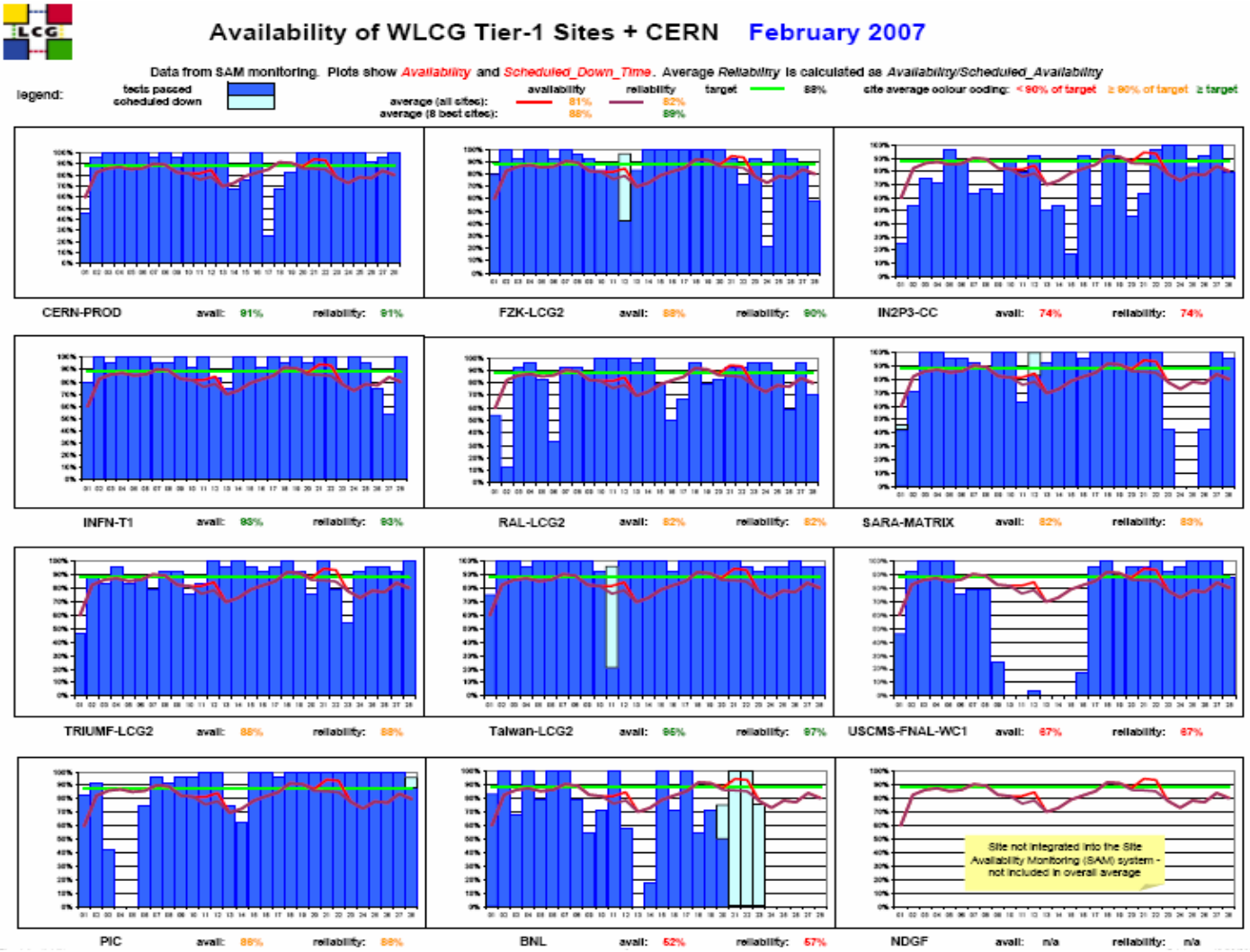
LCG Site Reliability Reports – February 2007

