GES 2007 German e-Science

Available online at http://www.ges2007.de

Challenges of the LHC Computing Grid by the CMS experiment

A. Scheurer[1], M. Ernst[3], A. Flossdorf[3] C. Hof[2], T. Kress[2], K. Rabbertz[1], G. Quast[1]

[1] Institut für Experimentelle Kernphysik, Universität Karlsruhe,
Wolfgang-Gaede-Str. 1, 76131, Karlsruhe, Germany
[2] III. Physikalisches Institut, RWTH Aachen, 52056 Aachen, Germany
[3] DESY, Notkestrasse 85, 22607 Hamburg, Germany

Abstract

This document summarises the status of the existing grid infrastructure and functionality for the high-energy physics experiment CMS and the expertise in operation attained during the so-called ”Computing, Software and Analysis Challenge” performed in 2006 (CSA06). This report is especially focused on the role of the participating computing centres in Germany located at Karlsruhe, Hamburg and Aachen.

1 Introduction

In preparation for the enormous amounts of data expected from the future large hadron collider (LHC) presently under construction at the European laboratory for particle physics, CERN, a grid infrastructure is being set up by the physics communities in the participating countries. The world-wide LHC Computing Grid (WLCG) is based on tier-structured layers of one Tier0 at CERN and several Tier1, Tier2 and Tier3 centres distributed all around the globe. The German part of the CMS grid infrastructure relies heavily on the two Helmholtz centres Forschungszentrum Karlsruhe as Tier1 and DESY Hamburg as well as on installations at the universities Karlsruhe and Aachen. The emphasis of this report lies on so-called service challenges performed in 2006 where the full functionality of the whole chain from data processing at the Tier0, data distribution to the Tier1 and Tier2 centres up to organised data re-processing and more chaotic physics analyses by single users was exercised.

2 The CMS Computing Model

The CMS computing model [1] relies on the hardware infrastructure and the services offered by the world-wide LHC Computing Grid (WLCG) [2, 3]. It is based on four hierarchically-ordered layers of one Tier0 situated at CERN, seven Tier1 centres in Europe, Northern America and Asia, about 30 Tier2 and numerous Tier3 centres distributed equally around the world, see figure 1. The data flow from the source, the CMS detector, to the desktops of the analysing physicists has to pass many different stages:

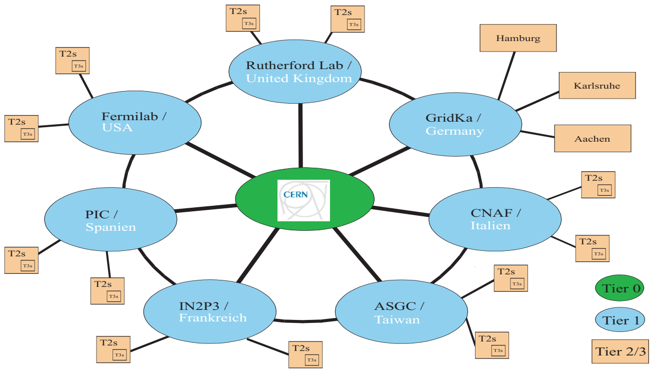

Figure 1: Overview of the WLCG/CMS tier structure, including Tier0, all seven Tier1 and several Tier2 and Tier3 sites.

- The Tier0 at CERN collects the data accepted by the trigger system of the experiment and performs a first reconstruction pass at a rate of 40 Hz using the new CMS software framework described in section 3.4. The processed data is stored permanently on magnetic tapes and temporarily on disk-pools for further distribution.
- These data are distributed to the regional Tier1 centres, where a second, distributed copy is archived on magnetic tape for future reprocessing. The Tier1 also receives a full copy of the particle information as reconstructed in the detector, on which most physics analyses will be performed.
- Calibration jobs are executed at the Tier0, Tier1 and Tier2 centres.
- Re-reconstruction of existing datasets and skim jobs (generation of smaller data samples matching certain selection criteria) are performed at the Tier1s.
- Selected data are exported to the Tier2 centres, where analysis jobs by individual physicists are executed.
- The Tier3 centres are not explicitly defined in the computing models. They provide local resources for the analysing physicists, in particular for interactive analysis, software development, testing and grid access and the final interpretation of the analysis results.

