WLCG Service Report

Harry.Renshall@cern.ch

~ ~ ~

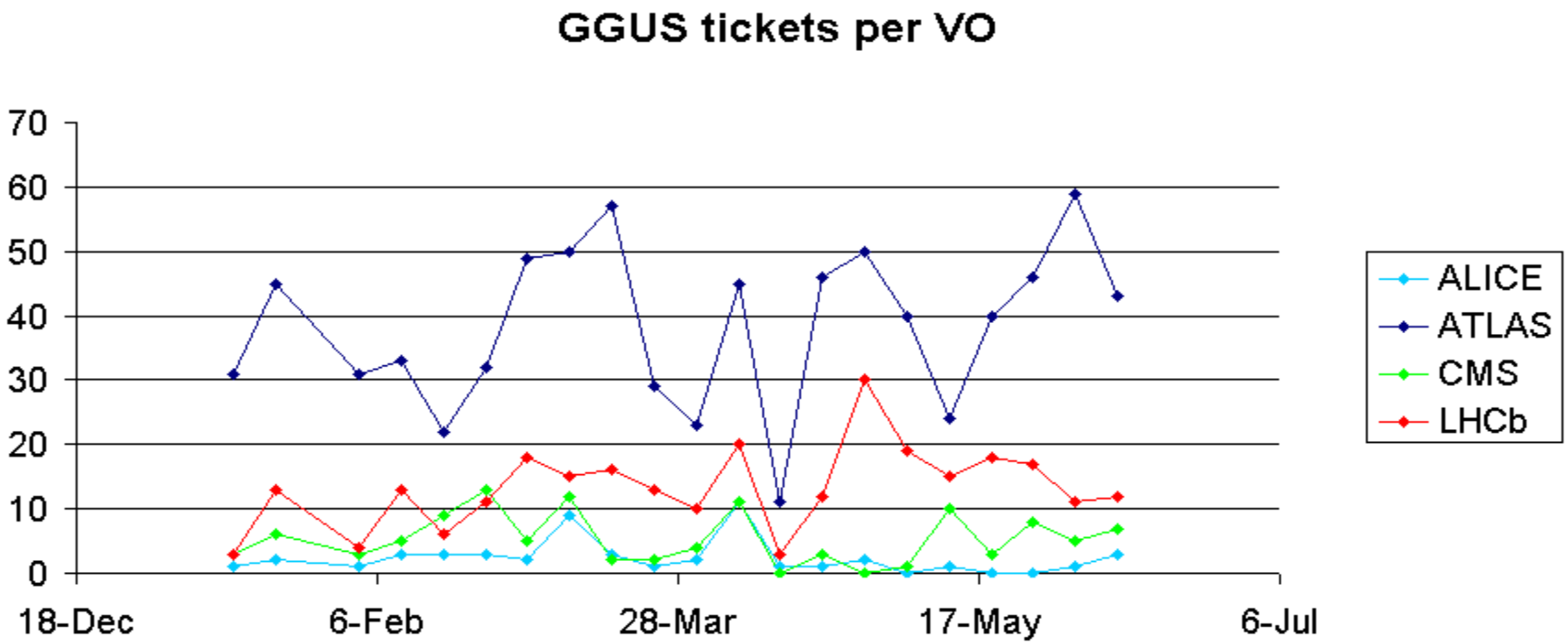
WLCG Management Board, 9th Jun 2009

Introduction

- This report covers the service for the two weeks period 24 May to 6 June so includes the first week of Step'09
- GGUS ticket rate normal and no alarm tickets.
- Three Central Service incidents:
  - A bug in the upgrade script from CASTOR 2.1.7 to CASTOR 2.1.8, carried out for LHCb on 27 May, and coupled with an earlier redefinition of LHCb pool attributes resulted in the loss of some 7500 files (of reference MonteCarlo data).
  - Port failure on 1 June in router connecting RAC5 to GPN cut off production databases for several hours.
  - Following the 27 May LHCb incident a further 6500 files were lost on 5 June during a manual operation to try and enable migrations for tape0disk1 files in order for them to become tape1disk1 and which exposed a pre-existing CASTOR bug.

