WLCG Service Report

Harry.Renshall@cern.ch

~ ~ ~

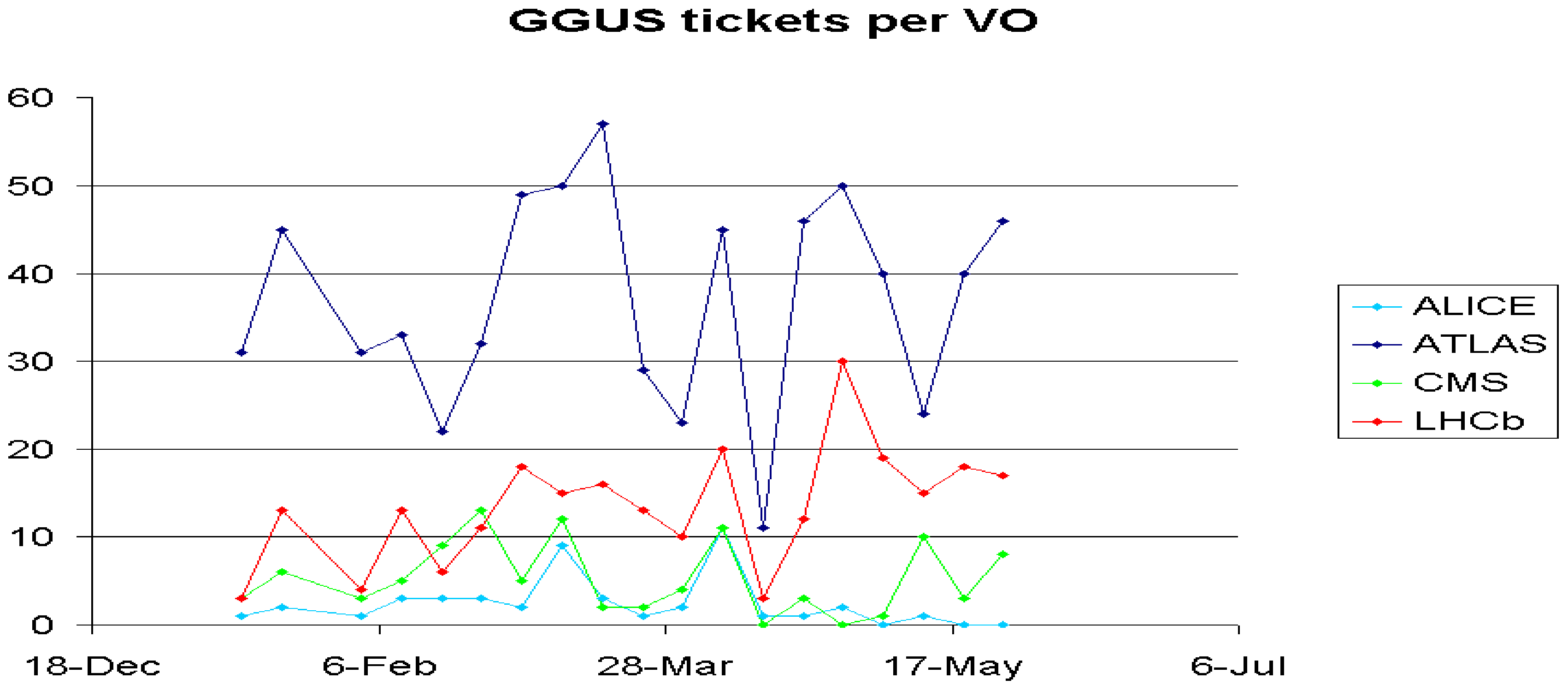
WLCG Management Board, 26th May 2009

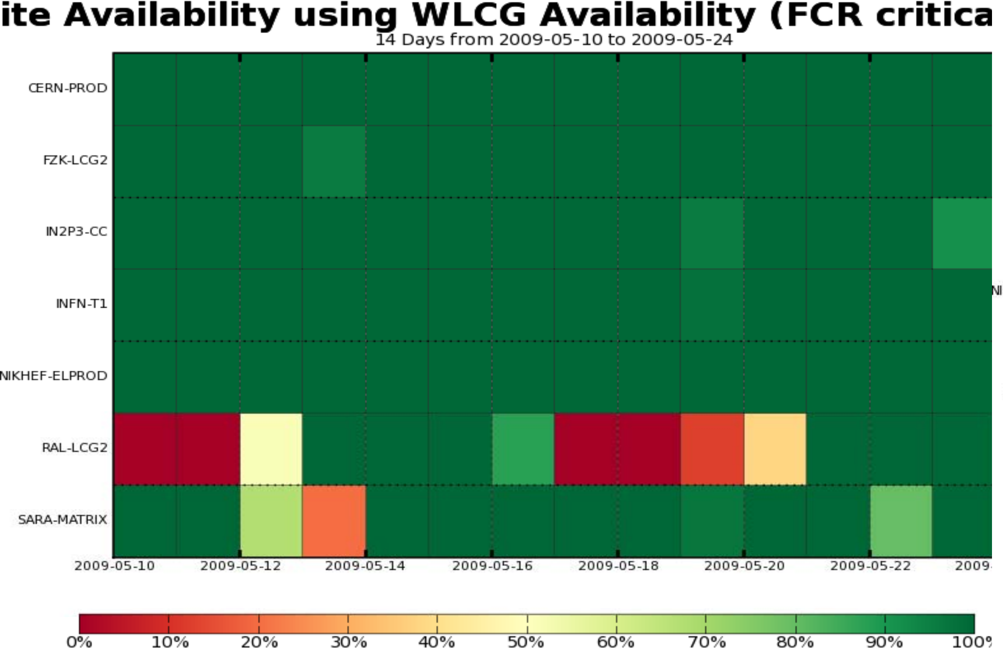
Introduction

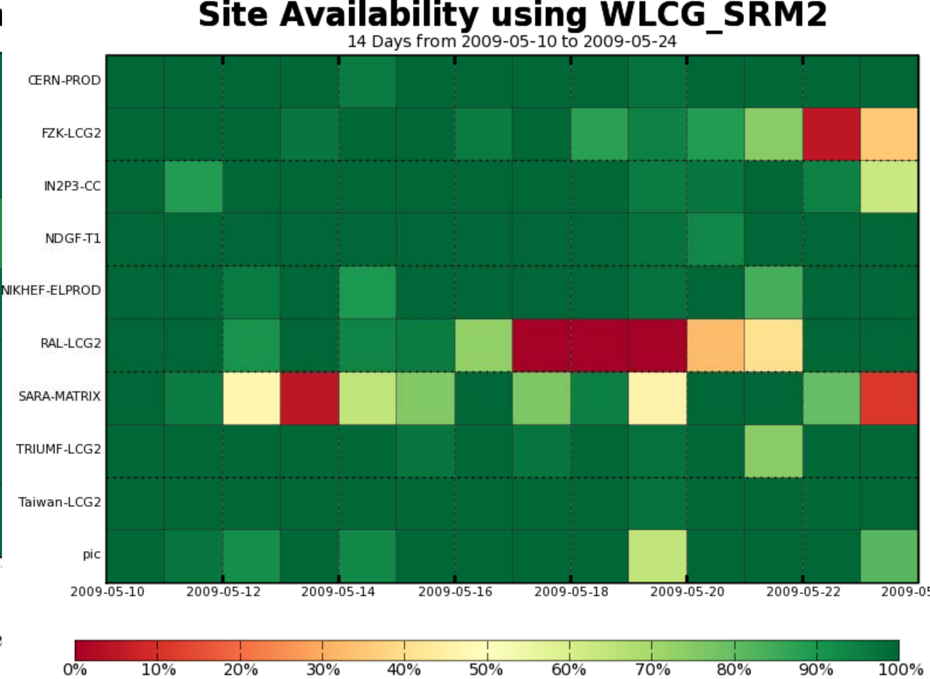
- This report covers the service for the two weeks period 10 to 23 May
- GGUS ticket rate normal and one alarm ticket.
  - 19 May There was a GPN network connectivity failure to CERN from just before 10.00 to just after 11.00 due to a GEANT reconfiguration which cut off CERN addresses. CMS started losing production jobs and raised a CERN alarm ticket which was not escalated to the data operations piquet as per procedure as it did not concern the services of castor, fts, lfc or srm. In this case the ticket only arrived when the GPN came back. How to handle such cases to be followed up.
- Three Central Service incidents: failure to publish SAM test results for several hours after the above WAN incident, lost batch jobs after Linux upgrade and SRM upgrade rolled back after causing multi-file 'get's to fail.
- One incident leading to SIR submitted to WLCG
  - PIC 14 May 5 hours cooling down

