LHCb Computing Resources: 2012, 2013 re-assessment and 2014 request

LHCb Public Note

Issue:	V1
Revision:	0
Reference:	LHCb-PUB-2012-014
Created:	9th Sep 2012
Last modified:	18th Sep 2012
Prepared By:	LHCb Computing Project R. Graciani/Editor

Abstract

This document presents an update of the computing resource estimates for LHCb. It presents a re-evaluation of the 2012-13 requirements in view of the updated LHC schedule, and the new request for 2014. This document should be considered an update of LHCb-PUB-2011-009, in which 2012 resource requests were initially presented, and LHCb-PUB-2012-004, where the estimates were updated in view of the 2011 experience.

The model presented will follow the nominal LHCb computing model, in which all real-data processing activities take place at the Tier0/1s while the Monte Carlo simulation is taken care of by Tier2s and any other computing resources to which we can get access. When necessary these roles can be mixed to dedicate larger amounts of CPU power to any of these main activities.

Differences in the model and deviations from previous estimates are described.

Table of Contents

List of Figures

List of Tables

1. Introduction

This document summarizes the last updates of LHCb computing resource usage estimates. It covers the period 2012-2014[1] and is based on the latest LHCb running experience and the latest updates to the LHC schedule. For 2012 it is an update, including the consequences of the changes in the LHC schedule as well as recent modifications to the LHCb running parameters. These updates significantly affect previous estimates for 2013. For 2014 this document should be considered the best forecast that we are able to do given the current knowledge of the LHC schedule of that year.

During 2012 LHCb has been smoothly taking data at a rate slightly above the initial estimates. The rate was increased by approximately 10%, up to 5 kHz as shown in LHCb-PUB-2012-013. Additionally CERN decided to extend the pp running period for 2 extra months in order to increase the total integrated luminosity of the experiments before the 2013/14 shutdown. Finally, LHCb has decided to take part in the pA run now scheduled for January/February 2012. Overall all these changes imply give rise to over 40% extra more data in 2012, while the new schedule also imposes extra constraints for the processing of this data.

This document is organized as follows: a brief summary of the LHCb computing model is presented in Section 2, relevant changes to the activities and processing conditions for 2012 and 2013 are given in Section 3. In Section **Error! Reference source not found.** we very briefly review the assumptions made concerning the LHC startup after the end of Long Shutdown 1. The resulting updated requests are given in Section 4. A summary is given in Section 5.

2. Summary of the LHCb Computing Model

A detailed description of the LHCb Computing Model can be found in LHCb-PUB-2012-04 and LHCb-PUB-2011-09. We summarise here the salient points relevant for this document.

- The LHCb detector operates at constant instantaneous luminosity and a roughly constant High Level Trigger rate. The volume of RAW data produced by LHCb is to first order proportional to the time spent in stable beams, regardless of the instantaneous luminosity evolution in the GPD interaction regions.
- The RAW data are sent to the CERN MSS, CASTOR, and from there, a second copy is distributed to one of the LHCb Tier1s Tape systems (CNAF, GRIDKA, IN2P3, NL-T1, PIC and RAL). The share between Tier1s is proportional to the CPU pledges at those sites, in order to even the prompt processing and reprocessing activities.
- A prompt Reconstruction of these data takes place at CERN and at Tier1s, producing reconstructed data that is kept in the MSS of the production site. Each Tier1 processes around 80% of the RAW data it received, the 20% left being processed at CERN.
- Reconstructed data is stored locally and then processed by Stripping jobs that run selection algorithms on the full data sample and produce a number of streams in DST (~100 kB/event) and MDST (~10kB/event) formats.
- These streams are the input to the physics analysis and therefore are replicated to several Tier1s and kept Online on disk. An archive copy is also kept on Tape at CERN and at one Tier1.

[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.