3 The Computing, Software and Analysis Challenge 2006

In order to test the readiness of CMS for the upcoming startup of LHC, a so-called Computing, Software and Analysis Challenge (CSA06) was performed in autumn 2006. More information than given below can be found in the CSA06 Summary Report [4].

3.1 Motivation and Goals

Since the buildup processes of the CMS computing resources are still ongoing and far from being finished, CSA06 was designed to operate at a level of 25% of the resources foreseen for 2008 during four weeks of operation. Large datasets had to be simulated as input for the Tier0 reconstruction. The success of CSA06 was measured by pre-defined metrics (the parenthesised numbers indicate the threshold for being considered a success):

- Automatic data transfer from Tier0 to Tier1 and Tier1 to Tier2 via the CMS data placement tool PhEDEx [5], as well as automatic transfer of skim job output data to the Tier2 centres.
- Read and write access to the offline database at CERN (Oracle) via the database caching service Frontier/Squid [6, 7], and running of re-reconstruction jobs generating datasets based on updated database information.
- Submission of user jobs via the grid job submission tool CRAB [8] and using the CMS-developed Dataset Bookkeeping Service (DBS) and the Data Location Service (DLS).
- Number of participating Tier1 sites: 7 (5) with an uptime > 90%.
- Number of participating Tier2 sites: 20 (15).
- Weeks of running at the sustained rates: 4 (2).
- Tier0 efficiency during the best 2 weeks: 80% (30%).
- Number of running grid jobs at Tier1 and Tier2 centres per day, each job running about 2 hours: 50 000 (30 000).
- Grid job efficiency: 90% (70%).
- Data serving capability for the jobs at each participating site: 1 MB/sec per execution slot (Tier1: 400 MB/sec, Tier2: 100 MB/sec).
- Data transfer rates from Tier0 to Tier1 tape system individually vary between 10 - 50 MB/sec (5 - 25 MB/sec).
- Data transfer rates to the Tier2 centres: 20 MB/sec (5 MB/sec).

The tools mentioned here are described in section3.5.

3.2 Participating Sites

Tier0: Based at CERN, the Tier0 is the core centre within the CMS computing model. As mass storage it provided two standard Castor pools, an input pool with about 64 TB and an export pool with 155 TB of disk space. 700 dual-CPU worker nodes with a system clock of 2.8 GHz and 2 GB of memory formed the batch system running under LSF (Load Sharing Facility) [9] on SLC4 (Scientific Linux CERN version 4).

Tier1: All seven CMS Tier1 centres participated in CSA06 and provided 25% of the resources descibed in the Memorandum of Understanding (MoU) [3]. The estimated requirements to reach the goals of the service challenge were about 300 CPUs for processing and about 70 TB of storage capacity for a nominal Tier1.

Tier2: In total 27 Tier2 centres participated in CSA06. The estimated requirements for a nominal Tier2 were about 100 CPUs and 10 TB of disk space. The variance between the Tier2 centres, however, is rather wide.

3.3 German Sites

With GridKa, the German Tier1 centre, and a federated Tier2 consisting of installations at DESY and RWTH Aachen, the German groups will be well positioned to competitively contribute to the physics analysis. CSA06 used gLite version 3.0 as grid middleware, and the CMS-specific services described in section 3.5 were installed on the operating system Scientific Linux CERN version 3 (SLC3) at each site.

Tier1 GridKa: GridKa provided about 70 TB of storage space and 250 CPUs during CSA06. The storage system is based on dCache [10] with four pool nodes dedicated to CMS. The network connection between CERN and GridKa consisted of one 10 GBit link via optical fiber. As the CMS computing model also foresees data transfers to any Tier2 from Tier1 centres, corresponding links to numerous Tier2 centres via standard Internet connections have been established.

