PROCEEDINGS OF SCIENCE

PoS(ISGC2015)005

The LHCb Distributed Computing Model and Operations during LHC Runs 1, 2 and 3

Stefan ROISER [a], Adrian CASAJUS[b], Marco CATTANEO[a], Philippe CHARPENTIER[a], Peter CLARKE[c], Joel CLOSIER[a], Marco CORVO[d], Antonio FALABELLA[e], Jose FLIX MOLINA[fg], Joao Victor De FRANCA MESSIAS MEDEIROS[h], Ricardo GRACIANI DIAZ[b], Christophe HAEN[a], Mikhail HUSHCHYN[i], Cinzia LUZZI[a], Zoltan MATHE[a], Andrew MCNAB[j], Raja NANDAKUMAR[k], Stefano PERAZZINI[l], Daniela REMENSKA[m], Vladimir ROMANOVSKIY[n], Michail SALICHOS[a], Renato SANTANA[h], Mark SLATER[o], Luca TOMASSETTI[d], Andrei TSAREGORODSEV[p], Andrey USTYUZHANIN[i], Vincenzo VAGNONI[l], Aresh VEDAEE[q] and Alexey ZHELEZOV[r]

[a]*European Organization for Nuclear Research, CERN*
Geneva, Switzerland
[b]*University of Barcelona*
Barcelona, Spain
[c]*University of Edinburgh*
Edinburgh, United Kingdom
[d]*University of Ferrara and INFN*
Ferrara, Italy
[e]*INFN CNAF*
Bologna, Italy
[f]*Port d'Informació Científica (PIC), Universitat Autònoma de Barcelona*
Bellaterra (Barcelona), Spain
[g]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, CIEMAT*
Madrid, Spain
[h]*CBPF - Brazilian Center for Physics Research*
Rio de Janeiro, Brazil
[i]*Yandex School of Data Analysis*
Moscow, Russia
[j]*University of Manchester*
Manchester, United Kingdom
[k]*STFC - Rutherford Appleton Laboratory*
Harwell Oxford, United Kingdom
[l]*Universita e INFN*
Bologna, Italy
[m]*Nikhef National institute for subatomic physics*
Amsterdam, The Netherlands
[n]*Institute for High Energy Physics*
Protvino, Russia
[o]*University of Birmingham*
Birmingham, United Kingdom
[p]*Centre National de la Recherche Scientifique*
Marseille, France
[q]*CC-IN2P3 - Centre de Calcul*
Lyon, France
[r]*Ruprecht-Karls-Universitaet*
Heidelberg, Germany
E-mail:
Stefan.Roiser@cern.ch, Adria@ecm.ub.edu, Marco.Cattaneo@cern.ch,
Philippe.Charpentier@cern.ch, Peter.Clarke@ed.ac.uk,
Joel.Closier@cern.ch, Marco.Corvo@cern.ch,
Antonio.Falabella@cnaf.infn.it, Jose.Flix.Molina@cern.ch,
Joao.Victor.Medeiros@cern.ch, Graciani@ecm.ub.edu,
Christophe.Haen@cern.ch, Mikhail.Hushchyn@cern.ch,
Cinzia.Luzzi@cern.ch, Zoltan.Mathe@cern.ch, Andrew.Mcnab@cern.ch,
Raja.Nandakumar@cern.ch, Stefano.Perazzini@cern.ch,
DanielaR@nikhef.nl, Vladimir.Romanovskiy@cern.ch,
Michail.Salichos@cern.ch, Renato.Santana@cern.ch,
Mark.Slater@cern.ch, Luca.Tomassetti@unife.it, ATsareg@in2p3.fr,
Andrey.Ustyuzhanin@cern.ch, Vincenzo.Vagnoni@bo.infn.it,
Aresh.Vedaee@cern.ch, Zhelezov@physi.uni-heidelberg.de

PROCEEDINGS OF SCIENCE

LHCb is one of the four main high energy physics experiments currently in operation at the Large Hadron Collider at CERN, Switzerland. This contribution reports on the experience of the computing team during LHC Run 1, the current preparation for Run 2 and a brief outlook on plans for data taking and its implications for Run 3. Furthermore a brief introduction on LHCbDIRAC, i.e. the tool to interface the experiment distributed computing resources for its data processing and data management operations is given.

During Run 1 several changes in the online filter farms had impacts on the computing operations and the computing model such as the replication of physics data, the data processing workflows and the organisation of processing campaigns. The strict MONARC model originally foreseen for LHC distributed computing was changed. Furthermore several changes and simplifications in the tools for distributed computing were taken e.g. for the software distribution, the replica catalog service or the deployment of conditions data. The reasons, implementations and implications for all these changes will be discussed.

