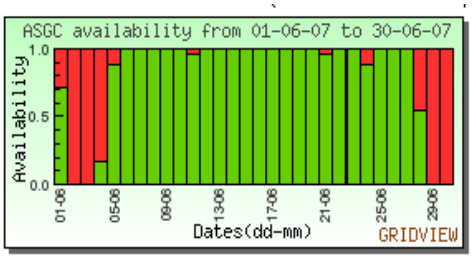
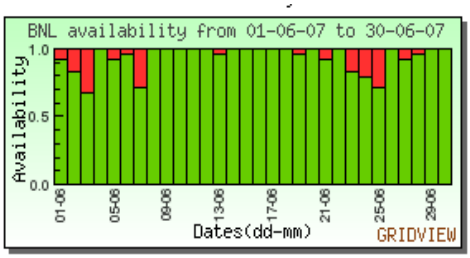
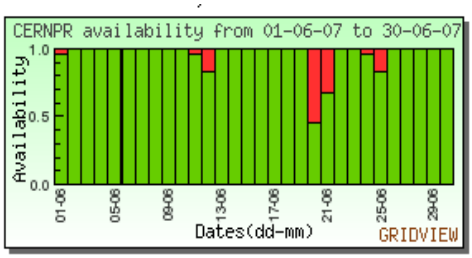
# Site Availability Reports
# June 2007

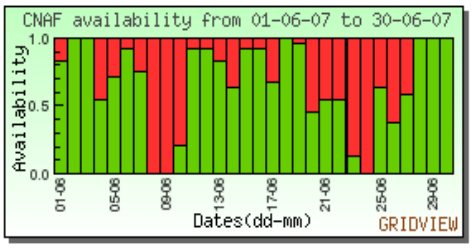
- *Please review and complete the Site Reports below.*
  *Deadline Friday6 July 2007.*
  *Edit your section and mail the document back to A.Aimar.*
- *There examples of good reports (BNL, CERN, RAL) and report completely missing at the OPS meeting (ASGC, CNAF, IN2P3) for the whole month.*
- *Please pass this information to your representative at the Operations meeting so that we do not need to complete them at the end of each month.*

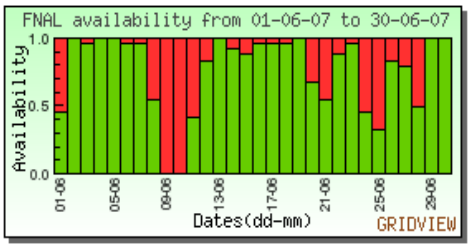
Sites availability, for all details see here:
http://lcg.web.cern.ch/LCG/MB/availability/site_reliability.pdf

