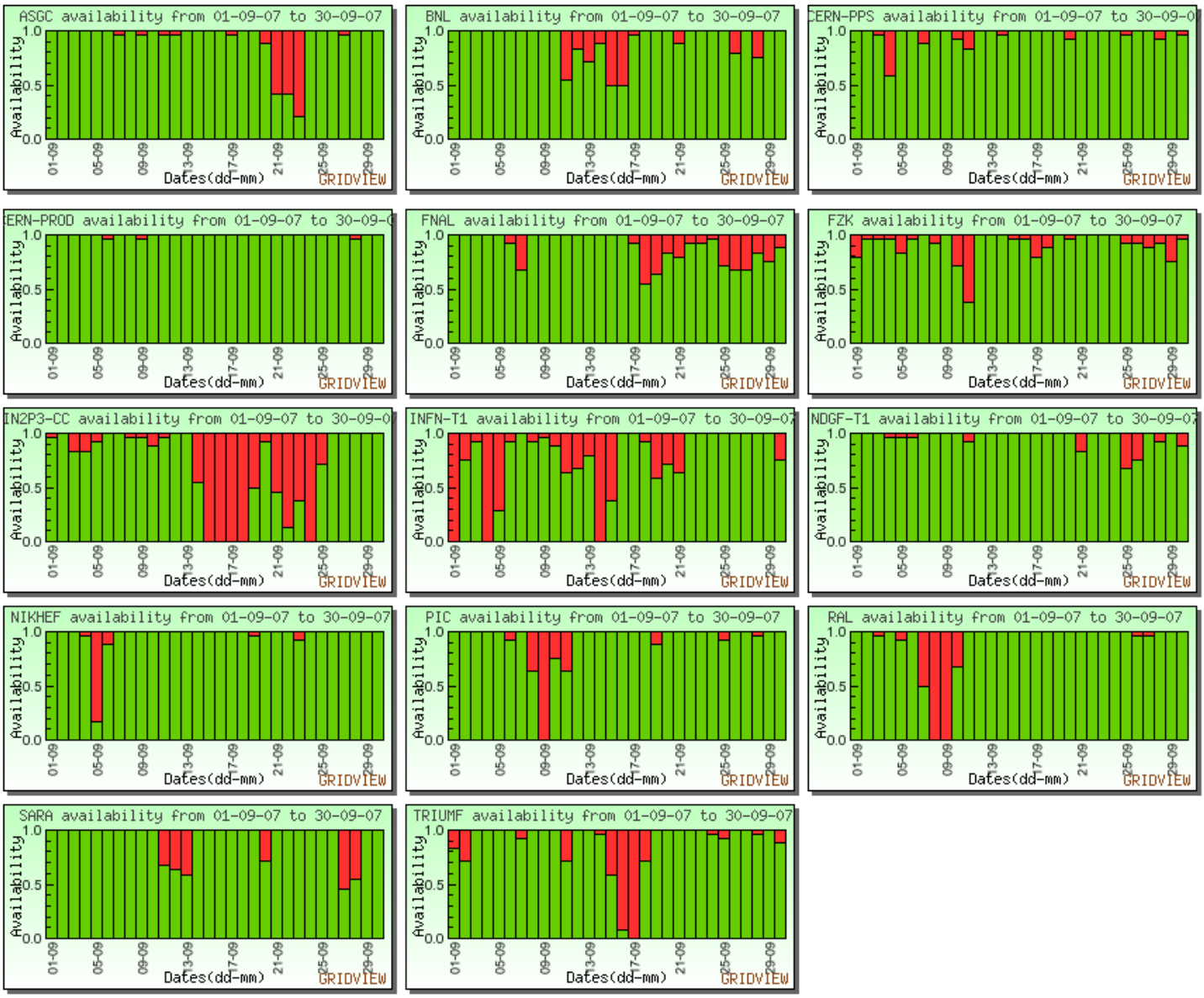
WLCG Site Reliability Reports
September 2007

