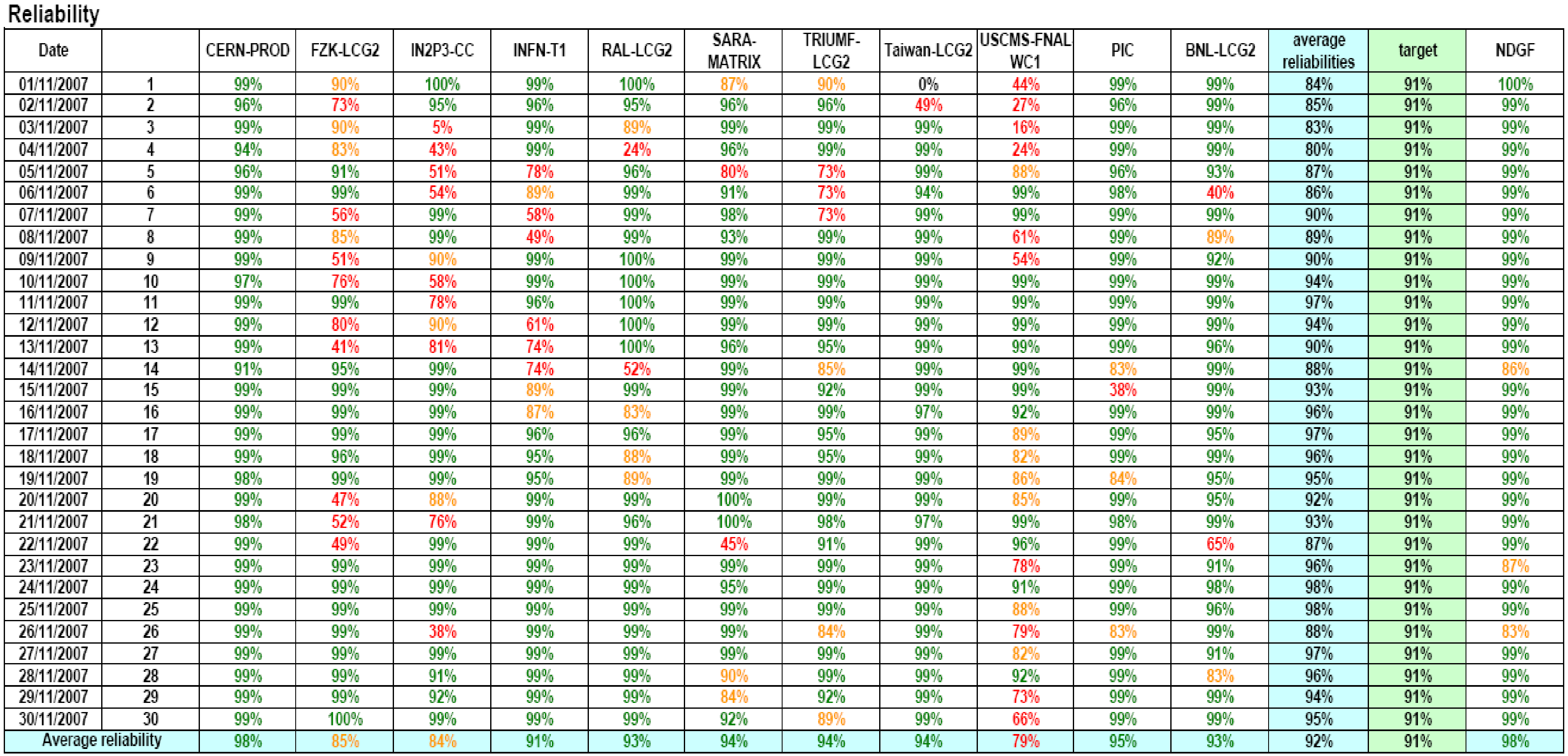
WLCG Site Reliability Reports
November 2007

- *Please review and complete the Site Reports below. Edit your section and mail the document back to A.Aimar.*
- *Deadline: Monday 10 December 2007*