For Run 2 the running conditions of the LHC will change which will also have an impact on the distributed computing as the output rate of the high level trigger (HLT) approximately will double. This increased load on computing resources and also changes in the high level trigger farm, which will allow a final calibration of data will have a direct impact on the computing model. In addition more simplifications in the usage of tools are foreseen for Run 2, such as the consolidation of data access protocols, the usage of a new replica catalog and several adaptions in the core the distributed computing framework to serve the additional load. In Run 3 the trigger output rate is foreseen to increase. One of the changes in HLT, to be tested during Run 2 and taken further in Run 3, which allows direct output of physics data without offline reconstruction will be discussed.

LHCb also strives for the inclusion of cloud and virtualised infrastructures for its distributed computing needs, including running on IaaS infrastructures such as Openstack or on hypervisor only systems using Vac, a self organising cloud infrastructure. The usage of BOINC for volunteer computing is currently in preparation and tested. All these infrastructures, in addition to the classical grid computing, can be served by a single service and pilot system. The details of these different approaches will be discussed.

PoS(ISGC2015)005

International Symposium on Grids and Clouds 2015
15-20 March 2015
Academia Sinica, Taipei, Taiwan

	Run 1	Planned for 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
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

This paper describes activities of the LHCb experiment [16], currently in operation at the Large Hadron Collider at CERN, Switzerland, in the realm of distributed computing. The LHC project is designed to be operated several decades and the activities are divided into data taking (Run) and maintenance, upgrade (Shutdown) periods. This paper covers the LHCb activities during Run 1 from fall 2010 until early 2013, the upcoming Run 2 foreseen for mid 2015 until mid 2018 and Run 3 planned to start in early 2020. This paper is divided into four major sections which correspond to the major systems and activities LHCb relies on for distributed computing. For each section the status as at the end of Run 1, the outlook for Run 2 and where applicable future perspectives looking at Run 3 and beyond will be described. In section 2 the conditions and the evolution of the LHC machine, delivering colliding bunches to the LHCb detector and the event filtering are described. Section 3 describes the major data processing workflows executed on distributed computing resources and their evolution. In section 4, the data management, both in the areas of data access and data transfers are described. Section 5 describes the evolution of services on which the experiment relies for its distributed computing operations, but they are not owned or developed by LHCb. Finally section 6 provides a summary of this paper. A dedicated overview of LHCb and of the other 3 main LHC experiments in the view of Run 2 is provided in [7]. For its interaction with distributed computing resources, LHCb has developed the LHCbDIRAC middleware [22], which is based on the DIRAC interware project [24, 25].

2. LHC Conditions and Online Activities

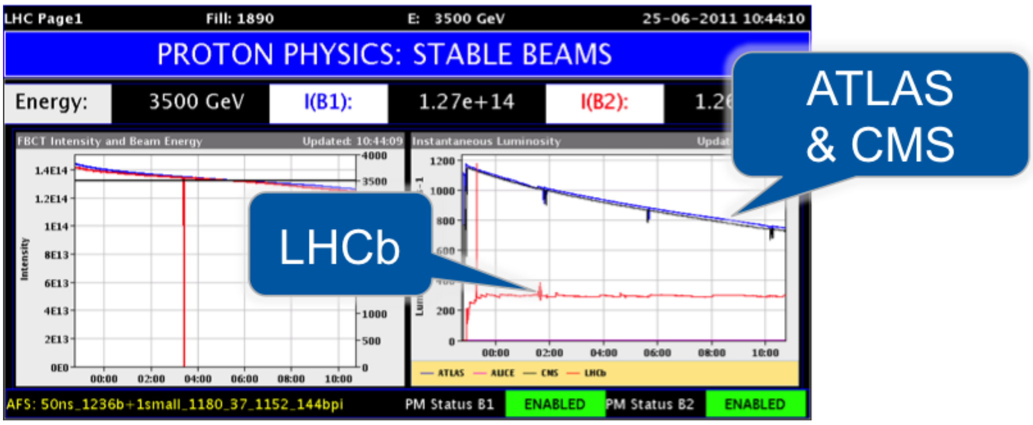
The LHC optimal beam conditions during Run 1 and the planned conditions for Run 2 are summarised in Table 1.

Apart from the increased beam energy, the most important parameter which changes during Run 2 is the decrease of bunch spacing. The beam crossing frequency will double with respect to Run 1 at the interaction points. The changes in beam conditions result in an increase of the maximum instantaneous luminosity of the LHC by one order of magnitude.

The increase of maximum luminosity will be mainly observed by the general purpose detectors at the LHC, ATLAS and CMS. As can be seen in Table 2, the maximum instantaneous luminosity will stay below these maximum rates and constant throughout LHC Runs 1 and 2 for the LHCb

	Run 1	Planned for Run 2
Maximum instantaneous luminosity	$4 * 10^{32}$ cm^{-2}s^{-1}	$4 * 10^{32}$ cm^{-2}s^{-1}
Average number of collisions per bunch crossing (μ)	1.6	1.2

Table 2: The LHC beam conditions observed at the LHCb interaction point.



