PROCEEDINGS OF SCIENCE

LHCb experience during the LHC 2015 run

Christophe HAEN[*] on behalf of the LHCb collaboration
CERN
E-mail: christophe.haen@cern.ch

LHCb is one of the four high energy physics experiments currently in operation at the Large Hadron Collider at CERN, Switzerland. After a successful first running period (Run1 from 2011 to 2013), the LHC just entered the second exploitation phase (Run2, 2015-2018).
The technical break between these two running periods, known as Long Shutdown 1 (LS1), was the opportunity for LHCb to adapt, among other area of development, its data acquisition and computing models.
The operational changes on the data acquisition aspect include a clear split of the High Level Trigger (HLT) software in two distinct entities, running in parallel and in an asynchronous mode on the filtering farm, allowing a higher output rate to the final offline storage for further physics analysis. A very challenging and innovative system performing full calibration and reconstruction in real time has been put in place. Thanks to this system, a fraction of the output of the HLT can be used directly for physics, without any intermediate step: this output is named “Turbo stream”.
Many changes were operated on the offline computing side as well. Besides the use of more modern and/or more scalable tools for the pure data management aspect, the computing model itself and the processing workflow were revisited in order to cope with the increased load and amount of data. The new Turbo stream requires new operational management compared to the other “standard” streams. The clear separation between the different levels of Tiers (0, 1 and 2) has been abandoned for a more flexible, dynamic and efficient “Mesh” processing model, in which any site can process data stored at any other site. Validation and probing procedures were established and automatized before the start of massive Monte Carlo Simulation. This paper presents the changes that were operated, and gives some feedback on their usage during the running period in 2015.

International Symposium on Grids and Clouds 2016
13-18 March 2016
Academia Sinica, Taipei, Taiwan

[*] Speaker.

	Run 1	Run 2
Maximum beam energy	4 TeV	6.5 TeV
Transverse beam emittance	1.8 μm	1.9 μm
Beam oscillation (β*)	0.6 m / LHCb 3 m	0.4 m / LHCb 3 m
Number of bunches	1374	2508
Maximum number of protons per bunch	1.7 * 10^{11}	1.15 * 10^{11}
Bunch spacing	50 ns	25 ns
μ	1.7	1.1
Maximum instantaneous luminosity	7.7 * 10^{33} cm^{-2}s^{-1}	1.6 * 10^{34} cm^{-2}s^{-1}

Table 1: The general LHC beam conditions during Run 1 and planned for Run 2.

1. Introduction

The Large Hadron Collider (LHC) at CERN is the largest particle collider in the world. Its activity is meant to last for several decades, and the data taking periods called "run" are interleaved with technical stops ("shutdown"). These shutdown periods are used to perform maintenance and upgrades, not only by the LHC collider, but also by the experiments exploiting it. Run 1 spawned from fall 2010 until early 2013, Run 2 has started mid 2015 and will last until mid 2018 and Run 3 is planned to start in early 2020. LHCb[1] is one of the four large experiments making use of the LHC collisions. While its major upgrade will happen between Run 2 and Run 3, numerous changes have been carried out between Run 1 and Run 2. In this paper, we will first mention the changes in the LHC running conditions that have an impact for LHCb. We will then focus on the changes that were carried out in the LHCb Online, before looking at the difference for the Offline computing model.

2. LHC conditions

The running conditions of the LHC are obviously driving constraints for the computing activities, both in the Online and the Offline world. For the Online, it reflects in terms of filtering and data acquisition capabilities, while for the Offline, it translates into storage and processing challenges. The evolution of the running conditions of the LHC between Run 1 and Run 2 are summarized in table 1.

As a direct consequence of the increased number of beam bunches, the bunch spacing is decreased, resulting in doubling the bunch crossing frequency in Run 2 compared to Run 1. This, together with the increased beam energy leads to doubling the instantaneous luminosity of the LHC.

It has however a controlled impact on LHCb since, contrary to the two other large LHC experiments (ATLAS and CMS), LHCb is able to adjust its instantaneous luminosity. This is achieved by a technique called luminosity levelling [2, 3], through which the beams at the LHCb interaction point will not collide head on but are slightly displaced. This technique is needed because LHCb would not be able to distinguish primary vertices at full luminosity. While the beam luminosity will decrease during an LHC fill, this displacement of beams will be constantly adjusted such that the instantaneous luminosity delivered at LHCb remains the same throughout the fill (see also Fig. 1). LHCb plans on having the same luminosity in Run 2 as in Run 1 ($4 * 10^{32} cm^{-2} s^{-1}$).

