Major Changes to the LHCb Grid Computing Model in Year 2 of LHC Data

L Arrabito[1], V Bernardoff[2], D Bouvet[1], M Cattaneo[2], P Charpentier[2], P Clarke[3], J Closier[2], P Franchini[4], R Graciani[5], E Lanciotti[2], V Mendez[6], S Perazzini[4], R Nandkumar[7], D Remenska[8], S Roiser[2], V Romanovskiy[9], R Santinelli[2], F Stagni[2], A Tsaregorodtsev[10], M Ubeda Garcia[2], A Vedaee[2], A Zhelezov[11]

[1] CC-IN2P3 Centre de Calcul, France
[2] CERN, European Organization for Nuclear Research, Switzerland
[3] University of Edinburgh, United Kingdom
[4] INFN, Istituto Nazionale di Fisica Nucleare, Italy
[5] University of Barcelona, Spain
[6] PIC, Port d'Informació Cientifica, Spain
[7] STFC, Science & Technology Facilities Council, Great Britain
[8] NIKHEF, Nationaal Instituut voor Subatomaire Fysica, The Netherlands
[9] Institute for High Energy Physics, Russia
[10] Universite d'Aix - Marseill II, France
[11] Ruprechts-Karls-Universit"at Heidelberg, Germany

Abstract. The increase of luminosity of the LHC in 2011 also introduced an increase of computing requirements for data processing. This paper describes the data processing operations during 2011 prompt reconstruction as well as the end of year re-processing of the full data sample. It further gives an outlook to next evolutionary steps in the LHCb computing model for 2012 data processing and beyond.

1. Introduction

The increase of luminosity in the LHC during its second year of operation (2011) was achieved by delivering more protons per bunch and increasing the number of bunches. Taking advantage of these changed conditions, LHCb ran with a higher pileup as well as a much larger charm physics introducing a bigger event size and processing times. These changes led to shortages in the offline distributed data processing resources, an increased need of cpu capacity by a factor 2 for reconstruction, higher storage needs at T1 sites by 70% and subsequently problems with data throughput for file access from the storage elements. To accommodate these changes the online running conditions and the Computing Model for offline data processing had to be adapted accordingly.

This paper describes the LHCb data processing work flows [6, 12] and the additional characteristics of the computing model in sections 2 and 3. It further describes operational issues discovered and solved during 2011 and presents the performance of the system in section 4 and concludes by lessons learned to further improve the data processing reliability and quality for the 2012 data re-processing in section 5.

2. Data Processing Workflow

Data collected from the LHCb detector is processed in several steps with the LHCbDirac framework [11] before it is made available to physicists for their final analysis work. This section will describe the different steps of this work flow, concentrating on the data processing itself, but also mentioning the needed data management (data transfer) for processing the data in each section where needed.

2.1. DataReconstruction

A fill of the LHC machine, lasting up to more than one day, is divided into several runs by the LHCb detector. Runs last usually around 1 hour. The data collected is processed by the high level trigger which finally produces around 250 RAW files, of 3 GB each, per run. As the trigger rate has significantly increased in 2012, but the other parameters stayed constant (file size, run duration, event size), the number of files per run has also increased in 2012. The files are transferred from the pit to the CERN tape storage (CASTOR [7]) and from there further replicated to one Tier1 centre again copying the file to the local tape system. Thus always two copies of a RAW file are available. All files of a given run are copied together to one T1 site. A DataReconstruction job (Fig 1) will stage a RAW file from the tape system onto a local disk pool and further download it to the worker node. The reconstruction program, will process the file and produce a SDST (Stripping Data Summary Tape) file of the same size, which contains the information about the reconstructed physics quantities plus some additional information of the RAW event. The SDST is again copied onto the local tape system. Reconstruction jobs last a minimum of one day.

2.2. DataStripping

A Stripping job (Fig 2) takes an SDST and RAW file as input and selects the reconstructed events into different physics stream which are groupings of event types. It needs to run at the same site as the reconstruction job, as this is the only place where the SDST is available. In the past stripping jobs accessed the data by protocol, i.e. remotely, but in the last year this has changed and also stripping jobs download their input data to the worker node. Stripping jobs will run 3 hours on one pair of input files and produce one file per demanded physics stream. The output of the Stripping job is uploaded to the local disk only storage. Usually one to two pairs of input files are processed by one stripping job.

