WLCG Service Report

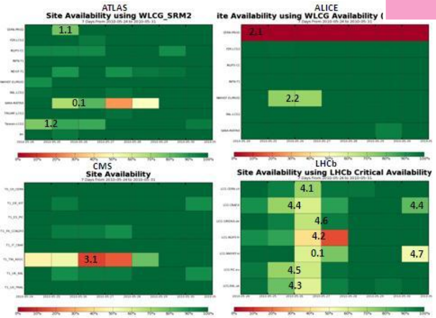
Harry.Renshall@cern.ch
Maria.Girone@cern.ch
Jamie.Shiers@cern.ch

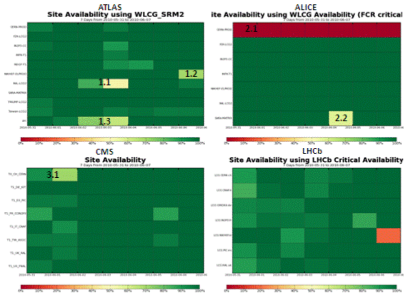
~~~

WLCG Management Board, 8th June 2010
~~~

WLCG Operations Report – Summary

KPI	Status	Comment
GGUS tickets	**End-end tests of alarm chain; 1 case (CMS CASTOR) where alarm appropriate (not used)**	**Drill-down on real alarms; comment on tests.**
Site Usability	Minor issues	Drill-down provided
SIRs & Change assessments	**Several SIRs**	Drill-down provided

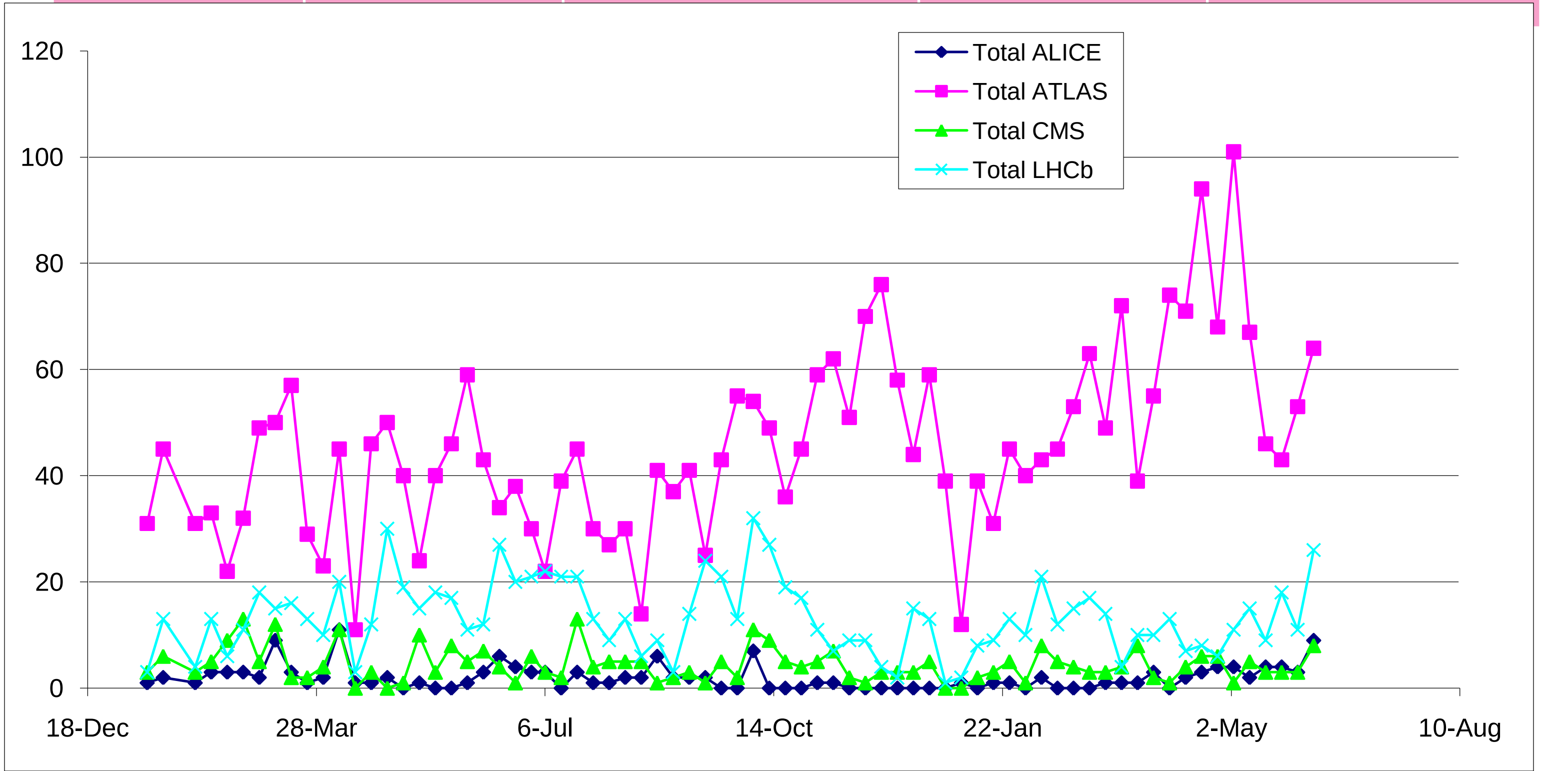
VO	User	Team	Alarm	Total
ALICE	5	0	7	12
ATLAS	40	64	11 + **2**	117
CMS	9	1	1	11
LHCb	3	27	6 + **1**	37
Totals	57	92	28	177





