EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-LHCb-DP-2019-001
July 17, 2019

Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC

Author list on the following page

Abstract

The LHCb collaboration has redesigned its trigger to enable the full offline detector reconstruction to be performed in real time. Together with the real-time alignment and calibration of the detector, and a software infrastructure to make persistent the high-level physics objects produced during real-time processing, this redesign enabled the widespread deployment of real-time analysis during Run 2. We describe the design of the Run 2 trigger and real-time reconstruction, and present data-driven performance measurements for a representative sample of LHCb's physics programme.

Submitted to JINST

R. Aaij[19], S. Akar[40,7†], J. Albrecht[11], M. Alexander[34], A. Alfonso Albero[24], S. Amerio[17], L. Anderlini[15], P. d'Argent[12], A. Baranov[22], W. Barter[26†,36], S. Benson[19], D. Bobulska[34], T. Boettcher[39], S. Borghi[37,26], E. E. Bowen[28†,b], L. Brarda[26], C. Burr[37], J.-P. Cachemiche[7], M. Calvo Gomez[24,c], M. Cattaneo[26], H. Chanal[6], M. Chapman[30], M. Chebbi[26†], M. Chefdeville[5], P. Ciambrone[16], J. Cogan[7], S.-G. Chitic[26], M. Clemencic[26], J. Closier[26], B. Couturier[26], M. Daoudi[26], K. De Bruyn[7†,26], M. De Cian[27], O. Deschamps[6], F. Dettori[35], F. Dordei[26†,14], L. Douglas[34], K. Dreimanis[35], L. Dufour[19†,26], G. Dujany[37†,9], P. Durante[26], P.-Y. Duval[7], A. Dziurda[21], S. Esen[19], C. Fitzpatrick[27], M. Fontanna[26], M. Frank[26], M. Van Veghel[19], C. Gaspar[26], D. Gerstel[7], Ph. Ghez[5], K. Gizdov[33], V.V. Gligorov[9], E. Govorkova[19], L.A. Granado Cardoso[26], L. Grillo[12†,26†,37], I. Guz[26,23], F. Hachon[7], J. He[3], D. Hill[38], W. Hu[4], W. Hulsbergen[19], P. Ilten[29], Y. Li[8], C.P. Linn[26†], O. Lupton[38†,26], D. Johnson[26], C.R. Jones[31], B. Jost[26], M. Kenzie[26†,31], R. Kopecna[12], P. Koppenburg[19], M. Kreps[32], R. Le Gac[7], R. Lefèvre[6], O. Leroy[7], F. Machefert[8], G. Mancinelli[7], S. Maddrell-Mander[30], J.F. Marchand[5], U. Marconi[13], C. Marin Benito[24†,8], M. Martinelli[27†,26], D. Martinez Santos[25], R. Matev[26], E. Michielin[17], S. Monteil[6], A. Morris[7], M.-N. Minard[5], H. Mohamed[26], M.J. Morello[18,a], P. Naik[30], S. Neubert[12], N. Neufeld[26], E. Niel[8], A. Pearce[26], P. Perret[6], F. Polci[9], J. Prisciandaro[25†,1], C. Prouve[30†,25], A. Puig Navarro[28], M. Ramos Pernas[25], G. Raven[20], F. Rethore[7], V. Rives Molina[24†], P. Robbe[8], G. Sarpis[37], F. Sborzacchi[16], M. Schiller[34], R. Schwemmer[26], B. Sciascia[16], J. Serrano[7], P. Seyfert[26], M.-H. Schune[8], M. Smith[36], A. Solomin[30,d], M. Sokoloff[40], P. Spradlin[34], M. Stahl[12], S. Stahl[26], B. Storaci[28†], S. Stracka[18], M. Szymanski[3], M. Traill[34], A. Usachov[8], S. Valat[26], R. Vazquez Gomez[16†,26], M. Vesterinen[32], B. Voneki[26†], M. Wang[2], C. Weisser[39], M. Whitehead[26†,10], M. Williams[39], M. Winn[8], M. Witek[21], Z. Xiang[3], A. Xu[2], Z. Xu[27†,5], H. Yin[4], Y. Zhang[8], Y. Zhou[3].

[1] *Université catholique de Louvain, Louvain, Belgium*
[2] *Center for High Energy Physics, Tsinghua University, Beijing, China*
[3] *University of Chinese Academy of Sciences, Beijing, China*
[4] *Institute of Particle Physics, Central China Normal University, Wuhan, Hubei, China*
[5] *Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, IN2P3-LAPP, Annecy, France*
[6] *Clermont Université, Université Blaise Pascal, CNRS/IN2P3, LPC, Clermont-Ferrand, France*
[7] *Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France*
[8] *LAL, Univ. Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, Orsay, France*
[9] *LPNHE, Sorbonne Université, Paris Diderot Sorbonne Paris Cité, CNRS/IN2P3, Paris, France*
[10] *I. Physikalisches Institut, RWTH Aachen University, Aachen, Germany*
[11] *Fakultät Physik, Technische Universität Dortmund, Dortmund, Germany*
[12] *Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany*
[13] *INFN Sezione di Bologna, Bologna, Italy*
[14] *INFN Sezione di Cagliari, Monserrato, Italy*
[15] *INFN Sezione di Firenze, Firenze, Italy*
[16] *INFN Laboratori Nazionali di Frascati, Frascati, Italy*
[17] *INFN Sezione di Padova, Padova, Italy*
[18] *INFN Sezione di Pisa, Pisa, Italy*
[19] *Nikhef National Institute for Subatomic Physics, Amsterdam, Netherlands*
[20] *Nikhef National Institute for Subatomic Physics and VU University Amsterdam, Amsterdam, Netherlands*
[21] *Henryk Niewodniczanski Institute of Nuclear Physics Polish Academy of Sciences, Kraków, Poland*
[22] *Yandex School of Data Analysis, Moscow, Russia*
[23] *Institute for High Energy Physics (IHEP), Protvino, Russia*
[24] *ICCUB, Universitat de Barcelona, Barcelona, Spain*

[25] *Instituto Galego de Física de Altas Enerxías (IGFAE), Universidade de Santiago de Compostela, Santiago de Compostela, Spain*
[26] *European Organization for Nuclear Research (CERN), Geneva, Switzerland*
[27] *Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[28] *Physik-Institut, Universität Zürich, Zürich, Switzerland*
[29] *University of Birmingham, Birmingham, United Kingdom*
[30] *H.H. Wills Physics Laboratory, University of Bristol, Bristol, United Kingdom*
[31] *Cavendish Laboratory, University of Cambridge, Cambridge, United Kingdom*
[32] *Department of Physics, University of Warwick, Coventry, United Kingdom*
[33] *School of Physics and Astronomy, University of Edinburgh, Edinburgh, United Kingdom*
[34] *School of Physics and Astronomy, University of Glasgow, Glasgow, United Kingdom*
[35] *Oliver Lodge Laboratory, University of Liverpool, Liverpool, United Kingdom*
[36] *Imperial College London, London, United Kingdom*
[37] *School of Physics and Astronomy, University of Manchester, Manchester, United Kingdom*
[38] *Department of Physics, University of Oxford, Oxford, United Kingdom*
[39] *Massachusetts Institute of Technology, Cambridge, MA, United States*
[40] *University of Cincinnati, Cincinnati, OH, United States*

[a] *Scuola Normale Superiore, Pisa, Italy*
[b] *Dunnhumby Ltd., Hammersmith, United Kingdom*
[c] *La Salle, Universitat Ramon Llull, Barcelona, Spain*
[d] *Institute of Nuclear Physics, Moscow State University (SINP MSU), Moscow, Russia*

[†] *Author was at institute at time work was performed.*

1 Introduction

The LHCb experiment is a dedicated heavy-flavour physics experiment at the LHC, focused on the reconstruction of particles containing c and b quarks. During Run 1, the LHCb physics programme was extended to electroweak, soft QCD and even heavy-ion physics. This was made possible in large part due to a versatile real-time reconstruction and trigger system, which is responsible for reducing the rate of collisions saved for offline analysis by three orders of magnitude. The trigger used by LHCb in Run 1 [1] executed a simplified two-stage version of the full offline reconstruction. In the first stage, only charged particles with at least ~ 1 GeV/c of transverse momentum (p_{T}) and displaced from the primary vertex (PV) were available; the p_{T} threshold was somewhat lower for muons, which in addition were not required to be displaced. This first stage reconstruction enabled the bunch crossing rate to be reduced efficiently by roughly one order of magnitude. In the following second stage, most charged particles with $p_{\mathrm{T}} \gtrsim 300\,\mathrm{MeV}/c$ were available to classify the bunch crossings (hereafter "events"). Particle-identification information and neutral particles such as photons or π^0 mesons were available on-demand to specific classification algorithms. Although this trigger enabled the majority of the LHCb physics programme, the lack of low-momentum charged particles at the first stage and full particle identification at the second stage limited the performance for c-hadron physics in particular. In addition, resolution differences between the online and offline reconstructions made it difficult to precisely understand absolute trigger efficiencies.