LHC Page1 Fill: 1890 E: 3500 GeV 25-06-2011 10:44:10

PROTON PHYSICS: STABLE BEAMS

Energy: 3500 GeV I(B1): 1.27e+14 I(B2): 1.2

FBCT Intensity and Beam Energy Updated: 10:44:09

Instantaneous Luminosity Updat

AFS: 50ns_1236b+1small_1180_37_1152_144bpi PM Status B1 ENABLED PM Status B2 ENABLED

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

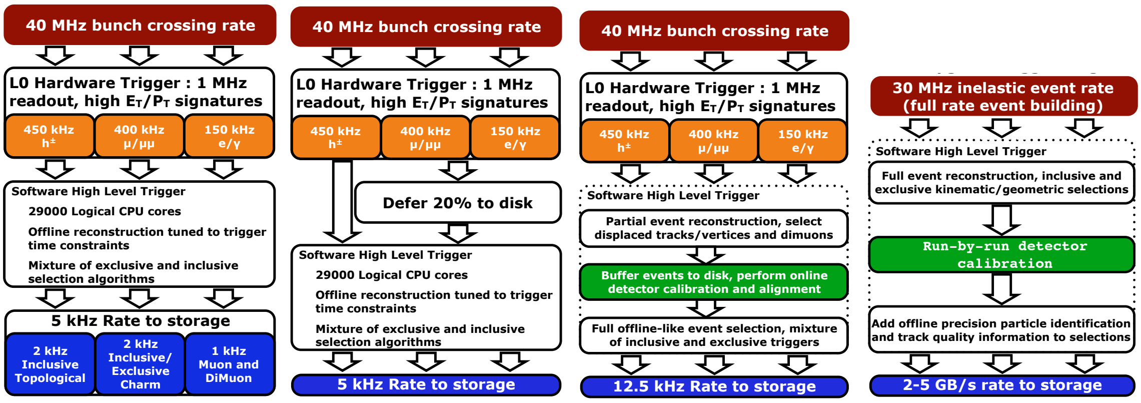
experiment. This is achieved by a technique called luminosity levelling [2, 20], through which the beams at the LHCb interaction point will not collide head on but are slightly displaced. While the absolute beam luminosity will decrease during an LHC fill, this displacement of beams will be reduced such that the instantaneous luminosity for LHCb will stay constant (see also Fig. 1).

In Run 2 the maximum LHC instantaneous luminosity will increase as a result of the reduced bunch spacing. As the instantaneous luminosity at LHCb will stay the same during Run 1 and Run 2, a reduction of simultaneous particle collisions per bunch crossing in Run 2 will occur. The estimate is that the complexity of events will be slightly reduced by the decrease of this “in-time” pile up but outweighed by the increase of “out-of-time” pile up, i.e. multiple collisions seen inside the detector originating from previous or subsequent events happening over time.

LHCb has been processing events at a rate of up to 20 MHz during Run 1 [1] and will process at 40 MHz during Run 2. These events are filtered (triggered) by various hardware and software triggers to a rate of several kHz which will finally be stored in "RAW" files and exported from the detector site to be further processed in a worldwide distributed computing environment, mainly organised via the Worldwide LHC Computing Grid (WLCG) [11] which will be discussed in sections 3 and beyond.

After the Level 0 trigger which is implemented in hardware, another important ingredient in the filtering of the events close to the detector is done in the high level trigger (HLT), which is a software trigger. During Run 1 the hardware trigger was decreasing the event rate from the original bunch crossing rate to 1 MHz and the HLT was further decreasing the event rate to up to 5 kHz (see Fig 2a). Also during Run 1 a deferral of event triggering was introduced (see Fig 2b). The deferred triggering allowed all events that could not be processed "live" during the LHC fill to be stored on

[1] The 40 MHz in Fig. 1 refer to the nominal LHC rates for all schemes


40 MHz bunch crossing rate
L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures
450 kHz $h^\pm$
400 kHz $\mu/\mu\mu$
150 kHz e/γ
Software High Level Trigger
29000 Logical CPU cores
Offline reconstruction tuned to trigger time constraints
Mixture of exclusive and inclusive selection algorithms
5 kHz Rate to storage
2 kHz Inclusive Topological
2 kHz Inclusive/ Exclusive Charm
1 kHz Muon and DiMuon
40 MHz bunch crossing rate
L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures
450 kHz $h^\pm$
400 kHz $\mu/\mu\mu$
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 Rate to storage
40 MHz bunch crossing rate
L0 Hardware Trigger : 1 MHz readout, high E_T/P_T signatures
450 kHz $h^\pm$
400 kHz $\mu/\mu\mu$
150 kHz e/γ
Software High Level Trigger
Partial event reconstruction, select displaced tracks/vertices and dimuons
Buffer events to disk, perform online detector calibration and alignment
Full offline-like event selection, mixture of inclusive and exclusive triggers
12.5 kHz Rate to storage
30 MHz inelastic event rate (full rate event building)
Software High Level Trigger
Full event reconstruction, inclusive and exclusive kinematic/geometric selections
Run-by-run detector calibration
Add offline precision particle identification and track quality information to selections
2-5 GB/s rate to storage

(a) Run 1 (b) Run 1 deferred (c) Run 2 (d) Run 3

Figure 2: LHCb Trigger schemes implemented in LHC Runs 1 and 2 and foreseen for Run 3.

local disk caches in the HLT farm and later be processed during the LHC inter-fill gaps.