	CERN-PROD	FZK-LCG2	IN2P3-CC	INFN-T1	RAL-LCG2	SARA-MATRIX	TRIUMF-LCG2	Taiwan-LCG2	USCMS-FNAL-WC1	PIC	BNL-LCG2
1	46%	79%	25%	79%	54%	42%	46%	75%	46%	83%	83%
2	96%	100%	54%	100%	13%	71%	88%	100%	92%	92%	100%
3	100%	92%	75%	96%	92%	100%	83%	100%	100%	42%	67%
4	100%	100%	71%	100%	96%	100%	96%	96%	100%	0%	100%
5	100%	100%	96%	100%	83%	96%	83%	100%	100%	0%	79%
6	100%	92%	88%	100%	33%	96%	88%	100%	75%	75%	100%
7	96%	100%	63%	96%	92%	92%	79%	100%	79%	96%	100%
8	100%	96%	67%	96%	92%	88%	92%	100%	79%	92%	79%
9	96%	92%	63%	100%	88%	100%	92%	100%	25%	96%	54%
10	100%	83%	88%	92%	100%	100%	75%	92%	0%	96%	71%
11	100%	88%	79%	100%	100%	63%	83%	21%	0%	100%	100%
12	100%	42%	92%	83%	100%	83%	100%	100%	4%	100%	58%
13	100%	83%	50%	75%	96%	92%	96%	100%	0%	75%	0%
14	67%	100%	54%	100%	100%	100%	100%	100%	0%	63%	17%
15	75%	100%	17%	100%	79%	100%	96%	100%	0%	100%	100%
16	100%	100%	92%	92%	50%	96%	92%	92%	17%	100%	71%
17	25%	100%	54%	100%	67%	100%	96%	100%	96%	96%	100%
18	67%	100%	96%	96%	96%	100%	100%	100%	100%	100%	54%
19	83%	100%	88%	100%	79%	100%	92%	100%	88%	100%	71%
20	100%	100%	46%	96%	83%	100%	75%	100%	96%	100%	50%
21	100%	92%	63%	100%	92%	100%	100%	100%	96%	100%	0%
22	100%	71%	96%	100%	92%	100%	79%	96%	100%	100%	0%
23	100%	92%	100%	88%	96%	42%	54%	92%	92%	100%	0%
24	100%	21%	100%	100%	96%	0%	92%	96%	96%	100%	0%
25	100%	100%	83%	96%	88%	0%	96%	96%	100%	100%	0%
26	92%	92%	92%	75%	58%	42%	96%	100%	100%	100%	0%
27	96%	88%	100%	54%	96%	100%	92%	96%	100%	100%	0%
28	100%	58%	79%	100%	71%	96%	100%	96%	88%	88%	0%
ility	91%	88%	74%	93%	82%	82%	88%	95%	67%	86%	52%



1 Sites Reliability and Availability Reports

1.1 ASGC – J.Shih

* Feb 24, 2007
 * 9 job submission error cause by nfs failure as well as maui daemon down. extra cron job have been add to make sure batch service functional all the time. sam functional testing jobs start passing at '24-Feb-2007 14:05:43' (time stamp associated are: 13:05:22, 11:05:18, 10:05:38, 09:05:12, 07:05:20, 05:05:19, 03:05:22, and 01:05:31)
 * HA solution of nfs have been proposed, and will be applied soon in last Q1.

Feb 26, 2007
 * replica management testing error observed, time stamp associated is 05:15:36.
 * sam event: 'lcg_cr: Communication error on send', the error arise from missing se info from site giis, and gstat also showing critical error message for the missing site resources from infosys. problem resolved after info sys entries back to normal, and root cause of this error remain unclear.

1.2 CC-IN2P3 – F.Hernandez

the average reliability of the site in February 2007 was 74%. According to GridView plots, the overall service availability of the site is dominated by the availability of the dCache-backed SRM service.

Besides installing the patches as recommended by the dCache/SRM experts, some other actions have been and will be taken: the main SRM server was replaced by a more powerful machine and the PNFS server will be replaced next tuesday March 20th.

The redundancy of the computing elements (3 hosts) allowed us to maintain a near 100% overall availability of this service, in spite of one of them being completely overloaded on February 1st due to a misconfigured resource broker. It took some time to diagnose the source of the problem with the prompt help of Maarten Litmaath: thanks to him.

We will improve the monitoring of the results of the SAM tests. As I suggested in several of my previous monthly reports, it is highly desirable that the error messages produced by the tests be more explicit and give some clues on where the problems may be. For instance, a recurrent failed test of a 'lcg-cr' command produces as the only message:

   "Timeout when executing test SRM-put after 600 seconds!"

I would like to be sure that this work on the sites for producing monthly reports is really helpful for improving the monitoring tools.

1.3 CERN – T.Cass

- An unalarmed condition caused the CERN CASTOR SRM v11 to be down leading to the SAM test failure for 17th-19th. The monitoring has been improved to catch the failure condition and we have also improved the configuration to prevent this problem.