Tier2 DESY and Tier2/3 Aachen: The German Tier2 is a federation of two independent grid installations at DESY, Hamburg and the university of Aachen. DESY provided 100 CPUs and 35 TB of dCache storage space. The spacial proximity of the Aachen Tier2 and Tier3 allows to gain from potential synergy effects due to sharing parts of the same hardware architecture. Aachen provided about 35 workernodes with 94 cores in mixed dual- and quad-core machines, 1 GB of RAM per core. The batch system is running Maui/Torque [11, 12]. The storage is handled by a dCache system with a capacity of more than 5 TB dedicated to CSA06.

Tier3 Karlsruhe: The Tier3 cluster at Karlsruhe is provided by the Institut für Experimentelle Kernphysik. The cluster structure is quite heterogeneous with an order of 30 worker nodes with a mixed single and dual CPU setup with about 1 or 2 GB of memory per core. They run under the Maui scheduler and the Torque batch system. The storage capacity adds up to 8 TB of nfs shared disk space for CMS.

3.4 The CMS Offline Software

The physics-related software of CMS, i.e. simulation and reconstruction code, has undergone heavy development during preparation of CSA06. The CMS software is managed centrally using grid jobs for installation and validation. Nonetheless, frequent software updates added some instability even during the main phase of CSA06.

3.5 Production and Grid Tools

CRAB - CMS Remote Analysis Builder: CRAB allows access to the remotely stored data and thus analysing it via the grid infrastructure. With the help of the CMS data management services DBS and DLS it determines the location of the data, prepares the needed user code for the analysis task and manages the interaction with the gLite middleware from job submission over job status queries up to job output retrieval.

DBS - Dataset Bookkeeping System: At CMS, each dataset is divided up in several data blocks, whereas each block contains on the order of some hundred files represented by Logical File Names (LFN). The Dataset Bookkeeping System is responsible for the proper mapping between files and blocks as well as providing some processing related information like the size or number of events for each block.

DLS - Data Location Services: Based on the Local File Catalog (LFC) infrastructure provided by the EGEE grid project [13], DLS is responsible for the mapping between data blocks and grid sites physically hosting them.

TFC - Trivial File Catalog: Since DBS and DLS only provide the locations of a specified data block in the grid and the LFNs of the contained files, the TFC is responsible for the mapping of each LFN to a physical filename on each sites storage system. The TFC is a xml-based mapping file hosted on each tier site.

PhEDEx - Physics Experiment Data Export: The CMS data placement tool is responsible for physically transferring the data between the different sites. Although PhEDEx is capable of using the basic grid file transfer protocols like gridFTP, only more advanced protocols like SRM (Storage Resource Manager) and FTS (File Transfer Service) have been used during CSA06.

3.6 Data Processing

The workflow at the Tier0 during CSA06 covered prompt reconstruction, creation of special data streams for detector alignment and calibration (AlCaReco) and data for analysis (AOD), merging them to larger output files, injection into the Dataset Bookkeeping System and PhEDEx as well as access to the alignment and calibration database via Frontier. The Tier0 ran stably and mainly unattended during the four weeks of the challenge with an uptime of 100%. Only scheduled operator inventions like changing the running reconstruction version, adjusting the rate of different physics channels or fixing trivial bugs in the Tier0 framework helped to achieve all proposed metrics of CSA06.

3.7 Data Distribution

The transfer rates to the Tier1 centres were expected at about 25% of the needs in 2008. Table 1 gives an overview of the goals, threshold values and achieved transfer rates for the seven Tier1s as well as information about network speed and provided storage capacity. During the whole period of CSA06, more than 1 PB was transferred among all participating tier sites. Figure 2 shows the cumulative transferred amount of data via the wide area network.

Tier1	Goal	Thres.	30day avg.	15day avg.	Storage	Network
Site	[MB/sec]	[MB/sec]	[MB/sec]	[MB/sec]	[TB]	[Gb/sec]
ASGC	15	7.5	17	23	48+(36)	2
CNAF	25	12.5	26	37	40	10+(10)
FNAL	50	25	68	98	700	11
GridKa	25	12.5	23	28	40+(30)	10
IN2P3	25	12.5	23	34	70	10
PIC	10	5	22	22	50	1
RAL	10	5	23	33	60+(20)	11

Table 1: Intended and achieved transfer rates from CERN to the Tier1 centres as well as the provided storage space and network capacity. The parenthesised numbers indicate installed components during CSA06.