GGUS summary (2 weeks)

VO	User	Team	Alarm	Total
ALICE	5	0	7	12
ATLAS	40	64	13	117
CMS	9	1	1	11
LHCb	3	27	7	37
Totals	57	92	28	177



Alarm tickets

- An updated list of the Critical Services at the T0 was is now available at https://twiki.cern.ch/twiki/bin/view/LCG/WLCGCriticalServices

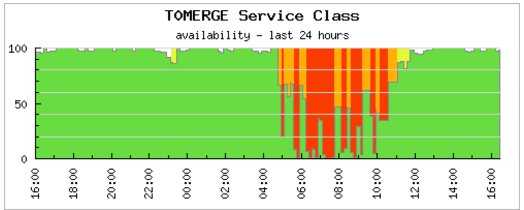
- There were 26 test ALARM tickets with the Tier0 as per the periodic action to test the full chain of reaction for Critical Services. Results were smooth ***with the exception of tickets related to the network and VOBoxes***. Actions and conclusions in https://savannah.cern.ch/support/?114705

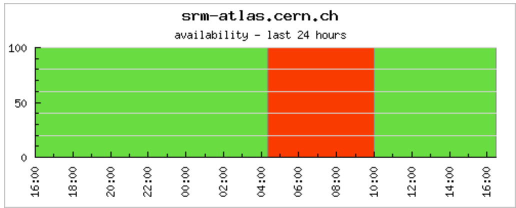
- Drills of real ALARMS (2 by ATLAS and 1 by LHCb this time) since last MB follow.

ATLAS ALARM->CERN CASTOR

What time	What happened
2010/05/25 7:03	GGUS ALARM ticket opened, automatic email notification to atlas-operator-alarm@cern.ch AND automatic assignment to ROC_CERN [SIR for this incident]
2010/05/25 7:17	Expert starts working on the problem.
2010/05/25 9:49	Expert diagnoses an intense pool activity and a stuck rsyslog. Problem 'solved'.
2010/05/25 15:32	Submitter agrees and 'verifies' the GGUS ticket.