For Run 2 the event output rate will increase to 12.5 kHz and the HLT trigger scheme will be further modified. The software triggering will be split into two parts (see Fig 2c). The HLT 1 will do a partial event reconstruction and reduce the event rate to O(80kHz). The deferral of events that cannot be processed live will happen after the HLT 1 step. These events are also used for the detector calibration and alignment which is supposed to take place in the order of minutes. After the calibration and alignment has been done, the new constants will be applied to all of the HLT processes for the rest of the LHC fill[2]. HLT 2 will do a full event reconstruction and apply a further selection down to the final HLT output event rate. This calibration and alignment is supposed to be the final one for any further data processing activities. Therefore the offline data processing, as described in section 3, is expected to be the final processing pass and analysts can use these data for physics studies right away. These changes in the HLT farm for Run 2 change also the offline data processing workflow in the sense that during Run 1 a first processing pass was done offline, the output of this pass was used to do the detector calibration and alignment and only towards the end of a calendar year a so-called reprocessing campaign was launched, which re-did the offline data processing workflow for all data collected so far. These re-processing activities were usually a huge load on storage systems and computing resources, as the whole dataset needed to be reprocessed in a short period of time and mostly also in parallel with ongoing data taking and first pass processing of new data from the LHC.

The average event size of a LHCb raw event when written to file after the HLT processing was of about 60 kBytes during Run 1, this size is expected to be the similar also during Run 2.

One more change introduced in the online data processing workflow for Run 2 is the Turbo Stream [5]. This concerns reconstruction data processing done within the HLT farm in contrast to "normal" offline data processing as described in section 3. The HLT farm in Run 2, and especially the HLT 2, reconstruct event data close to the quality of the offline processing workflow. This reconstruction quality allows to perform some of the physics selections directly in the HLT farm which constitute the Turbo Stream data. For those the output of the HLT2 will not need to be further

[2] A LHC fill lasts from time of new bunches injected into the LHC machine until the beam dump.

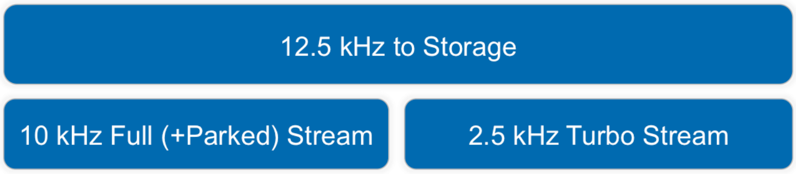


Figure 3: HLT output streams during Run 2.

processed. It is ready for physics analysis right away[3]. The different output streams coming from the HLT during Run 2 are summarised in Figure 3, where at least 10 kHz of events are processed in the standard offline mode. In case there are not enough computing resources available to process all of these, a fraction of them can be "parked", i.e. the RAW information stored and only processed after Run 2. The remaining 2.5 kHz are the Turbo stream as described above.

The concept of Turbo stream is envisaged to be taken further in Run 3, when even more RAW reconstruction, producing "ready for physics" data will be done at the HLT farm (see Fig 2c). In addition, the L0 (hardware) trigger in this run will be removed and the event filtering will be done only in software. LHCb is the first high energy physics experiment which plans to have a full software trigger which processes data at the beam crossing rate.

Executing offline workflows is another use of the HLT computing resources, e.g. for Monte Carlo Simulation. Because of the deferred trigger the HLT farm has reduced availabilities during data taking periods, therefore this opportunistic use for LHCb is restricted mainly to times where the LHC is not operating.

3. Offline Data Processing

Offline data processing activities can be divided into three major sections

- Real data processing, i.e. the processing of data as collected by the experiment from LHC collisions until its readiness for physics analysis.
- Monte Carlo simulation, mainly used for estimating systematics of the physics analysis.
- User analysis executed by individual physicists or groups to study the aforementioned data.

The data processing workflow from the point of receiving the RAW data from the detector until it is ready for physics analysis is described in Figure 4. The workflow starts from a buffer storage area at a given site where the RAW file has been replicated to.

1. The RAW file as exported from the LHCb detector site is replicated to a buffer disk storage on the processing site.

[3] In the beginning of Run 2 the RAW information of these Turbo events will be kept and stripped off in the offline processing. Later when confidence about the recorded data quality is achieved, this RAW information will not be kept anymore.