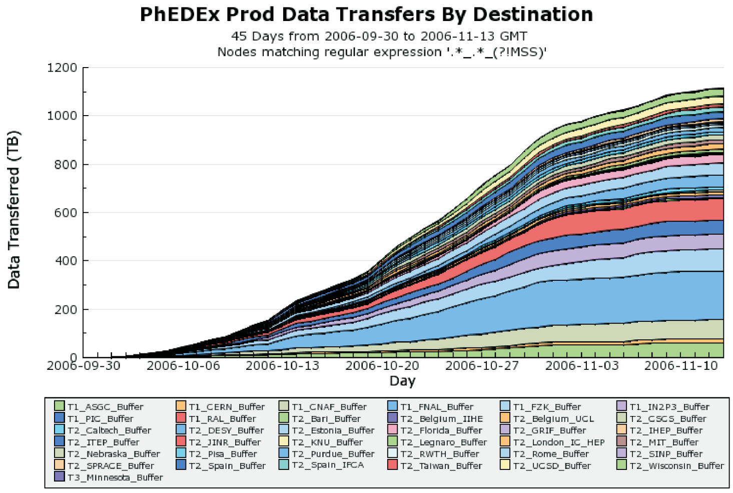

Figure 2: Cumulative data transferred during CSA06 between all participating tier sites in TB.

3.7.1 Tier0 to Tier1 Operations

During the first weeks of CSA06, the transfer rate between CERN and the seven Tier1 centres did not exceed the desired target rate of 150 MB/sec. The increase of the Tier0 reconstruction rate resulted in much higher transfer rates during the second half of the challenge. As can be seen in table 1, all centres reached the goal within a 15 day average of the challenge, whereas five Tier1 centres even succeeded during the whole 30 day period of CSA06. Figure 3 shows the transfer rates between CERN and the Tier1 sites. Due to some database problems, a higher amount of data was accumulated resulting in transfer rates of over 300 MB/sec. Towards the end of the challenge, outgoing rates of more than 350 MB/sec have been reached easily due to the higher Tier0 reconstruction rate. All shown transfer rates are averaged over 24 hours whereas the hourly average even exceeded 650 MB/sec. Most Tier1 centres performed at a very high level of quality. In case of GridKa, some stability problems with the initial version of dCache needed frequent manual intervention and made a scheduled upgrade inescapable. The subsequent reconfiguration and tuning of the system led to large inefficiencies during the last week of CSA06. Nevertheless, GridKa finished the challenge with only a few days of delay and record transfer rates of over 440 MB/sec over more than 8 hours.

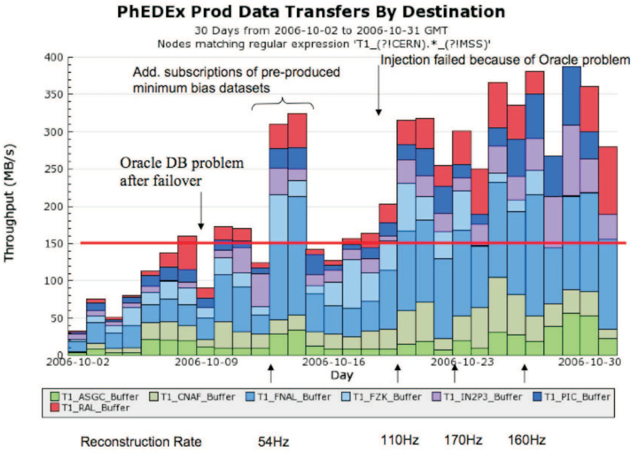

Figure 3: Data transfer rate between Tier0 and Tier1 centres during the first 30 days of CSA06 in MB/sec. The red line indicates the target rate for CSA06.

