Journal of Instrumentation

A comprehensive real-time analysis model at the LHCb experiment

To cite this article: R. Aaij *et al* 2019 *JINST* **14** P04006

Published by IOP Publishing for Sissa Medialab

Received: *March 5, 2019*
Accepted: *April 1, 2019*
Published: *April 15, 2019*

A comprehensive real-time analysis model at the LHCb experiment

R. Aaij,[a] S. Benson,[a] M. De Cian,[b] A. Dziurda,[c] C. Fitzpatrick,[d] E. Govorkova,[a] O. Lupton,[e] R. Matev,[d] S. Neubert,[f] A. Pearce,[d,1] H. Schreiner,[g] S. Stahl[d] and M. Vesterinen[e]

[a] *Nikhef National Institute for Subatomic Physics,*
Amsterdam, Netherlands
[b] *Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL),*
Lausanne, Switzerland
[c] *Henryk Niewodniczanski Institute of Nuclear Physics Polish Academy of Sciences,*
Kraków, Poland
[d] *European Organization for Nuclear Research (CERN),*
Geneva, Switzerland
[e] *Department of Physics, University of Warwick,*
Coventry, United Kingdom
[f] *Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg,*
Heidelberg, Germany
[g] *University of Cincinnati,*
Cincinnati, OH, United States

E-mail: alex.pearce@cern.ch

Abstract: An evolved real-time data processing strategy is proposed for high-energy physics experiments, and its implementation at the LHCb experiment is presented. The reduced event model allows not only the signal candidate firing the trigger to be persisted, as previously available, but also an arbitrary set of other reconstructed or raw objects from the event. This allows for higher trigger rates for a given output data bandwidth, when compared to the traditional model of saving the full raw detector data for each trigger, whilst accommodating inclusive triggers and preserving data mining capabilities. The gains in physics reach and savings in computing resources already made possible by the model are discussed, along with the prospects of employing it more widely for Run 3 of the Large Hadron Collider.

Keywords: Data processing methods; Data reduction methods; Trigger concepts and systems (hardware and software)

ArXiv ePrint: 1903.01360

[1] Corresponding author.

https://doi.org/10.1088/1748-0221/14/04/P04006

Contents

1 Introduction

Experimental tests of the Standard Model must become ever more precise if small effects due to new physics are to be observed. To meet this challenge, at the Large Hadron Collider (LHC) both the centre-of-mass energy of the colliding proton-proton beams and the instantaneous luminosity delivered to the experiments are periodically increased. The corresponding increase in signal production rate must be balanced against the availability of computational resources required to store the data for offline analysis. The disk space required is given by the product of the running time of the experiment and the trigger output bandwidth defined as

$$\text{Bandwidth}\,[\text{MB/s}] \propto \text{Trigger output rate}\,[\text{kHz}] \times \text{Average event size}\,[\text{kB}].$$

When the output rate of any given trigger is dominated by events containing signal processes, tightening the selection further to reduce the output rate is undesirable. The size on disk of the full raw detector information cannot be decreased beyond the standard zero-suppression and compression techniques. Therefore, reduced event formats must be employed instead, wherein a subset of the high-level reconstructed event information computed in the final software trigger stage is recorded and sent to permanent storage. As the rate of signal processes increases, so must the reliance on reduced event formats.

The CMS, LHCb, and ATLAS experiments utilised their own versions of reduced formats during Run 2 (2015–2018) [1–3]. Typically, the reduced format contains information pertaining only to the reconstructed physics objects which passed the relevant set of trigger selections, as well as some event summary information. For a jet trigger, for example, the objects may be the momentum vector and particle multiplicity value of the highest energy jet in an event. For a heavy flavour decay trigger it may be a set of four momenta and decay vertex positions. Such an approach provides a maximal reduction in persisted event size whilst still allowing for the analysis performed in the trigger to be continued offline. This allows for higher trigger rates for a given output bandwidth, extending the physics reach of an experiment within limited computational resources. However, it also restricts the utility of an event for broader analysis and data mining. Such a reduction is then unsuitable for inclusive triggers, which constitute a large fraction of trigger output rates, as well as for analyses in which other information may be required later, such as performing flavour tagging or isolation studies, the exact input to which is not well-defined at the time of the trigger decision.

The LHCb experiment has pioneered the widespread usage of a reduced event format since the beginning of Run 2. This has been driven by the desire to continue the rich charm physics programme started in Run 1 (2010–2012) in spite of the large charm production rate in the detector acceptance, being 25 times larger than that for beauty [4, 5]. During Run 2, almost all events selected by charm triggers at LHCb were persisted in a reduced format, enabling a broader physics programme which would otherwise not be possible within the available resources. In Run 3 (2021–2023) the instantaneous luminosity delivered to the LHCb experiment will increase by a factor of five. The rate of events processed by the trigger that contain charm or beauty hadrons will scale not only by this factor, but also by a factor of two due to the implementation of a higher-efficiency full software trigger [6, 7]. Coupled with the larger raw data size produced by the upgraded detector, it is necessary for some fraction of the beauty programme to migrate to reduced event formats in order to fit within available computational resources. Given the strong reliance the beauty programme has on inclusive triggers [8, 9], the style of reduced event formats previously described are not sufficient, and so this model must be extended if the charm and beauty programmes are to be sustained into the future.

