EGI-InSPIRE

File Catalogue and Storage Consistency

LHCb Tier1 Jamboree,
Lyon 7-8 March 2011

Elisa Lanciotti

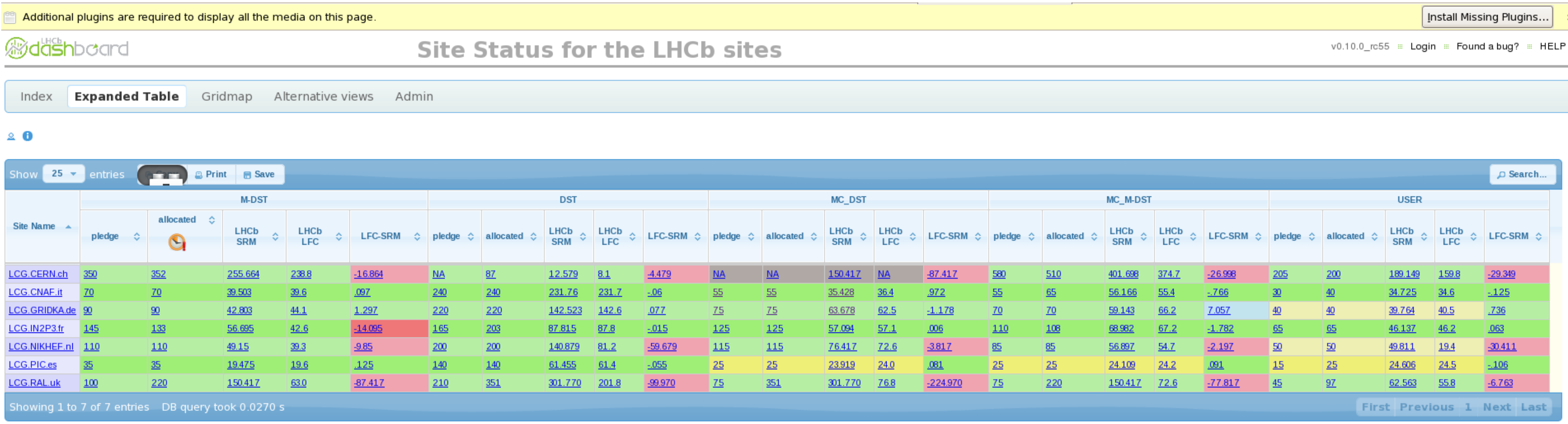
e-infrastructure

Overview

- Problem in space management: the data actually stored in the grid storage elements (SEs) do not always correspond to the data registered in the LFC: this is a considerable waste of disk space
- Overview of the current situation
- New developments in DIRAC DMS to provide tools to investigate these inconsistencies
- Analyzed the 'dark data' of NIKHEF-DST': found some possible reasons of the inconsistencies
- Actions to take:
  - For the past inconsistencies, how to fix them
  - For the future, try to prevent them

Current situation

A unified view of the storage resource usage available at URL http://dashb-lhcb-ssb.cern.ch/dashboard/request.py/siteview?view=spacetokens:



Site Status for the LHCb sites

Site Name	M-DST					DST				
	pledge	allocated	LHCb SRM	LHCb LFC	LFC-SRM	pledge	allocated	LHCb SRM	LHCb LFC	LFC-SRM
LCG.CERN.ch	350	352	255.664	238.8	-16.864	NA	87	12.579	8.1	-4.479
LCG.CNAF.it	70	70	39.503	39.6	.097	240	240	231.76	231.7	-.06
LCG.GRIDKA.de	90	90	42.803	44.1	1.297	220	220	142.523	142.6	.077
LCG.IN2P3.fr	145	133	56.695	42.6	-14.095	165	203	87.815	87.8	-.015
LCG.NIKHEF.nl	110	110	49.15	39.3	-9.85	200	200	140.879	81.2	-59.679
LCG.PIC.es	35	35	19.475	19.6	.125	140	140	61.455	61.4	-.055
LCG.RAL.uk	100	220	150.417	63.0	-87.417	210	351	301.770	201.8	-99.970

