The LHCb Computing Model and Real Data

Philippe Charpentier
PH Department, CERN, CH-1211 Genève 23
Philippe.Charpentier@cern.ch
On behalf of the LHCb Collaboration

Abstract. After several years of experience with Grid production and Analysis dealing with simulated data, the first LHC collision data (as of March 2010) have confronted the LHCb Computing Model with real data. The LHCb Computing Model is somewhat different from the traditional MONARC hierarchical model used by the other LHC experiments: first pass reconstruction, as well as further reprocessing passes, are performed at a set of 7 Tier-1 sites (including CERN), while Tier2 sites are used mainly for simulation productions. User analysis is performed at LHCb Analysis Centers for which the baseline is the 7 Tier1s. Analysis relies on the concept of reduced datasets (so-called stripped datasets) that are centrally produced at the 7 Tier-1’s and then distributed to all the analysis centers. We shall review the performance of this model with the 2010 real data, and give an outlook for possible modifications to be put in place for the 2011 run.

1. Introduction

The LHCb experiment [1] at the Large Hadron Collider (CERN) is dedicated to the study of CP violation in the sector of the beauty (b) quark. Its characteristics are quite different from the General Purpose Detectors (ATLAS, CMS): its geometry is not surrounding the interaction point but the sensitive detectors are arranged in such a way as to detect the particles produced in the forward direction, i.e. at a small angle with respect to the proton beam. LHCb searches for rare decays of particles containing the b-quark that live for a very short time (~1.5 ps) and therefore decay in flight a few millimeters away from the proton-proton interaction point. The baseline design of the LHCb detector assumes a low average number of collisions per beam crossing (µ) of around 0.4, in order to avoid confusing secondary interactions with b-quark decays.

2. The LHCb Computing Model [2]

2.1. Basic data characteristics

Thanks to the specific geometry of the LHCb detector, the small number of collisions per beam crossing, as well as an aggressive compression of raw data, the expected nominal event size is only 35 kB. The nominal reconstruction work is 12 HS06.s per event.

As LHCb is studying rare decays of the b-quark, an aggressive event selection has to take place in real time through hardware and software triggers. However, in order to stay within the time constraints
of real time while having a reasonable trigger efficiency, the rate of event data taking has to be of the order of 2000 events per second [3].

Before physics analysis takes place, the number of events has to be further reduced by a factor around 10. This “stripping” process takes place after the full reconstruction of the events, and produces a set of a dozen “streams” of analysis datasets (called DSTs). The work needed to select the events in the stripping process is around 1 to 1.5 HS06.s, i.e. 10 times faster than the event reconstruction.

2.2. The Role of Grid Tiers in LHCb Computing
Despite the small event size and fast reconstruction, and due to the high event rate, the total amount of CPU work per second is of the order of 26 kHS06.s, corresponding to about 2000 modern CPU cores. Because this power is large compared to the expected resource pledges to LHCb, the Computing TDR proposes to not only use the Tier0 resources for first pass reconstruction, but to profit from the distribution of raw data to Tier1s for distributing the reconstruction work. The Tier0 represents about 20% of the total resources.

From LHCb’s experience of distributed computing on the Grid, it appeared clearly that data management and data access is the biggest challenge. Since the LHCb operations team is quite small (around 12 people including the Tier1 contacts), it was decided to also use the Tier1 computing resources for data analysis, and to assign the heavy task of event simulation to all other sites on the Grid (Tier2s and other opportunistic resources). LHCb requires a large number of simulated events in order to identify all possible background sources as well as evaluate the signal efficiencies. These needs for CPU resources match well the non-Tier1 available resources for LHCb.

LHCb gets resources from 6 Tier1s: CNAF, KIT, IN2P3, NL-T1, PIC and RAL. Raw data are collected on the CERN Tier0 and immediately distributed to the Tier1s (one copy across all sites), using the CPU pledges of the sites for determining the apportionment. Whole data taking runs (about one hour in time, thus about 250 GB) are assigned to a given Tier1 site, such that after reconstruction and streaming the DSTs can be merged into large files (5 GB) at a single site. Reconstruction jobs are generated automatically and submitted to Tier1s as soon as raw datasets have been distributed. Each Tier1 reconstructs a fraction of its raw datasets only in order to take into account that some runs are reconstructed at CERN.