- *Please review and complete the Site Reports below. Edit your section and mail the document back to A.Aimar.*
- *Deadline Monday Morning 8 October 2007.*
- *No reports from ASGC, CNAF*

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



	ASGC Taiwan-	BNL-LCG2	CNAF INFN	CERN-PROD	FNAL-USCMS	FZK-LCG2	IN2P3-CC	NDGF-T1	PIC	RAL-LCG2	SARA-MATIRX	TRIUMF-LCG2	
	93%	91%	81%	100%	89%	91%	70%	97%	93%	90%	92%	96%	
9/1/2007	100%	100%	0%	100%	100%	79%	96%	100%	100%	100%	100%	83%	9/1/2007
9/2/2007	100%	100%	75%	100%	100%	96%	100%	100%	100%	100%	100%	71%	9/2/2007
9/3/2007	100%	100%	100%	100%	100%	96%	83%	100%	100%	100%	100%	100%	9/3/2007
9/4/2007	100%	100%	100%	100%	100%	96%	83%	96%	100%	100%	100%	100%	9/4/2007
9/5/2007	100%	100%	79%	100%	100%	83%	92%	96%	100%	92%	100%	100%	9/5/2007
9/6/2007	100%	100%	92%	96%	92%	96%	100%	96%	92%	100%	100%	100%	9/6/2007
9/7/2007	96%	100%	100%	100%	67%	100%	100%	100%	100%	50%	100%	92%	9/7/2007
9/8/2007	100%	100%	92%	100%	100%	92%	96%	100%	63%	0%	100%	100%	9/8/2007
9/9/2007	96%	100%	96%	96%	100%	100%	96%	100%	0%	0%	100%	100%	9/9/2007
9/10/2007	100%	100%	88%	100%	100%	71%	92%	100%	75%	67%	100%	100%	9/10/2007
9/11/2007	100%	54%	63%	100%	100%	38%	96%	92%	63%	100%	67%	71%	9/11/2007
9/12/2007	96%	83%	67%	100%	100%	100%	100%	100%	100%	100%	63%	100%	9/12/2007
9/13/2007	100%	71%	79%	100%	100%	100%	100%	100%	100%	100%	58%	100%	9/13/2007
9/14/2007	100%	88%	0%	100%	100%	100%	54%	100%	100%	100%	100%	96%	9/14/2007
9/15/2007	100%	50%	38%	100%	100%	96%	0%	100%	100%	100%	100%	100%	9/15/2007
9/16/2007	100%	50%	100%	100%	100%	96%	0%	100%	100%	100%	100%	100%	9/16/2007
9/17/2007	96%	96%	100%	100%	92%	79%	0%	100%	100%	100%	100%	100%	9/17/2007
9/18/2007	100%	100%	92%	100%	54%	88%	0%	100%	100%	100%	100%	84%	9/18/2007
9/19/2007	100%	100%	58%	100%	63%	100%	50%	100%	100%	100%	100%	100%	9/19/2007
9/20/2007	88%	100%	71%	100%	83%	96%	92%	100%	100%	100%	71%	100%	9/20/2007
9/21/2007	42%	88%	63%	100%	79%	100%	46%	100%	100%	100%	100%	100%	9/21/2007
9/22/2007	50%	100%	100%	100%	92%	100%	13%	100%	100%	100%	100%	100%	9/22/2007
9/23/2007	42%	100%	100%	100%	92%	100%	38%	100%	100%	100%	100%	100%	9/23/2007
9/24/2007	100%	100%	100%	100%	96%	100%	0%	100%	100%	100%	100%	96%	9/24/2007
9/25/2007	100%	100%	100%	100%	71%	92%	71%	67%	92%	100%	100%	92%	9/25/2007
9/26/2007	100%	79%	100%	100%	67%	92%	100%	75%	100%	96%	100%	100%	9/26/2007
9/27/2007	96%	100%	100%	100%	67%	88%	100%	100%	100%	96%	46%	100%	9/27/2007
9/28/2007	100%	75%	100%	96%	83%	92%	100%	92%	96%	100%	54%	96%	9/28/2007
9/29/2007	100%	100%	100%	100%	75%	75%	100%	100%	100%	100%	100%	100%	9/29/2007
9/30/2007	100%	100%	75%	100%	88%	96%	100%	88%	100%	100%	100%	88%	9/30/2007