- The Reconstruction and Stripping passes are repeated when improved calibrations/alignment are available or when the algorithms have been improved. This implies a recall from Tape of the RAW data sample (for re-reconstruction) or reconstructed data (for re-stripping)
- Analysis is running mostly at Tier1s, jobs running at one of the sites where the (M)DST has been replicated. Possibly some simulation studies are carried out by users at Tier2s.
- Tier2 sites are mainly dedicated to producing simulated data (requiring no input) that is uploaded for analysis to an associated site (CERN or Tier1). After merging the resulting DSTs are replicated to CERN and a few Tier1s. An archive copy is also kept on Tape at one of the CERN or Tier1 sites.

The implementation of this model is flexible, and allows Tier2 sites to contribute to real data processing when extra CPU resources are needed due to concurrent activities (e.g. prompt reconstruction and reprocessing). Tier1s may also be used to produce simulated data when they are not dedicated to their main activity (real data processing and analysis).

3. Changes in the Computing Model and 2012 data taking

3.1. New reconstructed data format

A major change in the processing model comes from the experience of the last re-Stripping campaigns that took place during the first months of 2012 and experience from the 2012 prompt processing. Stripping requires initially access to reconstructed data (SDST) and then to the RAW data for the selected events. For the Stripping passes that immediately follow a reconstruction this is not a big problem since both files are already on disk caches. However, for standalone Stripping passes it requires the synchronized staging of two sets of independent files, which in general are located on different sets of physical tapes (since they were created at different times). The experience shows that keeping this synchronization when several hundreds of terabytes of data needs to be staged is very demanding for the tape system of the sites while trying at the same time to obtain the maximum throughput. As of July 2012 LHCb has decided to produce a new data format after a reconstruction pass, FULL.DST that includes in a single file the SDST and the corresponding RAW. This halves the number of tape recall operations but effectively stores a third copy of the RAW data on tape. To compensate for the extra amount of tape that this requires, the archive replicas of other derived data samples (DST, MDST and MC) have been reduced from two to one.

3.2. 2012 Data Taking conditions

In addition there were many changes in the data-taking conditions:

- CERN has decided to extend by 2 months the pp data-taking in 2012, until mid of December.
- LHCb has been taking data at a higher effective trigger rate than originally estimated.
- LHCb has decided participate in the Heavy Ion run of LHC in early 2013.

These changes significantly increase the physics potential of LHCb but will require additional resources for its exploitation. The expected increased size of the 2012 RAW and Reconstructed formats (tape) are as follows:

- RAW data: 1.7 PB, 43% increase with respect to previous estimates (1.2 PB).
- Reconstructed data: 3.1 PB, 120 % increase with respect the previous estimate (1.4 PB).
- Heavy Ion RAW data: 100 TB.

The disk resident formats (DST and MDST) used for physics analysis are expected to increase by 43 %.

3.3. Plans for 2012 data processing

The extension of the LHC running together with the requirement to have the largest possible fraction of 2012 data fully reprocessed before the 2013 winter conferences imposes new constraints. Different solutions have been evaluated and the following compromise has been achieved to optimally use available resources during an extended simultaneous running of prompt reconstruction and full reprocessing. Making use of the flexibility provided by our processing tools we will proceed as follows:

- Prompt reconstruction will be allocated to CERN but only 50% of the collected data will be promptly reconstructed. This should be enough to validate the data quality and provide new calibration constants for the full reprocessing. In order to free a fraction of the Tier0 capacity for analysis about 20% of this activity will be sub-contracted to Tier2s (downloading the RAW data from CERN).
- Full reprocessing, starting by mid September, will take 5 months and will be allocated to the Tier1s (where a full copy of the RAW data is present) and to CERN after data taking finishes. As for the Tier0, 20% of the reconstruction will be sub-contracted to Tier2s in order to free some CPU resources at the Tier1s for analysis.
- In order to make space for the extra size of the new FULL.DST format, reconstructed data samples (SDST and FULL.DST) produced by the prompt reconstruction of 2012 data are currently being removed from the Tape systems at the sites. This implies that no further stripping of the 2012 data will be possible until the reprocessed data becomes available.
- A new full reprocessing of 2011 data is planned starting in March 2013, once the full reprocessing of 2012 data is over.
- Another stripping pass of 2012 data is foreseen in spring 2013 to provide samples for new analyses not included in the reprocessing.
- The rest of the schedule for further processing of 2011 and 2012 is maintained. In the Autumn of 2013 we plan a new stripping of the samples and during 2014 a full reprocessing of both data samples should provide the ultimate version of these data.
- Given the significant increase of the physics samples, a 40% increase for the simulation needs has been included in the model.

