

The LHCb computing data challenge DC06

Authors : Raja NANDAKUMAR (Science and Technology Facilities Council), Sergio GOMEZ JIMENEZ (University Rovira i Virgili), Marco ADINOLFI (H. H. Wills Physics Laboratory, Bristol), Roland BERNET (Universitat Zurich), Johan BLOUW (Physikalisches Institut, Heidelberg), Daniela BORTOLOTTI (Universita & INFN, Bologna), Angelo CARBONE (Universita & INFN, Bologna), Besma M'CHAREK (Vrije Universiteit, Amsterdam), Davide Luigi PEREGO, (INFN sez. Milano-Bicocca), Andrew PICKFORD (University of Glasgow), Cedric POTTERAT (LPHE-IPEP, Lausanne), Marcos SECO MIGUELEZ (Universidad de Santiago de Compostela), Marianne BARGIOTTI (CERN), Nick BROOK (University of Bristol), Adria CASAJUS (Universitat de Barcelona), Gianluca CASTELLANI (CERN), Philippe CHARPENTIER (CERN), Carmine CIOFFI (University of Oxford), Joel CLOSIER (CERN), Ricardo GRACIANI DIAZ (Universitat de Barcelona), Gennady KUZNETSOV (Rutherford Appleton Laboratory), Stuart PATERSON (CERN), Roberto SANTINELLI (CERN), Andrew CAMERON SMITH (University of Edinburgh), Andrei TSAREGOROTSEV (Universit d'Aix - Marseille II)

E-mail: r.nandakumar@r1.ac.uk

Abstract. The worldwide computing grid is essential to the LHC experiments in analysing the data collected by the detectors. Within LHCb, the computing model aims to simulate data at Tier-2 grid sites as well as non-grid resources. The reconstruction, stripping and analysis of the produced LHCb data will primarily place at the Tier-1 centres. The computing data challenge DC06 started in May 2006 with the primary aims being to exercise the LHCb computing model and to produce events which will be used for analyses in the forthcoming LHCb physics book. This paper gives an overview of the LHCb computing model and addresses the challenges and experiences during DC06. The management of the production of Monte Carlo data on the LCG was done using the DIRAC workload management system which in turn uses the WLCG infrastructure and middleware. We shall report on the amount of data simulated during DC06, including the performance of the sites used. The paper will also summarise the experience gained during DC06, in particular the distribution of data to the Tier-1 sites and the access to this data.

1. Introduction

The LHCb experiment [1] is the Large Hadron Collider beauty experiment at CERN, primarily intended for precise measurements of CP violation and rare decays. Along with the remaining LHC experiments, LHCb expects to start taking data in mid-2008. The data rate is expected to be about 1 PetaByte per full year of running. This amount of data will need large computing resources at a scale in which the only reasonable solution is the worldwide computing grid.

LHCb uses the worldwide computing grid to perform most of the large scale computing tasks. The next section describes the LHCb computing model [2] in more detail. The computing data challenge DC06 is intended to produce data for the LHCb physics book and realistically test the computing model while challenging the LHCb grid submission and production service. This includes Monte Carlo simulation, reconstruction, stripping and analysis of data with the consequent testing of the data storage and transfer frameworks

2. The LHCb computing model

The LHCb computing solution as proposed in the computing model relies on operating in the WLCG (Worldwide LHC Computing Grid) framework with a possibility of expanding into other frameworks as and when resources become available to LHCb and their usage becomes feasible. LHCb uses the DIRAC (Distributed Infrastructure with Remote Agent Control) technology to submit jobs to sites in the WLCG. DIRAC uses the pull method of submitting pilot agents to sites and is further described in [3]. The analysis user interface is provided by Ganga (Gaudi / Athena and Grid Alliance, described in [4]) while DIRAC is directly used for production jobs.

To summarise the LHCb computing model, CERN is a Tier-0 centre (also a Tier-1 centre). There are 6 Tier-1 centres at RAL (UK), PIC (Spain), CNAF (Italy), GridKa (Germany), NIKHEF (The Netherlands) and IN2P3 (France). The primary sources of storage will be at CERN and the Tier-1 centres. The raw data will have two copies – one at CERN and one at one of the Tier-1 centres. The Monte Carlo simulation will be done at any Tier-2 / Tier-1 / Tier-0 centre available for use by LHCb. The reconstruction will be done at one of the Tier-1 / Tier-0 centres and the output rDST will be stored at the local storage element (storage system at a given location). The stripping of the reconstructed data will also be done at the same Tier-1 centre, but with the output being redistributed across all the Tier-1 / Tier-0 centres. Analysis is projected to run at present only at the Tier-1 / Tier-0 centres, but with a possibility to run on a large enough Tier-2 centre in future.