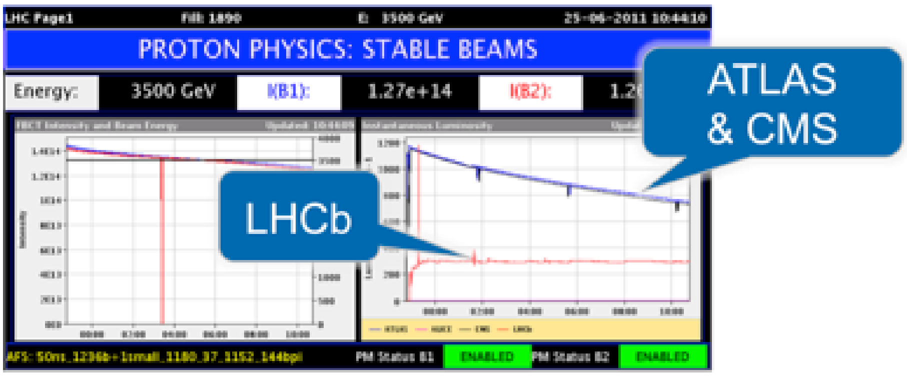


Figure 1: LHC monitoring page of the beam luminosity during Run 1.

μ is defined as the number of simultaneous collisions during a single bunch crossing. Given that the LHC maximum instantaneous luminosity will largely increase due to the decreased bunch spacing, the luminosity leveling of LHCb at the same level as Run 1 will result in a smaller μ. The consequence of this smaller value is a reduced complexity for the events, and hence a smaller event size, compensated by an increase due to the higher energy.

3. Online: trigger and data acquisition

The number of events produced by the LHC collisions is way too large to store all of them. Moreover, only a very limited fraction are interesting from the physics point of view. For these reasons, the DAQ chain is composed of several filtering layers called "Triggers". The LHCb trigger strategy has changed significantly between Run1 and Run2.

3.1 Run 1

The Data Acquisition Chain (DAQ) of LHCb as it was during Run 1 is described in Fig. 2.

The first level, called "L0" is a hardware trigger based on FPGA. Since it is a synchronous trigger, it is cadenced at the beam crossing frequency (40MHz), pipelined, and has a latency of 4 μs. Given its real time nature, the L0 trigger only looks at a fraction of the information of an event to make a decision, namely data coming from the calorimeters and the muon. This trigger allows to reduce the event rate down to 1 MHz.

The output of the L0 trigger is fed into a second level called "High Level Trigger" (HLT), implemented in software application. Several instances of this application are run on each machine

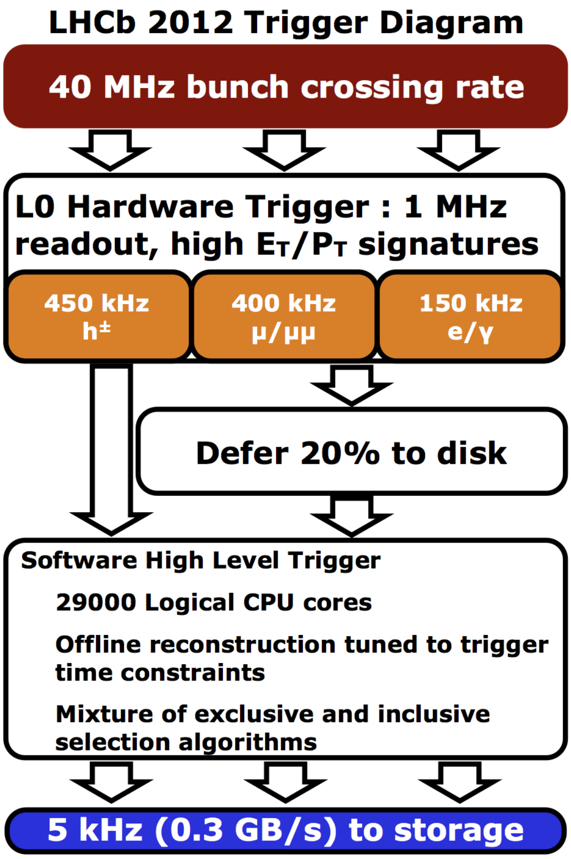

LHCb 2012 Trigger Diagram
40 MHz bunch crossing rate
L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures
450 kHz h±
400 kHz μ/μμ
150 kHz e/γ
Defer 20% to disk
Software High Level Trigger
29000 Logical CPU cores
Offline reconstruction tuned to trigger time constraints
Mixture of exclusive and inclusive selection algorithms
5 kHz (0.3 GB/s) to storage