A significant evolution of the reduced event model is proposed here, in which additional reconstructed objects in the event are selected and persisted after the trigger decision has been made. When implemented with sufficient flexibility, this allows for fine-grained tuning between trigger output bandwidth and event utility offline. Since mid-2017, such a scheme has been adopted in the LHCb trigger, resulting in substantial bandwidth savings without a loss of physics reach. The reduction in bandwidth has allowed for the introduction of new trigger selections, increasing the physics scope of the experiment. It is foreseen that the reduced event format will be adopted for a majority of the LHCb physics programme in Run 3 [10]. The latest evolution, presented here, now caters to all use cases.

The rest of this paper is organised as follows. The LHCb detector and trigger system is described in section 2, and the need for a fully flexible reduced event format in Run 2 is motivated quantitatively. A technical overview of the new format is given in section 3. The benefits already achieved in Run 2 and the prospects for Run 3 are presented in sections 4 and 5. A summary is given in section 6.

2 The LHCb detector and trigger strategy

The LHCb detector is a forward-arm spectrometer designed to measure the production and decay properties of charm and beauty hadrons with high precision [11, 12]. Such objects are predominantly produced at small angles with respect to the proton beam axis [13]. A tracking system is used to reconstruct the trajectories of charged particles. It consists of a silicon vertex detector surrounding the interaction region (VELO), a silicon strip detector located upstream of a dipole magnet (TT), and three tracking stations downstream of the magnet. The latter each consist of a silicon strip detector in the high-intensity region close to the beamline (IT) and a straw-tube tracker in the regions further from the beamline (OT). Neutral particles are identified with a calorimeter system made of a scintillating pad detector (SPD), an electromagnetic calorimeter (ECAL) preceded by a pre-shower detector (PS), and a hadronic calorimeter (HCAL). Charged particle identification is provided by combining information from the ring-imaging Cherenkov detectors (RICH1 and RICH2), wire chambers used to detect muons, and the calorimeter system. Shower counters located in high-rapidity regions either side of the main detector can be used to veto events with particles produced at a very low angle to the beam, mainly used for studies of central exclusive production [14]. The LHC provides proton-proton collisions to the experiment at a rate of 30 MHz during most run periods. The detector also operates during special physics runs, such as heavy ion collisions with lead and xenon nuclei, and in a fixed target mode where a noble gas is injected in the interaction region [15, 16].

During proton-proton operation of the LHC, an interesting physics signal process occurs at a rate of around 10 Hz. To filter out the large majority of collisions that do not contain interesting information, and to fit the experimental output rate into the available computing resources, a three-stage trigger system is employed [9], illustrated in figure 1. In total, it reduces the event rate by three orders of magnitude, with the output being sent to permanent storage. The level-0 hardware trigger (L0) uses information from the calorimeter and muon systems to compute a decision using field-programmable gate arrays within a fixed latency of 4 µs. Events are selected by the L0 at a rate of about 1 MHz, and are sent to an Event Filter Farm (EFF) where they are processed by the two-stage High Level Trigger (HLT) on commodity processors. The first stage, HLT1, uses tracking and calorimetry information to perform a partial reconstruction of charged particles, and writes the raw information for each passing event to a 10 PB disk buffer at a rate of around 110 kHz. Asynchronously to HLT1, the data in the buffer are used to compute alignment and calibration constants which, if significantly different from previous runs, are saved to a conditions database [17]. The second stage of the software trigger, HLT2, performs a full event reconstruction using the latest alignment and calibration constants and all available detector information. Selections in HLT2 span a spectrum from *inclusive* selections, which require the presence of a heavy flavour decay signature such as a displaced multi-body vertex or a high transverse momentum lepton, and *exclusive* selections, which fully reconstruct signal decays. Events are written to offline storage from HLT2 at a rate of around 12.5 kHz.

Within the HLT, an event comprises a set of so-called sub-detector raw banks, each containing the zero-suppressed readout of a given sub-detector. Each trigger stage adds a set of trigger raw banks to the event which summarise how the event passed that stage. Selected events are persisted to a set of *streams* in permanent storage. During this *streaming*, different raw banks may be kept or removed depending on the stream the event is sent to. Each trigger selection in HLT2 is associated to a particular stream, and an event is sent to a stream if it passes at least one associated selection. An event may be sent to multiple streams, and different sets of raw banks for that event can be saved in each stream. The LHCb physics trigger rate is distributed over three streams:

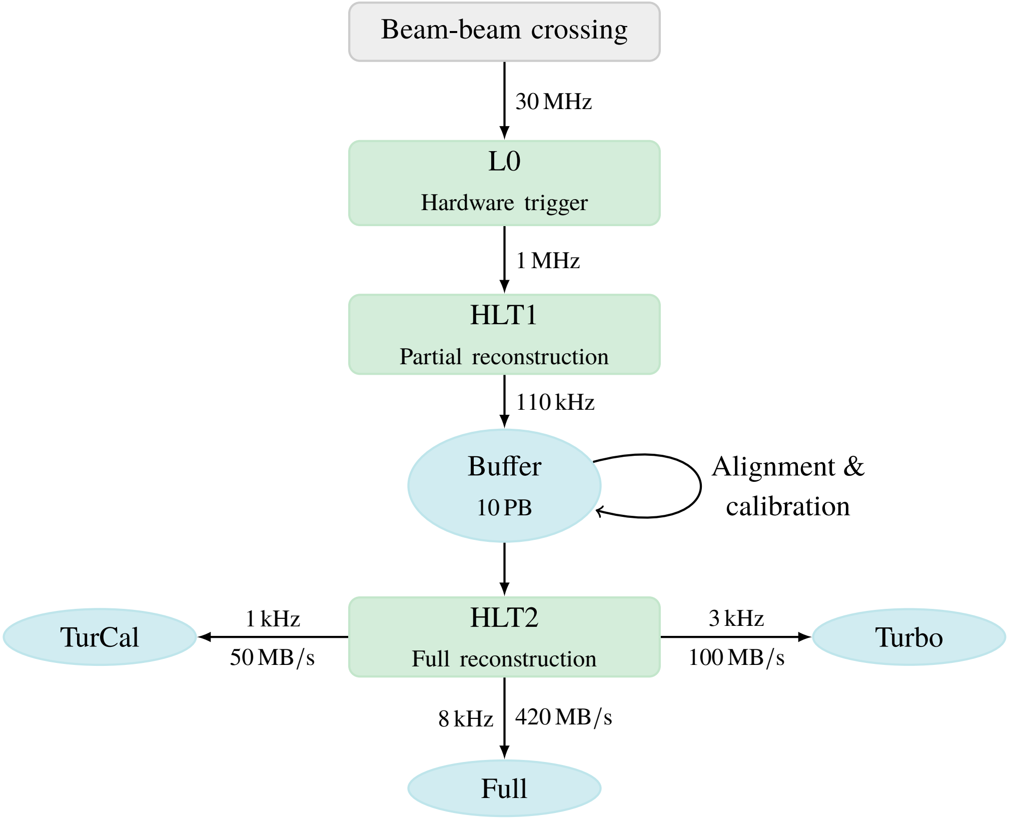

Figure 1. An overview of the LHCb trigger scheme in Run 2 [9]. The green boxes represent trigger stages, the blue ellipses represent storage units, and arrows between elements represent data flow, labelled with approximate flow rates. Events that reach a terminal storage unit are kept for offline analysis.

1. Around half of all triggered events enter the *full* stream, which contains the full set of sub-detector and trigger raw banks, with the trigger reconstruction being discarded. Events in the full stream are reconstructed by a separate application offline, followed by a set of physics selections that reconstruct thousands of decay chains for direct analysis. This scheme is used in many other experiments, where it is common for the trigger reconstruction to be of poorer quality to that of the offline reconstruction in order to fit within the timing constraints of their online systems.

2. Around a third of triggered events are sent to the *Turbo* stream, where a reduced event format is persisted in a dedicated raw bank, described in detail in section 3.

3. The remaining trigger rate is captured by the *TurCal* calibration stream, where both the reduced and full formats are kept.

The motivation for creating the reduced format used in the Turbo and TurCal streams now follows.

2.1 Real-time analysis

Here, ‘real time’ is defined as the interval between a collision occurring and the point at which the corresponding event must be either discarded forever or sent offline for permanent storage. In most high-energy physics experiments, this interval cannot exceed the time between collisions as their trigger systems are synchronous. A multi-stage trigger increases the available event processing time from one stage to the next; however each stage requires more computing power than the previous one. The addition of a disk buffer between two stages increases the effective computation time allowed in the latter stage as that stage can then also operate during periods when data are not being taken. The buffer also permits the execution of tasks that provide additional input to the latter trigger stage, such as the running of alignment and calibration algorithms.

During Run 1, LHCb installed a disk buffer such that 20 % of the L0 output was deferred to disk. This increased the effective processing power of the software trigger by allowing it to run both during beam operation and when the LHC was in downtime. Between Run 1 and Run 2, the buffer size was increased to 10 PB, with events being buffered between the two software trigger stages. As the buffer is so large, ‘real time’ can be up to two weeks during nominal data-taking [9]. The possibility to buffer the data for such a long period has allowed for the execution of the aforementioned real-time detector alignment and calibration on the data in the buffer. The increase in computing power has permitted the implementation of the full offline reconstruction in HLT2. With these substantial additions, the HLT2 reconstruction is of equal quality to what is achieved in the offline processing, such that full physics selections are performed in real time in the trigger without a loss of precision. This permits trigger selections to be much closer or identical to those applied in the offline analysis, as there are no resolution effects between online and offline to be accounted for. Such a scheme reduces experimental systematic uncertainties and saves money by making tighter trigger selections acceptable, therefore reducing the rate and output bandwidth.

With an offline-quality reconstruction in the final trigger stage (HLT2), it is no longer necessary to run another reconstruction offline. Instead, the objects created by trigger selections are written out to the permanent storage directly. Physics measurements are performed on these objects. This technique almost eliminates processing requirements offline and reduces output bandwidth, if the relevant subset of the reconstruction is smaller than the raw event. Offline analysis of trigger-level information has been used at the CMS experiment [1], called ‘scouting’, and at the ATLAS experiment [3], called ‘trigger-object level analysis’. Although the method has extended their physics reach, in both cases only object summaries are available for analysis, rather than offline-equivalent constructs, and the quality of the reconstruction is worse than that achieved offline. Since 2015, the LHCb experiment has employed a persistence model called ‘Turbo’. It allows the offline-equivalent information computed in HLT2 to be saved, sacrificing neither physics performance nor analyst convenience, as existing tools are used to process the data.

Until recently, the reduced event formats employed by the LHC experiments have described only the objects that enter into some trigger selection, such as a jet or a charm hadron and its decay products. These formats are best used to reduce the output bandwidth of events captured by exclusive trigger selections. In contrast, an inclusive trigger selection, which does not necessarily consider all the information on the physics process of interest in its decision, cannot be accommodated. In order to be able to cater for such use cases, an advanced reduced persistency model has recently been developed as an evolution of the Turbo model, and is described in the following section.