- https://gus.fzk.de/ws/ticket_info.php?ticket=58474





ATLAS ALARM->CERN AFS

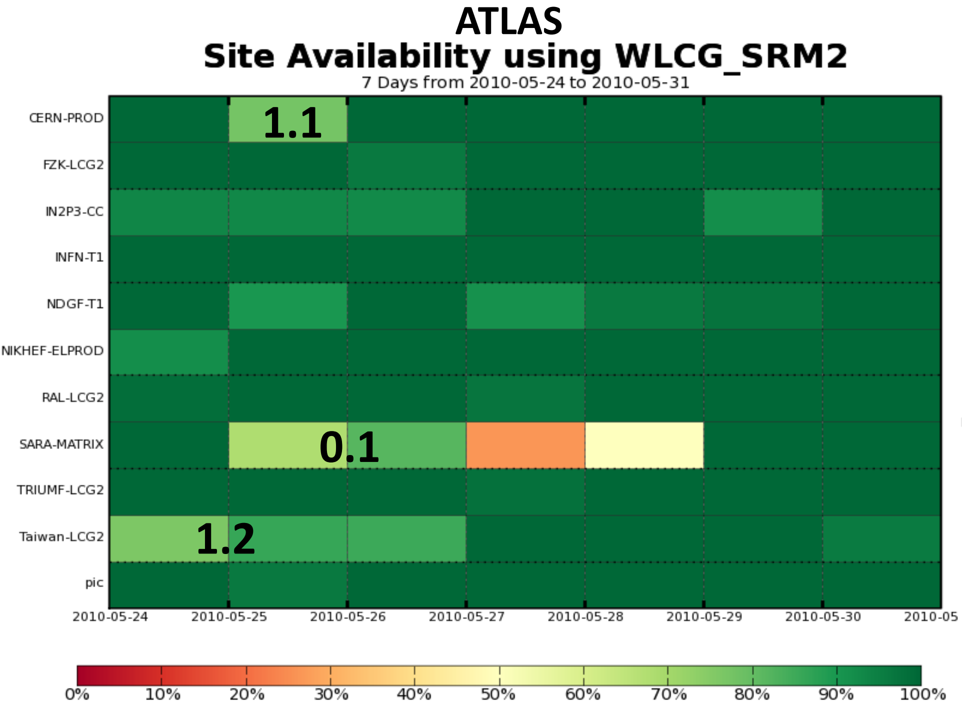
What time	What happened
2010/05/28 17:06	GGUS ALARM ticket opened, automatic email notification to atlas-operator-alarm@cern.ch AND automatic assignment to ROC_CERN ATLAS Tier-0 production is **crippled** and almost at a **standstill**, because this server hosts the volumes where job logfiles/scripts (necessary to run jobs on LSF) reside: /afs/cern.ch/atlas/project/tzero/prod1/log1, log2, log3
2010/05/28 17:44	Expert starts working on the problem. Operator asks for tel. number for offline communication to debug. (Replication mechanism had not been triggered for ATLAS s/w releases – once done + other actions to reduce load things "OK") **[Follow-up with AFS experts in IT-DSS?]**
2010/05/31 14:05	Expert diagnoses a configuration error in the afs access scripts and a server overload. Problem 'solved'.
2010/06/07 06:51	Submitter agrees and 'verifies' the GGUS ticket.

Information missing in GGUS thread, e.g. what was done on ATLAS side to reduce load!

LHCB ALARM->CERN AFS

What time	What happened
2010/05/31 8:51	GGUS ALARM ticket opened, automatic email notification to lhcb-operator-alarm@cern.ch AND automatic assignment to ROC_CERN
2010/05/31 9:25	Expert replies this was not an afs problem. It looked like a rack power failure.
2010/05/31 9:46	Submitter agrees and 'verifies' the GGUS ticket.