http://lcg.web.cern.ch/LCG/MB/availability/site_reliability.pdf

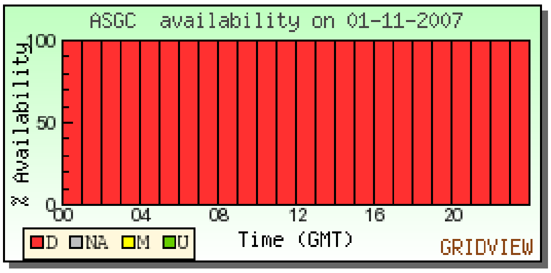


Reliability

Date		CERN-PROD	FZK-LCG2	IN2P3-CC	INFN-T1	RAL-LCG2	SARA-MATRIX	TRIUMF-LCG2	Taiwan-LCG2	USCMS-FNAL-WC1	PIC	BNL-LCG2	average reliabilities	target	NDGF
01/11/2007	1	99%	90%	100%	99%	100%	87%	90%	0%	44%	99%	99%	84%	91%	100%
02/11/2007	2	96%	73%	95%	96%	95%	96%	96%	49%	27%	96%	99%	85%	91%	99%
03/11/2007	3	99%	90%	5%	99%	89%	99%	99%	99%	16%	99%	99%	83%	91%	99%
04/11/2007	4	94%	83%	43%	99%	24%	96%	99%	99%	24%	99%	99%	80%	91%	99%
05/11/2007	5	96%	91%	51%	78%	96%	80%	73%	99%	88%	96%	93%	87%	91%	99%
06/11/2007	6	99%	99%	54%	89%	99%	91%	73%	94%	99%	98%	40%	86%	91%	99%
07/11/2007	7	99%	56%	99%	58%	99%	98%	73%	99%	99%	99%	99%	90%	91%	99%
08/11/2007	8	99%	85%	99%	49%	99%	93%	99%	99%	61%	99%	89%	89%	91%	99%
09/11/2007	9	99%	51%	90%	99%	100%	99%	99%	99%	54%	99%	92%	90%	91%	99%
10/11/2007	10	97%	76%	58%	99%	100%	99%	99%	99%	99%	99%	99%	94%	91%	99%
11/11/2007	11	99%	99%	78%	96%	100%	99%	99%	99%	99%	99%	99%	97%	91%	99%
12/11/2007	12	99%	80%	90%	61%	100%	99%	99%	99%	99%	99%	99%	94%	91%	99%
13/11/2007	13	99%	41%	81%	74%	100%	96%	95%	99%	99%	99%	96%	90%	91%	99%
14/11/2007	14	91%	95%	99%	74%	52%	99%	85%	99%	99%	83%	99%	88%	91%	86%
15/11/2007	15	99%	99%	99%	89%	99%	99%	92%	99%	99%	38%	99%	93%	91%	99%
16/11/2007	16	99%	99%	99%	87%	83%	99%	99%	97%	92%	99%	99%	96%	91%	99%
17/11/2007	17	99%	99%	99%	96%	96%	99%	95%	99%	89%	99%	95%	97%	91%	99%
18/11/2007	18	99%	96%	99%	95%	88%	99%	95%	99%	82%	99%	99%	96%	91%	99%
19/11/2007	19	98%	99%	99%	95%	89%	99%	99%	99%	86%	84%	95%	95%	91%	99%
20/11/2007	20	99%	47%	88%	99%	99%	100%	99%	99%	85%	99%	95%	92%	91%	99%
21/11/2007	21	98%	52%	76%	99%	96%	100%	98%	97%	99%	98%	99%	93%	91%	99%
22/11/2007	22	99%	49%	99%	99%	99%	45%	91%	99%	96%	99%	65%	87%	91%	99%
23/11/2007	23	99%	99%	99%	99%	99%	99%	99%	99%	78%	99%	91%	96%	91%	87%
24/11/2007	24	99%	99%	99%	99%	99%	95%	99%	99%	91%	99%	98%	98%	91%	99%
25/11/2007	25	99%	99%	99%	99%	99%	99%	99%	99%	88%	99%	96%	98%	91%	99%
26/11/2007	26	99%	99%	38%	99%	99%	99%	84%	99%	79%	83%	99%	88%	91%	83%
27/11/2007	27	99%	99%	99%	99%	99%	99%	99%	99%	82%	99%	91%	97%	91%	99%
28/11/2007	28	99%	99%	91%	99%	99%	90%	99%	99%	92%	99%	83%	96%	91%	99%
29/11/2007	29	99%	99%	92%	99%	99%	84%	92%	99%	73%	99%	99%	94%	91%	99%
30/11/2007	30	99%	100%	99%	99%	99%	92%	89%	99%	66%	99%	99%	95%	91%	99%
Average reliability		98%	85%	84%	91%	93%	94%	94%	94%	79%	95%	93%	92%	91%	98%

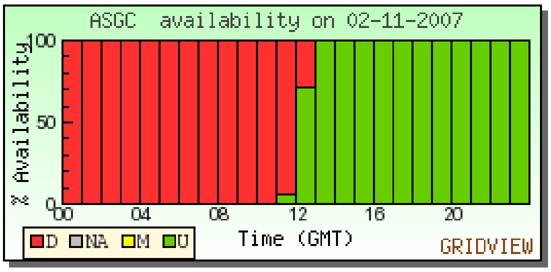
TW-ASGC

⇒ 01 Nov 2007



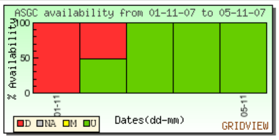
Title: job submission failure on w-ce01
Date: start from 01-Oct-2007 07:17:34, and expect to end at '01-Oct-2007 12:44:03.
Reason: Cannot read Job Wrapper output, both from Condor and from Maradona.
Severity: the impact of the failure affect all monitoring jobs if keep sending to the same problem WN.
Solution: due to the missing NFS mount pt of exp s/w scratch, all gLite Environment settings are not able to load if accepting monitoring jobs as pool account, say opssgm, that the job wont be able to update the execution status or downloading also the input sandbox. have remount the NFS mount pt, and rescanning all backend WN to make sure NFS mount pt are attached correctly.

⇒ 2 Nov 2007

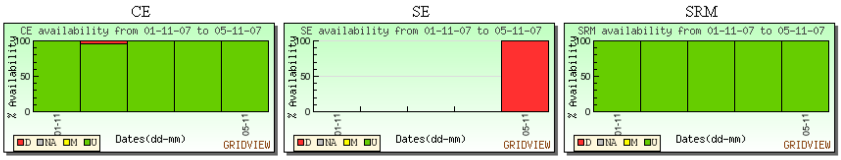


Title: SAM job submission failure
Date: 02-Oct-2007 03:11:05
Reason: Got a job held event, reason: Unspecified gridmanager error
Severity: the impact are severe, seems not only SAM monitoring jobs submit by ops VO. but also the other production jobs sent by at/cm are all fail at the problem WN, if scheduler dispatches the job on it. Hereafter also the statistic of mom cpyfile failure extract from system log. the root cause remains unclear, and cluster.out in WMS working dir, under globus tmp have occupied more than 60gb disk space, while the other production jobs are not having the same behavior.
Solution: force removing the globus tmp dir help solving the pb.

Overall Service Availability for Site:ASGC VO:OPS (Daily Report)



Individual Service Availability for site:ASGC VO:OPS (Daily Report)

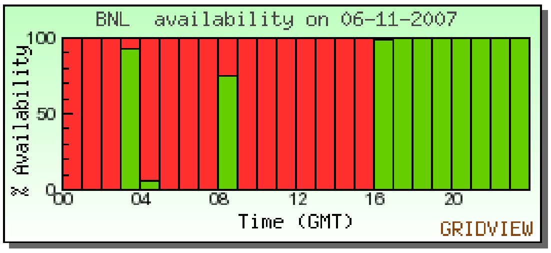


Except for this event, seems SAM calculation have confused with site overall site available metric. As shown above, we have extract 5 day time span site availability from GridView, and confirm that SAM SRM testing does pass on one of site SE (this is also the same SE adopt for SAM CE testing, say dpm01.grid.sinica.edu.tw, and the overall metric missing at Nov. 1st, and also half day of Nov 2nd.

All these two events shown above shall have been fixed shortly, and shall all of them referring to one of the CE only, rather than all CEs fail with SAM CE testing. We check our event log and couldn't see any particular event related.

