# WLCG Site Reliability Reports - December 2007

- *Please review and complete the Site Reports below. Edit your section and mail the document back to A.Aimar.*
- *Deadline: Monday 14 January 2008*

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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01/12/2007 | 1 | 100% | n/a | 100% | n/a | 100% | 100% | 92% | 100% | 75% | 100% | 100% | 97% | 91% | 100% |
| 02/12/2007 | 2 | 100% | n/a | 100% | n/a | 100% | 100% | 91% | 100% | 73% | 100% | 100% | 96% | 91% | 100% |
| 03/12/2007 | 3 | 100% | 0% | 59% | 82% | 64% | 76% | 100% | 100% | 87% | 100% | 100% | 81% | 91% | 100% |
| 04/12/2007 | 4 | 100% | 100% | 0% | 100% | 48% | 0% | 100% | 100% | 95% | 100% | 91% | 78% | 91% | 100% |
| 05/12/2007 | 5 | 100% | 91% | 12% | 100% | 55% | 73% | 100% | 100% | 61% | 84% | 100% | 81% | 91% | 100% |
| 06/12/2007 | 6 | 100% | 100% | 92% | 100% | 53% | 30% | 85% | 100% | 68% | 100% | 100% | 86% | 91% | 100% |
| 07/12/2007 | 7 | 100% | 100% | 82% | 100% | 100% | 0% | 100% | 92% | 92% | 100% | 100% | 89% | 91% | 100% |
| 08/12/2007 | 8 | 100% | 100% | 100% | 100% | 100% | 0% | 100% | 100% | 56% | 100% | 100% | 88% | 91% | 100% |
| 09/12/2007 | 9 | 100% | 97% | 100% | 100% | 100% | 0% | 100% | 100% | 87% | 88% | 100% | 89% | 91% | 100% |
| 10/12/2007 | 10 | 100% | 100% | 100% | 100% | 100% | 0% | 100% | 100% | 95% | 100% | 100% | 91% | 91% | 100% |
| 11/12/2007 | 11 | 100% | 100% | 100% | 100% | 98% | 7% | 100% | 100% | 81% | 93% | 100% | 90% | 91% | 100% |
| 12/12/2007 | 12 | 100% | 100% | 95% | 100% | 82% | 0% | 95% | 100% | n/a | 100% | 100% | 88% | 91% | 100% |
| 13/12/2007 | 13 | 100% | 100% | 100% | 100% | 100% | 25% | 100% | 100% | 100% | 100% | 100% | 94% | 91% | 100% |
| 14/12/2007 | 14 | 100% | 100% | 100% | 100% | 100% | 96% | 100% | 62% | 87% | 87% | 58% | 91% | 91% | 100% |
| 15/12/2007 | 15 | 100% | 100% | 100% | 100% | 96% | 97% | 100% | 100% | 92% | 100% | 0% | 90% | 91% | 100% |
| 16/12/2007 | 16 | 100% | 100% | 100% | 100% | 100% | 76% | 100% | 100% | 83% | 100% | 0% | 88% | 91% | 100% |
| 17/12/2007 | 17 | 100% | 100% | 100% | 93% | 100% | 53% | 100% | 100% | 100% | 100% | 0% | 87% | 91% | 100% |
| 18/12/2007 | 18 | 100% | 100% | 100% | 91% | 90% | 36% | 100% | 100% | 80% | 82% | 0% | 82% | 91% | 100% |
| 19/12/2007 | 19 | 100% | 95% | 100% | 100% | 100% | 44% | 100% | 100% | 100% | 74% | 0% | 84% | 91% | 100% |
| 20/12/2007 | 20 | 100% | 87% | 100% | 100% | 94% | 90% | 100% | 100% | 92% | 77% | 0% | 87% | 91% | 100% |
| 21/12/2007 | 21 | 100% | 54% | 100% | 66% | 100% | 100% | 84% | 100% | 100% | 100% | 4% | 84% | 91% | 100% |
| 22/12/2007 | 22 | 100% | 39% | 100% | 88% | 100% | 100% | 100% | 100% | 100% | 100% | 0% | 86% | 91% | 100% |
| 23/12/2007 | 23 | 100% | 100% | 100% | 100% | 100% | 95% | 100% | 100% | 100% | 95% | 0% | 91% | 91% | 100% |
| 24/12/2007 | 24 | 100% | 100% | 100% | 100% | 100% | 93% | 100% | 100% | 100% | 93% | 0% | 91% | 91% | 100% |
| 25/12/2007 | 25 | 100% | 100% | 100% | 100% | 100% | 97% | 97% | 100% | 100% | 100% | 0% | 91% | 91% | 100% |
| 26/12/2007 | 26 | 100% | 82% | 100% | 100% | 100% | 100% | 93% | 100% | 100% | 98% | 0% | 89% | 91% | 100% |
| 27/12/2007 | 27 | 100% | 87% | 100% | 84% | 100% | 13% | 98% | 100% | 51% | 100% | 0% | 78% | 91% | 100% |
| 28/12/2007 | 28 | 100% | 91% | 100% | 84% | 100% | 42% | n/a | 100% | 94% | 100% | 0% | 83% | 91% | 100% |
| 29/12/2007 | 29 | 100% | 100% | 100% | 94% | 45% | 0% | 59% | 100% | 84% | 100% | 0% | 74% | 91% | 100% |
| 30/12/2007 | 30 | 100% | 100% | 100% | 97% | 100% | 0% | 100% | 100% | 92% | 100% | 0% | 82% | 91% | 100% |
| 31/12/2007 | 31 | 100% | 100% | 100% | 100% | 96% | 0% | 100% | 100% | 100% | 100% | 0% | 83% | 91% | 100% |
| Average reliability | | 100% | 90% | 92% | 96% | 91% | 50% | 96% | 99% | 88% | 96% | 44% | 87% | 91% | 100% |

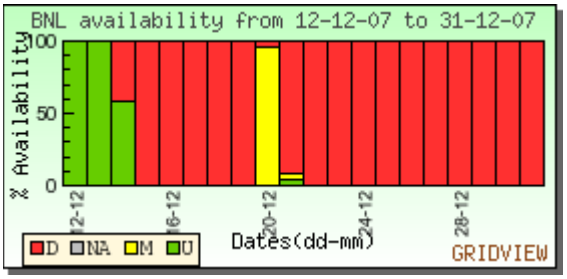
## TW-ASGC

⇒ **14 Dec 2007**

Title: SAM lcg-rep failures due to missing CERN DPM SE from asgc bdii
Date: start from 14-Dec-2007 09:10:27 and end at 14-Dec-2007 17:11:15, at least 8 events have been detected
Reason: seems the problem arises from time out of the bdii query, looks like the ldif wasn't able to complete before the timeout. Somehow the bdii query timeout have been reduce to 30s only, I am extending the criterion to 120s, and also for the breathe timeout; it have been confirm that we're able to query relevant end point from gfal. Have confirm that latest SAM result start passing at '14-Dec-2007 18:32:50'. I double check the smokeping monitoring page, and confirm the problem wasn't related to network, except for generic timeout error due to the ldapsearch query from bdii.
Severity: the impact is severe, and at least 6 sites in APROC are affected, and result in SAM lcg-rep testing failures.
Solution: by extending the timeout limit in bdii, we're able to fix the problem, but root cause remains unclear since the same time out have been applied since Jun this year, and the timeout of bdii query found since 7pm (UTC) today.