3.7.2 Tier1 to Tier2 Operations

The goals for CSA06 were reached by 20 out of the 27 participating Tier2 sites, whereas one other centre reached the threshold anyhow. Figures 4 and 5 show the data transfer rates between Tier1 and Tier2 centres and the transfer quality for these transfers, respectively. Since the Tier2 sites are only expected to transfer the data in bursts, the rates are not as continuous as they are for the Tier1s. The transfer quality varied widely among the participating Tier2 centres. Figure 6 shows that the German Tier2 sites performed extremely well during the whole period of CSA06 with successful transfers from GridKa to DESY at rates exceeding 140 MB/sec.

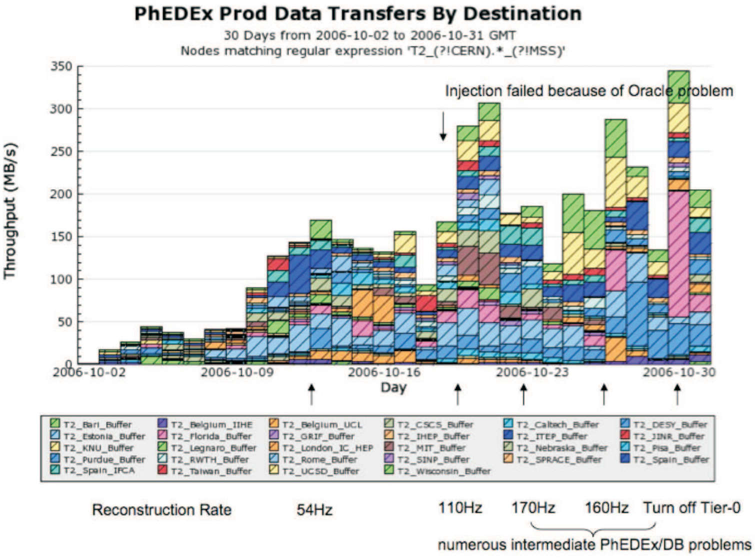

Figure 4: Data transfer rate between Tier1 and Tier2 centres during the first 30 days of CSA06 in MB/sec.

3.8 Data Re-Processing

The goal was to test the re-reconstruction of existing datasets with new calibration constants from the offline database at an order of 100 000 events at each Tier1. Due to problems with the yet untested reconstruction code, not all Tier1 centres achieved this goal immediately. After skimming and running over the resulting smaller data samples re-reconstruction succeeded at all Tier1 centres.

4 Monte Carlo Production

In preparation for CSA06, more than 67 million events had to be simulated beforehand and then were transferred to the Tier0. The grid-based Monte Carlo

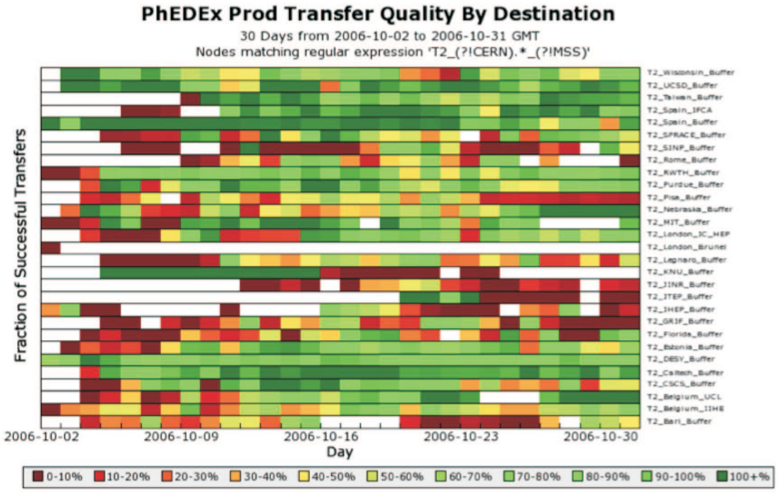

Figure 5: Data transfer quality between Tier1 and Tier2 centres during the first 30 days of CSA06.