GGUS Summaries – 2 weeks

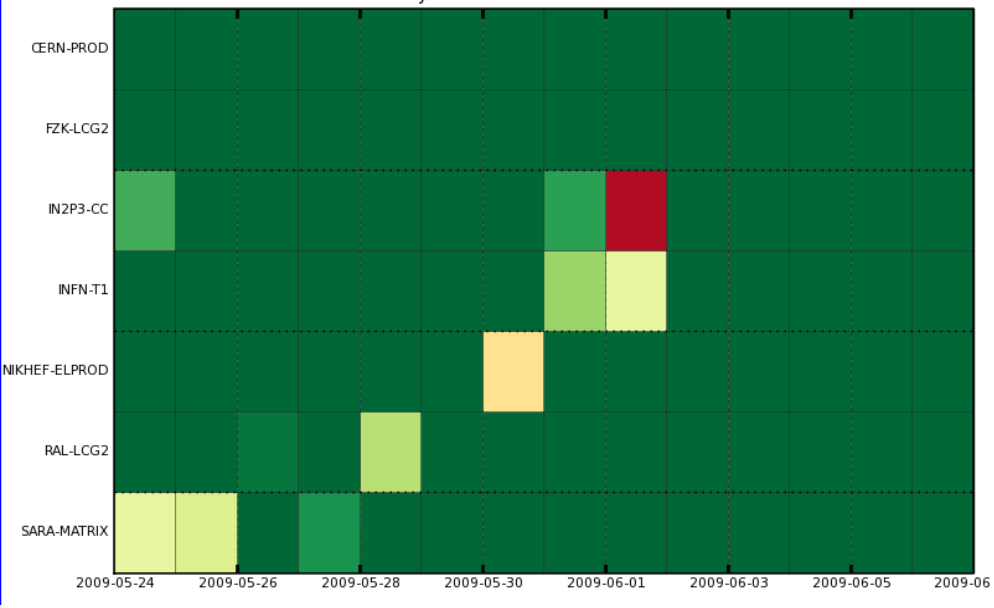
VO concerned	USER	TEAM	ALARM	TOTAL
ALICE	4	0	0	4
ATLAS	47	55	**0**	102
CMS	10	2	0	12
LHCb	3	20	**0**	23
Totals	64	77	**0**	141