## US-T1-BNL

⇒ **15 Dec – 31 Dec 2007**




From P.Nyczyk:
I checked carefully SAM DB for information related to BNL SE(s), and I find the following (all results for OPS VO):

There are two machines dcsrm.usatlas.bnl.gov and dcsrmv2.usatlas.bnl.gov

The first one (dcsrm) was passing the tests since the beginning of December, but on 14th Dec it was removed from BNL-LCG2 site. Later on it was still tested with failures between 17th and 20th (no information in BDII) and occasional failures after that. However it didn't contribute to BNL availability since 14th Dec.

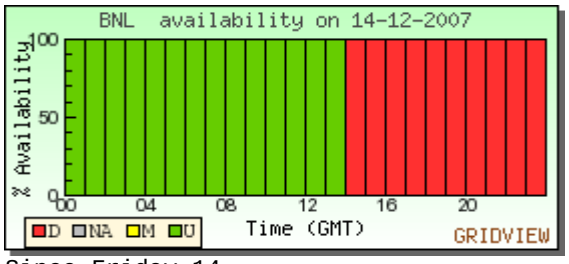
The second machine (dcsrmv2) was failing the tests since the beginning with the following error message:
Exception thrown by diskCacheV111.services.authorization.GPLAZMALiteVORoleAuthzPlugin: Permission Denied: Cannot determine Username from grid-vorolemap for DN /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=samoper/CN=582979/CN=Judit Novak and role /ops/Role=lcgadmin/Capability=NULL

To sum up, until 14th December BNL was available according to SAM/GridView as it had one good SE (dcsrm) and one "bad" (dcsrmv2). After that date only the "bad" one remained, and consequently the availability dropped to 0.

I don't see any failures related to the directory permission problem you are referring to. Anyway as for SAM tests, they are never "choosing" the directory on SE. The test just depends on lcg-utils which use the discovery mechanism and take the directory

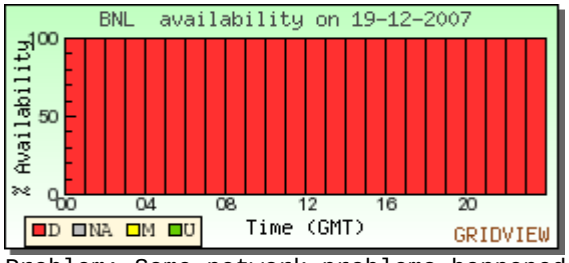
assigned for the VO from the BDII. So if there are any problems like using wrong directory you should rather look at your site BDII and which directory you are publishing there for OPS VO.

### ⇒ 14 Dec 2007



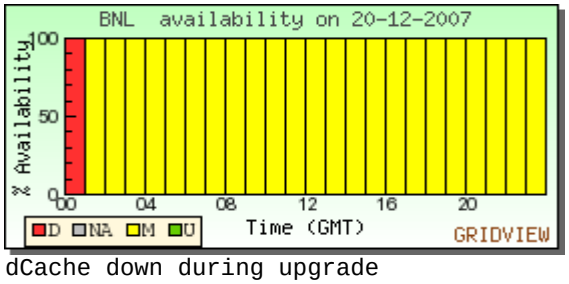

Since Friday 14
BNL is reported as down even though the SE is working properly
Cause: there seem to be a number of causes: the first one, which is now solved,was that the SAM tests were trying to write in the wrong directory. Currently,even though the SAM tests are passing, BNL is still reported as down.
Severity: dCache reported as down in GridView, even though the system is working correctly
Remediation: still under investigation

### ⇒ 19 Dec 2007




Problem: Some network problems happened on Panda production servers.
Cause: There is no specific cause for the problem yet: Our speculation is that BNL dCache sends high volume of data traffic to BNL firewall. In the mean time, we observed that Panda servers has network connection problem.
Impact: the USATLAS/Triump/IN2P3 production services are impacted.
Solution: we have to redirect the traffic to the GridFtp doors instead of firewalls, and we are speeding up the Panda relocation to a new subnet which does not suffer the firewall problems as bad as the current subnet which the Panda servers reside.

### ⇒ 20 Dec 2007

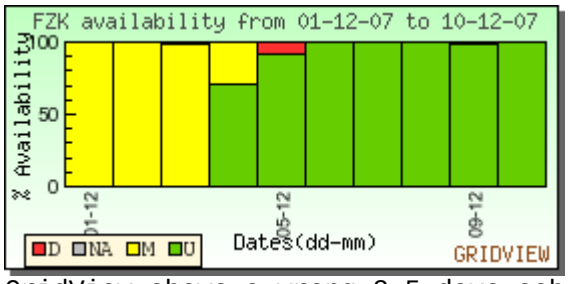



dCache down during upgrade
Cause: planned outage
Severity: system down

---

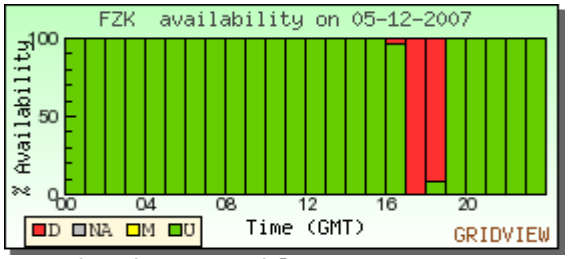
## DE-KIT

### ⇒ 01 Dec 2007 – 04 Dec 2007



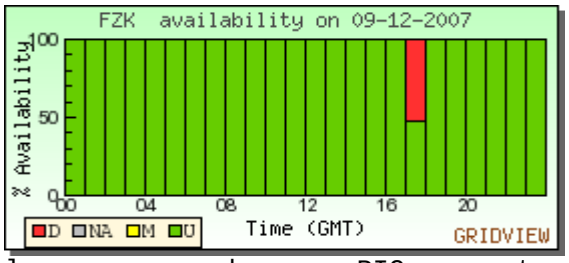

GridView shows a wrong 3.5 days scheduled downtime due to a GridView summarizer bug after changing downtime info in GOCDB. See
GGUS ticket #29977
Savannah bug #31877
The bug was fixed by GridView developers soon after bug report submission.