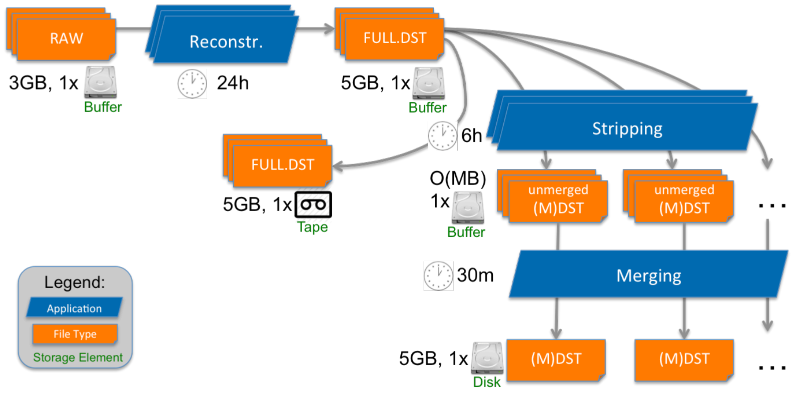


Figure 4: LHCb offline data processing workflow.

2. The reconstruction application (Brunel) will process the RAW file and return a FULL.DST file on the same buffer storage. The FULL.DST contains the reconstruction output together with necessary raw information from the input file for the subsequent steps. Once the reconstruction application has run successfully, the input RAW file is removed from buffer storage.

    (a) The FULL.DST file is copied to tape storage asynchronously.

3. The "stripping" application (DaVinci) will process the FULL.DST input file and select events according to physics selections into different streams. The output are "unmerged DST" files, again stored on the same buffer storage. Once the stripping application has run successfully, the FULL.DST input file is removed from buffer.

    (a) The aforementioned stripping step is repeated for all files of a LHCb Run[4], producing several small unmerged DST files for each physics stream.

4. Once a collection of unmerged DST files within a single LHCb Run either reaches a combined threshold of 5 GB or all files of a LHCb Run have been processed, a merging application will concatenate these unmerged DST files into a single merged DST file. The input unmerged DST files are removed from buffer once the merging has run successfully. The output DST file is put onto permanent disk storage for physics analysis.

There are two possible formats for the physics analysis events. The DST file contains multiple attributes characterising the event while the micro DST (MDST) format only contains a subset of the DST information, which should still have enough information for physics analysis. LHCb streams are mostly converted to MDST format. The difference in size ranges from 120 kB to 10 kB per event respectively.

After the completion of the data processing workflows the output (M)DST files will be replicated to multiple storage sites, which will be discussed in detail in section 4.

Initially during Run 1 the data processing workflow was executed once in a "first pass" processing to investigate and confirm the quality of the reconstructed data as well as to determine the

[4] A LHC fill is partitioned into "LHCb Runs" which can last to up to one hour.

detector calibration and alignment constants. Towards the end of the data taking year, the whole workflow was repeated in a so-called reprocessing campaign for all of the recorded data with the obtained and improved calibration and alignment measurements together with possible improvements of the software applications. For what concerns the location of data processing, LHCb initially was following the MONARC model [6] and the data processing workflows were executed at CERN (T0) and T1 sites. From this starting point several improvements were done:

- The initial idea that the first pass processing would be sufficient for physics analysis was proven to be incorrect. Therefore only a subset of each LHCb Run was processed in the first pass processing step, which was enough for the data quality, calibration and alignment work to be executed.
- Reprocessing usually started in fall of each year while data taking continued until mid December. Therefore during a certain period both processing passes were executed on the distributed computing resources producing additional load on the originally foreseen T0 and T1 centres. In order to reduce this load, a subset of T2 centres were "attached" to T1 sites for the reconstruction step of the processing workflow. The T2 site downloaded the RAW input file from the T1 storage, processed the file locally and subsequently uploaded the FULL.DST output to the T1 storage area, where it was further stripped and merged. This introduced a one-to-many relation between a certain T1 storage and helper T2 sites.
- The concept of T2 site attachment is further extended for Run 2 where the one-to-many is modified to a many-to-many relation, i.e. a given T2 site can download input files from any T1 storage and upload the output to the same T1 storage area from where the input was taken (see Fig. 5). This allows more flexibility in the workflow execution. Initially it is foreseen to use this concept only for the reconstruction step, but the implementation of the model will also allow to use this feature for other steps, e.g. stripping.
- The calibration and alignment, moved from offline processing to the HLT farm (see section 2), will also be used for offline processing and is seen as the final calibration. This has several other consequences for the data processing:
  - The offline data processing executed on distributed computing will be the final processing pass. No reprocessing of the data is foreseen until the end of Run 2.
  - The stripping retention is expected to be increased as the HLT 2 farm will have more information for its data selection. This increase in data stored will be partially damped by moving to a wider set of MDST formats for several physics streams.

