WLCG Service Report

Harry.Renshall@cern.ch

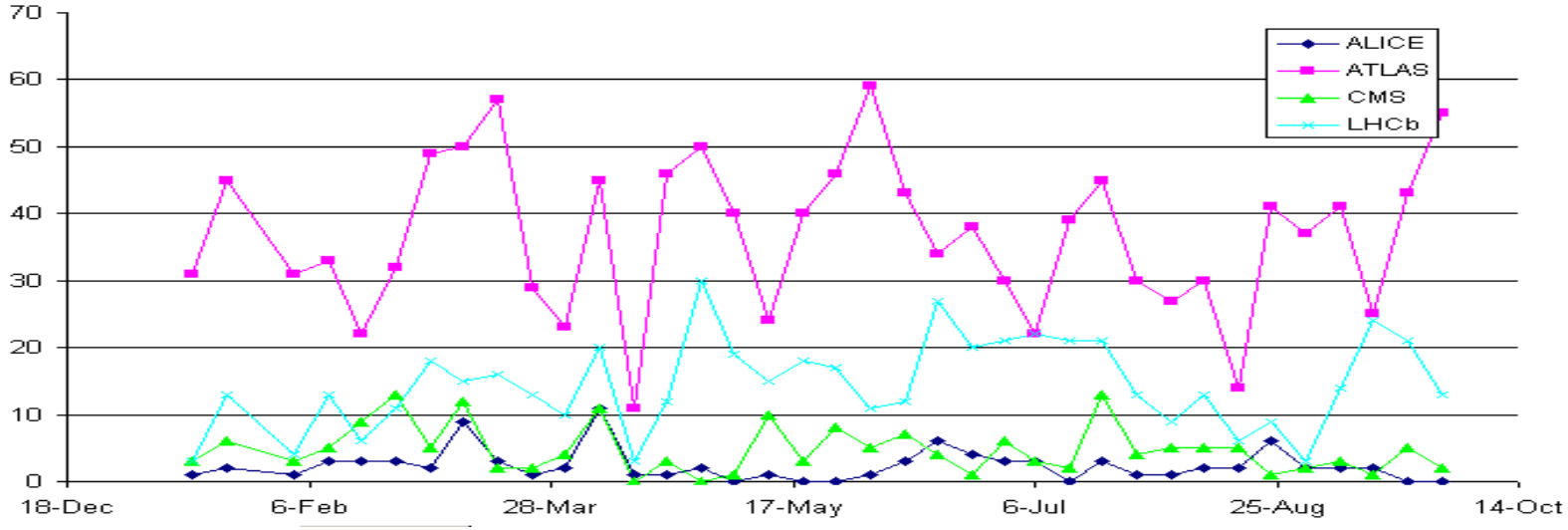
~~~

WLCG Management Board, 29th September 2009
~~~