ASGC

> 20-23 Sep 2007 :

Problems: SAM job submission testing failure at site CE w-ce01.grid.sinica.edu.tw
Date: start from 15:06:51, and end at 17:19:43
Reason: the problem encountered when SAM ops job execution change the DN, before the first err encountered, say '20-Sep-2007 15:06:51', all SAM functional testing jobs are submit by Judit, but later with Piotr's DN, all lcmaps fail to redirect the mapping to proper opssgm user/group.
Severity: the severity limited to one of the site CEs, but SAM job submission testing should still pass at the other CEs. The overall impact are limited but only w-ce01.
Solution: have forced updating the gridmapfile of lcmaps that will change to single user mapping for sgm role of ops rather than pool account mapping.

Problems: Job submission failure of site CE, lcg00125.grid.sinica.edu.tw
Date: job submission - from '23-Sep-2007 05:12:16' and end at '23-Sep-2007 07:17:02', and replica management testing – from '23-Sep-2007 07:15:19' and end at '23-Sep-2007 17:07:22'
Reason: we have maintenance starting from '22-Sep-2007 22:16:32' and end at '23-Sep-2007 05:12:13', and replica management start from '23-Sep-2007 07:15:19' and end at '23-Sep-2007 17:07:22'.
Severity: all these events are related to power maintenance we have, and all of them are discovered after power recovery, the event are also related to next event stated below.
Solution: latest SAM event pass at '23-Sep-2007 18:07:08'

Problems: SAM job submission failure of site CE, w-ce01.grid.sinica.edu.tw and quanta.grid.sinica.edu.tw
Date: start from '22-Sep-2007 22:38:23' and end at '23-Sep-2007 18:24:22'
Reason: associated SAM events referring to replica management testing failure starting from '23-Sep-2007 10:11:55', and end at '23-Sep-2007 17:09:39'. The root cause is that network interface fail to startup normally after power maintenance, and the maintenance stuck with VLAN settings from the beginning, but later discover that rebooting management blade from chassis help re-enable the network interfaces for those blades passing to VLAN up linking to the other edge switch.
Severity: due to the failure of network failure, all SAM testing not able to pass due to the failure of lcg-rm, and dpm01 will be the only serving ops monitoring services.
Solution: latest SAM testing pass around the same time we fixed the dpm01 network problem and the time stamp of the event carried out around the same time with respect to the other two CEs, say '23-Sep-2007 18:09:10'

BNL-LCG2

> 11-13 Sep 2007

Tuesday + Wednesday, September 11-12
dCache not available
Cause: Scheduled network downtime for firewall maintenance

Thursday September - 13
Problem: Occasional glitches in dCache availability
Cause: HPSS upgrade and related activities
Severity: No data can be staged in and out of HPSS.
Solution: HPSS was upgraded during the week. After the HPSS upgrade was complete, the glitches went away.

> 14 Sep 2007 :

Problems: dCache has bad local account mappings for some critical users. It caused USATLAS/ATLAS data transfer failures.

Cause: We did an in-place GUMS (Grid user management system) upgrade on Thursday afternoon. Some critical users were mapped to wrong accounts due to the configuration file problem. This problem was undetected because GUMS still provided mapping service, and the generated grid map appeared to be "OK" while it was not. The symptom did not show up several hours later until midnight when dCache regenerated its map file from the GUMS server, and experienced data transfer failures thereafter. Even the GUMS update had been properly announced, we might not be able to discover this type of problem.

