September 2006

Review of WLCG Site Monitoring and Operation in Service Challenge 4 (Service Phase)

Observations

This document reviews the state of site monitoring and operations in the Service Phase of WLCG Service Challenge 4 (SC4). It covers the period June 1st 2006 – September 15th 2006 (SC Tech Day and start of CMS CSA06), whereas the Service Phase in principle runs until the end of September.

We highlight in particular the following issues:

1. We are still not able to demonstrate full nominal Tier0-Tier1 transfer rates (1.6GB/s) over extended periods, let alone recovery rates (targeted at twice nominal);
2. However, experiment-driven data transfers (ATLAS and also CMS) achieved rates close to the target of full nominal rates (see table 1 below) for a single experiment (about half of the total rate for all experiments) under much more realistic conditions than for previous DTEAM transfers. For this reason, this is considered a positive result;
3. In addition, both ATLAS and CMS have managed to export over 1PB of data (1 PB of data per month for CMS over a 90-day period, 1.25 PB of data for ATLAS in the two-month period starting 19th June);

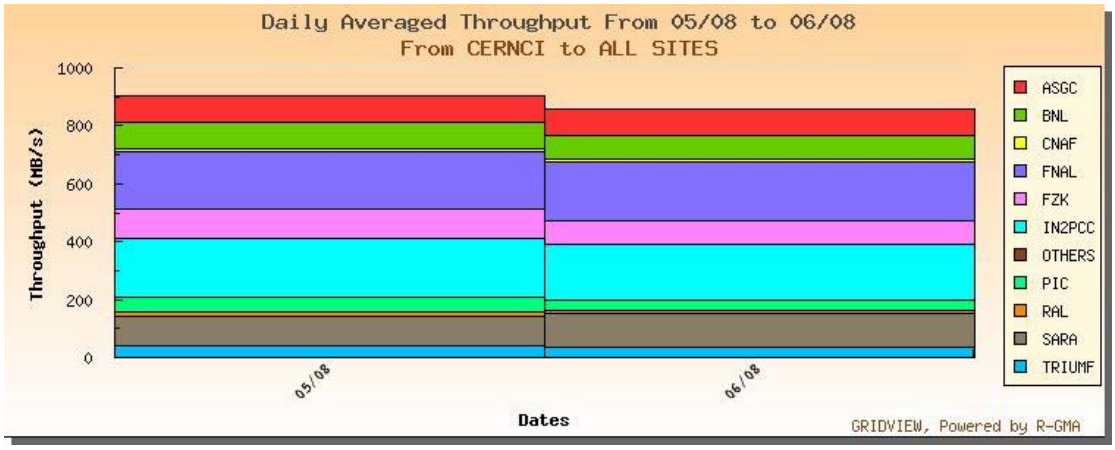


Figure 1 - ATLAS & CMS Driven Transfers (Weekend 5-6 August)

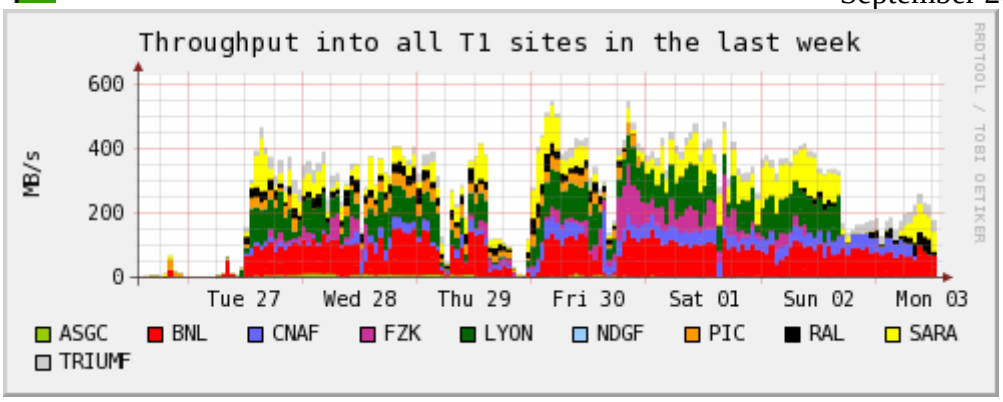


Figure 2 - ATLAS T0-T1 Transfers (Last Week of Tests)

