# WLCG Service Report

Harry.Renshall@cern.ch
Jamie.Shiers@cern.ch
Jean-Philippe.Baud@cern.ch

~~~

**WLCG Management Board, 2nd March 2010**
~~~

# Introduction

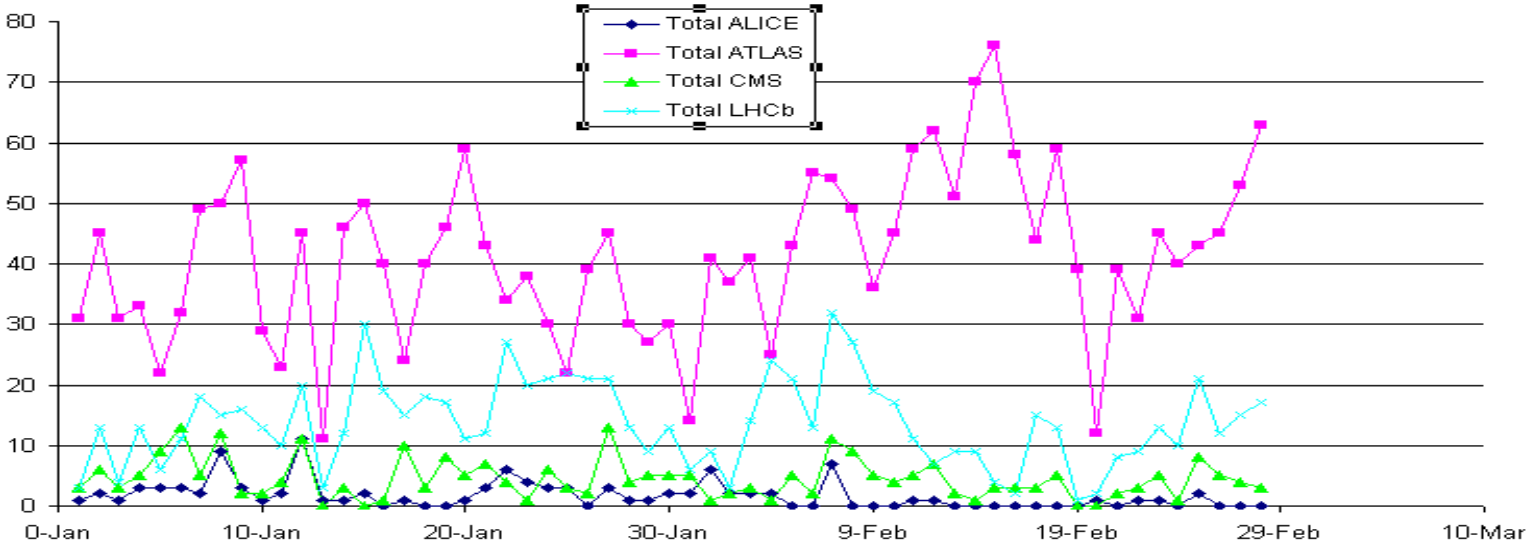
- Covers the three weeks 8 to 28 February.
- Usual rich mixture of problems
- 6 alarm tickets:
  - 4 were GGUS tests
  - 2 were ATLAS CERN xrootd server failures:
    - Xroot access to atlascerngroupdisk failing. Ticket submitted at 08.10 Monday 15 Feb, solved at 09.15 by daemon restart.
    - Same problem on 18 Feb reported at 10:14 and 2-3 other times later. Marked as 'solved by installing a new version' at 10.30 on 19th. NB rfcp was working in the meantime – a partial workaround.
- Incidents leading to (eventual) service incident reports
  - PIC on 10 Feb: Complete blackout for 7 hours of services involving grid certificates either personal or host from Spanish CA at CERN: Affected Voms, FTS, etc.
  - IN2P3 on 15 Feb: Local worker nodes lost network connectivity for 4 hours
  - ASGC on 25th Feb: Power outage caused by testing of new backup supply

# Meeting Attendance Summary (Last week only)

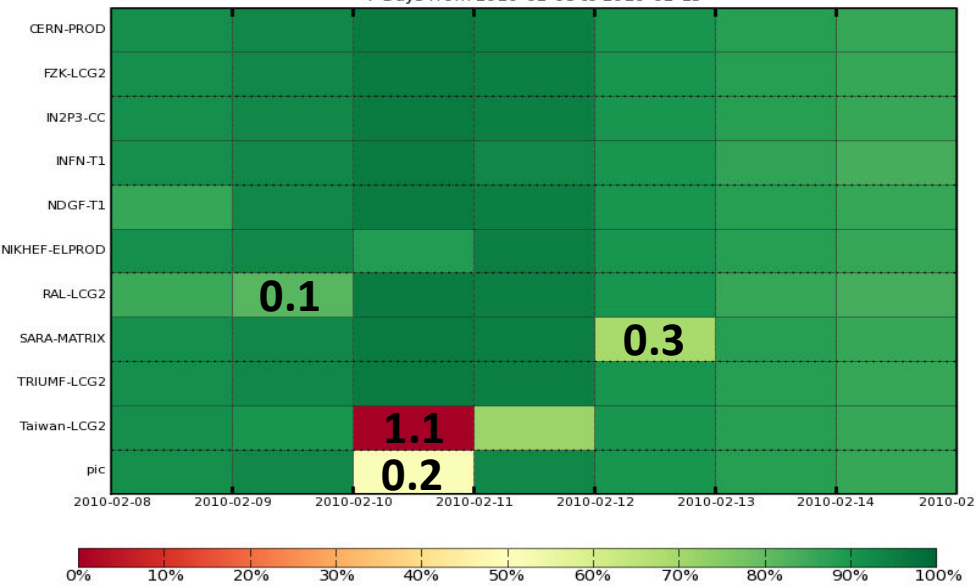
| Site | M | T | W | T | F |
|---|---|---|---|---|---|
| CERN | Y | Y | Y | Y | Y |
| ASGC | Y | Y | Y | Y | Y |
| BNL | Y | Y | Y | Y | Y |
| CNAF | | | | | Y |
| FNAL | Y | Y | Y | Y | Y |
| KIT | Y | Y | Y | Y | Y |
| IN2P3 | Y | Y | Y | Y | Y |
| NDGF | Y | | | Y | Y |
| NL-T1 | Y | Y | Y | Y | |
| PIC | Y | Y | | Y | Y |
| RAL | Y | Y | Y | Y | Y |
| TRIUMF | | | | | |