Cause: A fraction of USATLAS data transfer failed between midnight and 10:30AM, Friday monitoring.
Solution: Short term, we corrected the configuration file errors, and let dCache regenerate dCache map file, recovered the data transfer problem. In order to fix this problem and prevent the future occurrence, we will do two improvements: 1) The dCache team might want to consider regenerating the grid map file during prime business hour, i.e. 9:00AM, and 3:00PM. 2) We will develop Nagios probes to validate the certificate mapping for the critical users: Nurcan''''s production certificate, Hiro''''s data transfer certificates. (Please let us know any other critical certificates). Main symptom was a slower throughput to HPSS. Files kept being flushed to HPSS and stayed precious.

Cause: due to a Solaris/Linux difference, the script that stages files to HPSS is not recognizing flushed files correctly and was not working reliably on the Thumper Severity: Apart from a minor slowdown, few files got corrupted. In the case that a file was written, delete, and then rewritten with the same name, it could happen that the old copy was the one actually kept in HPSS.
Remediation: Thumper assigned only as a read node Long term solution: need to do more work to guarantee that Thumpers in the write pool work reliably

> 15 Sep 2007 :

Friday 14 - Saturday 15 Production had problems reading files.

Cause: no pool was actually assigned to production, because of the reconfiguration done prior to the HPSS upgrade Severity: USATLAS production affected
Remediation: reset the configuration dCache went down

Cause: unknown - still investigating Severity: site down
Remediation: restart dCache core servers Saturday 15 User client hang because of suspended requests

Cause: due to HPSS upgrade instability, many requests were suspended Severity: USATLAS production affected
Remediation: retry the requests Ongoing File disappear. More so during HPSS upgrade.

Cause: still unknown - user activity is primary suspect Severity: some files are lost
Remediation: manually retrieve the list of file lost, and clean up the data catalog so users do not request files that are not available

> 17 Sep 2007 :

Monday 17 - Thursday 20 Users job get stuck

Cause: too few movers available on some pool nodes Severity: one user affected
Remediation: increase the number of movers

> 19 Sep 2007 :

Monday 17 - Thursday 20 Users job get stuck

Cause: too few movers available on some pool nodes Severity: one user affected
Remediation: increase the number of movers

> 20 Sep 2007 :

Monday 17 - Thursday 20 Users job get stuck

Cause: too few movers available on some pool nodes Severity: one user affected
Remediation: increase the number of movers Friday 21 dCache was down

Cause: power failure in the facility brought down, and the UPS servicing that rack was not working properly. One of the machines in the rack was the PNFS servers. Severity: system down for less than an hour
Remediation: system restarted Long term solution: fix the UPS

> 21 Sep 2007 :

Friday 21 dCache was down

Cause: power failure in the facility brought down, and the UPS servicing that rack was not working properly. One of the machines in the rack was the PNFS servers. Severity: system down for less than an hour
Remediation: system restarted Long term solution: fix the UPS

> 22 Sep 2007 :

problem: The second BNL OSG gatekeeper was down

Cause: Authentication test failure. The error message is : authentication with the remote server failed" Impact: Gatekeeper was down for a few hours.
Solution: The problem went away after a few hours. Problem: One of USATLAS production panda monitoring system was frozen.

Cause: It has ran out of memory, this triggered alarms, Impact: There was less than half an hour outage for monitoring.
Solution: USATLAS production group reacted, the problem went away

> 24 Sep 2007 :

Monday 24 During the night, dCache was not responding

Cause: likely the high rate of prestage requests overloaded the PoolManager
Severity: system not responding
Remediation: PoolManager restarted Long term solution: submit prestage requests at a lower rate (1 per second)

> 25 Sep 2007 :

Wednesday 25 FTS down for upgrade problem: Panda monitor was un available for several minutes, 09-25-2007 04:38:21

Cause: Network maintenance caused the glitch. Impact: A few minutes outage. Not noticed by any user.
Solution: Connection was reestablished after the network maintenance was finished.