2.3. Merging

Merging jobs are the last jobs in the data processing chain. They collect files of a given physics stream and concatenate them to one bigger file, usually around 5GB. The output of the merging jobs, the merged DST (Data Summary Tape) files which contain physics events of one stream.. They are copied again to disk storage and are further replicated several times to other Tier1 disk storages, where one copy is always kept at CERN. In addition two copies of a merged DST file are kept on custodial storage for archiving. Usually merging jobs last less than 30 minutes.

2.4. Additional Workflows

In addition to the “standard” data processing work flow as described above, another set of additional productions have been setup in 2011/12.

2.4.1. Calo Femto Stripping This work flow (Fig 3) consists of a stripping and merging step (Fig 3) which uses different software than the Data Stripping. It also produces smaller files than the normal Stripping and will run faster, therefore up to 5 pairs of input files can be used for these productions. Technically the steps work the same way as above, ie. stripping runs

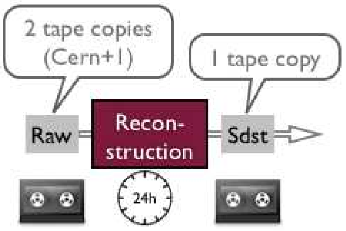

Figure 1. Reconstruction Workflow

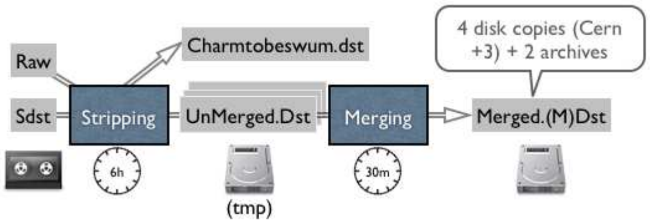

Figure 2. Stripping + Merging Workflow

on RAW/SDST pairs and produces one unmerged DST file which is subsequently merged. The production output itself is used for calibration of the Calorimeter subdetector.

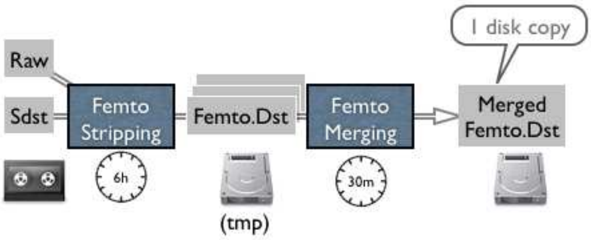

Figure 3. Calo Femto Workflow

2.4.2. DataSwimming DataSwimming [8] consists of several steps. The “Moore” step is running the trigger application on one selected output stream of the standard Stripping, which is followed by a Stripping and Merging step. The Moore step is exercising one event up to 10 times, testing the detector acceptance in the vertex locator. This is very CPU intensive but has only little I/O and will last around 1 working day.

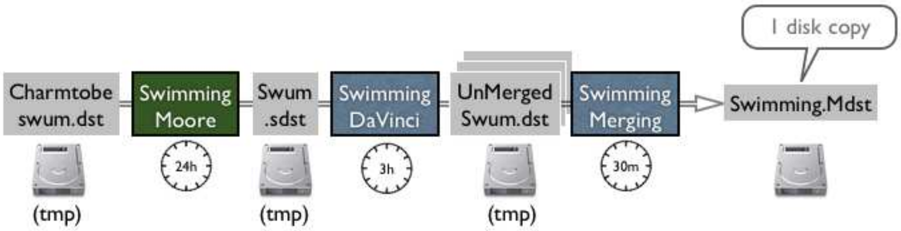

Figure 4. Swimming Workflow

2.4.3. Working Group Productions are being setup in an organised production manner to select large quantities of analysis data. In the past individual physicists were submitting these large productions to the grid with their own credentials, storing the output in their own storage area. With the production team handling these productions they can be more easily integrated with the other work flows and priorities are better handled. Also the storage is not attributed to an individual user but to the general grid storage for production data, thus making the handling easier for the data managers. By using working group productions the meta-data of this data can easily be shared between physicists, which otherwise is not easily achievable.

2.4.4. Simulation jobs consist of several steps for the actual particle generation, detector response simulation, digitisation, high level trigger and offline analysis. The main activities carried out at T2 sites are simulation activities. Recently a “stripping” step has been commissioned for the simulation work flow where the reconstructed particles can be selected in the same way as in prompt reconstruction/stripping. This will further help to minimise the storage needs for LHCb.

