Recent and planned changes to the LHCb computing model

M Cattaneo\textsuperscript{1}, P Charpentier\textsuperscript{1}, P Clarke\textsuperscript{2} and S Roiser\textsuperscript{1}

\textsuperscript{1} European Organization for Nuclear Research (CERN), Geneva, Switzerland
\textsuperscript{2} School of Physics and Astronomy, University of Edinburgh, Edinburgh, U.K.

E-mail: marco.cattaneo@cern.ch

Abstract. The LHCb experiment \cite{1} has taken data between December 2009 and February 2013. The data taking conditions and trigger rate were adjusted several times during this period to make optimal use of the luminosity delivered by the LHC and to extend the physics potential of the experiment. By 2012, LHCb was taking data at twice the instantaneous luminosity and 2.5 times the high level trigger rate than originally foreseen. This represents a considerable increase in the amount of data which had to be handled compared to the original Computing Model from 2005, both in terms of compute power and in terms of storage. In this paper we describe the changes that have taken place in the LHCb computing model during the last 2 years of data taking to process and analyse the increased data rates within limited computing resources. In particular a quite original change was introduced at the end of 2011 when LHCb started to use for reprocessing compute power that was not co-located with the RAW data, namely using Tier2 sites and private resources. The flexibility of the LHCbDirac Grid interware allowed easy inclusion of these additional resources that in 2012 provided 45% of the compute power for the end-of-year reprocessing. Several changes were also implemented in the Data Management model in order to limit the need for accessing data from tape, as well as in the data placement policy in order to cope with a large imbalance in storage resources at Tier1 sites. We also discuss changes that are being implemented during the LHC Long Shutdown 1 (LS1) to prepare for a further doubling of the data rate when the LHC restarts at a higher energy in 2015.

1. Introduction
This paper describes the evolution of the LHCb computing model that has been necessary in order to accommodate within limited computing resources the computing requirements of the expanding physics programme of the LHCb experiment. After a brief summary of the evolution of LHCb data taking conditions in section 2, section 3 summarises the computing model as had been foreseen at the time of the Computing Technical Design Report (TDR) \cite{2} and highlights some shortcomings of this model. The changes that have been made or that are foreseen to address these shortcomings are described in sections 4 and 5 for data processing and data management models respectively.

2. LHCb data-taking conditions
The LHCb experiment has taken data between December 2009 and February 2013. The data taking conditions and trigger rate were adjusted several times during this period to make optimal use of the luminosity delivered by the LHC and to extend the physics potential of the experiment.
Table 1 shows the evolution of the data-taking parameters that have a direct influence on the computing resources required to store and process the data. The assumptions made for the 2015 forecast are described in [3].

|                          | Design (TDR) | 2012 (actual) | 2015 (forecast) |
|--------------------------|--------------|---------------|-----------------|
| Instantaneous luminosity (cm$^{-2}$s$^{-1}$) | $2 \times 10^{32}$ | $4 \times 10^{32}$ | $4 \times 10^{32}$ |
| Mean visible p-p interactions/crossing ($\mu$) | 0.4 | 1.7 | $\sim 1.0$ |
| Raw event size (kB) | 25 | 60 | 60 |
| HLT output rate (kHz) | 2.0 | 5.0 | 12.5 |

The LHCb trigger rate is kept constant throughout all fills of protons in the LHC. This is because LHCb operates at a constant instantaneous luminosity, achieved by continuously adjusting the overlap between the colliding beams (luminosity levelling).

3. The LHCb computing model

3.1. The Technical Design Report (TDR) computing model

The computing model described in the LHCb Computing TDR [2] has the following features:

- RAW data recorded by the experiment is copied to the CERN Tier0 site and written to tape. A second copy is distributed to the six LHCb Tier1 sites (CC-IN2P3, CNAF, FZK, NIKHEF, PIC, RAL) with as equal as possible shares at each site.
- A first pass reconstruction is executed shortly after the data is taken. In this activity, CERN is considered as a Tier1, and therefore processes one seventh of the RAW data. The remaining reconstruction jobs are executed at the Tier1 sites holding the corresponding RAW data.
- A reprocessing reconstruction of the whole year’s data is carried out at the end of each year of data-taking, and should be completed within two months. The jobs are shared as for the first pass reconstruction, but with the possibility also to execute a fraction of the jobs on the High Level Trigger (HLT) farm.
- A “stripping” pass, run at the same site as the reconstruction, follows each reconstruction pass. Several hundreds of different physics group selections are executed; the selected events are written to one or more of about ten output streams. Further “stripping” passes are carried out as needed to accommodate new physics selections. A retention fraction of the order of 10% is expected.
- The stripped datasets are replicated to all six Tier1 sites plus CERN. They are the input to user analysis and to further centralized processing by physics working groups. Users do not have access to RAW data nor to non-stripped reconstruction output.
- All jobs execute at sites that hold the input data. Simulation jobs can run anywhere, including Tier2 sites where there is no disk. All other job types (user analysis, reconstruction, stripping) can run at Tier1s or CERN but not at Tier2s.
- Simulated datasets are replicated to CERN and to three of the six Tier1 sites.

