May 2006, Jamie Shiers

Review of Tier0 and Tier1 Site Monitoring and Operation In Service Challenge 4

Change Log

Changes since the Management Board discussion on May 16 are high-lighted in yellow. In addition, the colour scheme of table 1 has been modified such that:

1. Green – always meets targets;
2. Bright green – usually meets targets;
3. Yellow – sometimes meets targets;
4. Red – rarely meets targets.

This is to emphasise that – given the nature of the Grid with its inherent need for fault-tolerance and recovery – having all sites ***"usually meeting targets"*** can be considered success.

Executive Summary

This document reviews the state of site monitoring and operations in the Tier0 – Tier1 disk – disk and disk – tape Throughput Tests carried out in April 2006.

The focus of this report is not on the transfer rates achieved, but on the production-readiness of the support infrastructure at the various sites – including the Tier0 – as well as the ability of sites to monitor and respond to problems and general operational procedures.

We highlight in particular the following critical issues:

1. Several sites took a long time to ramp up to the performance levels required, despite having taken part in a similar test during January. This appears to indicate that the data transfer service is not yet integrated in the normal site operation;
2. Monitoring of data rates to tape at the Tier1 sites is not provided at many of the sites, neither 'real-time' nor after-the-event reporting. This is considered to be a major hole in offering services at the required level for LHC data taking;
3. Sites regularly fail to detect problems with transfers terminating at that site – these are often picked up by manual monitoring of the transfers at the CERN end. This manual monitoring has been provided on an exceptional basis 16 x 7 during much of SC4 – this is not sustainable in the medium to long term;
4. Service interventions of some hours up to two days during the service challenges have occurred regularly and are expected to be a part of life, i.e. it must be assumed that these will occur during LHC data taking and thus sufficient capacity to recover rapidly from backlogs from corresponding scheduled downtimes needs to be demonstrated;
5. Reporting of operational problems – both on a daily and weekly basis – is weak and inconsistent. In order to run an effective distributed service these aspects must be improved considerably in the immediate future.

Recommendations

6. All sites should provide a schedule for implementing monitoring of data rates to input disk buffer and to tape. This monitoring information should be published so that it can be viewed by the COD, the service support teams and the corresponding VO support teams. (See June internal review of LCG Services.)
7. Sites should provide a schedule for implementing monitoring of the basic services involved in acceptance of data from the Tier0. This includes the local hardware infrastructure as well as the data management and relevant grid services, and should provide alarms as necessary to initiate corrective action. (See June internal review of LCG Services.) [Action – J. Shiers: follow-up with I. Bird regarding service / operational requirements on middleware components.] [3,4]
8. A procedure for announcing scheduled interventions has been prepared and is pending Management Board approval (since agreed with small changes). [9]
9. All sites should maintain a daily operational log – visible to the partners listed above – and submit a weekly report covering all main operational issues to the weekly operations hand-over meeting. It is essential that these logs report issues in a complete and open way – including reporting of human errors – and are not 'sanitised'. Representation at the weekly meeting on a regular basis is also required. (No strong support – particularly for daily logs. Concerns on repeated reporting – can this be done once and fed to all interested parties? Further discussion at joint operations workshop in June? (With usual caveat that there is no time for any non-trivial development prior to WLCG service in October))
10. Recovery from scheduled downtimes of individual Tier1 sites for both short (~4 hour) and long (~48 hour) interventions at full nominal data rates needs to be demonstrated. Recovery from scheduled downtimes of the Tier0 – and thus affecting transfers to all Tier1s – up to a minimum of 8 hours must also be demonstrated. A plan for demonstrating this capability should be developed in the Service Coordination meeting before the end of May. (Accepted.)
11. Continuous low-priority transfers between the Tier0 and Tier1s must take place to exercise the service permanently and to iron out the remaining service issues. These transfers need to be run as part of the service, with production-level monitoring, alarms and procedures, and not as a "special effort" by individuals. (Accepted.)

Metrics

In order to measure site production readiness, we propose the following metrics:

12. Ability to ramp-up to nominal data rates – see results of SC4 disk – disk transfers [2];
13. Stability of transfer services – see table 1 below;
14. Submission of weekly operations report (with appropriate reporting level);
15. Attendance at weekly operations meeting;

16. Implementation of site monitoring and daily operations log;
17. Handling of scheduled and unscheduled interventions with respect to procedure proposed to LCG Management Board.
18. Four service levels are defined:
    1. Excellent – consistently meets targets;
    2. Good – normally meets targets;
    3. Average – sometimes meets targets;
    4. Poor – rarely meets targets.