For these reasons, the LHCb trigger system was redesigned during 2013–2015 to perform the full offline event reconstruction. The entire data processing framework was redesigned to enable a single coherent real-time detector alignment and calibration, as well as real-time analyses using information directly from the trigger system. The key objectives of this redesign were twofold: firstly, to enable the full offline reconstruction to run in the trigger, greatly increasing the efficiency with which charm- and strange-hadron decays could be selected; and secondly, to achieve the same quality of alignment and calibration within the trigger as was achieved offline in Run 1, enabling the final signal selection to be performed at the trigger level.

A schematic diagram showing the trigger data flow in Run 2 is depicted in Fig. 1. The LHCb trigger is designed to allow datataking with minimal deadtime at the full LHC bunch crossing rate of 40 MHz. The maximum rate at which all LHCb subdetectors can be read out is imposed by the bandwidth and frequency of the front-end electronics, and corresponds to around 1.1 MHz when running at the designed rate of visible interactions per bunch crossing in LHCb of $\mu = 0.4$. During Run 2 LHCb operated at $\mu = 1.1$ in order to collect a greater integrated luminosity, which limited the actual readout rate to about 1 MHz. A system of field-programmable gate arrays with a fixed latency of 4 μs (the `L0` trigger) determines which events are kept. Information from the electromagnetic calorimeter, hadronic calorimeter, and muon stations is used in separate `L0` trigger lines.

The High Level trigger (HLT) is divided into two stages, HLT1 and HLT2. The first level of the software trigger performs an inclusive selection of events based on one- or two-track signatures, on the presence of muon tracks displaced from the PVs, or on dimuon

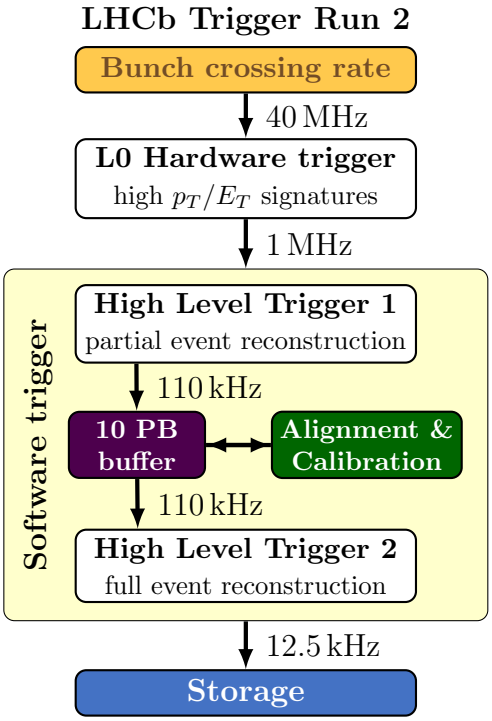

Figure 1: Overview of the LHCb trigger system.

combinations in the event. Events selected by the HLT1 trigger are buffered to disk storage in the online system. This is done for two purposes: events can be processed further during inter-fill periods, and the detector can be calibrated and aligned run-by-run before the HLT2 stage. Once the detector is aligned and calibrated, events are passed to HLT2, where a full event reconstruction is performed. This allows for a wide range of inclusive and exclusive final states to trigger the event and obviates the need for further offline processing.

This paper describes the design and performance of the Run 2 LHCb trigger system, including the real-time reconstruction which runs in the HLT. The software framework enabling real-time analysis ("TURBO") has been described in detail elsewhere. The initial proof-of-concept deployed in 2015 [2] allowed offline-quality signal candidates selected in the trigger to be written to permanent storage. It also allowed physics analysts to use the offline analysis tools when working with these candidates, which was crucial in enabling LHCb to rapidly produce a number of publications proving that real-time analysis was possible without losing precision or introducing additional systematics. Subsequent developments [3] generalized this approach to allow not only the signal candidate but also information about other, related, particles in the event to be saved. These developments also transformed the proof-of-concept implementation into a scalable solution which will now form the basis of LHCb's upgrade computing model [4].

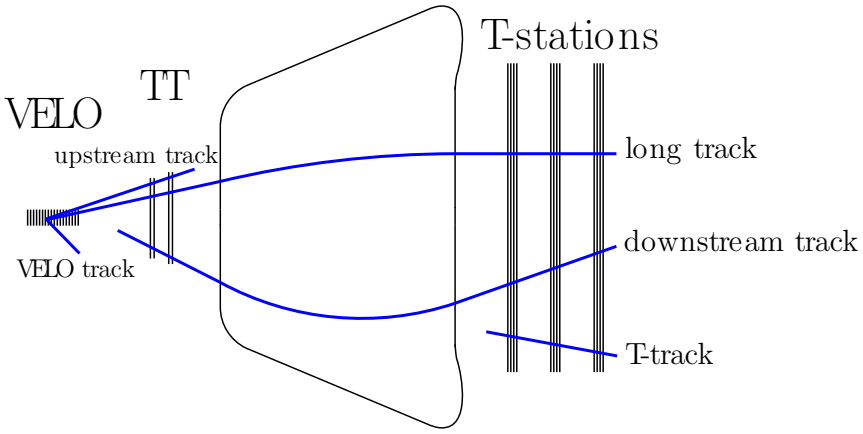

Figure 2: Sketch of the different types of tracks within LHCb.

2 The LHCb detector

The LHCb detector [5,6] is a single-arm forward spectrometer covering the pseudorapidity range $2 < \eta < 5$. The detector coordinate system is such that z is along the beam line and x is the direction in which charged particle trajectories are deflected by the magnetic field. The detector includes a high-precision tracking system consisting of a silicon-strip vertex detector (VELO) surrounding the pp interaction region [7], a large-area silicon-strip detector (TT) located upstream of a dipole magnet with a bending power of about 4 Tm, and three stations of silicon-strip detectors (IT) and straw drift tubes [8] (OT) placed downstream of the magnet. These are collectively referred to as the T-stations. The tracking system provides a measurement of momentum, p, of charged particles with a relative uncertainty that varies from 0.5% at low momentum to 1.0% at 200 GeV/c. A sketch of the various track types relevant in LHCb is shown in Fig. 2.

The minimum distance of a track to a PV, the impact parameter, is measured with a resolution of $(15 + [29\,\text{GeV}/c]/p_{\text{T}})\,\mu\text{m}$. Different types of charged hadrons are distinguished using information from two ring-imaging Cherenkov detectors [9]. Photons, electrons and hadrons are identified by a calorimeter system consisting of scintillating-pad (SPD) and preshower detectors (PS), an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL). Muons are identified by a system composed of alternating layers of iron and multiwire proportional chambers (MUON) [10].

The LHCb detector data taking is divided into fills and runs. A fill is a single period of collisions delimited by the announcement of stable beam conditions and the dumping of the beam by the LHC, and typically lasts around twelve hours. A fill is subdivided into runs, each of which lasts a maximum of one hour. The downtime associated with run changes is negligible compared to other sources of downtime.

Detector simulation has been used in the tuning of most reconstruction and selection algorithms discussed in this paper. In simulated LHCb events, pp collisions are gener-

ated using PYTHIA [11] with a specific LHCb configuration [12]. Decays of hadronic particles are described by EVTGEN [13], in which final-state radiation is generated using PHOTOS [14]. The interaction of the generated particles with the detector, and its response, are implemented using the GEANT4 toolkit [15] as described in Ref. [16].

3 Data acquisition and the LHCb trigger

All trigger systems consist of a set of algorithms that classify events (or parts thereof) as either interesting or uninteresting for further analysis, so that the data rate can be reduced to a manageable level by keeping only interesting events or interesting parts of them. It is conventional to refer to a single trigger classification algorithm as a "line", so that a trigger consists of a set of trigger lines.

3.1 Hardware trigger

The energies deposited in the SPD, PS, ECAL and HCAL are used in the L0-calorimeter system to trigger the selection of events. All detector components are segmented transverse to the beam axis into cells of different size. The decision to trigger an event is based on the transverse energy deposited in clusters of 2×2 cells in the ECAL and HCAL. The transverse energy of a cluster is defined as

$$E_{\mathrm{T}} = \sum_{i=1}^{4} E_i \sin\theta_i \ , \tag{1}$$

where E_i is the energy deposited in cell i and θ_i is the angle between the z-axis and a line from the cell centre to the average pp interaction point (for more details, see Ref. [1]). Additionally, information from the SPD and PS systems is used to distinguish between hadron, photon and electron candidates.

