# Migration plans

## ADC Jamboree

Cédric Serfon

CERN, PH-ADP-CO

December 10th 2012

# Contents

# Introduction of scope

- See Vincent's talk for more info.
- Scope introduced in DQ2 (both Central Catalog and Site Services) :
  - All new datasets/containers produced get a scope corresponding to the project (i.e. first field) of the dataset/container name.
  - All new files get a scope corresponding to the project of the dataset they were first added.
- Campaign to add scope to all existing datasets is being run. Almost all datasets (i.e. more than 99%) have now a valid scope. Remaining datasets are corrected manually.
- Change completely transparent for users/sites...

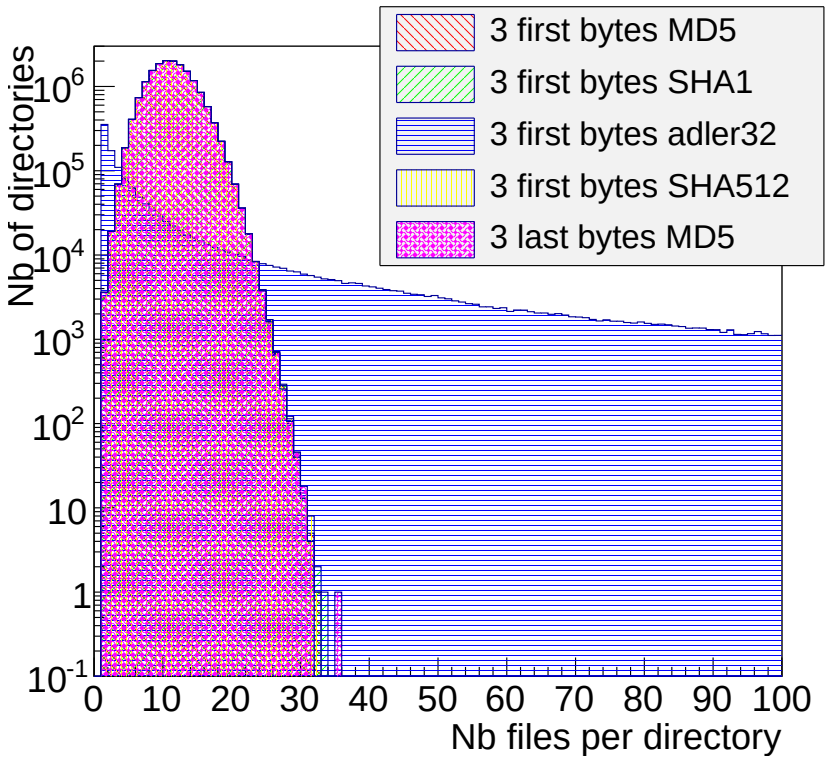
# New naming convention (1)

- Rucio relies on a new convention to have a deterministic path :
  - No need to rely on a external catalog like LFC.
  - Faster lookup.

  $\rightarrow$ Much easier for federation.
- The proposed convention for a file with LFN=scope:filename is the following : basepath/scope/L1/../LN/filename, where L1, ..., LN are hashes from 2 hexadecimal digits (i.e. from 00 to FF).
  - At most $256^N$ directories are created by scope, for N=2 (resp. 3) that allows to store $\sim$2B (resp. $\sim$500B) files (assuming 32k as the maximum number of files per directory).
  - Thanks to the hash, we can have the files evenly balanced between the directories. But we will have an increase of the number of directories.
- The naming convention doesn't rely on the dataset name since a file can belong to different datasets.