In the table below tentative service levels are given, based on the experience in April 2006. It is proposed that each site checks these assessments and provides corrections as appropriate and that these are then reviewed on a site-by-site basis. These metrics will be measured regularly and reported to the Management Board, with a clear goal that all sites should reach "excellent – consistently meets targets" (or good with demonstrated ability to recover?) prior to the end of the SC4 service phase in September 2006.

The importance of each column is certainly not uniform. Some issues – such as weekly reporting and scheduled interventions – are relatively trivial to fix (but not without effort). On the other hand, ramp-up, stability and monitoring / operations are likely to require non-negligible work to resolve.

Site	Ramp-up	Stability	Weekly Report	Weekly Meeting	Monitoring / Operations	Interventions	Average
CERN	2-3	2	3	1	2	1	2
ASGC	4	4	2	3	4	3	3
TRIUMF	1	1	4	2	1-2	1	2
FNAL	2	3	4	1	2	3	2.5
BNL	2	1-2	4	1	2	2	2
NDGF	4	4	4	4	4	2	3.5
PIC	2	3	3	1	4	3	3
RAL	2	2	1-2	1	2	2	2
SARA	2	2	3	2	3	3	2.5
CNAF	3	3	1	2	3	3	2.5
IN2P3	2	2	4	2	2	2	2.5
FZK	3	3	2	2	3	3	3

Table 1 - Summary of Site Production Readiness from SC4 Disk-Disk Throughput Phase

In the above table, non-EGEE sites are somewhat unfairly treated, as reporting procedures for such sites are still not fully established / agreed. Similarly, the

reporting procedure for Northern European sites (NDGF/SARA) needs to be clarified. However, in both cases this still highlights an area of concern for WLCG as a whole. In any event, the un-weighted average, as shown in the last column, is almost certainly too simplistic a measure of the overall state of site production-readiness. Finally, this table does not cover other highly commendable work – such as the excellent report on SC4 produced by IN2P3.