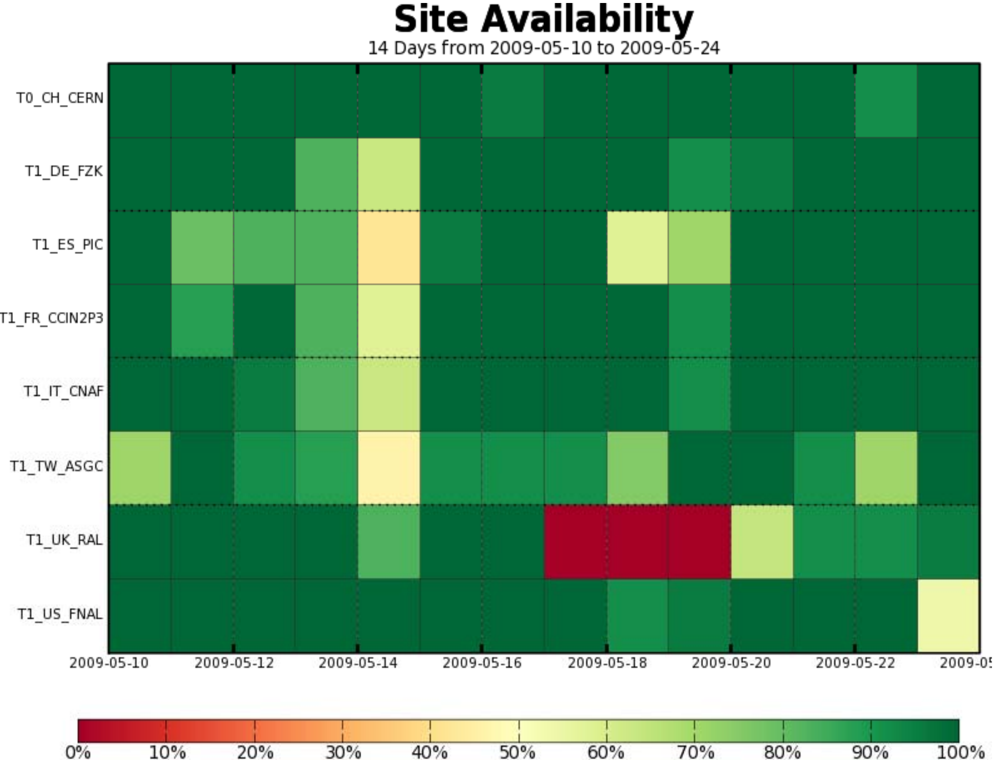
GGUS Summaries – 2 weeks

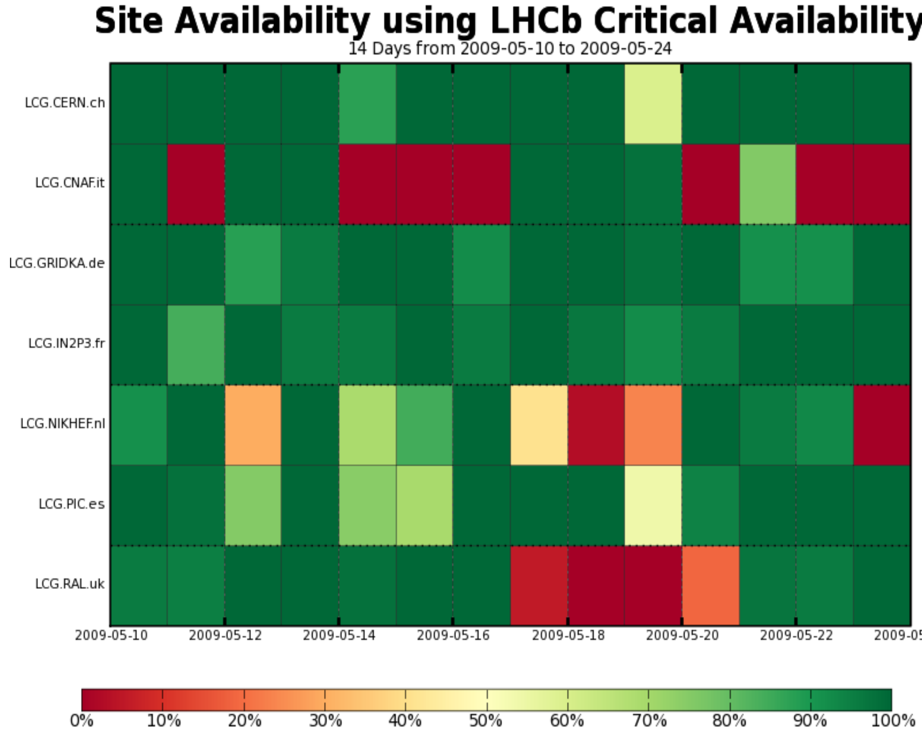
VO concerned	USER	TEAM	ALARM	TOTAL
ALICE	1	0	0	1
ATLAS	46	64	**0**	110
CMS	18	0	1	19
LHCb	5	46	**0**	51
Totals	70	110	**1**	181