The L0-muon trigger searches for straight-line tracks in the five muon stations. Each muon station is sub-divided into logical pads in the x-y plane. The pad size scales with the distance to the beam line. The track direction is used to estimate the p_{T} of a muon candidate, assuming that the particle originated from the interaction point and received a single kick from the magnetic field. The p_{T} resolution of the L0-muon trigger is about 25 % averaged over the relevant p_{T} range. The trigger decision is based on the two muon candidates with the largest p_{T}: either the largest p_{T} must be above the L0Muon threshold, or the product of the largest and second largest p_{T} values must be above the L0DiMuon threshold. In addition there are special trigger lines that select events with low particle multiplicity to study central exclusive production and inclusive jet trigger lines for QCD measurements.

To reduce the complexity of events and, hence, to enable a faster reconstruction in the subsequent software stage, a requirement is placed on the maximum number of SPD hits in most L0 trigger lines. The L0DiMuon trigger accepts a low rate of events, and therefore, only a loose SPD requirement is applied, while no SPD requirement is applied

Table 1: The L0 thresholds for the different trigger lines used to take the majority of the data for each indicated year. Technical trigger lines and those used for special areas of the physics programme are excluded for brevity. The Hadron, Photon, and Electron trigger lines select events based on the E_{T} of reconstructed ECAL and HCAL clusters. The Muon, Muon High, and Dimon trigger lines select events based on the p_{T} reconstruced MUON stubs, where the Dimuon selection is based on the product of the largest and second largest p_{T} stubs found in the event. As some of the subdetectors also read out hits associated to other bunch crossings, the use of bandwidth is further optimised in most of the L0 lines by rejecting events with a large E_{T} ($> 24\,\mathrm{GeV}$) for the previous bunch crossing [17].

L0 trigger	$E_{\mathrm{T}}/p_{\mathrm{T}}$ threshold			SPD threshold
	2015	2016	2017	
Hadron	> 3.6 GeV	> 3.7 GeV	> 3.46 GeV	< 450
Photon	> 2.7 GeV	> 2.78 GeV	> 2.47 GeV	< 450
Electron	> 2.7 GeV	> 2.4 GeV	> 2.11 GeV	< 450
Muon	> 2.8 GeV	> 1.8 GeV	> 1.35 GeV	< 450
Muon high p_{T}	> 6.0 GeV	> 6.0 GeV	> 6.0 GeV	none
Dimuon	> 1.69 GeV^2	> 2.25 GeV^2	> 1.69 GeV^2	< 900

in the high p_{T} L0Muon trigger used for electroweak production analyses in order to avoid systematic uncertainties associated with the determination of the corresponding efficiency. The thresholds used to take the majority of the data are listed in Table 1 as a function of the year of data taking. Note that while the use of SPD requirements selects simpler and faster-to-reconstruct events, it does not result in a significant loss of absolute signal efficiency compared to a strategy using only E_{T} and p_{T} requirements. This is because the L0 signal-background discrimination deteriorates rapidly with increasing event complexity for all but the dimuon and electroweak trigger lines. Note that the 2017 thresholds are looser than the 2016 thresholds because the maximum number of colliding bunches, and hence, the collision rate of the LHC was significantly lower in 2017, due to difficulties with part of the injection chain. The optimization of the L0 criteria is described in more detail in Sec. 6.

3.2 High level trigger

Events selected by L0 are transferred to the Event Filter Farm (EFF) for further selection. The EFF consists of approximately 1700 nodes, 800 of which were added for Run 2, with 27000 physical cores. The EFF can accommodate ≈ 50000 single-threaded processes using hyper-threading technology.

The HLT is written in the same framework as the software used in the offline reconstruction of events for physics analyses. This allows for offline software to be easily incorporated into the trigger. As detailed later, the increased EFF capacity and improvements in the

software allowed the offline reconstruction to be performed in the HLT in Run 2.

The total disk buffer of the EFF is 10 PB, distributed such that farm nodes with faster processors get a larger portion of the disk buffer. At an average event size of 55 kB passing HLT1, this buffer allows for up to two weeks of consecutive HLT1 data taking before HLT2 has to be executed. Therefore, it is large enough to accommodate both regular running (where, as we will see, the alignment and calibration is completed in a matter of minutes) and to serve as a safety mechanism to delay HLT2 processing in case of problems with the detector or calibration.

Around 40% of the trigger output rate is dedicated to inclusive topological trigger lines, another 40% is dedicated to exclusive c-hadron trigger lines, with the rest divided among dimuon lines, trigger lines for electroweak physics, searches for exotic new particles, and other exclusive trigger lines for specific analyses. There are in total around 20 HLT1 and 500 HLT2 trigger lines.

3.3 Real-time alignment and calibration

The computing power available in the Run 2 EFF allows for automated alignment and calibration tasks, providing offline quality information to the trigger reconstruction and selections, as described in Ref. [18, 19]. A more detailed description of this real-time alignment and calibration procedure will be the topic of a separate publication.

Dedicated samples selected by HLT1 are used to align and calibrate the detector in real time. The alignment and calibrations are performed at regular intervals, and the resulting alignment and calibration constants are updated only if they differ significantly from the current values.

The major detector alignment and calibration tasks consist of:

- the VELO alignment, followed by the alignment of the tracking stations;
- the MUON alignment;
- alignment of the rotations around various local axes in both RICH detectors of the primary and secondary mirrors;
- global time calibration of the OT;
- RICH gas refractive-index calibration;
- RICH Hybrid Photon Detectors calibration;
- ECAL LED (relative) and π^0 (absolute) calibrations.

Each of these tasks has a dedicated HLT1 trigger line which supplies it with the types of events required. When the required sample sizes have been collected, the selected events are saved to the disk buffer of the EFF, and calibration and alignment tasks are performed in parallel within the EFF. A schematic view of the alignment and calibration procedure is shown in Fig. 3, together with the time when the tasks are launched and the typical time taken to complete them.

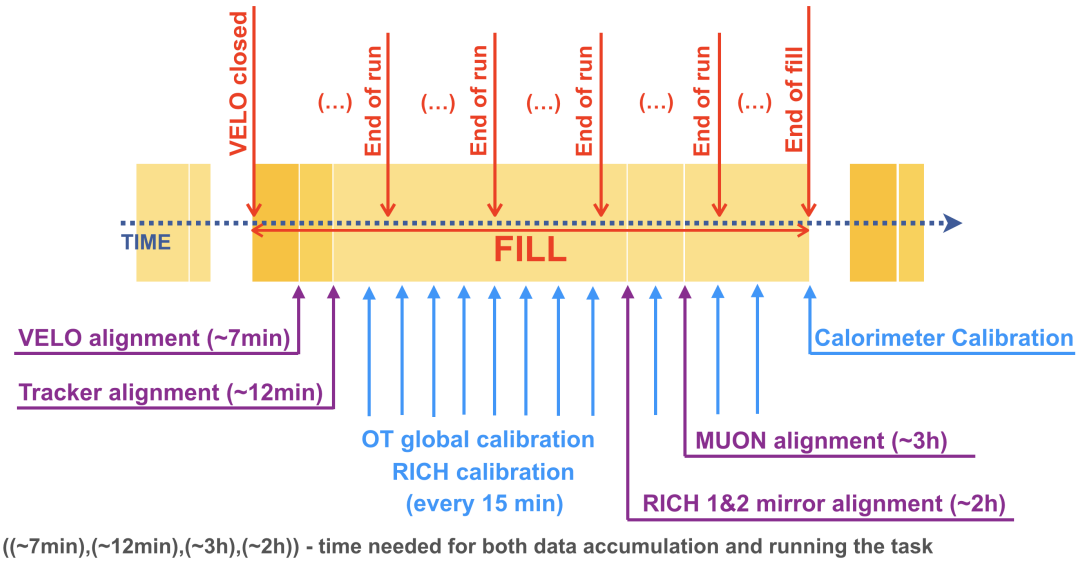

Figure 3: Schematic view of the real-time alignment and calibration procedure starting at the beginning of each fill, as used for 2018 data taking.

4 HLT1 partial event reconstruction

HLT1 reconstructs the trajectories of charged particles traversing the full LHCb tracking system, called long tracks, with a p_{T} larger than 500 MeV/c. In addition, a precise reconstruction of the PV is performed. The details of both steps are presented in Sec. 4.1.

Tight timing constraints in HLT1 mean that most particle-identification algorithms cannot be executed. The exception is muon identification, which due to a clean signature produced by muons in the LHCb detector can be performed already in HLT1, as described in Sec. 4.2. As discussed in Sec. 6.7.3, a subset of specially selected HLT1 events serves as input to the alignment and calibrations tasks.