![Figure 1. Schematic description of the LHCb Computing Model.](image-url)
2.3. The LHCb Grid Experience

Up to now, the only experience on distributed computing on the Grid was coming from simulated data. LHCb have developed a fully integrated Grid solution called DIRAC. DIRAC is based on the existing middleware and Grid infrastructures. It integrates a Workload Management System (WMS) as well as a Data Management System (DMS).

Several lessons had been learnt from the LHCb experience of several years and weaknesses of the Grid computing were mitigated by several adaptations of the DIRAC system:

- Workload Management System: the still large failure rates of jobs on the Grid was mitigated by the usage of “pilot jobs”, building a resource overlay through late binding of computing slots to the actual jobs. LHCb used on all Tier1s so-called “multi-user pilot jobs” that allow each pilot to execute in turn several tasks, e.g. reconstruction and user analysis. This paradigm was extremely successful and effective in terms of required resources.

- Data access: reconstruction jobs last many hours, and therefore job failures may occur from loss of connection between the worker node and the file server. This problem was mitigated by downloading the input file to the worker node and not using remote file access during the whole processing step.

3. The 2010 Real Data

3.1. The 2010 Data Taking Conditions

Due to various limitations in the capabilities of the LHC, it was decided to achieve high luminosity by increasing the number of protons per bunch as well as the focusing of the beams, rather than increase the number of proton bunches in the machine.

The immediate consequence was that the average number of collisions per crossing greatly exceeds the nominal $\mu=0.4$ for LHCb. However in order to take advantage of the delivered luminosity, it has been decided to take data at the maximum luminosity allowed by the trigger system, i.e. up to $\mu=2.5$.

The high value of $\mu$ has two primary consequences: an increase in event size (due to higher pile-up of collisions) and an increase in processing time (due to the higher complexity of events).

![Figure 2. Event size as a function of the average number of collisions per crossing $\mu$.](image)

![Figure 3. Average CPU work (in HS06.s) as a function of $\mu$.](image)
event size and processing time are well in line with the expectations. The raw event size was between 1.5 and 2 times larger than the nominal size.

The DST event size followed the same trend as the raw event size and was around 130 kB on average for an expected size of 85 kB at nominal conditions.

3.2. Adaptability of the Computing Model

During this first year of data taking, the beam conditions have been varying substantially. In an initial phase, the collision rate was smaller than the data taking capacity of 2000 events/s, therefore a very loose trigger was applied and all collisions were recorded (minimum bias data). As the luminosity and collision rate increased tighter and tighter trigger selections were applied in order to remain within 2000 events/s.

For the minimum bias data (first 14 nb⁻¹), no stripping could be applied, as the aim was to study the characteristics of proton-proton collisions at 7 TeV. Therefore, despite the low integrated luminosity, the disk space used by these datasets was quite large.

As soon as a trigger was applied and $\mu$ was increasing, a stripping selection was required. However, due to the unforeseen large multiplicity in the events, the early stripping selections were quite ineffective at selecting events and were requiring a much larger work than anticipated (up to 50 times more for a retention close to 100%). In order to provide data for physics analysis as well as for developing more powerful selections, the data were reconstructed in quasi real time and processed by the stripping and streaming application. This was made possible by the availability of all disk pledges for 2010, but is not sustainable in the long term.

Because of the large demands in CPU work, and so that jobs processing a single raw file could fit in the work budget allowed by the Grid queues at Tier1s, LHCb had to decrease the raw data file size from the nominal 3 GB to 1 GB. This reduction increases the number of jobs to be submitted and monitored as well as the number of files to be handled by a factor 3.

The size of the DST datasets made it necessary to continuously monitor and manage the disk storage space at Tier1s, e.g. by reducing the number of replicas from the nominal 7 down to 4 or 3 depending on the processing passes.

Despite the prompt actions to reduce the file sizes and to release improved versions of the applications, LHCb suffered during summer 2010 from a large number of job failures due to CPU or memory resource excess. However the LHCb production system, based on DIRAC, re-created new jobs for processing those files that couldn’t be processed in a first instance and eventually after several retries all files were processed without any human intervention.

The datasets derived from this first processing were used to produce improved versions of the applications that were eventually used in the last part of data taking for the prompt reconstruction and stripping. For safety, the small raw file size of 1 GB was kept throughout the whole data taking.

