LHCb Computing Resources: 2013 re-assessment, 2014 request and 2015 forecast

LHCb Public Note

Issue: 1
Revision: 1

Reference: LHCb-PUB-2013-002
Created: 5th Feb 2013
Last modified: 26th Feb 2013

Prepared By: LHCb Computing Project
M. Cattaneo/Editor

Abstract

This document presents the computing resource estimates from LHCb. It should be considered an update of the documents LHCb-PUB-2012-014 and LHCb-PUB-2012-015 submitted to the C-RSG in September 2012.

Table of Contents

List of Tables

1. Introduction

This document summarizes the most recent updates of LHCb computing resource usage estimates. It covers the period 2013-2015[1] and is based on the latest measurements of the LHCb computing parameters and latest updates to the LHCb analysis plans.

The estimates for 2013 have been updated relative to those presented in LHCb-PUB-2012-014 by taking into account the measured volumes of data collected in the 2012 pp run and 2013 p-Pb/Pb-p run and the DST sizes arising from the reprocessing and restripping of the 2011 and 2012 datasets that was completed in December 2012 (for the 2012 data) and in February 2013 (for the 2011 data).

The request for 2014 has been refined following a reassessment of simulation needs.

A preliminary request for 2015 is presented, taking into account our best knowledge of data taking parameters for the 2015 run; it should be stressed however that during the first half of 2013 LHCb is reviewing its HLT strategy for 2015 and its computing model, so significant changes may take place in the request to be presented at the October 2013 C-RRB.

2. Recent changes in the LHCb Computing Model

A detailed description of the LHCb Computing Model can be found in LHCb-PUB-2012-014 and LHCb-PUB-2011-009. We summarize here changes that have taken place since September 2012.

- LHCb suffered a serious shortage of tape during the autumn of 2012, due to the combined effects of a 2012 total pledge lower than requested, an increase by 40% in the length of data-taking in 2012, and the introduction of the FULL.DST format as output of the reconstruction. Part of the shortage was absorbed by a number of sites that installed the 2013 pledges ahead of schedule, but part had to be mitigated by reducing the number of archive copies of processed data (Simulated and Real Data DSTs) from two copies to one copy.
- The reprocessing of 2012 data (Reco14) ran entirely outside CERN, because CERN was dedicated to prompt processing of new data. To speed up the reprocessing, extensive use was made of Tier2 sites that were associated to a Tier1 site from which they received the RAW data and to which they uploaded the FULL.DST. Since CERN was not involved in the reprocessing, it did not receive any of the Reco14 FULL.DST and will therefore not participate in future restrippings of these data.
- We have added provision for 1TB of disk per user in CERN-EOS, for storage of user data with world-wide interactive access. This is intended as an eventual replacement of CERN-CASTOR home directories. This is in addition to the already existing user grid storage accessible from grid jobs.

[1] For the purpose of this document a given year always refers to the period between April 1st of that year and March 31st of the following year.

3. Model for 2013-2014 resources estimates

3.1. Measurements of 2011 and 2012 data processing

The data collected by LHCb in 2011 and 2012 were fully reprocessed between September 2012 and February 2013, using consistent reconstruction and detector calibration and alignment models for the two years (Reco14). Following the LHCb computing model, the output of the reconstruction (FULL.DST) was stored as a single copy on tape. The data was then stripped to select events to be used by subsequent physics analysis (Stripping20). Since these datasets are destined for final analyses, they were replicated on four disk copies on the Tier1 sites. A single archive copy is kept on tape.

Table 3-1 shows the resources used by the Reco14 and Stripping20 processing passes. These measurements are used to estimate the resources needed to further process these data during 2013 and 2014.

Processing pass	CPU	Tape	Disk
	kHS06*y	TB	TB
Reco14, 2011	11.0	1500	
Reco14, 2012	19.3	2790	
Stripping20, 2011	4.1	100	400
Stripping20, 2012	6.4	222	888

Table 3-1: Resources used for 2011 and 2012 data reprocessing in 2012-2013

In addition, LHCb collected heavy ion collisions (p-Pb) data[2] during January and February 2013. These data were fully reconstructed but not yet stripped. The resources used for the p-Pb reconstruction are shown in Table 3-2

Processing pass	CPU	Tape
	kHS06*y	TB
p-Pb Reco, 2013	2.0	290

Table 3-2: Resources used for prompt reconstruction of 2013 p-Pb data

3.2. Measurements of 2011 simulation

At the end of 2012, we can consider that the analysis of the $1pb^{-1}$ of data collected in 2011 and reprocessed at the end of 2011 (Reco12) is approaching completion. We can therefore consider that the MonteCarlo simulation of the 2011 detector, consistent with the Reco12 dataset, is also completed. We refer to this simulation as MC11a.