⇒ **05 Dec 2007**



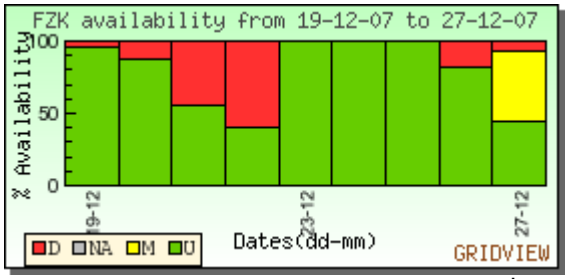

srm database problem. A restart was needed.

⇒ **09 Dec 2007**




lcg-cr errors because PIC was not reachable. This is a bug. Ticket was opened.

⇒ **20 Dec 2007 – 27 Dec 2007**

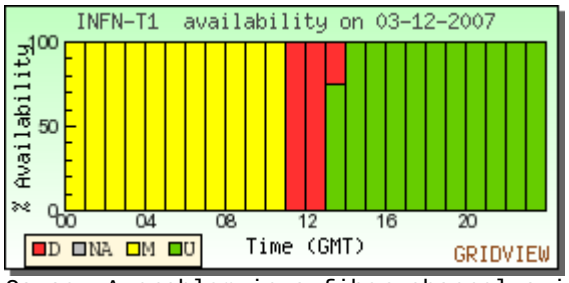



20 Dec 2007 – 22 Dec 2007: SRM data base became extremely slow due to massive usage. Data base was dropped on 23rd.

26 Dec 2007 – 27 Dec 2007: Lost one interface (hardware) in the GridKa backbone and a respective failover mechanism didn't work as expected. Error source is analyzed and corrected.

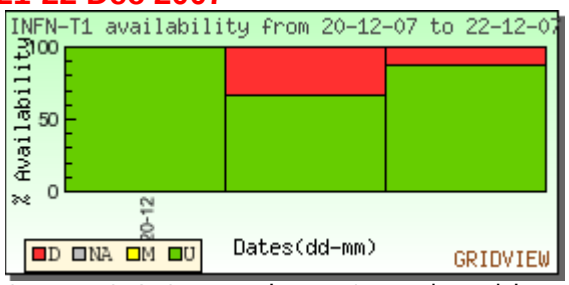
## IT-INFN-CNAF

⇒ **3 Dec 2007**



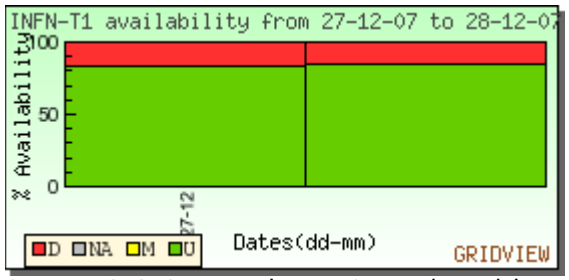

Cause: A problem in a fiber channel switch was found and a fix was applied.
Severity: some storage subsystems and farm production queues

⇒ **21-22 Dec 2007**




Cause: CASTOR services stopped working (but apparently were up). A restart fixed them
Severity: CASTOR services unavailable for all LHC experiments (except ATLAS D1T0)

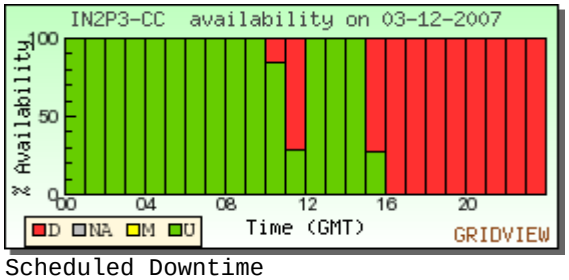
⇒ **27-28 Dec 2007**




Cause: CASTOR services stopped working (but apparently were up). A restart fixed them
Severity: CASTOR services unavailable for all LHC experiments (except ATLAS D1T0)

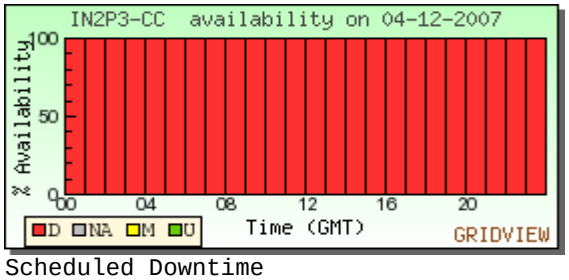
## FR-CCIN2P3

⇒ **03 Dec 2007**



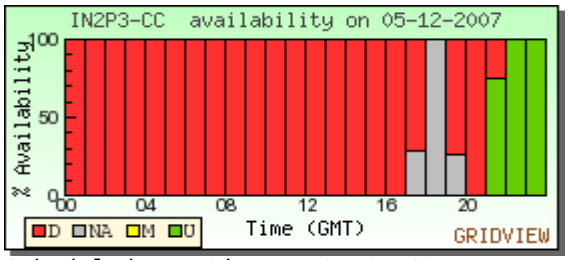

Scheduled Downtime

⇒ **04 Dec 2007**



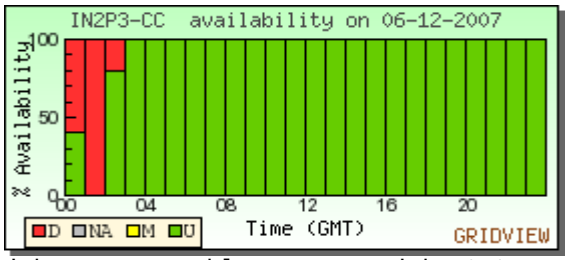

Scheduled Downtime

⇒ **05 Dec 2007**



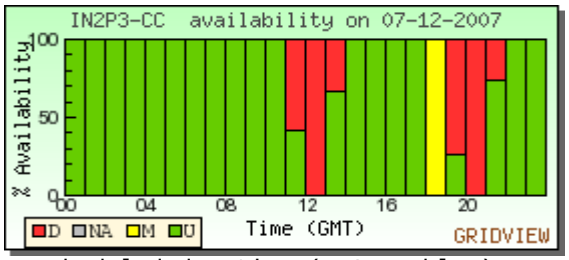

Scheduled Downtime up to 05:00PM.
from 07:00 PM to 08:30 SRM problem

⇒ **06 Dec 2007**



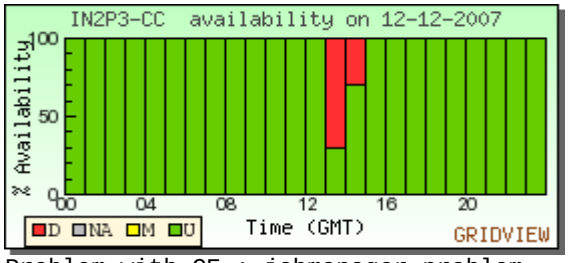

jobmanager problem : wrong job status reported

⇒ **07 Dec 2007**




Unscheduled downtime (AFS problem)