Figure 2: LHCb DAQ chain during Run 1.

of a large computing farm of about 1700 nodes, which amounts to about 29000 cores. The HLT is asynchronous with respect to the LHC clock, and takes of the order of 20 ms to process an event. The rate of accepted events after the HLT selection was around 5 kHz during Run1. These 5 kHz of events, or about 300 MB.s^{-1}, were aggregated to produce 3 GB "RAW" files at the central online storage, before being transferred to Tier0 for Offline processing.

As of 2012, an extra step called "Deferred triggering" was added to the HLT. The idea was to overcommit by 20 % the processing capability of each node of the computing farm. The events entering the HLT that could not be processed immediately would be dumped on the local disk of the node. The time between two LHC fills would be used to process these extra events, and hence increase the total recorded luminosity. While very simple from the principle point of view, this change implied non trivial modifications, in particular in the control system of the detector, as well as operational challenges in terms of system and disk administration.

3.2 Run 2

Many changes, visible in Fig. 3 have been brought to the LHCb DAQ chain during the long shutdown.

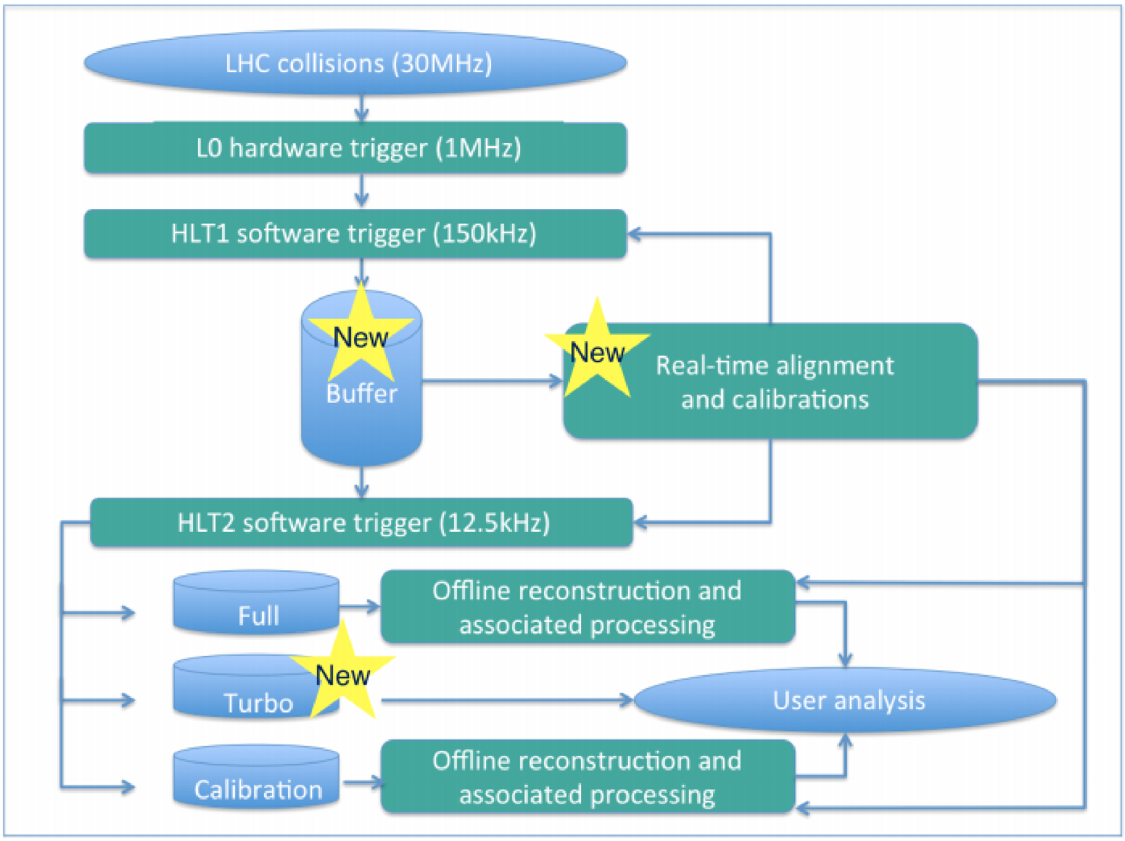

LHC collisions (30MHz)
L0 hardware trigger (1MHz)
HLT1 software trigger (150kHz)
New
Buffer
New
Real-time alignment and calibrations
HLT2 software trigger (12.5kHz)