Table 3-3 shows the resources used by the MC11a simulation campaign. These measurements are used to estimate the resources needed for simulation during 2013 and 2014.

[2] The heavy ion collision data was taken in both proton-ion (p-Pb) and ion-proton (Pb-p) configurations. In this note, whenever we refer to p-Pb data we implicitly mean both configurations

Processing pass	CPU	Tape	Disk
	kHS06*y	TB	TB
MC11a	47.4	586	1780

Table 3-3: Resources used for the MC11a simulation campaign

3.3. Plans for real data processing in 2013 and 2014

For the purposes of this document, we assume that data taking after Long Shutdown 1 (LS1) will not resume before April 2015. Thus there will be no new RAW data collected during 2013 and 2014.

The Reco14 FULL.DST will be the starting point for all Stripping passes that select data to be made available on disk for physics analysis. A first Stripping pass (Stripping20) on the Reco14 data was performed immediately following the reprocessing campaign. Further stripping passes are foreseen during 2013 and 2014; since the reconstruction will not change, these stripping passes will be *incremental*, meaning that they will add new selections but will not supersede existing ones. Thus any such restripping will add to the disk and tape space required for the analysis of the Reco14 data. For planning purposes we assume two restrippings in 2013 (spring and autumn) and one in spring 2014. We assume that each restripping will add approximately 20% to the volume of the original Stripping20 dataset. Each restripping requires running over the complete FULL.DST, that needs to be staged to disk from tape, and consumes an approximately constant amount of CPU time (assumed to be equal to the Stripping20 processing time), independent of the number of selections. The elapsed time for the restripping campaign is dominated by the rate at which the FULL.DST can be staged from tape.

During 2013, work will continue to improve the reconstruction and the detector calibration and alignment. This work will culminate in a definitive reconstruction for the Run 1 data, which will be used for a complete reprocessing of the 2011-2012 data and is intended to be the long-term replacement of all previous processings of the Run 1 data. We plan to execute this reprocessing (Reco15) during the second half of 2014, followed by a new stripping (Stripping21) that will eventually replace the Stripping20 datasets. The resources required for this reprocessing and restripping are assumed to be identical to those required for the Reco14 and Stripping20 processing passes. Since these data will become available for analysis late in 2014, any disk space freed by the removal of previous strippings will not be available before 2015. However, the Reco15 FULL.DST will replace the Reco14 FULL.DST, so the tape used by Reco14 can be recovered.

The p-Pb data taken at the beginning of 2013 will be reprocessed as soon as calibrations are available. A stripping must also be provided before these data can be analysed. By analogy with the 2011 and 2012 datasets, we assume that the stripping time per event will be one third of the reconstruction time, and that a factor of ten can be achieved as reduction factor for the size of one copy of the stripped dataset relative to the FULL.DST size.

In 2012 we introduced a data “Swimming” activity to understand lifetime biases in the charm analyses (see CERN-LHCb-PUB-2012-004). Contrary to expectations in that document, Swimming productions have not required any significant amount of additional CPU resources and this activity can be neglected for the purposes of this document.

3.4. Plans for simulation in 2013 and 2014

The analysis of the Reco14 dataset will require a corresponding MonteCarlo simulation (which we refer to as Sim08). By analogy with the analysis of the Reco12 dataset (which used the MC11a simulation), we consider that the number of simulated events required in Sim08 will be equivalent to the MC11a dataset for the 2011 analysis, and twice the number of MC11a events for the 2012 analysis (since the 2012 real data was twice the 2011 data, $2pb^{-1}$).