4.1 Track and vertex reconstruction in HLT1

The sequence of HLT1 algorithms which reconstruct vertices and long tracks is shown in Fig. 4. The pattern recognition deployed in HLT1 consists of three main steps: reconstructing the VELO tracks, extrapolating them to the TT stations to form upstream tracks, and finally extending them further to the T stations to produce long tracks. Next, the long tracks are fitted using a Kalman Filtering and the fake trajectories are rejected. The set of fitted VELO tracks is re-used to determine the positions of the PVs.

4.1.1 Pattern recognition of high-momentum tracks

The hits in the VELO are combined to form straight lines loosely pointing towards the beam line [20]. Next, at least three hits in the TT are required in a small region around a

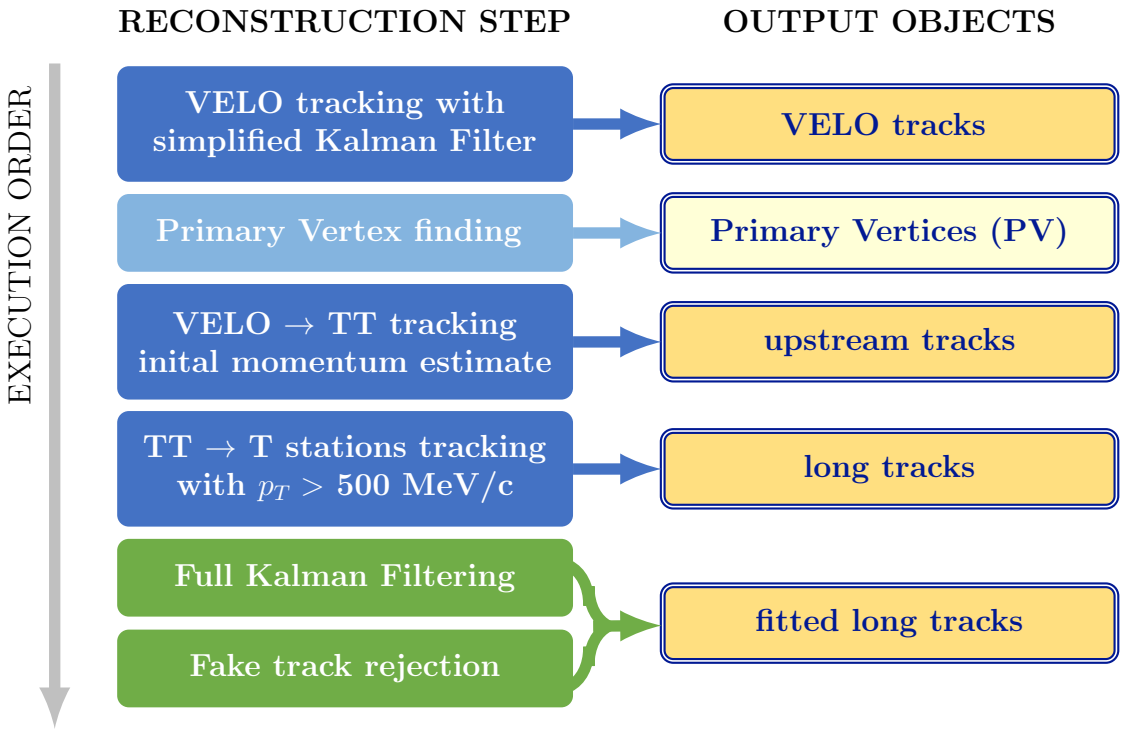

Figure 4: Sketch of the HLT1 track and vertex reconstruction.

straight-line extrapolation from the VELO [21] to form so-called upstream tracks. The TT is located in the fringe field of the LHCb dipole magnet, which allows the momentum to be determined with a relative resolution of about 20%. This momentum estimate is used to reject low p_{T} tracks. Matching long tracks with TT hits additionally reduces the number of fake VELO tracks. Due to the limited acceptance of the TT, VELO tracks pointing to the region around the beampipe do not deposit charge in the TT; therefore, they are passed on without requiring TT hits.

The search window in the IT and OT is defined by the maximum possible deflection of charged particles with p_{T} larger than 500 MeV/c. The search is also restricted to one side of the straight-line extrapolation by the charge estimate of the upstream track. The use of the charge estimate is new in Run 2 and enabled the p_{T} threshold to be lowered from 1.2 GeV/c to 500 MeV/c with respect to Run 1. For a given slope and position upstream of the magnet and a single hit in the tracking detectors downstream of the magnet, IT and OT, the momentum is fixed and hits are projected along this trajectory into a common plane. A search is made for clusters of hits in this plane which are then used to define the final long track [22]. In 2016 two artificial neural nets were implemented to increase the purity and the efficiency of the track candidates in the pattern recognition [23].

4.1.2 Track fitting and fake-track rejection

Subsequently, all tracks are fitted with a Kalman filter to obtain the optimal parameter estimate. The settings of the online and offline reconstruction are harmonised in Run 2 to obtain identical results for a given track. Previously, the online Kalman filter performend only a single iteration which did not allow the ambiguities from the drift-time measurement of OT hits on a track to be resolved. In Run 2 the online fit runs until convergence or maximally 10 iterations. Furthermore the possibility to remove up to two outliers has been added in the online reconstruction. The offline reconstruction is changed to use clusters which are faster to decode but have less information, and to employ a Kalman filter that utilizes a simplified geometry description of the LHCb detector. This significantly speeds up the calculation of material corrections in the filtering procedure. For Run 2 the calculation of material corrections due to multiple scattering has been improved. The new description is additive for many small scatterers resulting in more standard normal pull distributions. The changes made to the offline Kalman filter enable running the same algorithm in both the HLT and offline. These changes neither affect the impact parameter resolution nor the momentum resolution as shown in Sec. 5.1.2.

Since 2016 the fake track rejection, described in details in Sec. 5.1, has been used in HLT1 reducing the rate of events passing this stage by 4%.

4.1.3 Primary vertex reconstruction

Many LHCb analyses require a precise knowledge of the PV position and this information is used early in the selection of displaced particles. The full set of VELO tracks is available in HLT1. Therefore, the PVs in Run 2 are reconstructed using VELO tracks only, neglecting the additional momentum information on long tracks which is only available later. This does not result in a degradation in resolution compared to using a mixture of VELO and long tracks. Furthermore, this approach produces a consistent PV position from the beginning to the end of the analysis chain which reduces systematic effects.

As there is no magnetic field in the VELO, the Kalman filter for VELO tracks uses a linear propagation, allowing for a single scattering at each detector plane, tuned using simulation. This simplification results in no loss of precision compared to a more detailed material description, but significantly reduces the amount of time spent in the filtering phase, as no expensive computations are necessary. A byproduct of this simpler track fit is that the PV covariance matrix is more accurate than that used offline in Run 1, with pull distributions more compatible with unit widths in all three dimensions.

The PV resolution is obtained by randomly splitting the input VELO tracks into two subsets. The PV algorithm is executed independently on each subset, and the PVs found in each subset are matched based on the distance between them. The width of the distribution of the difference of matched PV positions in a given dimension, corrected by a factor of $\sqrt{2}$, gives the corresponding PV position resolution. The resolution of the PV reconstruction for Run 2 is shown in Fig. 5 compared to the Run 1 (2012) offline reconstruction algorithm. The new algorithm performs equally well for the x (y) coordinate, while with respect to Run 1 the resolution on the z coordinate is improved by about 10%.

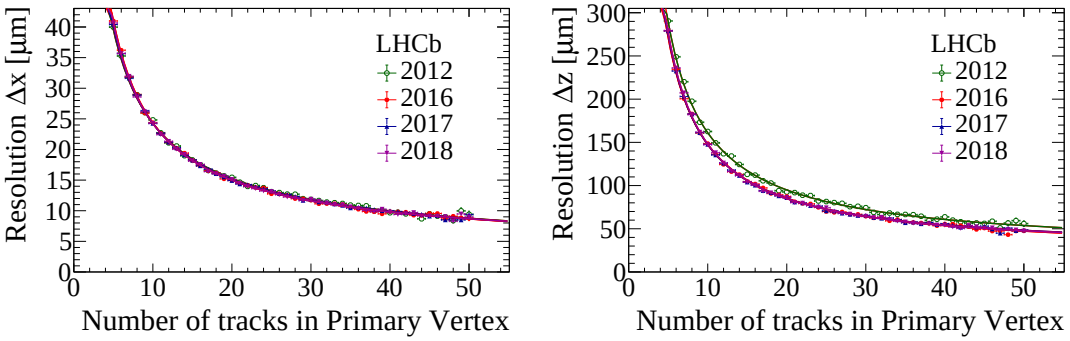

Figure 5: The PV x (left) and z (right) resolution as a function of the number of tracks in the PV for the Run 1 offline and Run 2 (used both offline and online) PV reconstruction algorithms.