2.4.5. User Jobs are jobs launched by individual physicists of the collaboration e.g. for data analysis. Jobs in this group mostly work on merged DST files which are replicated several times on the different storage elements. Jobs of this kind are not controlled in any way by the production team.

Table 1. Site usage by job types (**Y**)es, (**N**)o, (**P**)ossible (not tested)

Job Type	Tier 0/1	Tier2[0]
Reconstruction	Y	Y
Stripping	Y	P[1]
Merging	Y	N
Swimming	Y	Y
Working Group	Y	N
Simulation	Y[2]	Y
User Jobs	Y	Y[3]

[0] including unpledged resources
[1] only with data download
[2] if site not used by any other job type
[3] if not using input data

Table 2. Storage characteristics

Data Name	Storage Type	Copies	Archive (Number)
RAW	tape	2	Y(2)
SDST	tape	1	N
DST	disk	4	N

3. The LHCb Computing Model

This section describes the main characteristics of the LHCb Computing Model providing the infrastructure for the work flows described above and as described in the experiment’s Computing Technical Design Report (TDR) [3, 9].

- The computing resources for LHCb are spread over sites in several tiered levels
  - Tier0 - CERN
  - 6 Tier 1 sites - CNAF, PIC, NL-T1, GRIDKA, RAL, IN2P3
  - 120 Tier 2 sites with pledged resources for LHCb
  - several unpledged sites providing CPU for LHCb
  - computing resources in the Event Filter Farm (HLT) which can be utilised for data processing during “no-beam”
- Storage elements are only used at T0 and T1 sites for both disk-only and tape storage
- Different Tier levels are generally used for different kind of activities (see Table 1)

- Several copies of different data file types are being kept at the different storage elements (see Table 2)

4. Major Data Processing Activities in 2011/12

This section describes the main processing activities carried out in 2011 and what has already been done in the first quarter of 2012.

4.1. Prompt Reconstruction '11

In 2011 a total of 1fb^{-1} data for physics has been reconstructed and made available to the collaboration. This data was processed at T0 and T1 sites only. Towards the end of data taking the prompt reconstruction was only carried out at CERN while the T1s were dedicated to data reprocessing (see next section). During that time the processing at CERN was accumulating as the site was barely able to cope with the load.

4.2. Data Reprocessing '11

In October the reprocessing of data collected in 2011 started. During that time the prompt reconstruction was still ongoing. In order to make the handling of the processing activities easier it was decided to continue with the prompt reconstruction only at CERN until the end of data taking while the reprocessing was started at all Tier 1 sites. For the first time Tier 2 sites were also involved in this activity. The selected Tier 2 sites, attached to the topologically closest T1 storage, provided 25 % of processing power. It is worth mentioning that also the prompt reconstruction at CERN was helped by selected Tier 2 sites in or near Switzerland. After data taking from LHC had finished and the accumulated backlog of prompt reconstruction was finished, also CERN and "its" Tier 2 sites were involved in the reprocessing activities.

The processing followed the usual reconstruction / stripping / merging work flow, where reconstruction jobs were executed both at Tier1 and Tier2 sites and the remaining stripping / merging at Tier1 sites. Reconstruction downloaded the input data to the worker nodes, stripping was using protocol access, merging again download.

During this operation care was taken not to overload the storage elements as the full data set needed to be recalled from tape onto temporary disk pools. In order to do this the run ranges for the reconstruction and subsequent stripping were extended carefully by the production team.

LHCb did not actively monitor network performance or bandwidth.

4.3. Data Restripping '12

Beginning of 2012 a Re-Stripping of the 2011 data set was launched with improved stripping code, based on the latest SDST from 2011 fall reprocessing. As most computing resources for LHCb were idle at that time (no prompt reconstruction was going on) the restripping was carried out at T0 and T1 sites only. The input RAW and SDST files needed to be recalled from tape for the whole data set. As stripping jobs are now also downloading data to the worker nodes it would have been possible to extend the activities also to T2 sites if necessary. In previous stripping exercises often failures were observed in the staging of the input files. As both RAW and SDST files need to be staged the stage pools easily became overloaded by the amount of input data to be staged. For this re-stripping a new process for staging was introduced where the amount of data to be staged during a given period of time was partitioned into "slices" which the site storage element could withstand. After some initial tuning of the slice sizes this new process mitigated the above problem perfectly. The re-stripping operation was finished before the start of 2012 data taking.