⇒ **12 Dec 2007**




Problem with CE : jobmanager problem

## CERN

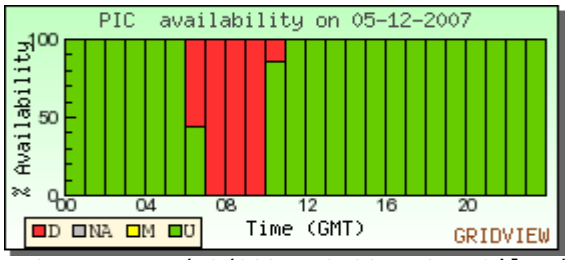
No periods below target.

## NDGF

No periods below target.

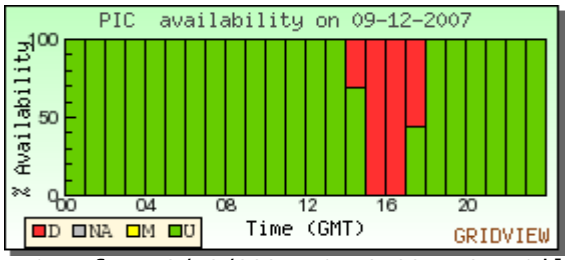
## ES-PIC

### ⇒ 05 Dec 2007



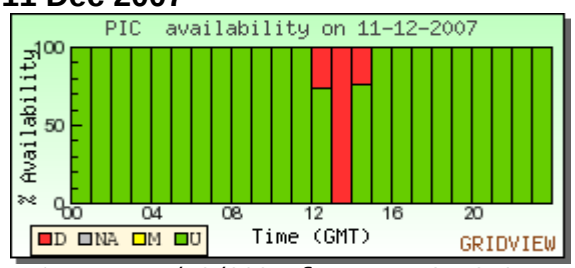

Date: From 4/12/2007 19:00 UTC until 5/12/2007 9:00 UTC
Problem: All the dcache gridftp doors (9 nodes) crash (kernel panic) due to overload. Too many transfer streams simultaneously and the processes ran out of memory.
Severity: High. The SRM service is unavailable during the failure time.
Solution: The server was restored after rebooting gridftp hosts. The limits on the max number of concurrent streams per door was lowered to avoid this high load to be reached again.

### ⇒ 09 Dec 2007




Date: from 8/12/2007 at 18:00 UTC until 9/12/2007 at 15:00 UTC
Problem: The SOA DNS for pic.es has an outage. The 21hrs of DNS outage finally result in about 3hrs of service interruption, since the TTLs of the hostnames was set up to 18hrs.
Severity: High. During about 3hrs, the services at PIC were unreachable due to DNS resolution not working.
Solution: The DNS server was restarted. For the moment we have increased the TTL to 36hrs. A deeper DNS robustization is ongoing.

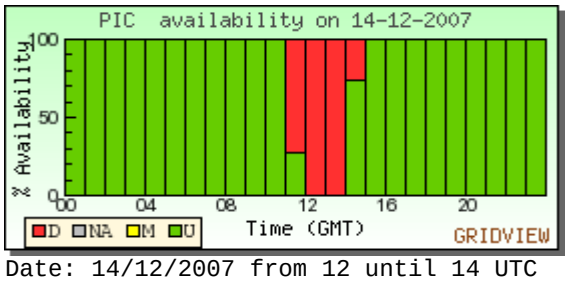
### ⇒ 11 Dec 2007




Date: on 11/12/2007 from 11:43:53 CET until 21:58:53 CET
Problem: OPN failure. Outage on the Dark Fiber between Madrid and Geneva.
Severity: Medium. It affects only to part of the PIC services. Those in the new

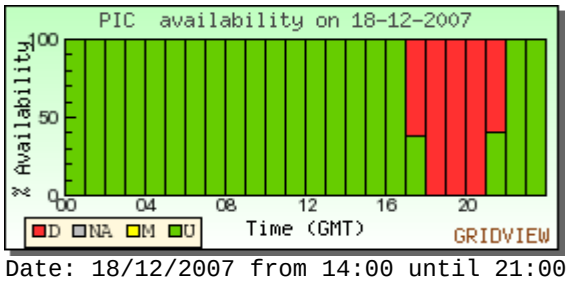
IP range (this is the SRM-disk service)
Solution: GEANT reported that they solved the problem. Still waiting for a complete explanation.

## ⇒ 14 Dec 2007



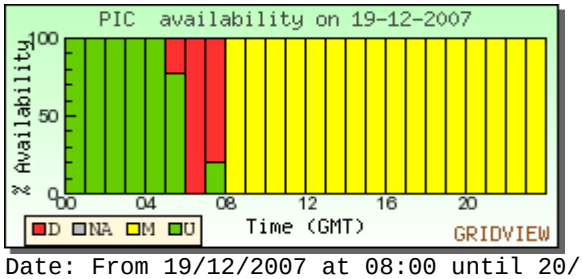

Date: 14/12/2007 from 12 until 14 UTC
Problem: A migration of h/w of the PNFS server that should have taken few minutes, took longer than expected.
Severity: Medium. Some pools took up to 2 hours to become operative again.
Solution: None.

## ⇒ 18 Dec 2007



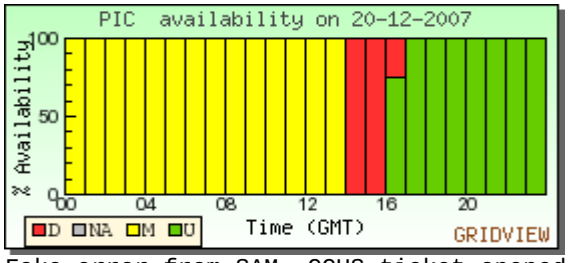

Date: 18/12/2007 from 14:00 until 21:00 aprox.
Problem1: A problem in the yaim configuration of the CEs breaks the authentication for OPS after reconfiguration for a glite update.
Solution1: Correct the edg-mkgridmap.conf for the OPS VO.
Problem2: A problem in the configuration of the site-bdii broke the information published for the castorsrm service.
Solution2: The previous site-bdii configuration was restored.
Severity: Low. The Problem1 only affected the OPS VO, and the problem2 only affected the castor service, which is being deprecated.