ite Availability using WLCG Availability (FCR critica
14 Days from 2009-05-10 to 2009-05-24
CERN-PROD
FZK-LCG2
IN2P3-CC
INFN-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
2009-05-10
2009-05-12
2009-05-14
2009-05-16
2009-05-18
2009-05-20
2009-05-22
2009
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


Site Availability using WLCG_SRM2
14 Days from 2009-05-10 to 2009-05-24
CERN-PROD
FZK-LCG2
IN2P3-CC
NDGF-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
TRIUMF-LCG2
Taiwan-LCG2
pic
2009-05-10
2009-05-12
2009-05-14
2009-05-16
2009-05-18
2009-05-20
2009-05-22
2009-05
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


Site Availability
14 Days from 2009-05-10 to 2009-05-24
T0_CH_CERN
T1_DE_FZK
T1_ES_PIC
T1_FR_CCIN2P3
T1_IT_CNAF
T1_TW_ASGC
T1_UK_RAL
T1_US_FNAL
2009-05-10
2009-05-12
2009-05-14
2009-05-16
2009-05-18
2009-05-20
2009-05-22
2009-05
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


Site Availability using LHCb Critical Availability
14 Days from 2009-05-10 to 2009-05-24
LCG.CERN.ch
LCG.CNAF.it
LCG.GRIDKA.de
LCG.IN2P3.fr
LCG.NIKHEF.nl
LCG.PIC.es
LCG.RAL.uk
2009-05-10
2009-05-12
2009-05-14
2009-05-16
2009-05-18
2009-05-20
2009-05-22
2009-05
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

Experiment Site Availability Issues (1/2)

- **ALL: saw the extended RAL downtime moving backend Oracle databases to new hardware**
- **ATLAS:**
  - **ASGC now considered fully functional after local router hardware problem was resolved. ASGC still suffering from Oracle BigID bug - workaround in preparation.**
  - **LFC server at PIC become overloaded during a centrally run file deletion operation although this rate of deletion has been done before. May be some strange LFC entries. PIC to do manual deletion.**
  - **This happened at the weekend and it was nor clear at PIC who to contact in ATLAS to exchange information and plan actions. ATLAS are suggesting a new sites to experiment mailing list that would have restricted posting (say 20 names) and end up sending an SMS to the ATLAS expert on duty.**

Experiment Site Availability Issues (2/2)

- **ALICE: Smooth production running.**

- **CMS: GPN to CERN failure 19 May stopped Phedex transfers and lost some batch jobs.**

- **LHCb: Issue of SRM not returning Turls at FZK now understood. Dirac does not always set the ‘done’ state after a get request causing this instability.**

- **AT CNAF a host certificate expired on an LHCb storm diskserver. Ticket routing was somehow not optimal and will have taken 5 days to fix.**