3 The Turbo data processing model

The Turbo data processing model has evolved considerably over the course of Run 2. The initial prototype, established in 2015 and described in detail in ref. [2], saved the set of objects associated to individual reconstructed decay cascades extracted from events fully reconstructed in the trigger. This exploits the event topology characteristic of hadron colliders wherein only a small subset of objects produced in the collision are relevant for analysis offline. In a hard proton-proton scatter at LHCb, on average 40 tracks are associated to the resulting primary vertex, whereas only 2–6 tracks are required to reconstruct a typical heavy flavour decay. Significant savings in persisted event size are then possible by discarding reconstructed objects not needed in the offline analysis. In this section, three new developments to the Turbo processing model are described, which together allow for all LHCb analyses to be accommodated.

In order to perform physics analyses with the output of the trigger reconstruction, decay candidates must appear in the same format used by existing analysis tools. Containers of physics object classes are serialised per event into raw banks, as illustrated in figure 2, in order to conform to the trigger output format. This format is optimised for simple event-by-event concatenation, rather than the heavily compressed format used offline. The initial prototype of the Turbo model [2] serialised only the candidates that enter the trigger decision. To allow for additional objects to be persisted, the serialisation framework used in the offline infrastructure was adapted to work within the online system. This increases the compatibility between analyses using the online and offline reconstructions, and requires that only one serialisation framework be maintained. Furthermore, trigger selection configurations which use the Turbo model were extended to allow for a set of ‘additional selections’ to be run *after* the trigger decision has been made. These allow for any reconstructed object to be captured. At the end of each event processing, the set of all C++ physics

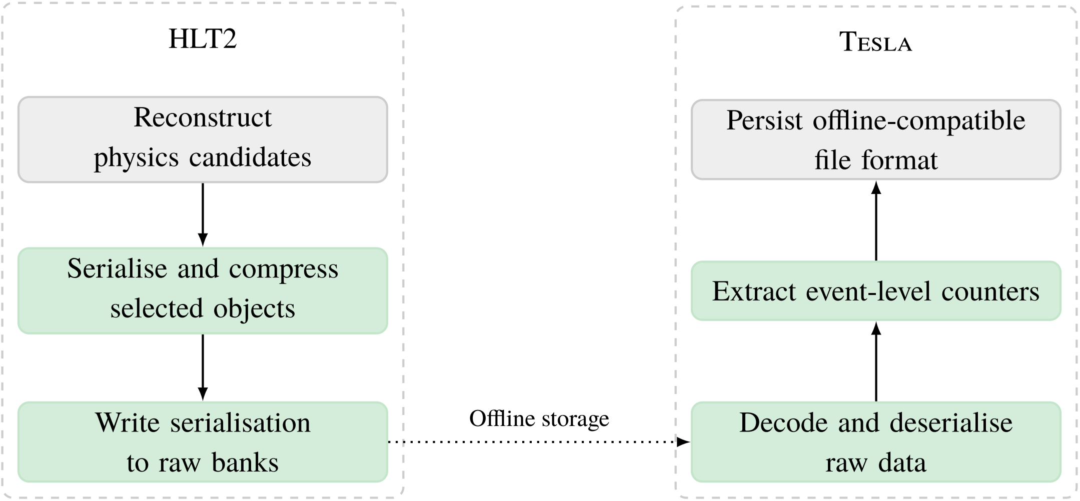
HLT2
Reconstruct physics candidates
Serialise and compress selected objects
Write serialisation to raw banks
Offline storage
Tesla
Persist offline-compatible file format
Extract event-level counters
Decode and deserialise raw data

Figure 2. Data flow in saving and restoring trigger objects from online (left) to offline (right) [2].

objects selected by all trigger lines using the Turbo model is copied to a common location in memory. The instances in the copied location are then compressed and serialised into raw banks [18], suitable for transfer within and out of the online system.

A dedicated application, called Tesla [2], runs offline to transform the HLT2 output into a format that is ready for analysis. This involves a file conversion from serialised raw data to a more compressed format used by the offline storage, as well as the computation and storage of information necessary for luminosity determination. Tesla also ensures that additional information calculated in the online reconstruction is accessible to standard analysis tools, for example event-level sub-detector occupancies and information calculated using the whole reconstructed event. When processing simulated events, the application also matches reconstructed objects to true simulated objects and stores the information as relations tables. In comparison with the traditional model of an additional offline reconstruction, the processing cost of running Tesla is negligible. In principle, some or even all parts of the work done by Tesla can be moved into the trigger itself in the LHCb upgrade, such as if compressed object containers were written directly out of HLT2 instead of encoding them into raw banks.

With the serialisation and file preparation frameworks in place, different levels of granularity on what physics objects to select are now available. In the following, the resulting flexibility is explained, such that all measurements can take advantage of the real-time model.

3.1 Standard Turbo model

A majority of trigger lines based on the Turbo model define exclusive selections where the full decay is completely specified, and no additional objects from the event are required for subsequent analysis. In this model, the objects saved are:

- The reconstructed decay chain that fired the line, which comprises:
  - The set of all tracks and neutral objects, calorimeter and PID information relating to those objects, and decay vertices that form the candidate.
  - The tracking detector clusters associated to the candidate tracks, such that the tracks can be re-fitted offline.
- All of the reconstructed primary vertices (PVs) in the event, which are necessary to perform PV mis-association studies offline.

Other reconstructed objects in the event as well as the raw data are not kept for offline processing. That allows for a significant reduction of the event size and hence also trigger output bandwidth. Offline disk space and CPU processing time are also saved as the offline reconstruction step is omitted. This model has been operational since the beginning of data-taking in 2015, and enabled the first LHCb Run 2 measurements to be presented 18 days after the data were collected [4, 5].