3.2. Shortcomings of the TDR model

The TDR processing model is rather inflexible in the use of CPU resources. In particular, the Tier1 CPU needs to be sized to accommodate the peak power required to complete the end of year reprocessing within two months, as shown in figure 1. In the TDR it was foreseen to smooth these peaks by making use of the HLT farm for the reprocessing, but this turned out to be impractical because the reprocessing schedule was incompatible with the need of maintenance of the online infrastructure during the short winter shutdowns, and because the network and disk topology of the
online farm is not adapted to uploading the input RAW data to the worker nodes from the Tier0 storage.

![Figure 1: Simulation of processing power required at Tier1 as a function of time. The large peaks correspond to reprocessing campaigns, their size is proportional to the accumulated luminosity in a given year.](image)

The TDR data management model is very demanding on storage space, because all sites are treated equally regardless of available space. The requirement to have complete copies of all datasets at all analysis sites is wasteful, because it does not take into account the popularity of the datasets. The requirement to have all disk resources located at Tier1 or at CERN limits the ability of LHCb to attract investment in disk, in particular from countries that do not host a Tier1 site.

4. Changes to the LHCb data processing model

4.1. Changes implemented in 2012

In order to take advantage of processing power at Tier2s, and therefore reduce the peak power needed at Tier1s, it was decided in 2011 to “attach” Tier2 sites to Tier1 storage elements. The concept, tested in the 2011 end of year reprocessing as described in [4], consists of downloading each RAW file (3 GB) from a Tier1 storage element to a Tier2 site, running the reconstruction job at the Tier2 site (lasting about 24 hours of wall-clock) and then uploading the reconstruction output file to the same Tier1 storage.

In 2012, this concept made it possible to introduce a new strategy for producing the first full processing. Originally, the first pass “prompt” processing had been conceived for making all of the data available for analysis within a few days of data-taking, but in practice only limited datasets were used to rapidly produce fast preliminary results in flagship analyses. The bulk of the high precision measurements waited for the end of year reprocessing and a more definitive calibration and alignment. Therefore in September 2012 it was decided to reduce the initial prompt processing to reconstruct only sufficient data to determine the calibration and alignment needed for the full processing reconstruction and to assess the quality of the data. For this initial prompt processing only about 30% of the RAW was reconstructed, and this was done exclusively at CERN. The calibration and alignment procedures were sufficiently mature to allow the generation of definitive constants within 2-4 weeks of data-taking, after which delay the full reprocessing could be run on the entire RAW dataset, using the resources available at Tier1 and Tier2 centres. In this way the complete fully processed and calibrated 2012 dataset was available to physicists within days of the end of the 2012 data-taking period. Since a further end of year reprocessing of the 2012 data was then no longer necessary, the resources freed were used to reprocess again the 2011 data, using the same version of the reconstruction software used for the 2012 data. The profile of CPU usage for the reprocessing activities described above is shown in figure 2. Figure 3 shows the share of CPU resources used for reprocessing at the different production sites.
Figure 2: Distribution of CPU usage for reprocessing activities from September 2012 to February 2013. The first peak in September/October 2012 corresponds to the full processing of 2012 data taken until then. Subsequent peaks correspond to the release of new calibration constants 2-4 weeks after data-taking. The peak in February 2013 corresponds to the reprocessing of 2011 data. The power required for the continuous reprocessing of 2012 data during October to December never exceeded the power required for the 2011 reprocessing. This was achieved without the use of CERN resources (purple in the 2011 reprocessing peak).

Figure 3: Distribution of CPU usage by site for the reprocessing activities included in figure 2. 45% of the CPU work was provided by sites not holding the RAW data (Tier2+Yandex).

4.2. Changes foreseen for 2015
The architecture of the HLT is undergoing a major redesign in preparation for the restart of data-taking in 2015 after Long Shutdown 1 (LS1) [5]. The HLT dataflow is being split into two parts. HLT1 (consisting of a partial event reconstruction, selecting displaced tracks/vertices and dimuons) will run in real time. The selected events will be stored on the local disks of the HLT farm for several hours, after which the HLT2 (full offline-like event selection, mixture of inclusive and exclusive triggers) will be executed. A continuous calibration will run in the Online farm, producing calibration constants that can be used by the Particle Identification (PID) algorithms running in HLT2.