Site Availability

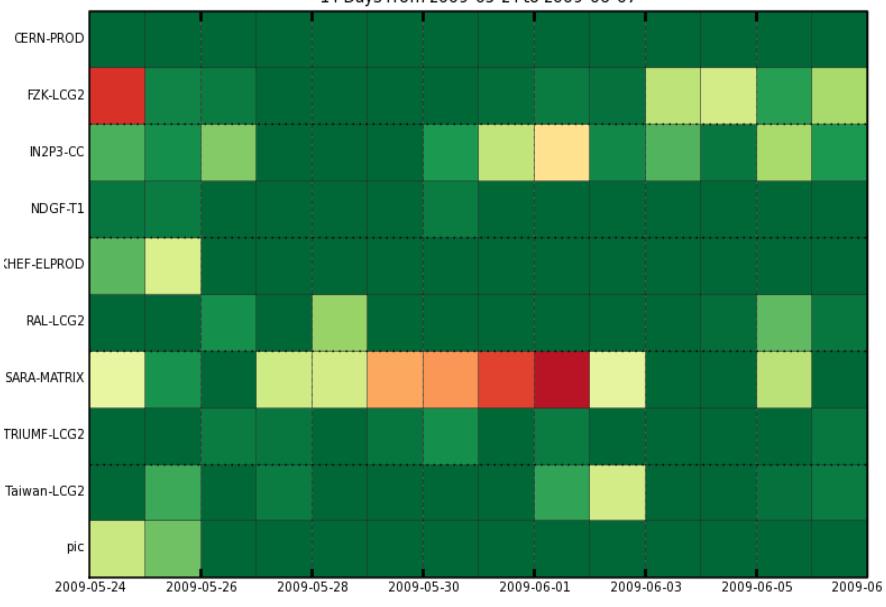
ALICE

14 Days from 2009-05-24 to 2009-06-07



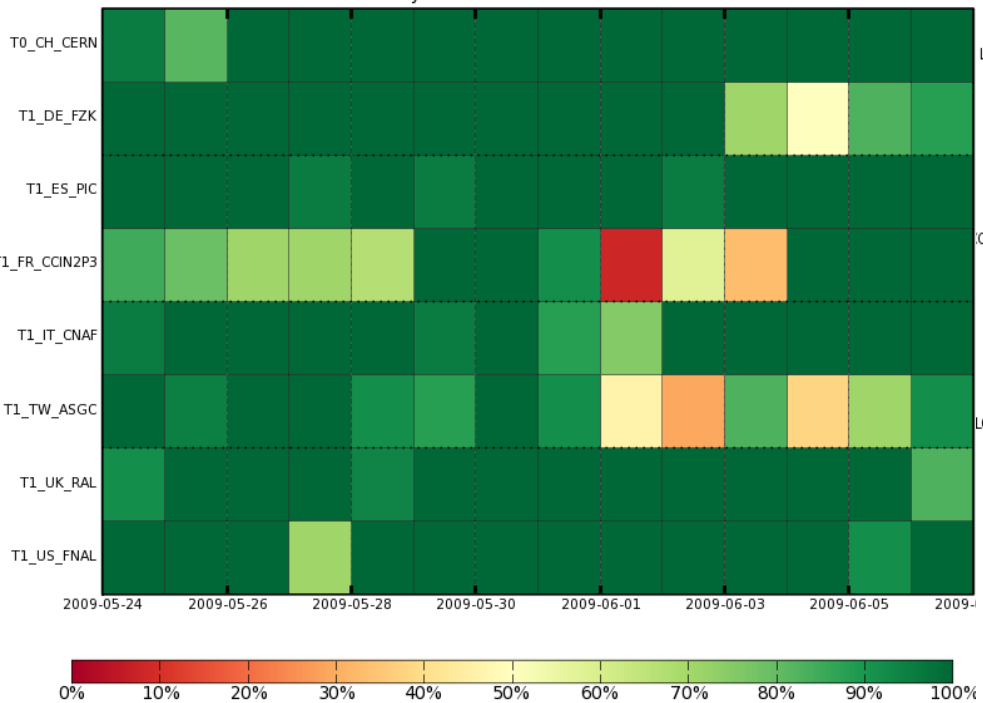
ATLAS

14 Days from 2009-05-24 to 2009-06-07



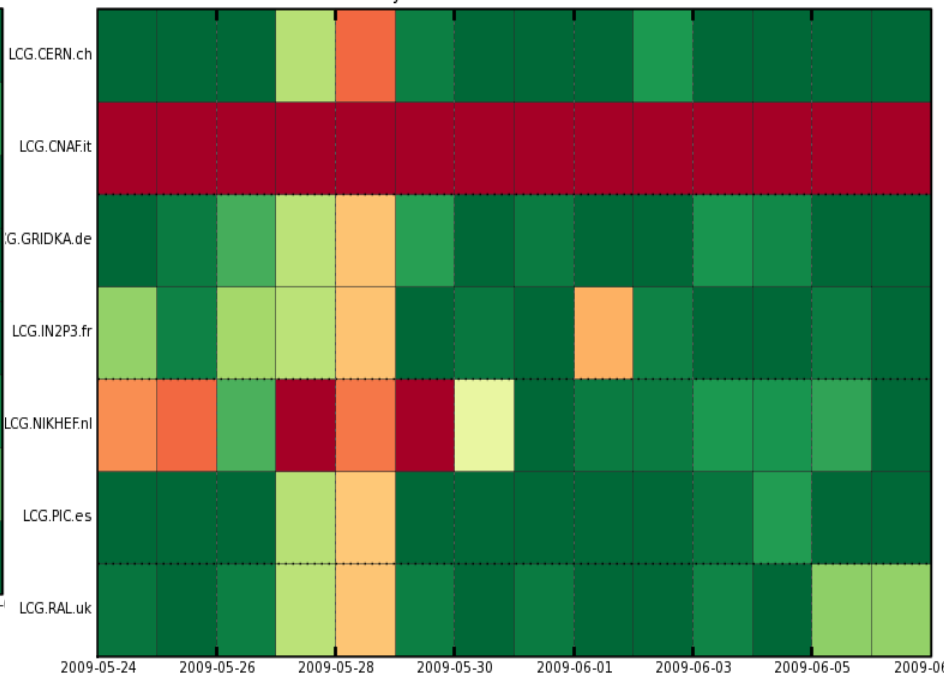
CMS

14 Days from 2009-05-24 to 2009-06-07



LHCb

14 Days from 2009-05-24 to 2009-06-07



Experiment Site Availability Issues (1/4)