# GGUS summary (3 weeks)

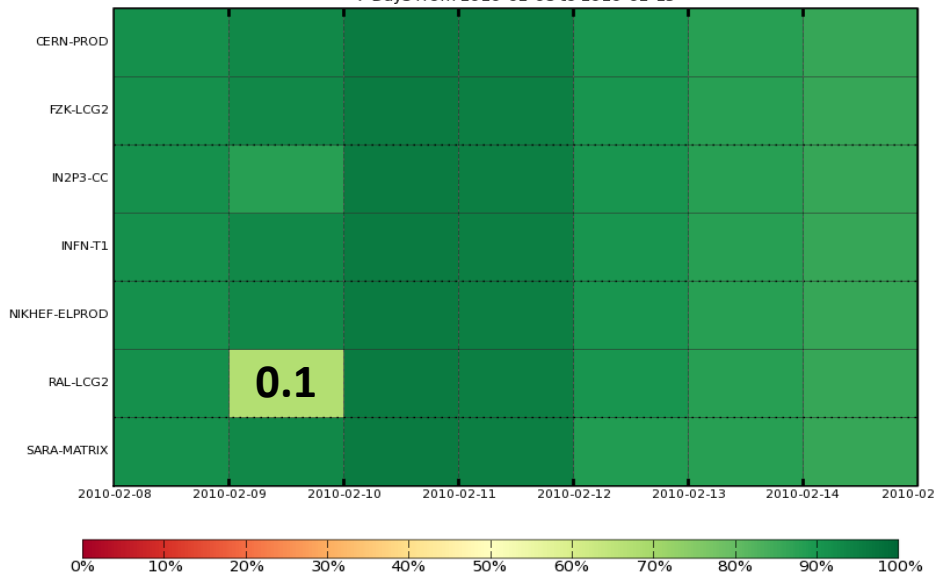
| VO | User | Team | Alarm | Total |
|---|---|---|---|---|
| ALICE | 0 | 0 | 0 | 0 |
| ATLAS | 36 | 119 | 6 | 161 |
| CMS | 8 | 4 | 0 | 12 |
| LHCb | 2 | 42 | 0 | 44 |
| Totals | 46 | 165 | 6 | 217 |



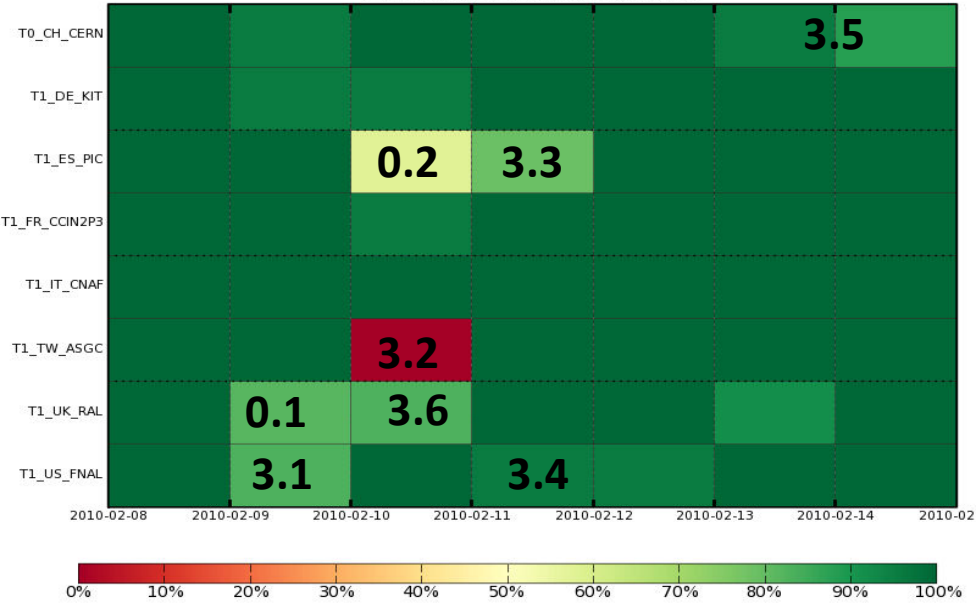
## ATLAS
### Site Availability using WLCG_SRM2