> 26 Sep 2007 :

Wednesday 26 From Wed at 4 pm till Thu 9:30 high load on PNFS (RDO01)

Cause: user activity Severity: performance severely degraded
Remediation: load went down when user activity terminated

> 28 Sep 2007

Problem: Some read pools in dCache were offline.

Cause: High load on Location Manager, still under investigation

Severity: some read pools are not accessible

Solution: Investigation.

CERN-PROD

Never below 91% target.

FZK-LCG2

> 1 Sep 2007

Some hanging gridftp doors of dCache that affected SRM response. Severity: low.

> 06 Sep 2007 :

Problems ocurrred in data management and transfer SRM had to be restarted several times this week because of memory problems. FTS did not work reliably because the machine running the underlying database became unresponsive. Both problems are under investigation. Severity for both problems: low. Transfers will pick up again after restart of systems

> 10 Sep 2007 :

SRM locked up because of memory problems. Reason is still unknown and being investigated by developers (dCache). Severity: high. All data transfers (in/out) stalled.

> 11 Sep 2007 :

Between 15:00 and 22:00 UTC GridView reports errors for SRM functionality. We have not observed any problems and transfers continued. Supposedly the GridView probes returned false signals.

> 17 Sep 2007 :

SRM lockups. Possibly caused by memory subsystem interaction with the java vm.
Severity: moderate

> 25 Sep 2007 :

lcg-rm errors because of instable gridftp doors on dcache. Severity: low

> 26 Sep 2007 :

lcg-rm errors because of instable gridftp doors on dcache. Severity: low

> 27 Sep 2007 :

lcg-rm errors because of instable gridftp doors on dcache. Severity: low

> 29 Sep 2007

lcg-rm errors because of instable gridftp doors on dcache. Severity: low

---

INFN-T1

> 1-2 Sep 2007

Cooling problems. Affected CASTOR system (and hence all srm endp-points) and farm.

> 5 Sep 2007

Scheduled intervention on the cooling system to fix a problem not addressed in the previous intervention (August 28-29). By mistake the intervention was noted on the gocdb as unscheduled (and it is not possible to modify).

> 10-15 Sep 2007

Instabilities on CASTOR. All production srm end-points affected.

> 19-21 Sep 2007

Scheduled intervention on storage to upgrade CASTOR to version 2.1.3-24 on September 19. All production srm end-points affected.

> 30 Sep 2007

Problems with 2 of the 3 LSF license servers needed for the farm and CASTOR (but only CASTOR was affected). All production srm end-points (based on CASTOR) were affected.

IN2P3-CC

> 3-4 Sep 2007

Some instabilities with the dCache-based SRM service for non-LHC experiments. This affected the score of the computing elements.

> 14-19 Sep 2007

From 14/09 to 17/09, instabilities with the dCache-based SRM service for non-LHC experiments.

From 17/09-8:00 to 19/09-12:00, CCIN2P3 was in scheduled downtime due to different maintenance operations, including electrical outage on 18/09. Computing resources were drained 24h before (17/09 at 8:00). The availability scores still do not correctly take into account the scheduled downtime of the whole site.

> 21 Sep 2007 :

6:00 - 10:00 [GMT] : AFS problems, connexions are sometimes impossible and requests to this service take a long time. Main local services have been affected. 10:00 – 11:00 [GMT] : Local information system not responding 13:00 - 14:00 [GMT] : short network outage - site unavailable form the outside 17:00 - 00:00 [GMT] : AFS problems, connexions are sometimes impossible and requests to this service take a long time. Main local services have been affected. We have migrated data on another AFS server. During this operation maintenance, jobs have

been locked in queue and CE have been unvavailable. This has affected the SRM SE also in the same time.

> 22-25 Sep 2007

From 22/09-06:30 to 22/09-19:32, CEs in downtime due to AFS problem again. All the WNs had to be restarted.