The PID calibration algorithms will be adapted from those that were running offline in the latter part of 2012, thus the quality of the PID in HLT2 will be equivalent to what can be achieved offline. This removes the need for running a first pass offline reconstruction as was the case in 2010-2012.
automatic validation of the output of the continuous calibration and of the stability of the detector alignment is sufficient to give the green light for an offline reconstruction whose quality will be equivalent to an end of year reprocessing, which is therefore no longer needed.

The HLT2 output will be split into three streams:
• “Full” stream, 5-10 kHz, to be fully reconstructed. It is foreseen to buffer the RAW data of this stream for up to two weeks on Tier1 disk, so as to allow time to correct any problems flagged by the automatic validation before launching the reconstruction and stripping. There will be no reprocessing of this RAW data, but it is foreseen to restrip all the reconstruction output at the end of the data-taking year, which can be an opportunity to apply more refined PID calibration constants to the datasets used for final analysis.
• “Parked” stream, 0-5 kHz. If there are insufficient resources on the Grid to reconstruct the full 10 kHz bandwidth of the Full stream within the maximum delay of two weeks, while leaving sufficient resources for user analysis and the highest priority simulation, up to 5 kHz will be moved to the “Parked” stream. Data in this stream will not be looked at until the next long shutdown. Current estimates of data rates and computing resources indicate that this stream will not be needed before 2017, but it will nevertheless be commissioned for the 2015 startup.
• “Turbo” stream, 2.5 kHz. Events in this stream do not need to be reconstructed offline; the analysis will exclusively use information from the HLT reconstruction, that will be added to the RAW data (adds ~10 kB/event to the RAW data size), and which will be stripped off for the offline analysis by a dedicated stripping application.

4.3. Going beyond the Grid paradigm
The changes described in this section have blurred the distinction between the roles of different Tiers, since several types of processing activities can now take place at all Tiers. Currently, production managers decide which sites should participate in a given type of production activity and manually update the DIRAC [6] configuration systems to attach or detach sites accordingly. In future, we imagine a system whereby sites declare their availability for a given activity and provide the corresponding computing resources.

The flexibility of the LHCbDirac Grid interware [7] has allowed LHCb to easily include additional computing resources from sites outside WLCG (such as the LHCb HLT farm) or even from outside of High Energy Physics (such as the Yandex™ Internet company [8]). From January to September 2013, about 20% of CPU resources consumed by LHCb were provided by the HLT farm, and 6.5% by Yandex.

Additional infrastructures are being deployed as well:
• Virtual machines running on cloud infrastructures collecting jobs from the LHCb central task queue [9].
• Virtual machines created and contextualised for virtual organisations by remote resource providers (vac concept) [10].
• Volunteer computing using the BOINC infrastructure enabling payload execution on arbitrary compute resources [11].

5. Changes to the LHCb data management model
The increased trigger rate and expanding physics programme of LHCb are putting strong pressure on storage resources.
For tape, shortages have been mitigated by reducing the volume of tape dedicated to long-term archive for analysis preservation. In particular, all archives of derived data (stripping output, simulation output) have been reduced to a single tape copy. In principle these datasets could be regenerated, but in practice this becomes operationally unaffordable for datasets more than a few years old and some risk of data loss has to be accepted.
The rest of this section outlines measures that have been taken or are foreseen to address disk shortages.

5.1. Changes to data formats
The highly centralized LHCb data processing model allows optimizing data formats for operation efficiency. All datasets are stored as ROOT files, but the compression level can be optimised according to the function of the dataset: intermediate temporary files use the fastest compression algorithm available in ROOT (ZLIB:1); for the permanent files that are disk resident for analysis use the more efficient LZMA:6 compression is applied [12].

Generalised use of the Gaudi framework [13] for all stages of the LHCb data processing (from HLT through to the final analysis) shields all the data processing algorithms from the implementation of the persistent data classes. A set of classes highly optimised for storage has been introduced, with an automatic conversion to/from the transient classes seen by the applications, where all floating point numbers are saved as integers with the adequate range of precision.

There is a strong drive to optimise the event data content of the different datasets according to the foreseen usage. The reconstruction output (FULL.DST) now contains a copy of the RAW data to avoid staging separate files when adding RAW information to events selected by stripping as was foreseen in the original computing model. The DST originally foreseen for most analyses contains the same information as FULL.DST for the selected events, as well as additional information stored by the stripping algorithms. Many exclusive analyses however only require information relating to the few reconstructed particles of the selected decay tree. A microDST format has been developed containing this reduced information. Many iterations were required to get the content correct, but the gain in space is considerable, from ~120 kB/event on DST to ~13 kB/event on microDST.

