Site Availability Reports
May 2007

	CERN-PROD	FZK-LCG2	IN2P3-CC	INFN-T1	RAL-LCG2	SARA-MATRIX	TRIUMF-LCG2	Taiwan-LCG2	USCMS-FNAL-WC1	PIC	BNL-LCG2
1	100%	100%	100%	100%	100%	100%	96%	100%	96%	100%	92%
2	100%	92%	100%	96%	100%	100%	96%	96%	63%	100%	96%
3	100%	88%	92%	67%	100%	100%	100%	100%	0%	100%	96%
4	100%	92%	100%	100%	92%	100%	75%	100%	0%	100%	100%
5	96%	100%	100%	100%	100%	100%	100%	100%	0%	33%	100%
6	100%	96%	100%	100%	100%	100%	92%	96%	0%	0%	100%
7	100%	92%	100%	100%	100%	100%	75%	100%	0%	63%	96%
8	88%	96%	100%	100%	100%	100%	96%	100%	0%	96%	100%
9	83%	96%	79%	100%	100%	100%	88%	100%	83%	96%	100%
10	100%	92%	88%	83%	83%	100%	67%	100%	92%	100%	100%
11	100%	92%	96%	42%	71%	100%	100%	100%	92%	96%	100%
12	100%	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%
13	100%	83%	100%	92%	96%	96%	92%	100%	100%	96%	100%
14	100%	75%	96%	96%	96%	79%	100%	100%	100%	83%	100%
15	100%	80%	100%	100%	93%	100%	100%	100%	100%	93%	100%
16	92%	83%	100%	83%	100%	100%	100%	100%	100%	100%	88%
17	100%	54%	100%	71%	100%	93%	100%	100%	100%	96%	92%
18	100%	58%	100%	92%	100%	92%	100%	100%	88%	100%	96%
19	100%	83%	100%	100%	100%	100%	100%	100%	96%	100%	96%
20	100%	71%	100%	100%	100%	100%	92%	100%	100%	100%	83%
21	96%	63%	100%	67%	100%	100%	100%	100%	100%	100%	100%
22	63%	29%	96%	92%	100%	96%	100%	100%	100%	100%	100%
23	54%	58%	100%	96%	100%	100%	100%	100%	100%	71%	100%
24	100%	92%	100%	100%	88%	100%	100%	100%	96%	54%	100%
25	83%	71%	100%	100%	0%	100%	100%	100%	100%	0%	100%
26	100%	46%	100%	88%	17%	100%	100%	92%	100%	0%	100%
27	100%	50%	100%	100%	96%	100%	100%	100%	100%	0%	96%
28	75%	100%	54%	58%	88%	100%	100%	100%	100%	46%	100%
29	58%	79%	21%	100%	79%	100%	100%	100%	83%	92%	100%
30	8%	100%	100%	21%	79%	100%	100%	100%	100%	71%	100%
31	79%	33%	100%	54%	4%	100%	88%	67%	100%	100%	100%
ability	90%	79%	94%	87%	87%	99%	95%	98%	77%	77%	98%

ASGC

> 31.05.2007

Reason: the SAM error referring generic replica management problem, that the CE-sft-rm fail in timeout problem when dealing with supertest of replica management. Further to this SAM events, we found SAM fail at ASGC due to the regular permission error, which referring to wrong acl configuration. We patch the script recommend by Maarten that will update acl on all 'geneerated' dir in all VO configured. We're then able to solve the long standing problem, that lcg-rm start faling periodically either with supertest or permission error.

Severity: except ASGC, majority of the regional centers fail with same problem, due to the acl settings, that ops monitoring jobs fail to register file into into 'generated' directory if base on other dn with same opssgm vo membership mapping. The severity is limited, since generic vo mapping wont have same acl problem, say dteam, cms, or atlas. And site dpm services dedicated for SAM testing only, the other two srm fabrics dedicated for atlas and cms wont suffering from current problem anyway.