For 2013 and beyond the extra CPU work needed for the new MC simulation production will be accommodated by using the High Level Trigger (HLT) farm and by using the Tier0/Tier1 resources outside the periods in which they are not dedicated to real data processing.

3.4. Data taking after LS1

After the long shutdown LS1, LHC is expected to resume operations at the end 2014. For the purpose of this document we assume one month of data taking before the end of March 2015, at a data rate out of the HLT farm similar to that of the 2012 run. This data and the corresponding reconstruction and stripped samples provide the current best estimates of the needs for this period.

4. Updated estimates

4.1. CPU resources

Including all the updates described in the previous sections and feeding them into our computing model simulation tool we come out with a table of required CPU work for the different activities. Table 4-1 presents those estimates for the period 2012-14.

Work	2012		2013		2014	
	kHS06*y	%	kHS06*y	%	kHS06*y	%
MC	101	46	142	60	142	65
Analysis	40	18	39	17	39	18
Reco	29	13			8	4
Repro	49	23	55	23	27	13
Total	**219**	**100**	**236**	**100**	**217**	**100**

Table 4-1: Estimated CPU work needed for the different activities.

When these activities are assigned to the different Tiers we come out with a requirement on the CPU power capacity that needs to be available at each of them. This is presented in Table 4-2 that derives from the detailed profile presented in Figure 4-1. Note that we intend to use the HLT farm to provide some of the required Tier2 capacity.

Power	**Pledge 2012**	**2012**		**2013**		**2014**	
	kHS06	**kHS06**	**%**	**kHS06**	**%**	**kHS06**	**%**
Tier0	**34**	**34**	**18**	**34**	**16**	**34**	**16**
Tier1	**91**	**110**	**58**	**110**	**52**	**110**	**52**
Tier2	**47**	**46**	**24**	**46**	**22**	**46**	**24**
HLT farm				**20**	**10**	**20**	**10**

Table 4-2: Estimated CPU power needed at the different Tier levels.

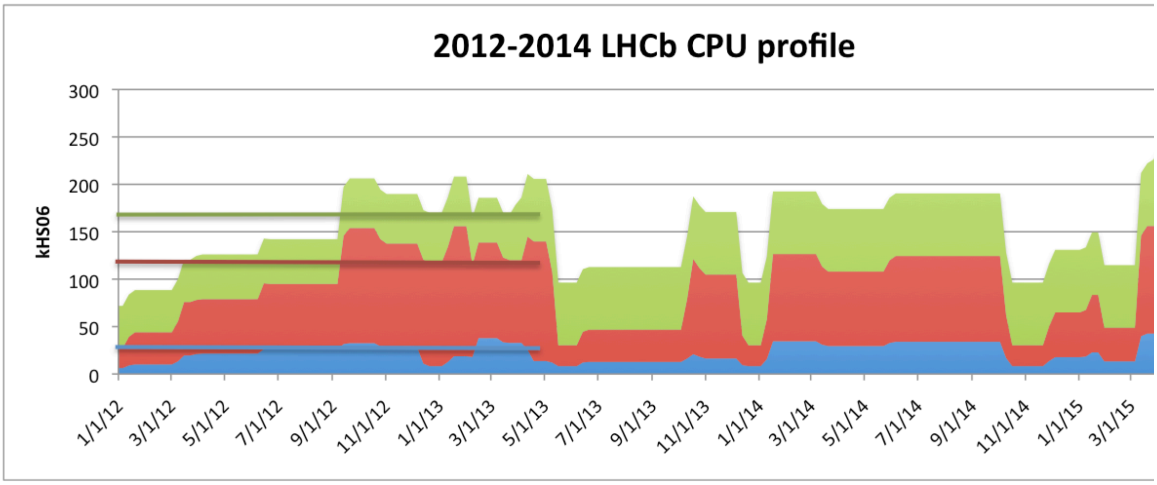

Figure 4-1: Estimated CPU profile for the reported period. Tier0/1s holes are not filled up with simulation to appreciate the contributions form the different real data processing activities.