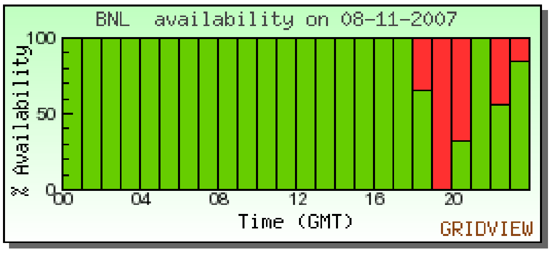
US-T1-BNL

⇒ 06 Nov 2007



SRM crashed
Cause: too many requests from FTS that couldn't be completed (permission problems)
Severity: production affected
Resolution: fix the upgraded DQ2 code

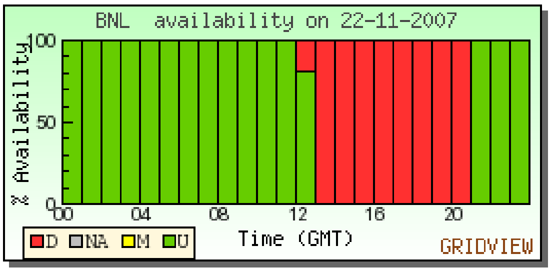
⇒ 08 Nov 2007



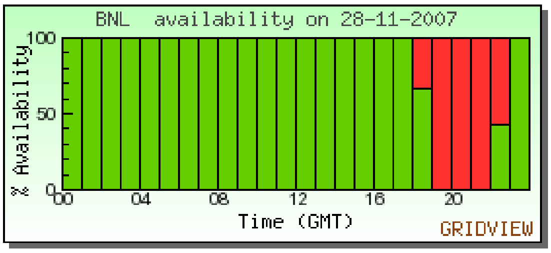
Thursday Nov 8
Srmcp didn't work
Cause: pin manager didn't work properly
Severity: minor, production wasn't affected
Resolution: restart the pin manager
problems:
USATLAS production monitor has constant time out.
Cause:
massive packet loss was observed due to high traffic volume through the RHIC/ATLAS Firewalls. The high traffic was caused by large number of RHIC jobs running on ATLAS farm.
Impact:

User can not use the USATLAS production monitoring. The affects to the USATLAS production is still under investigation.
Solution:
Shutdown the RHIC jobs to avoid high traffic volume through the RHIC/USATLAS firewall. After that, USATLAS production monitor came back.
Gatekeeper gidgk02 was briefly down
cause: OSG upgrade
Resolution: service was restored after upgrade was completed.
Problem: Panda monitor on gridui02 had several short interruptions of service (Continuation of problem from previous day)
cause: Network problem, still investigated.
Impact: Panda monitor pages on gridui02 were not available
Resolution: Problems went away, exact cause is still under investigation.

⇒ 22 Nov 2007



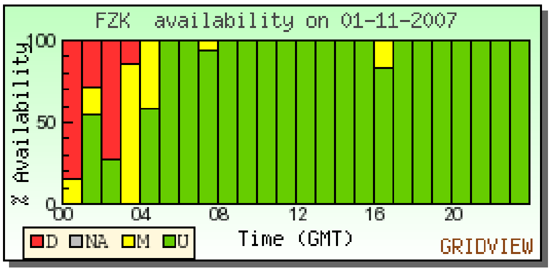
⇒ 28 Nov 2007



Problem: Panda monitor machine gridui01 crashed for 2 hours
Cause: High memory/high load caused the machine to go down
Solution: Machine rebooted. High memory usage must be adressed by developers

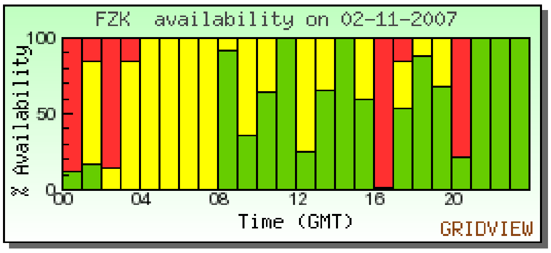
DE-KIT

⇒ 01 Nov 2007



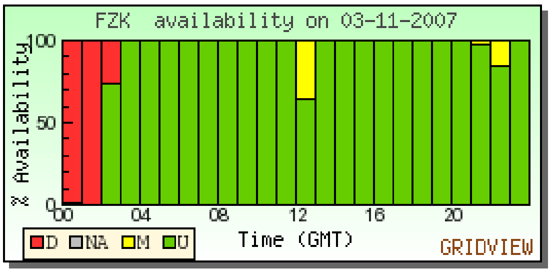
SRM: instabilities
CE: CEs on very high load, dropping out of InfoSystem from time to time
    -> one CE in scheduled downtimes (at risk) for load reasons
       one CE error reasons: unspecified gridmanager errors (only ops affected)

⇒ 02 Nov 2007



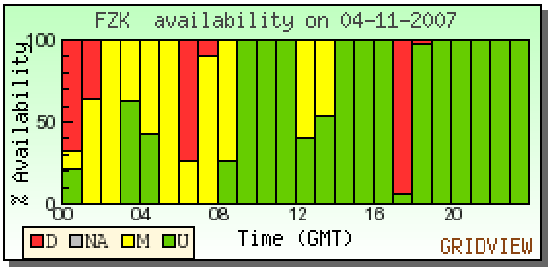
CE instabilities that have been tracked to a erroneous entry of the batch system
configuration.
Impact: severe, all vo''''s
SRM instabilities continue but have been partly caused by nightly database backups that stalled processing.
Impact: moderate, all vo''''s but transfers are restarted

⇒ 03 Nov 2007



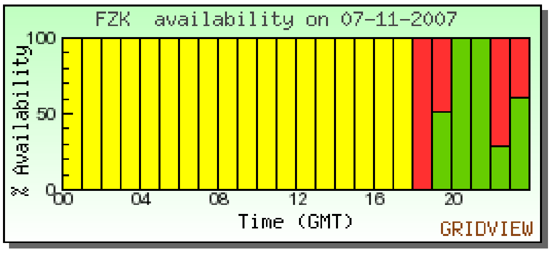
SRM: instabilities
CE: CEs under very high load

⇒ 04 Nov 2007



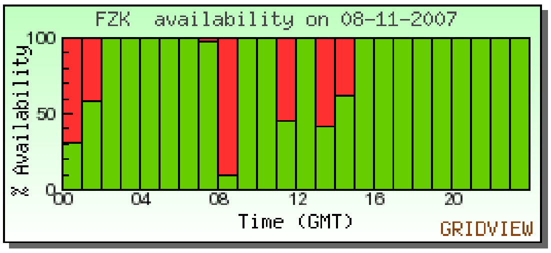
SRM: instabilities
CE: CEs on high load, dropping out of InfoSystem from time to time
    -> one CE in scheduled downtimes (at risk) for load reasons
       one CE error reasons: unspecified gridmanager errors (only ops
affected)

⇒ 07-10 Nov 2007

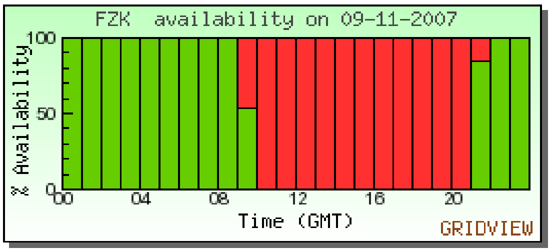


Upgrade to dCache 1.8 extended downtime because some needed features that are working in version 1.7 stopped working in 1.8. Part of it was solved during

the
extended downtime. Waiting for dcache development.