Solution: the workaround have been discussed in rollout, and Sohpie remind that the permission error could arise from wrong acl settings. And problem resolved after patching up the acl of all generated dir.

BNL

> 16.05.2007

Problem:
A user complained slow response in accessing data out of dCache between Tuesday afternoon and Wednesday afternoon.
Cause: many AOD files were purged out from disk area to make space available for new files. Users requested to pin AOD data files into the disks. We chose a cost module to choose pools with available disk space. The cost module we used did not balance well between available disk space and CPU load, it resulted that many busy read pools were chosen to serve users data access requests.
Severity: most users will experience slow response and time out.
Solution:
We rebalanced the cost model for dCache among CPU load and available disk space.

> 20.05.2007

Problem: HPSS was offline early in the moring.
Cause: The HPSS core server process crashed.
Severity: tape storage was offline for an hour.
Solution: restarted the HPSS core server.

> 27.05.2007

Problem: the dCache Nagios script (/usr/lib/nagios/plugins/check_dcache_client.py) on acas farm made by the Linux farm group sent out false alarms while all hosts reported with errors appear perfect normal.

Cause: the script makes attempt to test the existence of dCache process during the host startup process even before the dCache component is started.

Severity: false alarms were sent to the farm group. No function and performance are affected.

Solution: disable the alarm of scripts until the farm group fixes the scripts.

GridFtp server problem:
The server appears to be powered off around 12:00, at least there is no video signal
From the remote console. The remote access controller card
appears to be malfunctioning on this server,

Cause: The system logs show that a hardware failure may have triggered the server to shutdown itself:

(May 27 11:57:07 dcdoor03 kernel: Uhhuh. NMI received for unknown reason 21 on CPU 0.
May 27 11:57:07 dcdoor03 kernel: Dazed and confused, but trying to continue
May 27 11:57:07 dcdoor03 kernel: Do you have a strange power saving mode enabled?
May 27 12:08:09 dcdoor03 kernel: Uhhuh. NMI received for unknown reason 31 on CPU 0.
May 27 12:08:09 dcdoor03 kernel: Dazed and confused, but trying to continue
May 27 12:08:09 dcdoor03 kernel: Do you have a strange power saving mode enabled?
May 27 12:18:40 dcdoor03 shutdown: shutting down for system halt
May 27 12:18:41 dcdoor03 init: Switching to run level:
)

Some mailing list posts indicate that this might be a memory problem:
http://lists.us.dell.com/pipermail/linux-poweredge/2005-January/018477.html.

Severity: 14% of dCache connectivity is lost for two hours.

Solution: reboot the server.

CERNPR

> 8-9.05.2007

08/05/2007:
- 3 h unavailability of CE (20-22), CE-sft-lcg-rm
09/05/2007:
- 4 h unavailability on CE, CE-sft-lcg-rm

An unusual high number of SRM query requests for LHCb files degraded both SRM endpoint srm.cern.ch and the Castorlhcb service. Problems started late Tuesday evening (8th May), when the machine hosting the SRM request spool developed a hardware problem (degraded RAID array).

On Wednesday morning (9th May), request spool server was replaced by its hot-standby, and normal service was restored at 11:00.

> 16.05.2007

- DATE: Wednesday 16th May
- PROBLEM: 2 h unavailability, 8-9.00 am in the CEs, with the error CE-sft-lcg-rm-cr ERROR
Cause: scheduled intervention (DB hardware upgrade) on the castor instance that supports dteam/ops transfers. See the broadcast sent in advance:

CASTORPUBLIC intervention next Wednesday, May 16
Published on : 2007-05-11

Published by David Collados (CERN)

> 22-23.05.2007

22/05: 9 h unavailability on the CEs, CE-sft-lcg-rm-cr failures
- cause: Castorpublic (MSS backend for *all* OPS SE and SRM tests)
degraded for 12 hours because of overload. Please note that
this did not affect SE availability for LHC VO's
- broadcast that day: All gLite WMS 3.[0-1] at CERN need to be rebooted

23/05: 11 h unavailability, from 12.00 till 23.00, CE-sft-lcg-rm-cr failures
- We have seen this outage, and it was investigated but no error on the back end side was found.