Monte Carlo Simulation is executed in several steps from the event generation, through detector response, digitisation and trigger decisions, followed by the normal offline processing workflow described in Figure 4. The simulation may be executed in a rejecting and non-rejecting mode, where either the trigger and stripping decisions will slim down the number of events written or not. As interactions in the LHCb detector happen at a constant instantaneous luminosity, there is no need to simulate different in-time and out-of-time event pile-up situations. Usually the output of the final simulation step will be provided for physics analysis but any intermediate format can

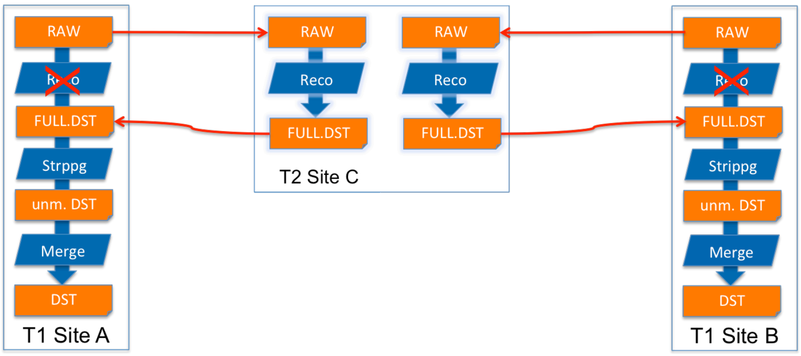


Figure 5: LHCb offline data processing workflow execution on T0, 1 and 2 sites during Run 2.

	Run 1	Run 2
Data Processing	T 0 / 1	T 0 / 1 / 2
Monte Carlo Simulation	T 2	T 2
	can also be executed on T 0 / 1 if resources available	
User Analysis	T 0 / 1	T 0 / 1 / 2D
	can also be executed on T 2 if no input data	

Table 3: LHCb workflow execution on MONARC Tier levels during the Runs 1 and 2.

also be used. A new development in the realm of simulation comes from "elastic" Monte Carlo jobs [21]. For each Monte Carlo production, the CPU time per event is calculated upfront on a test farm. With this information it is possible to calculate the number of events that can be produced within a certain time frame. In the case of grid jobs or volunteer computing jobs, where the remaining time for execution of a job can become limited, this information can be used to start the job with only a limited number of events, such that it fits into the remaining time slot.

User analysis jobs account for around 10 % of the executed work on distributed computing resources but have the highest priority in the queue when dispatching individual jobs to sites.

There have also been changes concerning the location of workflow execution on different Tier levels. Initially the MONARC model was strictly adhered to and the data processing and user analysis were executed at T0/1 sites only. Monte Carlo Simulation was processed mainly at T2 sites and in case of available resources also on T0/1 sites. During the course of Run 1, and especially during Run 2, this model has been and will be further relaxed. The evolution can be seen in Table 3 where data processing, as discussed above, is now possible also on T2 sites, and user analysis with input data on T2 D sites (which will be explained in section 4).

Apart from WLCG sites representing the Tier levels above, LHCb also leverages other computing resources. The details, split into virtualised and non-virtualised resources, are given in Table 4. In addition to the classic grid resources, there are non-pledged computing resources provided to LHCb by commercial companies such as search engine providers. Other resources include high performance computing centres and the HLT farm, as already discussed in section 2. Virtualised resources can be leveraged by LHCb mainly via two systems, i.e. Vac [17–19] where the virtual machines will be deployed on hypervisors and self-manage their load, or via Infrastructure as a Ser-

Non-virtualised resources	Virtualised resources
Classic Grid (CE, batch-system)	Vac (self managed cloud resources)
Non-pledged (commercial sites, HPC)	Vcycle (IaaS managed cloud resource)
HLT (LHCb filter farm)	BOINC (volunteer computing)

Table 4: LHCb computing resource types.

vice (IaaS) resources such as Openstack, where a system and implementation called Vcycle [17,19] will interact with the IaaS resource and spawn and tear down virtual machines according to the load in the LHCb central task queue. The BOINC [3] infrastructure is yet another possibility to execute workflows in a virtualised environment, e.g. on individual computers of volunteer contributors. All the resources mentioned above can be addressed within the same pilot framework spawned from within LHCbDIRAC [23].

4. Data Management

The LHCb data management in LHCbDIRAC [14] is divided into two main areas, data storage and data access. On each of the sites providing storage, LHCb owns one or more of the following space tokens

- "LHCb-Disk" is a disk-only resident area which holds production data available for physics analysis, usually produced by the production workflows described in section 3. In addition this space token also holds smaller areas for temporary buffers or failover areas which hold data which initially could not be replicated to the desired destination storage.
- "LHCb-Tape" is the tape-only space token. Data on this storage area has either only few accesses, e.g. raw data, or is used for archival of derived production data.
- "LHCb_USER" is the space token available to LHCb physicists to store the output of their analysis jobs.