Site Name	MC_DST					MC_M-DST				
	pledge	allocated	LHCb SRM	LHCb LFC	LFC-SRM	pledge	allocated	LHCb SRM	LHCb LFC	LFC-SRM
LCG.CERN.ch	NA	NA	150.417	NA	-87.417	580	510	401.698	374.7	-26.998
LCG.CNAF.it	55	55	35.428	36.4	.972	55	65	56.166	55.4	-.766
LCG.GRIDKA.de	75	75	63.678	62.5	-1.178	70	70	59.143	66.2	7.057
LCG.IN2P3.fr	125	125	57.094	57.1	.006	110	108	68.982	67.2	-1.782
LCG.NIKHEF.nl	115	115	76.417	72.6	-3.817	85	85	56.897	54.7	-2.197
LCG.PIC.es	25	25	23.919	24.0	.081	25	25	24.109	24.2	.091
LCG.RAL.uk	75	351	301.770	76.8	-224.970	75	220	150.417	72.6	-77.817

Site Name	USER				
	pledge	allocated	LHCb SRM	LHCb LFC	LFC-SRM
LCG.CERN.ch	205	200	189.149	159.8	-29.349
LCG.CNAF.it	30	40	34.725	34.6	-.125
LCG.GRIDKA.de	40	40	39.764	40.5	.736
LCG.IN2P3.fr	65	65	46.137	46.2	.063
LCG.NIKHEF.nl	50	50	49.811	19.4	-30.411
LCG.PIC.es	15	25	24.606	24.5	-.106
LCG.RAL.uk	45	97	62.563	55.8	-6.763

Showing 1 to 7 of 7 entries DB query took 0.0270 s

Sources of information:

- Information about space usage of the **LFC** got through DIRAC tools (StorageUsage service)
- Information about total allocated space and used space from the **SRM** via **lcg_util API**
- The script feeding the Dashboard page will be soon migrated to a new version (some corrections will be applied, see RAL)

Considerations on the current situation

- Big discrepancies (>10TB) observed only for some space tokens and sites
- In case of inconsistency, it's always more data in SE than in LFC (good!). The opposite case has been observed in extremely rare cases
- Small discrepancies displayed in the table are not alarming: delays in the registration may happen, and the information from the LFC has a 12h latency (polling time of DIRAC agent which feeds the backend DB).

Summary by space token

ST	Not in LFC	Total pledge	percentage
USER	60TB	450TB	13%
M-DST	44TB	890TB	5%
DST	64TB	1175TB	4%
MC-M-DST	29TB	1000TB	3%
MC-DST	4TB	470TB	<1%

Globally, the inconsistency is not dramatic (except for USER), but for some particular sites it is: i.e. NIKHEF-DST 60/200 = 30% of pledge space used by 'invisible' data. This affects the efficiency of the site, since jobs go where data are

This is a considerable source of inefficiency for the site, and a waste of resources for the experiment.

Cleaning campaigns to remove old data cannot recovery it, as this data is not in the LFC

➡ New development in DIRAC Data Management System to provide tools to address this problem

DIRAC DMS tools to identify the inconsistency

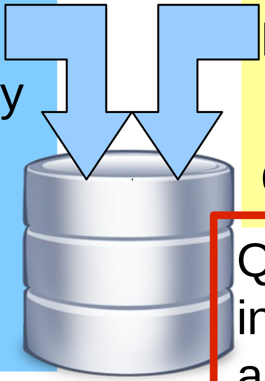
StorageUsage

an agent browses the content of the LFC every 12h

A summary by directory is stored in a DB

DIRAC tools allow to query it

Already in production.



New development: SE Usage Agent

Reads the content of the SE

Check whether the directory is in the LFC

If yes, stores it in a table for replicas, if not in a table for 'dark data'

Path, size, files, and insertion time are stored.

Development phase done, to be put in certification

Querying the DB we know which data are inconsistent, where and when the inconsistency appeared

Problem: how to feed the agent. Not possible to get the information interrogating the SRM (i.e. gfal API): Too high load on SRM interface

Weak point of this system: NEED to ask sites to periodically provide a dump of the content of the space tokens (file path, size, time stamp). Weekly frequency is fine. Can sites provide this information on a regular basis?

How to do consistency checks

Objective: answer to the question: what is this data? What to do with it?