- **LHCb: Since some time the visualisation of the LHCb critical availability at CNAF was showing either the T1 or T2 randomly. This now shows the T1 only which exposed that there were two tests which could not both succeed hence the site was always red. This will be corrected. NB – all green in later SAM talk !**
- **ATLAS: ASGC Issues - Conditions Database streams replication to ASGC was re-established synchronised on the Rutherford instance on 4 June. However, since then read-only access to the ASGC conditions database from the worker nodes has been failing stopping STEP'09 activities. It was realised that in the long period following Oracle reconfiguration at ASGC and the fire the port numbers for Oracle access had been changed without directly informing ATLAS operations. Correcting this would mean recompiling/reconfiguring ATLAS modules during STEP09. As a workaround they are trying to set up a second Oracle 'listener' but as yet ASGC has hardly participated in STEP09 at the halfway stage. News Tuesday morning: jobs have started working and a bulk pre-staging has been triggered – high failure rate reported. Steady data flow from CERN at 70-80 MB/sec.**

Experiment Site Availability Issues (2/4)

- **ATLAS: FZK – have announced it is unlikely their tape layer will be able to participate in STEP09 though they are working hard on their problems. ATLAS have been working with disk based data but last week their FZK SE was working slowly and was in serious trouble over last weekend. Transfer rates are commensurate with a 5% Tier1, not 10%, with 850 datasets (around 30TB) to catch up which will be challenging before the end of STEP09. FZK latest report is that they have built an alternative SAN fabric using old (decommissioned) switches. With this SAN fabric their tape connection to dCache is quite stable, however they cannot achieve the rates necessary for STEP09.**
- **LHCb+ATLAS: NL-T1 – upgraded dcache to 1.9.2-5 25 May then found gsidcap doors failing with large memory consumption (8GB). Eventually downgraded back to dcache 1.9.0-10 on Wednesday 27 May.**

Experiment Site Availability Issues (3/4)

- **ATLAS: IN2P3 – HPSS migration finished well as planned with the scheduled intervention over at 18.00 on Thursday 4 June and they are running a more recent version of HPSS on more powerful hardware.**
  - **ATLAS local experts then did manual tests to restore tape files and hit the bug already seen by LGCb that srm-ls reports wrong locality for a tape file implying the file is not available. They concluded it was not possible to read data back from tape and ATLAS did not schedule any tape based activities over that weekend as a result.**
  - **IN2P3 since demonstrated that lcg-cp will restore a file and that the issue was unrelated to the HPSS upgrade.**
  - **IN2P3 requested bulk prestaging by ATLAS (hence reprocessing) be delayed till Tuesday (today) to give time to ramp up a new component responsible for scheduling the tape staging requests sent to HPSS by dcache.**
  - **ATLAS perception was that a 4 day scheduled downtime lasted for 9 days**
  - **Follow up: improve communication between IN2P3 site operations and experiment operations.**

Experiment Site Availability Issues (4/4)

- **ATLAS: T1-T1 transfer tests showed many FTS timeouts for the new large (3-5 GB) merged AOD files. FTS timeouts have been systematically increased at all sites to typically 2 hours with good success.**
- **ATLAS: CERN – small 'user' files are still being written to tape via the 'default' pool/service class.**
- **ALICE: Batch jobs were failing to be scheduled at IN2P3 – traced to a bdii publishing failure at GRIF which hosts the French WMS for ALICE. This in turn has been reported as due to a site cooling problem and an alternative bdii is now being used.**
- **LHCb: Had an NFS problem at NL-T1 affecting LHCb software access on worker nodes after a user somehow caused NFS locks (which should not happen). NL-T1 had to reboot the lock server plus the concerned worker nodes.**
- **ALL: would like to understand what multi-VO activity is happening at sites and what are its affects.**