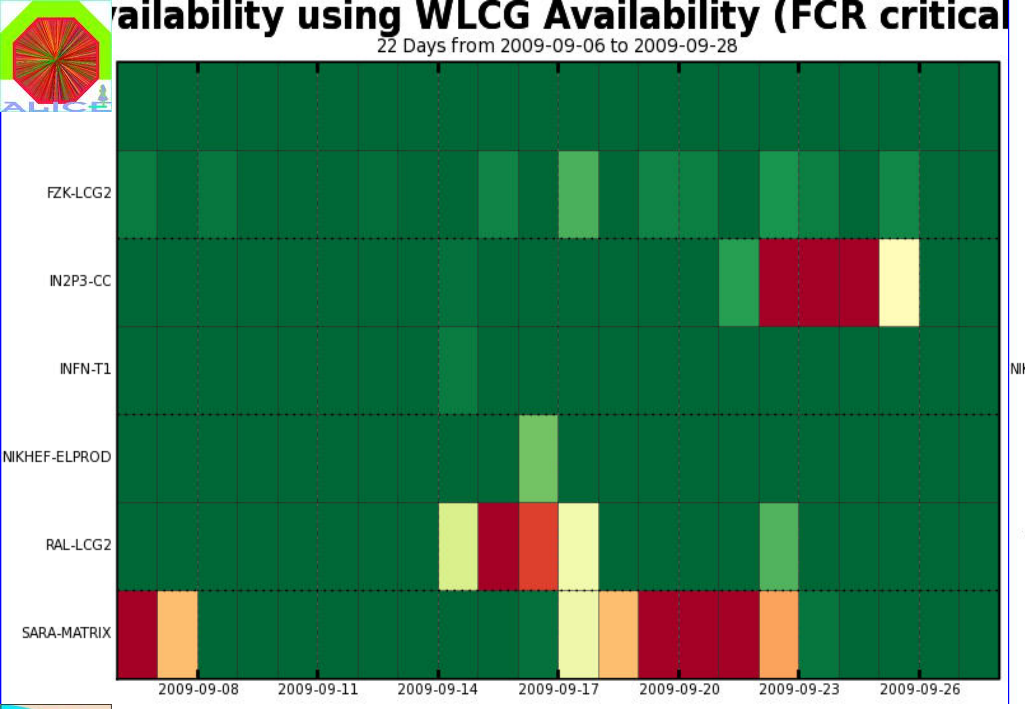
Introduction

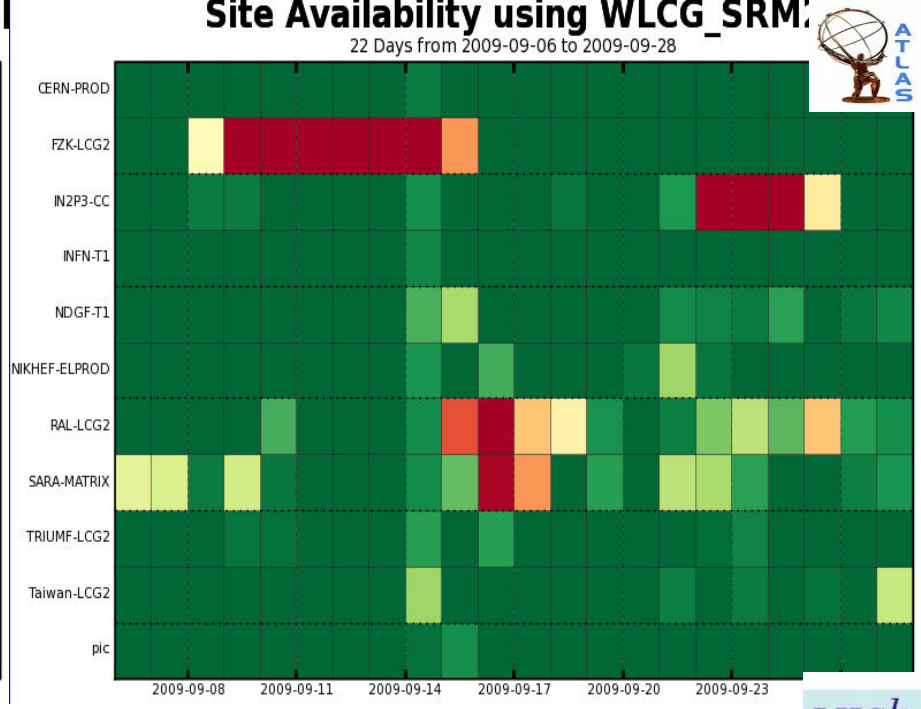
- Covers the three weeks 6th to 27th September
- Mixture of problems – mostly database related
- One test alarm ticket: from CERN to BNL-T1.
  - Acknowledged by BNL but did not automatically open an OSG ticket. New global tests this week.
- Incidents leading to service incident reports
  - Two separate LHCb CERN CASTOR blockages 7 and 8 September.
  - Access to ATLAS DB at FZK blocked on 7 September with too many open sessions. Degraded from 8 to 16th.
  - RAL Disk to Disk transfers failing from 15-17 Sep during a planned upgrade to the CASTOR Name Server.
  - ATLAS Replication Tier0->Tier1 down on 21 Sep.