The parameters of the MC11a simulation (processing time per event, event size) are similar to those measured for the latest simulation; furthermore, the mix of event types requested is expected to be similar to MC11a (processing time per event in the event generation phase is strongly dependent on the event type) as is the fraction of productions for which we will choose to store only those events which pass the trigger and stripping emulation. For these reasons, we believe that our estimate of resources needed for Sim08 can be based on a simple scaling of the luminosity, i.e. a factor 3 in CPU, disk and tape relative to MC11a. Since the Reco14 data will be analysed during 2013, we hope to produce the bulk of this simulation during 2013.

During 2013 we also plan various evolutions of the simulation software:

- There is a strong activity to retune the simulation to better match the real data. This will culminate in a new version of the simulation (Sim10) towards the end of 2013
- The simulation will be used to optimise detector geometries in view of the TDRs for the LHCb Upgrade to be installed in LS2
- Simulations will be required to prepare the physics analyses and HLT tuning for the new LHC energy after LS1

Sim10 is expected to have similar CPU and storage requirements to Sim08, and can be considered a replacement of that simulation. We hope that it will come online before the completion of Sim08, so the number of events to be generated by the two simulations combined will not be doubled, we assume a factor of 2 relative to MC11a in addition to the Sim08 estimate. The Upgrade simulations will take more CPU per event and produce larger events (due to the higher pileup conditions in the Upgrade), whereas the post-LS1 simulations should produce smaller events (reduced pileup relative to 2012); overall we allocate another one MC11a equivalent for these two simulations

The relative priorities of the above simulations will be adjusted according to the priorities of the physics analysis. For the purposes of this estimate we assume a constant production rate throughout the years 2013-2014 using any CPU capacity not dedicated to processing of real data (reprocessing, restripping) or to user analysis.

We consider that, by 2014, any analyses still using MC11a will be completed. Thus the disk space occupied by MC11a will be released in 2014.

3.5. User analysis

The total CPU used for user analysis has been rather constant over the last two years, albeit with peaks in the usage corresponding to the run up to major conferences. For these estimates with assume a constant CPU requirement for user analysis at the same level as in 2011 and 2012, and a moderate increase of grid-enabled disk usage, at a rate of 20% per year. In addition, we are making provision for 1TB per user of disk in CERN-EOS, to provide user storage accessible from interactive usage worldwide, as an eventual replacement for CERN-Castor home directories

4. Model for 2015 resources estimates

4.1. Simulation in 2015

In 2015 we expect most analyses of 2011-2012 data to be in an advanced state and to have satisfied most of their simulation needs. Simulation efforts are likely to concentrate on further simulations for the LHCb Upgrade studies, and on tuning the simulation to the observed 2015 data-taking conditions. We expect to start a massive production for analysis of 2015 data in January 2016. For the purposes of this estimate we will consider that the CPU resources dedicated to simulation in 2015 will be two thirds of those foreseen for

2013 and 2014, with a corresponding increase in the tape archive needs. The increase in disk will be smaller, since by then it should be possible to reduce the number of disk resident copies of Sim08.

4.2. Data taking in 2015

There are still many unknowns in respect of the running conditions for 2015.

According to current planning, LHC is expected to resume operation for physics in the spring of 2015. As was already the case during the 2010 startup, LHCb plans to run with an open trigger initially, which will be tightened when the machine reaches the nominal LHCb luminosity. Because of luminosity levelling at the LHCb interaction region, it is expected to reach the nominal LHCb luminosity rather quickly.

Therefore, for the purpose of this document, we will assume that LHCb will take data at its nominal rate throughout the 2015 running period, which we assume to be a total of $5 \cdot 10^6$ seconds of stable beam stretching over approximately 7 month of machine operation.

In order to estimate the LHCb RAW data rate in 2015 we need to take into account the many changes that will take place compared to 2012:

- The average multiplicity of single pp interactions will be larger, due to the higher collision energy
- The effect of pileup on RAW event size is difficult to predict. If the LHC operates at 25ns bunch spacing the multiple interaction rate will be lower, but the left over signal from out of time events will be greater.
- LHCb may run at a higher (up to 50%) luminosity, that may translate into a higher High Level Trigger (HLT) output rate
- The allocation of HLT bandwidth to different trigger selections will depend on studies to be performed during LS1. It is too early to say whether this will lead to a significant change in the HLT output rate.

For this document we assume that the combined effect of the above changes will lead to an output rate of 700 MB/s, a factor of 2 larger than the maximum rate observed during the 2012 run, and close to the current hardware limitations of the HLT farm.