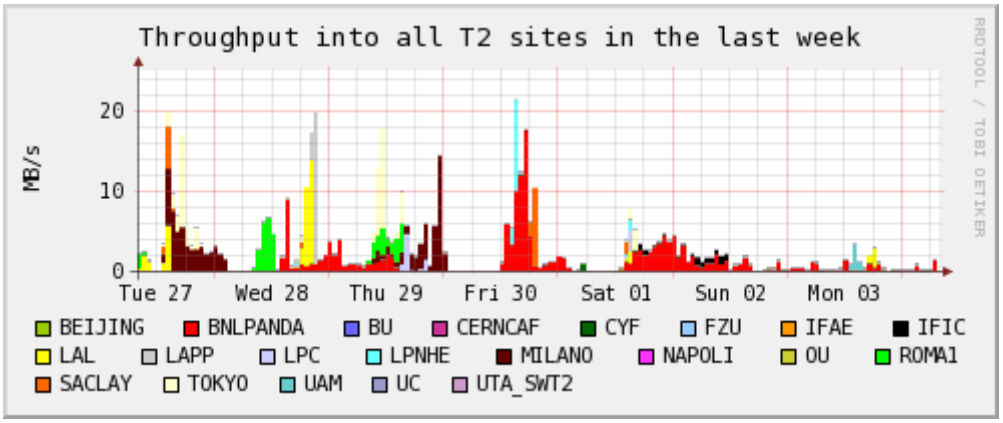


Figure 3 - Rates into ATLAS T2s

4. By definition, these activities tested site services, such as LFCs, VO boxes, and overall production readiness significantly more than the DTEAM-driven transfers. A number of issues have been found at a variety of sites and solutions have been found or are planned (see under the ATLAS section below). However, they underline the fact that certain sites / regions still have to make significant progress to achieve the required service level;
5. A particular effective model, as demonstrated by Lyon for ATLAS, is to have a contact person for the experiment both at the Tier0 and the Tier1;
6. Sites appear to be able to focus their full attention on a specific experiment or challenge for a few days only. This is clearly indicative of the high workload at the sites and should be built into the experiments' operational models (i.e. a few days at high priority per month per experiment already completely drains the sites involved);
7. Upgrades to CASTOR2 at a number of sites have led to further instabilities. Once all such migrations have been completed, a further test needs to be made to ensure that these sites can now meet both throughput and stability targets;
8. Several sites have experienced significant power and / or cooling problems, resulting in prolonged service downtime;

9. Several – if not many – sites appear to suffer from significant manpower shortages, which impacts both the service level that they are able to provide and the response time to requests (both "setup" and problem resolution);

10. Reporting to and attendance at the weekly Joint Operations Meetings[1] has improved since the previous report in May 2006 but still leaves considerable room for further improvement (reports are often written in a style that is clearly oriented at local consumption, some sites still do not provide reports on a regular basis, even though there is significant activity at that site);

11. Opportunistic use of resources – used or expected to be used by all experiments – may result in the use of CPU resources at sites with insufficient local storage. As an interim solution, unrestricted WAN access to the CERN SE has been provided, but this can result in poor and/or unpredictable network performance and result in problems that are highly complex to debug. It is considered important to clearly separate this opportunistic use of resources from the standard production model, where data is typically written to the local storage element (and eventually archived to the associated Tier1 site in the case of Monte Carlo production at Tier2s.);

12. A bug in Oracle 10.2.0.2 led to logical data corruption in the LFC and VOMRS instances at CERN. Once the problem had been sufficiently understood, it was successfully escalated to Oracle as a top priority issue. A work-around was put in place and the experiments and all outside sites were advised accordingly. At the time of writing a patch that passes all test cases has still not been received, although the workaround – effectively to turn off the faulty code path – solves most of the problems and eliminates the risk of further data corruption. This can be viewed as an important test case both of our ability to escalate such problems within the Oracle support structure as well as to handle bugs that potentially affect a large number of sites.