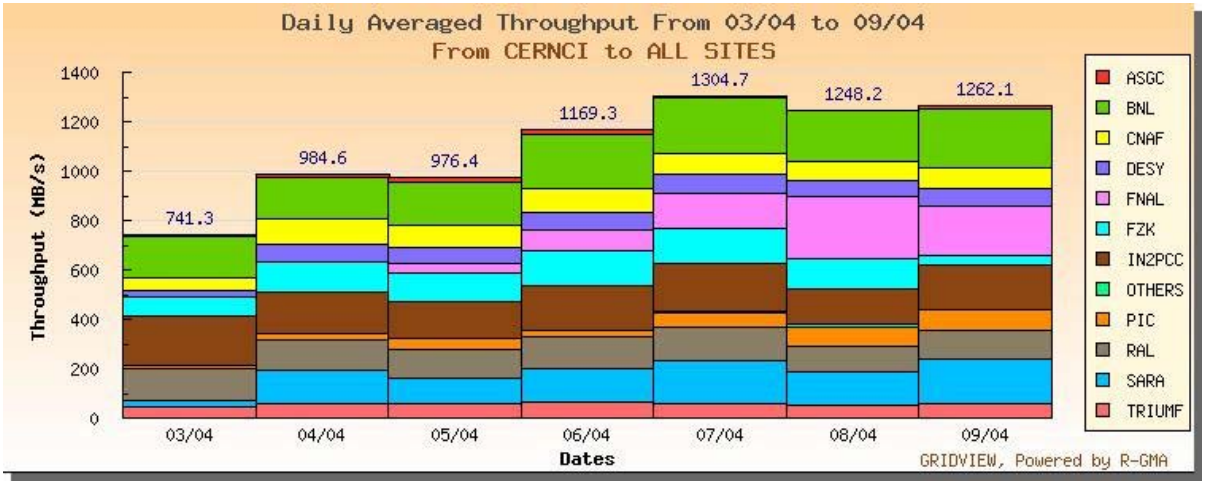

Figure 1 - Data Transfer Rates during Week 1 of S4 Disk - Disk Throughput Tests

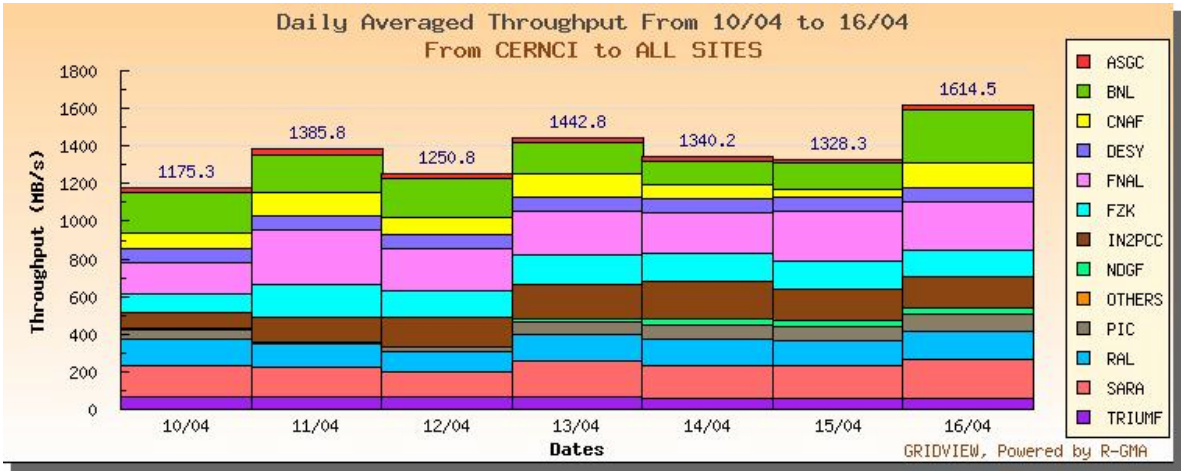

Figure 2 - Data Transfer Rates during Week 1 of SC4 Disk - Disk Throughput Tests

Site/Date	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Av. (Nom.)
ASGC	0	7	23	23	0	0	12	22	33	25	26	21	19	22	**17(100)**
TRIUMF	44	42	55	62	56	55	61	62	69	63	63	60	60	62	**58(50)**
FNAL	0	0	38	80	145	247	198	168	289	224	159	218	269	258	**164(200)**
BNL	170	103	173	218	227	205	239	220	199	204	168	122	139	284	**191(200)**
NDGF	0	0	0	0	0	14	0	0	0	0	14	38	32	35	**10(50)**
PIC	0	18	41	22	58	75	80	49	0	24	72	76	75	84	**48(100[1])**
RAL	129	86	117	128	137	109	117	137	124	106	142	139	131	151	**125(150)**
SARA	30	78	106	140	176	130	179	173	158	135	190	170	175	206	**146(150)**
CNAF	55	71	92	95	83	80	81	82	121	96	123	77	44	132	**88(200)**
IN2P3	200	114	148	179	193	137	182	86	133	157	183	193	167	166	**160(200)**
FZK	81	80	118	142	140	127	38	97	174	141	159	152	144	139	**124(200)**

Table 2 - Summary of Achieved Transfer Rates (MB/s)

Detailed Observations on Transfers (Maarten Litmaath)

ASGC

Still running CASTOR-1. Experimented with various upgrades to get improved performance per node: kernel from 2.4 to 2.6 for TCP,

ext3 to XFS file system, newer drivers for RAID controllers, TCP buffer and window sizes, latest CASTOR gridftpd, etc.

This did not converge in time, and in the end various components had to be downgraded again, after which stable running at 120 MB/s to disk was achieved for a few days, but followed by instabilities that are not all understood yet. The tests and investigations are useful preparations for the CASTOR-2 upgrade foreseen to take place in a few months.

BNL

Changed their dCache setup to allow for stable running at nominal rates using 3rd party transfers (i.e. deploying a sufficient number of powerful gridftp door nodes), not relying on srmCopy. SC4 exercise exposed a few dCache issues/bugs, e.g. a single stuck transfer can block all others, and recursive PNFS listings can cause a DoS.

In the end, with occasional admin interventions, met both disk-disk and disk-tape nominal rates for many days.

CNAF

Had a mixed CASTOR-1/-2 setup during April, suffering a lot from problems fixed in later CASTOR-2 versions. Upgraded to the latest release at the end of April and were finally able to run at 200 MB/s disk-disk for many hours in a row, but followed by new instabilities.