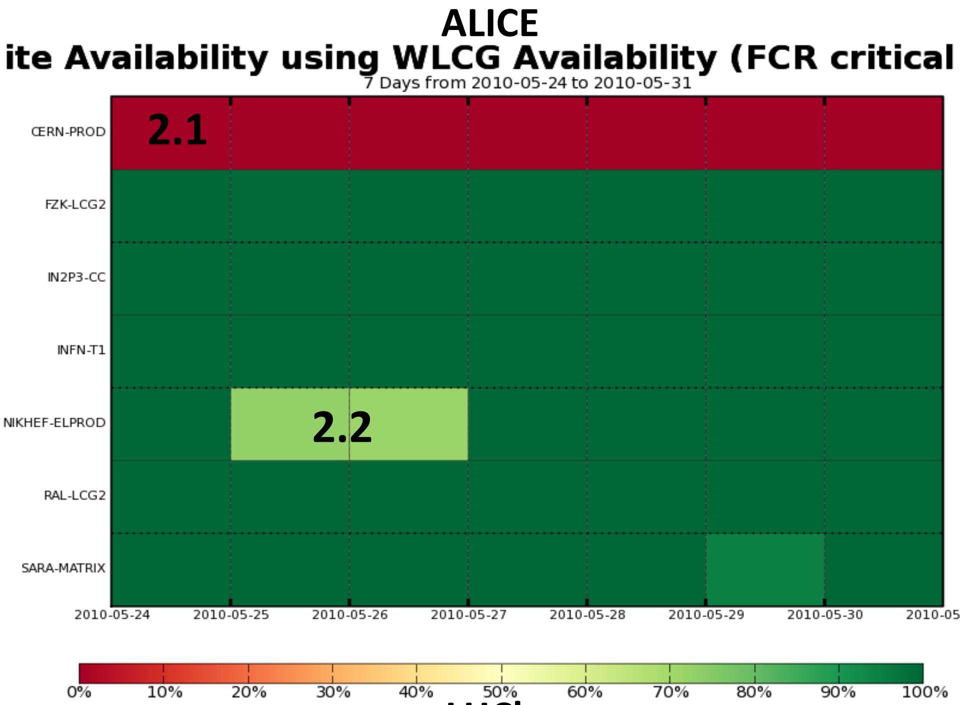
- https://gus.fzk.de/ws/ticket_info.php?ticket=58643