> 25.05.2007

25/05: 4 h unavailability in the Ces, 4-8 am, ERROR: CE-sft-lcg-rm-cr
- cause and solution: probably related to SAM unavailability

> 27-31.05.2007

From 25 May 12:00 until 29 May 13:00 - significant instability of SAM, significantly lowered frequency of tests results with rather quite low impact on availability numbers.

- 28/05: 6h unavailability, 19-24 h
- 29/05: 10 h unavailability, 0-5 and 20-24 h
- 30/05: 22 h unavailability, all day except 2-3 am
- 31/05: 5 h unavailability, 12-15 h and 21 h

CAUSE AND SOLUTION:
the unavailabilities in the last days are mainly related to performance issues with the LCG CEs. While until early this week the highest number of GRID jobs seen at CERN (running+pending) was around 12k we are passed 18k on Wednesday morning and right now we are approaching the same level again.
Our CEs cannot stand the load any more, and we just lost ce101 which shows weird HW problems and therefore is now being drained.

Moreover:
- the capapcity in SLC3 is decreasing, resulting in a lower job throughput for SLC3
- all gLite CEs are idle

Measures taken so far:

- we have 8 new CEs in the queue. Problem: useless queries of gridice daemons to the batch system which cannot be intercepted hammer the batch system. We are trying to find out if we can switch them off. Clearly, this does not improve the situation. I have put in the first one some minutes ago

- operator instructions have been updated: they are now allowed to reboot stalled CEs themself

- new actuators put in: dumping the process table of a hanging CE it is evident that it piles up gatekeeper processes. A new version of the monitoring agend (gridgris) introduced recently now also counts the number of gatekeeper processes, and restarts the gatekeeper if there are more than 25 of them. This has been rolled out last night (0h) on most CEs, and is now packaged and deployed on all of the CEs

- some tuning of LSF: allow more GRID jobs to run in parallel

We have not changed anything on the setup of our CEs since the last SW upgrade. To me these problems rather look like scalability issues of the CE software.

P.S.: see
http://lsfmon.cern.ch/lrf-lsf/queues.php?time=1&detailed=&auto_update=

CNAF

> 3.05.2007

Day: 03/05/07
Reason: instability of CASTOR services for some hours
Severity:
Solution: restart of the services

> 10-11.05.2007

Day: from 10/5/07 to 11/05/07
Reason: Problems with both CEs dedicated to LCG VOs: the LDAP service became not operational after the upgrade to the INFNGRID Special Update 21/22/23 of GLite 3.0. The problem was linked, among other things, to the configuration of the VOviews, which still requires manual intervention. The full dynamic of the problem is not completely clear since of the two redundant and identical CEs, at first (on May 9), only one of them (ce05-lcg.cr.cnaf.infn.it) failed while the other (ce06-lcg.cr.cnaf.infn.it) had the same symptoms on the following day only. It appears anyway that problems were introduced in some of the YAIM functions originally shipped (by EGEE, not by the INFN release team) with this update.
Severity: All LHC VOs were affected
Solution:

> 16-17.05.2007

Day: from 16/05/07 to 17/05/07
Reason: Problems with CASTOR: too many jobs stalled the LSF server (well known LSF plugin bug addressed in the new release to be installed). As a consequence, SAM tests for both SE and SRM were affected. Moreover a (non critical) failure, during the night of 17/05/07, on one of our core switches isolated part of the farm for some minutes. No jobs were lost.
Severity:
Solution:

> 21.05.2007

Day: 21/05/07
Reason: Problems with SRM server
Severity:
Solution: restart of the service

> 28.05.2007

Day: 28/05/07
Reason: Problems with OPS VO on SE and SRM servers (timeout).
Severity: No real services were affected.
Solution:

> 30-31.05.2007

Day: fro 30/05/07 to 31/05/07

Reason: During the night both the LDAP servers, used to grant local authorization, crashed stalling all the resources. This was caused by a very high number of requests.
Severity: Both the farm and the storage were affected.
Solution: After some hours, the servers were rebooted. A new authorization client is being deployed on wns and disk-servers in order to allow local authorization for standards accounts (e.g. root, monitoring accounts, pool accounts…) and minimize the load on LDAP servers.