3.2 Complete reconstruction persistence

In order to use the Turbo model for inclusive triggers, the ability to store all the reconstructed objects in the event was introduced in the beginning of 2016. It is made available on a per-selection basis by a user flag. When enabled, the full event reconstruction as performed in HLT2 is persisted in addition

Table 1. Average event sizes for trigger lines requesting varying levels of information persistence, measured on data collected in 2018.

Persistence method	Average event size (kB)
Turbo	7
Selective persistence	16
Complete persistence	48
Raw event	69

to the information described in section 3.1. In comparison to saving the raw event, this approach reduces disk space usage and requires no further processing offline. However, the information needed to re-run the reconstruction offline is discarded. This technique permits high-rate inclusive triggers that would otherwise not be feasible [19].

Persisting the whole reconstructed event is expensive in terms of event size in comparison with only saving the trigger candidate, as shown in table 1. In most cases, only a small fraction of the full reconstructed event is required in an offline analysis. Therefore, a middle-ground between persisting only the candidate or the whole reconstructed event is introduced.

3.3 Selective reconstruction persistence

Selective persistence allows for explicit specification of which information is stored on top of the trigger candidate itself. This permits a significant event size reduction without the usual sacrifice of allowing for exploratory analysis offline. At the beginning of 2017, trigger lines were augmented with additional selections that are executed after the trigger decision has been computed. A typical additional selection captures objects in the event that are somehow related to the trigger candidate. However, any selections are possible.

As an example, consider a trigger using the Turbo model that reconstructs and selects the $D^0 \to K^- \pi^+$ charm decay, as shown in figure 3, where the information specified in section 3.1 is saved. If complete reconstruction persistence is enabled for this line, the underlying reconstructed event will also be stored. With selective persistence, additional objects are instead specified explicitly, such as all charged pions that are associated to the same PV as the D^0 and that form a good-quality excited $D^{*\pm}$ candidate. The selection framework allows for requirements to be made on both the pions themselves and on the $D^0\pi^\pm$ combination, but then only the pions that pass these cuts are persisted. The $D^{*\pm}$ candidates are discarded, since it can be built again exactly in an offline processing if required. Similar selections can be added for other extra particles, such as kaons, photons, and hyperons to support a wide spectrum of charm spectroscopy measurements with a single trigger selection.

In the LHCb upgrade for Run 3, selective persistence is a key ingredient in the migration of the physics programme to the real-time analysis model [10]. One use-case is flavour tagging, a determination of the initial flavour of a beauty or anti-beauty meson. The decision of one or more tagging algorithms together with the probability of the assigned flavour being wrong is computed using a set of reconstructed objects in the event, namely objects that are related to the same PV as the signal decay. This set can be loosely defined upfront and added as additional selections to trigger lines which reconstruct the signal beauty hadron decays of interest. As the tagging