From 22/09-19:32 to 25/09: problems with two SE nodes that are still in the GOCDB but are decommissioned. They are still in the GOCDB because as a user, the current version of the GOCDB do not allow us to remove them.

NDGF-T1

> 17 Sep 2007 :

Problem: After enabling "passive pools", we started seeing "no route to host" errors in the FTS logs.
Solution: This was diagnosed as dcache not giving PORT commands in the appropriate port range for the dcsc_ku_dk firewall. "Passive pools" disabled for now.

> 19 Sep 2007 :

Problem: The pools at uio did not come back quite on time after the scheduled outaget.
Solution: Waited another 30 minutes, and they were back. Apparently an incompatibility between switch and the new 10GigE NICs, so the pool nodes were left on the old gigE.

> 25-26 Sep 2007

> 30 Sep 2007

---

pic

> 06 Sep 2007 :

The site-bdii problem (8.53am) was a problem with the network

> 8-10 Sep 2007

Date: From 8 Sep at 15:30 UTC til 10 Sep at 6:30 UTC aprox

Problem: SRM dCache not available.

Severity: High. All the transfers to/from PIC were failing because the SRM node hanged.

Solution: The SRM node was restarted on Monday morning.

> 11 Sep 2007

Date: 11 Sep from 14:30 UTC til 23:30 UTC aprox

Problem: We believe this was a false negative. All the SAM SRM tests were failing for PIC, but the services were working fine. We believe it was a SAM problem because at that time we saw other sites (eg, IFIC, LIP, CIEMAT) failing the SRM tests in the same way.

Severity: None.

Solution: None.

> 19 Sep 2007 :

Downtime for srm-disk.pic.es due to dCache intervention (connection to Enstore)

> 25 Sep 2007 :

- We failed a couple of tests due to a problem with the network. There was a problem with the router configuration and we remained for more or less one hour without network. This caused us to fail the tests on 25-09-07 18:31:34 for all of our CE''s.

> 28 Sep 2007 :

RAL-LCG2

> 05 Sep 2007 :

Problem: Network problem at RAL-LCG2
Solution: Problem was fixed by Site Networking team

> 07 Sep 2007 :

Problem: An incorrectly formatted certificate was used to replace an expiring certificate on a disk server, this lead to all gridftp transfers to this host failing, the host was only serving ops data, so no other vos were affected
Solution: The certificate was regenerated correctly and the gridftp service restarted, transfers succeeded from that point on.

> 08 Sep 2007 :

Problem: An incorrectly formatted certificate was used to replace an expiring certificate on a disk server, this lead to all gridftp transfers to this host failing, the host was only serving ops data, so no other vos were affected
Solution: The certificate was regenerated correctly and the gridftp service restarted, transfers succeeded from that point on.

> 09 Sep 2007 :

Problem: An incorrectly formatted certificate was used to replace an expiring certificate on a disk server, this lead to all gridftp transfers to this host failing, the host was only serving ops data, so no other vos were affected
Solution: The certificate was regenerated correctly and the gridftp service restarted, transfers succeeded from that point on.

> 10 Sep 2007 :

Problem: An incorrectly formatted certificate was used to replace an expiring certificate on a disk server, this lead to all gridftp transfers to this host failing, the host was only serving ops data, so no other vos were affected
Solution: The certificate was regenerated correctly and the gridftp service restarted, transfers succeeded from that point on.

> 26 Sep 2007 :

Problem: failed test CE-sft-lcg-rm on lcgce02. One of the gridftp door system hit a limit.
Solution: Restarted door system with higher limit.

SARA-LISA

> 01 Sep 2007 :

Problem: SAM test runs out of wallclocktime because the gfal_read of the SAM POSIX test times out. Because it runs out of wallclocktime the SAM is killed and the results are never published. Unable to figure out why gfal_read times out, since it works for our other clusters, no solution so far.

> 02 Sep 2007 :

Problem: SAM test runs out of wallclocktime because the gfal_read of the SAM POSIX test times out. Because it runs out of wallclocktime the SAM is killed and the results are never published. Unable to figure out why gfal_read times out, since it works for our other clusters, no solution so far.