Additionally, the parameters of the PV reconstruction have been retuned to give a higher efficiency and smaller fake rate [24]. The resulting improvement in efficiency of reconstructing PVs is 0.5% for PVs associated with the production of a b quark pair, 1.3% for those associated with the production of a c quark pair, and 6.6% for light quarks production. Simultaneously, the fraction of fake PVs, for example due to material interactions or the decay vertices of long-lived particles, is reduced from 3.5% to 1%.

4.2 Muon identification

The muon identification starts with fully fitted tracks. Hits in the MUON stations are searched for in momentum-dependent regions of interest around the track extrapolation. Tracks with $p < 3\,\text{GeV}/c$ cannot be identified as muons, as they would not be able to reach the MUON stations. Below a momentum of $6\,\text{GeV}/c$ the muon identification algorithm requires hits in the first two stations after the calorimeters. Between 6 and $10\,\text{GeV}/c$ an additional hit is required in one of the last two stations. Above 10 GeV, hits are required in all the four MUON stations. This same algorithm is used in HLT1, HLT2 and offline.

In HLT1, the track reconstruction is only performed for tracks with p_{T} above 500 MeV/c. For particles with lower p_{T}, a complementary muon-identification algorithm has been devised, which is more similar to the HLT1 muon identification performed in Run 1. Upstream track segments are extrapolated directly to the MUON stations, where hits are searched for around the track extrapolation. The regions of interest used in this search are larger than those used for otherwise-reconstructed long tracks. If hits are found in the muon system, the VELO-TT segment is extrapolated through the magnetic field using the momentum estimate and matched to hits not already used in the HLT1 long-track reconstruction. This procedure extends the muon-identification capabilities down to a p_{T} of $80\,\text{MeV}/c$ for a small additional resource cost, significantly improving the performance for lower-momentum muons which are important in several measurements [25].

The muon identification code has been reoptimized for Run 2, gaining significant efficiency, in particular at small p_{T}. This is demonstrated using LHCb simulation in Fig. 6. The performance of the muon identification in HLT1 is shown in Fig. 7 (left) as determined

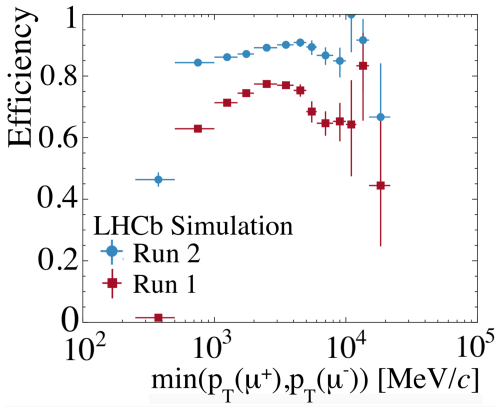

Figure 6: HLT1 dimuon efficiency as a function of the minimum p_{T} of the two muons. A large gain, especially at low p_{T}, can be seen from the comparison of the Run 1 and Run 2 algorithms.

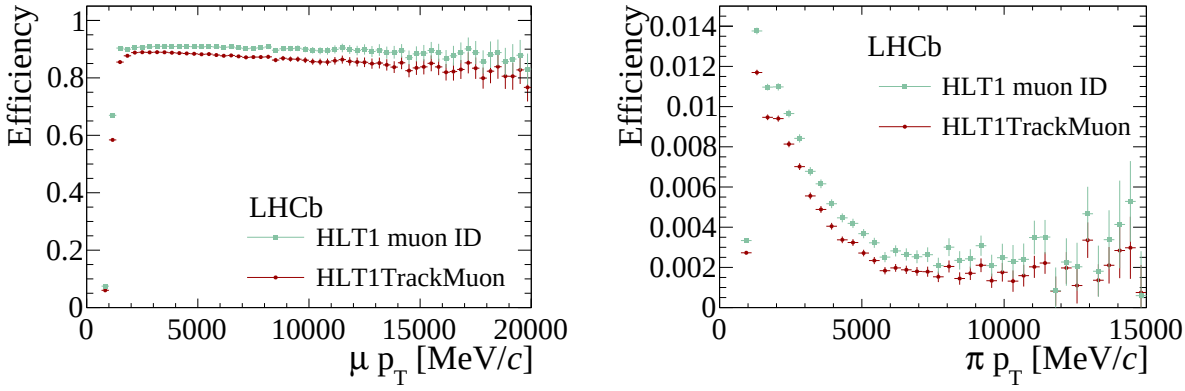

Figure 7: HLT1 muon identification efficiency for (left) muons from $J/\psi \to \mu^+\mu^-$ decays and (right) pions from $D^0 \to K^-\pi^+$ decays. Green circles show only the identification efficiency (HLT1 Muon ID) while red squares show the efficiency of the additional trigger line (named HLT1TrackMuon) requirements (see text).

from unbiased $J/\psi \to \mu^+\mu^-$ decays using the tag-and-probe method. This performance is obtained by studying the efficiency of the single-muon HLT1 trigger, which includes requirements on the displacement of the muon and on the minimum momentum (6 GeV/c) and p_{T} (1.1 GeV/c). Analogously in Fig. 7 (right) the misidentification efficiency with the same criteria is shown for pions from $D^0 \to K^-\pi^+$ decays. The muon identification efficiency of the single-muon HLT1 trigger is slightly reduced by the displacement and (transverse) momentum requirements, at the benefit of a lower misidentification probability.

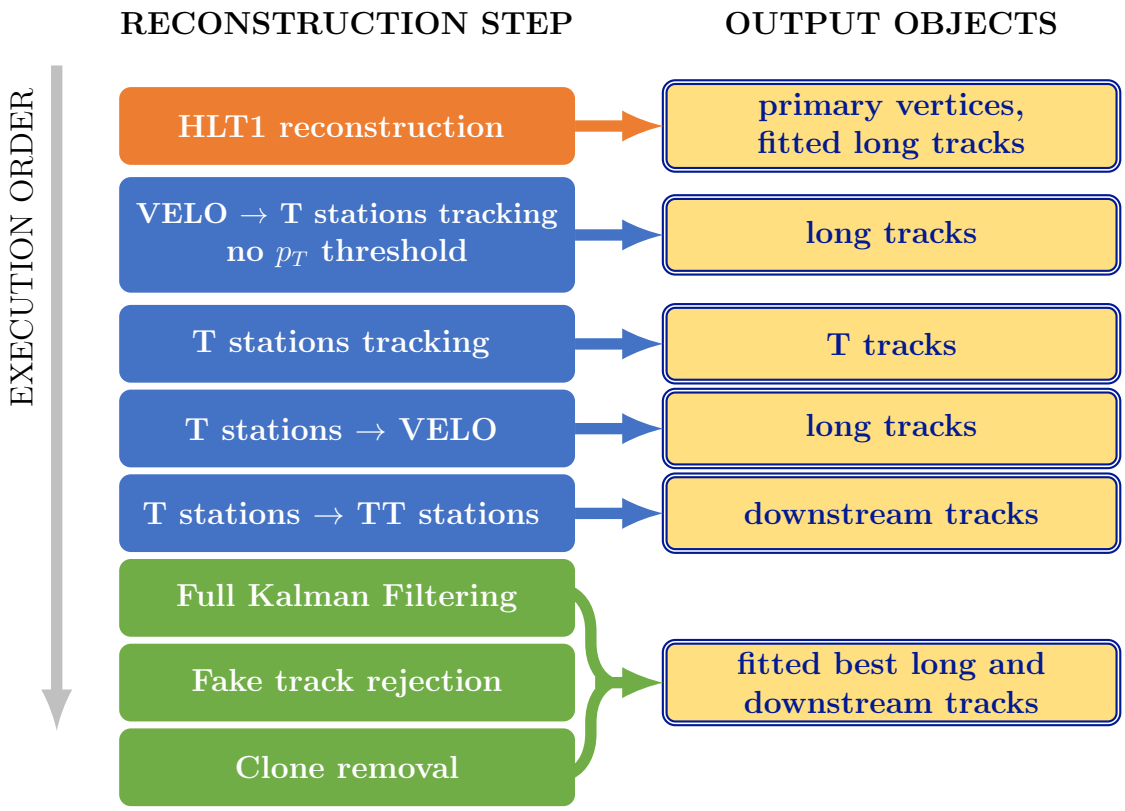

Figure 8: Sketch of the HLT2 track and vertex reconstruction sequence.