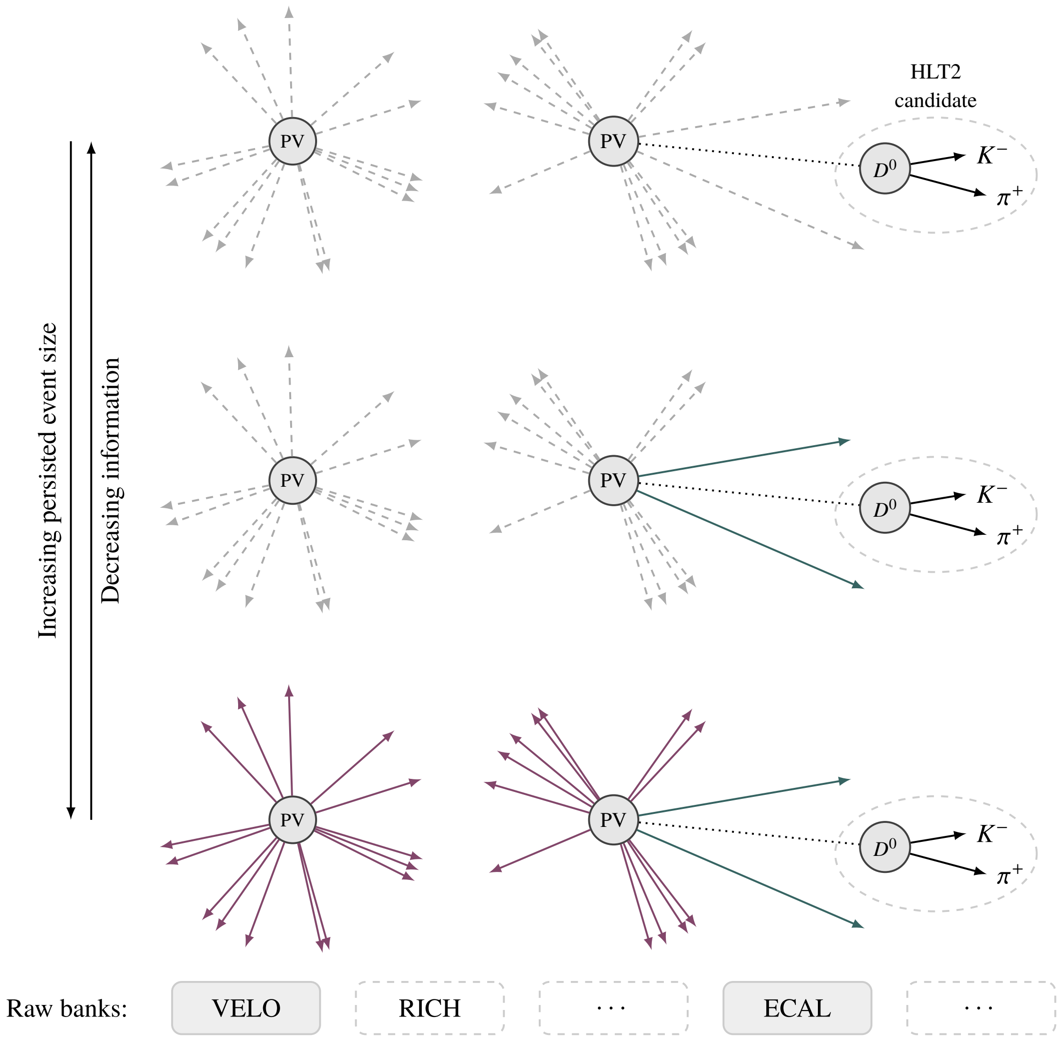

Figure 3. A cartoon of the same reconstructed event with varying levels of object persistence: Turbo (top); selective persistence (middle); and complete reconstruction persistence (bottom). Solid objects are those persisted in each case. A trigger selection may also ask for one or more sub-detector raw banks to also be stored, shown as solid rectangles.

algorithms undergo improvements during data-taking, the flavour tagging can be re-run offline using the information captured by the additional selections. Initial studies show that the set of reconstructed objects required as input to the tagging algorithms constitute only around 10 % of the space that would be required for persisting the full reconstruction.

3.4 Selective raw persistence

There are some cases in which saving all possible information from an event is required. One important example is for efficiency measurements on calibration samples. Today, detector-level efficiencies are determined from control channels whose trigger lines save the full raw detector information. The high bandwidth associated with doing this means that the choice and selection of control channels is severely restricted. This impacts the calibration sample size and leads to larger uncertainties

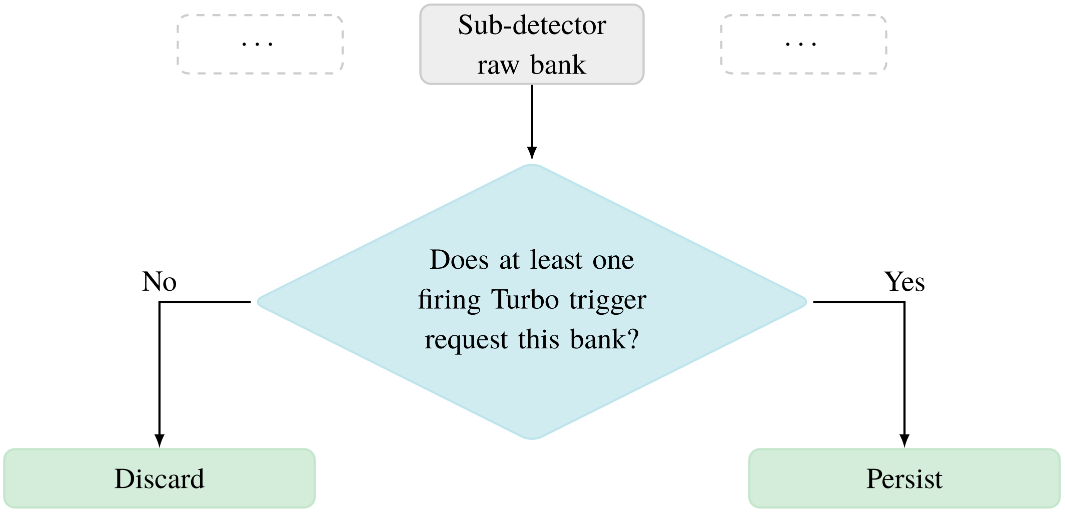

Figure 4. Algorithm flow to decide whether a given raw bank is persisted for the current event. The resulting list of banks persisted for a given event is the superset of those required by the individual lines.

on the efficiencies. The ability to study the efficiency of a given aspect of the reconstruction requires saving detector hits that were not used in the reconstructed object. However, the efficiency of reconstructing tracks with the OT, for example, would require saving only the raw banks associated with that sub-detector, and would not need the banks associated with the calorimeter or the RICH sub-detectors. If a more fine-grained, per-trigger specification of required raw banks was possible, the resulting bandwidth savings would allow the calibration samples to more than double in size, resulting in more precise efficiency determinations and therefore more accurate physics measurements.

In order to accomplish this, a new algorithm has been developed and deployed. On initialisation, it determines the list of trigger selections and their requested raw banks from the trigger configuration, as shown in figure 4. On a per-event basis, the decision of each trigger selection is examined and the superset of the required raw banks is persisted to the corresponding output stream. Raw banks not requested by any firing trigger line are discarded. This algorithm has been running in the LHCb trigger since 2018. With it, each trigger selection has complete control on additional information in the event is persisted, from any object created in the reconstruction, to the raw information created by any sub-detector. Therefore, the Turbo model is now able to cater for any use case.

4 Achievements in Run 2

Since the beginning of Run 2, a considerable fraction of triggered events have used the Turbo model. In 2018, the number of trigger lines using the Turbo model was around 50 % of the number using the traditional model, however the Turbo stream bandwidth was 25 % that of the full stream due to the reduced average event size. Physics measurements using the Turbo model include charm and J/ψ cross-sections [4, 5], the discovery of new ground-state and excited charm baryons [20, 21], searches for dark photons [22], and the characterisation of charmonium production within jets [19]. This section summarises how the new Turbo model has increased physics reach whilst reducing the event size since its introduction in 2016.