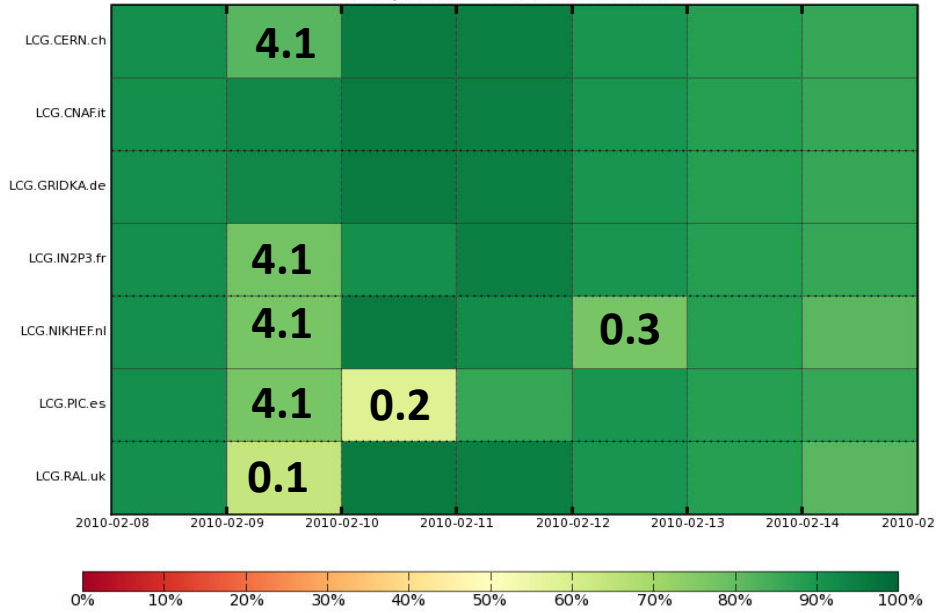

## ALICE
### ite Availability using WLCG Availability (FCR critical




## CMS
### Site Availability




## LHCb
### Site Availability using LHCb Critical Availability



# Analysis of the availability plots

**COMMON FOR THE ALL EXPERIMENTS**

**0.1 RAL:** Outage in the morning for the network reconfiguration: some test failures were expected

**0.2 PIC:** problem with CRL

**0.3 SARA:** Scheduled downtime for disk server firmware upgrade and new kernel

**ATLAS**

**1.1 Taiwan:** Scheduled Downtime: Install alternate power supply

**ALICE**

Nothing to report

**CMS**

**3.1 FNAL:** Power + cooling incident due to an EPO circuit problem. CMS services recovered in 2-3 hours. An authorization module failure which made some SAM jobs fail

**3.2 ASGC:** Scheduled downtime

**3.3 PIC:** Scheduled downtime for 1h (took 1h more). Problem 32/64bit issue with dCache upgrade. dCache team has been contacted. Jobs suspended during intervention - no jobs crashed