Full
Offline reconstruction and associated processing
New
Turbo
User analysis
Calibration
Offline reconstruction and associated processing

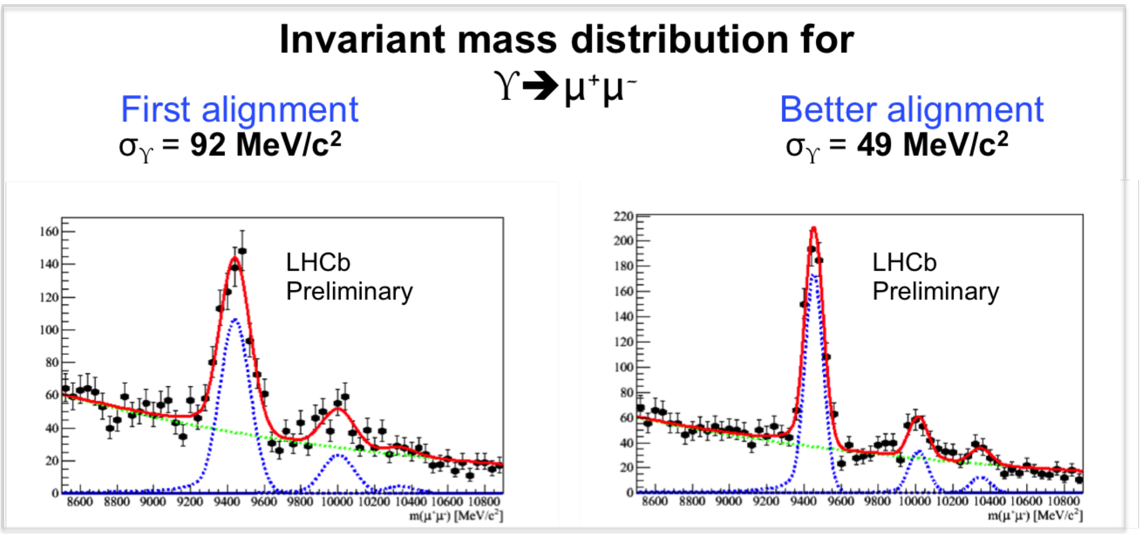
Figure 3: LHCb DAQ chain during Run 2.

The L0 trigger was mainly left unchanged, while the HLT trigger underwent heavy changes. In particular, the HLT software has been split in two separate entities, called HLT1 and HLT2. All the events coming out of the L0 trigger are now directly processed by HLT1: the deferred triggering of Run 1 does not exist anymore. HLT1 performs a partial event reconstruction, before writing the result on the local disk. The output rate of HLT1 is 150 kHz.

Before HLT2 can read the buffer and further process and filter the events, an extra step has to take place: the real-time alignment and calibrations. The physics performance relies on the spatial alignment of the detector and the accurate calibration of its subcomponents (see also Fig. 4). This procedure between HLT1 and HLT2 consists in performing nearly in real time and at regular interval these various tasks:

- RICH refractive index and HPD image calibration

- CALORIMETER calibration
- OT-t_0 calibration
- VELO and tracker alignment
- MUON alignment
- RICH mirror alignment
- Calorimeter π^0 calibration


Invariant mass distribution for $\Upsilon \rightarrow \mu^+\mu^-$

First alignment
σ_Υ = 92 MeV/c²

Better alignment
σ_Υ = 49 MeV/c²

Figure 4: Importance of the alignment.