The raw data and rDSTs are typically stored on tape, while the stripped DSTs are stored on the local disk. It may be mentioned here that for reconstruction, stripping and analysis, the jobs directly open the input files on the storage system without copying it to the local area. This dramatically reduces the amount of resources required on the worker nodes and the peak loads on the storage system itself, though the job now becomes a little more dependent on the long term stability of the storage system on the site.

3. Resources required in DC06

The resources needed for DC06 are given below. Also given are the different storage technologies used at the various Tier-1 centres.

- Disk space of 80 TB at CERN and each Tier-1
- Processing power of 2 MSI2K cpu-days/day
- A few more TB (primarily at CERN) for physics analysis
 - This is mainly to store root trees produced by physicists.
 - There will be a small need to store reduced private event datasets.
- Bandwidth of 10 MB/s from CERN to the Tier-1s
 - 1 MB/s from Tier-2s to Tier-1s
 - We need FTS channels set up to handle data transfers between CERN and the Tier-1s

| Site | Storage |
|--------|---------|
| CERN | CASTOR2 |
| CNAF | CASTOR2 |
| PIC | CASTOR1 |
| GridKa | dCache |
| IN2P3 | dCache |
| NIKHEF | dCache |
| RAL | dCache |

4. Computing Data Challenge DC06

DC06 involved a series of definite steps to challenge the computing model. These included Monte Carlo simulation, its transfer to the Tier-1 sites, reconstruction and stripping. The analysis of the stripped data is the final step in this and is not discussed here, as it is not performed by the production team. It may be mentioned that the analysis of earlier produced DC04 data took place independently, in parallel with the above steps.

4.1. Monte Carlo simulation

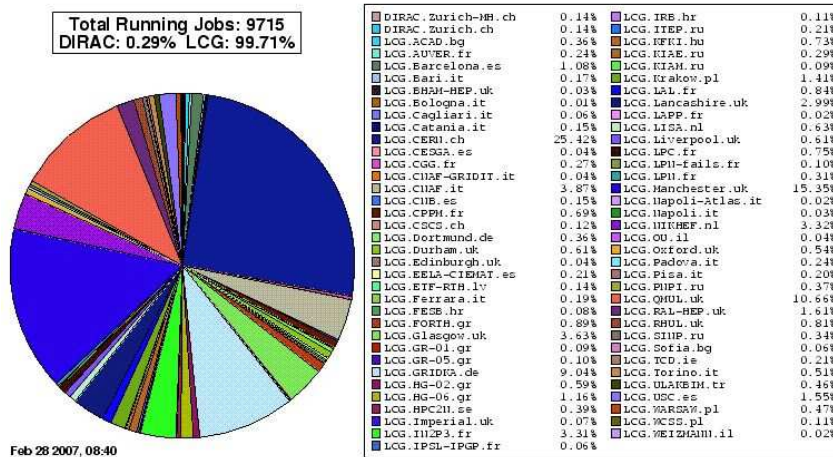


Figure 1. Figure showing the distribution of simulation jobs over various sites on the grid.

For any particle physics analysis the existence of large amounts of Monte Carlo is essential. During DC06, about 700 million events were simulated over the grid (about 120 different sites) in a span of about 475 days, corresponding to about 1.5 Million events per day. A maximum of about 10000 jobs ran simultaneously as can be seen in the figure 1. The limitation here was the availability of sites and the LHCb fairshare available there. All the initially requested simulations were finished by the end of March 2007.

4.2. Reconstruction and Stripping

The simulated data above need to be reconstructed using the standard reconstruction software to extract the particles. The reconstructed data is then stripped. In both these cases, data required is pre-staged before the job is submitted to the grid. This eliminates a large dead time of the job in the beginning, while waiting for data to be staged. Since these two steps involve the storage systems, they take place only at CERN and the Tier-1 sites.

About 100 Million events have been reconstructed in DC06, involving 200,000 file recalls from tape. 10,000 jobs were submitted and all the Tier-1 sites were used simultaneously. A snapshot of the reconstruction at a time when 441 jobs were running is shown in figure 2. Stripping involves creating a reduced data set by processing the data

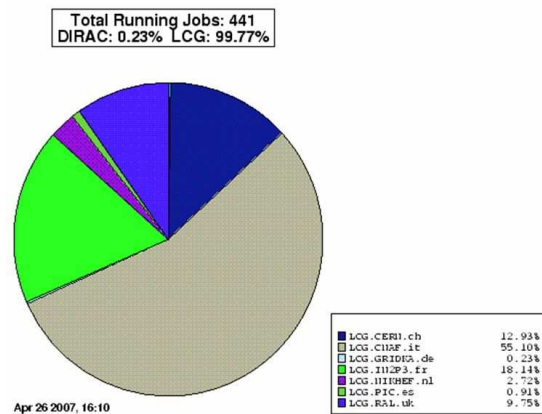


Figure 2. The distribution of reconstruction jobs over the various Tier-1 sites on the grid.

through a set of pre-selections. 10 Million events have been stripped so far. Both the reconstruction and stripping are works in progress.