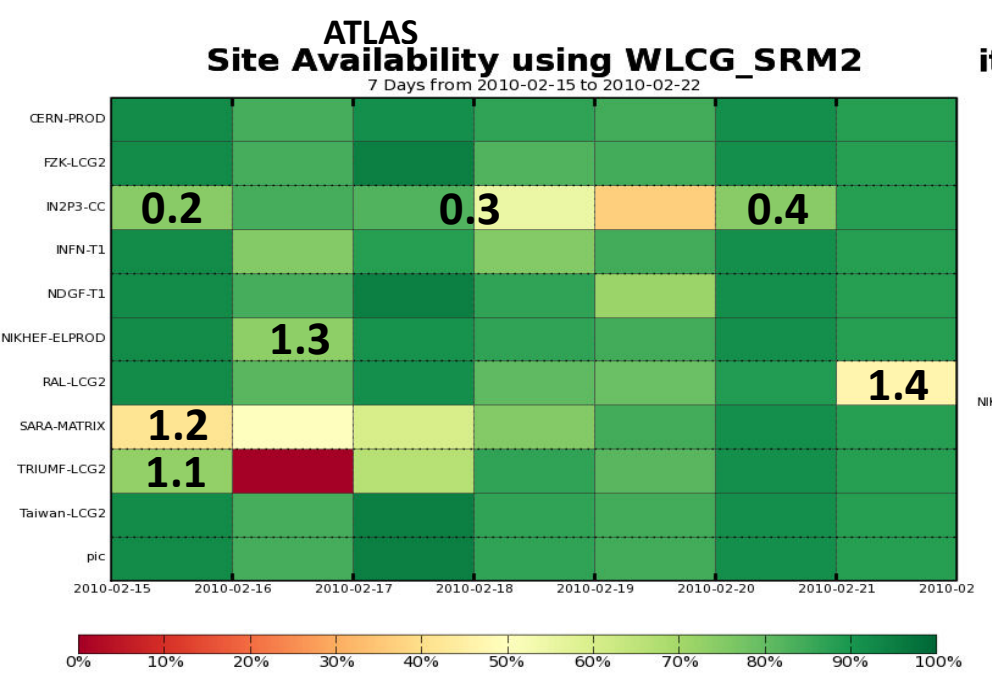
**3.4 FNAL:** Scheduled downtime

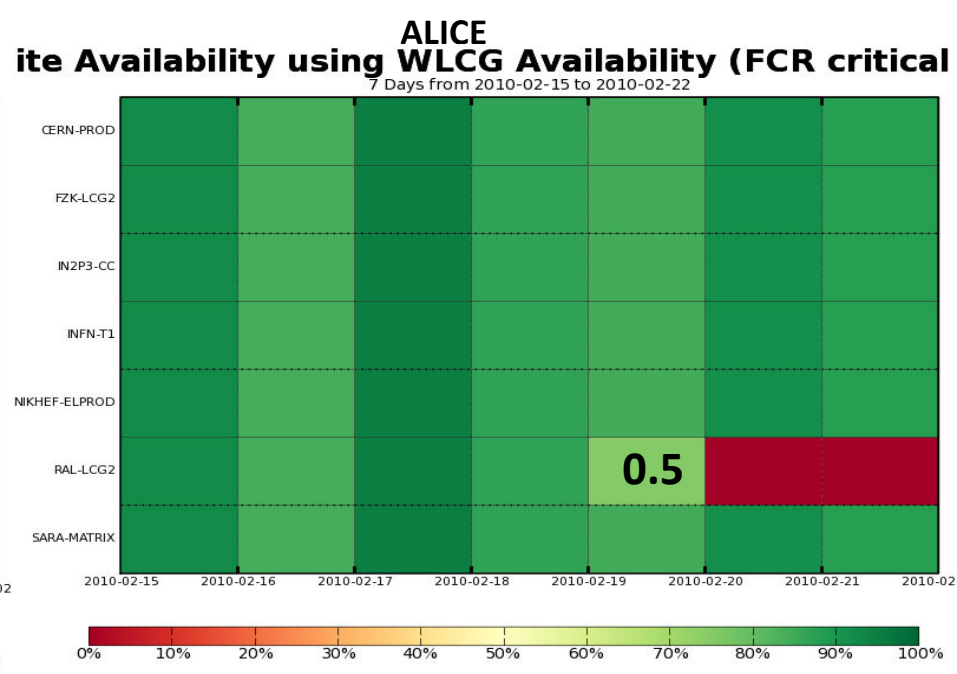
**3.5 CERN:** Some SRM test failures over the week-end

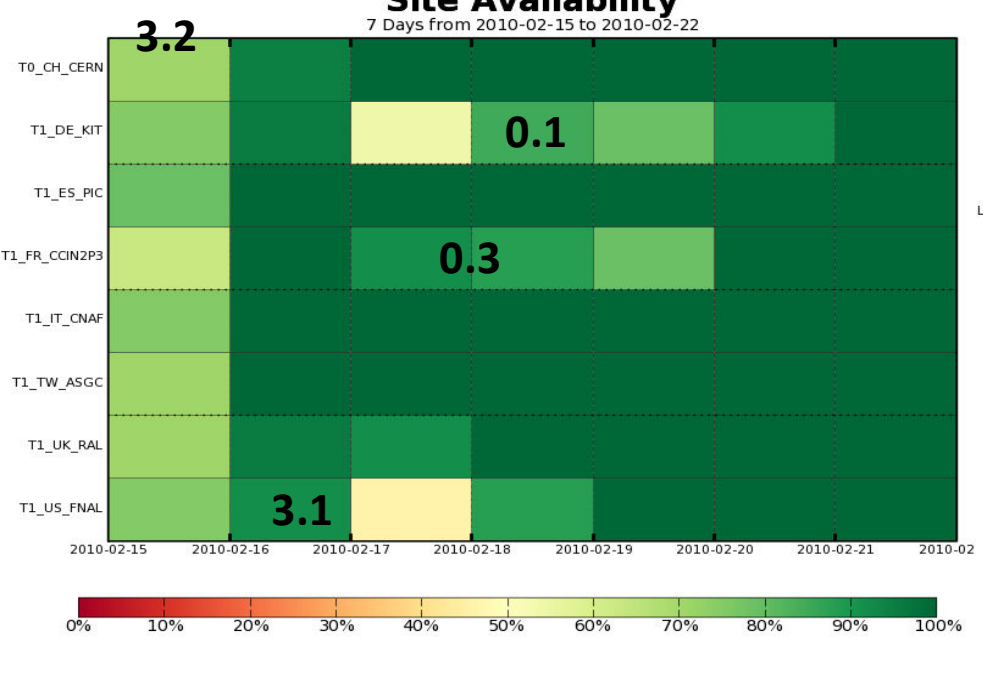
**3.6 RAL:** lcgadmin job stuck in queue blocking new sw installation jobs, fixed by site admins