All these points are performed using a dedicated data stream. By definition, some of these numbers are expected to be stable. Therefore, they are only monitored by a dedicated set of machines called the Calibration farm, and recomputed only if needed. The others require frequent updates, at each fill or run. To do so, the full processing power of the whole cluster is used to recalculate these values in real time. The computing time is in the order of 7 minutes. During Run1, all these alignments and calibrations were performed offline, with a computing time of about 1 hour, and applied to the reconstruction. It is to be noted that these same constants are also used for offline processing.

The HLT2 can then read the events from the disk buffer filled in by HLT 1, and perform a full reconstruction using these constants. LHCb is the first HEP experiment with a full calibration, alignment and reconstruction done in real time.

The output of the HLT2 is mainly composed of two streams.

The first and new stream compared to Run 1 is called "Turbo" [4]. This corresponds to the events that have been fully reconstructed using the aforementioned constants and can be used for physics almost directly out of the HLT. Another advantage of outputting reconstructed events is that their size is greatly reduced compared to the size of a raw event: 10kB instead of 70 kB. The drawback of this approach is that a mistake in the reconstruction would result in a net data loss, since the raw detector information is not preserved. The nominal output rate of the turbo stream is 2.5 kHz, which results in a throughput of 25 MB.s^{-1}. However, during 2015, and for the sake of commissioning this very ambitious change, the HLT 2 was adding to the Turbo event the corresponding RAW event to allow further offline crosscheck. This resulted in a much larger throughput of 187.5 MB.s^{-1}.

The second stream, called "FULL stream", corresponds to what was described for Run 1, that is RAW files that need further offline processing, as described in section 4. There is and always will be a need for such a stream to allow for exotic studies beyond the Standard Model. This stream outputs 10 KHz of 70 kB events i.e. 700 MB.s^{-1} to the storage.

To achieve the Turbo stream required a tremendous amount of work and greatly complicated some Online operational aspects – in particular the control system needed to be adapted to handle in parallel the asynchronous behavior of HLT 1, HLT 2 and the calibration procedure. However, the outcome is really worth the effort: the Turbo stream was used to perform the early measurement cross-section, and the results were presented only one week after the data acquisition (see for example the J/Ψ cross section measurement accepted by JHEP in Fig. 5)

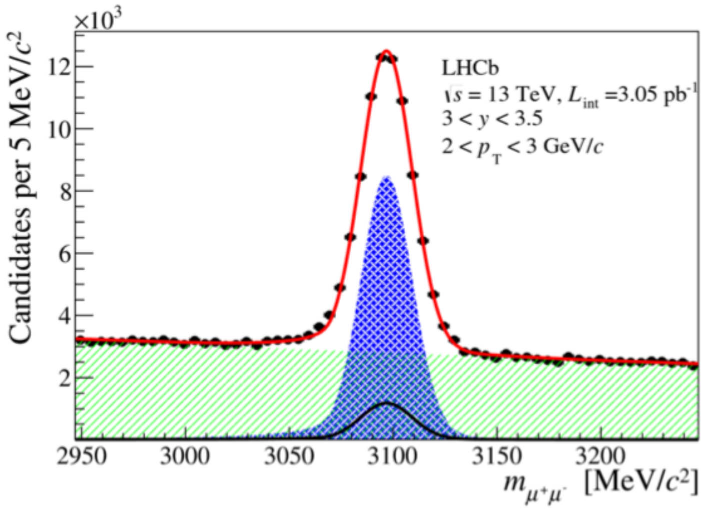


Figure 5: J/Ψ cross section measurement.

4. Offline computing

A number of changes in the Offline computing were carried out during the LHC long shutdown. Some of these changes are general improvements based on the experience gained during the Run 1 period, while some others are directly related to the changes in the Online.

The Offline processing during Run 1 comprises three major workflows:

- Real data processing, i.e. the processing the FULL stream collected Online from LHC collisions until its readiness for physics analysis.
- Monte Carlo (MC) simulation, which represents the biggest share.
- User analysis executed by individual physicists or groups to study the aforementioned data.