## ⇒ 19 Dec 2007



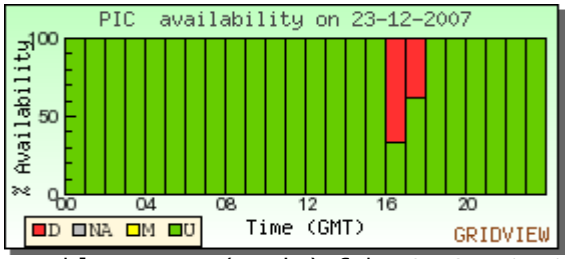

Date: From 19/12/2007 at 08:00 until 20/12/2007 at 14:00
Issue: Scheduled Downtime for upgrading the Storage Service from dcache-1.7 to dcache-1.8

## ⇒ 20 Dec 2007



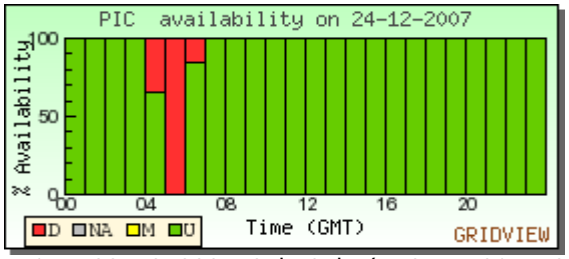

Fake error from SAM, GGUS ticket opened #30807
Problem: Fake failures
Solution: -
Severity: Very low

⇒ **23 Dec 2007**



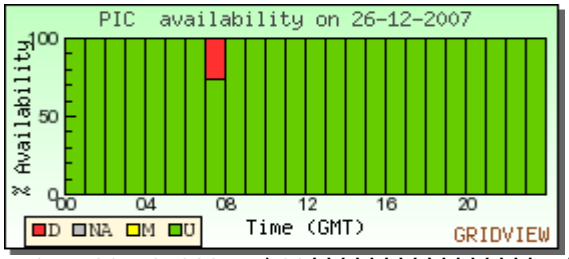

Problem:seems (again) fake CE SAM test failures.
Severity: Low
Solution: GGUS #30852

⇒ **24 Dec 2007**




Date: 23-12-2007 04h-07h (ce05,ce06,ce07)
Problem: Fake CE SAM test failures (Problem with the RB at CERN?)
Severity: Low
Solution: GGUS #30852

⇒ **26 Dec 2007**

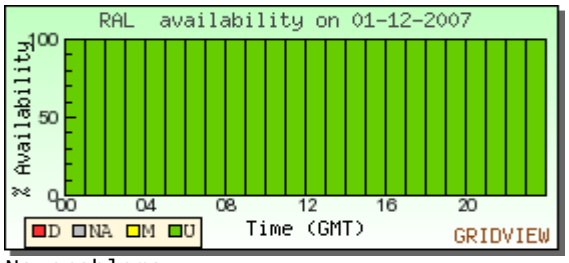



Date: 26-12-2007 7h09''''''''''''''''''-7h24''''''''''''''''
Problem: bdii query response time (ms): 0 NOTE: no reponse time collected bdii entries found: 0 ERROR: no bdii entries!
Severity: Low
Solution: Nothing - false positive ?

---

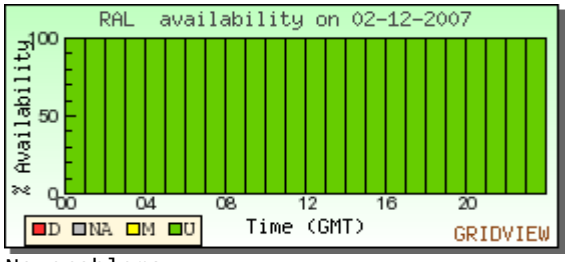
## UK-T1-RAL

⇒ **01 Dec 2007**



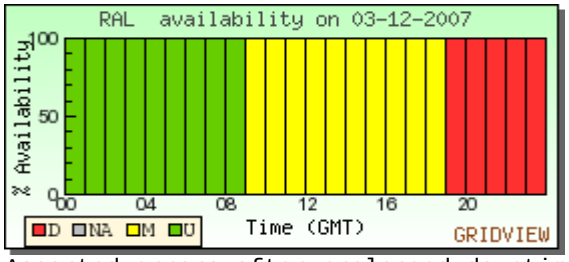

No problems

⇒ **02 Dec 2007**



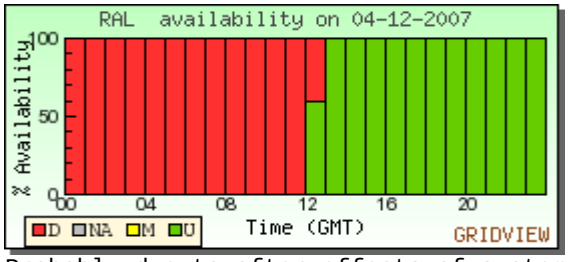

No problems

⇒ **03 Dec 2007**



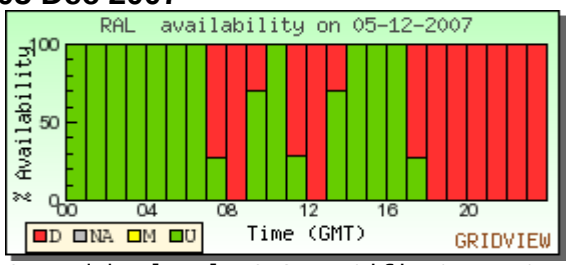

Assorted errors after prolonged downtime for maintenance, fixed during following day

⇒ **04 Dec 2007**



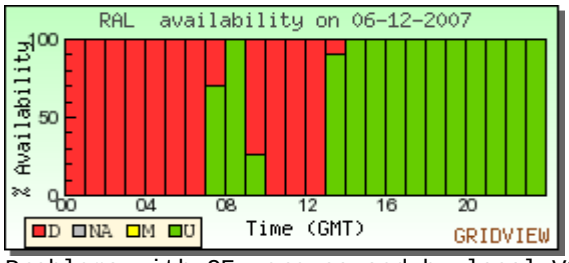

Probably due to after effects of system upgrades across the Tier1 on previous day (Monday 3rd December)