- Other major unavailability periods correlate with extremely high rates of queries to the CEs from middleware, most notably from two misconfigured test systems for the 14th/15th.

1.4 GridKa/FZK – H.Marten

Days in February with GridKa availability < 88%

01.2. 79% some intermittant rm failures because of timeouts (CERN BDII?)

10.2. 83% short SRM instability

12.2. 42% Scheduled downtime for hardware and software maintenance

13.2. 83% some intermittant rm failures because of timeouts; unclear reason

22.1. 71% SRM/SE connectivity was lost for a few hours due ti problems with data base of dCache (vaccuum full on pnfs db takes inordinate amount of resources and locks the db for the time of the vacuum)

24.2. 21% SRM part of dCache became unresponsive for many hours

28.2. 58% SRM part of dCache became unresponsive for many hours

The SRM instabilities showed up at the end of the month after the experiments started to put heavy load on the system, and they continued in March. In the mean time, the SRM/dCache team provided a patch that seems to have significantly improved the situation. Many thanks for this (although we certainly shouldn't celebrate too early).

1.5 INFN/CNAF – L.Dell'Agnello

the average availability and reliability of the site in February 2007 were both 93%.

For all the days with availability "below the green line" (01/02, 12-13/02 and 26-27/02) the problems, reported as related to srm, were due to overload of our CASTOR instance (mainly for occasional shortage of disk space in pools).

1.6 NDGF

1.7 PIC – G.Merino

The main contribution to the PIC unreliability in the month of February (about 10% monthly average) comes from the CE service. On the weekend of the 3,4 of February the only CE head node that we had at PIC got completely flooded with jobs coming from MCprod activities (mainly atlas and lhcb). This caused an extremely high load on this machine (around 120 seen) which effectively broke the service.
The recovery of this situation was very painful, and it took essentially all of the next monday (5-Feb) for the CE responsible to get it back to live. This is due to the fact that, even killing processes and rebooting the machine, the jobmanager processes re-appear when RBs retry contacting the CE. At the end, the only way to stabilise the machine turns out to be banning the authentication of those users causing the highest load for some days, which is of course a not-acceptable operation procedure (but the only one we have found).
Few days later on that same week, two extra CE head nodes were deployed in order to better load balance the service and have more room for dealing with problems appearing in any of them. Since then, the availability of the service has greatly improved.

1.8 RAL – J.Gordon

The reduction in availability on the 1st and 2nd of the month had several causes; there were a number of BDII timeouts, but the SAM tests also show a significant number of Unspecified Gridmanager errors. When we have investigated these jobs we found that the job is successfully submitted to the batch system but is then very shortly afterwards dequeued at the request of the CE, a GGUS ticket has been raised about this issue (18603). It appears that the job is in an unexpected state when queried by the CE causing the CE to ask for the job to be removed.

The SAM CE ops test show a large number of BDII timeouts between the 5th and the 7th, accounting for the reduction in availability over that period.

The reduction in availability over the period between the 14th and the 20th is due to a combination of BDII timeouts and also the RB unable to submit jobs to our CE. We discovered that the CE had fallen out the information system due to the scheduler daemon being sufficiently slow responding to the information plugin query for the list of queues that the request was timing out. The problem began occurring on the 16th, abated to some extent on the 18th, before reoccurring on the 19th. The scheduler was restarted on the 20th and the query was then able to successfully retrieve the information and the CE reappeared in the information system.

The reduction in availability through the 25th and 26th was due to a large logfile on the Castor SRM used for the OPS Computing Element tests causing

Replica Management test failures. Rotation of the log files has been improved to avoid this occuring again.

The reduction in availability on the 28th was due to the gatekeeper process on the Computing Elelemnt crashing, additional monitoring has been put in place to catch this more quickly.

On the 19th of February we upgraded our top-level BDII from a single system to two systems behind a round-robin DNS alias. After this, the incidence of BDII timeouts due to our top-level BDII has been much reduced.

1.9 SARA-NIKHEF – J.Templon

Sorry for the late report. The SARA reliability for february is reported as 83%, and the availability is 82%. the downtime is primarily from the following periods:

```
 1/2   44%   shared by almost all Tier-1s, probably a SAM problem
 2/2   71%   CE problem
11/2   63%   CE + SE problem
23/2   42%   CE problem
24/2    0%   CE problem
25/2    0%   CE problem
26/2   42%   CE problem
```

The SARA folks are looking through the records to find the source of these problems, but it's worth mentioning that the CE problem is a non-problem, since for the NL Tier-1 all the dedicated LHC resources are at NIKHEF. Unfortunately SAM and Gridview are not yet able to deal with this.