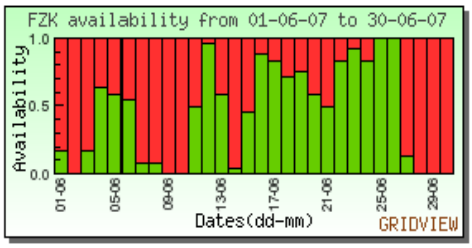


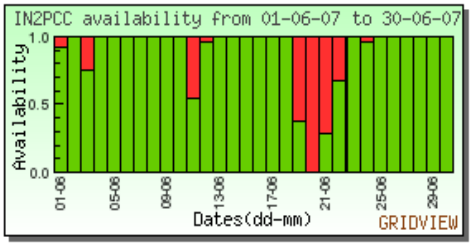



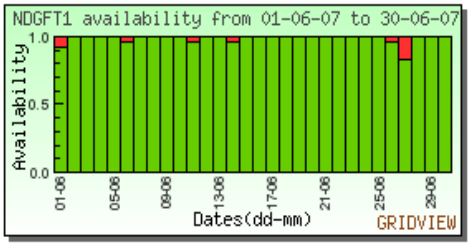



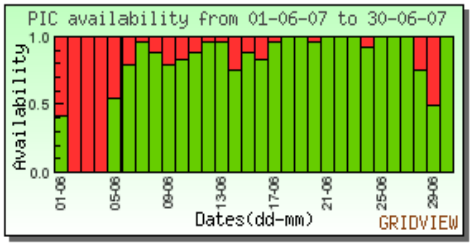



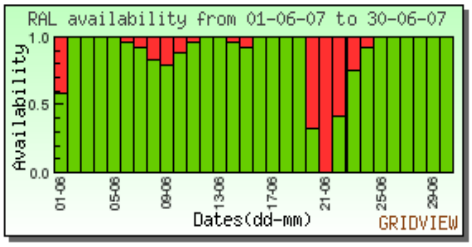



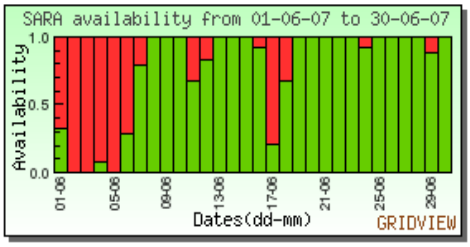



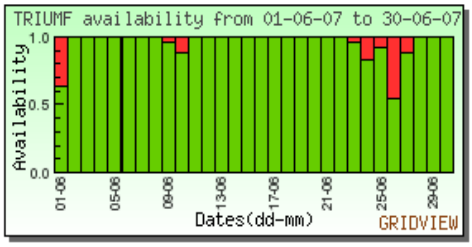



# ASGC

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

# BNL

## > Friday: June 01/2007

We continue to observe sluggish Tier 0 data export from CERN to several Tier 1 sites. Our PNFS should high load generated by USATLAS production and Tier 0 data exports. We saw many time-out in data transfer.

## > Saturday, June/02/2007

Problem: SRM server in BNL was off-line between 4:00PM and 7:00PM.
Cause: The SRM server, GridFtp door server, and write pool server certificates expired around 4:00PM.
Severity: the data transfer completely stopped during 4:00PM and 7:00PM. Tier 0 data export to BNL, USATLAS Production and AOD/NTUPLE data replication were impacted.
Solution: We renewed the certificates around 6:00PM.
Problem: USATLAS production has problems to write data into BNL dCache system at 8:48PM. No data can be written into the subdirectories of the dCache root directory: /pnfs/usatlas.bnl.gov/AOD01, RDO01, etc.
Cause:
We changed the ownership of directories to the production account ???usatlas1???. But did not attach storage tag, which means that there were no write pool resources assigned to these sub-directories. The production could not create subdirectories and write files.
Severity: The USATLAS production was affected for two hours.
Solution: attach the storage tag into subdirectories, as we agreed on the morning meetings.
Problem: The problem happened on Wednesday continued. USATLAS production manager reported the performance of the data export from BNL to some Tier 2 sites was sluggish.
Cause: Production requests data files that are only in HPSS, but not in disk areas. It takes long time to stage-in files into disks and transfer them to the remote sites.
Severity: many Tier 2 sites are running out of data files and waiting for data transfer since Tuesday morning. The resource utilization for USATLAS decreased.
Solution:
We reallocated 10 HPSS drives to speed up the data stage-in. The USATLAS production manager gave us a list of files. Our data management expert created a script to stage-in the input files from HPSS to disks.

## > Sunday: June/03/2007.

Problem:

The Panda monitor (USATLAS production monitoring server) crashed. The kernel crashed from a stack overflow (do_IRQ: stack overflow: 488), probably because it was not able to keep up with the increasing memory pressure.
Cause: from the message that is still visible on the console and the system logs, it looks like that apache was using too much memory, triggering the kernel\\\''\''s out-of-memory killer several times in the hour prior to its crashing:
Jun 3 00:12:33 gridui02 kernel: Out of Memory: Killed process 27405 (httpd).
Jun 3 00:29:04 gridui02 kernel: Out of Memory: Killed process 28348
there could be a memory leak in the panda monitor or the version of apache it is using.
Severity:
USATLAS production dashboard is off air for twelve hours. Production runs blindfolded.
Solution: Reboot the server, and add a memory warning watermark in BNL Nagios monitoring page.

Problem: USATLAS production manager reported the performance of the data export from BNL to some Tier 2 sites was sluggish.
Cause: Production requests data files that are only in HPSS, but not in disk areas. It takes long time to stage-in files into disks and transfer them to the remote sites.
Severity: many Tier 2 sites are running out of data files and waiting for data transfer since Tuesday morning. The resource utilization for USATLAS decreased.
Solution:No Solution was given yet.
problem: data export from Tier 0 and USATLAS production data transfer suddenly failed around 1:00PM.

Cause:
The dCache map file is updated around 1:00PM. Both update scripts and SRM read/write from/to the same dCache grid user map file. SRM reads only a fraction of the file around 1:00PM, and two accounts are missing from the file: USATLAS1 and USATLAS4
Severity: Both USATLAS production and Tier 0 data export have been off-line for twenty minutes.
Solution:
Our dCache administrators redirected the output of the dCache grid map file updating script to a temporary file. After the update is finished and validated, then the script swaps the new grid map file with the existing dCache grid map file.