Experiment Activities

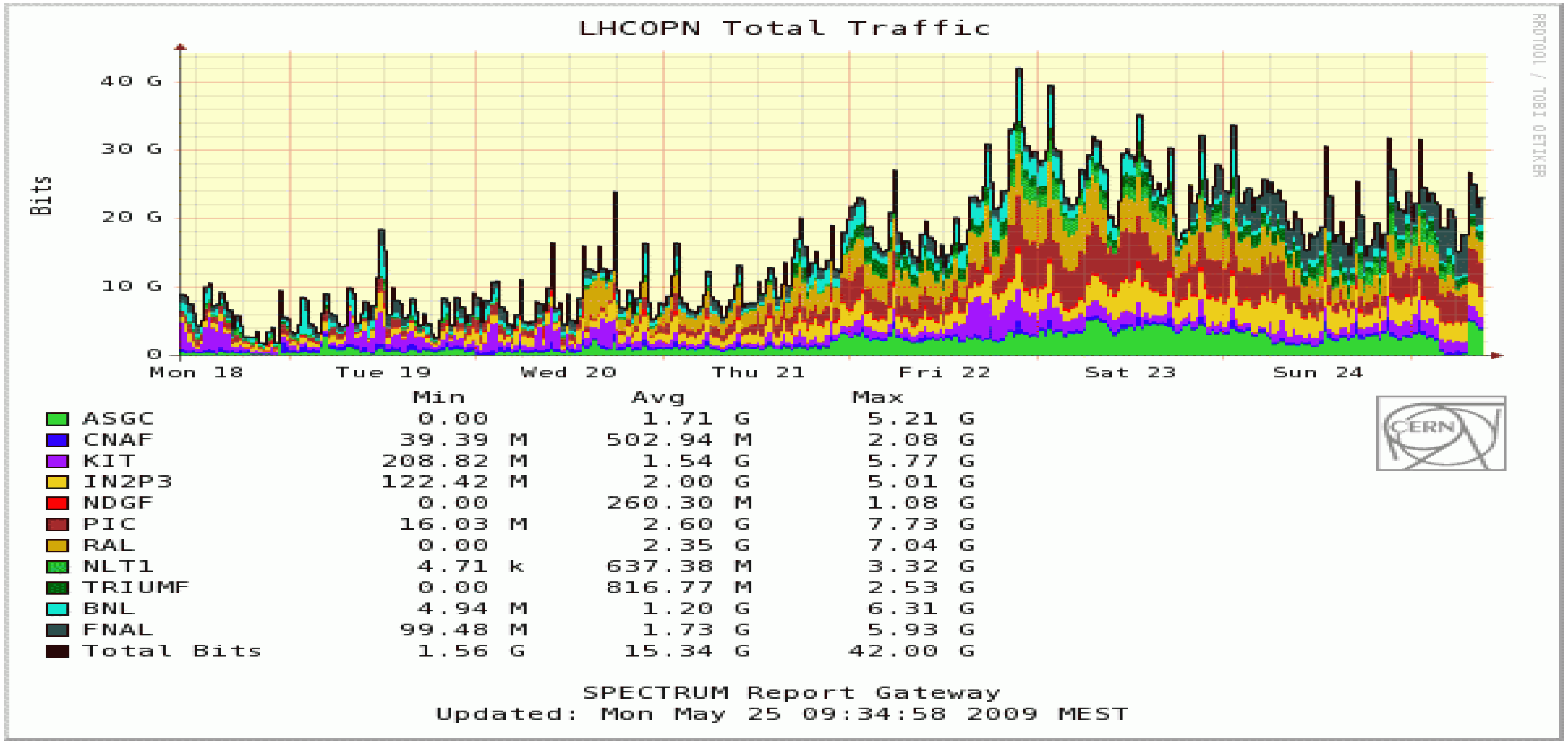
- ATLAS: Preparing for STEP'09 by distributing cosmics raw data to T1 (for reprocessing – all finished except ASGC), AOD (for T1-T1 tests and analysis at T2) and DPD (for analysis at T2 sites). All clouds involved. Ramp up of Step'09 from 2 June.
- LHCb: Week of 090511 was a FEST week.
- CMS: Good rate of ticket resolution in recent weeks – many Tier 2 are now able to reach 80% availability.
- CMS Step'09 planning meeting held with prestaging as main subject. Small scale prestaging tests at Tier 1 to be run this week.
- CRUZET (Cosmics at Zero Tesla) run at CERN next week

GPN WAN Incident

1 hour failure 10.00-11.00 19 May. Geant, the main Internet provider for most of the European research and education institutes, applied a wrong configuration to its routers that resulted in CERN being unreachable to all of their customers.

LHCOPN Traffic Increase

Weekly



LHCOPN Total Traffic

	Min	Avg	Max
🟩 ASGC	0.00	1.71 G	5.21 G
🟦 CNAF	39.39 M	502.94 M	2.08 G
🟪 KIT	208.82 M	1.54 G	5.77 G
🟨 IN2P3	122.42 M	2.00 G	5.01 G
🟥 NDGF	0.00	260.30 M	1.08 G
🟫 PIC	16.03 M	2.60 G	7.73 G
🟧 RAL	0.00	2.35 G	7.04 G
▨ NLT1	4.71 k	637.38 M	3.32 G
▨ TRIUMF	0.00	816.77 M	2.53 G
🟦 BNL	4.94 M	1.20 G	6.31 G
⬛ FNAL	99.48 M	1.73 G	5.93 G
⬛ Total Bits	1.56 G	15.34 G	42.00 G

SPECTRUM Report Gateway
Updated: Mon May 25 09:34:58 2009 MEST

Factor of 3 average traffic increase this last weekend - ATLAS distributing data for STEP’09