As an example of the gains that can and have been made, we consider a subset of trigger lines using the Turbo model that exclusively reconstruct and select Cabibbo-favoured decays of ground-state charm hadrons, such as $D^0 \to K^-\pi^+$ and $\Lambda_c^+ \to pK^-\pi^+$ and their charge conjugates, originating directly from beam-beam interactions. These trigger selections are intended for calibration studies, as the properties of these objects and decays are well known, as well as for charm spectroscopy, searching for and characterising excited charm states that cascade down to the ground states. In 2016, these trigger lines were enhanced by the addition of the complete reconstruction persistence, to allow for excited states to be reconstructed offline, and so the average event size increased from 7 kB in 2015 to 48 kB.[1] In 2017, the spectroscopy lines moved to the selective reconstruction persistence, reducing the average event size to 16 kB. In turn, the bandwidth used by the Turbo stream decreased from 139 MB/s to 79 MB/s. The newly-available bandwidth was then utilised for new inclusive charm baryon trigger lines that would have otherwise not been possible within the given computational resources. The selections used to reduce the set of persisted objects from the spectroscopy lines were aligned with the ones used for spectroscopy offline. The additional selections are then 100 % efficient with respect to the offline selections by definition.

5 Prospects for Run 3

The LHCb detector will be upgraded for Run 3, where the instantaneous luminosity will increase from that in Run 2 by a factor of five. With the corresponding increase in heavy flavour production rate, in addition to expected trigger efficiency gains and an increase in the raw event size, the trigger output bandwidth will increase by an order of magnitude from Run 2. Due to constraints on offline storage resources, the bandwidth is limited to a maximum of 10 GB/s [10]. The breadth of the current experimental physics programme can only survive within these resources if the fraction of trigger output rate sent to the Turbo stream increases by over a factor of two. The flexibility of the reduced event model described here has been designed to reach that goal without an inherent loss of physics performance or reach. This section briefly discusses some relevant techniques.

In principle, reducing the amount of persisted information comes with risk, as the set of information needed for a given measurement is not always known upfront, and indeed the analysis itself has often not been conceived. Information discarded in the trigger can be later required by some unforeseen analysis offline. However, the factor-five increase in instantaneous luminosity means a corresponding increase in the average number of visible beam-beam interactions per bunch crossing. Therefore, a relatively safe selection is to discard objects which are identified as originating from primary vertices other than those associated to the signal trigger object. This information is not relevant to an analysis of the trigger object as it is unrelated to the signal process, having been produced from independent parton-parton scatters.[2] Given a fully reconstructed signal candidate, its associated primary vertex can be defined as the one with the smallest impact parameter with respect to the signal momentum vector. Primary vertices from which associated information should not be

[1] The average size in 2016 across all events using the Turbo model was 42 kB, illustrating the dominating rate of these selected processes.

[2] Such a technique is similar in concept to jet grooming, in that one wishes to keep only the objects associated to the hard scatter which produced the heavy flavour quark-antiquark pair present in the event.

persisted can then be identified using a minimum impact parameter cut, the exact value of which can be tuned based on the expected resolution available to distinguish separate primary vertices.

Inclusive trigger selections of heavy flavour decays present a particularly challenging event size reduction problem as they have very high rates. The potentially incomplete reconstruction of the signal momentum vector reduces the accuracy of non-signal primary vertex suppression. Complimentary selections that reduce the persisted information in inclusively triggered events include:

- The rejection of tracks that form a very poor quality vertex with the inclusive candidate, as such tracks would not be used in an offline analysis; and

- The rejection of objects identified by a multivariate algorithm trained to distinguish uninteresting objects from those associated to the portion of the signal process captured by the inclusive selection.

Preliminary studies have shown such techniques can fully capture all signal objects in an inclusively triggered event with an efficiency of around 90 %, compared to saving the full reconstruction, whilst rejecting over 90 % of the unrelated objects. This reduces the bandwidth of the prototype inclusive beauty trigger under study from 7.5 GB/s to 1 GB/s. While promising, further work is needed to quantify any possible biases which may be associated to the signal decays which are not completely captured by the selective persistence.

6 Summary and conclusions

Traditional reduced event formats allow for a broader physics programme within available computational resources, but this is usually countered by a poorer quality reconstruction in the trigger and reduced data mining capabilities. Since 2015, a real-time alignment and calibration procedure between trigger stages has allowed the LHCb experiment to deploy and exploit its Turbo model. With this, offline-quality signal candidates are persisted directly from the trigger for later analysis. This has been crucial for the charm programme, which otherwise would have had to significantly compromise its reach and diversity.

Since 2017, the implementation of the Turbo model has been overhauled and extended to allow an arbitrary subset of the trigger reconstruction and raw sub-detector information to be persisted along with the trigger candidate. As such, the model is now capable of supporting the entirety of the experiment's broad research programme, and in particular the parts which rely on inclusive trigger selections. Given the large increase in instantaneous luminosity and trigger efficiency foreseen in Run 3, this evolution completes a crucial step in allowing the continuation of today's physics measurements into the future.

The updated Turbo model has already provided a 50 % reduction in bandwidth in comparison with saving the full reconstruction, and this saving has been exploited with the addition of new high-rate trigger selections. Even larger gains should be possible when applying similar techniques to the remaining set of trigger selections, and studies are ongoing into these avenues.