## > Monday: June/11/2007

Problem:
One of GridFtp servers node (dcdoor02) crashed on 6:00AM Monday morning.
Cause: an IRQ kernel error.
Severity: the GridFtp server was off-line for four hours and 14% connectivity was lost due to this problem.
Solution:
Our system administrator updated the server since it hadn't been updated in many months. Then, after the update, we
rebooted it a second time so it would be running the updated kernel.
Problem:
Two OSG gatekeepers were reported critical at 04:54:56 EDT 2007

Cause:
GUMS server went off-line (but log output had stopped. we restarted Tomcat and it appears to be functioning).
The new host certificate generated for the GUMS server was missing an attribute needed for a server. The old one had
X509v3 extensions:
Netscape Cert Type:

```
                SSL Client, SSL Server
while the new one had:
  X509v3 extensions:
             Netscape Cert Type:
                 SSL Client, S/MIME

Severity:
Two OSG gatekeepers were impacted for three hours before our administrator
intervention.
Solution:
We disabled the GUMS on two OSG gatekeepers and used the static Grid map files to
allow BNL CE to be accessible right away.
 In the mean time, we obtained a new host certificate with a proper server
attribute for our GUMS server.
```

### > Tuesday: June/12/2007

```
Maintenance:
Kernel update was performed on all seven GridFtp server nodes one by one.
Each GridFtp server has less than an hour downtime.   The update is transparent
to users.  No need to make announcement.
```

### > Wednesday: June/13/2007

```
Problem:
A fraction of data transfers from BNL to other ATLAS Tier 1 sites failed with
certificate mismatch errors.
Cause:
A fraction of our dCache read pool nodes have bad certificates that their DNs do
not match with their hostnames.
The error message is shown as follows:
-------------------------------------
06/07 12:01:04 Cell(SRM-dcsrm@srm-dcsrmDomain) : Authentication failed.
Caused by GSSException: Operation unauthorized (Mechanism level:
[JGLOBUS-56] Authorization failed. Expected
"/CN=host/acas0203.usatlas.bnl.gov" target but received
"/DC=org/DC=doegrids/OU=Services/CN=acas0399.usatlas.bnl.gov")
Severity:  This problem affects the data transfer directly between the remote
party and these affected dCache read pool nodes
 while the data transfer via (GridFtp servers) was not affected.   BNL to other
Tier 1 data transfer does not use GridFtp
server nodes; all data transfer from BNL affected nodes to Tier 1 sites
experienced transfer failures.
 USATLAS production is NOT affected by this problem.
Solution:
We replaced these bad certificates on Wednesday.  We notified the ATLAS data
operation team to confirm whether the lower
performance problem with the data transfers from other Tier 1 sites to BNL.
We will add scripts to validate the host certificates.
```

---

# CERN-PROD

### > Remark(s) on 2007-06-01

```
01/06: OK, only 1 h unavailability
```

### > - 12/06/2007, 4 h (7-10 am):

```
problem: replica management (CE-sft-lcg-rm ) test failure
cause: CASTORPublic Unavailable
```

broadcast sent in advance: The public CASTOR2 service at CERN will be unavailable from 07.00 to 16.0 UTC (09.00 to 18.00 CET) on Tuesday 12 June TODAY while a major upgrade is performed.
As it is seen from the SAM monitoring, the intervention only caused a 4 h interruption.
There was an scheduled downtime on the SE for this interrruption:
Upgrade of the public CASTOR2 stager at CERN 12nd June 2007 - 09:00 12nd June 2007 - 14:30
But as the failure was on the CE (replica management test) it was not taken into account by SAM.

### > Remark(s) on 2007-06-20

Problem: an attempt to fix the configuration for srm-durable-<VO>.cern.ch reported by LHCb was attempted at 11am. Unfortunately the fix introduced another misconfiguration for the non-durable SE srm.cern.ch.
Solution: the configuration for srm.cern.ch was fixed in the evening.
Problem: SAM tests not running since 18:00 and site availability continued to be flagged red for the whole night even if the SE problems were fixed in the evening
Solution: SAM team contacted in the morning and they admitted the problem and restarted the service

### > Remark(s) on 2007-06-21

Problem: SAM tests not running since 18:00 the day before and site availability continued to be flagged red for the whole night even if the SE problems were fixed in the evening
Solution: SAM team contacted in the morning and they admitted the problem and restarted the service