ATLAS

Site Availability using WLCG_SRM2

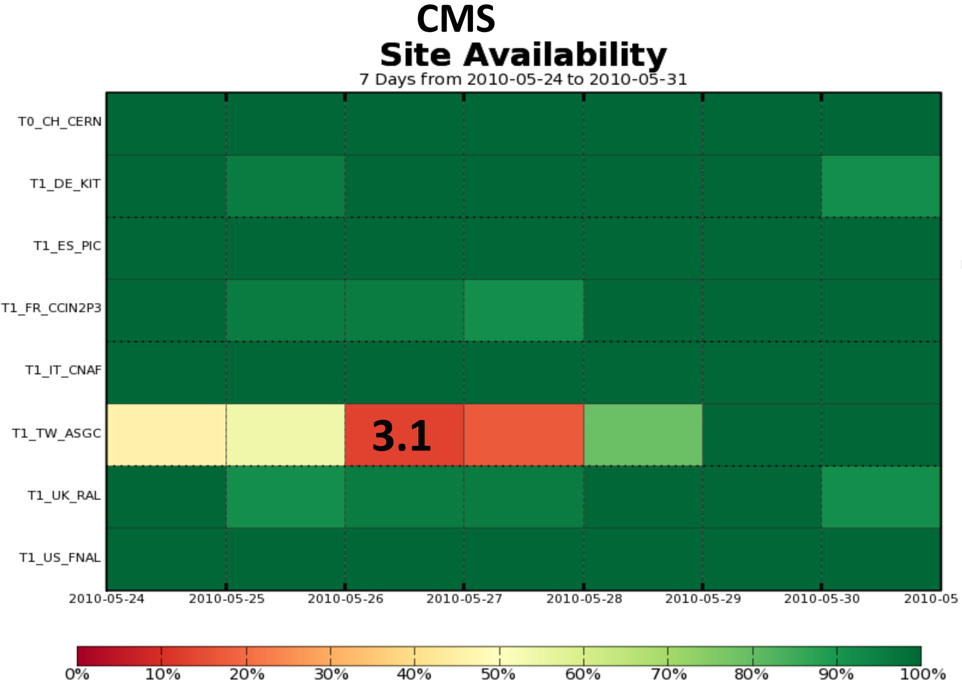
7 Days from 2010-05-24 to 2010-05-31



ALICE

ite Availability using WLCG Availability (FCR critical

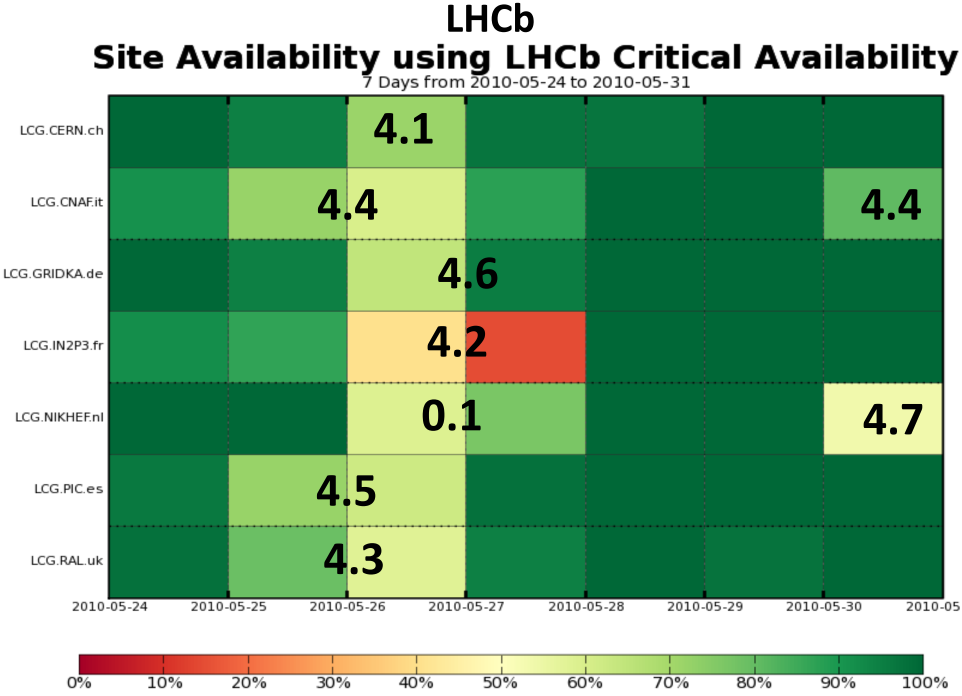
7 Days from 2010-05-24 to 2010-05-31



CMS

Site Availability

7 Days from 2010-05-24 to 2010-05-31



LHCb

Site Availability using LHCb Critical Availability

7 Days from 2010-05-24 to 2010-05-31

Analysis of the availability plots

COMMON FOR THE ALL EXPERIMENTS

0.1 NL-T1: migrating from 12 dcache pool nodes to a new 12. This required a dcache reconfiguration and restarts which caused some failures. The whole operation will take a few days and it was agreed to document the new procedure for other dcache sites (LHCb: SARA dCache is banned due to ongoing maintenance)

ATLAS

1.1 CERN: a groupdisk server has some inaccessible files and needs a file system repair, fixed

1.2 Taiwan: SRM tests failures (timeouts)

ALICE

2.1 CERN PROD: **vobox voalice11: the software area is not reachable, in progress**

2.2 NIKHEF: The local service responsible of the software installation (PackMan) was failing. The problem has been reported to the AliEn experts before warning the site. Solved

CMS

3.1 ASGC: intermittent SAM tests failures, savannah ticket opened, fixed after few days: maradona errors traced to a bad worker node (a restart had failed to mount a file system). Quickly verified - AFS repaired and problem fixed. File had become a directory.