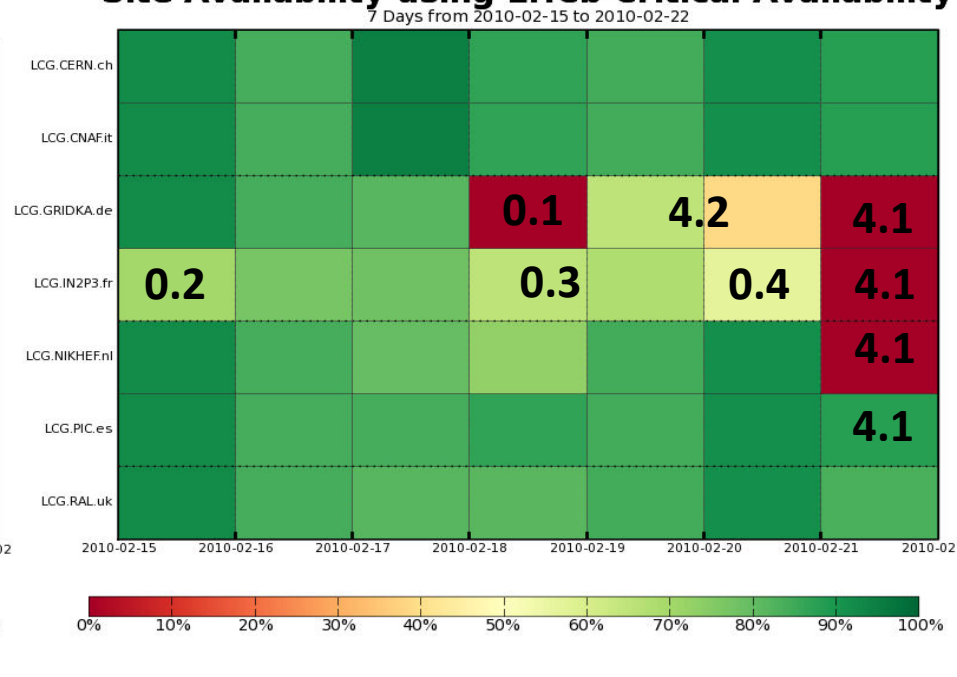
**LHCb**

**4.1** Internal problem with software installation modules: only CNAF among T1's seems to have the LHCB application properly installed. Some SAM tests were not publishing the information. LHCb experts fixed the problem


ATLAS
Site Availability using WLCG_SRM2
7 Days from 2010-02-15 to 2010-02-22
CERN-PROD
FZK-LCG2
IN2P3-CC
INFN-T1
NDGF-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
TRIUMF-LCG2
Taiwan-LCG2
pic
0.2
0.3
0.4
1.3
1.4
1.2
1.1
2010-02-15
2010-02-16
2010-02-17
2010-02-18
2010-02-19
2010-02-20
2010-02-21
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


ALICE
ite Availability using WLCG Availability (FCR critical
7 Days from 2010-02-15 to 2010-02-22
CERN-PROD
FZK-LCG2
IN2P3-CC
INFN-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
0.5
2010-02-15
2010-02-16
2010-02-17
2010-02-18
2010-02-19
2010-02-20
2010-02-21
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


CMS
Site Availability
7 Days from 2010-02-15 to 2010-02-22
T0_CH_CERN
T1_DE_KIT
T1_ES_PIC
T1_FR_CCIN2P3
T1_IT_CNAF
T1_TW_ASGC
T1_UK_RAL
T1_US_FNAL
3.2
0.1
0.3
3.1
2010-02-15
2010-02-16
2010-02-17
2010-02-18
2010-02-19
2010-02-20
2010-02-21
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


LHCb
Site Availability using LHCb Critical Availability
7 Days from 2010-02-15 to 2010-02-22
LCG.CERN.ch
LCG.CNAF.it
LCG.GRIDKA.de
LCG.IN2P3.fr
LCG.NIKHEF.nl
LCG.PIC.es
LCG.RAL.uk
0.1
4.2
4.1
0.2
0.3
0.4
4.1
4.1
4.1
2010-02-15
2010-02-16
2010-02-17
2010-02-18
2010-02-19
2010-02-20
2010-02-21
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

# Analysis of the availability plots

**COMMON FOR THE ALL EXPERIMENTS**

**0.1 KIT:** Planned downtime: OS update on router and file servers. Finished successfully.

**0.2 IN2P3:** Unscheduled outage Monday afternoon

**0.3 IN2P3:** H/w problem with machine running SRM on Wednesday. New unscheduled SRM downtime on Thursday: firmware update

**0.4 IN2P3:** SRM problem in the afternoon. A server in a rack had a power supply problem that tripped switches in the same rack. Recovered at about 18.00

**0.5 RAL:** An outage (from 19.02 to 23.02) on lcgce07 to allow for a drain and a disk replacement

**ATLAS**

**1.1 TRIUMF:** The migration to Chimera. Back in business on Wednesday

**1.2 SARA:** SRM timeouts due to server overload. Unscheduled downtime on Tuesday morning. SRM upgraded. DPM upgrade to gLite 3.2. Firmware upgrade for disk storage. SRM instabilities on Wednesday. On Thursday – the problem with one of dCache head nodes /var filled up. Fixed. SRM issues solved by increasing memory in the SRM server and improving the tuning of POSTGRES.

**1.3 NIKHEF:** Planned disk server firmware upgrade

**1.4 RAL:** MCDISK storage issues

**ALICE – Nothing to report**

**CMS** **3.1 FNAL:** Switch crash on Tuesday, back online within 2 hours. Significant building power issues on Wednesday, CMS out 04.30 to 19.00 local time when back with all services available.