Each of the space tokens is split into "Dirac storage elements", that are internal sub-divisions done within the LHCbDIRAC grid middleware. E.g. the LHCb-Disk space token is split into the "DST, MC_DST, BUFFER and FAILOVER" space tokens.

For what concerns data storage, LHCb initially started to have storage for both disk and tape only at T0 and T1 sites. During Run 1 the concept of Tier2D sites was developed which allowed a subset of Tier 2 sites to provide disk storage. These sites only provide LHCb-Disk space tokens and need to ramp up to a minimum of 300 TB of data storage.

Data access to the different space tokens is handled by the SRM protocol which is a front end to the storage, returning a transfer URL which provides the actual access to the storage area. With the start of Run 2 the access to disk storages (LHCb-Disk and LHCb_USER space tokens) will be executed without SRM interaction. LHCb will construct the xroot transfer urls needed for direct access. In a later stage the construction of http/webdav transfer urls is envisaged. Jobs with input data are deployed with local catalog information. The job is always sent to one of the locations of the input data and will try to access the local file replica first. If this fails and multiple replicas of the input file are available, the deployed catalog information will be used to try accessing another remote location of the input data. Input data for all production jobs is always downloaded first to the local worker node and processed from there. User analysis jobs which are the only workflows in LHCb which read input data remotely via natives protocols. Access to the LHCb-Tape space token was initially handled within LHCbDIRAC. As of Run 2 the interaction with tape storage will mainly happen via the "WLCG File Transfer Service (FTS3)" [4] which will be described in section 5. Initially, files recalled from tape storages were executed directly from disk caches in front of the tape systems. The workflows are now changed such that the recalled files will be copied to a BUFFER storage element (LHCb-Disk) and the files processed from there. This will have the advantage that recalled files will not be removed from the disk caches before being processed, because of cleaning policies of the tape systems. Furthermore the disk caches in front of tape can be reduced considerably, freeing space for other space tokens, as those are just "pass-through" areas now. For cataloging of files owned by LHCb, the experiment uses two catalogs:

- The Bookkeeping catalog is responsible for storing provenance information of data, such as the ancestors and descendants of files being processed. Furthermore, the Bookkeeping will store information about the characteristics of the data processing, such as site and worker node information where the file was produced, memory consumption, processor type, etc.
- The File Catalog stores information about replicas of a certain file. The implementation of the File Catalog was moved from the "LCG File Catalog (LFC)" to the "Dirac File Catalog (DFC)" [13] before the start of Run 2. Initially LHCb had LFC catalogs deployed on all T1 sites, each keeping a full copy of all replica information. During Run 1 it was shown that one read/write and another read-only instance at CERN were sufficient to sustain the load, therefore the T1 instances were switched off.

Since 2012, LHCb collects information from user analysis jobs about their accesses to input data. The aggregation of these accesses provides information about the popularity of data over time. This information is further used to optimise the number of replicas for datasets as they are used, and therefore optimise the amount of disk storage needed by the experiment [15].

5. Other Services

This section lists services which are not developed within LHCb but which the experiment relies on for its distributed computing activities.

- The WLCG File Transfer Service (FTS3) is used for WAN file transfers. LHCb uses the system for data replication of the production data and user output data. As of Run 2 FTS3 will also be used for bulk interactions with tape systems, such as pre-staging of input data for major data processing activities.
- The CernVM file system (CVMFS) [1, 9] is a world-wide read-only distributed file system which allows easy deployment of application software to the grid sites. It replaces previously used shared file systems deployed within the sites for which extra "software installation jobs"

needed to be run. CVMFS is organised in tier levels where the librarian deploys the software on a Stratum 0 server which will be replicated automatically to a set of Stratum 1 servers, which serve worker nodes with the needed files via squid proxies.

- CernVM [8, 10] is the chosen technology by LHCb for virtual machine images. The latest version, CernVM 3, is a very light image of a few MB which bootstraps itself with the help of CVMFS. All the needed software for LHCb applications will be deployed via CVMFS.

- WLCG monitoring provides several monitoring systems on a grid-wide level in addition to the LHCbDIRAC monitoring and accounting infrastructure. The provided services include site worker node and storage monitoring, network monitoring and file access monitoring.

- The LHCbDIRAC middleware infrastructure is deployed on special nodes via the VOBox service provided by CERN and T1 sites.

- LHCbDIRAC relies in several areas on databases which are provided by the CERN database service. Available and used technologies are MySQL and Oracle. In the future, NoSQL stores are envisaged to be used for e.g. monitoring purposes.

- The Http Federation [12] is a service provided by WLCG on top of http/webdav access to the storage areas. It provides seamless access to the whole LHCb namespace via the http and https protocols.