# New naming convention (2)

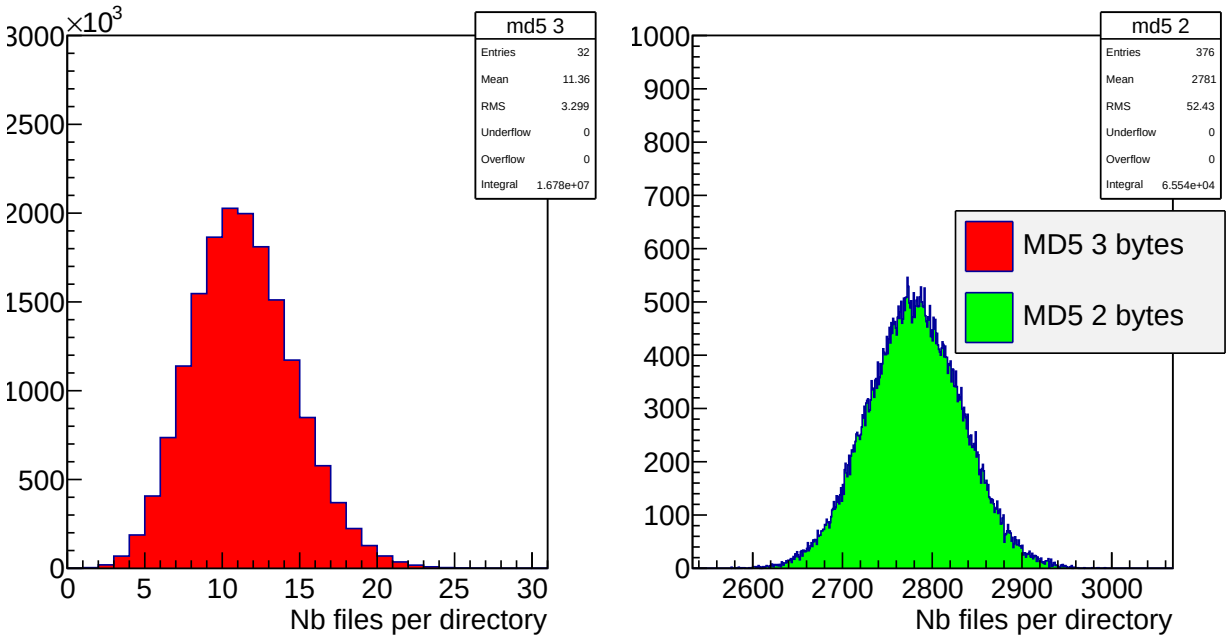
- Different hashes have been tested on a subset (only took one replica for a given file) of the files registered on the central LFC (∼250M). In total 177M files for this analysis.




- Study done assuming one uniq scope, and all files on a single site (worst case scenario).
- Results very similar between all algorithms (except adler32).

# New naming convention (3)

- Comparison of 2 vs 3 levels of hash.




- With the current statistic, the mean number of directories for 2 levels of hash peeks at 2800 (assuming only 1 scope).
- Even with an order of magnitude more files (we want that Rucio scales for at least the next 10 years), we stay under 32k files.

# New naming convention (4)

- Proposal : 2 levels of hash that are the first 2 bytes of md5, e.g. :
  - For file user.jdoe:004406.EXT0._00011.root :
    - scope=user.jdoe
    - md5(user.jdoe:004406.EXT0._00011.root)= 35be9fb53d01500d33011414abccde53
    - PFN : user/jdoe/35/be/004406.EXT0._00011.root
  - For file data11_7TeV:AOD.491965._0042.pool.root.1 :
    - scope=data11_7TeV
    - md5(data11_7TeV:AOD.491965._0042.pool.root.1)= 43635c43b1a59b446bf71272b5c1352c
    - PFN : data11_7TeV/43/63/AOD.491965._0042.pool.root.1
- There is the possibility to customize the deterministic function for each RSE (Rucio Storage Element) in case of technical constraints like e.g. too high number of directories, BUT the simpler, the better.
- Do the Storage providers see any potential issues with this new convention ?

# New naming convention and Site Services

- The DQ2 Site Services are now scope aware.
- There is the possibility to enable the new naming convention on a list of sites.
- This will only apply for newly produced files. The plan is to switch to the new naming convention beginning of next year (Jan./Feb.) on DISK sites.
- For already existing files, we need to migrate the files from the old to the new naming convention. We are talking of $\sim$350M files !

# Migration strategies

- 2 possible scenarii :
  - Physical migration, i.e. Move physically files using the current infrastructure :
    - \+ Infrastructure already there.
    - \+ Already have some experience from MCDISK to DATADISK migration.
    - \- Volume to migrate ∼120 PB is too high and will be difficult to handle in term of Storage capacity...
    - \- and bandwidth.
  - Renaming, i.e. rename files both in LFC and on the Storage :
    - \+ No data movement
    - \- No infrastructure there...
    - \+ but what will be developed for the migration could be used for Rucio to replace Site Services.
- Decided to go to central renaming. No action needed from the sites except to provide some protocol to allow renaming (e.g. WebDAV, xROOTd).