Recommendations and Actions

13. Streamlining of reporting to the weekly combined operations meeting – now held on Wednesdays at 16:00 Geneva time – and the various LCG coordination meetings (LCG Resource Scheduling Meeting Mondays at 15:00, LCG Service Coordination Meeting Wednesdays at 10:00) has been proposed to the WLCG Management Board and has been put in place;

14. The use of the EGEE broadcast tool for announcing both scheduled and unscheduled interruptions has greatly improved. Improvements in the tool to clarify broadcast targets are underway. Sites are requested to ensure the nature and scope of the event are clear both from the subject and text of the announcement (and are not, for example, inferred from the e-mail address of the sendee);

    a. Tape robot maintenance at CERN 10.30-16.00 Thursday 13 July

    b. Tape access interrupted

[1] See http://agenda.cern.ch/displayLevel.php?fid=258 to access agendas, reports, action items and minutes.

15. Site monitoring of local services still needs considerable further improvement – many issues that could be spotted locally are still first found by the central Service Coordination Team or – worse still – by the users;
16. Sites are encouraged to share their monitoring tools and experience. To this end, a focussed discussion on monitoring is foreseen at the Service Challenge Technical Day, September 15th at CERN.
17. Problem resolution – and reporting – needs to be improved, particularly in the case of complex problems which require a range of expertise and / or sites to resolve (see below);
18. Regular reviews of open tickets and identification of complex / unresolved problems are held with escalation (depending on exact problem) as required. This has proved successful in the resolution of chronic LHCb problems as well as the CMS CSA06 preparation.
19. Phone and / or physical participation of the experiments in the CERN daily operations meeting[2] (~10-15' starting at 09:15) is encouraged to highlight new problems and ensure that there is adequate information flow. These meetings are also be open to external sites wishing to participate;
20. A WLCG "Service Dashboard", allowing both supporters and production managers to clearly see the status of critical components (CASTOR@CERN, FTS, network transfers etc.) should be implemented as soon as possible to replace the laborious manual expert intervention – typically scanning log files – that is currently required;
21. A "Service Coordinator (On Duty – SCOD)" – a rotating, full-time activity for the length of an LHC run (but almost certainly required also outside data taking) should be established as soon as possible. The person assuming this activity would, for their period on duty:
    a. Attend the daily and weekly operations meetings, relevant experiment planning and operations meetings, CASTOR deployment meetings;
    b. Liaise with site and experiment contacts;
    c. Maintain a daily log of on-going events, problems and their resolution;
    d. Act as a single point of contact for all immediate WLCG service issues;
    e. Escalate problems as appropriate to sites, experiments and / or management;
    f. Write a detailed 'run report' at the end of the period on duty.
22. It is proposed that this rota be staffed by the Tier0 and Tier1 sites, each site manning ~2 2-week periods per year (or 4 1-week periods);
23. A regular (quarterly?) WLCG Service Coordination meeting, where the Tier0 and all Tier1+Tier2 federations as well as the experiments are represented, should be established. This should review the services delivered by that

[2] These meetings are typically held in the "openspace" in B513, except when this room is needed for a VIP visit. Dial-in access is via +41 22 767 6000 access code 0175012.

federation, main issues encountered and plans to resolve them, possibly following the model used by GridPP for their collaboration meetings (see, for example Deployment Metrics and Planning, presented at GridPP16). It should also cover the experiments' plans for the coming quarter in more detail than can be achieved at the weekly joint operations meetings (which nevertheless could cover any updates). This meeting should not require physical presence, but would require the reports / presentations to be submitted in advance;

ATLAS

24. The overall plan for the ATLAS SC4 exercise was to send data out to all ATLAS Tier1 sites at the full nominal rate expected for that site during LHC pp running. The data rates per site are shown in the table below.
25. Whilst these data rates were not achieved for the target of one week, this exercise uncovered a number of problems – many of which have since been resolved – and was clearly an important step towards reaching full nominal rates under realistic conditions.
26. Key accomplishments were:
    a. Ran a full-scale exercise, from EF, reconstruction farm, T1 export, T2 export with realistic data sizes, complete flow
    b. Included all T1s sites in the exercise from first day
    c. Included ~ 15 T2s sites on LCG by the end of the second week
    d. Maximum export rate (per hour) ~ 700 MB/s (Nominal rate ~ 780 MB/s (with NGDF))
    e. ATLAS regional contacts were actively participating in some of the T1/T2 clouds
    f. Put in place monitoring system allowing sites to see their rates (disk/tape areas), data assignments, errors in the last hours, per file, dataset, …
    g. FTS channels in place between T0 and T1 and now progressing between T1 and T2s
    h. Exported a total of 1PB of data by Sunday August 6th
27. Problems with VO box load have been identified and resolved, whereas adequate monitoring of LFC services at Tier1 sites remains an outstanding issue;
28. Major concerns include communication issues with the sites and the serious lack of manpower globally;