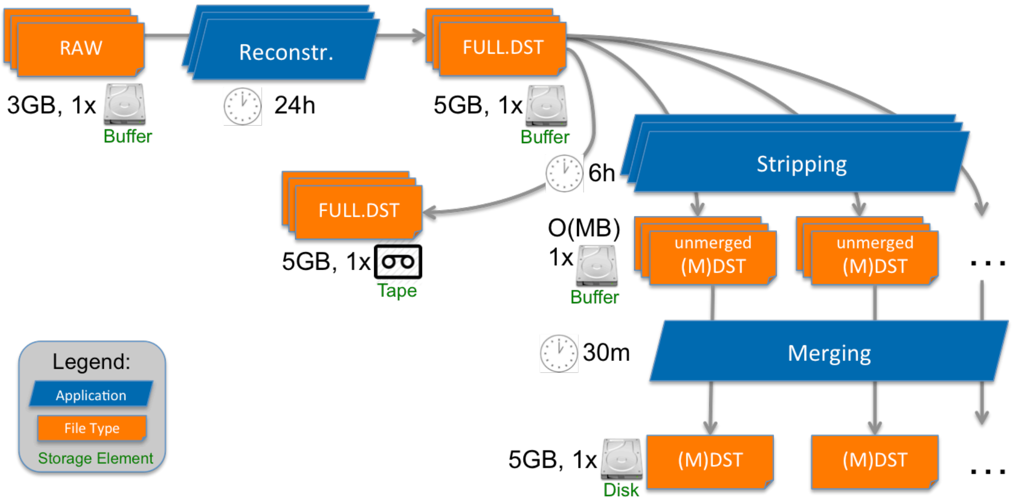


Figure 6: FULL stream offline data processing workflow.

Before being available to the physicists for studies, the RAW files produced by the FULL stream need to go through several centrally managed production steps, be it for Run 1 or Run 2:

- Reconstruction: this step performs the full reconstruction of the physics events from the raw detector information. It typically takes around 24 hours to reconstruct a 3GB RAW file. The output file is of type FULL.DST.
- Stripping: it consists in streaming the events into different "buckets" corresponding to different physics use cases, defined by the physics working groups. An event that would not match any of the bucket will simply be discarded. An event might belong to several buckets, but the overlap is minimized and is currently in the order of 10%. For one input file, a stripping job produces as many output files as buckets, that is 13. These files are called "unmerged DST" files. A slimmed down version of the DST file format containing reduced information is sometimes produced instead of the DST: the MDST files.
- Merging: A merging job takes as input unmerged DST files from the same bucket, and merge as many of them as needed to produce an output file of 5GB. These output files are the DST files used by the physicists to perform their analysis.

The Offline computing team sometimes performs "re-processing" campaigns:

- Re-reconstruction: following changes in the algorithms or constants, in order to improve the reconstruction quality.
- Re-stripping: imrpoved or new definition of a bucket.
- Incremental stripping: small addition to the previous complete stripping.

The common point between these productions is that their input files are stored on tapes, and thus need to be staged on disk before being processed.

During Run 1, a re-processing production would be started, which translates in many jobs being created. As the input files for these jobs are unavailable at the time of their creation, they would be put on hold, while a stager system integrated to DIRAC [5, 6, 7, 8, 9] –the GRID middleware used by LHCb – would trigger and monitor bring-on-line operations for the required files. Once an input file has been staged, the stager would issue a callback to the matching job. This job will then download the file for processing on the local worker node from the tape cache. The major flaw of this approach is that there are many jobs competing to have their file staged, resulting in the garbage collector of the tape system possibly removing a file from the cache between the moment a job would have received the callback and the moment it would download the file. Such a job would need to go through the whole process again. This was a source of inefficiency and operational burden.

As of 2015, a new operational model has been put in place. Prior to starting the re-processing production, the data management team replicates all the necessary files to a disk based storage using the FTS3 services [10]. This offload all the complexity of managing the bring-on-line requests to this service. The jobs are then copying from the disk storage to the worker node, with no risk of seeing the file deleted. Once the production is finished and has been validated, the data management team takes care of removing the disk replicas. This procedure shows more efficient, less error prone, and is more permissive in case of problems with the production definition. It is to be noted that thanks to the Turbo stream, no re-reconstruction is needed anymore as of 2015.

The new Turbo stream implemented in the Online requires a special treatment from the Offline point of view. Although the content of the Turbo files is fully reconstructed events ready to be used for physics analysis, their format is Online specific. The events thus need to be converted in a more commonly used format, compatible with ROOT.