Experiment Activities: STEP'09 Summary of Goals

Experiment	Summary
ALICE	T0-T1 data replication (100MB/s) Reprocessing with data recalled from tape at T1
ATLAS	Parallel test of all main ATLAS computing activities at the nominal data taking rate -Export data from T0 -Reprocessing and reconstruction at T1 ,tape reading and writing, post-reprocessing data export to T1 and further toT2 -Simulation at T2 (real MC) -Analysis at T2 using 50% of T2 CPU, 25% pilot submission, 25% submitted via WMS
CMS	-T0 multi-VO tape recording -T1, special focus on tape operations data archiving & pre-staging. -Data transfer -Analysis at T2
LHCb	-Data injection into HLT -Data distribution to Tier1 -Reconstruction at Tier1

STEP'09 – Activities During 1st Week

- ***"Though there are issues discovered on daily basis, so far STEP09 activities look good"***
  - IT-GS "C5" report

- Tier0 multi-VO tape writing by ATLAS (320 MB/sec), CMS (600 MB/sec) and LHCb (70 MB/sec) happening now since a few days then CMS stop today (for weekly mid-week global run). CMS resume midnight 11 June, ALICE (100 MB/sec) should start then also then ATLAS stop midnight Friday 12 June. Would have liked a longer overlap but FIO report success so far.
- See GDB report tomorrow for detailed performance.

CASTOR Oracle BigID Issue

- Oracle has released a first version of the patch for the “BigId” issue (cursor confusion in some condition when parsing a statement which had already been parsed and for which the execution plan has been aged out of the shared pool). It was immediately identified that the produced patch is having a conflict with another important patch that we deploy. Oracle is working on another “merge patch” which should be made available in the coming days. In the meantime a work-around has been deployed in most of the instances at CERN as part of CASTOR 2.1.8-8.

Central Service Outages (1/2)

- A bug in the upgrade script from CASTOR 2.1.7 to CASTOR 2.1.8 has converted Disk 1 Tape 1 (D1T1) pools to Disk 0 Tape 1 (D0T1) during the upgrade performed on 27 May. The garbage collection has been activated in pools that had a tape copy of the disk data. Unfortunately, one of the diskpools of LHCb happened to contain files with no copy on tape, as these files were created at a time when the pool was defined as Disk 1 Tape 0 (D1T0 – Disk only). As a consequence, 7564 files have been lost. A post mortem from FIO is available as mentioned in their C5 report (https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem20090603)

Central Service Outages (2/2)

- On 5 June a prepared script was run to convert the remaining misidentified LHCb tape0disk1 files to become tape1disk1 files and hence be migrated to tape. Due to wrong logic in the check for the number of replicas triggered by this operation 6548 tape0disk1 files that had multiple diskcopies in a single service class have been lost.
- Probably triggered by the same operation, corrupted subrequests blocked the three instances of the JobManager. For this reason the service was degraded from 11:30 to 17:00.
- Post Mortem is at https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem20090604
- As a follow up whenever a bulk operation is to be executed on a large number of files, the standard practice should be to first run it on a small subsample of files (a few 100s), which have first been safely backed up somewhere outside or inside (as a different file) castor.

Central Service Outages (3/3)

- On Monday June 1st, around 8:20 am, the XFP (hot pluggable optical transceiver) in the router port, which connects the RAC5 public switch to the General Purpose Network, failed causing unavailability of several production databases including: ATLR, COMPR, ATLDSC and LHCBDSC. Also data replication from online to offline (ATLAS) and from tier0 to tier1s (ATLAS and LHCb) was affected. The hardware problem was resolved around 10am and all aforementioned databases became available ~15 minutes later. Streams replication was restarted around 12:00.
- During the whole morning some connection anomalies were also possible in the case of ATONR database (ATLAS online) which is connected to the affected switch for monitoring purposes. The XFP failure caused one of the Oracle listener processes to die. The problem was fixed around 12:30.
- Detailed post-mortem has been published at https://twiki.cern.ch/twiki/bin/view/PSSGroup/StreamsPostMortem.

WLCG Service Summary

- Three Tier 1 did not fully participate in week 1 of STEP09. IN2P3 now fully functional and ASGC has just restarted. FZK tape layer working in degraded mode.
- Many STEP09 activities going well.
- Communications routes from Tier 1/2 site operations to experiment local and central operations needs to be documented.
- We need to be able to identify multi-VO activity at sites and any possible interference between VOs. Summarise at Post Mortem workshop.
- Serious data loss at CERN for LHCb due to complex bugs. Vital to carefully pre-test any bulk operations.
- Week 2 should now be maximum overlap of all VOs with maximum load from multiple activities.