FNAL

> 2-9.05.2007

Day: Several days of this week
Reason: Configuration error in information system
Severity: Caused SE tests to fail for OPS VO
Solution: Solved information system problems

> 22.05.2007

We were in an official downtime on May 24. FNAL downtimes are conducted approximately once per month

> 29.05.2007

We were operational; we believe the down time can be attributed to failures in the test

> 31.05.2007

We had a unscheduled cooling outage. I registered a downtime, You need a different color for registered downtimes.

FZK

> 11.-17.05.2007 (weekly report)

Inadvertent power cut of main administrative rack and follow-up problems later that day resulted in a significant number of failures on the CE.
Errors on CE because of previous day power cut.

> 18.-24.05.2007 (weekly report)

One of the (2) CEs locked up before the weekend. The failure was not solved till 21/5. Reason for the lockup is unknown. Procedure to detect and fix this particular failure class is being developed. The SE via the SRM became inoperative several times. Reason for failure: unknown. Course of action: unknown
Overall availability improved after the CE and SRM instabilities earlier this week although occasionally the SRM is still unresponsive.

> 3.05.2007

Day: 3.5.2007
Reason: sporadically failing replica tests with timeouts
Severity: low (temporary)
Solution: restarting of gridftp doors (and dcap door once that day)

> 13-23.05.2007

Day: 13.5.2007
Reason: mixture of short sBDII problems, sporadically failing replica tests
Severity: low
Solution: temporary

Day:14.5.2007
Reason: more BDII problems, reason seems to be have been implemented changes to DNS to except load balancing at CERN BDII.
Severity: medium
Solution: rectify DNS

Day: 15.5.2007
Reason: sporadically failing replica tests with timeouts
Severity: low, but persistent
Solution: restarting gridftp doors

Day:16.5.2007
Reason: inadvertent power cut on one rack manager
Severity: medium
Solution: power came back immediately, but followup problems later

Day:17.5.2007 (public holiday)
Reason: CEs locked up
Severity: high
Solution: responsible site admin not available (public holiday)

Day:18.5.2007 (not a working day at FZK)
Reason: CEs still locked up (one completely, the other improving over the day)
Severity: high
Solution: responsible site admin not available (not a working day)

Day:19.5.2007 (Saturday)
Reason: one CE still locked, the other working again
Severity:medium
Solution: responsible site admin not available (not a working day)

Day:20.5.2007 (Sunday)
Reason: one CE still locked, the other working again, SRM problems at the end of the day
Severity:highest
Solution: responsible site admin not available (not a working day)

Day:21.5.2007
Reason: still problems the early hours of that day
Severity: high
Solution: restarted locked CE, restarted SRM

Solution after long weekend (4 days) with very low availability: improve functioning of weekend on-duty monitoring service

Day:22.5.2007
Reason: SRM became unstable again
Severity: highest
Solution: attempt to install latest dCache patch failed, SRM restarted

Day:23.5.2007
Reason: SRM became again unstable during the night, later PNFS crashed
Severity: high
Solution: SRM restarted, reboot (PNFS)

> 25.-27.05.2007

Day: 25.05.2007
Reason: SRM problems
Severity: medium
Solution: utility domain restarted "out of memory"

Day:26.05.2007
Reason: replica management timeouts (related to SAM overload at that time or already CE slowing down?)
Severity: low
Solution: suspected SAM problem as broadcast

Day:27.05.2007
Reason: missing SAM tests on both CEs
Severity: low
Solution: inform SAM team

> 29.05.2007

Day:29.5.2007
Reason: SRM instabilities, both CEs running under very high load
Severity: medium
Solution: pool and gridftp restarts (SRM), CE under observation

> 31.05.2007

Day: 31.5.2007
Reason: CEs overloaded, reason is under investigation
Severity: high
Solution: stop queues to drain CE, plan for more powerful hardware

IN2PCC

> 27.04 to 03.05

1) Too much timeout with the new Top Regional BDII :
A lot of CE-sft-lcg-rm failures was due to Top BDII timeout. After having monitored the requests made to our regional Top BDII, indexes were added to the LDAP databases. The problem seems to be fixed now.