LHCb

4.1 CERN: shared software area problems

4.2 IN2P3: shared area issues

4.3 RAL: lost a disk server. Files have been recovered

4.4 CNAF: bug in Storm

4.5 PIC: PIC-USER space token is full

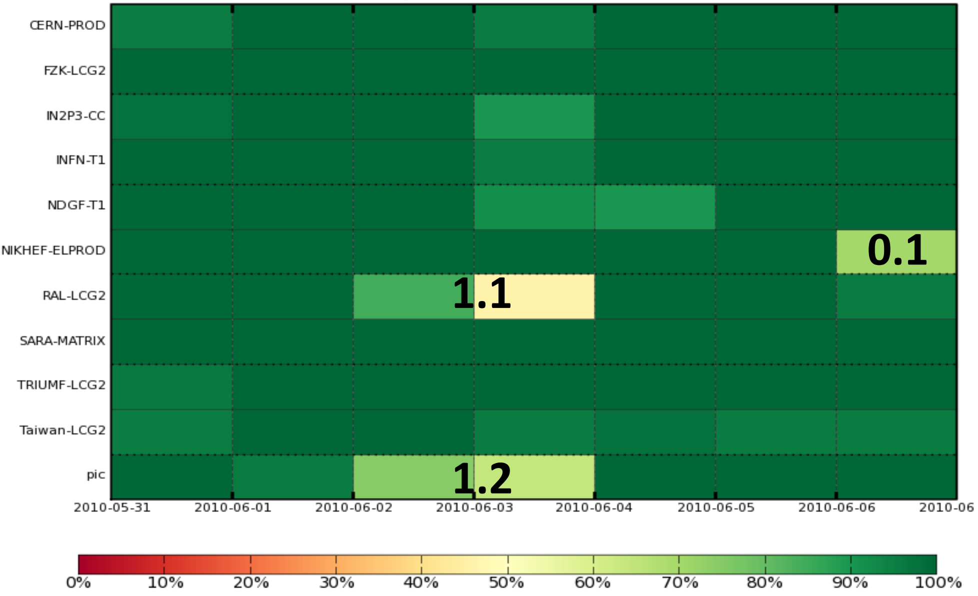
4.6 KIT: shared software area issue

4.7 NIKHEF: Grid ftp server: one CERN CA certificate expired

ATLAS

Site Availability using WLCG_SRM2

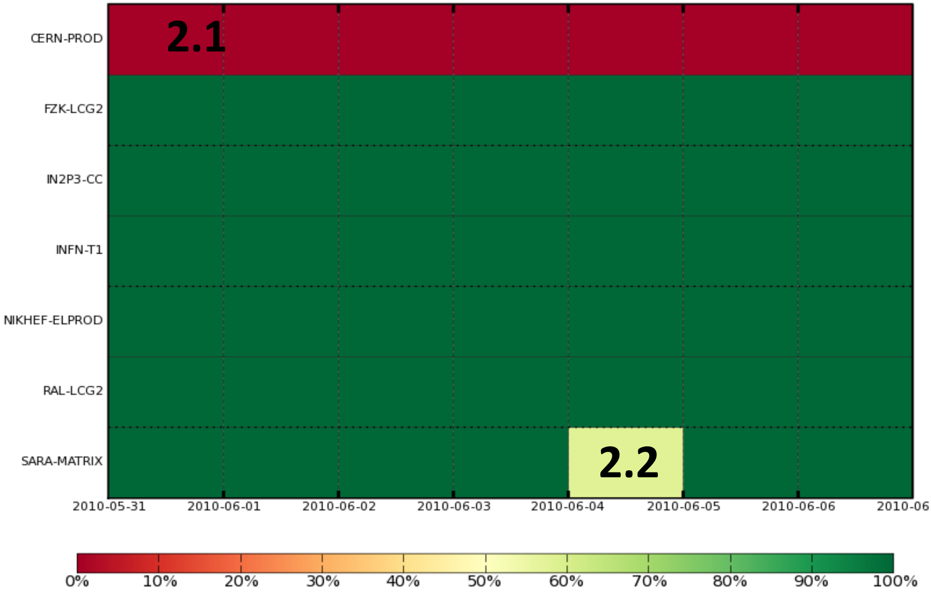
7 Days from 2010-05-31 to 2010-06-07



ALICE

ite Availability using WLCG Availability (FCR critical

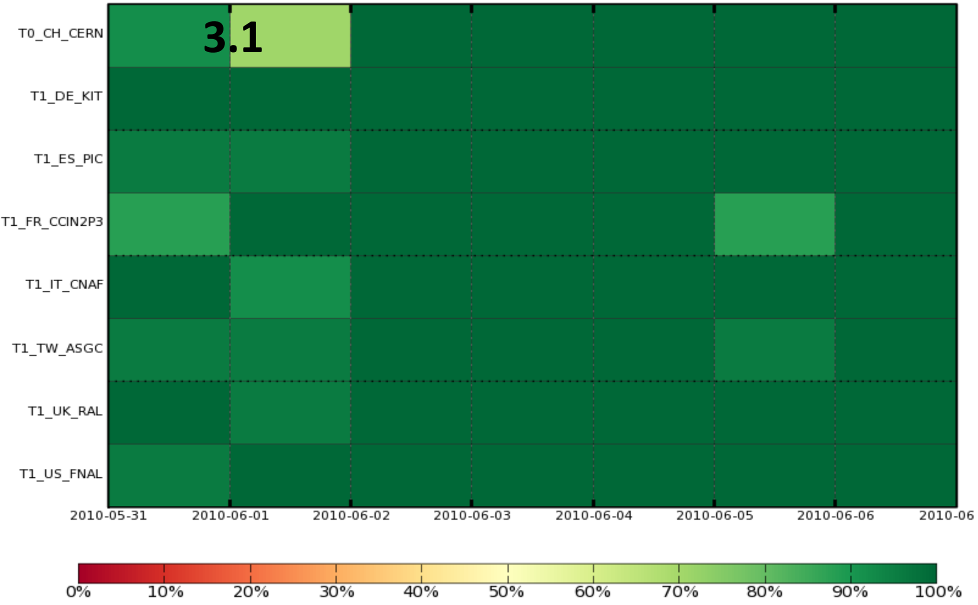
7 Days from 2010-05-31 to 2010-06-07



CMS

Site Availability

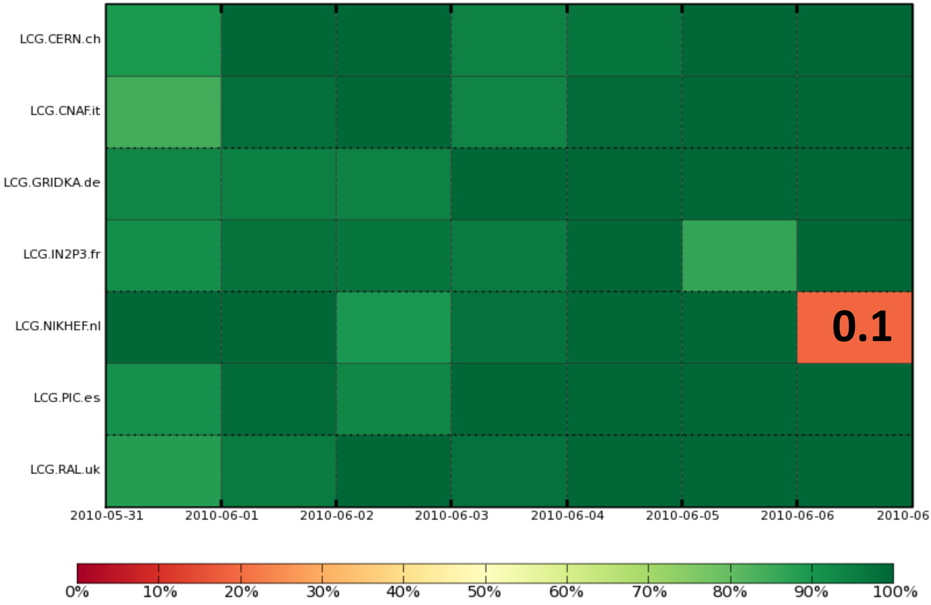
7 Days from 2010-05-31 to 2010-06-07



LHCb

Site Availability using LHCb Critical Availability

7 Days from 2010-05-31 to 2010-06-07



Analysis of the availability plots

COMMON FOR THE ALL EXPERIMENTS

0.1 NL-T1: SRM problem - dCache bug which will be reported

ATLAS

1.1 RAL: **unscheduled outage: scheduled upgrade on DBs behind LFC and FTS for April CPU patch ran into problems and have led to outage on these services**

1.2 PIC: temporary SAM tests failures

ALICE

2.1 CERN PROD: **vobox voalice11: the software area is not reachable, in progress**

Trying to solve this persistent test failure (by making s/w area reachable on this node…)

2.2 SARA-MATRIX: temporary SAM tests failures

CMS

3.1 CERN **CMS Castor instance was unable to record new files, fixed (jobs may have failed during this period with timeouts)**