### > Remark[s] on 2007-06-26

Problem: wrong SE records published, breaking lcg-utils, during 3 hours in the afternoon.
Solution: roll back to previous version of Castor SRM information provider

---

## CNAF

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

---

## FNAL

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

### > Remark(s) on 2007-06-01

unscheduled cooling outage

### > Remark(s) on 2007-06-03

fully operational, test defect

### > Remark(s) on 2007-06-06 to 07

fully operational, test defect

### > Remark(s) on 2007-06-08 to 11

Failure of DNS at CERN

### > Remark(s) on 2007-06-12 to 24

USCMS was fully operational, test defect

### > Remark[s] on 2007-06-25

USCMS was operational defective tests for single bars

Disk failure on CRL squid cache led to authentication problems.
Squid cache now a critical service so we will be alerted of problem.

### > Remark[s] on 2007-06-26

Disk failure on CRL squid cache led to authentication problems.
Squid cache now a critical service so we will be alerted of problem.

### > Remark[s] on 2007-06-27

SRM database failure, dropped all SRM tables to recover

### > Remark[s] on 2007-06-28

USCMS was operational, defective tests

### > Remark[s] on 2007-06-29

Scheduled Downtime, should not be red.

---

## FZK-LCG2

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

### > Remark(s) on 2007-06-21

Almost all difficulties this (and last week) stem from stability problems on the CE\''s. More specific, the info provider system (gris) sometimes returns erroneous data (i.e. no data). Consequently the job requests fail. We are investigating and have setup more extensive monitoring of all relevant activity on the CE. However for the time being the situation remains unsatisfiable.

### > Remark[s] on 2007-06-23

instabilities of CE. (infosystem) Under investigation

### > Remark[s] on 2007-06-24

instabilities of CE. (infosystem) Under investigation

### > Remark[s] on 2007-06-25

instabilities of CE. (infosystem) Under investigation

### > Remark[s] on 2007-06-28

Scheduled downtime

### > Remark[s] on 2007-06-29

Scheduled downtime

---

## IN2PCC

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

---

## NGDF

Availability report : NDGF-T1

### > Remark(s) on 2007-06-01

Problem: Sam test issue, most likely
Soltion: ignore it and it fixed itself

### > Remark(s) on 2007-06-06

Problem: SRM door on srm.ndgf.org had hung.
Solution: Restarted shortly afterwards and worked since.

---

## PIC

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

### > Remark[s] on 2007-06-25

Two CE-sft-lcg-rm spurious errors at 00h and 08h due to one misconfigured WN pointing to a default_SE recently migrated, with the Information Provider still to be configured. The bad WN was correctly configured on Monday 25 June at 11h.

### > Remark[s] on 2007-06-29

Problem with maui''''''''''''''''''s reservations. The ops jobs finished in an incorrect WN. Problem solved.

---

# RAL-LCG2

Availability report : RAL-LCG2

### > Remark(s) on 2007-06-01

Problem: high load on CE
Solution: problem eventually resolved itself after rebooting

### > Remark(s) on 2007-06-06

Problem: top-level BDII timed out
Solution: none (load on BDII not generally a problem)

### > Remark(s) on 2007-06-07

Problem: Scheduled maintenance + during reconfiguration a change was introduced that meant the OPS tests were mapped to accounts which could not run jobs successfully.
Solution: The reconfiguration was backed out, only OPS jobs would have been affected

### > Remark(s) on 2007-06-08

Time: 1300 GMT
Problem: Network Problem caused by switches going into odd state
Solution: Switches were reset.
Time: 2000 GMT until 0000 GMT
Problem: After reconfiguration on previous day, a local change to the job manager was overwritten causing the CE to remove
jobs it found in a Waiting State, many jobs at RAL-LCG2 enter this state but then go on to run successfully, so the local
change to the job manager to ignore waiting jobs was reintroduced on Monday 11th.

### > Remark(s) on 2007-06-09

Time: 0100 GMT until 0900 GMT
Problem: After reconfiguration on previous day, a local change to the job manager was overwritten causing the CE to remove
jobs it found in a Waiting State, many jobs at RAL-LCG2 enter this state but then go on to run successfully, so the local
change to the job manager to ignore waiting jobs was reintroduced on Monday 11th.