4.2. Storage resources

Following the same model we have estimated the corresponding Storage requirements. They are determined based on the peak values observed in the simulated usage profiles. These profiles take into account both the new samples being produced and their replicas, as well as the schedule for removal of replicas of older processings The resulting profile is shown in Figure 4-2, Table 4-3 and Table 4-4.

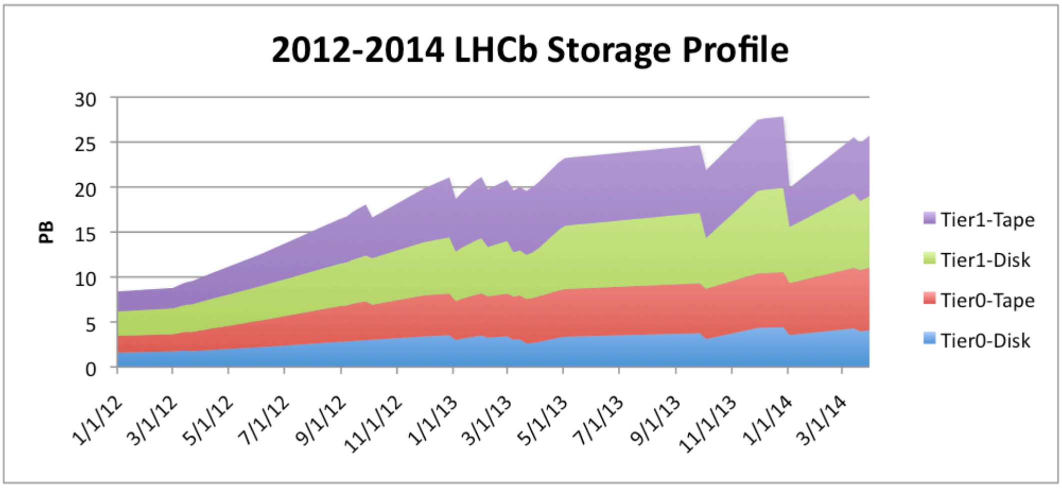

Figure 4-2: Estimated Storage profile for the reported period. Notice that cleanups (the dips) have been carefully tuned to delay as much as possible the increase of resources needed due to the increase of data samples in 2012.

Disk	2012		2013		2014	
	PB	%	PB	%	PB	%
Tape cache	1.0	10	1.0	8	1.0	6
RAW + FULL.DST	0.1	1	0.1	0.8	0.1	0.6
DST+MDST	4.8	49	5.8	45	6.5	41
MC	3.0	31	5.0	38	7.0	44
User	0.9	9	.1.1	9	1.3	8
Total	9.8	100	13.0	100	15.9	100

Table 4-3: Break down of Disk Storage estimates for each of the different data categories that are included in the model.

Tape	2012		2013		2014	
	PB	%	PB	%	PB	%
RAW	5.9	36	5.9	34	6.1	32
FULL.DST	5.8	36	5.8	34	6.2	32
DST+MDST (archive)	3.3	20	3.8	22	4.5	23
MC (archive)	1.2	7	1.8	10	2.4	13
Total	16.2	100	17.3	100	19.2	100

Table 4-4: Break down of Tape Storage estimates for each of the different data categories that are included in the model.

Making use of the above results and allocating them amongst the Tier levels results in updated estimates of the storage requirement for 2012-14 summarized in Table 4-5 and Table 4-6.

Note that the resulting Disk estimates for 2012 are below the “official” pledge by the sites. This is the result of the mitigating actions taken to respond to the limited actual available disk (not all of the pledge is in place) and the delays in the processing steps introduced by the new LHC schedule. It is important to note that an important fraction of the 2013 resources will be necessary already in the first half of the year in order to accommodate the new samples produced by the re-processing and re-stripping campaigns taking place during those months.