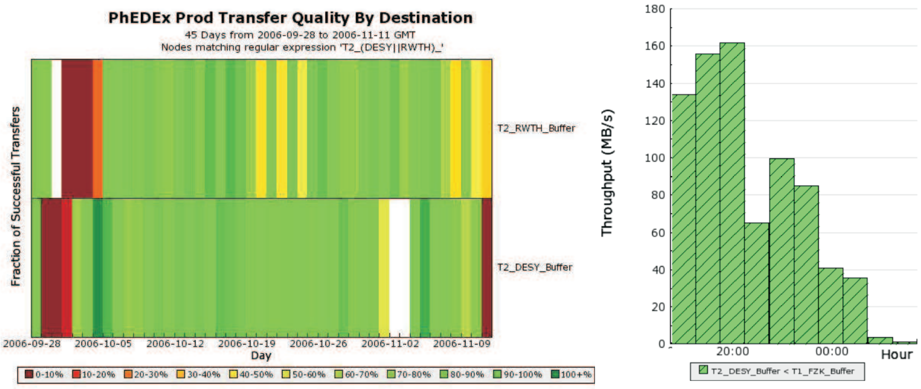

Figure 6: Data transfer quality for the DESY and Aachen Tier2 during CSA06 and an example of transfer rates between GridKa and DESY.

production started in mid July and was conducted by four operations teams from the Tier2 centres CIEMAT, DESY/RWTH, INFN/Bari, and University of Wisconsin, Madison. Although primarily a task of the Tier2s, Tier1 centres aided in this exercise. The number of events produced at the German sites adds up to 14% of the whole CMS Monte Carlo Production.

5 Physics Analyses

The data distributed during CSA06 was not only used for pure testing, but was at the same time used for physics studies by individual users. In addition, a "job robot" secured a base load of grid jobs being run at each centre. Here, the performance of the resource broker was not sufficient to keep all available CPUs busy.

6 Conclusions

The Computing, Software and Analysis challenge of the CMS collaboration performed in the fall of the year 2006 demonstrated the base functionality of the grid infrastructure, grid services and the experimental software at a level of 25% or more of the requirements foreseen for 2008. In particular the grid storage elements showed some fragility under heavy load and required substantial efforts on the operational side. The German CMS sites, the Tier1 centre GridKa and the federated Tier2 centre DESY/RWTH Aachen, successfully participated at all stages of the challenge. Much experience was gained with the execution of complicated workflows on the grid infrastructure, and performance bottlenecks and instabilities were identified. Further service challenges at a larger scale will be needed to guarantee the functionality of the whole system at the scale required for the start-up of data-taking at the LHC.

References

1. CMS Collaboration, "The Computing Project, Technical design report", CERN/LHCC 2005-023
2. J. Knobloch, L. Robertson, et al. "LHC Computing Grid - Technical Design Report". LCG-TDR-001 (2005). CERN-LHCC-2005-024. ISBN 92-9083-253-3.
3. LHC and WLCG Collaboration "Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid". CERN-C-RRB-2005-01/Rev. http://lcg.web.cern.ch/LCG/C-RRB/MoU/WLCGMoU.pdf.
4. CMS Collaboration "CMS Computing, Software and Analysis Challenge in 2006 (CSA06) Summary". CMS NOTE 2007/006 (2007). CERN/LHCC 2007-010. LHCC-G-128.
5. CMS PhEDEx Webpage. http://cmsdoc.cern.ch/cms/aprom/phedex/.
6. Frontier Webpage. http://lynx.fnal.gov/ntier-wiki/.
7. Squid Webpage. http://www.squid-cache.org/.
8. CMS CRAB Webpage. http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/.
9. Load Sharing Facility Webpage. http://www.platform.com/Products /Platform.LSF.Family/Platform.LSF/Home.htm.
10. dCache Development Webpage. http://dcache.desy.de/.
11. Maui Scheduler Webpage. http://mauischeduler.sourceforge.net/.
12. Torque Resource Manager Webpage. http://www.clusterresources.com /pages/products/torque-resource-manager.php.
13. Enabling Grids for E-Science Project Webpage. http://public.eu-egee.org/.