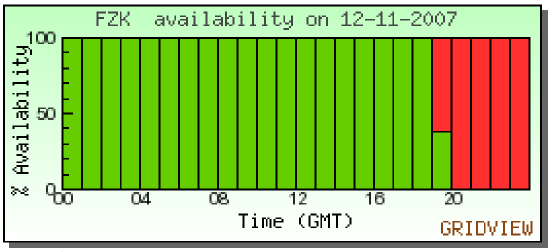


SE problem caused by wrong config entry fixed at 12:00 CET.
SRM instabilities caused by various restarts of dcache pools and servers.



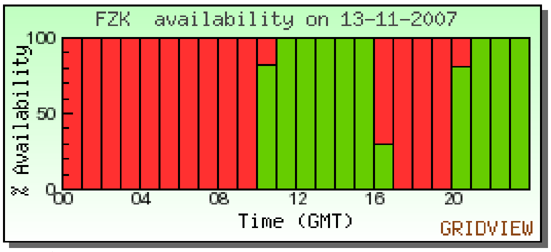
SRM/dCache instability following dcache update

⇒ 12 Nov 2007



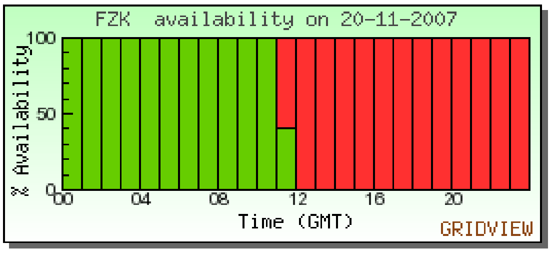
SRM/dCache instability following dcache update

⇒ 13 Nov 2007 :



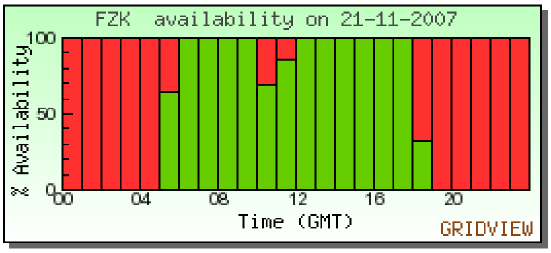
SRM/dCache instability following dcache update

⇒ 20 Nov 2007 :



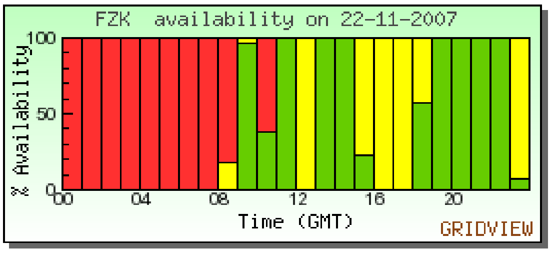
```
SRM instabilities because of hardware problems with a disk system
Severity: low. only the sfts are affected.
```

⇒ 21 Nov 2007 :



```
dCache SRM bug workaround caused disk space overflow
Severity: all VOs affected
```

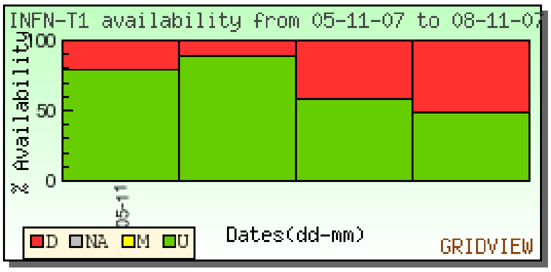
⇒ 22 Nov 2007 :



```
Fix for dCache deployed. A number of files (at least of the ATLAS VO)
were lost.
```

IT-INFN-CNAF

⇒ 05-08 Nov 2007



On **November 5**, we had some problems with our LSF system due to an high load on the LSF sw shared area.

On **November 6**, late afternoon, we noticed many GRID jobs failed due to the error:

```
[...]
fetch-crl[9108]: 20071106T182318+0100 RetrieveFileByURL: download no data
from http://gridca.hpcc.nectec.or.th/pub/crl/cacrl.crl
fetch-crl[9108]: 20071106T182318+0100 downloaded file from
http://gridca.hpcc.nectec.or.th/pub/crl/cacrl.crl is not a valid CRL file
fetch-crl[9108]: 20071106T182318+0100 Could not download any CRL from :
[...]
```

To prevent other failures we put in inactive status our grid queues (queues accepted new jobs but took them in pending status).

As a consequence SAM test jobs were not able to run marking us as "not available".

After a few hours we were able to download manually a valid CRL file from the server above and hence we reactivated the queues.

Starting **from the late afternoon of November 7 to the morning of 8**, SAM tests for srm had been failing at IT-INFN-CNAF. The problem was common to all our castor srm end-points, so our suspects were appointed on CASTOR itself. After a careful inspection no major issues were found except for a long queue in tape migration particularly for dteam VO.

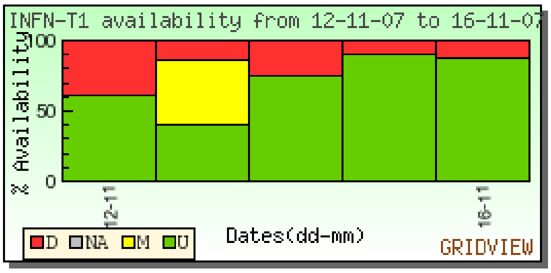
Moreover we observed a flooding of requests for CASTOR from dteam (a factor 20 respect to usual).

Therefore we suspect that the SAM tests have been biased by this.