5 HLT2 full event reconstruction

The HLT2 full event reconstruction consists of three major steps: the track reconstruction of charged particles, the reconstruction of neutral particles, and particle identification (PID).The HLT2 track reconstruction exploits the full information from the tracking sub-detectors, performing additional steps of the pattern recognition which are not possible in HLT1 due to strict time constraints. As a result high-quality long and downstream tracks are found with the most precise momentum estimate achievable. Similarly, the most precise neutral cluster reconstruction algorithms are executed in the HLT2 reconstruction. Finally, in addition to the muon identification available in HLT1, HLT2 exploits the full particle identification from the RICH detectors and calorimeter system. The code of all reconstruction algorithms has been optimized for Run 2 to better exploit the capabilities of modern CPUs. Together with the algorithmic changes described in the following sections, this results in a two times faster execution time while delivering the same or in several cases better physics performance than that achieved offline in Run 1.

5.1 The track reconstruction of charged particles

A sketch of the track and vertex reconstruction sequence in HLT2 is shown in Fig. 8. The goal is to reconstruct all tracks without a minimal p_{T} requirement. This is particularly important for the study of the decays of lighter particles, such as charmed or strange hadrons, whose abundance means that only some of the fully reconstructed and exclusively selected final states fit into the available trigger bandwidth. Often, not all of the decay products of a charm- or strange-hadron decay pass the $500\,\mathrm{MeV}/c$ p_{T} requirement of HLT1, particularly for decays into three or more final-state particles. Therefore, to efficiently trigger these decays, it is necessary to also reconstruct the lower-momentum tracks within the LHCb acceptance.

In a first step, the track reconstruction of HLT1 is repeated. A second step is then used to reconstruct the lower-momentum tracks which had not been found in HLT1 due to the kinematic thresholds in the reconstruction. Those VELO tracks and T-station clusters used to reconstruct long tracks with a good fit quality in the first step are disregarded for this second step. A similar procedure as in HLT1 is employed: the remaining VELO tracks are extrapolated through the magnet to the T-stations using the same algorithm, where the search window is now defined by the maximal possible deflection of a particle with p_{T} larger than 80 MeV/c. No TT hits are required for the second step to avoid the loss of track efficiency induced by acceptance gaps in the TT. The new Run 2 track finding optimization results in 27% fewer fake tracks and a reconstruction efficiency gain of 0.5% for long tracks. In addition, a standalone search for tracks in the T stations is performed [26], and these standalone tracks are then combined with VELO tracks to form long tracks [27, 28]. The redundancy of the two long-track algorithms increases the efficiency by a few percent.

Tracks produced in the decays of long-lived particles like Λ or K^0_{S} that decay outside the VELO are reconstructed using T-station segments that are extrapolated backwards through the magnetic field and combined with hits in the TT. For Run 2, a new algorithm was used to reconstruct these tracks [29]. It uses two multivariate classifiers, one to reject fakes, and another to select the final set of hits in the TT in case several sets are compatible with the same T-station segment. In combination with other improvements, this results in a higher efficiency and a lower fake rate compared to the corresponding Run 1 algorithm.

The next step in the reconstruction chain is the rejection of fake tracks. These fakes result from random combinations of hits or a mismatch of track segments upstream and downstream of the magnet. They are reduced using two techniques. First, all tracks are fitted using the same Kalman filter. In Run 1, the only selection to reject fake tracks was based on a reduced χ^2. For Run 2, the upper limit on this χ^2 selection was increased to allow for a better overall track reconstruction efficiency. To offset the corresponding increase in the number of fake tracks, a neural network was trained using the TMVA [30,31] package to efficiently remove these tracks [32]. Its input variables are the overall χ^2 of the Kalman filter, the χ^2 values of the fits for the different track segments, the numbers of hits in the different tracking detectors, and the p_{T} of the track.

A previous version of the neural network was widely and successfully used in Run

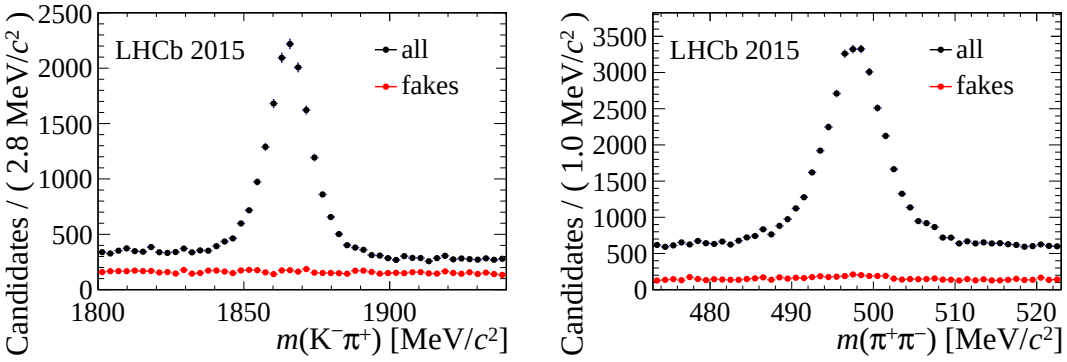

Figure 9: Performance of the fake-track classifier on (left) $D \to K^-\pi^+$ and (right) $K^0_S \to \pi^-\pi^+$ decays. For these plots, the clones have been removed.

1 to discard fake tracks at the analysis level. The use of a different set of variables, whose computations are less time consuming, together with optimization of the code made it possible to deploy this classifier in the Run 2 trigger without any significant impact on the execution time. The evaluation uses only about 0.2% of the total CPU budget. Furthermore, the better performance of this fake track rejection in both stages of the HLT leads to 16% less CPU consumption in the entire software trigger. The neural network was trained on simulated tracks. The working point was chosen such that it rejects 60% of all fake tracks, while maintaining an efficiency of about 99%.

The performance of the fake track removal was validated on first collision data in 2015 to ensure a uniform response over a large area of the phase space. As an example, the performance for $D^0 \to K^-\pi^+$ and $K^0_S \to \pi^-\pi$ decays is shown in Fig. 9.

After the removal of fake tracks, the remaining tracks are filtered to remove so-called clones. Clones can be created inside a single pattern-recognition algorithm or, more commonly, originate from the redundancy in the pattern-recognition algorithms. Two tracks are defined as clones of each other if they share enough hits in each subdetector. Only the subdetectors where both tracks have hits are considered. The track with more hits in total is kept and the other is discarded. This final list of tracks is subsequently used to select events as discussed in Sec. 6.

5.1.1 Tracking efficiency

The track reconstruction efficiency is determined using a tag-and-probe method on $J/\psi \to \mu^+\mu^-$ decays that originate from the decays of b-hadrons [33]. One muon is reconstructed using the full reconstruction, while the other muon is reconstructed using only specific subdetectors, making it possible to probe the others. For Run 2, the track reconstruction efficiency is determined in HLT2 using the data collected by specific trigger lines, see Sec. 6.8.3. The performance compared to Run 1 is shown in Fig. 10. Given that the OT has a readout window which is larger than 25 ns and therefore is prone to spillover effects when reducing the bunch spacing from 50 ns to 25 ns, a small reduction in the track

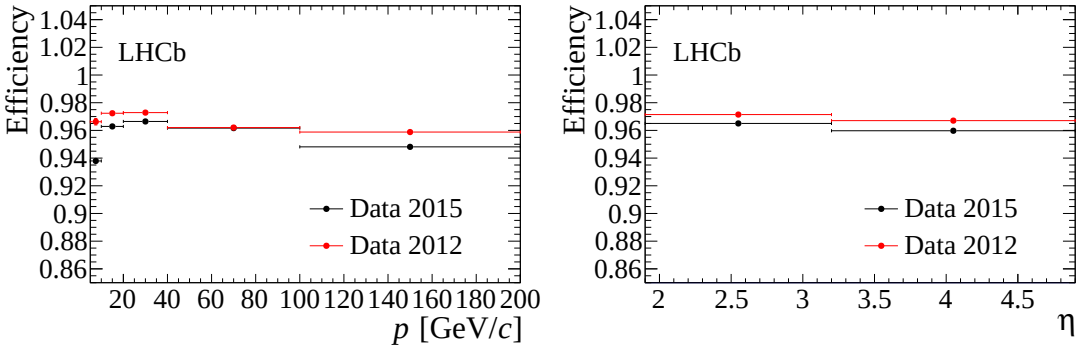

Figure 10: Comparison of the track reconstruction efficiency in 2015 and 2012 data as a function of the momentum (left) and pseudorapidity (right).