3.3. The 2010 Data Sample

Figure 4 shows as a function of time the integrated luminosity delivered to and recorded by LHCb until the CHEP10 conference. Out of 21.9 pb⁻¹, LHCb recorded 19.7 pb⁻¹, corresponding to an overall efficiency of 91.2%. Most data were collected in the last few weeks of data taking.
Up to the time of the conference, 65.7 TB of raw data were distributed to the Tier1s. Slightly more data were recorded in Castor at CERN, due to some calibration runs, not distributed. One can see (Figure 5) that the amount of data does not follow the luminosity, on the contrary, the trigger selection improves with time and therefore the amount of data per unit of luminosity decreases with time. However the processing time does depend on the amount of data and not on the luminosity.

The share of distribution of raw data to Tier1s corresponds to the published CPU pledges at the beginning of 2010. When a Tier1 is temporarily unavailable, its share is set to zero and it recovers its fair share when coming back into production.

Figure 5. Amount of raw data collected on the Tier0 (left) and distributed to each of the 6 Tier1s (right).

4. LHCb Grid Usage with 2010 Data

4.1. Global Grid Usage

Figure 6 shows since the start of the LHC to the time of the conference the amount of CPU work (not normalized) in hours per second (i.e. a value of 1 corresponds to 3600 CPU cores at work). Monte-
Carlo simulation (outside Tier1s) represents about 50% of the resources, while user analysis (29%) is larger than data reconstruction. The peak structure corresponds to periods of active data taking and to simulation requests from the physics groups. It is clear that Tier1 usage is increasing with time, due to the increase in amount of collected data.

4.2. Reconstruction and Stripping Jobs
As can be seen on Figure 8 and Figure 9, reconstruction and stripping jobs have been following very closely the data taking, as the peaks do represent bunches of real data. During end of August and beginning of September a full reprocessing of all existing data was taking place, corresponding to the largest peak.
4.3. User analysis on the Grid
In LHCb over 250 physicists have been running analysis jobs on the Grid, with a very large spectrum in the number of jobs and amount of CPU resources used. Analysis took place at all Tier1 sites as well as CERN, as shown on Figure 9. All LHCb Grid Analysis jobs have been submitted to DIRAC through the ganga tool.

4.4. Success rates
Overall 81% of the jobs submitted through DIRAC to the Grid were successful. It should be noted that in case of failure, jobs are automatically resubmitted, unless the failure comes from an irrecoverable error (e.g. corrupted data or application bug).

The main causes of failures are:
- Simulation jobs may enter under certain rare conditions into an infinite loop within the Geant4 stepping code. This causes the job to exhaust CPU resources and fail. New jobs were then created to provide the requested amount of simulated events.
- Reconstruction jobs exceeding CPU resources or memory limit due to the much higher multiplicity in the detector. Those jobs were resubmitted with larger requirements.
- Data access problems from within the application may cause them to fail. This occurs due to authentication problems or temporary unavailability of files (e.g. when a disk server becomes unavailable).

4.5. Further adaptations of the Computing Model
It was foreseen in the LHCb Computing Model to use larger Tier2s also for user analysis. Some test setups have been put in place, but no real need occurred during 2010.

It was also realized that in order to free CPU resources at Tier1s it might be more advantageous to run reconstruction at Tier2s. Although they do not hold data, it is very fast to download the input data from a Tier1 and upload the output to that same Tier1. Since data accesses are one of the large failure causes on the Grid, using the Tier2s for much more controlled activities may reduce the needs in manpower for solving those problems.

5. Conclusions
Despite several difficulties mainly related to its unforeseen running conditions, LHCb made best use of the Grid resources it was allocated. The biggest challenges remain the data management (monitor and control disk space usage) and data access (due to temporary failures in storage systems). Running jobs in a controlled environment (production) for real data reconstruction and simulation is now routine work, although reaching a full 100% success rate requires still a substantial effort from the LHCb Operations team.

References
[1] The LHCb Collaboration: The LHCb Detector at the LHC, URL http://dx.doi.org/10.1088/1748-0221/3/08/S08005.
[2] LHCb Computing TDR, CERN/LHCC 2005-019, URL http://doc.cern.ch/archive/electronic/cern/preprints/lhcc/public/lhcc-2005-019.pdf
[3] LHCb Trigger TDR, CERN/LHCC 2003-031, URL http://lhcb-doc.web.cern.ch/lhcb-doc/progress/Source/LHCC/Sep_2003/trigtdr.pdf
[4] Tsaregorodtsev A. et al, 2008 J. of Phys.: Conf. Ser. 119 062048 URL http://stacks.iop.org/1742-6596/119/i=6/a=062048
[5] Maier A. et al, 2010 J. Phys.: Conf. Ser. 219 072008 URL http://dx.doi.org/10.1088/1742-6596/219/7/072008