Acknowledgments

We express our gratitude to our colleagues in the CERN accelerator departments for the excellent performance of the LHC. We thank the technical and administrative staff at the LHCb institutes. We acknowledge support from CERN and from the national agencies: CAPES, CNPq, FAPERJ and FINEP (Brazil); MOST and NSFC (China); CNRS/IN2P3 (France); BMBF, DFG and MPG (Germany); INFN (Italy); NWO (Netherlands); MNiSW and NCN (Poland); MEN/IFA (Romania); MSHE (Russia); MinECo (Spain); SNSF and SER (Switzerland); NASU (Ukraine); STFC (United Kingdom); NSF (USA). We acknowledge the computing resources that are provided by CERN, IN2P3 (France), KIT and DESY (Germany), INFN (Italy), SURF (Netherlands), PIC (Spain), GridPP (United Kingdom), RRCKI and Yandex LLC (Russia), CSCS (Switzerland), IFIN-HH (Romania), CBPF (Brazil), PL-GRID (Poland) and OSC (USA). We are indebted to the communities behind the multiple open-source software packages on which we depend. Individual groups or members have received support from AvH Foundation (Germany); EPLANET, Marie Skłodowska-Curie Actions and ERC (European Union); ANR, Labex P2IO and OCEVU, and Région Auvergne-Rhône-Alpes (France); Key Research Program of Frontier Sciences of CAS, CAS PIFI, and the Thousand Talents Program (China); RFBR, RSF and Yandex LLC (Russia); GVA, XuntaGal and GENCAT (Spain); the Royal Society and the Leverhulme Trust (United Kingdom); Laboratory Directed Research and Development program of LANL (USA). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 747822.

References

[1] CMS collaboration, *Data Parking and Data Scouting at the CMS Experiment*, CMS-DP-2012-022 (2012).

[2] R. Aaij et al., *Tesla: an application for real-time data analysis in High Energy Physics*, *Comput. Phys. Commun.* **208** (2016) 35 [arXiv:1604.05596].

[3] ATLAS collaboration, *Trigger-object Level Analysis with the ATLAS detector at the Large Hadron Collider: summary and perspectives*, ATL-DAQ-PUB-2017-003 (2017).

[4] LHCb collaboration, *Measurement of forward J/ψ production cross-sections in pp collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **10** (2015) 172 [*Erratum ibid.* **1705** (2017) 063] [LHCb-PAPER-2015-037] [CERN-PH-EP-2015-222] [arXiv:1509.00771].

[5] LHCb collaboration, *Measurements of prompt charm production cross-sections in pp collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **03** (2016) 159 [*Erratum ibid.* **1609** (2016) 013] [LHCb-PAPER-2015-041] [CERN-PH-EP-2015-272] [arXiv:1510.01707].

[6] LHCb collaboration, *Framework TDR for the LHCb Upgrade: Technical Design Report*, CERN-LHCC-2012-007 (2012).

[7] LHCb collaboration, *LHCb Trigger and Online Technical Design Report*, CERN-LHCC-2014-016 (2014).

[8] R. Aaij et al., *The LHCb Trigger and its Performance in 2011*, 2013 *JINST* **8** P04022 [arXiv:1211.3055].

[9] R. Aaij et al., *Performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC*, arXiv:1812.10790.

[10] LHCb collaboration, *Computin Model of the Upgrade LHCb experiment*, CERN-LHCC-2018-014 (2018).

[11] LHCb collaboration, *The LHCb Detector at the LHC*, 2008 *JINST* **3** S08005.

[12] LHCb collaboration, *LHCb Detector Performance*, *Int. J. Mod. Phys.* **A 30** (2015) 1530022 [arXiv:1412.6352].

[13] LHCb collaboration, *Measurement of $\sigma(pp \to b\bar{b}X)$ at $\sqrt{s} = 7$ TeV in the forward region*, *Phys. Lett.* **B 694** (2010) 209 [CERN-PH-EP-2010-029] [LHCb-PAPER-2010-002] [arXiv:1009.2731].

[14] K. Carvalho Akiba et al., *The HeRSCheL detector: high-rapidity shower counters for LHCb*, 2018 *JINST* **13** P04017 [arXiv:1801.04281].

[15] C. Barschel, *Precision luminosity measurement at LHCb with beam-gas imaging*, Ph.D. thesis, RWTH Aachen University (2014).

[16] LHCb collaboration, *Precision luminosity measurements at LHCb*, [LHCb-PAPER-2014-047] [CERN-PH-EP-2014-221] 2014 *JINST* **9** P12005 [arXiv:1410.0149].

[17] G. Dujany and B. Storaci, *Real-time alignment and calibration of the LHCb Detector in Run II*, LHCb-PROC-2015-011 (2015).

[18] J. Albrecht, P. Billoir, D. H. Campora Perez, M. Cattaneo, C. Marco, B. Couturier et al., *Upgrade trigger & reconstruction strategy: 2017 milestone*, LHCb-PUB-2018-005 (2018).

[19] LHCb collaboration, *Study of J/ψ Production in Jets*, *Phys. Rev. Lett.* **118** (2017) 192001 [LHCb-PAPER-2016-064] [CERN-EP-2017-006] [arXiv:1701.05116].

[20] LHCb collaboration, *Observation of five new narrow Ω_c^0 states decaying to $\Xi_c^+ K^-$*, *Phys. Rev. Lett.* **118** (2017) 182001 [LHCb-PAPER-2017-002] [CERN-EP-2017-037] [arXiv:1703.04639].

[21] LHCb collaboration, *Observation of the doubly charmed baryon Ξ_{cc}^{++}*, *Phys. Rev. Lett.* **119** (2017) 112001 [LHCb-PAPER-2017-018] [CERN-EP-2017-156] [arXiv:1707.01621].

[22] LHCb collaboration, *Search for Dark Photons Produced in 13 TeV pp Collisions*, *Phys. Rev. Lett.* **120** (2018) 061801 [LHCb-PAPER-2017-038] [CERN-EP-2017-248] [arXiv:1710.02867].