First analysis on the data provided by SARA: NIKHEF-DST space (60 TB of dark data)

Procedure followed:

1-The agent processes the input files provided by the site and stores in the DB the problematic directories

2-Query to the DB: list of inconsistent directories → sample to be analyses on the basis of the path, creation time and other meta-data (retrievable from the Bookkeeping) to understand the origin of the inconsistency and classify the data into categories. Mainly two:

Case 1- Data that were copied correctly to the SE but failed to be registered in the LFC

Case 2- Data that were copied and registered correctly. Later, an attempt to remove them partially failed: they were removed from the LFC but not from the SE

(and many sub cases).

For every category, define an **action** (remove from SE or register to LFC) to fix the inconsistency

Determine if the problem is still happening, or if the inconsistency we observe now only happened in the past

First analysis on NIKHEF-DST

Case 1: failure at replication time when registering the file to the LFC

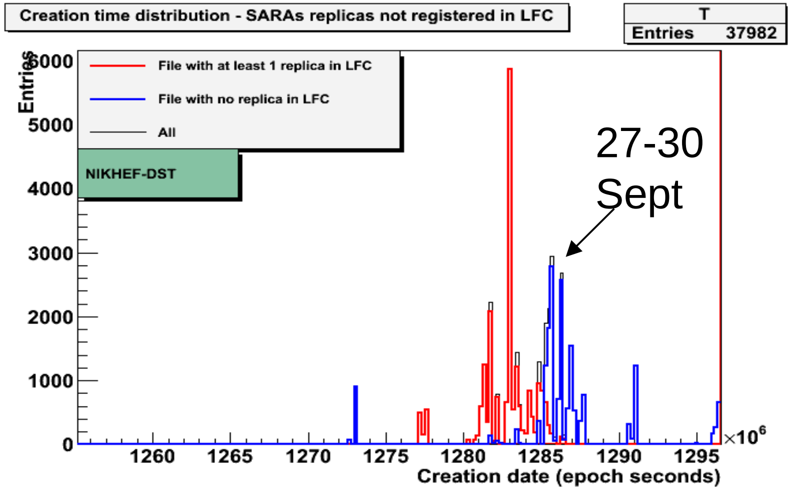
-> the files creation data distribution should show some structure correlated with some outage of the SRM or of the LFC

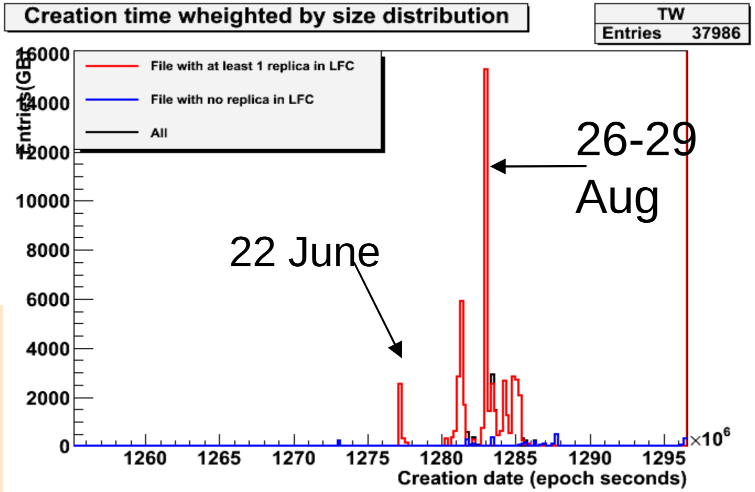
The biggest peak -> time bin (27-29 Aug 2010): ELOG: SARA SRM down.

