LHCb at IN2P3

- Occasional issues
- Recurrent issues
- Critical issues:
  - dCache Transfer
  - Short Pilots
  - Corrupted Files

OCCASIONAL ISSUES

Internal:

- files not available: due to a tape reboot issue (2011-11)
- job backlog (pool to pool replication): added last week two new LHCb staging pools in our dCache instance. (2011-11)
- Replication to IN2P3 LFC not working: upgrade to Oracle 11g (2012-02) + patch (16.03.2012 et 20.04.2012)
- lcg_utils cant contact IN2P3 SRM endpoint: Problem with the the authorization module in dCache (gPlazma). restarted it and SRM server. (2012-01)
- SRM Authentication failed Problem fixed (2012-02)
- Ghost pilots: misconfig on cccreamceli06, some missing users in /etc/shadow (2012-02)
- Pilots aborted: due to cccreamceli05 overload, causing input sandbox scp failure. (2012-05)

External:

- SAM tests on VO sw dir timing out: 'ls -R' filling the cache with unnecessary information (2012-01)
- Low level of running jobs at IN2P3-CC(-T2): pinning was 24 hours and not 6. (2012-03)
- FTS transfers from CERN to IN2P3: problem between RENATER GEANT and CERN. (2012-02)

RECURRENT ISSUES

SHARED SW AREA

- CVMFS repositories were not properly mounted on those machines (2011-11)

- on some hosts related to cache quotas, apparently caused by a bug (2011-11)

- SharedArea (CVMFS) problem at IN2P3-CC(-T2): out of production to repair de CVMFS service. (2012-03)

LHCb JOBS ISSUES

- Stalled jobs due to memory consumption. As agreed with LHCb, we still supply 5GB by job on the verylong queue (2012-02) jobs consume 1GB (dirac) + 4GB while VMEM limit at IN2P3 is for the process group was 4GB (2012-03)

- Pilot jobs of LHCb can be submitted on queue long. (previously only verylong queue) (2012-04)

- Failed Pilots failed exceeded either CPU/memory limit queue long (4G vsize limit) (2012-05)

- Pilots aborted: limit MaxNumberOfJobs = 1500 pear each CE (3000 jobs running+waiting per user) (2012-04)

- reco13 72h reconstruction/reprocessing jobs, set up its queues 'long' (48h) and 'verylong' (72h) + right access to the users lhcb049 et lhcb099 (2012-04).

CRITICAL ISSUES

SHORT PILOTS (2012-03-13)

short pilots (80% jobs consume less than 400s CPU) -> scheduling pass can take too much (+2900s!)

Solution:

- Andrei's pilots for the dirac certification

- reduction objective: from 3200 slots to 1600.

- matching issues (after 5 tries it ends: delay time patch for the pilot) + timeleft not supported

GFTP (23.03.2012)

- very high activity of LHCb export transfers (*after power outage!*) e.g. about 2100 GFTP connections in queue on the 7 export pools. GTFP doors overloaded and other VOs will be impacted.

- local copy on WN improves job efficiency (failure rates and CPU/wall clock time), protocol used is not dcap (dccp) but gftp via lcg-cp (that adds an additional step: from import to export buffer)

Solution:

1. decreased the number of jobs (1000 instead of 1200) and merging jobs (70 from 100)

2. maybe unfortunate time (power outage) when many jobs started at the same time and therefore were overloading the storage.

3. Strange: there were 2100 connections from max 1200 jobs, maybe some connections not properly closed?

CORRUPTED FILES (2012-03-17)

- GGUS reopend 2012-05-16 corrupted files (for the previous one the solution: remove the files)

- PROBLEM: out of 1153050 lhcb files (from DCache 05/15/2012 02:44 PM) there are

lhcb files with lcg/lfc checksum mismatch: 5853 (0.5%)

adler32: 678, 0.058% (prod: 204, 0.017%)

md5: 4405 (prod: 4405)

hex: 518 (prod: 142)

short: 252 (prod: 121)

. effectively corrupted (only 7 are not in LFC)
. same filesize both on LCG and LFC
. 144 belong to 2012 (141 to March 2012) and 531 to 2009
. those of March 2012 distributed over 26 different pools

Note that:

"prod" means files marked neither as "test" nor as user" ;

"adler" means that both the lfc and the lcg checksums are in adler32 format (e.g. lcg:ec3093e3 lfc:13cf6c1d)

"md5" means the lfc checksum has an md5 format (e.g. lcg:f1f3f4ab lfc:81cd33b07d79aa3ca935df576c0a761b)

"hex" means that the lfc checksum starts with an 'x' (e.g. lcg:06a82da5 lfc:x6a82da5)

"short" means that the lfc checksum is 7 digits because the lcg checksum is truncated of the leading '0' (e.g. lcg:0003816d lfc:003816d)

- Investigations:

1. IN2P3 task force + Philippe to get more details from LHCb point of view

1.1 LHCb no checksum verification neither for the job's output transfer, nor for the FTS transfer (verified on FTS)

1.2 LHCb use lcg-cp that by default doesn't calculate checksums (instead of lcg-cr to make sure that corrupted data could not be stored).

1.3 Dirac has the Adler function (DIRAC/Core/Utilities/Adler.py), instead of a checksum comparison, DIRAC/Resources/Storage/SRM2Storage.py compares only file size.

2. check if other experiments calculate checksum (Atlas does it, CMS uses Phedex'TDBM)

3. check interventions at IN2P3 on GOCDB and Elog (in particular see March 2012)

2012-03-23: power outages

2012-03-05 and 2012-03-15: 2 quick interventions (10min) on GE problems (lhcb jobs id 99xxxx failed).

(other interventions didn't impact LHCb or were harmless on this issue)

- Conclusions:

No date correlation between in2P3 downtimes and file corruption

No pool specific

No WN specific

No silent corruption (when you copy file locally you get same checksum)

- Todo:

FTS stats (e.g. rate of FTS transfer failures due to checksum inconsistency)

Network team feedback

Late file stats for inconsistency check