The total reconstruction time of events in LHCb has been shown to roughly scale with the total size of the RAW data sample, so for this estimate we use the same reconstruction time per MB as measured on 2012 data. As a first approximation the same retention rates as in 2012 are assumed for the Stripping.

According to current plans, we will implement during LS1 an automated calibration and alignment system for the detector that will allow us to have a “prompt” reconstruction performance of sufficient quality for the first publications, removing the need for an end of year reprocessing and the corresponding peak in CPU power. This has already been experimented successfully in 2012, when the data taken after September were reconstructed within two weeks of data-taking with a detector calibration of the same quality as used to reprocess the earlier data. We will however foresee a full restripping of the data at the end of data-taking, since the stripping selections can only be correctly tuned once we have seen the data at the new energy, new beam conditions and retuned HLT.

5. Updated estimates

5.1. CPU resources

Table 5-1 presents, for the different activities, the CPU work estimates for the period 2013-15 when applying the models described in the previous sections. Note that in this table we do not apply any efficiency factors: these are resource requirements assuming 100% efficiency in using the available CPU.

Work	2013		2014		2015	
	kHS06*y	%	kHS06*y	%	kHS06*y	%
Prompt Reconstruction					39	
Prompt Stripping					13	
Reprocessing Reconstruction	2		30			
Reprocessing Stripping	1		10			
Restripping	21		10		13	
Simulation	142		142		95	
User Analysis	11		11		11	
Total	**177**		**205**		**170**	

Table 5-1: Estimated CPU work needed for the different activities.

In the current computing model, User Analysis and Stripping can run at Tier0 or Tier1s where the input data are located. Simulation can run anywhere (including the HLT farm). Reconstruction is normally targeted at the Tier1 that holds the RAW data, with a 20% fraction targeted to CERN, but can also be assigned to Tier2 according to the availability of the different computing resources. Because of this, and because of the suppression of the end of year reprocessing from the computing model, the peak power required at the Tier1s is no longer determined by the necessity to complete the reprocessing in a relatively short amount of time, and CPU usage can be smoothed throughout the year by adjusting the relative priorities of the different activities, in particular by limiting the CPU allocated to simulation during reprocessing and restripping campaigns.

Table 5-1 shows the CPU work requirements integrated over a whole year. In order to derive a CPU power, we assume that CPU can be used at 85% efficiency averaged over the whole year. This includes both the efficiency of single jobs at keeping the CPU busy, and the operational efficiency of keeping the job queues full at all times. The resulting requirement is shown as the “requirement” row of Table 5-2.

As can be seen in Table 5-2, the pledged resources in 2013 are insufficient to satisfy the estimated needs, However, this is not a problem, since during 2013 and 2014 we can rely on the HLT farm to supply the missing CPU power. Similarly, our request for 2014 takes into account the resources available in the HLT farm and other sites such as Yandex that do not pledge resources to LHCb through WLCG, but which LHCb can use. In 2015, when data taking resumes, the HLT farm is no longer available, but the preliminary estimate suggests that the CPU increase requested for 2014, would be sufficient to satisfy also the 2015 needs.

Power	Pledge 2013	Available 2013	Request 2014	Estimate 2015
	kHS06	KHS06	kHS06	kHS06
Required		**209**	**241**	**201**
Tier0	34	34	40	40
Tier1	92	92	110	110
Tier2	47	47	47	47
Total WLCG	**173**	**173**	**197**	**197**
HLT farm		**30**	**30**	
Unpledged		**10**	**10**	**10**

Table 5-2: Estimated CPU power needed at the different Tier levels.

5.2. Storage resources

Table 5-3 presents, for the different data classes, the additional disk space required for period 2013-2015 when applying the models described in the previous sections. Note that in this and the next table we do not apply any efficiency factors: these are resource requirements assuming 100% efficiency in using the available disk.

The user disk increment for 2014 includes a request for an allocation of 200TB in CERN-EOS for user space, which should be on secure disk.

Disk	2013	2014	2015
	PB	PB	PB
Prompt Stripping			1.8
Reprocessing Stripping	0.1	2.1	-1.0
Restripping	0.5	0.3	1.8
Simulation	5.3	3.5	0.0
User	0.1	0.3	0.1
Total	**6.0**	**6.2**	**2.7**

Table 5-3: Break down of estimated Disk Storage increments for each of the different data categories that are included in the model.

