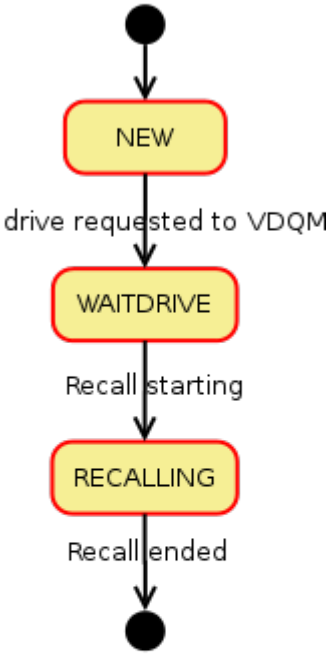
```
[root@c2cmssrv301 ~]# printrecalluser
    USER/GROUP (UID/GID) RECALLGROUP LASTEDITOR          LASTEDITION
-------------------------------------------------------------------
  phedex:zh (22014:1399)         vip       root 08-Oct-2012 16:31:52
   stage:st (14029:1474)         vip       root 08-Oct-2012 16:31:47
  cmsprod:zh (5410:1399)         vip       root 08-Oct-2012 16:31:42
cmsprod2:zh (50726:1399)         vip       root 08-Oct-2012 16:31:37
  relval:zh (31275:1399)         vip       root 08-Oct-2012 16:31:32
  wildish:zh (2907:1399)         vip       root 09-Oct-2012 16:28:47
castorc3:c3 (20395:1028)   immediate       root 01-Nov-2012 18:28:07
     iven:c3 (6925:1028)   immediate       root 01-Nov-2012 18:27:43
```

- enter/delete/modify/printrecallgroup
- enter/delete/printrecalluser

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The recall Mount

- Represents a mounted tape for recall
- Creation
  - Via DataBase job running every minute
  - Using the RecallGroup criteria
    - Checking all recallGroups
    - Using only RecallJobs for files not already handled

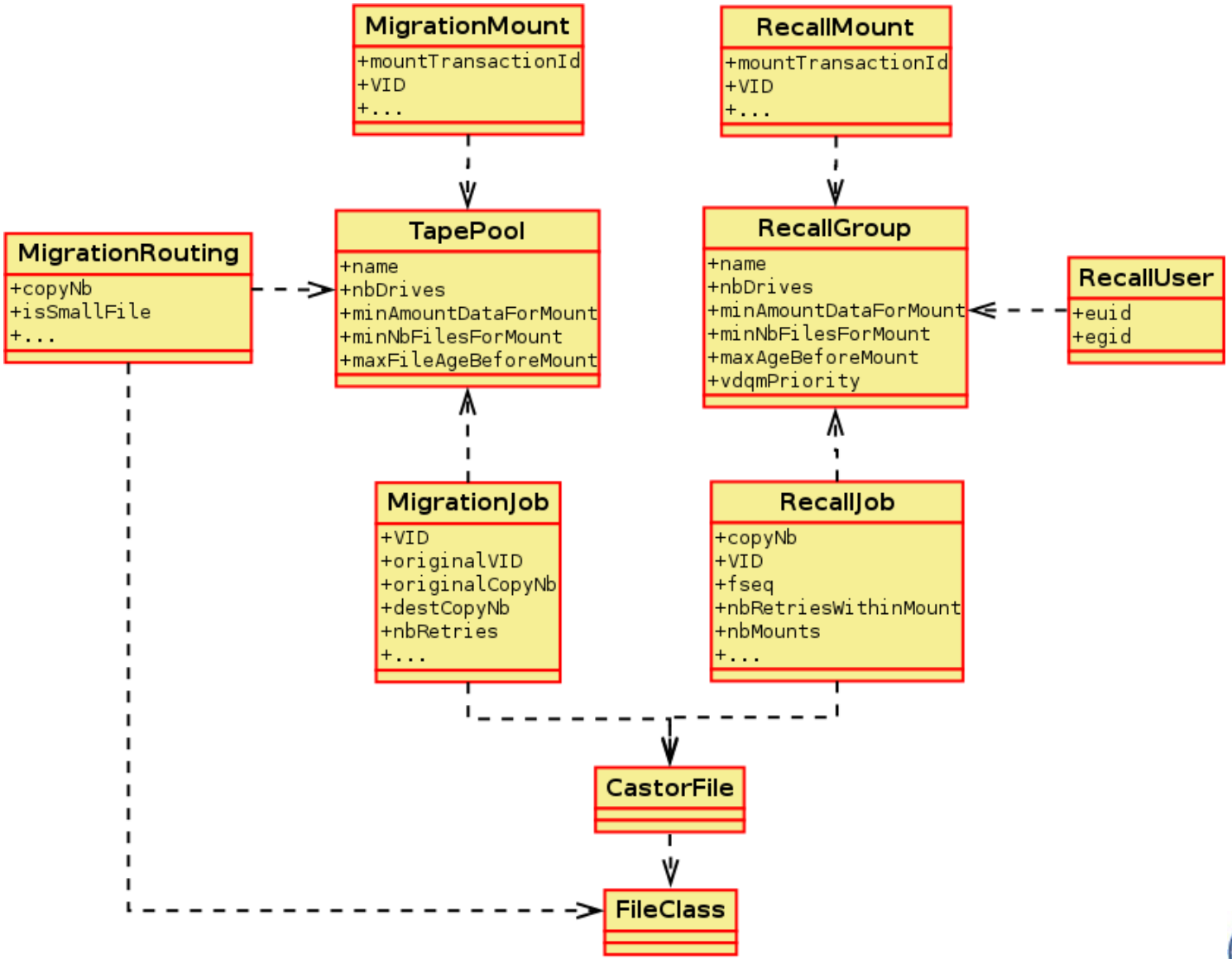


RecallMount

+id
+mountTransactionId
+VID
+label
+density
+recallGroup
+startTime
+status
+lastVDQMPingTime
+lastProcessedFseq

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

New stager DB schema for tape



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Miscellaneous

- Mounts can be forgotten by VDQM
  - So there are regularly checked
    - LastVDQMPingTime is used (see Migration/RecallMount)
  - A tapegateway thread is handling this

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it