Update on Storage Elements and File Catalogue consistency

Tier1 Service Coordination Meeting

03.11.2011

Elisa Lanciotti

Overview

- Grid storage elements (SEs) are decoupled from the file catalogue (FC) => inconsistency can arise, mainly of two types:
  - Data in the SEs, but not in the FC ('dark data') -> waste of disk space
  - Data in the FC, but not in the SEs (lost/corrupted files) -> serious operational problems (jobs failing etc..)
- Experiments have developed specific tools to handle these inconsistency. An overview for all experiments reported at T1SCM of 21.04.2011
- ATLAS, CMS, LHCb perform consistency checks SE vs FC based on storage dumps provided by sites. In order to streamline the operations from the site side, there was an initiative to define common formats and procedures to produce storage dumps. Last update given at T1SCM of 01.09.2011
- Today's update:
  - Update on storage dumps formats and procedures
  - Current situation of SE vs LFC consistency for LHCb

Storage dumps format

Some standard formats (like XML syncat for dCache) had already been defined and used for some time, some other have been defined recently (for StoRM and Castor) coordinating with sites and experiments (ATLAS,CMS,LHCb).

- The required information is: Space Token, LFN (or PFN), file size, file creation time, checksum (optional), storage dump creation date
- Sites can choose the more convenient option between text format and XML format. Each of these 2 formats is well documented, with instructions and examples
- The frequency to produce the storage dump (daily/weekly/monthly or on demand) should be agreed between experiment and site

Instructions and examples provided in a twiki:
https://twiki.cern.ch/twiki/bin/view/LCG/ConsistencyChecksSEsDumps#Format_of_SE_dumps

Procedure for different storage systems

DPM: XML format can be obtained via a script that directly queries the DPM DB

dCache: two possible ways

- File system dump: XML format can be obtained with **chimera-dump/pnfs-dump**. Documentation and examples available. Text format can be obtained with chimera-dump with “-a” option
- dCache SRM database dump: Text format dump can be obtained querying the SRM DB (srmspacefile table). Advantage: provides information about space token. A script provided by SARA is deployed at several sites for LHCb.

  Downside: often file system and SRM DB are not in synch! (More on this later).

Castor:

- Text format (Castor nameserver DB dump, plus some parsing). Performed weekly at RAL for LHCb data: ~930K files, takes approx. 6h.
- Same tool can maybe be used at CERN. To be evaluated by Castor/CERN, as the bigger amount of data could cause scalability issues

StoRM:

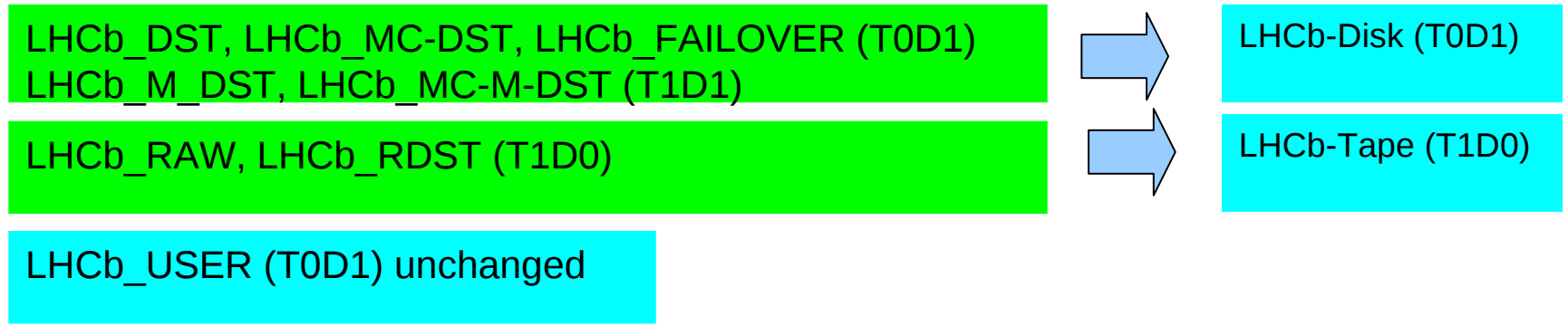
- Storage dumps produced with a GPFS file system dump (text format).
- CNAF (V.Vagnoni) has implemented a script to create a daily dump for LHCb. Can be implemented also for CMS and ATLAS

LHCb use case

For dCache sites the SRM db dump (containing info about space tokens) has been preferred to a file system dump.

In addition to consistency checks, storage dumps are also used for monitoring space usage in 'released' space tokens.

Very useful for LHCb after migration to new space tokens schema (Apr 2011), as the 'old' space tokens have been released (though they still contain data) and their space usage cannot be obtained through SRM.



dCache sites

Some issues with dCache sites:

- Often inconsistency arise in the system: files residing in Pnfs/Chimera file system and content of Postgres dCache SRM DB (srmspacefiles table) do not match
- Both cases observed:
  - files in pnfs/Chimera are not registered in the dCache SRM DB
  - files that have been physically deleted from the file system are still registered in the SRM db

It would be a great help if sites could periodically check the internal consistency of their storage systems. Experiments should only ensure the consistency between storage elements and their file catalogues.

Some questions related to this issue:

What are the consequences if a file is not registered in dCache SRM database? Does it have any consequence on the file management?

LHCb: SE vs LFC consistency

Summary relative to production files on LHCb-Tape and LHCb-Disk space tokens (all files except users').

NRD=size of data at the site's storage which is not registered in LFC

✓ CNAF – StoRM: NRD < 0.5 TB (~0.05% of total data), mostly test files

GRIDKA – dCache: NRD ~ 2-3 TB. Some internal inconsistency detected. To be fixed. Followed up through GGUS ticket 72114

IN2P3 – dCache: some internal inconsistency detected. To be fixed. Followed up through GGUS ticket 75158

PIC – dCache: checks are not performed on a regular basis. A preliminary check in Sept showed no big problem. Regular checks will start soon.

✓ RAL- Castor: NRD < 1TB (~0.1% of total data)

✓ SARA – dCache: NRD < 1TB (~0.1% of total data)

CERN – Castor: no check currently performed

Small discrepancy are not alarming:

- Checks are based on the LHCbDirac service which performs a summary of LFC at directory level and runs every 12h

- Some delay between storage dump creation and the checks

Summary

- Defined common formats and procedure for storage dumps suitable for the three experiments, ATLAS, CMS, LHCb: Space token (if available), LFN (or PFN), size, creation date, checksum (if available). XML or text format.
- System for consistency checks SE vs LFC is now in place also for LHCb
  - CNAF, SARA, GRIDKA, IN2P3, RAL are providing storage dumps.
  - Good consistency at CNAF, RAL, SARA
  - At GRIDKA, IN2P3 some internal inconsistency at the site to be fixed
  - In general, no big amount of 'dark data' found so far.
  - LHCbDirac data management tools for data replication and removal are consistent
- Next steps:
  - Consolidate the experiments' tools and implement also checks in the other direction (LFC vs SE)
  - Define a procedure to remove the non-registered replicas at sites