**3.2** Internal CMS problem, not affecting the sites. Fixed quickly.

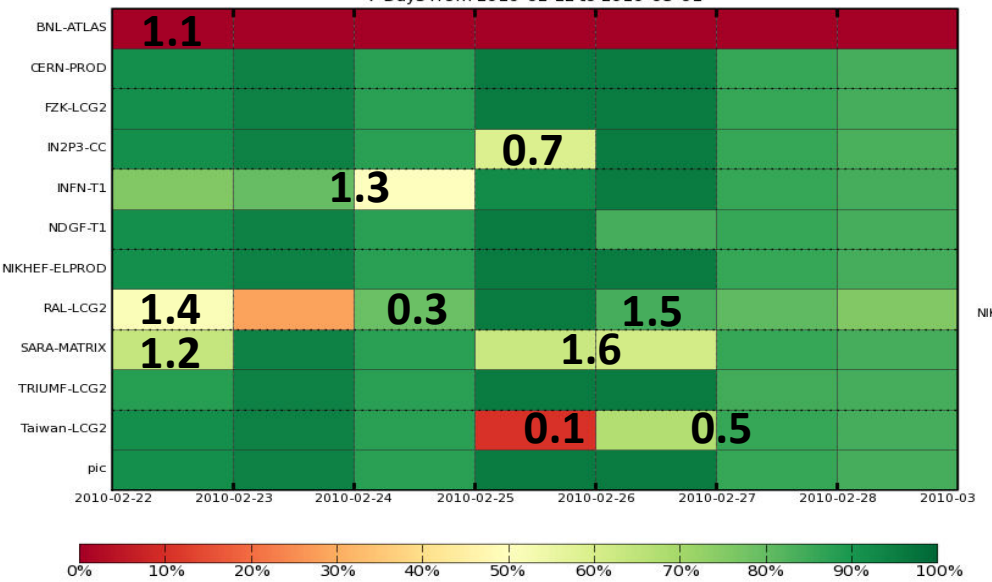
**LHCb**

**4.1** LHCb SAM tests status configuration changes for dCache sites: from 'info' to 'error', to track the fileAccess error more efficiently. May appear in next days. Will not appear in the new version of ROOT

**4.2 GRIDKA:** SQLite locks issues (NFS locking) at the site (GGUS ticket submitted)

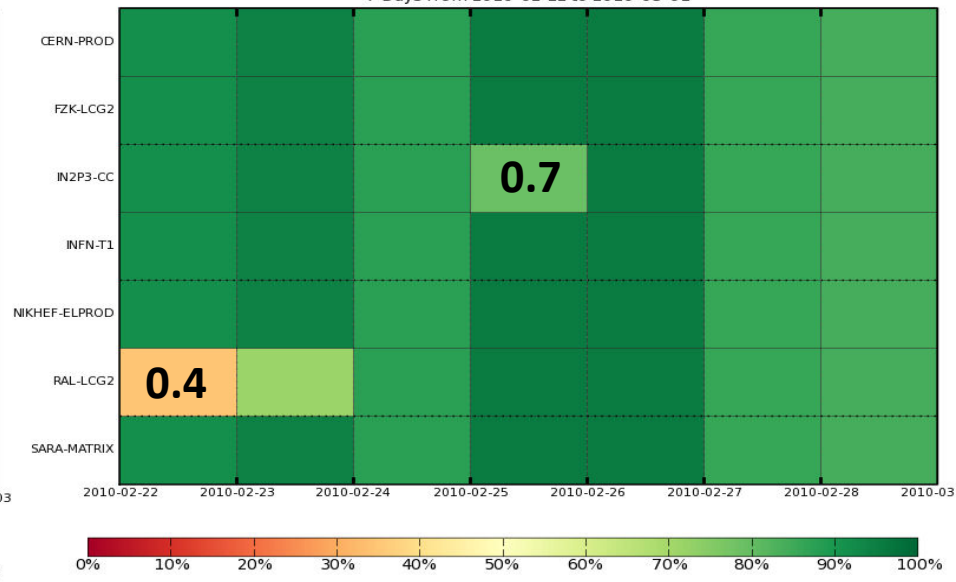
## ATLAS
## Site Availability using WLCG_SRM2