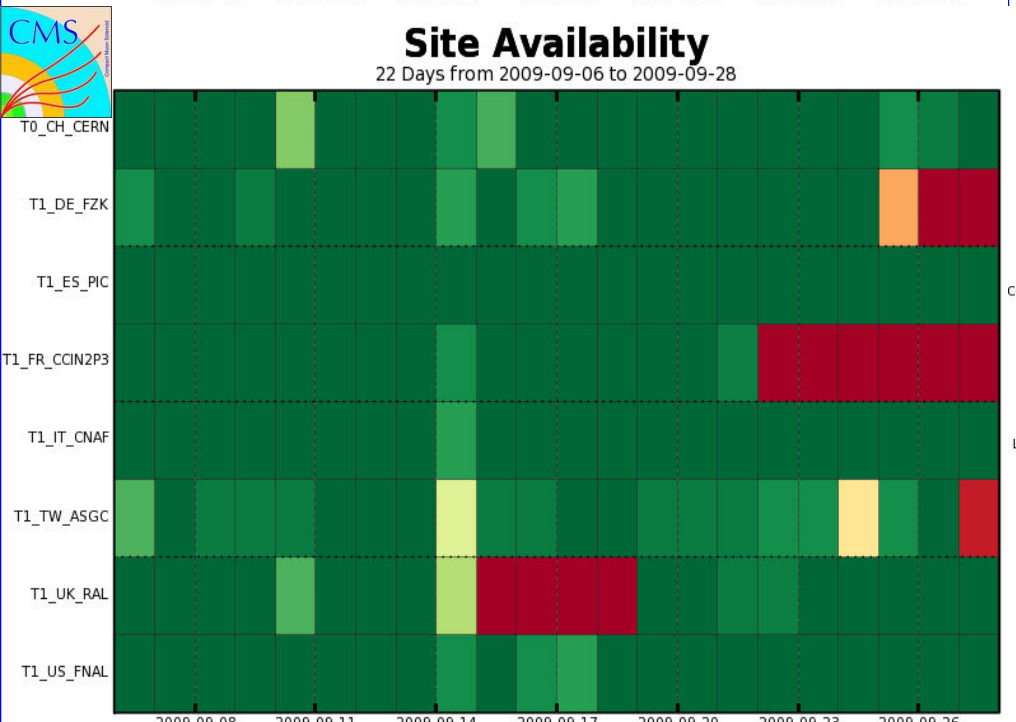
GGUS summary (3 weeks)

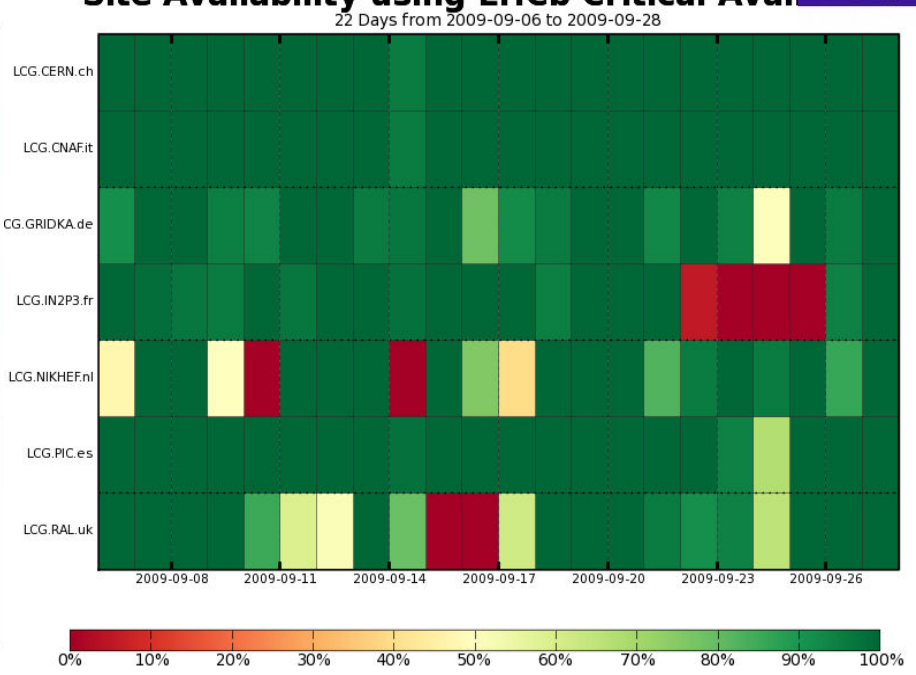
VO	User	Team	Alarm	Total
ALICE	2	0	0	2
ATLAS	41	81	1	123
CMS	8	0	0	8
LHCb	2	56	0	58
Totals	52	137	1	191




vailability using WLCG Availability (FCR critical
22 Days from 2009-09-06 to 2009-09-28
FZK-LCG2
IN2P3-CC
INFN-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
2009-09-08
2009-09-11
2009-09-14
2009-09-17
2009-09-20
2009-09-23
2009-09-26