This is referring to CE-sft-lcg-rm failures reported during the period: from 27.04 to 03.05. We had a lot of rm failures when we started to use our own Top BDII (instead of the CERN Top BDII). But, thanks to availability computation, the numerous failures are not visible within the availability graphic. However, jobs execution was certainly quite disturbed by this random behavior.

2) CEs unavailability during 2 hours due to an update of the LRMS client. Problem was quickly identified and CEs'' jobmanager was consequently modified.

This is referring to CE-sft-job failures occurring on 02.05 during 2 hours for 2 CEs (/3 CEs). As one of our CEs didn’t fail the test, the site appeared available.

> 9-10.05.2007

* May 4th, 9th and 10th: Unexplained increase of CE-sft-job failures. It seems that sometimes RBs (rb113, rb115, rb127, ...) could not select IN2P3-CC's CEs. It sounds like a problem of CERN Top BDII request problems.
* May 9th from 14:00 to 18:00: electrical power outage.
* Atlas production job submissions:
1) As Atlas production jobs are directly submitted through globus (bypassing RB use), CE''''s /home repository is intensively used. By the way, we discovered that /home/atlas050 (mapping of production role) contained ~320000 directories and ~604000 files. In the meantime, most of Atlas production jobs were lost (atlas efficacity: 1%) even though those jobs were considered ok by site (running and ending without any detected error). We changed the mapping to use atlas048 (/home/atlas048 was unused), and the job submission efficacity grown up to 97% during the week-end.
2) A lot of Atlas job failures due to memory size exceeded. The problem comes from the fact that memory limitation are not specified by queue. Memory size is unfortunately published only once at SubCluster level, and YAIM does not allow to configure several SubClusters by CE. So, as all queues are linked to the same SubCluster, all queues are supposed to provide the same amount of memory (a maximum). By the way, a job might be submitted to a queue that does not provide sufficient memory. This is what happened with atlas jobs. We will try to hack the CE information provider to get several SubCluster, but it would be better if this becomes a YAIM feature.

> 28-29.05.2007

Day: 28-29.05.2007
Reason: A Downtime was scheduled from 28.05 to 29.05, and it seems that it was not taken into account by the Availabity metrics system.
So, we are wondering whether the delay between announcement and downtime is now taken into account, or not, for the site availability computation? It might possibly explain this "unavailability" despite the scheduled downtime.
Unfortunately, we cannot remember, neither retrieve from GOC DB, the announcement date.
Severity: Not relevant
Solution: Ask the project to clarify why our scheduled downtime was not taken into account.

PIC

> 5-7.05.2007

-SRM-tape service:
 +A problem in the central NIS service caused the castorsrm service to fail from 5-May at around 8h. The NIS problems were solved on monday 7-May at around 9h.
 +Several problems with the robotic arm in the STK tape library generated serious problems in the tape migration and recovery service (started on 4-may at 5h and extended intermitently until 11-may at 17h)

-SRM-disk service: No major isssues.

-lcg-CE service: SAM tests failrures for the lcg-CE service between the 5th May in the morning until the 7th May in the morning, due to the NIS problem affecting the castorsrm service described above. That problem caused the lcg-CE replica management SAM test to fail because we had the CloseSE configured as castorsrm. This was changed on the 7th May for two reasons: castor1 is being deprecated at PIC as tape management system, and also it makes more sense that the default storage area is a disk-based one rather than a tape one.

-Currently upgrading the configuration system to the new yaim infrastructure. We have already tested it in the PIC-PPS site and we are almost ready to deploy it in the production site.