⇒ **05 Dec 2007**



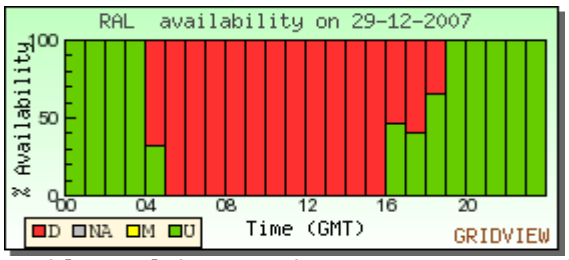

Caused by local VOMS certificates not at latest release; now fixed

⇒ **06 Dec 2007**




Problems with CE were caused by local VOMS certificates not at latest release

⇒ **29 Dec 2007**

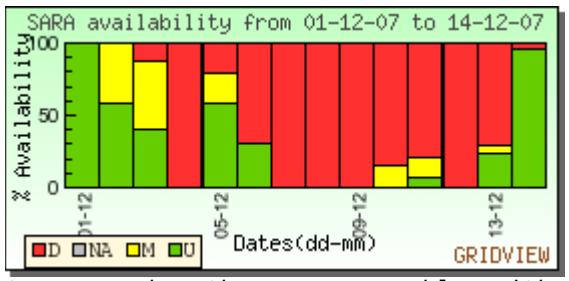



Problem: globus-gatekeeper process on the CE died, preventing job submissiion via the grid
Solution: restarted process

---

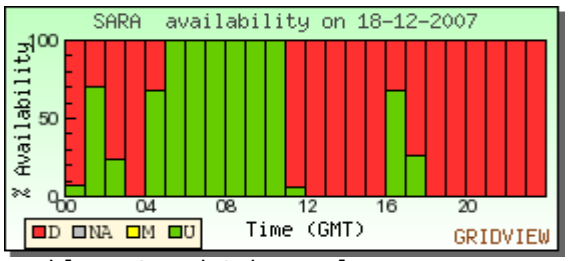
## NL-T1

⇒ **4-13 Dec 2007**



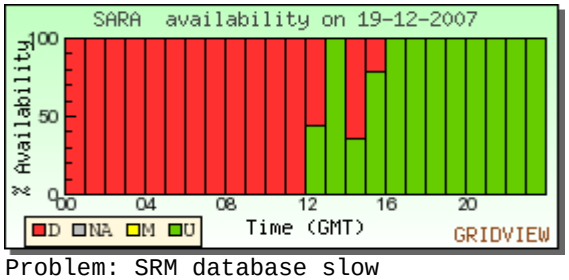

On 4 December there was a problem with the site BDII, the file system was corrupted enough that the BDII no longer worked, but not badly enough to trigger the failover mechanism. 6 December was the start of a long series of problems with Dcache; they started with too many postgres threads in the pnfs database. A kill was needed, which unfortunately resulted in a corrupt database. This started a several-day journey into the bowels of postgres, to recover the database.

⇒ **18 Dec 2007**



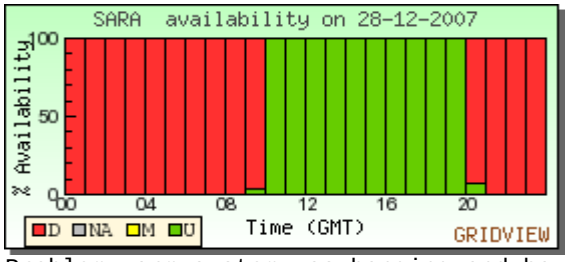

Problem: SRM database slow
Solution: Database cleanup

⇒ **19 Dec 2007**



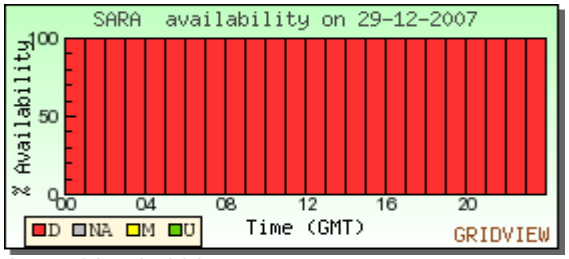

Problem: SRM database slow
Solution: Database cleanup

⇒ **28 Dec 2007**




Problem: srm system was hanging and having a very high load.
Solution: Probably the nscd daemon went crazy forking itself all the time due to a faulty configuration in /etc/nscd.conf. This has been fixed.

⇒ **29 Dec 2007**

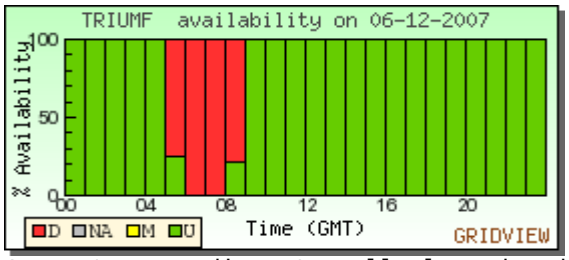



See: 28-12-2007

---

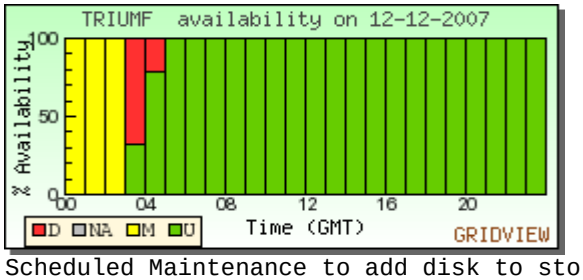
## CA-TRIUMF

⇒ **06 Dec 2007**



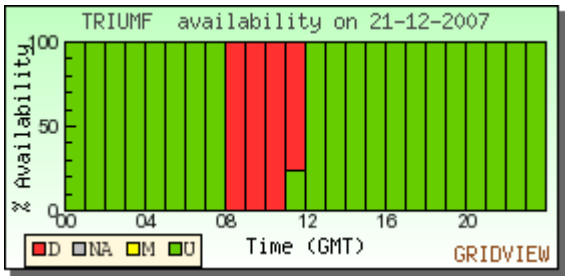

SRm not responding. On-call alarmed and did
/opt/d-cache/bin/dcache-srm stop
ps -ef |grep tomcat
kill -9
ps -ef |grep tomcat
/opt/d-cache/bin/dcache-srm start

⇒ **12 Dec 2007**

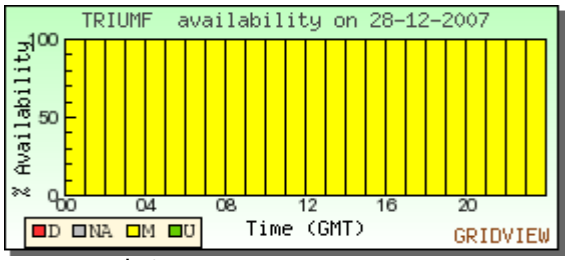



Scheduled Maintenance to add disk to storage system. Re-configred LFC mysql Db to use multiple files. SL3 updates to 3.0.9. Slight overrun on downtime, and forgot to restart FTS channels immediately.
Reduced to 1 slot per core due to Panda efficient usage. Previously 5 per 4 core due to jobs spenidn time on stage-in/out.