4.3. Data distribution

Within LHCb, there are two use cases for data distribution. The first is in the redistribution of data that exists at one site to other sites, as needed by the situation. The other is the upload of data from the worker nodes to the relevant storage element. These cases are described below.

The Monte Carlo raw data was redistributed among the various Tier-1 sites for reconstruction in autumn 2006, as a part of WLCG SC4. This was done using the FTS, implemented by a Data Management Service built around DIRAC. Though good throughput was seen (figure 3), there were many instabilities in the various storage elements, with the result that simultaneous transfers to all six Tier-1 sites never took place. The reconstruction job creation and submission was data driven during this period.

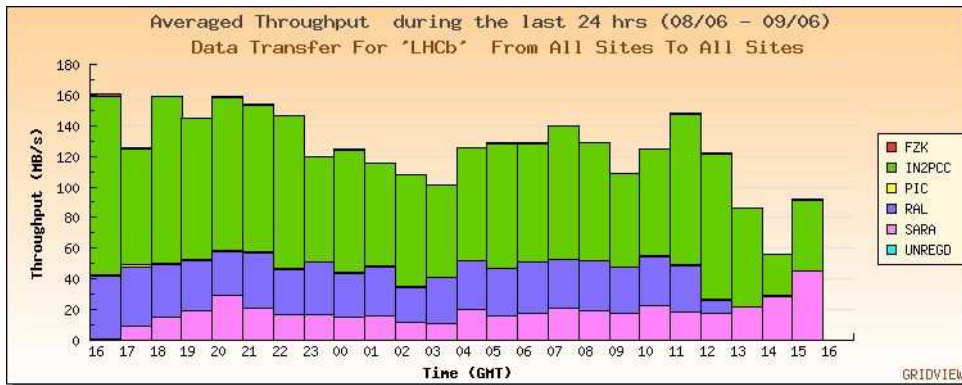


Figure 3. The transfer of data from CERN to the Tier-1 sites in autumn 2006, as a part of WLCG SC4.

The second use case is the upload of data to various storage elements by jobs on the worker nodes. These jobs use LCG utilities to upload data to one of various sites, depending on both, job requirements and the availability of the site. They may then set a request on a VO-box depending on the following conditions :

- (i) The destination(s) required are not available. In this case, the upload of data is essentially a fail-over mechanism for data upload.
- (ii) Multiple destinations are required. In this case, the data is uploaded to one site, and the VO-boxes take care of replicating it to the other destinations as specified in the uploaded request.

Sustained total transfer rates of more than 50 MB/s were seen (figure 4) though the problem of site stability and availability affected this use case also.

5. Issues faced

Many problems have been seen during the operation of DC06. Most of the problems were caught by the DIRAC system, while a large number needed dedicated effort, sometimes over long periods. The more important of these problems are discussed below.

5.1. Data access on the worker nodes

As mentioned before, LHCb does not copy data on to the worker node, but accesses it directly from the storage element, using root. There were many incompatibilities seen in the interactions

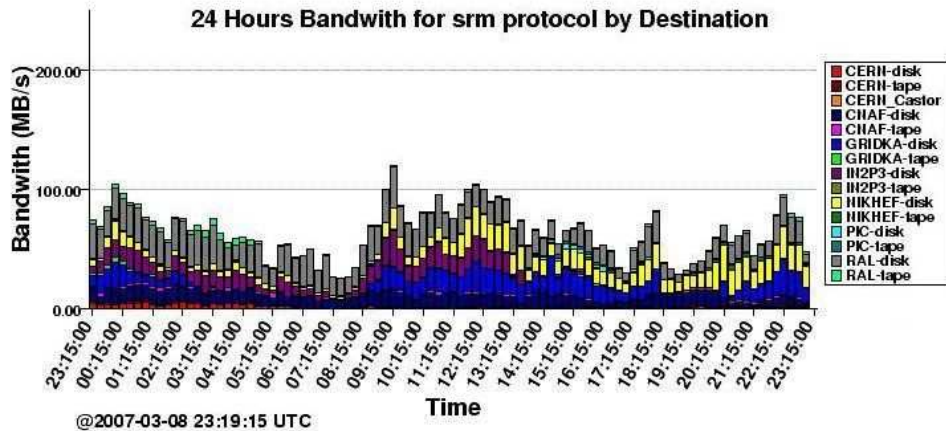


Figure 4. Upload of data from the worker nodes to the various storage elements.

between root and the storage element, especially when a dCache SE does not support the dcap protocol. It was also found necessary to pre-stage the files on the tape system to reduce latency in the job running. This pre-staging also gives us an early warning of storage problems when the time to stage files becomes excessively long.