6. Summary

This paper provides information about the evolution of the LHCb distributed computing environment since the start of LHC Run 1 (2009) until now and provides an outlook on the upcoming Run 2 and, where applicable, possible scenarios for LHC Run 3. During Run 1 already some consolidation of services has happened (e.g. reducing the number of LFCs, reducing the number of file replicas) which has been taken further until the start of Run 2 (e.g. direct file access instead of SRM, virtualisation). For Run 3, to cope with the increased amount of data to be processed and with limited funding, several more optimisations are foreseen, e.g. the wider use of the Turbo Stream, which will be first executed and tested during Run 2.

7. Acknowledgements

The authors would like to acknowledge the help of bodies, which enable the computing activities described above. The funding agencies, personnel at the computing sites and developers and maintainers of services and middleware for distributed computing.

References

[1] C. Aguado Sanchez, J. Bloomer, P. Buncic, L. Franco, S. Klemer, and P. Mato. Cvmfs-a file system for the cernvm virtual appliance. In *Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research*, volume 1, page 52, 2008.

[2] R. Alemany-Fernandez, F. Follin, and R. Jacobsson. The LHCB Online Luminosity Control and Monitoring. (CERN-ACC-2013-0028):3 p, May 2013.

[3] D. P. Anderson. Boinc: A system for public-resource computing and storage. In *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on*, pages 4–10. IEEE, 2004.

[4] A. Ayllon, M. Salichos, M. Simon, and O. Keeble. Fts3: New data movement service for wlcg. In *Journal of Physics: Conference Series*, volume 513, page 032081. IOP Publishing, 2014.

[5] S. Benson, M. Vesterinen, V. Gligorov, and M. Williams. The lhcb turbo stream. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[6] I. Bird. Computing for the large hadron collider. *Annual Review of Nuclear and Particle Science*, 61(1):99–118, 2011.

[7] I. Bird, P. Buncic, F. Carminati, M. Cattaneo, P. Clarke, I. Fisk, M. Girone, J. Harvey, B. Kersevan, P. Mato, R. Mount, and B. Panzer-Steindel. Update of the Computing Models of the WLCG and the LHC Experiments. Technical Report CERN-LHCC-2014-014. LCG-TDR-002, CERN, Geneva, Apr 2014.

[8] J. Blomer, D. Berzano, P. Buncic, I. Charalampidis, G. Ganis, G. Lestaris, R. Meusel, and V. Nicolaou. Micro-cernvm: slashing the cost of building and deploying virtual machines. In *Journal of Physics: Conference Series*, volume 513, page 032009. IOP Publishing, 2014.

[9] J. Blomer, P. Buncic, I. Charalampidis, A. Harutyunyan, D. Larsen, and R. Meusel. Status and future perspectives of cernvm-fs. In *Journal of Physics: Conference Series*, volume 396, page 052013. IOP Publishing, 2012.

[10] P. Buncic, C. A. Sanchez, J. Blomer, L. Franco, A. Harutyunian, P. Mato, and Y. Yao. Cernvm–a virtual software appliance for lhc applications. In *Journal of Physics: Conference Series*, volume 219, page 042003. IOP Publishing, 2010.

[11] C. Eck, J. Knobloch, L. Robertson, I. Bird, K. Bos, N. Brook, D. Düllmann, 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. CERN, Geneva, 2005.

[12] F. Furano, S. Roiser, and A. Devresse. Seamless access to http/webdav distributed storage: the lhcb storage federation case study and prototype. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[13] 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. to be published.

[14] C. Haen, A. Tsaregorodtsev, and P. Charpentier. Data management system of the dirac project. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[15] M. Hushchyn, M. Cattaneo, P. Charpentier, and A. Ustyuzhanin. Disk storage management for lhcb based on data popularity estimator. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[16] LHCb. Lhcb technical proposal. *CERN/LHCC*, 4, 1998.

[17] A. McNab, P. Love, and E. MacMahon. Managing virtual machines with vac and vcycle. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[18] A. McNab, F. Stagni, and M. U. Garcia. Running jobs in the vacuum. In *Journal of Physics: Conference Series*, volume 513, page 032065. IOP Publishing, 2014.

[19] A. McNab, F. Stagni, and C. Luzzi. Lhcb experience with running jobs in virtual machines. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[20] 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. (CERN-ATS-2011-147):3 p, Sep 2011.

[21] F. Stagni and P. Charpentier. Jobs masonry with elastic grid jobs. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[22] 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*, volume 396, page 032104. IOP Publishing, 2012.

[23] F. Stagni, C. Luzzi, A. McNab, and A. Tsaregorodtsev. Pilots 2.0: Dirac pilots for all the skies. Computing in High Energy Physics, Okinawa, Japan, April 2015. to be published.

[24] 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*, volume 119, page 062048. IOP Publishing, 2008.

[25] 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, pages 19–25, Washington, DC, USA, 2004. IEEE Computer Society.