Other peak 22 June: ELOG and GGUS ticket: SARA SRM problems

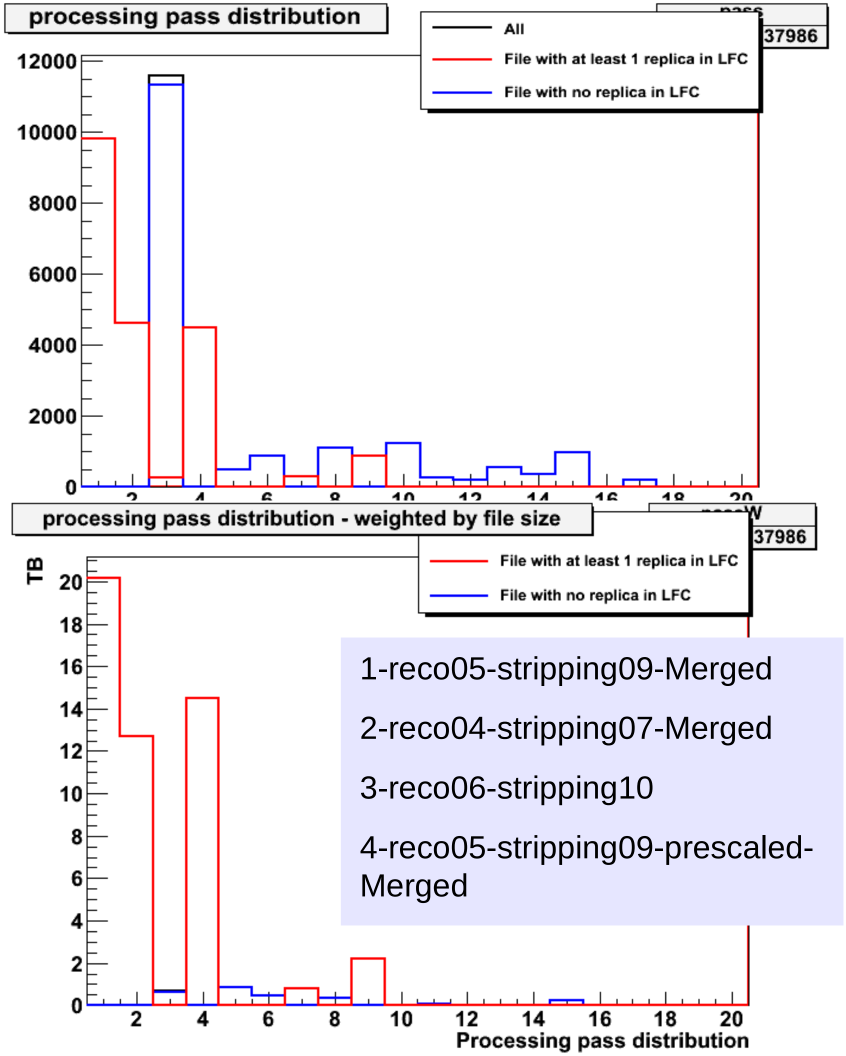
27-30 Sept: ELOG about failed replication of BHADRON.DST, CHARM.DST because of file type missing in the Bookkeeping.

Either a pure SRM problem or a combination of SRM problem and DIRAC client bug to handle the error in case of SRM instability at the origin of the inconsistency?





Analysis of NIKHEF-DST (II)



1-reco05-stripping09-Merged

2-reco04-stripping07-Merged

3-reco06-stripping10

4-reco05-stripping09-prescaled-Merged

Case 2: failure at removal time

Since cleaning is done per processing pass, some correlation should be observed with the processing pass of dark data

Data are concentrate mainly in 3-4 processing passes:

-reco4-strip7 should have been removed from SARA (ELOG Sept 27th), but some of them are still on the SE (though the replicas are NOT in the LFC): to be removed

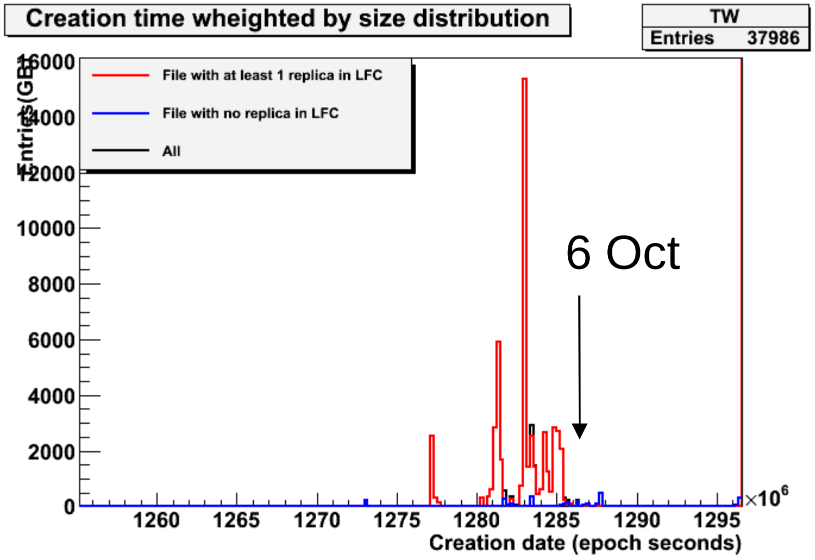
-reco06-strip10: to be removed

-reco05-strip09-Merged and reco05-strip09-prescaled-Merged : have these replicas been removed from SARA? No mention found in ELOG

Note: non-merged data of old processing have been removed from all SEs and from LFC

Merged files replicas in some case have been removed only from some sites (see in the plot that all *-Merged data have other replicas that are correctly registered)

Is the problem still happening?