Centre	ATLAS SC4	Nominal (pp) MB/s (all experiments)
ASGC	60.0	100
CNAF	59.0	200
PIC	48.6	100

IN2P3	90.2	200
GridKA	74.6	200
RAL	59.0	150
BNL	196.8	200
TRIUMF	47.6	50
SARA	87.6	150
NDGF	48.6	50
FNAL	-	200

Table 1 - Target Rates by Site (40% to tape, 60% to disk)

29. The site by site summary of this exercise are listed in the table below.

CERN	
ASGC	after VO BOX upgrade, went very well. 100 MB/s when ATLAS runs; 40~50 MB/s when CMS runs (should be 60 MB/s); communication problems during start-up of exercise
BNL	not using realistic tape area; suffering from read/write contention when using 'production' areas (as opposed to SC4 /dev/null area); very good support for ATLAS
CNAF	unstable Castor-1; now fighting Castor-2 installations. Needs re-evaluation during next phase
LYON	very good service T0->T1 and T1->T2! The only site that was constantly part of the exercise (except for scheduled downtimes).
FZK	after VO BOX upgrade, went better. Still very unstable service (in/out of the exercise all the time)
PIC	stable service; dCache disk area and Castor tape area occasionally suffering some timeouts/overload issues
RAL	not stable; difficult to understand status; could not sustain rate for a few hours. See the LCG Quarterly Report for Q2 2006 for further details of on-going storage issues at RAL.
SARA	very stable service overall
TRIUMF	remains stable; network distance leads to occasional LFC connection glitches

Table 2 - ATLAS Tier1 Summary

CMS

30. The main activity during this period was preparation work for CMS CSA06. This involved debugging of data rates into and out of CERN (using PhEDEx

over FTS), clarification of FTS channel setup, monitoring and operations and testing of the gLite RB;

31. Problems resolved include poor transfers both into and out of CERN (related to the use of the loopback interface for SRM transfers and to incorrect handling at the SRM level of duplicate nameserver entries. Once these problems were resolved, and following tuning at the PhEDEx level, CMS were able to drive transfers at the target rate for CSA06 of 150MB/s (1/4 of the nominal rate);

32. Following this successful debugging exercise, an attempt to run at 500MB/s out of CERN for at least 3 days was made. Whilst this target was not reached, the 'threshold' of 300MB/s was attained, with a daily average of 450MB/s on 8th August, with ATLAS and other transfers proceeding in parallel.