We are configuring an additional set of sanity checks at the batch system level to help preventing this kind of situations

On **November 8**, apparently by mistake, a job with the flag #BSUB -n 129198776 was submitted: 129198776 pending jobs where created on our batch system (LSF) causing dteam jobs not being executed until the above mentioned job was killed.

⇒ 12-16 Nov 2007

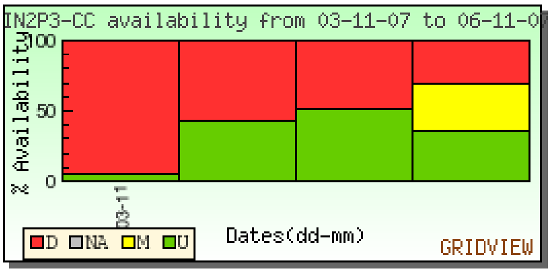


November 13-16 – CASTOR upgrade from 2.1.3-24 to 2.1.4-10. After the upgrade, a bug was discovered on the stager. Two hot-fixes, released by the CERN Castor development support, have been applied.

All VOs were affected.

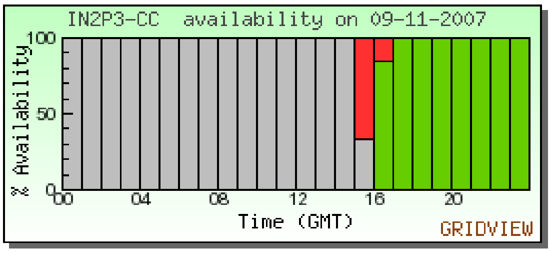
FR-CCIN2P3

⇒ 03-06 Nov 2007



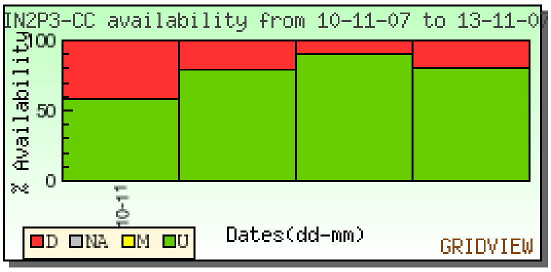
AFS problem. Jobs locked in queue from 00:00PM to 11:30PM on 03/11, from 08:00AM to 02:00PM on 06/11.

⇒ 09 Nov 2007



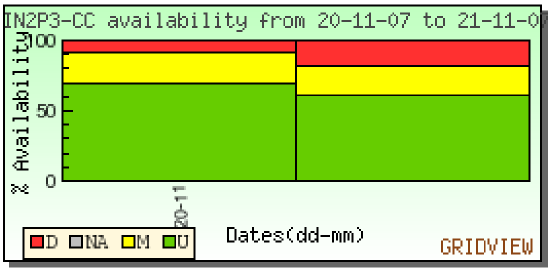
- sometimes timout with local SE SRM : ccsrm.in2p3.fr during transfer. This makes CE lcg-utils tests failed.

⇒ 10-13 Nov 2007



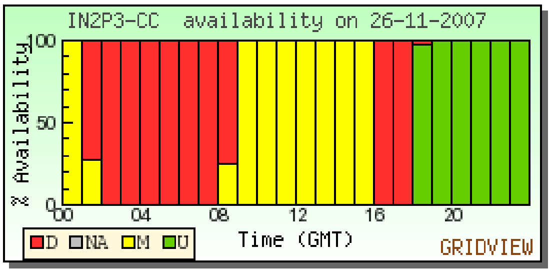
Oracle service was overloaded on 12/11.
Workers became blackholes due to "low memory saturation" on SL4 32b.

⇒ 20-21 Nov 2007



SD for BQS service from 08:30AM to 02:00PM on 20/11.
Cooling outage from 10:00AM to 05:30PM on 21/11

⇒ 26 Nov 2007



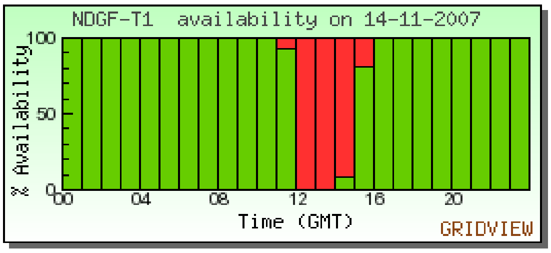
Unscheduled downtime from 07:00PM on 25/11 to 05:PM on 26/11 : AFS problem on SL4 32b machines and an AFS server shutdown.
dCache Upgrade on 26-27/11.

CERN

No periods below target.

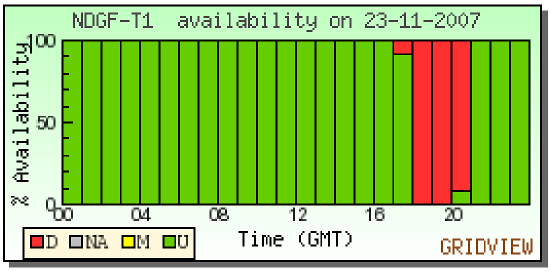
NDGF

⇒ 14 Nov 2007



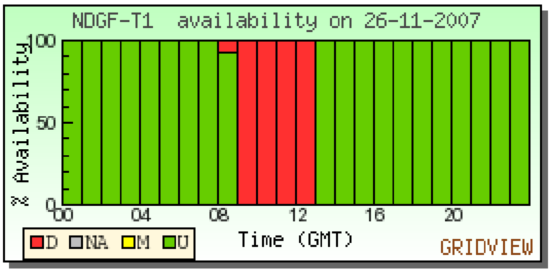
Typo in SAM sensor led to all CE:s failing SAM test.

⇒ 23 Nov 2007



SAM tests could not be submitted for a brief period, due to the dying specialized GIIS server. WLCG VOs do not use this server, thus actual Tier1 services were not interrupted - only SAM tests. The reason was a faulty disk; it is replaced now and a fall-back GIIS is being deployed in order to alleviate such problems in future.

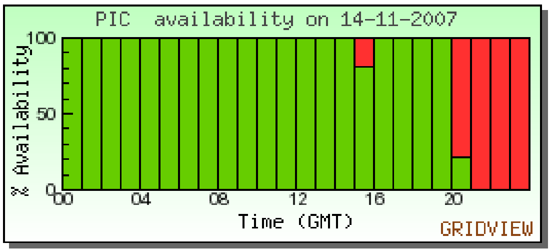
⇒ 26 Nov 2007



SAM tests could not be submitted for a brief period, due to the dying specialized GIIS server. WLCG VOs do not use this server, thus actual Tier1 services were not interrupted - only SAM tests. The reason was a faulty disk; it is replaced now and a fall-back GIIS is being deployed in order to alleviate such problems in future.