5.2. SRM endpoint problems

All the sites have been affected by SRM endpoint problems at various times. The issues include bad srm configuration, server overload, storage space availability, etc. These have been the predominant problem seen in DC06 and these endpoint problems have caused drops in job efficiency or have led to sites being completely out of action. In a few cases, server overload caused namespace corruption – the system believed that the file existed (in the namespace) while it did not really exist in the storage, or vice-versa.

5.3. Problems with computing elements

The computing elements are of course crucial, since the jobs are routed by them to the worker nodes. Their configuration is hence very important and problems here can lead to wrong environments on the worker nodes - examples include jobs failing due to wrong / incompatible versions of libraries being picked up. Batch queue manager problems have also been seen at various sites. On occasion, a single node – or a few nodes – can act as a black hole for various reasons. In such a case, the node keeps pulling in jobs and failing them, causing the LHCb queue to drain without any useful output being produced. This is a problem when a CE manages a heterogenous set of machines with different hardware specifications and the LHCb jobs fail on only a few of these cpu-s.

5.4. Grid middleware and services

The grid middleware and services are well behind schedule. Thus, the needed functionality is either in early stages, not robust enough, or simply does not exist. BDII server overloads, failures in staging of files and firewall problems happen occasionally and have mostly been understood and fixed. Proxy lifetime issues, and in general problems with VOMS have led to access permissions errors, unexpected role changes and job failures, and continue to be a recurrent theme during DC06. And then there are the painful and sometimes unforeseen problems including fires and cooling breakdowns, etc happening in the wrong place at the wrong time.

6. Summary

The LHCb computing model has been successfully exercised. All the requested Monte Carlo has been successfully simulated using many (> 120) sites on the worldwide computing grid. The profile of LHCb running jobs over the grid, during the last year is shown in the figure 5.

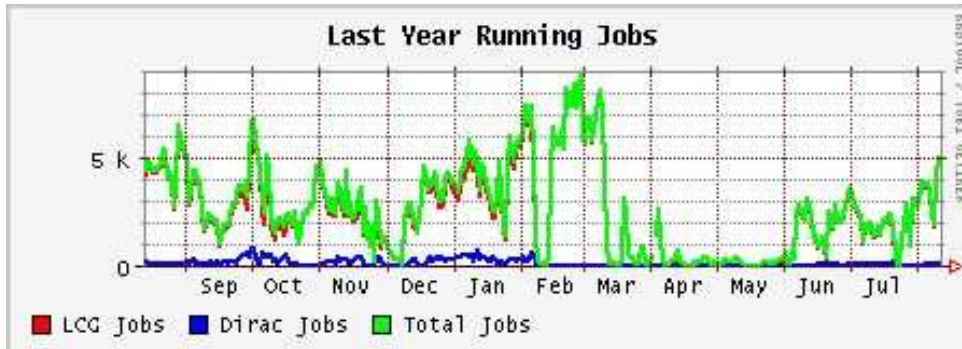


Figure 5. The history of total LHCb jobs on the grid in the last year.

Reconstruction and stripping of this data is currently going on at the various LHCb Tier-1 sites. Satisfactory data rates for transfers between various sites have also been observed, both using FTS for transfers between Tier-1 sites and for srm transfers from Tier-2 to Tier-1 sites. However, many problems affecting the above have been seen, with quite a few important issues still unsolved as yet.

References

- [1] LHCb Collaboration [S. Amato et. al.], LHCb Technical Proposal, CERN / LHCC / 98-4
- [2] LHCb Computing Model [N. Brook et. al.], LHCb/2004/119; CERN/LHCb/2004-119; 16 Dec 2004
- [3] DIRAC – the LHCb Data Production and Distributed Analysis system [A. Tsaregorodtsev et. al.], Proc. 2006 Conference for Computing in High Energy and Nuclear Physics, Mumbai, India, 2006
- [4] Ganga: a Grid user interface for distributed analysis [K. Harrison et. al.] Proc. Fifth UK e-Science All-Hands Meeting, Nottingham, UK, National e-Science Centre, Sept 2006, pages 518-525