Site Availability using WLCG_SRM
22 Days from 2009-09-06 to 2009-09-28
CERN-PROD
FZK-LCG2
IN2P3-CC
INFN-T1
NDGF-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
TRIUMF-LCG2
Taiwan-LCG2
pic
2009-09-08
2009-09-11
2009-09-14
2009-09-17
2009-09-20
2009-09-23


Site Availability
22 Days from 2009-09-06 to 2009-09-28
T0_CH_CERN
T1_DE_FZK
T1_ES_PIC
T1_FR_CCIN2P3
T1_IT_CNAF
T1_TW_ASGC
T1_UK_RAL
T1_US_FNAL
2009-09-08
2009-09-11
2009-09-14
2009-09-17
2009-09-20
2009-09-23
2009-09-26


Site Availability using LHCb Critical Avai
22 Days from 2009-09-06 to 2009-09-28
LCG.CERN.ch
LCG.CNAF.it
CG.GRIDKA.de
LCG.IN2P3.fr
LCG.NIKHEF.nl
LCG.PIC.es
LCG.RAL.uk
2009-09-08
2009-09-11
2009-09-14
2009-09-17
2009-09-20
2009-09-23
2009-09-26
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

Experiment Availabilities Reports

FZK ATLAS 3D DB degradation from 7-16 Sep – see SIR

RAL nameserver upgrade 15-17 Sep lead to disk-to-disk copy failures 15-17 Sep for all VOs. See SIR.

NIKHEF had scheduled network maintenance on morning of 22nd. Three dcache-SRM crashes in one week – debugging with developers. SAN problem on 21st which cut off their bdii.

IN2P3 scheduled Electrical Power upgrades from 22-24 September extended to at risk for 2 days (followed now by dcache migration to chimera file system 28-30 Sep.)

CERN CASTORLHCb outages

7 September CASTORLHCB stager was blocked from 05.00 for 3 hours when a tablespace could not be extended. This should be automatic but in fact an associated control file file-system was filled by Oracle dumps put there by a known bug in the clusterware. The lemon metric to warn of the full file-system was unfortunately not correctly configured. This has now been fixed and DES group is checking Oracle patch availability to prevent further core files.

8 September one of the nodes of the RAC cluster of LHCb became blocked (login could not complete) with console logs showing errors on a disk partition. Clusterware should have ejected the node but did not the suspicion being that it was partially responding. DB monitoring did detect the problem and the node was rebooted.

ATLAS 3D DB RAC at GridKa degradation Sep 7-16, 2009.

Due to too many open sessions on the first of two ATLAS RAC nodes we rebooted it on Sep 7. After rebooting the DB was not properly open (was to be seen only in alertlog). We tried to restart the instance several times, but not successfully. From Sep 8, we stopped the 1-st instance and the DB worked on only the 2-nd node. Next day we opened an ORACLE Service Request.

Streams replication of the ATLAS conditions to FZK was hence down from 7 to 8 September.

On Sep 16 the 2-nd node rebooted for unknown reasons (maybe network). Surprisingly, afterwards the DB started properly on both nodes and everything seems fine now. Therefore, on Sep 17 we closed the SR.

From 8 to 16 Sep the service was at risk from a second failure and possible overloading (activity was in fact low during this period).

Follow up - Not clear, but we suspect this problem to be an unexpected consequence of the Oracle July CPU patches that we applied on Aug 11. By a reboot of the nodes (one after the other) directly after patching maybe we would have noticed this issue already on Aug 11. We are discussing to re-open the SR to find the real cause (during the SR, the Oracle support was not very useful).

RAL Disk to Disk transfers failing 15-17 Sep

Disk to Disk (D2D) transfers started failing during a planned upgrade to the Name Server and were down for all VOs during 44 hours.

After applying a scheduled upgrade of the Castor NS from version 2.1.7-27 to version 2.1.8, during testing it was found that Disk to Disk (D2D) copies started failing across all instances. Any possible link with the NS upgrade was ruled out. Investigations carried out with the assistance of CASTOR developers at CERN revealed that there was an LSF job scheduler problem resulting in D2D transfer jobs failing. After LSF was restarted, both on the central servers and on all disk servers, the D2D transfer problem disappeared.