### ⇒ 21 Dec 2007

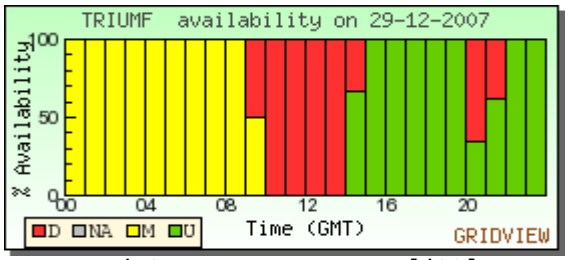



### ⇒ 28 Dec 2007




Power maintenance

### ⇒ 29 Dec 2007

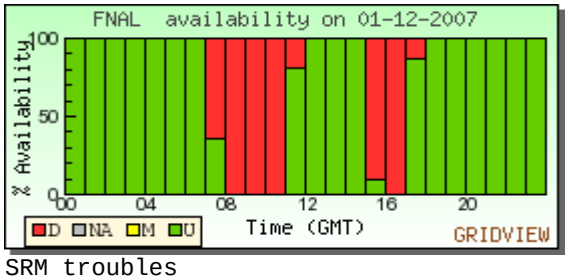



Power maintenance overran a little

---

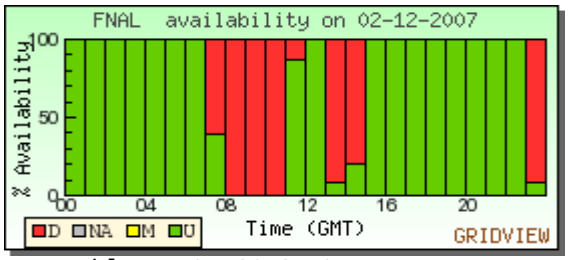
## US-FNAL-CMS

### ⇒ 01 Dec 2007



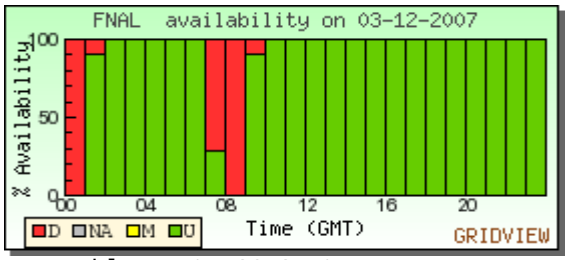

SRM troubles

⇒ **02 Dec 2007**



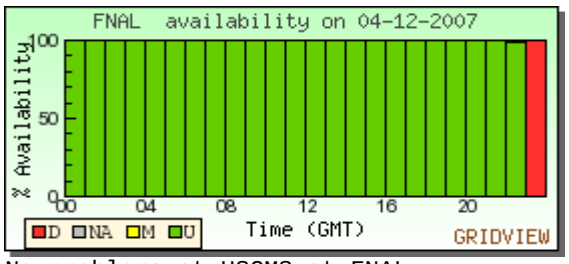

No problems at USCMS at FNAL

⇒ **03 Dec 2007**



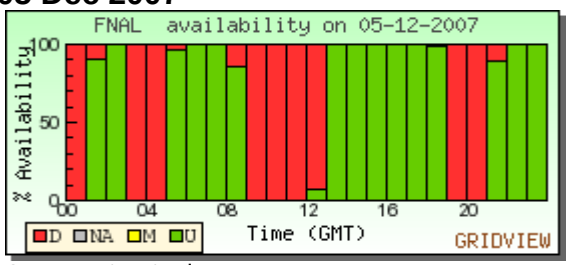

No problems at USCMS at FNAL

⇒ **04 Dec 2007**



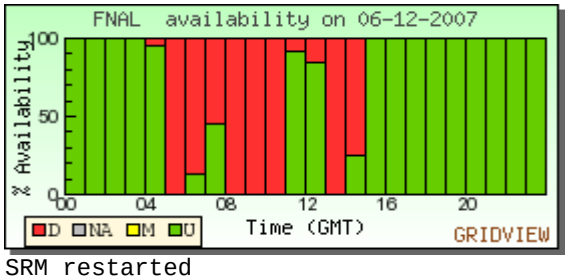

No problems at USCMS at FNAL

⇒ **05 Dec 2007**



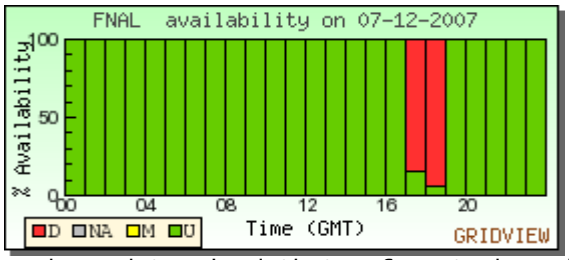

SRM restarted

⇒ **06 Dec 2007**



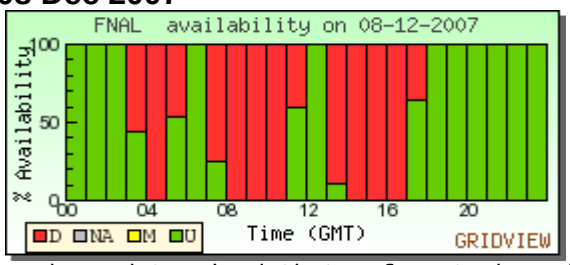

SRM restarted

⇒ **07 Dec 2007**



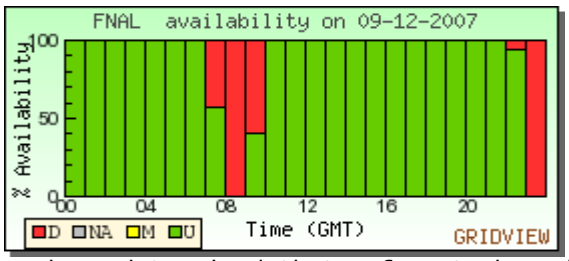

We have determined that pnfs gets huge backlogs (8000-10000) files in the pnfs manager during certain parts of the day. It continues to work, but takes more than 30 seconds to respond to srm queries. The SRM times out and fails your transfer. Other non-sam transfers retry an succeed.

⇒ **08 Dec 2007**



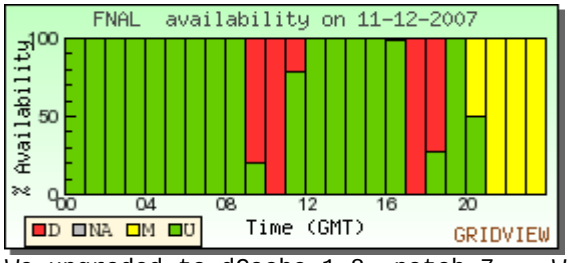

We have determined that pnfs gets huge backlogs (8000-10000) files in the pnfs manager during certain parts of the day. It continues to work, but takes more than 30 seconds to respond to srm queries. The SRM times out and fails your transfer. Other non-sam transfers retry an succeed.