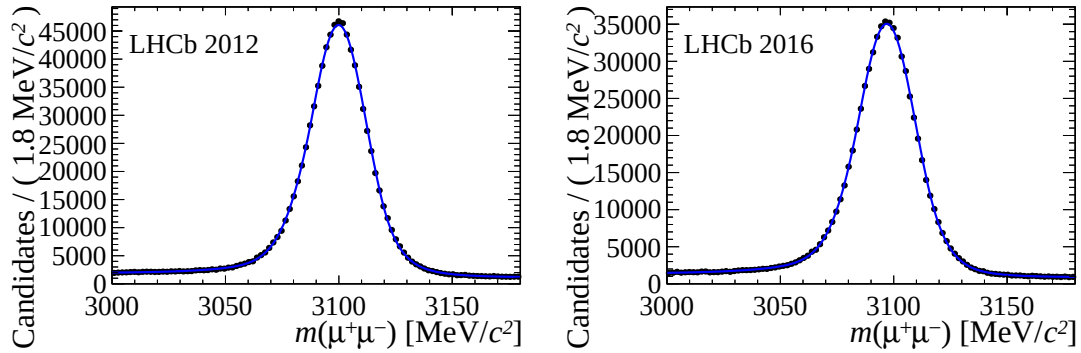

Figure 11: Comparison of the invariant mass distributions for a subset of the 2012 (left) and 2016 (right) data set, using $J/\psi \to \mu^+\mu^-$ decays, with the J/ψ originating from a b-hadron.

reconstruction efficiency is observed in 2015.

5.1.2 Invariant mass resolution

The invariant mass resolution is determined on a sample of $J/\psi \to \mu^+\mu^-$ decays, where the J/ψ originates from a b-hadron decay. The dimuon invariant mass distribution is modelled using a double Crystal Ball function [34], where the weighted mean of the standard deviations of the two Gaussian components is used to estimate the resolution. The distributions for subsamples of the 2012 and 2016 data can be seen in Fig. 11, the resolutions are 12.4 MeV/c^2 for the 2012 data sample and 12.7 MeV/c^2 for the 2016 data sample. The difference comes from a slightly higher-momentum spectrum in 2016, due to the larger beam energy in Run 2, and a small degradation in the performance due to the use of a simplified description of the detector geometry throughout Run 2.

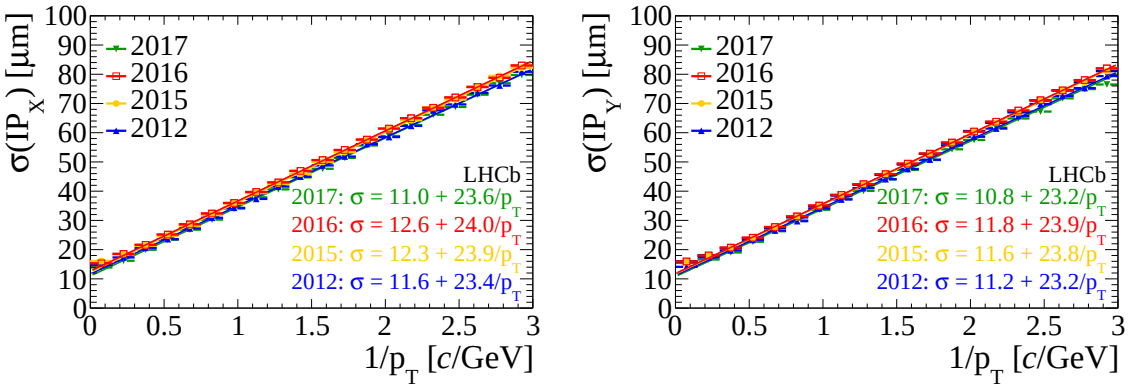

Figure 12: Resolution of the x (left) and y (right) components of the impact parameter comparing the 2012 (blue), 2015 (orange), 2016 (red) and 2017 (green) data-taking periods. The resolution as a function of p_{T} is given in the bottom right corner.

5.1.3 Impact parameter and decay time resolutions

The impact parameter and decay time resolutions are extracted with data-driven methods which are described in more detail elsewhere [6]. The impact parameter is defined as the distance between a particle trajectory and a given PV. It is one of the main discriminants between particles produced directly in the primary interaction and particles originating from the decays of long-lived hadrons. The impact parameter resolution as a function of $1/p_{\mathrm{T}}$ is shown in Fig 12. Only events with one reconstructed PV are used, and the PV fit is rerun excluding each track in turn. The resulting PV is required to have at least 25 tracks to minimise the contribution from the PV resolution. Multiple scattering induces a linear dependence on $1/p_{\mathrm{T}}$. For high p_{T} particles, the impact parameter resolution is roughly 12 μm in both the x and y directions. The observed improvement of about 1 μm in 2017 data taking is due to the use of an updated VELO error parametrisation.

The decay time of a particle is determined from the distance between the PV and the secondary decay vertex. An excellent decay time resolution is a key ingredient of time-dependent mixing and CP violation measurements. The resolution is determined from J/ψ decays combined with two random tracks which mimic $B_s^0 \to J/\psi\,\phi$ decays. In the absence of any impact parameter requirements these combinations come mainly from prompt particles and, therefore, the expected decay time is zero. The width of the distribution is thus a measure of the decay time resolution. A comparison of the decay time resolution as a function of momentum for Run 1, 2015, and 2016 data taking is shown in Fig. 13. For Run 2 the average resolution is about 45 fs for a 4-track vertex.

5.2 Muon reconstruction

As mentioned in Sec. 4.2, the same muon identification algorithm is used in HLT2 and HLT1, apart from the fact that the HLT2 algorithm takes as its input the full set of fitted tracks available after the HLT2 reconstruction.

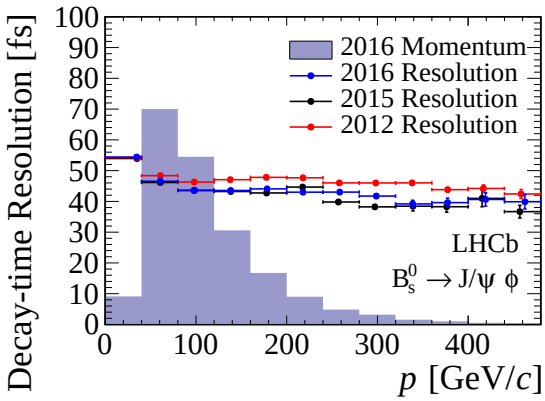

Figure 13: Decay time resolution for $B_s^0 \to J/\psi\,\phi$ decays (in their rest frame) as a function of momentum. The filled histogram shows the distribution of B_s^0 meson momenta, to give an idea of the relative importance of the different resolution bins for the analysis sensitivity.

5.3 RICH reconstruction

The identification of different particle species is crucial across LHCb's physics programme. The RICH detectors provide the main discrimination between deuterons, kaons, pions, and protons. Cherenkov light is emitted in a cone around the flight direction of a charged particle, where the cone width depends on the velocity of the particle. The photon yields, expected Cherenkov angles, and estimates of the per-track Cherenkov angle resolution are computed under each of the deuteron, proton, kaon, pion, muon and electron mass hypotheses. The RICH reconstruction considers simultaneously all reconstructed tracks and all Cherenkov photons in RICH1 and RICH2 in each event. The reconstruction algorithm provides a likelihood for each mass hypothesis. As the RICH reconstruction consumes significant computing power, it could not be run for every track in the Run 1 real-time reconstruction. Improvements in the HLT and in the RICH reconstruction itself made it possible, however, to run the full algorithm in the Run 2 HLT2. The performance of the RICH particle identification is shown in Fig. 14 for the 2012 and 2016 data. A small improvement is obtained in Run 2 for particles below 15 GeV/c of momentum.

5.4 Calorimeter reconstruction

The reconstruction of electromagnetic particles (photons, electrons, and π^0 mesons) is performed by the calorimeters. A cellular automaton algorithm is used to build clusters from the energy deposits in the different calorimeter subsystems, which are combined to determine the total energy of each particle [35]. Neutral particles are then identified according to their isolation with respect to the reconstructed tracks. Electron identification is also provided by combining information from the isolation of clusters in the ECAL, the presence of clusters in the PS, the energy deposited in the HCAL, and the position of possible Bremsstrahlung photons.