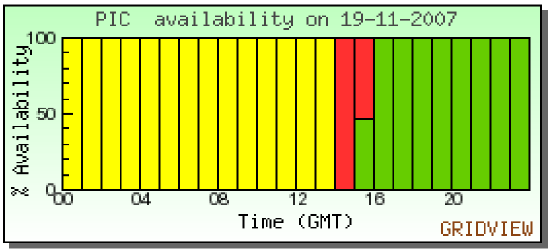
ES-PIC

⇒ 14-15 Nov 2007



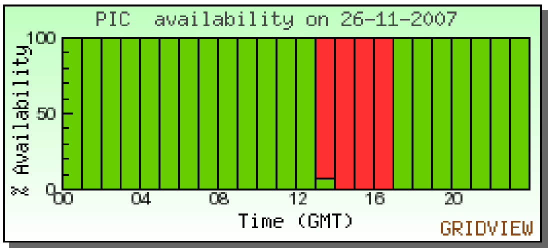
Date: From 14/11/07 at 18:43 UTC until 15/11/07 at 14:17 UTC
Problem: The SRM service stops working. The cause is a log file of a dCache service (pnfsd) reaching 2GB size and not rotating properly.
Severity: High. The SRM service was unavailable during this time.
Solution: The rotation of the log file was manually forced. It took longer
than
expected due to simultanous travels of four of the system experts. A
permanent
correction of this problem is being implemented as part of the robustisation
of
the service.

⇒ 19 Nov 2007



Date: 19/11/2007 15:31
Problem: We had a Sched. Downtime this day until 14:00 UTC. The batch queues were enabled at PIC at 15:51 local time (14:51 UTC), so we reopened the
queues
a bit too late.
Severity: Low. The Sched. Downtime lasted about one hour more than planned.
Solution: None.

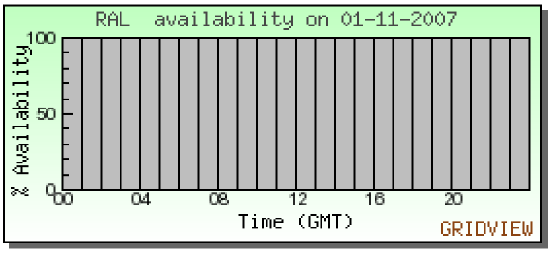
⇒ 26 Nov 2007



```
Date: 26/11/2007 from 12:40 UTC until 15:40 UTC
Problem: A failure in the internal pro-active monitoring system (Ingrid)
caused
the site-bdii.pic.es to fail during some hours.
Severity: Medium. lcg-utils commands failed, since SEs were not in the
infosys.
Solution: Restarting the site-bdii.
```

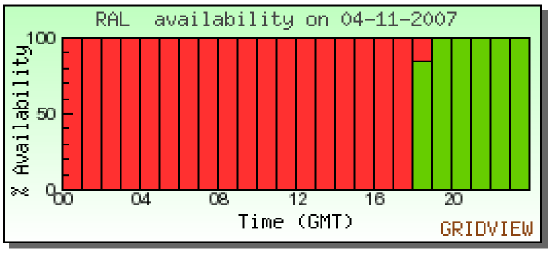
UK-T1-RAL

⇒ 01 Nov 2007 :



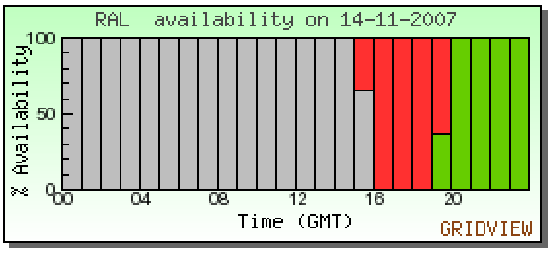
GGUS ticket 28520 raised as tests were unavailable at remote site. Tests have restarted at circa 04:00 on Friday 2nd Nov. RAL site available during this period confirmed with Grid services group.

⇒ 03-04 Nov 2007



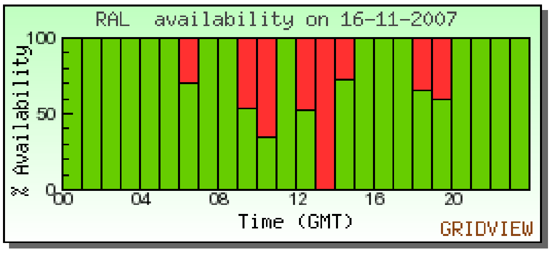
Problem: Java memory problem on dCache SE used for CE tests
Solution: server was restarted and the memory limit was increased; a Nagios test will be added shortly to catch it again

⇒ 14 Nov 2007



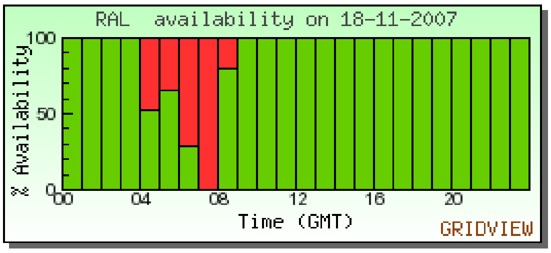
External Gstat problems account for sBDII failures.

⇒ 16 Nov 2007



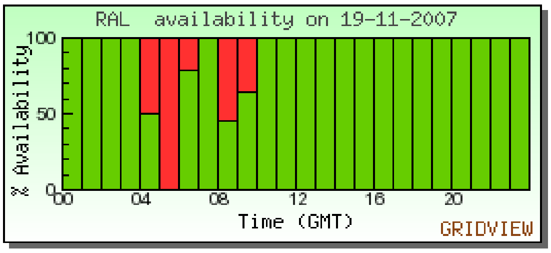
dCache timeouts.

⇒ 18 Nov 2007 :



dCache timeouts.