Table 5-4 combines the increments presented in Table 5-3 with the current disk usage, to show the total forecast requirement for disk for the years 2013-2015.

Disk	2013	2014	2015
	PB	PB	PB
Disk Buffer	0.4	0.4	0.4
Real Data (Stripped DST)	5.0	7.3	9.9
Simulation	8.2	11.7	11.7
User	0.6	0.9	1.0
Other (Freezer, Debug)	0.1	0.1	0.1
Total	**14.3**	**20.4**	**23.1**

Table 5-4: Break down by type of data of estimated Disk Storage requirements for 2013-2015.

Table 5-5 presents, for the different data classes, the additional tape space required for period 2013-2015 when applying the models described in the previous sections.

Tape	2013	2014	2015
	PB	PB	PB
Raw Data			7.4
Prompt Reconstruction			5.6
Prompt Stripping			0.4
Reprocessing Stripping	0.0	0.5	
Restripping	0.1	0.1	0.4
Simulation	1.8	1.8	1.2
Total	**1.9**	**2.3**	**15.0**

Table 5-5: Break down of estimated Tape Storage increments for each of the different data categories that are included in the model.

Table 5-6 combines the increments presented in Table 5-5 with the current tape usage, to show the total forecast requirement for tape for the years 2013-2015.

Tape	2013	2014	2015
	PB	PB	PB
Raw Data	6.1	6.1	13.4
FULL.DST	3.9	3.9	9.4
Archive	5.3	7.7	9.7
Total	**15.3**	**17.6**	**32.6**

Table 5-6: Break down by type of data of estimated Tape Storage requirements for 2013-2015.

The disk and tape estimates shown in Table 5-4 and Table 5-6 have to be broken down into fractions to be provided by CERN and the Tier1s. According to the LHCb computing model, CERN holds one disk copy of all analysis datasets (i.e. one quarter of the total real data DST and one third of the simulation DST). It also holds a single copy of special datasets used for development and debugging of software ("Freezer"). For tape, the model assumes one copy of the RAW data at CERN and one copy distributed at the Tier1s; the FULL.DST should be shared according to the CPU shares so that re-stripping takes place in a balanced way at all Tier0 and Tier1 sites. Archive can be anywhere, we share it also in a balanced way. The results of this sharing are shown in Table 5-7 and Table 5-8, where the storage pledges for 2013 are also shown

Disk	2013 Pledge	2013	2014	2015
	PB	PB	PB	PB
Tier0	4.0	4.4	6.4	7.1
Tier1	7.0	9.9	14.0	16.0

Table 5-7: Disk estimates for each Tier level.

Tape	2013 Pledge	2013	2014	2015
	PB	PB	PB	PB
Tier0	6.5	5.5	6.6	11.8
Tier1	9.5	9.7	11.1	20.8

Table 5-8: Tape estimates for each Tier level.

Note that the tape estimates for 2013 and 2014 have decreased relative to those presented in LHCb-PUB-2012-014, due to the halving of the tape needed for archive, as discussed in Section 2

5.3. Tape staging bandwidth

The most demanding activity for what concerns recall of data from tape is the re-stripping. It requires staging in all the FULL.DST present at a site and processing them, either locally or at an associated Tier2. The processing rate of a stripping job is of the order of 4 events per second, i.e .500 kB/s.

Table 5-9 shows the amount of FULL.DST from the latest reprocessing of 2011 and 2012 data at each site. The rate is indicated for a re-stripping campaign lasting 60 days, together with the corresponding number of concurrent stripping jobs. The staging throughput and the number of jobs scale in inverse proportion to the duration of the campaign.

60 days			
Tape SE	**Total TB**	**Rate (MB/s)**	**# of jobs**
CERN-RDST	259.4	50	100
CNAF-RDST	794.6	153	306
GRIDKA-RDST	643.5	124	248
IN2P3-RDST	696.8	134	268
PIC-RDST	201.4	39	77
RAL-RDST	575.8	111	222
SARA-RDST	537.7	104	207

Table 5-9: Average tape reading bandwidth required for running the 2011+2012 re-stripping in 60 days.