> 14.05.2007

+Site-Bdii service: A couple of sporadic timeouts generated SAM unavailability for PIC:
 -On 14-May from 14h till 19h (3 intermitent errors)
 -On 15-May from 15h till 16h (1 isolated error)

+SRM-disk service: Some problems on the 14-May, failing SAM tests from 8:15 till 11:15. The gridftp doors started failing all put/get request. The dCache head node log file contained an abnormally large number of "java.lang.OutOfMemoryError". A restart of dCache in the head node solved the problem. The origin of it is not known (quite usual, unfortunately).

+SRM-tape service: The hardware problems with the STK robot have disappeared since the support people changed one of the robotic hands on May 11th.

+lcg-CE service: No major incidences.

> 23-28.05.2007

Day: Problem with the top-bdii service. It started on 23/05/2007 @8:00 and it was solved on 24/05/2007 @8:00.

Reason: 15h unavailability in the lcg-CE service, caused by CE-sft-lcg-rm-cr failures. The reason for this failures was a misconfiguration in the local top-bdii service, which made the local SE intermittently disappear from the top-bdii info. This problem at the top-bdii was caused by a mistake when applying a patch in the top-bdii to introduce some indexes in the LDAP DB.

Severity: Medium. The problem was actually in the top-bdii, so users seeing problems when contacting the default local top-bdii should be able to try another top-bdii and use it. The main services CE and SE were working.

Solution: The problem was solved by correctly applying the index patch in the top-bdii on /05/2007 @8:00.

--

Day: Problem with the lcg-CE service. It started on 24/05/06 at 21:00 and was solved on 28/05/06 at 8:00.

Reason: A manual change in the CEs configuration made on 24-May in the evening made all submitted jobs to fail from then on. The problem took long time to be spotted mainly due to two things: a) the jobs were failing in a quite strange way, since from the local batch scheduler the exit code was OK. The only symptom we could in the end observe was the short time taken by all the jobs to end. b) The SAM framework instabilities from 24-29 May prevented us to use the lcg-CE SAM monitoring to confirm the malfunctioning of the service (since from the local batch scheduler, no error messages appeared).

Severity: High. The lcg-CE service is a critical one, and it was unavailable during this period, since it was not possible to run any job.

Solution: The cause of the problem was spotted and solved on Monday 28/05/06 at 8:00, even if The SAM unavailability extends until 12:00 (probably due to the SAM framework problems from 25-29 May)

> 30.05.2007

Day: SAM failures in the lcg-CE service on 30/05/2007 from 17:00 until 23:00.

Reason: CE-sft-lcg-rm-cr failures. We do not understand the cause of these failures. The SRM-disk service, to which the CE-sft-lcg-rm-cr test tries to write, was working ok. The top and site bdiis were ok as well.

Severity: Low. The main services CE SE and site-bdii were working ok during this period, so we believe this could be a false alarm. It seems that only 1 SAM test was launched from 17h until 23h, so this could have fake the severity of a spurious error.

Solution: The error disappeared without any intervention.

RAL

> 4.05.2007

Date: Friday 4th May 2007
Reason: A rogue job filled the /tmp directory on one batch worker, the two SAM jobs attempted to run on that node and failed.
Severity: Major - all jobs to that node would fail, potentially draining farm.
Solution: The node was removed from the batch system shortly after this by our monitoring and the runaway job was cleaned up.

> 10-11.05.2007

Date: Thursday 10th May 2007
Reason: Castor suffered a pileup of jobs due to Atlas transfers.
Severity: Critical - all users of Castor at RAL affected
Solution: Atlas was blocked from Castor and the existing jobs were drained.