> 03 Sep 2007 :

Problem: SAM test runs out of wallclocktime because the gfal_read of the SAM POSIX test times out. Because it runs out of wallclocktime the SAM is killed and the results are never published. Unable to figure out why gfal_read times out, since it works for our other clusters, no solution so far.

> 04 Sep 2007 :

Problem: SAM test runs out of wallclocktime because the gfal_read of the SAM POSIX test times out. Because it runs out of wallclocktime the SAM is killed and the results are never published. Unable to figure out why gfal_read times out, since it works for our other clusters, no solution so far.

> 05 Sep 2007 :

Problem: SAM test runs out of wallclocktime because the gfal_read of the SAM POSIX test times out. Because it runs out of wallclocktime the SAM is killed and the results are never published. Unable to figure out why gfal_read times out, since it works for our other clusters, no solution so far.

> 06 Sep 2007 :

Problem: SAM test runs out of wallclocktime because the gfal_read of the SAM POSIX test times out. Because it runs out of wallclocktime the SAM is killed and the results are never published. Unable to figure out why gfal_read times out, since it works for our other clusters, no solution so far.

> 31 Aug 2007 :

Problem: NAT connectivity malfunction
Solution: Network department fixed switch configuration

SARA-MATRIX

> 11 Sep 2007 :

Problem: Authentication error.
Solution: Problem went away by itself, we suspect that the automatic generation of the dcache.kpwd file did not work but at the moment we don''''''''t know why.

> 12 Sep 2007 :

Problem: lcg-cr returns "protocol not supported"
Solution: Due to a configuration error on our side the plugin and provider scripts of the information system did not work. The configuration error has been fixed.

> 13 Sep 2007 :

See 2007-09-12

> 20 Sep 2007 :

Problem: Information system (GRIS) for the srm endpoint did not work.
Solution: This problem was caused by the fabric management which caused some rpms to be uninstalled which broke the information system. we have fixed this problem.

> 27 Sep 2007 :

Problem: Power outage which was followed by substantial network problems
Solution: The network problems were fixed and several nodes were rebooted. After that it seems to work OK again but we are still monitoring the situation for irregularities.

TRIUMF-LCG2

> 01 Sep 2007 :

SRM problems - gridftp accidentally started on node without hostcert.

> 02 Sep 2007 :

SRM timeout

> 07 Sep 2007 :

replication error to pic using fzk LFC. Authorization problme with LFC.

> 11 Sep 2007 :

SAM test user cert changed to dtem but attempted ops operations. We reported the error, and compensated temporarily. The temporary fix, then broke the test once the correct credential was used. Throughout, there was no affect on prod operations. Not our fault.

> 18 Sep 2007

SAM test failures following a scheduled downtime by a few hours. The SRM/dCache service was not brought back into service in time (before the end of the scheduled downtime). We had issues in publishing into the information system.

> 30 Sep 2007

SRM/dCache service problem. Fixed by restarting a few services.

USCMS-FNAL-WC1

> 06 Sep 2007 :

trouble with downtime

> 07 Sep 2007 :

(a)Downtime extended beyond plan (b) 2nd group was false - due to 2nd FTS server registered in GOC was not in production

> 17 Sep 2007 :

Test shows SE and SRM down, but this is not true. Many ongoing transfers.

> 18 Sep 2007 :

Test shows SE and SRM down, but this is not true. Many ongoing transfers.

> 19 Sep 2007 :

Test shows SE and SRM down, but this is not true. Many ongoing transfers.

> 20 Sep 2007 :

Test shows SE and SRM down, but this is not true. Many ongoing transfers.

> 21 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.

> 22 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.

> 23 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.

> 24 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.

> 25 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.

> 26 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.

> 27 Sep 2007 :

Test reported SRM/SE not working, but our SRM/SE was working fine. Too busy to investigate false alarm this week.