5.2. Data placement strategies
It was quickly realised that the data replication policy foreseen in the computing TDR is too generous, and fewer replicas are required for operational efficiency. A data driven procedure is now in place that, when a new analysis dataset is produced, automatically archives a copy on tape, and makes a total of 4 disk replicas of real data and 3 disk replicas for simulated data. The disk replication sites are chosen at random using the remaining free disk space as a weight; for real data whole runs are kept together at the same site, whereas for simulated data the choice of site is per file. The replication algorithm prevents disk space at a given site from becoming saturated, with an exponential fall-off of free space, assuming of course that there are sufficient sites with some available space.

The computing model specifies policies for the retention on disk of previous versions of datasets (previous reconstruction, previous stripping pass). Defining the current processing as version “n”, for version “n-1” the disk replicas are reduced to 2. The choice of replica to remove is random, but there is the possibility to preferentially remove replicas from sites with less free space. For version “n-2” only the tape archive replica is kept.

5.3. Data popularity
The data replication policy described in the previous section only takes into account the processing history and not how popular a given dataset is for analysts. Any deviations from the standard policy need to be carefully prepared with physics groups to identify datasets that are no longer of general interest. Since all data accesses by Grid jobs take place via the DIRAC framework, it was possible to instrument the framework to record the data accesses of all jobs. Since May 2012, the following information has been recorded for all jobs: dataset path, number of files for each job, storage element used. This information currently allows identification of unused datasets by visual inspection.

It is planned to enhance this system by establishing, for each dataset, the last access date and the number of accesses in last (n) months (1<n<12), normalising the number of dataset accesses to the dataset size. Summary tables of accesses per dataset will be prepared that, together with the
knowledge of the storage usage at each site, will allow to trigger replica removals when space is required.

5.4. Use of disk at Tier2 sites
In the LHCb computing model, user analysis jobs requiring input data are executed at sites holding the data on disk. So far Tier1 sites (and CERN) were the only ones that LHCb asked to provide storage and computing resources for user analysis jobs. Early in 2013 we introduced the concept of Tier2D: these are a limited set of Tier2 sites that are allowed to provide disk capacity for LHCb, where a replica of physics analysis files can be stored, thus enabling user analysis jobs to run at these sites. This blurs even more the functional distinction between Tier1 and Tier2: a large Tier2D is equivalent to a small Tier1 without tape.

The minimal requirement for such a site has initially been set at 100 TB of disk, but this will ramp up to 300 TB per site in 2014. Currently two sites are in production and a further two are in the final stages of commissioning. A fifth site is being negotiated. An added benefit of this change to the computing model is that countries that do not host a Tier1 can now pledge disk to LHCb.

6. Conclusions
The LHCb computing model has evolved to accommodate within a constant budget for computing resources the expanding physics programme of the experiment. The model has evolved from the hierarchical model of the TDR to a model based on the capabilities of different sites. Further adaptations are planned for 2015, but no revolutionary changes should be needed to the model or to the frameworks (Gaudi, DIRAC) to accommodate the computing requirements of LHCb during Run 2.

References
[1] Augustor Alves Jr A et al 2008 The LHCb detector at the LHC J. Instrum. 3 S08005
[2] Antunes Nobrega R et al 2005 The LHCb computing technical design report CERN/LHCC 2005-019
[3] Albrecht J et al 2013 Implications of post-LS1 running conditions on LHCb's data processing CERN/LHCb-PUB 2013-008
[4] Roiser S et al 2012 Major changes to the LHCb grid computing model in year 2 of LHC data J. Phys.: Conf. Ser. 396 032092
[5] Neubert S et al 2013 The LHCb trigger architecture beyond LS1 These proceedings
[6] Tsaregorodtsev A et al 2008 DIRAC: a community grid solution J. Phys.: Conf. Ser. 119 062048
[7] Stagni F et al 2012 LHCbDirac: distributed computing in LHCb J. Phys.: Conf. Ser. 396 032104
[8] http://company.yandex.com/
[9] Ubeda Garcia M et al 2013 Integration of Cloud resources in the LHCb Distributed Computing These proceedings
[10] McNab A 2013 Running jobs in the vacuum These proceedings
[11] http://boinc.berkeley.edu/
[12] Canal P et al 2012 ROOT I/O improvements J. Phys.: Conf. Ser. 396 052017
[13] Barrand G et al 2001 Gaudi – a software architecture and framework for building HEP data processing applications Comp.Phys.Commun. 140 45-55