7 Days from 2010-02-22 to 2010-03-01



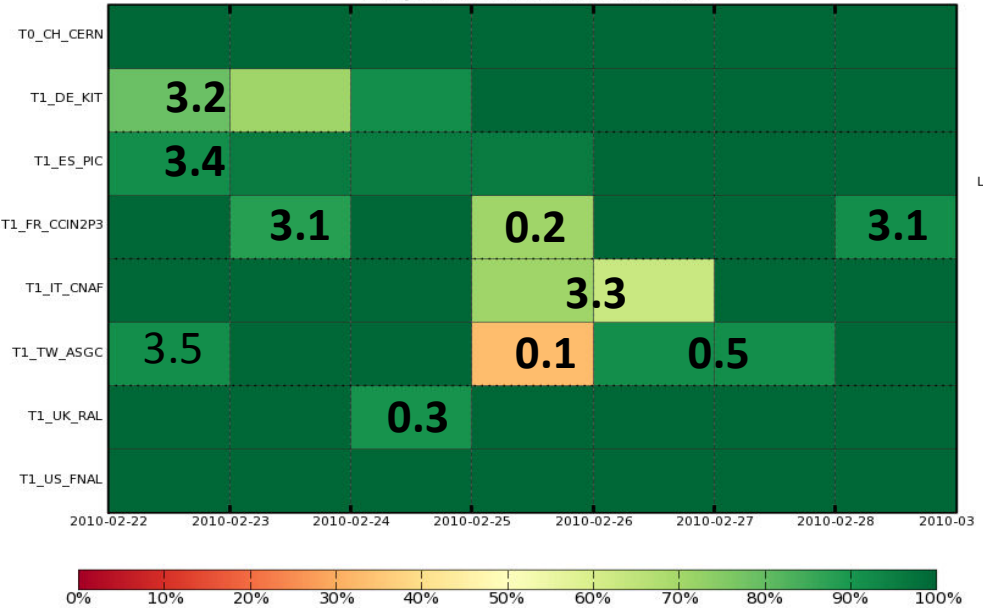

## ALICE
## ite Availability using WLCG Availability (FCR critical

7 Days from 2010-02-22 to 2010-03-01



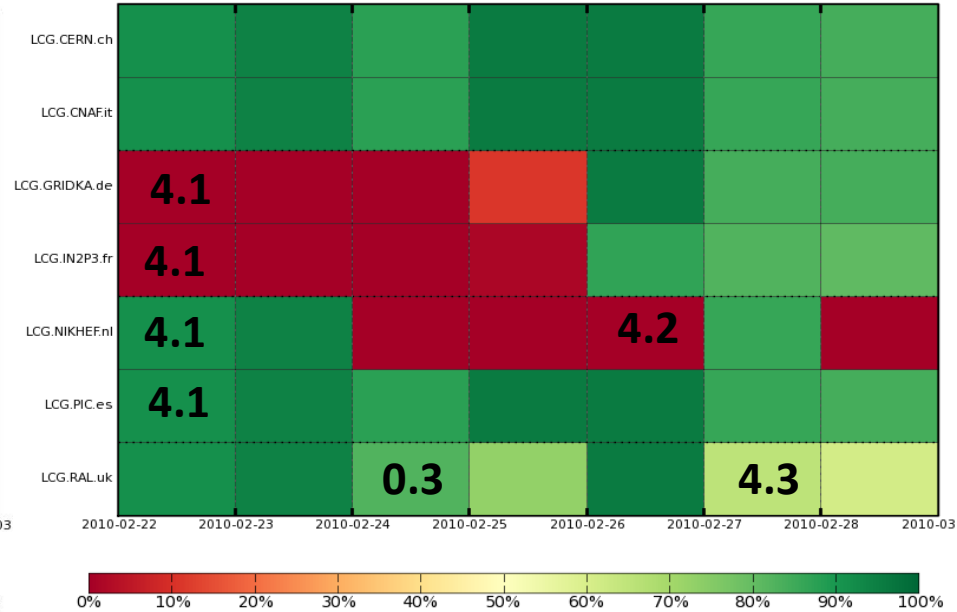

## CMS
## Site Availability

7 Days from 2010-02-22 to 2010-03-01




## LHCb
## Site Availability using LHCb Critical Availability

7 Days from 2010-02-22 to 2010-03-01



# Analysis of the availability plots

**COMMON FOR THE ALL EXPERIMENTS**

**0.1 TAIWAN:** Scheduled downtime to test the DC generator with the power source. Some problems during the downtime lead to power failure at 3:00 UTC. Squids down after power cut, restarted. About 5 hours total downtime on Thursday. SRM agent restarted, recovered

**0.2 IN2P3:** LFC is not accessible from outside. Proxy emitted by lcg-voms.cern.ch were unauthorized to access lfc. Fixed

**0.3 RAL:** Scheduled outage on Wednesday morning, went well.

**0.4 RAL:** An outage (from 19.02 to 23.02) on lcgce07 to allow for a drain and a disk replacement

**0.5 ASGC:** Problems during the WE, FTS/SRM issues

**ATLAS**

**1.1 BNL:** CE tests are not yet ready for OSG

**1.2 SARA:** Bad SRM performance was improved by tuning of the backend database server. One pool node went down and had to be restarted

**1.3 INFN:** Errors while upgrading to FTS 2.2.3 from FTS 2.1 on 23rd Feb, solved

**1.4 RAL:** MCDISK storage issues

**1.5 RAL:** Disk server problem

**1.6 SARA**: Problems in exporting data, fixed

**ALICE**

Nothing to report

**CMS**

**3.1 IN2P3:** SRM issues (same as last week), disappeared

**3.2 KIT:** Batch system problem

**3.3 CNAF:** Transient CRL errors on SRM machine – solved

**3.4 PIC:** JobRobot failures, MARADONA errors in the beginning of week

**LHCb**

**4.1** LHCb SAM tests status configuration changes for dCache sites: from 'info' to 'error', to track the "fileAccess" error more efficiently. Migrated to the new version of ROOT on Thursday - OK.

**4.2 NIKHEF:** "FileAccess" problem (some libraries missing), ggus ticked opened, under investigation

**4.3 RAL:** Problems with VOMS, ggus ticket opened, under investigation

# SIRs (1/3)

- Expiration of Spanish-CA CRLs at CERN (10th. Feb 2010)
- **Incident Start:** 10/02/2010 at 14:30 local time (UTC+1)
- **Incident End:** 10/02/2010 at 21:30 local time (UTC+1)
- **Description: 10 Feb** after 2:00PM (CET) services relying on personal or host proxies related to Spanish CA (Rediris) were failing:
- Users couldn't create proxies depending on VOMS servers located at CERN
- Data transfers coming to PIC from CERN for all LHC VOs failed
- This was only happening at CERN, since transfers from other T1s to PIC were working fine and certificates being authenticated in VOMS servers outside CERN were also working fine.
- SAM tests failed for PIC.
- By 18:30 CERN experts found they could download the Rediris CRLs from any other host *except* from the proxy server they use to concentrate all the CRLs – turned out it had been blacklisted in Rediris in January.
- Blacklisting removed by 19.30, CRLs were updated and by 21:30 PIC was again passing SAM tests.

# SIRs (2/3)

**IN2P3 Worker Node connectivity loss 15 Feb from 14:12**

All worker nodes lost network connection at nearly the same time.