Difficult to say on the basis on the information available.

If the new DIRAC agent would be running on a regular basis, we would be able to spot new inconsistency when they arise.

On the basis of the information available now, we can only say that at replication time no error has occurred since Oct 6th 2010

About errors at removal time: we should find whether and when reco05-strip09-Merged and reco05-strip09-prescaled-Merged have been removed from SARA

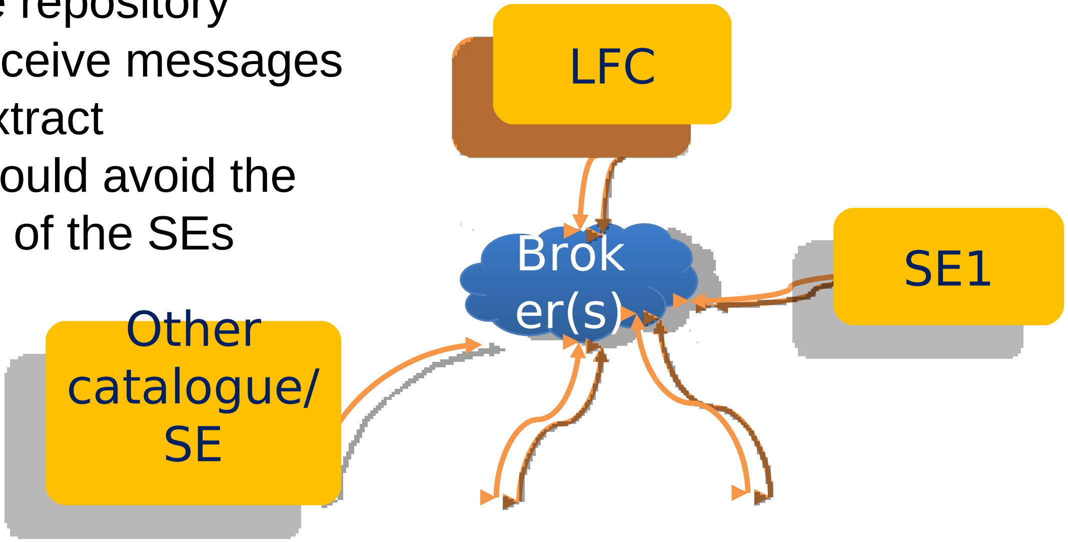
If we could find a recent case of inconsistency, it would be easier to find the cause, and fix it.

New developments in the middleware

- The same problem affects all experiments: keeping catalogues and SEs in synchronization is a general issue
- New developments in the middleware: SynCAT: a messaging system to make various catalogues and SEs talk to each and keep them synchronized
- Various SE/Catalogues can subscribe to send/receive messages

Possibility to implement a message repository which subscribes to the broker to receive messages and provide it with an interface to extract notifications published by SEs -> would avoid the need to ask sites to provide a dump of the SEs content

Very interesting perspective but no time line defined yet → still necessary to address this problem from the experiment side



Developed by F.Furano (IT/GT group)

Summary

- Data not registered in the LFC cause a considerable waste of disk space. And the cleaning campaigns to remove old data cannot recovery it
- Objective is to reduce the inconsistency to a reasonable level, even if it will never be zero. Small discrepancies are not alarming (provided there are more data in the SE than in the file catalogue, and not viceversa).

Origin of the inconsistency: Several possible reasons, not easy to find the exact reason for every file/directory (too many data, and log files of transfers not available..) but on the basis of what observed for NIKHEF-DST most inconsistency originated when trying to remove/replicate data when the SRM was unstable

For most of the data, a reasonable solution is to remove the data from storage

New tools developed in DIRAC-DMS to address this problem: For the time being, it can only work with the collaboration of sites who should provide a list of files per space token. In the future, new solutions from the middleware will help

Next to do: study other cases, especially urgent USER space (often close to get full in many sites)