PIC Cooling Incident (1/3)

- Useful SIR for other sites notably on restart issues and follow up

- On the 14th May 2009, around 15h, there was an emergency at the PIC machine room because the cooling equipment stopped working due to a power cut. At that time we didn't have any information about the root cause of that electrical incident, so we took the decision to stop the whole PIC site in order to avoid an ungraceful power off on our equipment.
- One hour and a half later, about 16:30, the electrical maintenance supervisor validated the power stability on the building infrastructure equipment, so we proceeded to start PIC site. About 19:00 all PIC services where up again.
- The cause of this electrical incident was due to an error in the installation supplying power to the building infrastructure. This problem was permanently solved on Friday (15h May)

PIC Cooling Incident (2/3)

- Most of the services could be restarted normally however some issues were found:
- Oracle: databases were stopped gracefully but the cabin power off was not clean. Because of that, we had to configure a raw device manually for FTS database before starting that service.
- Storage: Some problems to start Enstore service because tape servers were started back before our IBM robot was completely up. It was necessary to reboot both services.
- Computing: 16 workernodes from the same HP c7000 Blades initiated a reinstallation when started due to an human error. On two of them, the instalation finished unsuccesfully and acted as black holes causing errors in SAM tests until 2:00.
- Bacula backups: Some manual intervention was required because the storage daemon couldn't dismount a tape before the power off.

PIC Cooling Incident (3/3)

- **Follow-up**
- Accelerate the deployment of a tool to stop PIC services as fast as possible in a graceful way.
- MoD laptop didn't have good coverage to connect to UAB wifi. We should ensure that the actual provider has coverage in our offices in order to send the broadcast as soon as possible.
- Collect information about the UPS in order to distinguish if an electric incident has any impact on the machines room or in the cooling system.
- Review if it's possible to trigger an alarm when there are worker nodes acting as black holes.

Dcache at NL-T1

- NL-T1 reported this week that their 20 May upgrade of dcache to level 1.9.2-5 has been causing problems firstly with memory issues with a java virtual machine then with hung servers. They are in touch with dcache developers but are preparing to downgrade to their previous 1.9.0-10 release in order to participate in STEP’09.

CASTOR at CERN

- Last major CASTOR upgrade to 2.1.8-7 for LHCb this week (Wednesday).
- Minor (should be transparent) upgrades to 2.1.8-8 for ALICE, ATLAS and c2cernt3 (CMS and ATLAS analysis stager) also Wednesday.

Central Service Outages (1/2)

- The external network connection problems from Tuesday morning 19th affected all experiments to some degree, but CMS seems to have had the biggest problems due to the distributed nature of the Phedex service, which relies on the GPN. CMS also lost batch jobs.
- SAM test results not published on Tuesday for several hours due to problems with the DB and tomcat on the SAM servers. This seems to have been caused by the network problem and lasted until mid-afternoon.
- 12 May a scheduled Linux kernel upgrade missed having an LSF configuration applied. As a consequence, the SLC4 BATCH nodes threw away hundreds of PENDING jobs (already-RUNNING jobs were not affected). SLC5 BATCH was apparently not affected. Due to the LSF "blackhole" detection, (several tens of) SLC4 LXBATCH nodes were automatically removed from scheduling. This had not been detected during the testing phase since the configuration was applied.

Central Service Outages (2/2)

- May 13 the SRM 2.7-17 upgrade on all production services contained a bug which caused
  - **a major outage for LHCb's job processing (several hundred failures per hour) since they were unable to access (via RFIO) their data files (during their FEST week).**
  - **a 100% outage on the CERN->FNAL data export for CMS**

for a period of around 1 day. The bug was for the case of multiple files in an srm_get and was not found in the PPS testing.

- May 22 the router hosting the main CERN Firewall crashed at 09:59. The traffic moved automatically to the secondary path and returned to the main path when the router recovered. Traffic was then forced to the secondary path to prevent more crashes. This operation might have caused disruptions for long-time connections not using the HTAR, being almost unnoticeable for Web Internet traffic.

WLCG Service Summary

- Another Tier 1 power/cooling incident but a quick recovery.
- A GPN failure to which CMS is particularly vulnerable. Reporting mechanism is unclear.
- Mechanism required for site operations to communicate with the right people in experiment operations (not new).
- Mixture of hardware failures and scheduled upgrades throwing up problems not seen in testing.
- More instability than we would like at this stage
- STEP’09 activities start next week – analysis and data movement (at least) – can we expose more the planned activities and metrics svp.