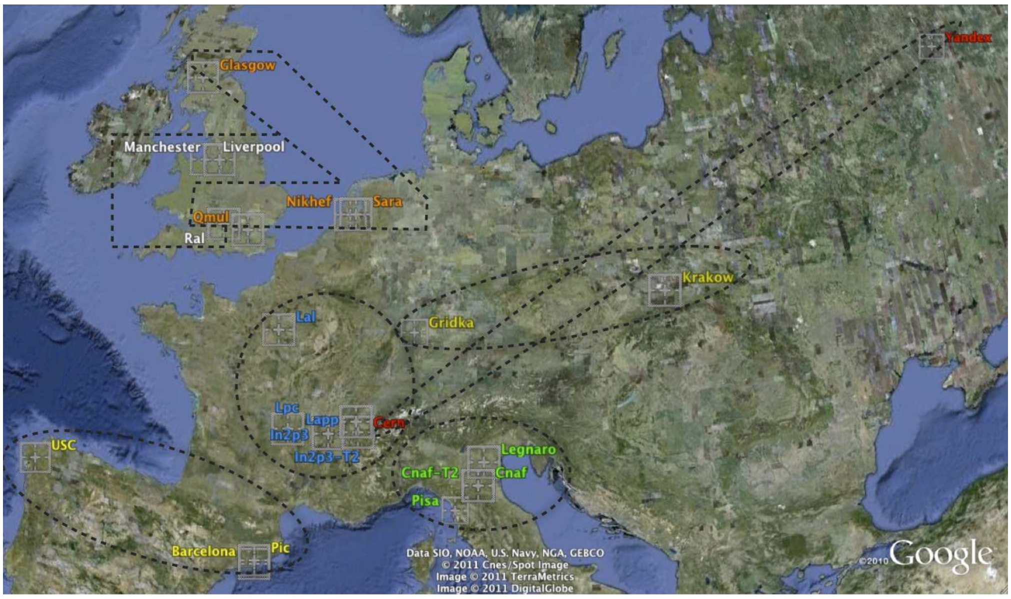

Figure 5. Grid sites of the 2011 data reprocessing

4.4. Prompt Reconstruction '12

Prompt reconstruction was launched at the start of 2012 data taking. A first set of data was reconstructed and stripped in order to re-align the detector. Once this new alignment was ready the first data was completely re-processed (reconstruction / stripping) together with the ongoing prompt reconstruction.

5. Possible Future Evolution of the LHCb Computing Model

As the CPU needs for data processing will further increase over the next years, new ways of assignment of CPU resources to processing tasks will need to be developed, similar to what has been done in other experiments [2, 4]. With the 2011 data reprocessing a first step into the direction of "attaching" T2 sites to T1 storage and their usage as "co-processor" was introduced, going beyond the initial tiered Monarc model [1, 5] of the Grid. The model has proven to work but was still thought to be too rigid for any further absorption of peak loads that might arise.

In this respect the next evolutionary step in the data processing of LHCb will be a "free" attachment of T2 sites to any T1 storage. This will also overcome the situation where a given storage element still has lots of data to process while data at other storage elements are fully processed. A T2 site with CPU capacity will then be able to get any input files which are available, process them and upload them again.

In the case of Stripping a further simplification will be introduced where the part of the RAW file will be copied into the reconstruction output SDST (now LDST). As such only a single file needs to be staged for stripping and will further simplify the work flow.

The staging itself will also become easier as a new "BUFFER" disk-only storage element will be introduced which will work as a temporary disk buffer for input / output data of the different work flows. For the time being the output of the reconstruction jobs (LDST) will be copied over to BUFFER and also migrated to tape. At the time when all stripping work flows have used a specific LDST, the file will be removed from the BUFFER storage. The same will happen for intermediate unmerged DST files as output of stripping jobs. The usage of EOS [10] for the BUFFER storage is envisaged at CERN.

In this scenario several rules will still need to be obeyed.

- Any site which is “allowed” to process remote data shall be able to do so. Usually the model results in a T2 processing data on a T1 storage element, but it should also include the possibility of a T1 to process data from another T1 storage element if CPU capacities are idle.
- A priory every site shall be able to process data from any T1 storage. In case the bandwidth is a limitation the site shall be disabled to download/upload data from that storage element. The file transfer rate between storage site and processing site needs to be monitored in order to not overload the Tier2 network and Tier1 storage.

This kind of processing shall be made available a priory for peak of data processing activities, e.g. data reprocessing or restripping of large data sets, which usually happen only a few times per year.