Why this problem started affecting services after the NS upgrade is so far unknown as it has not been possible to reproduce. The D2D transfer problem had also been detected on the certification instance immediately prior to this upgrade during NS testing. Since D2D transfers are unrelated to the NS, it was decided to proceed with the upgrade. We now believe that the problem was caused by a wrong procedure for stopping castor services and LSF prior to the NS upgrade, and bringing them up afterwards.

Follow-up: Certification instance must always be fully functional, and no future upgrades should proceed if anything is broken. A clearly defined startup and shutdown sequence for the LSF scheduler (both central LSF master and on daemons on all disk servers) and other CASTOR services needs to be written and tested, and should be used during future upgrades.

ATLAS Replication Tier0->Tier1 down Sep 21st

- Capture aborted after ORA-01280: Fatal LogMiner Error. Failure occurred at 8:05 and was not noticed by shifter until 18:00 due to lack of email notification from monitoring. Lack of email notification from the monitoring was caused by the overload of sendmail, which is sending a number of emails at 8:00. Because of the overload message from monitoring hasn't been sent. In the mean time monitoring web page was not checked by shifter. Replication to Tier1 sites was completely reestablished around 18:00.
- ATLAS dbas noticed that the capture process was down but sent a private email to a person on holidays. Mail should have been sent to the physics database support email list.

Follow up:

- Shifter now obliged to check monitoring web page. Monitoring is being reviewed to make sure that email notifications will be sent each time an abort occurs. For each message that monitoring was not able to send in 3 tries, a thread will be created and it will try to send the message until success.
- We are in contact with Oracle support in order to identify which is the cause for the LogMiner error.

Miscellaneous Reports (1/3)

- CREAM-CE's at CERN failing for ALICE job submission from 7 to 14 September with 'blparser service is not alive'. Needed detailed investigation/understanding.
- ASGC had problems migrating CMS data to tape over last week – tracked down to badly configured cartridges in the CMS pool. Planning to go from 6 to 24 tape drives mid-October so tape performance will be limited till then.
- LHCb Dirac has now been certified for SL5. CERN LHCb SRM upgraded to 2.8 to properly support xroot TURL's.
- Many WLCG staff at EGEE'09 in Barcelona last week giving presentations and demonstrations (at the WLCG booth) and attending EGI/HEP SSC meetings.
- RAL had problems with their Maui/Torque batch scheduler since converting to SL5. Suspected due to running 32-bit server version but cleared when server and worker node clients were brought up to the same software level.

Miscellaneous Reports (2/3)

- CMS plan to hold what they call the October Exercise involving all physics groups where they pretend to act as if they are trying to push out the first physics papers in a hurry after they have data. It starts on October 5th and lasts for 2 weeks with intensive grid job submission via their CRAB (CMS Remote Analysis Builder) servers.

- For this they needed to increase their number of Grid pool accounts at CERN (cmsxxx) from the current 200 to 999 and this has been done (thanks to AIS and FIO). These accounts appear on lxplus, lxbatch and cms VO-boxes but are not in AFS. They request in fact to go to 2000 for a long term solution.

- To appear in ldap they also have to have a NICE account hence a mail account. An individual CCID can only have 1023 mail accounts so a second service provider ‘CMS GRID-USER2’ has been setup in CRA by AIS group.

Miscellaneous Reports (3/3)

- ATLAS will perform a throughput test of data distribution from Tier-0 to their Tier-1 for 5 days from 5 October and using version 2.2 of the FTS software.

- ATLAS user analysis test for 21-23 October (or following weekend if not ready in time) to get many user analysis jobs running over the world-wide resources. Users are meant to run their normal analysis jobs during this test. This is a follow-on test from STEP-09 and the last before data taking.

- Plan is for users to run analysis jobs on a set of large AOD datasets for the first two days with the third day for copying output to Tier 3s or local disks.

- In preparation for this test ATLAS are distributing 5 large containers of AOD (each of 100M events = 10 TB) at two to each Tier 2 cloud then expert users will run over 3 containers so as to exercise at least two Tier 2 clouds with their analysis (these details subject to change).

Summary

- Hard to understand Oracle bugs/features/side-effects at CERN-LHCb, FZK-ATLAS and CERN-ATLAS.
- CASTOR/LSF startup/shutdown sequence issues at RAL – correct procedures should be tested and documented.
- Extended electrical power downtime at IN2P3 followed by Chimera migration – experiments did not want to restart then stop production.
- Each SIR can still teach us something new – often how independent faults conspire together. Please keep them coming.