5.4. Shortcomings of the pledging process

We would like to signal a long-term issue that tends to result in LHCb always receiving a total pledge that is less than its requirements. In the LHCb computing model, disk and tape resources are located only at Tier1 sites, plus CERN. We therefore express our requirements in terms of a CERN share (20-30% of the total) and a Tier1 share, and assume that the total of the Tier1 pledges satisfies the Tier1 request. Some countries pledge for a fixed share of the request; this would be OK if there were a mechanism to ensure that the total of the shares equals 100% of the Tier1 request. Other countries pledge according to their share of LHCb physicists; this can never result in a 100% total pledge since not all LHCb physicists come from countries with a Tier1 site. Because of these different pledging procedures, and because actual usage at the end of a year cannot exceed the total pledge, countries that pledged a fixed share of the request for a given year find at the end of that year that they are providing a higher share of the used resources than the share used to calculate the pledge. Understandably, this can lead to a reduction of their pledge for the following year in order to recover their nominal share of the total resources, thus exacerbating the shortage.

We understand that there is no existing guideline in WLCG that clarifies the expectation on Tier1s in this situation and would welcome a C-RSG recommendation on how to solve this issue.

As has been pointed out in the 2012 resources usage report, LHCb-PUB-2013-001, the change of share of resources between Tier1s in 2012 with respect to 2011 was an issue because the 2011 data had been distributed with a different share and therefore the CPU and tape resources had difficulties to cope with the already existing data. The lesson for the future is that data already accumulated at sites should be taken into account for the pledges of resources at Tier1s, for example by pledging for a fraction of the requested increment in resources, rather than a fraction of the total resources.

5.5. Consequences of storage shortages during 2013 and beyond

In this document we have presented a model whereby we take advantage of all computing resources available to LHCb during LS1 (including the HLT farm) to produce all the MonteCarlo samples needed for the final analysis of Run 1 data, and to prepare large samples for the initial analysis of Run 2 data and for detailed physics studies for the LHCb upgrade. The success of this strategy will depend on the availability of disk on which to store the simulated data to be analysed. As can be seen in Table 5-7, we predict a shortfall of disk already in 2013 and request a significant increase in disk for 2014. We are working on a new, reduced, DST format for simulation that is expected to reduce event sizes by approximately 30%, that should alleviate the shortage, but ultimately any unsatisfied shortage will lead to a reduction of the MonteCarlo production rate, thus leaving available CPU resources idle. These delayed productions will have to take place later, perhaps in 2015, where for the time being we have assumed that the majority of CPU resources will be used for real data processing. If we need to reserve more CPU for simulation in 2015, we will have to either increase the CPU request or foresee to "park" a significant fraction of the new data, to be processed only several years later.

Note also that the provision of a large amount of disk in 2014 reduces the need for a large step in disk provision for 2015. This is because simulated data that will have been analysed during 2013 and 2014 can be removed from disk in 2015, making room for the new real data. On the other hand, as can be seen in Table 5-8, we will require a large step in tape provision for the new data in 2015; this is largely incompressible: in Table 5-6 one can see that the major part of the increase is due to RAW data that, if not recorded, is lost. "Parking" of some fraction of this raw data will only postpone the need for the corresponding fraction of tape for FULL.DST.

6. Summary

This document has presented updated resource requests for 2013-2014 and a preliminary request for 2015. The estimates for CPU are given in Table 5-2, and the estimates for storage in Table 5-7 and Table 5-8. We give an indication of the maximum tape staging bandwidth requirement in Table 5-9.

The model we have presented results in a large investment in disk during 2013 and 2014, a moderate increase in CPU resources in 2014, and a large investment in tape during 2015. This allows us to spread the cost of the increase in capacity required for the recording and analysis of the initial Run 2 data, while ensuring that the existing CPU capacity is fully utilised throughout LS1 to complete the analysis of the Run 1 data. Any delay in the provision of the new disk resources will result in a false economy, since they will in any case be required in 2015; it would mean that existing CPU resources could not be fully utilised during LS1, thus delaying the completion of the Run 1 analyses into 2015, when they would compete for resources with the processing of the Run 2 data.

It is stressed that estimates for 2015 are very preliminary. These will be refined for the October 2013 C-RRB.