In addition to this approach other additional usage of computing resources is currently being evaluated, like the event filter farm with 12k compute nodes, whole node processing and vitalisation.

6. Conclusion

This paper presented the different work flows and the computing model of the LHCb experiment at CERN, and presented different past scenarios of major data processing activities. With the help of the Dirac framework for distributed computing it was possible to adapt the LHCb computing model to the different situations that occurred when shortages of resources appeared or major activities needed to be carried out. The co-processing of T2 sites during reprocessing and the offloading of cpu intensive tasks to T2 sites being a notable example. The relaxation of the original rigid model of separation of tasks to different Tier levels is currently under detailed discussion within the Dirac development and operations teams.

7. References

[1] Michael Aderholz, K Amako, E Aug, G Bagliesi, L Barone, G Battistoni, M Bernardi, M Boschini, A Brunengo, J J Bunn, J Butler, M Campanella, P Capiluppi, F Carminati, M D’Amato, M Dameri, A Di Mattia, A E Dorokhov, G Erbacci, U Gasparini, F Gagliardi, I Gaines, P Glvez, A Ghiselli, J Gordon, C Grandi, F Harris, K Holtman, V Karimki, Y Karita, J T Klem, I Legrand, M Leltchouk, D Linglin, P Lubrano, L Luminari, A L Maslennikov, A Mattasoglio, M Michelotto, I C McArthur, Y Morita, A Nazarenko, H Newman, Vivian O’Dell, S W O’Neale, B Osculati, M Pep, L Perini, James L Pinfold, R Pordes, F Prelz, A Putzer, S Resconi, L Robertson, S Rolli, T Sasaki, H Sato, L Servoli, R D Schaffer, T L Schalk, M Sgaravatto, J Shiers, L Silvestris, G P Siroli, K Sliwa, T Smith, R Somigliana, C Stanescu, H E Stockinger, D Ugolotti, E Valente, C Vistoli, Ian Malcolm Willers, R P Wilkinson, and D O Williams. Models of networked analysis at regional centres for LHC experiments (monarc), phase 2 report, 24th march 2000. Technical Report CERN-LCB-2000-001. KEK-2000-8, CERN, Geneva, Apr 2000.

[2] K Bloom, G Bagliesi, D Bonacorsi, C Brew, I Fisk, J Flix, P Kreuzer, and A Sciaba. CMS resource utilization and limitations on the grid after the first two years of LHC collisions. In *Computing in High Energy and Nuclear Physics*, New York City, NY, USA, May 2012.

[3] N Brook. LHCb computing model. Technical Report LHCb-2004-119. CERN-LHCb-2004-119, CERN, Geneva, Dec 2004.

[4] S Campana, I Ueda, F Barreiro Megino, C Serfon, and S Jezequel. Evolving ATLAS computing for today’s networks. In *Computing in High Energy and Nuclear Physics*, New York City, NY, USA, May 2012.

[5] M Campanella and L Perini. The analysis model and the optimization of geographical distribution of computing resources: a strong connection. CERN, 1998. Monarc Note 98/1.

[6] LHCb Collaboration. http://lhcb.web.cern.ch/lhcb/.

[7] S Ponce et al. Status and evolution of CASTOR (Cern Advanced STORage). In *Computing in High Energy and Nuclear Physics (CHEP)*, New York City (NY), USA, May 2012.

[8] M Gersabeck, H Dijkstra, G Raven, V Belyaev, and R Aaij. ”swimming” : a data driven acceptance correction algorithm. In *Computing in High Energy and Nuclear Physics*, New York City, NY, USA, May 2012.

[9] Ricardo Graciani Diaz. LHCb computing resource usage in 2011. Technical Report LHCb-PUB-2012-003. CERN-LHCb-PUB-2012-003, CERN, Geneva, Feb 2012.

[10] J Iven and M Lamanna. Overview of storage operations at CERN. In *Computing in High Energy and Nuclear Physics*, New York City, NY, USA, May 2012.

[11] A Tsaregorodtsev, M Bargiotti, N Brook, A C Ramo, G Castellani, P Charpentier, C Cioffi, J Closier, R G Diaz, G Kuznetsov, Y Y Li, R Nandakumar, S Paterson, R Santinelli, A C Smith, M S Miguelez, and S G Jimenez. DIRAC: a community grid solution. *Journal of Physics: Conference Series*, 119(6):062048, 2008.

[12] A Tsaregorodtsev and A Zhelezov. Managing large data productions in LHCb. In *CHEP 2009*, Mar 2009.