Storage Elements and File Catalogue consistency

Tier1 Service Coordination Meeting

01.09.2011

Elisa Lanciotti

Overview

- Grid storage elements (SEs) are decoupled from the file catalogue (FC) => inevitably some inconsistency can arise:
  - 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..)
- The first type of inconsistency ('dark data') is handled by the experiments performing consistency checks on the basis of full storage dumps (see summary reported at T1SCM of 21.04.2011)
- In order to streamline the operations from the site side, there was an initiative to define one common format and procedure to produce storage dumps, suitable for the three experiments
  - Status of the activity
- Closer look to the situation of LHCb

Storage dumps format

- Three experiments need storage dumps to perform consistency checks SE vs FC (ATLAS, CMS, LHCb)
- Agreed a common format for dCache, StoRM, Castor for the three experiments (in collaboration with Natalia Ratnikova (CMS) and Cedric Serfon (ATLAS))
- The required information is: Space Token, LFN (or PFN), file size, file creation time, checksum (optional)
- 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 storage dump should be provided on a weekly/monthly basis (or on demand)
- The way it is made available should be decided by the VO:
  - On the site's VO-box if the checks are performed locally at the site
  - Uploaded to a web server, or uploaded to a grid SE if checks are performed remotely

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:
  - 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 or with a query to the SRM DB (an example script has been provided by Onno Zweers (SARA))
- For Castor: No default procedure established yet.
  - First proposal from CERN Castor was to use the stager dumps produced daily. But this is useless for T1D0 service class, as they only contain files currently on the disk cache
  - Second proposal is to use nsls/nsfind to retrieve the full list of files from Castor nameserver. Under implementation at RAL. Downside: will not provide information about service class (space token)
- StoRM
  - text format seems preferable.
  - Storage dumps produced with a GPFS file system dump
  - CNAF (V.Vagnoni) has implemented a script to create a daily dump with the requested format for LHCb. Soon will be implemented for CMS and ATLAS

Contact with sites

Production of storage dumps for the LHCb experiment:

- SARA: has provided them since March 2011, now in the process of updating the format (after the space token configuration change in April 2011)
- CNAF: provided with right format on a daily basis since Aug 30th. Checks are fine.
- CERN: requested last month:
  - First proposal was to use stager dumps: not valid.
  - Second proposal: using nsls/nsfind: under study. Hopefully the same tool developed at RAL can be used.
- RAL: asked through the LHCb contact person at the site, they will be provided soon
- IN2P3, GRIDKA, PIC: asked through GGUS ticket on 31.08. Waiting for feedback.

LHCb: current situation

- Since Apr 2011, at each T1 sites + T0, 3 space tokens:
  - LHCb-Disk and LHCb_USER (T0D1)
  - LHCb-Tape (T1D0)
- Current situation for **LHCb_USER: OK**
  - LFC: space usage got from a LHCbDirac DB updated every 12h on the basis of LFC (a summary at directory level of the LFC)
  - SRM: space usage from lcgutil python API (lcg_util.lcg_stmd(SpaceToken, EndPoint))

site	CERN	CNAF	GRIDKA	IN2P3	PIC	RAL	SARA	LHCb_USER space
SRM (TB)	115.6	29.4	27.3	29.2	20.9	33.7	19.9	
LFC(TB)	99.9	28.3	27.2	29.0	20.1	32.2	20.0	
LFC-SRM	-15.6	-1.1	-0.1	-0.2	-0.8	-1.5	0.1	

Current situation for **LHCb-Disk** can't be stated for dCache sites as some data is still present in the 'old' space token and they are not accessible through lcg_util.lcg_stmd (the space tokens have been released, even if the data is still there and accessible)

LHCb space usage at CNAF

CNAF provides storage dumps daily.

Checks are done centrally with LHCbDirac DataManagement system tools.

Good agreement with LFC! Small discrepancies (O(1TB)) are not a real problem. They can be due to a delay between uploading to the SE and registration to LFC and delay to refresh the information in the LHCbDirac DB containing the LFC summary.

Preliminary results:

- LHCb-Disk: LFC-SRM~1.5TB (~0.4%)
  - DST + M-DST: LFC 176, SE 175
  - MC-DST + MC-M-DST: LFC 206, SE 204
  - FAILOVER: LFC 0.16, SE 0.16
- LHCb-Tape: LFC-SRM~2TB (~0.6%)
  - RAW: LFC 93.8, SE 93.8
  - RDST: LFC 94.5, SE 94.0
  - ARCHIVE: LFC 193.3, SE 191.5

The objective is to have a similar summary for all 6 Tier1 and Tier0

Summary

- Inconsistency between FC and grid SEs are a common problem for almost all the experiments (ATLAS, CMS, LHCb). As a consequence, they periodically need storage dumps to perform consistency checks against the FC-> identifying a common format and procedure to produce the dumps would make the task easier for sites
- Defined a common format suitable for the three experiments: Space token, LFN (or PFN), size, creation data, checksum (if available). Detailed instructions are provided.
- Common procedures have been defined for DPM, dCache, and StoRM. Not yet for Castor.
- Current status for LHCb:
  - Good consistency SE vs LFC for LHCb_USER space at all sites (except CERN)
  - For LHCb-Disk and LHCb-Tape: can't be stated without storage dumps from sites
  - CNAF is already providing storage dumps. SARA has provided them for months, RAL will provide them soon. For PIC, IN2P3, GRIDKA they have just been requested
- Next steps:
  - Wait for a feedback from all Tiers1
  - Consolidate the experiments' tools for processing the storage dumps and doing the checks
- Sites have been very cooperative (many thanks especially to O. Zweers (SARA), V.Vagnoni (CNAF))