Removing the "CE" problem periods, the reliability goes up to about 96%. This is assuming that the downtime on 1/2 is not a SAM problem, and also that the SE problem on 11/2 is serious enough that SAM would decide SARA was down.

I will send the explanations as soon as they are available.

1.10 TRIUMF – R.Tafirout

The reliability for TRIUMF in February was reasonably good at 88%.
Most of the failures have been due to our SRM being overloaded at times and the CE (mostly in early February and the majority related to replica management tests with BDII Connection Timeout with lcg-bdii.cern.ch).

In the beginning of February we upgraded our dCache version from 1.6.6-5 to 1.7.0 (which was a major upgrade). That dCache version is known to have some SRM transactions inefficiencies leading to very slow responses of the system forcing an SRM service restart. Last week (March 22) we've upgraded to a new patch level which is supposed to fix these SRM issues.

In February we were still using CERN's top level bdii so the CE failures in early February could be due to that. Since mid-March we are now using our own top level BDII service.

1.11 US ATLAS/BNL – M.Ernst

Since the SRM/dCache installation at BNL was upgraded to version 1.7.0 on 31 January stability problems were observed at the SRM and GridFTP door level on a regular basis. This is the main reason that has contributed to reduced availability throughout the month. ATLAS driven activities regarding Production, AOD replication and Tier-0-Tier-1 distribution tests have increased the number of transfer requests to a level the SRM server implementation was unable to cope with. Out-of-Memory situations in the related JVM were observed followed by the SRM server becoming unresponsive. The load factor of the machine hosting the SRM server (and the underlying DB maintaining the transfer state) was observed at 15 (average) increasing at times to >25.

During the second week of March the problem above was intensely studied by the SRM/dCache developers in close cooperation with the dCache team at BNL. Patches were provided and the SRM transfer state DB was offloaded to a separate host. Both measures have improved the stability and the performance significantly. All activities mentioned above can now be handled by the BNL SE at an excellent stability level and at the expected performance. Code

improvements as they were developed during the second week were provided to other Tier-1 centers (e.g. GridKa).

The ATLAS Linux farm was upgraded to SL4 on 21 February. Also Condor was upgraded to patch a critical bug in the current version used by the USATLAS Condor pool. The farm OS upgrade along with the installation of the Condor patch broke BNL's LCG CE.

Therefore tests associated with the SAM test suite failed since then for the following reason.

Following the upgrade of BNL's worker nodes to SL4, job submission through our lcg-CE continued to work, but the replica management commands began failing. We installed and configured the latest relocatable tarball (Feb 19) version of the "Combined gLite and LCG standard Worker Node".

After installing and configuring this, job submission no longer appeared to work. (The replica management commands, however, did work when run manually.) We did some initial troubleshooting but didn't have any success. While it may be the case that we could eventually get this configuration working, we decided not to pursue the matter for two reasons:

1) The EGEE/gLite middleware does not officially support Scientific Linux 4 yet. According to the CERN EGEE Twiki, the plan is to support SL4 (and only SL4) in gLite 3.1. gLite 3.0 officially supports only SL3.

https://twiki.cern.ch/twiki/bin/view/EGEE/SL4Planning

2) The initial gLite CE release did not officially support Condor as a Local Resource Management System (LRMS) so we stayed with the lcg-CE (version 3.0.9) already installed. While the glite-CE (supposedly) does now offer support for Condor (as of mid-January) we would need to install it on an SL3 server. This would then leave us needing to do an entirely new OS+CE installation when the SL4-compatible glite-CE is released.

Our current intention is to await full SL4 and Condor LRMS support in gLite 3.1.

1.12 US CMS/FNAL – I.Fisk

During the month of February the availability of the Tier-1 center in the US had two periods outside the target ranges.

The first is an extension of the hardware instability on the gatekeeper that was observed in January, which was fixed in the beginning of February.

The second is related to csh. The cluster was upgraded during a scheduled downtime and all systems were reinstalled for security and kernel patches. During this upgrade csh was mis-configured by the cluster management software. Grid workflows from CMS typically use the bash shell, and the problem was only observed in the SAM testing. csh was properly installed and the cluster has had high availability since the fix.