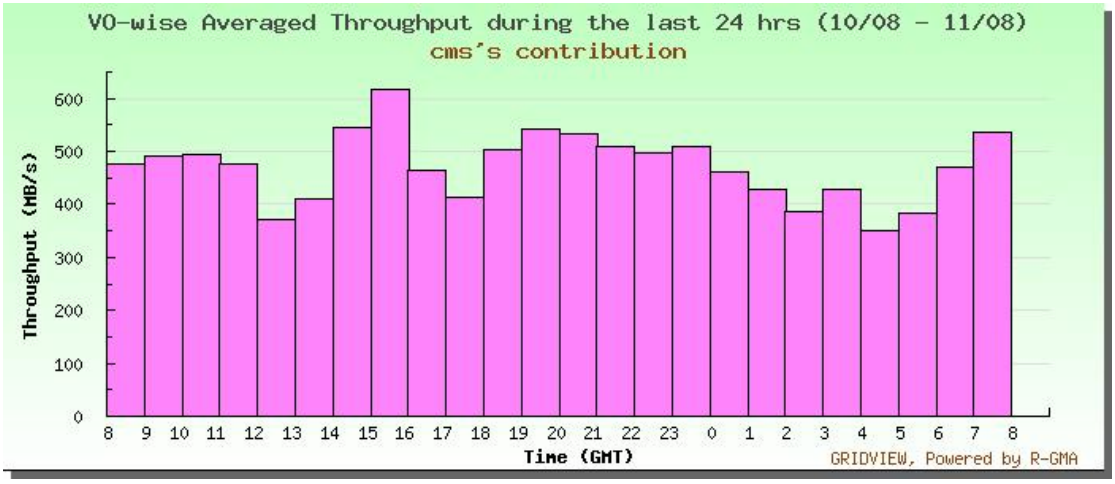
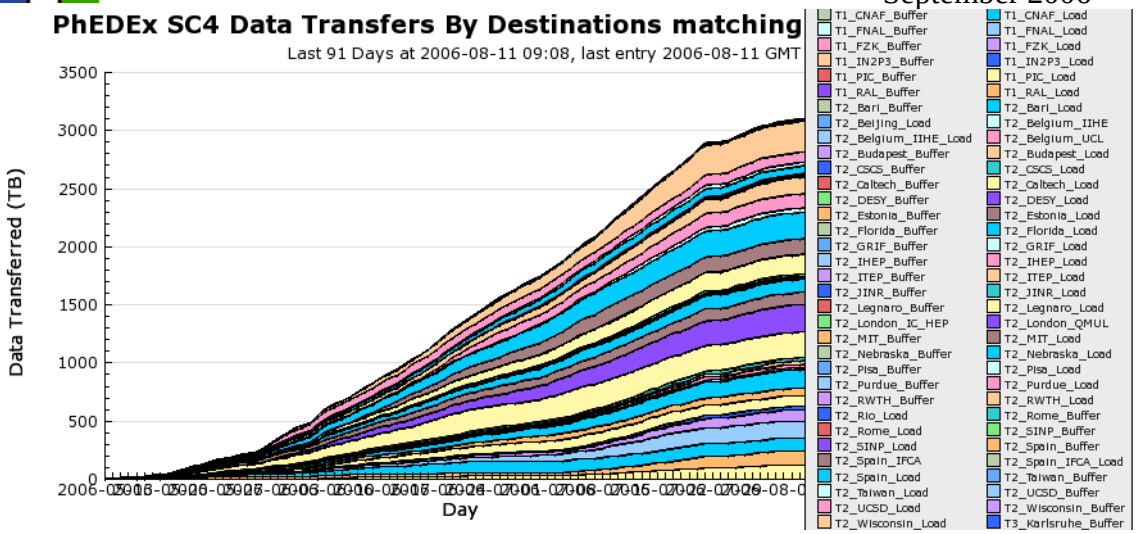
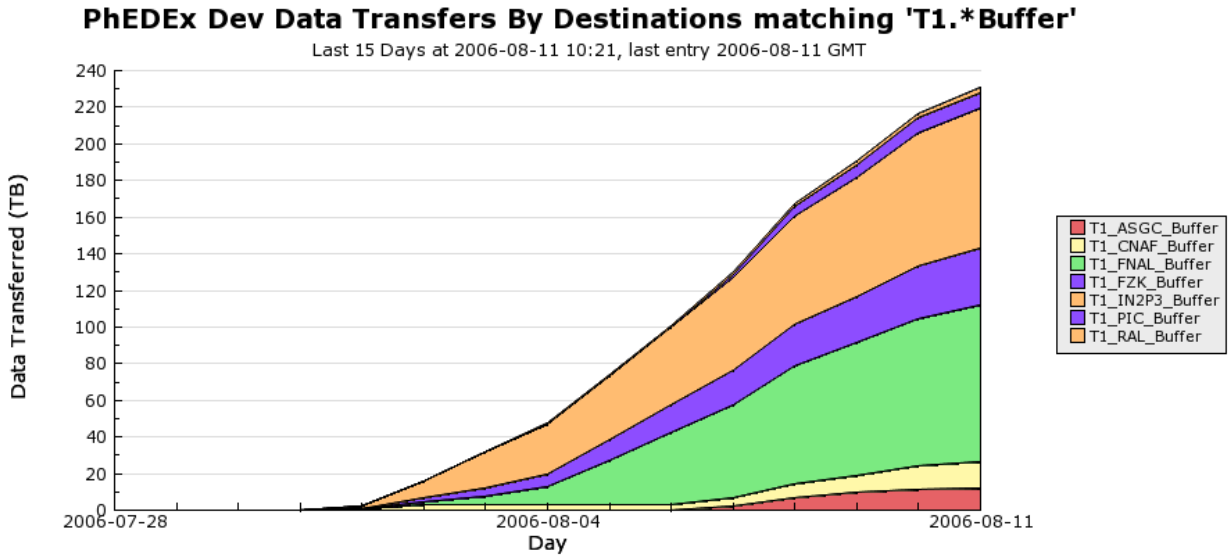