[1] The agreed target for PIC is 60MB/s, pending the availability of their 10Gb/s link to CERN.

DESY

An excellent reference site. During April their rate was limited to a constant 70 MB/s disk-disk by their network. Then they upgraded their uplink from 1 to 10 Gbps, which boosted their rate to a constant 170 MB/s.

FNAL

At this time the only site that ran with srmCopy, exposing a few bugs in the FTS that subsequently got fixed. Reached the highest rates of all sites, peaking at 450 MB/s, but needing a large number (70-160) of concurrent transfers and a large number (20) of streams per file.

They are using some thirty-odd ordinary machines to receive the data.

Easily reached stable nominal rates for disk-disk and disk-tape, but also suffered from a recursive PNFS listing.

FZK/GridKa

For all of April were limited to about 150 MB/s disk-disk for reasons that are not completely understood, though a significant problem in the disk striping configuration was only fixed early May, along with one or two other changes. The rates then finally managed to exceed the nominal 200 MB/s disk-disk, but there were new instabilities, e.g. a dip every 6 hours for reasons not understood. Switched to tape last week, using a single drive for now, not quite stable yet.

IN2P3

The very first site that was set up and ready even before SC4 started.

They came very close to their nominal 200 MB/s for many days in a row, but only came to steadily exceed that rate during the last two weeks, after having switched back from tape to disk, with some configuration changes. They now have met both disk-disk and disk-tape target rates.

NDGF

Started halfway through SC4, but immediately wrote all data to tape, even during the disk-disk phase, and very quickly reached their target rate of 50 MB/s, usually doing 60.

PIC

Still running CASTOR-1. Limited by their 1-Gbps shared uplink to some 70 MB/s disk-disk, in the end also reached during the disk-tape phase.

Had quite a few instabilities, but demonstrated about the maximum that could be achieved with their current setup. Now focusing on the upgrade to CASTOR-2. Network uplink to be upgraded in September.

RAL

Still running dCache instead of CASTOR-2. Using their production setup, shared with other users. Network situation was not completely clear, lightpath capacity got doubled in the midst of SC4, but did not have a very noticeable effect. Their disk-disk target rate was 150 MB/s, which was just met most of the time, but never exceeded. Tape target rate was 50 MB/s, which is met most of the time, but with high error rates.

SARA

Suffered some problems with their dedicated 10-Gbps link to CERN. Exceeded their nominal 200 MB/s disk-disk only for 2 days in a row. Discovered some bottleneck in their SAN configuration during the tape phase, limiting them to not much more than 30 MB/s. Switched their channel off to investigate short- and long-term solutions.

TRIUMF

Very stable reference site. Easily reached their 50 MB/s target rate both disk-disk and disk-tape.

Summary

The key operational problems encountered during Service Challenge 4 are reviewed. A concrete list of recommendations is proposed. Assuming agreement by the Management Board, the timeline for implementing these recommendations should be established and monitored.

Appendix

[1] The Service Challenge 4 'blog' - https://twiki.cern.ch/twiki/bin/view/LCG/ServiceChallengeFourBlog.

[2] Service Challenge 4 Disk-Disk transfer results - https://twiki.cern.ch/twiki/bin/view/LCG/AprilDiskDiskTransferTargetsAndStatus.

[3] The GridPP wiki - https://wiki.gridpp.ac.uk/wiki/Main_Page.

[4] TRIUMF SC3 disk – tape status page: http://grid.triumf.ca/status/sc4/sc4-disktape.html.

[5] BNL tape plots - http://www.atlasgrid.bnl.gov/dcache_tapewrite_monitoring/plots/.

[6] IN2P3 tape plots - http://netstat.in2p3.fr/weathermap/graphiques/lcgmss.html.

[7] IN2P3 preliminary analysis of SC4 throughput (Lionel Schwarz) - http://agenda.cern.ch/fullAgenda.php?ida=a057189.

[8] FNAL tape plots - http://cmsdcam.fnal.gov:8090/dcache/outplot?lvl=1&filename=billing-2006.04.daily.bwrsc3.png.

[9] Scheduling of Service Interruptions at WLCG Sites - http://agenda.cern.ch/askArchive.php?base=agenda&categ=a061501&id=a061501s0t11%2Fmoreinfo%2FSC4-scheduled-maintenance-May16.pdf. Can also be found via https://twiki.cern.ch/twiki/bin/view/LCG/TalksAndDocuments.