### > Remark(s) on 2007-06-10

Time: 0100 GMT until 2100 GMT
Problem: After reconfiguration on previous day, a local change to the job manager was overwritten causing the CE to remove
jobs it found in a Waiting State, many jobs at RAL-LCG2 enter this state but then go on to run successfully, so the local
change to the job manager to ignore waiting jobs was reintroduced on Monday 11th.

### > Remark(s) on 2007-06-11

Time: 0300 GMT
Problem: After reconfiguration on previous day, a local change to the job manager was overwritten causing the CE to remove
jobs it found in a Waiting State, many jobs at RAL-LCG2 enter this state but then go on to run successfully, so the local
change to the job manager to ignore waiting jobs was reintroduced on Monday 11th.

### > Remark(s) on 2007-06-12

n/a

### > Remark(s) on 2007-06-13

n/a

### > Remark(s) on 2007-06-14

Problem: OPS tests were not being scheduled with sufficient priority to run soon after submission
Solution: Ops tests were previously running using dteam pool accounts and queues but have now been switched to a separate
set of pool accounts and queue and the priority for this group has been increased.

### > Remark(s) on 2007-06-20

Problem: OPN link to CERN went down, causing CE replication and SE and SRM tests to fail
Solution: Link was repaired 22nd June approximately 14:00(BST)

### > Remark(s) on 2007-06-21

Problem: OPN link to CERN went down, causing CE replication and SE and SRM tests to fail.
Solution: Link was repaired 22nd June approximately 14:00(BST)

### > Remark[s] on 2007-06-23

Time : Midnight-13:00
Problem : OPN connection to CERN was down
Solution : OPN connectivity was restored at 13:00

### > Remark[s] on 2007-06-24

Time : 09:00 - 13:00 & 15:00 - 17:00
Problem : Timeouts in contacting the top-level BDII
Solution : BDIIs will be updated with new release to reduce load

---

## SARA - MATRIX

### > Remark(s) on 2007-06-02 to 07

Problem: problems with sgm mappings and change in SAM test to run as ops sgm
Solution: fixed now.

### > Remark(s) on 2007-06-11

Problem: Problems with site bdii due to a misconfiguration.
Solution: misconfiguration has been corrected

### > Remark(s) on 2007-06-12

Problem: Srm problems.
Solution: Service which was very slow has been restarted.
Problem: Problems with the oracle LFC which was very slow.
Solution: Restarted the service with an increased number of threads which fixed the problem.

### > Remark(s) on 2007-06-17

Problem: We have had problems with dcache pools running out of disk space on their root file systems. This problem was caused by idle gridftp doors generating lots messages stating that it has nothing to do. This generated huge log files.
Solution: Removed the gridftp logs and restarted the gridftp door and pools. In addition we have tightened the logrotate rules so that the dcache log files are not only rotated and compressed each day but also when they exceed a 2 GB limit.

### > Remark(s) on 2007-06-18

See above.
Availability report : SARA-LISA

### > Remark(s) on 2007-06-20

problem: fileserver/storage crash
solution: reboot of cluster

### > Remark(s) on 2007-06-21

problem: nat to internet from WNs broke
solution: network department fixed the issue
Availability report : SARA-MATRIX

### > Remark[s] on 2007-06-25

Problem: Could not replicate file to CERN
Solution: Problem dissappeared by itself.

---

## TRIUMF-LCG2

- *Please report on the site unavailability period(s), see page 1*

Day:
Reason:
Severity:
Solution:

### > Remark(s) on 2007-06-01

GC CA crl was not updated due the CA operator error.

### > Remark(s) on 2007-06-09

```
SRM directory creation failed. Unknown cause.
```

### > Remark(s) on 2007-06-10

```
CE, SE, BDII
delete failed
2000 jobs in queue due to ATLAS job piority problem - possible load
issue.
```

---

# SAM unavailability

- 11 June, 16:00-17:00 Piotr testing firewall script
- 20 June 20:00 - 21 June 10:00 tomcat down
  - reason: memory leak in the software
  - workaround: cron job
  - resolved: when Piotr is back
- 20 June: rb118 WMS down during night (rb108 was working)
- 15 June - 20 June: GOCDB syncronization was stopped (mistake) => Downtimes problem
- 25 June: ~14:00-15:30 due to a failure on the SAM SE the tests were using srm.cern.ch, which also showed failures. This way some sites got fake Replica Management test alarms.
- 26 June: ~8:00-~10:00 tomcat down, no test results could be published
- 28 June: ~16:00-~18:00 GOCDB3 migration => bug in the syncronization script => test submission disabled for the time of debugging