Figure 4 - CMS Transfers

33. In the 3 month period ending mid-August CMS transferred over 3.3 PB in wide-area transfers between storage systems. Of this, disk-to-disk SC4 transfers account for just over 3 PB and our recent two high-throughput Tier-0/Tier-1 disk-to-disk tests for most of the rest. This translates to an achieved rate of ~1 PB/month in CMS world-wide.





34. Specific problems encountered during these tests include various CASTOR2 bugs, such as the fact that CASTOR's reply to the stager_qry command was an arbitrary string that the PhEDEx stager agent had no chance to interpret in a sense that it could determine whether the requested file was on disk or on tape. Therefore it did what it was supposed to do, it submitted a stager_get request for that file. This resulted in a very large number (40K) of stager requests which rapidly overloaded the system. Thanks to Sebastien Ponce and his team the problem was quickly analyzed and a temporary fix was made available to CMS yesterday noon. The permanent fix is expected to be rolled-out by mid September;

35. Both CMS and LHCb experienced poor transfer rates into CERN (LHCb from worker nodes used opportunistically, CMS during the centralization of MC data as preparation for CSA06). These problems were eventually traced to the HTAR and have now been resolved. However, the intervention on the HTAR that led to these problems did not follow the agreed procedure for scheduling and announcing such changes and it is imperative that these procedures are rigorously followed in the future;

CERN	
ASGC	
FNAL	
CNAF	
LYON	
FZK	
PIC	
RAL	
SARA	

36. Work on patching and tuning the gLite RB as preparation for CSA06 (in collaboration with ATLAS) has been successful. Thus the CMS requirement to handle 50K jobs / day on less than 10 RBs can be met.

LHCb

37. The goals of the LHCb DC06 activity are as follows:
    a. Distribution of RAW data from CERN to Tier-1's
    b. Reconstruction/stripping at Tier-1's including CERN
    c. DST distribution to CERN & other Tier-1's
38. Simulated data are shipped to the 6 T1s + CERN with a share that depends on the computing power and status of the site. The amount of data processed is correlated to the amount of integrated data transferred out of CERN to various T1. So far the integrated rate is small (but close to a final draft of the LHCb computing model : ~3MB/s to each T1).
39. Problems at NIKHEF/SARA (dcap callback mechanism incompatible with network setup – resolved in a beta version of dCache) and at Lyon (use of gsidcap not yet supported by a production version of ROOT) impacted production, although temporary workarounds were found in both cases. For the above reason, the NIKHEF/SARA share is set to 0;

CERN	ran smoothly its share of jobs during the first month. Some issues with the AFS area serving the Software Installation Area that currently prevents to install jobs through a normal grid job. Problems with the Castor storage in uploading files from simulation jobs running on the small centers (due to the HTAR configuration) and also in the grid mapfile creation that seems to be uncorrelated to VOMS/LDAP mechanism as it happens somewhere else. Flickering behavior of the Information System.
CNAF	potentially CNAF is the largest center and could process the largest share of data. However it suffered a long standing problem with Castor2 stager. Basically at CNAF are using a different configuration to at CERN where for each VO there is a dedicated instance of the DB and LSF. There are several reasons behind:

	1. The single disk server serving the LHCb requests from LSF was not enough. There was also a limit on the max number of jobs per disk server increased to 300. (Fixed) 2. The DB is overloaded (deadlocks) and all the requests to the stager are stuck (fixed) 3. The pure disk pool (no Garbage Collector) seems to have problem in accessing files in case it becomes full (with consequent pending jobs overloading the LSF queue) Now CNAF should be OK.
LYON	ran smoothly DC06. Some minor issues due to the storage. They are using at Lyon the disk only storage instead of the tape endpoint (this last supporting only gsidcap protocol). Length of the largest queue doesn't fit with the LHCb Simulation jobs. Flickering Information System also experienced there.
FZK	Poor the usage of GridKA for reconstruction jobs of this DC06 (because it prevents to access data directly from the application), it has been rather used for production. The main problem (under investigation) seems related to their gridftp daemons that decide to close their sockets from time to time.
PIC	some issues with the storage; recent issue with pilot jobs that were not picking up any production (either reconstruction or simulation) job. PIC ran its share without any other major problem.
RAL	also ran smoothly DC06 reco jobs without major issues. Experienced a slowness accessing data at some point and problem fixed by adding another disk server.
SARA	NIKHEF/SARA never used for reconstruction: it is currently impossible accessing (through Root) data stored in the WAN connected Storage at SARA from WN via dcache.A patched version of the dCache client has been released for test. This version doesn't require Inbound connectivity on the WN because it wouldn't require calls client back. Site admins at NIKHEF are very collaborative and are pushing for testing/certifying new dcache libraries needed by LHCb. Once there will be prove that new clients are working fine they will install in their nodes without waiting official release of LCG. Experiment side also tests with the application are ongoing. They didn't yet manage to access file with gsidcap and these new dCache clients. Until further news, NIKHEF sits out DC06 activity.

ALICE

40. ALICE targeted FTD over FTS driven transfers at the full nominal rates expected during heavy ion running (i.e. HI rates spread out over the expected 4 months of the machine shutdown – 300MB/s out of CERN). Optimisations in

the interaction between FTD and FTS were required as initially the failure of a single site could block transfers to all sites.

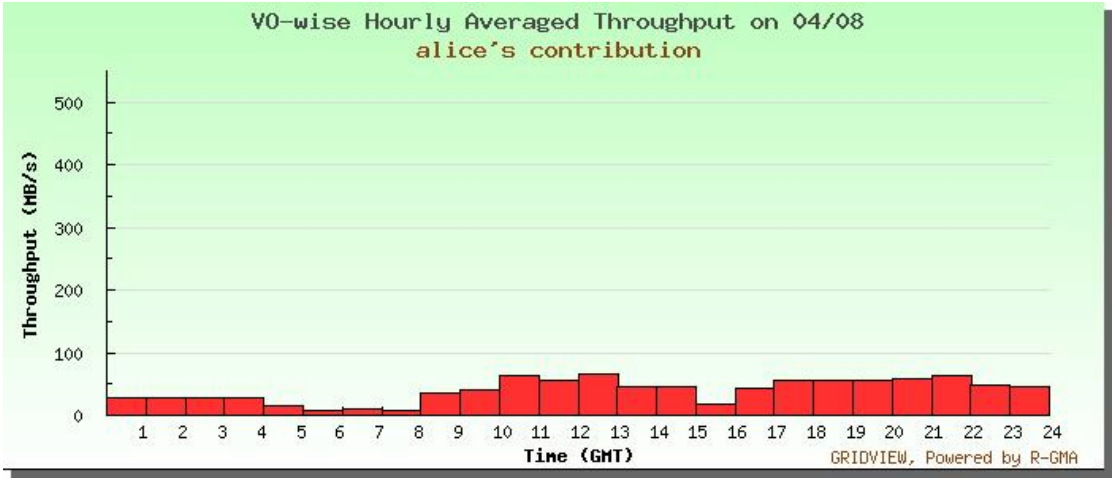


Figure 5 - ALICE driven FTD/FTS Transfers

CERN	
CNAF	
LYON	
FZK	
PIC	
RAL	
SARA	

[Summary of export exercise plus site-by-site review]

Summary

41. The importance of adequate preparation by the experiments has been clearly demonstrated, as has the need for constant activity to exercise the global WLCG services;
42. Significant improvements in site-local monitoring and services in general is required to reach the WLCG MoU targets;
43. Further improvements in experiment-driven Tier0-Tier1 data transfers are required to reach the nominal and recovery rates;
44. Sites of particular concern include NDGF (not able to participate to this phase of SC4); FZK (unstable service, particularly during ATLAS' activity); CNAF (unstable service – hopefully improved once CASTOR2 migration is fully completed, many expert interventions required, manpower concerns); RAL (effectively unable to participate due to disk controller problem described in LCG Q2 report);

45. ASGC, which achieved transfer rates significantly below its nominal target during the April SC4 DTEAM throughput tests, have shown significantly better results in the ATLAS and CMS driven transfers.