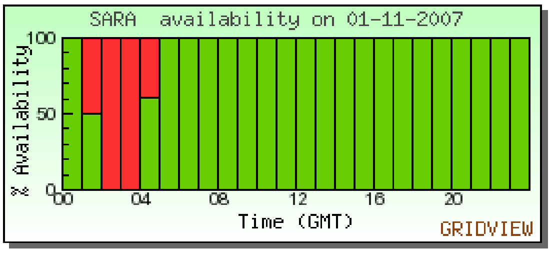
⇒ 19 Nov 2007 :



CE tests failing due to dCache timeouts; switched default OPS VO to CASTOR endpoint, ralsrma.rl.ac.uk.

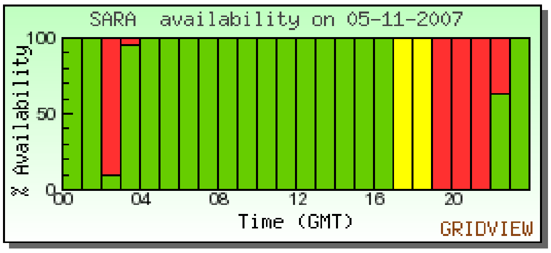
NL-T1

⇒ 01 Nov 2007



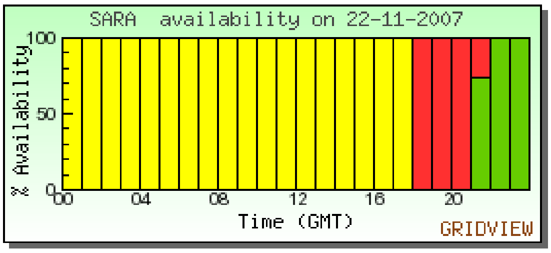
lcg-cr timeout after 600 seconds again.

⇒ 05 Nov 2007



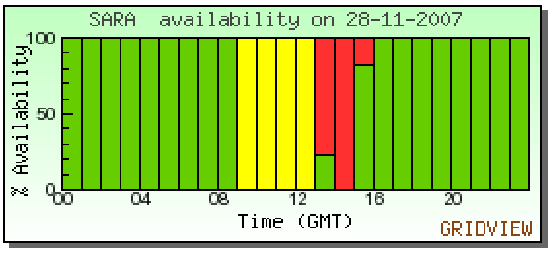
Problem: Cannot download testjob.tgz from
gsiftp://rb113.cern.ch/var/edgwl/SandboxDir/Gm/https_3a_2f_2frb113.cern.ch_3a9000_2fGmb34LoktRocc7BY-gzjKA/input/
Caused by a problem on the remote site. Not sure why, but some other sites also
had this problem.
Problem2: lcg-cr -v --vo ops -d srm.grid.sara.nl -l
lfn:sft-lcg-rm-cr-wn01.gina.sara.nl.071105133246
file:///home/opsgm/opsgm01/gram_scratch_RGu1Gunp0a/work/testjob/nodes/ce.gina.sara.nl/sft-lcg-rm-cr.txt
send2nsd: NS002 - send error : No valid credential found
Crondeamon had crashed and therefore the certificate revocation list was out of date

⇒ 22 Nov 2007



See 2007-11-19. The red is due to the fact that the scheduled downtime fell a little bit short.

⇒ 28-29 Nov 2007

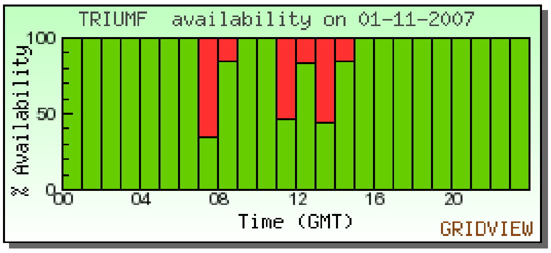


Maintenance due to necessary immediate upgrade of dCache.

The red is due to SAM problems.

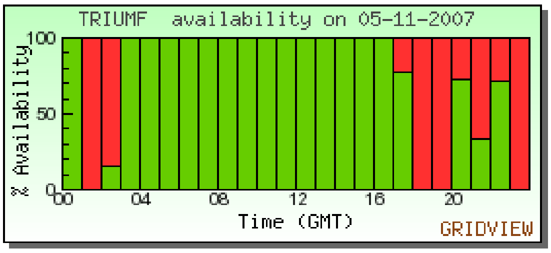
CA-TRIUMF

⇒ **01 Nov 2007**



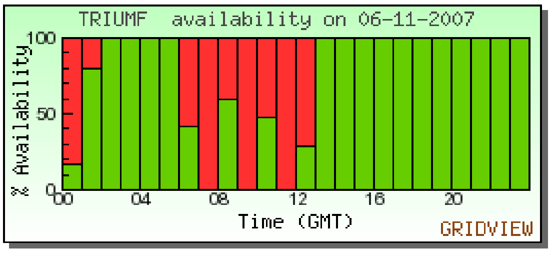
2 tests ran at same time on same node - SAM bug

⇒ **05 Nov 2007 :**



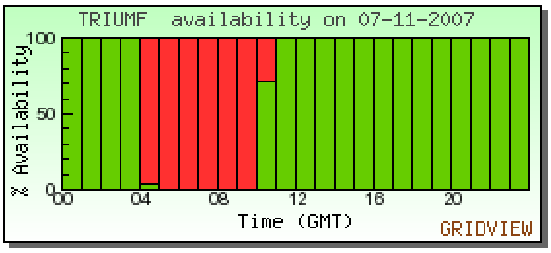
srm overload due to heavy productions use

⇒ **06 Nov 2007 :**



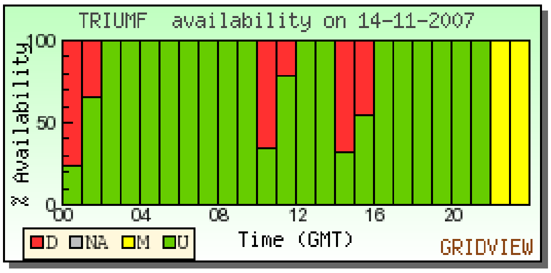
More overload. switch production from lcg-cp to dccp

⇒ **07 Nov 2007 :**

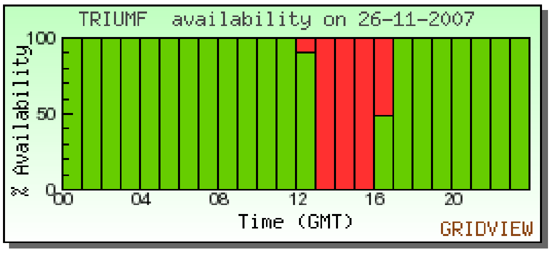


Site-wide power cut. 19:00-21:00. Site remained on ups but generator failed so
netwokk went down. On-call person automatically SMSd and went in with torch.
Workers powered down to prevent overheating as AC off.
All back by 01:00.