As described in section 3, the complexity of the turbo stream led us to carry extensive certification for it. In order to validate the good behavior of it, the raw detector information were kept alongside the fully reconstructed event. A complex offline production was then performing the traditional workflow applied to the full stream on the raw detector information, and comparing its output to the online reconstructed events. This procedure is expected to end mid-2016.

One of the major changes in the Offline computing model is the way to use the grid [11]. During Run 1, the different tiers levels were dedicated to specific tasks:

- Tier 0 and Tier 1: these are the major sites with big storage capacity. Any type of jobs can run there, but the centralized real data productions were run exclusively on these sites.
- Tier 2 and Tier 2D: the Tier 2 only provide computing resources, while Tier 2D also provide a small amount of storage resources. They were used to run MC simulation and user jobs.

This original design resulted from the network connections planned between sites (LHCONE and LHCOP). However, the connections and networks perform beyond expectations, and such a strict subdivision is not necessary anymore.

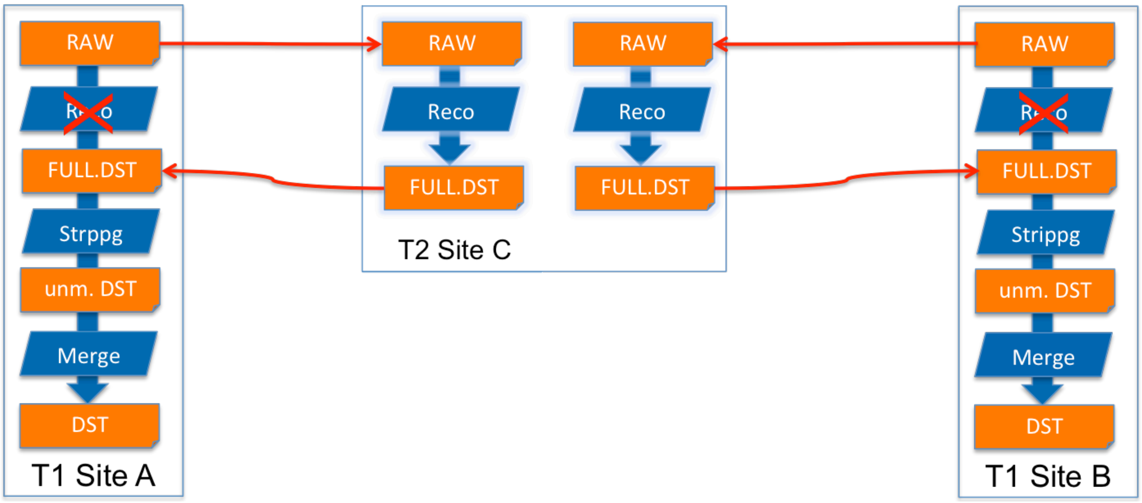


Figure 7: LHCb offline data processing workflow execution on T0, 1 and 2 sites during Run 2.

As of 2015, the LHCb computing model gave up the MONARC system [12] and adopted the so called "Mesh processing" approach, where analysis jobs would run in priority where the data is sitting, but any site (Tier 1 or Tier 2) can process data from any other site. This shows useful in particular when a site is lacking behind the others. Note that the job brokering remains data driven in a first instance.

Another big change that was put in production in 2015 is the way the very large MC productions are handled. Previously, a Monte Carlos simulation request was done by a working group, and once a production manager would have approved the request – meaning that there were resources available – the jobs would be created and the production would be processed till the end. After the operational burden and resources waste provoked by a huge buggy production, it was decided to change this workflow.

The new procedure works as follow:

1. A working group asks for a Monte Carlo production, just as before.

2. Once approved by a production manager, a small number of jobs are created, generating a small number of events. All these jobs will run at a defined site. This is called the "validation production".

3. The output of these jobs is analyzed.

4. If and only if the output is satisfactory can the big scale production start.

This procedure avoids wasting time and resources in producing useless output. While the principle is very simple, implementing it is fairly complex. It makes an extensive use of the so called "Elastic Monte Carlo jobs" [13]. These jobs have the possibility to decide at run time how many MC events will be produced, primarily to maximize the number of events produced while optimizing the resource usage and avoiding being killed by the batch system.