⇒ **09 Dec 2007**



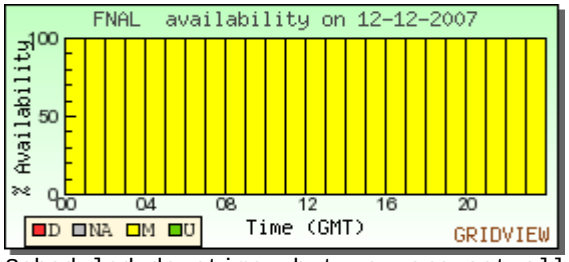

We have determined that pnfs gets huge backlogs (8000-10000) files in the pnfs manager during certain parts of the day. It continues to work, but takes more than 30 seconds to respond to srm queries. The SRM times out and fails your transfer. Other non-sam transfers retry an succeed.

⇒ **11 Dec 2007**



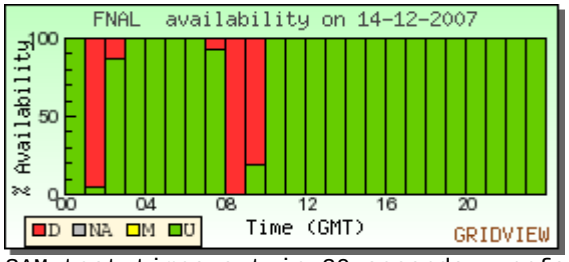

We upgraded to dCache 1.8, patch 7. Went smoothly, finished in about 4 hours, only about half-dozen minor issues. We made the system available to users, but kept the full downtime going in case we had to make changes.

## ⇒ 12 Dec 2007



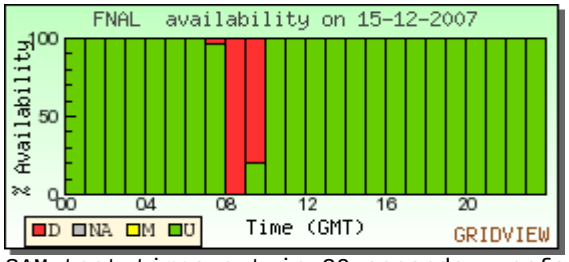

Scheduled downtime, but we were actually fully operational from the users perspective.

## ⇒ 14 Dec 2007



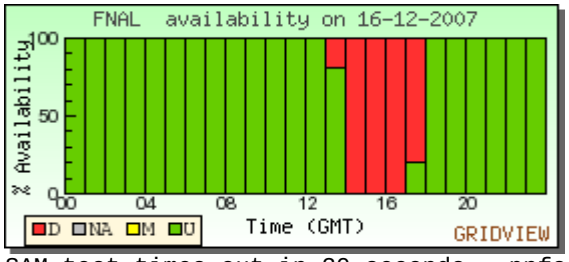

SAM test times out in 30 seconds - pnfs needs 60 seconds during busy periods.

## ⇒ 15 Dec 2007



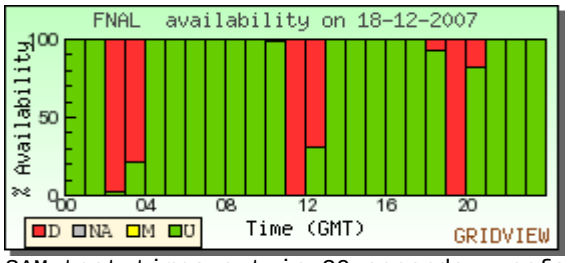

SAM test times out in 30 seconds - pnfs needs 60 seconds during busy periods.

## ⇒ 16 Dec 2007



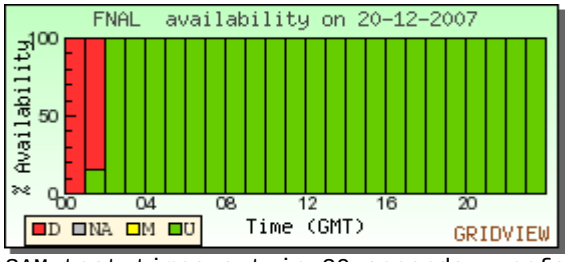

SAM test times out in 30 seconds - pnfs needs 60 seconds during busy periods.

⇒ **18 Dec 2007**



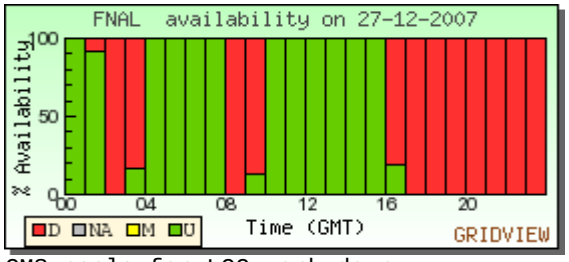

SAM test times out in 30 seconds - pnfs needs 60 seconds during busy periods.

⇒ **20 Dec 2007**



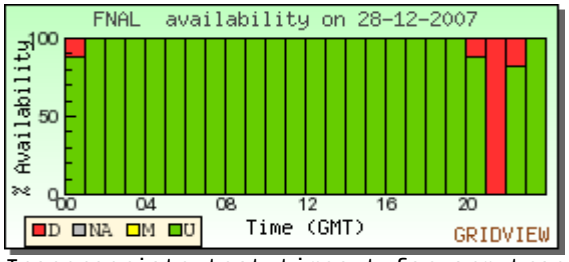

SAM test times out in 30 seconds - pnfs needs 60 seconds during busy periods.

⇒ **27 Dec 2007**



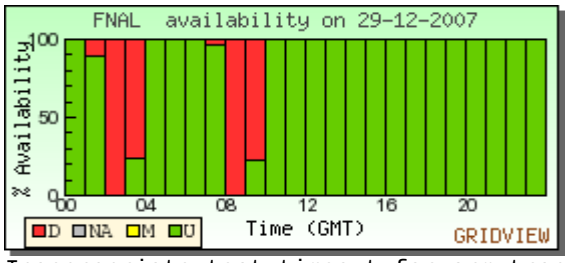

CMS pools for LCG work down
Pools for normal work were working, hence site was working

⇒ **28 Dec 2007**




Inappropriate test timeout for srm transfers during busy pnfs periods

⇒ **29 Dec 2007**




Inappropriate test timeout for srm transfers during busy pnfs periods