LHCb

nothing to report

SIRs

- Maintained here with external references, e.g. for CASTOR, Databases and Streams
  - A summary table is included in the WLCG QRs
- 3 DB-related and 3 CASTOR-related SIRs during this period
- Pending SIRs: DE DNS problem
- Analysis and follow-up still a point that could be improved:
  - What are the lessons learned?
  - How to avoid similar problems in the future? (Also for other sites…)
  - Is their agreement on the issues? Open questions?
- Something additional to follow-up at daily meetings & MB reports needed in some cases – e.g. external reviews

SIR Summary – CASTOR

What	When	Follow-up
stuck rsyslog affected ATLAS T0Merge. T0MERGE & SRM-ATLAS were unavailable from 4:30 to 9:30.	25 May	OPEN – ALARM – ATLAS raw data recording impacted (that's what it says in the SIR). **Negatively, one assumes…**
LSF reconfiguration after node move affected CASTORPUBLIC	31 May	LSF configuration change on C2PUBLIC (standard procedure; remove diskservers) lead to LSF becoming unavailable.
Writing into CASTOR CMS blocked – **NO TEAM NOR ALARM TICKET!** 00:58 OPS call PK 07:10 SLS goes green	1 June	Problem confirmed in the jobmanager code where an inconsistency in the data for the jobmanager was causing a 'no requests' return. Developer produces online code change and service resumes. The root cause will be investigated under the Castor savannah ticket http://savannah.cern.ch/bugs/?68205.

SIR Summary – DBs

What	When	Follow-up
CMSR node broken - CMSR instance 3 crashed around 9:20 am. It was caused by a hw problem related to a memory module failure.	26 May	Issue resolution and expected follow-up: Hardware problem escalated with Dell. On Wednesday 02.06, memory was exchanged and the node was added back to CMSR cluster. (Vendor should have replaced memory in 12 working hours.)
Database issues during patching – details in notes. Affected: CMSONR, CMSR, LCGR, ATLR.	May 31st - June 2nd	Following up quattor certificates problem with quattor support. We are investigating the possible cause of the strange behavior observed during the patching. FS label misconfiguration: ticket open with sysadmins, list of affected machines provided.
PSU APR 2010 patch […] is showing not to be suitable for production on ATONR, ATLR and LHCBR production databases.	June 2 - 3	More in notes: Recommendation: Tier1s roll-back if likely to be affected. (where auditing is enabled and COOL or similar (multiple sessions connected to one server process) is used to access the database)

Summary

- A number of additional issues are covered in the minutes of the daily operations call but…

- The number of Tier0 SIRs particularly high during this period, which also included an LHC machine stop

➢ **Some systematic review of SIRs is still needed**
  - Panel (experiment + service reps) report at GDB?
  - This could then feed into QRs (Table of SIRs already provided)

- End-end alarm tests worked (mostly): useful exercise

- But we could do better.

BACKUP

T0 ALARM tests full chain

T1 Service Coordination Meeting

2010/06/03

Aim of the exercise

To test the total workflow for GGUS ALARMs tickets, namely:

- **Email notification reception by the experiment experts, members of <LHCVOname>-operator-alarm@cern.ch**
- **Email notification reception by the CERN operator on duty and existence of procedures per WLCG Critical Service.**
- **Quick and correct assignment in CERN Remedy PRMS to the right category.**
- **Acknowledgment and ticket update by the CERN IT Service manager.**

Outcome

- The exercise remained unclear for some experiment members till the end despite the documented steps-to-follow and the agreed CriticalServices twiki.
- Test ALARM tickets for the ‘network’ service reclassified by ROC_CERN (now IT/PES) to IT Services-Network-Netcom-All in PRMS still remain ‘in progress’ with NO update from the service.
- Test ALARM tickets for the ‘VOboxes’ service showed the lack of operators’ procedures.

Conclusions

- IT services participating in the WLCG daily meeting were aware of the exercise and alert to respond and close tickets as ‘solved’.
- It is worth to check that full procedures do exist for all services and for real cases.
- grid-cern-prod-admins was added in the 4 <LHCVOname>-operator-alarm e-groups for faster notification of the service managers.