**Timeline**

- 14:30 LBMS main server found to have a huge amount of pending connections
- 14:37 The automatic take-out mechanism for stalled workers starts to eliminate the first workers from the list of available machines.
- 14:45 System administrators signal a large amount of worker nodes having lost sshd
- 15:45 Downtime declaration
- 15:50 Mass reboot of worker nodes starts
- 18:00 After various verifications, about 90 percent of the worker nodes are back to production
- 18:30 End of downtime

**Analysis**

The logs of every impacted worker node showed that between 14:12 and 14:23 a signal USR1 has been sent to a bunch of processes, namely acpid, rsshd, sshd, atd, crond, ypbind, bqs, ipmievd and others.

This looks like either a direct human error or an indirect one via an ill parametrised automatic procedure.

**Impact**

All jobs running at the moment of the incident considered to be lost. No new jobs scheduled before about 16:00.

An unscheduled downtime had been declared from 3:45pm to 6:30pm for the tier-1 (IN2P3-CC) and the tier-2 (IN2P3-CC-T2).

**Corrective actions**

Reboot of all worker nodes. This is an isolated incident but with an important impact. Actions in the long term may include finer logging, especially of cluster wide tools like rssh, and deployment of a monitoring and restart system for basic system processes.

# SIRs (3/3)

- ASGC: Suffered a power outage at 03:00 UTC on 25 February.
- Two weeks earlier they had installed a parallel power supply so they did not have to experience scheduled power cuts during twice-yearly safety tests.
- The vendor started to test this parallel system and a wrong wiring on a control line, not found during earlier checks, lead to a power failure.
- Total down time was about 5 hours.
- SIR awaited.

# Miscellaneous Reports (1/4)

- 7/8 Feb: Transfers into the CMS t0export pool were failing with 'terminated by service administrator'.
  Jobs being executed by one disk server (lxfssl1201) were terminating with a SIGBUS error. This machine was still on the network and responding to some requests. This was causing a 2% failure rate on transfers in the pool which corresponds to about 1-2 per minute. Console of the machine was full of messages regarding I/O errors on all the filesystems consistent with a RAID controller failure. Follow Up:
  - Why did the 3ware alarm not produce an operator alarm?
  - How can Castor protect itself against black holes such as this ?
- 8 Feb Atlas problem at T0: Some jobs were started in machines with the wrong architecture and failed.
- This was caused because an ATLAS VOC opened an ITCM ticket to upgrade some lxbatch/atlasrtt boxes from SLC4 to SLC5 directly to the sysadmins for an intervention in a machine that runs a service provided by IT.
- Follow up**:** Clarify which machines are managed by the VOCs (done)

# Miscellaneous Reports (2/4)

- Xrootd at CERN: affected ATLAS analysis in last two weeks
  - Daemon died 15 Feb – restarted
  - Same problem 3-4 times on 18th Feb - urgent software fix on 19th:
    - All threads were locked up in an unterminated SSL handshake
    - The daemon was running but not responding to requests. This meant that the lemon monitoring did not alarm the problem.
    - Following developer analysis it appears that there was a thread blockage if the communication with the client was severed.
  - New client hang on 21 Feb lead to another urgent fix put in on 25th Feb:
    - The xrootd service for Atlas has restarted several times due to a bug with the SSL plugin. This update contains the fix for the core dump, so that the daemon will not terminate, along with a leak which has caused a previous blockage of the daemon.
  - Similar problem on 26th. Working with ATLAS to move the SSL activity to a different redirector to isolate the root cause and limit the impact.

# Miscellaneous Reports (3/4)

- APEL accounting database: See http://goc.grid.sinica.edu.tw/gocwiki/ApelIssues-Jan_Feb_2010

- Problems started 4 Feb: There is an APEL (accounting DB) problem at the moment where the repository broke. Data is arriving but cannot be examined at the portal and SAM tests are failing. The repository is being reloaded from backup but the 100GB restore will take about 30 hours.

- Service was restored on 18th Feb: the APEL central database is now back online, and data are updated on the accounting portal. The latest data appearing there are from Monday 15th Feb. Recently published data will pop up regularly now. Sync pages and SAM tests are updated as well. After our integrity checks, we can guarantee that 99.8% of the data have been restored, and we have no reason to think that the remaining 0.2% have been lost. We however advise all sites to check their data on the accounting portal and report any inconsistencies through a GGUS ticket. A broadcast has been sent to the EGEE community.

- Service Providers Summary: It took a huge amount of time to restore the system. Most of this time has been used for checking data integrity and ensuring that no data had been lost. We now need to check out our backup procedures to reassure us of their integrity and put us in the position where we can always bring a backup into service quickly.

# Miscellaneous Reports (4/4)

- FTS 2.2.3 Migration status survey by A.Sciaba:
- Upgrade complete: BNL, KIT
- CERN: Have both FTS 2.1 and FTS 2.2.3 endpoints. Will plan for removal of FTS 2.1.
- Scheduled upgrades to FTS 2.2.3: FNAL on 1 March, PIC on 9 March, NL-T1 on 10 March, CNAF during week of 8 March, NDGF in the coming weeks.
- In testing before scheduling upgrade: RAL, TRIUMF, ASGC

# Summary/Conclusions

- Experiment site availability failures are now correlating better with reported site incidents.

- Keeping WLCG services running at the required level is succeeding but is labour intensive. Some of the problems would be helped by better change management coupled with risk assessment as recommended at the recent LHCC review. Another element is to set more realistic expectations between service providers and consumers. The new regular Tier0/1 Service Coordination meeting is proving effective in driving such improvements.

- Beam operations restarted over the weekend. Anticipating 18+ months of running with 1-2 months stoppage end of 2010.