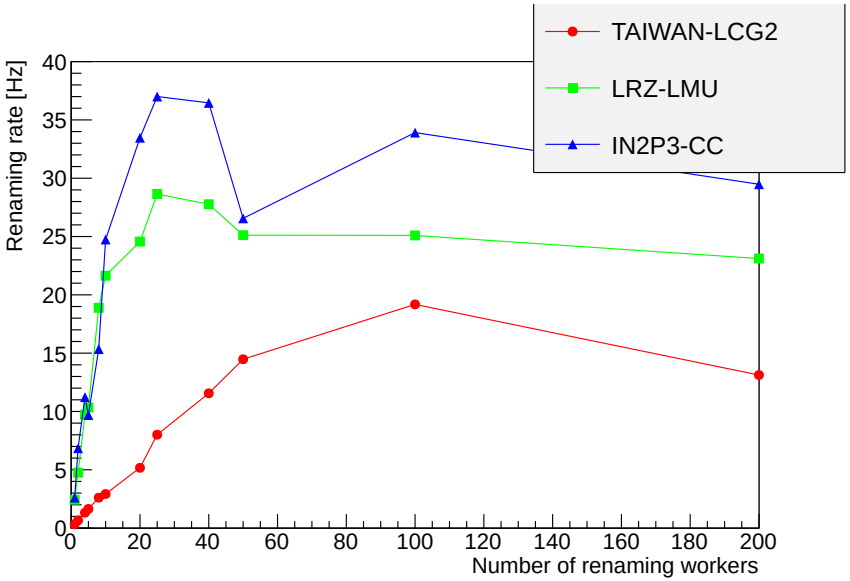
# WebDAV tests and results

- WebDAV is one protocol that can be used to perform remote renaming.
- Supported by dCache, DPM, i.e. 75% of the sites.
- Rucio interface to WebDAV already available.
- Tests on a few sites :

| Site | Storage type | RTT |
|---|---|---|
| TAIWAN-LCG2 | DPM (1.8.2-5) | ∼280 ms |
| LRZ-LMU | dCache (1.9.12-21) | ∼25 ms |
| IN2P3-CC | dCache (1.9.12-16) | ∼3 ms |

# WebDAV tests and results

- Big differences between SE related to RTT (each rename needs 4 interactions : 3 MKCOL, 1 MOVE)




- Encouraging renaming rate : At this speed we can rename most of the sites in a few days.
- For high number of concurrent renaming agents we probably suffer from contention.
- More sites need to be tested, but commissioning takes time and sometimes requires help from the site. Help from squads would be welcome.

# Renaming infrastructure and workflow

- Started to work on a renaming infrastructure based on :
  - gearman `http://gearman.org` for the workflow management
  - supervisor `http://supervisord.org/` for processes management
  - graphite `http://graphite.wikidot.com/` for monitoring
- Workflow :

**for** datasets **in** site **do**
    files = DQ2Lookup(datasets)
    **for** chunk **in** files **do**
        replicasAtSite = LFCLookup(chunk)
        newfiles = AddReplicaNewConventionInLFC(replicasAtSite)
        StorageRename(replicasAtSite, newfiles)
        DeleteOldReplicas(replicasAtSite)
    **end for**
**end for**

# Renaming infrastructure and workflow

- Each task (LFCLookup, AddReplicaNewConventionInLFC, ...) is realized by a set of gearman Workers (independant processes). supervisor controls these different processes.
- Each worker sends information on number of input processed, duration... to graphite.
- Test on a sample of 10759 files in Taiwan :

| Task | Number of workers | Time to process |
|---|---|---|
| DQ2Lookup | 2 | 1.16s |
| LFCLookup | 2 | 22.09s |
| LFCAddReplicas | 4 | 1min31.96s |
| StorageRename | 30 | 25min47.04s |
| LFCCleanup | 4 | 55.36s |

- Larger scale tests (more files/sites) to be conducted.

# Tentative schedule

- January 2013 :
  - Migration to the new naming convention for new produced files on DISK sites.
  - When it is done all files on PRODDISK will follow the new naming convention after a few days $\rightarrow$ Possibility for production jobs to skip the LFC look-up (can reduce significantly the load on the LFC). Requires some changes in the Panda server and in the pilot.
- Before end of April : migration of all old files to the new convention on a few 10s sites to test the renaming infrastructure. Validation of the migration procedure.
- From May to end of the year : bulk migration.

# Conclusions

- DQ2 is ready for switching to the new naming convention.
- The new naming convention will have no impact for users using DQ2 tools.
- The renaming of already existing files will be done centrally in a transparent way.
- This central migration can only be done if the sites provide protocols that allow renaming.

## Open questions

- Renaming of files on TAPE.
- Does the new naming convention fit for all Storage technologies ? Constraints for each Storage summarized in :

`https://docs.google.com/document/d/1zwgPV7s9N6jlXDuQDOrdoGJj2y24EiYDpSBiJb40onI/edit#.`