The nice aspect of all this procedure to validate the MC requests is that it is fully automated. The danger with such a system is that the physics working groups would start offloading the testing of their software to the operation team.

5. Summary

Many changes were carried during the long shutdown, both in the Online world and in the Offline realm. Some of them where driven by the changes in the LHC running conditions, while others ensue from the experience gained during Run 1. 2015 was the year of the restart, and the year to probe, test and certify all these changes, some of which being very heavy. LHCb has been astonishingly successful with all these improvements. All the modifications described in this paper work as expected, sometimes even beyond expectations. It is very encouraging, since they pave the way for our major upgrade before Run 3.

References

[1] LHCb, “Lhcb technical proposal,” *CERN/LHCC*, vol. 4, 1998.

[2] G. Papotti, R. Alemany, R. Calaga, F. Follin, R. Giachino, W. Herr, R. Miyamoto, T. Pieloni, and M. Schaumann, “Experience with Offset Collisions in the LHC,” p. 3 p, Sep 2011.

[3] R. Alemany-Fernandez, F. Follin, and R. Jacobsson, “The LHCB Online Luminosity Control and Monitoring,” p. 3 p, May 2013.

[4] S. Benson, M. Vesterinen, V. Gligorov, and M. Williams, “The lhcb turbo stream,” Computing in High Energy Physics, (Okinawa, Japan), April 2015. http://stacks.iop.org/1742-6596/664/i=8/a=082004.

[5] F. Stagni, P. Charpentier, R. Graciani, A. Tsaregorodtsev, J. Closier, Z. Mathe, M. Ubeda, A. Zhelezov, E. Lanciotti, and V. Romanovskiy, “Lhcbdirac: distributed computing in lhcb,” in *Journal of Physics: Conference Series*, vol. 396, p. 032104, IOP Publishing, 2012.

[6] A. Tsaregorodtsev, V. Garonne, and I. Stokes-Rees, “Dirac: A scalable lightweight architecture for high throughput computing,” in *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, GRID ’04, (Washington, DC, USA), pp. 19–25, IEEE Computer Society, 2004.

[7] A. Tsaregorodtsev, M. Bargiotti, N. Brook, A. C. Ramo, G. Castellani, P. Charpentier, C. Cioffi, J. Closier, R. G. Diaz, G. Kuznetsov, *et al.*, "Dirac: a community grid solution," in *Journal of Physics: Conference Series*, vol. 119, p. 062048, IOP Publishing, 2008.

[8] C. Haen, A. Tsaregorodtsev, and P. Charpentier, "Data management system of the dirac project," Computing in High Energy Physics, (Okinawa, Japan), April 2015. http://stacks.iop.org/1742-6596/664/i=4/a=042025.

[9] C. Haen, P. Charpentier, A. Tsaregorodtsev, and M. Frank, "Federating lhcb datasets using the dirac file catalog," Computing in High Energy Physics, (Okinawa, Japan), April 2015. http://stacks.iop.org/1742-6596/664/i=4/a=042025.

[10] A. Ayllon, M. Salichos, M. Simon, and O. Keeble, "Fts3: New data movement service for wlcg," in *Journal of Physics: Conference Series*, vol. 513, p. 032081, IOP Publishing, 2014.

[11] C. Eck, J. Knobloch, L. Robertson, I. Bird, K. Bos, N. Brook, D. DÃijllmann, I. Fisk, D. Foster, B. Gibbard, C. Grandi, F. Grey, J. Harvey, A. Heiss, F. Hemmer, S. Jarp, R. Jones, D. Kelsey, M. Lamanna, H. Marten, P. Mato-Vila, F. Ould-Saada, B. Panzer-Steindel, L. Perini, Y. Schutz, U. Schwickerath, J. Shiers, and T. Wenaus, *LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005)*. Technical Design Report LCG, Geneva: CERN, 2005.

[12] I. Bird, "Computing for the large hadron collider," *Annual Review of Nuclear and Particle Science*, vol. 61, no. 1, pp. 99–118, 2011.

[13] F. Stagni and P. Charpentier, "Jobs masonry with elastic grid jobs," Computing in High Energy Physics, (Okinawa, Japan), April 2015. http://stacks.iop.org/1742-6596/664/i=6/a=062060.