Disk	**Pledge**	**2012**		**2013**		**2014**	
	PB	**PB**	**%**	**PB**	**%**	**PB**	**%**
Tier0	**3.5**	**3.5**	**36**	**4.4**	**34**	**5.5**	**35**
Tier1	**7.2**	**6.3**	**64**	**8.6**	**66**	**10.4**	**65**

Table 4-5: Disk estimates for each Tier level.

Tape	**Pledge**	**2012**		**2013**		**2014**	
	PB	**PB**	**%**	**PB**	**%**	**PB**	**%**
Tier0	**6.4**	**6.2**	**38**	**6.5**	**38**	**7.3**	**38**
Tier1	**5.3**	**10.0**	**62**	**10.8**	**62**	**11.9**	**62**

Table 4-6: Tape estimates for each Tier level.

4.3. Evolution of Tier1 shares

In the LHCb computing model, jobs are executed at the Tier0 or Tier1 sites that host the required input data. In a well-balanced system, the sharing of storage resources among sites should match the sharing of CPU resources. If the storage share provided by a given site reduces compared to the CPU share, it becomes difficult to use that site efficiently for data processing.

Shares out of Tier1 totals						
2011 CPU share (%)	27.0%	15.9%	11.5%	21.6%	6.5%	17.4%
2011 Disk share (%)	27.2%	16.2%	12.1%	20.4%	6.6%	17.5%
2012 CPU share (%)	21.0%	21.6%	18.6%	15.6%	6.5%	16.7%
2012 Disk share (%)	14.3%	22.7%	19.8%	11.4%	6.8%	25.0%

Table 5-7 shows the evolution of the disk pledge at Tier1 sites in the last two years, as reported in the REBUS tables, together with the percentage of the Tier1 share of disk and CPU provided by each site in the two years.

Disk pledge	**FR-CCIN2P3**	**DE-KIT**	**IT-INFN-CNAF**	**NL-T1**	**ES-PIC**	**UK-T1-RAL**
2011(PB)	1010	600	450	757	244	651
2012(PB)	1010	1610	1400	810	485	1767
% change	**0%**	**168%**	**211%**	**7%**	**99%**	**171%**
Shares out of Tier1 totals						
2011 CPU share (%)	27.0%	15.9%	11.5%	21.6%	6.5%	17.4%
2011 Disk share (%)	27.2%	16.2%	12.1%	20.4%	6.6%	17.5%
2012 CPU share (%)	21.0%	21.6%	18.6%	15.6%	6.5%	16.7%
2012 Disk share (%)	14.3%	22.7%	19.8%	11.4%	6.8%	25.0%

Table 5-7: Tier1 Disk Pledges and CPU and disk shares in 2011 and 2012, as reported in REBUS.

One can clearly see from Table 5-7 that the disk share at two sites has fallen well below the CPU share. The mismatch is then compounded by the fact that new data can only be placed where there is space available; if, in a given year, a site provides less additional space than required to maintain its share, then the share of new data received is proportionally less.

This is best illustrated by Table 5-8, which shows the results of a simulation of how the physics analysis DST produced by the 2012+2011 reprocessing campaign will be distributed. Despite preferentially cleaning old replicas from sites below share, the fraction of new DST that can be replicated at the sites low on disk is well below their CPU share. This has a long term impact on the usability of CPU resources at those sites, because analysis jobs will be dispatched to sites according to the distribution of the data to be analysed.

Distribution of 2011 and 2012 reprocessed DST (TB)						
CERN	**FR-CCIN2P3**	**DE-KIT**	**IT-INFN-CNAF**	**NL-T1**	**ES-PIC**	**UK-T1-RAL**
614	113	491	455	111	53	450

Table 5-8: Distribution of 2011 and 2012 reprocessed DST (TB)

5. Summary

This document has presented updated resource requests for 2012-2014.

These estimates have evolved since the previous ones due to an increased LHCb physics trigger rate, the extended LHC schedule, and mitigating actions we have taken to adapt to actual resources available in 2012. Overall we expect an increase of ~40% more data in 2012 than originally foreseen leading to a commensurate increase in storage and CPU requirement.

The final estimates for CPU are given in Table 5-2, and the final estimates for storage in Tables 5-5 and 5-6