Date: Friday 11th May 2007
Reason: This downtime was was due to Castor problems at RAL, reserved space was not being freed and this meant that new writes could not succeed. Castor was taken into downtime once this problem was discovered and the Default SE was made dcache.gridpp.rl.ac.uk. Castor was returned to service on the 17th.
Severity: Critical – Castor unavailable for storing files
Solution: As a workaround, dcache.gridpp.rl.ac.uk was set to be the default SE for the Ops vo, to allow CE tests to pass, Castor was put into downtime while the Castor team deployed patches to fix the problem

> 24.05.2007

Date: Friday 24th May 2007
Reason:Very high load due to job submission caused the information provider on the CE to be unable to respond to the ldap query by
the site bdii, which lead to the CE being dropped from the information system. After investigations failed to identify a single cause for the load, the CE was rebooted.

> 25-31.05.2007

-> Remark on 2007-05-25
Problem with high load on CE. Cause unknown. No resolution yet.

-> Remark on 2007-05-26
Problem with high load on CE. Cause unknown. No resolution yet.

-> Remark on 2007-05-27
Problem with high load on CE. Cause unknown. No resolution yet.

-> Remark on 2007-05-28
Problem with high load on CE. Cause unknown. No resolution yet.

-> Remark on 2007-05-29
Problem with high load on CE. Cause unknown. No resolution yet.

-> Remark on 2007-05-30
Problem with high load on CE. Cause unknown. No resolution yet.

-> Remark on 2007-05-31
Problem with high load on CE. Cause unknown. No resolution yet.

SARA - MATRIX

> 14.05.2007

Day: 14.05.2007
Reason: migration of DNS servers was not transparent
Severity: reverse lookups failed causing numerous failures
Solution: migration was completed.

> 16-17.05.2007

Day: 16 and 17 may 2007
Reason: two reasons : first an internal network problem cut off access to the user database, which resulted in failing file transfers. Secondly there were disk problems with the Oracle server
Severity: failing file transfers and reduced availability for Oracle-based services
Solution: fix the disk problem, and set up a 2nd user DB server until the network problems were solved.

TRIUMF

> 4.05.2007

Day: May 4 2007
Reason: SRM service choked, port 8443 not listening anymore
Severity: critical, no access to storage resources
Solution: restart of SRM service only on our dCache

> 7.05.2007

Day: May 7 2007
Reason: Globus gatekeeper crashed on LCG CE, perhaps due to high WMS load (tests)
Severity: critical, no access for jobs to the CE / CPU resources.
Solution: service restarted

> 9-10.05.2007

Day:
Reason: The failures on May 9-10 2007 are also due to our CE Gatekeeper connection problem (similar to May 7). We've added a cron job to check for the service as a preventive method.

Severity:
Solution:

> 31.05.2007

Day: May 31 2007
Reason: Grid Canada CRL expired
Severity: critical, no access to Canadian Grid resources
Solution: update of CRL from certificate authority (automatic web update failed)

Grid Canada CRL expired at May 31 21:37:32 2007 GMT. All users and services with GC credential were dead because of this. GC was notified by mail and voicemail but It was past 5pm in Ottawa and we did not have 24hr contact info - we will remedy this. Luckliy got hold of GC person and problem was fixed. It was due to the publish step to the webserver failing for last 2 weeks. We now check crl and will get a warning 2 weeks prior to expiry. Problem fixed within 3hrs and service crl updated manually, but clients could suffer up to 6hrs more until cron updates.

SAM unavailability

Fom **24 May 14:00 until 25 May 12:00** - major unavailability of SAM, which effectively means very low frequency of results or no results at all. This could make availability calculation (GridView,

SLS) not reliable (due to outdated results).

From **25 May 12:00 until 29 May 13:00** - significant instability of SAM, significantly lowered frequency of tests results with rather quite low impact on availability numbers.

(Note: SAM results were always reliable in the sense that whatever was published was always the correct result. The unavailability of SAM means rather that a fraction of results were not published at all, which effectively means that the frequency of results got degraded.)

Between **30 May, 2007 15:00 and 17:00** due to certificate expiration. Apart from that we still experienced occasional unavailability periods (~10 minutes). Reason seems to be a misconfiguration of the automatic firewall update script. Temporary fix is provided (so SAM is stable since), but we still investigate the issue.