⇒ **14 Nov 2007**

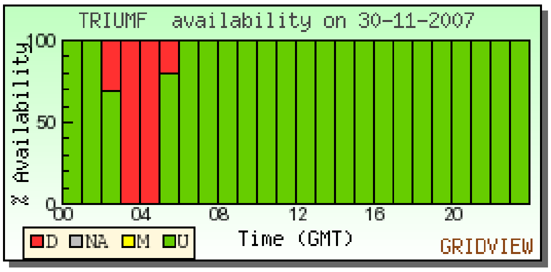


⇒ **26 Nov 2007**



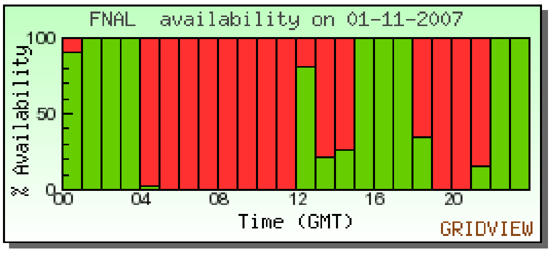
SRM trouble.

⇒ **30 Nov 2007**



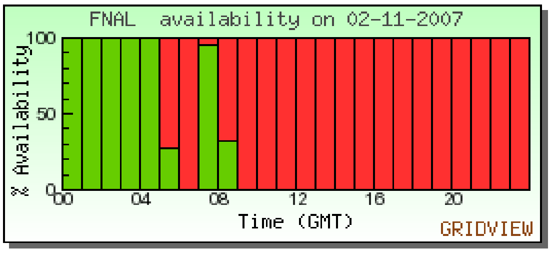
US-FNAL-CMS

⇒ 01 Nov 2007 :



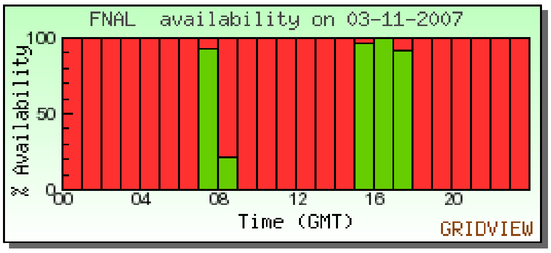
Scheduled downtime,
GOC downtime tool didn''t work - permission denied

⇒ 02 Nov 2007 :



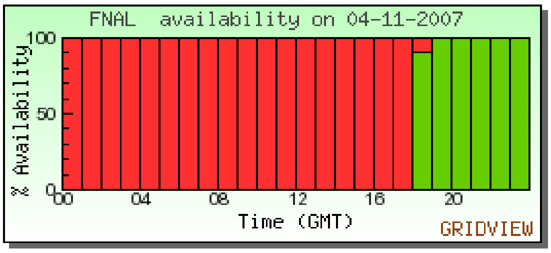
The host cert for the cmssrm was replaced, but the srm was not restarted, so it
was using the old expired cert. This was fixed by a srm restart.

⇒ 03 Nov 2007 :



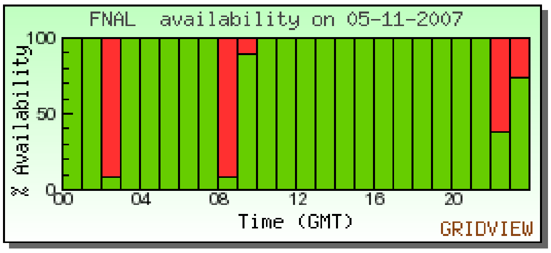
The host cert for the cmssrm was replaced, but the srm was not restarted, so it
was using the old expired cert. This was fixed by a srm restart.
The system was fully function from 08 onwards, test defect.

⇒ 04 Nov 2007 :



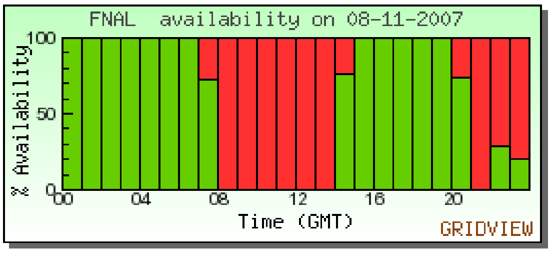
The system was fully functional, test defect.

⇒ 05 Nov 2007 :



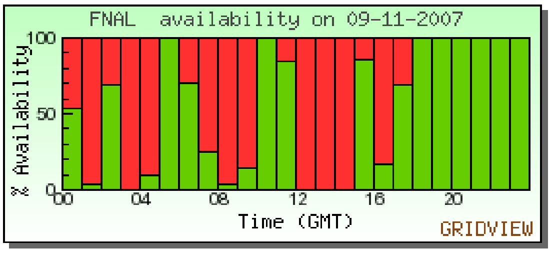
The system was fully functional, test defect.

⇒ 08 Nov 2007 :



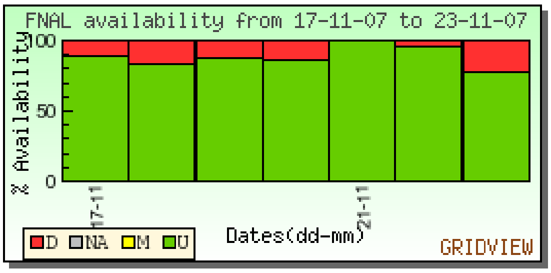
The system was fully functional, test defect.

⇒ 09 Nov 2007 :

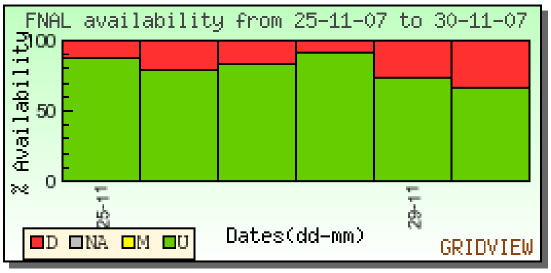


No problems at FNAL.

⇒ 17-23 Nov 2007



⇒ 25-30 Nov 2007


FNAL availability from 25-11-07 to 30-11-07
% Availability
Dates(dd-mm)
D
NA
M
U
GRIDVIEW