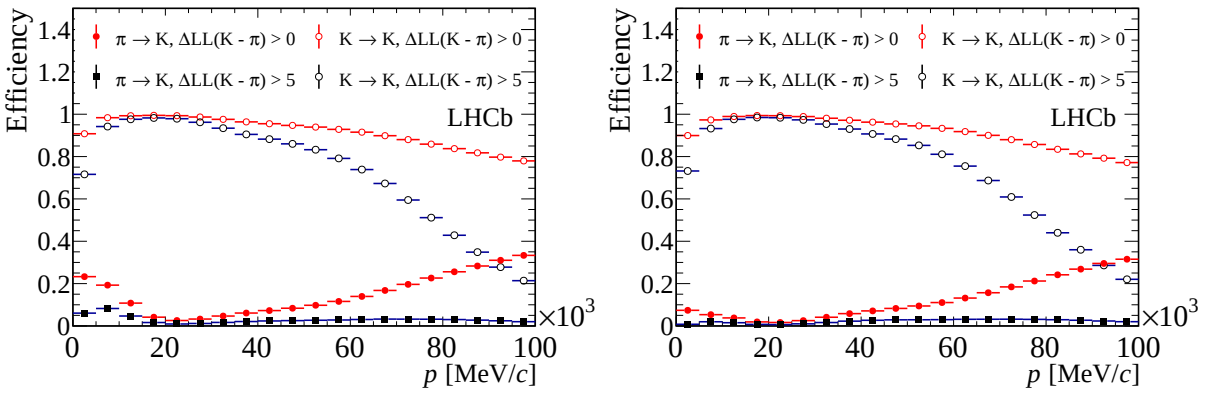

Figure 14: Efficiency and fake rate of the RICH identification for the 2012 (left) and the 2016 (right) data.

High-E_{T} π^0 mesons and photons are indistinguishable at the trigger level, as they both appear as a single cluster, while low-E_{T} π^0 mesons are built by combining resolved pairs of well-separated photons. The neutral-cluster reconstruction algorithm run in HLT2 is the same as that run offline.

The identification of these clusters as either neutral objects or electrons uses information from both the PS/SPD detectors, and a matching between reconstructed tracks and calorimeter clusters. Early in Run 2 this online identification was not identical to the offline version because the HLT did not reconstruct T-tracks (see Fig. 2), since these are not directly used by physics analyses. They are, however, relevant for neutral-particle identification. This misalignment was gradually reduced as Run 2 progressed, first by adding the reconstruction of T-tracks and then by subsequently applying a Kalman filter to them to align the algorithm to the offline reconstruction sequence.

A fully automated ECAL calibration was introduced in 2018. The automatic LED calibration is performed for fills longer than 3.5 hours as indicated in Fig. 3, while the absolute π^0 calibration is processed once per month when sufficient data (amounting to 300M events) is collected. The performance of the calorimeter reconstruction is shown in Fig. 15 using $B^0 \to (K^+\pi^-)\gamma$ decays. The invariant mass resolution has been improved with respect to Run 1 from about $91\,\mathrm{MeV}/c^2$ to $87\,\mathrm{MeV}/c^2$.

6 Trigger performance

The LHCb trigger performance is optimized around two key metrics: the `L0` kinematic and occupancy thresholds for each of the main trigger lines (muon, dimuon, electron, photon, and hadron); and the optimization of the HLT timing budget, which defines the maximum allowed HLT1 output rate. An automated procedure is used to divide the `L0` bandwidth among a set of representative signal channels. It has been significantly improved with respect to Run 1 and is described here. The procedure for determining the

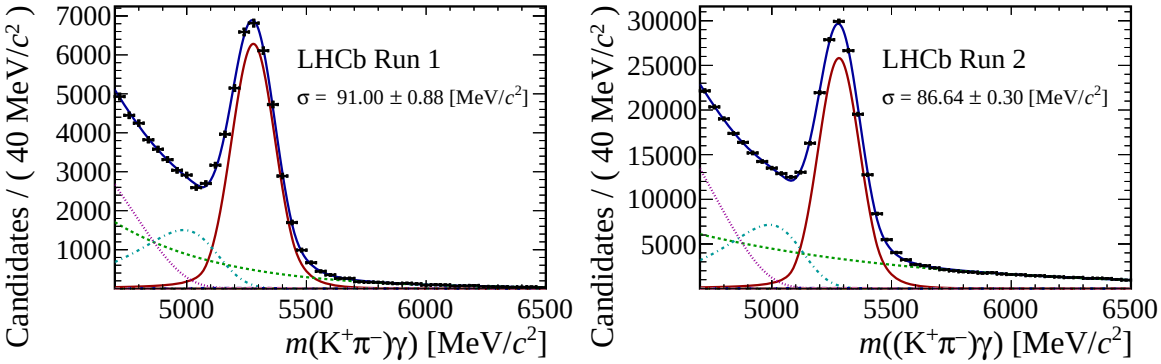

Figure 15: Invariant mass of $B^0 \to (K^+\pi^-)\gamma$ candidates in Run 1 (left) and Run 2 (right). The fit model includes the (red) signal component, (dashed green) combinatorial background, (dot-dashed turqoise) misidentified physics backgrounds (e.g. $B_s^0 \to \phi\gamma$ where a kaon is misidentified as a pion) and (dotted magenta) partially reconstructed physics backgrounds.

HLT processing budget is also described, and the L0 and HLT performance is evaluated using a tag-and-probe approach on Run 2 data.

6.1 L0 bandwidth division

The relative simplicity of the information available for the L0 trigger decision means that the trigger lines listed in Table 1 cover the majority of the LHCb physics programme. Out of these, the high-p_{T} muon trigger consumes a relatively negligible rate and is insensitive to the running conditions. The remaining five trigger lines: hadron, muon, electron, photon and dimuon, must have their p_{T} and E_{T} thresholds tuned to maximize signal efficiency under different LHC conditions. In particular, during the luminosity ramp-up of the LHC, signal efficiencies can be improved by reducing the thresholds to maintain an L0 output rate close to the maximal readout rate of 1 MHz.

Once the LHC reaches its nominal number of colliding bunches in a given year, determining the optimal division of rate between the L0 channels is important for achieving the physics goals of the experiment. This so-called "bandwidth division" is performed using a genetic algorithm to minimise the following pseudo-χ^2 for a broad range of simulated signal samples that are representative of the LHCb physics programme:

$$\chi^2(r) = \sum_i^N w_i \times \left(1 - \frac{\varepsilon(r)_i}{\varepsilon(r)_i^{\max}}\right)^2 . \tag{2}$$

The sum is over the N signal samples, $\varepsilon(r)_i$ is the efficiency including detector dead time of the i^{th} data set, and $\varepsilon^{\max}$ is the efficiency including detector dead time when all of the bandwidth is allocated to this data set. The ratio of dead-time-corrected efficiencies is designed to ensure that inefficient signal samples contribute more to the χ^2, *i.e.* the algorithm prioritizes improving the efficiency of signal samples which start with a low

absolute efficiency over making identical absolute efficiency improvements for signals with high efficiencies to begin with. The weight assigned to each data set, w_i, is predetermined by the LHCb collaboration and is designed to grant more bandwidth to higher-priority physics channels.

The dead-time correction to the signal efficiency acts as a rate limiter and is dependent upon the filling scheme:

$$\varepsilon(r) = \epsilon \times \left[1 - \delta^{\text{phys}}(r)\right] \times \left[1 - \delta^{\text{tech}}(r)\right]. \tag{3}$$

Here ϵ is the overall L0 signal efficiency and r is the retention of collisions collected using random trigger lines (henceforth "nobias"). The physics dead time $\delta^{\text{phys}}(r)$ is zero below $r_{\text{limit}} = 1.1$ MHz, which is the maximum HLT1 throughput, and r/r_{limit} above this. The technical dead time, $\delta^{\text{tech}}(r)$ is determined from a model trained on a filling-scheme dependent emulation of the detector readout dead time.

The results of the L0 bandwidth optimization are shown in Fig. 16 for the 2016 and 2017 data-taking conditions. The different optimal points are mostly connected to the different LHC running conditions in the two years, in particular to problems in 2017 which limited the maximum possible number of bunches in the LHC and hence led to lower trigger thresholds.

6.2 Measuring the HLT processing speed

Trigger configurations are tested for processing speed and memory usage on 13 TeV nobias collected at the same average number of visible interactions per bunch crossing as in regular data taking. As the L0 trigger conditions affect the complexity of events processed by the HLT, these tests are repeated for each L0 configuration. Nobias events passing the L0 configuration are processed by a dedicated "average" EFF node loaded with the same number of total tasks as in the online data-taking configuration. The timing is measured separately for HLT1 running on events passing L0, and HLT2 running on events passing both L0 and HLT1. Each HLT1 and HLT2 task processes an independent sample of around 10,000 events during this test, with the number of events chosen to balance robustness and turnaround speed. The processing speed of each of the individual HLT1 and HLT2 tasks is then averaged in order to remove fluctuations due to the limited number of test events, and these values are compared to the available budget. In addition, the memory usage is plotted as a function of the event number to verify that there are no memory leaks.

These tests give confidence that the HLT is running within its budget, and they are particularly important for spotting problems after any major changes are made to a stable configuration. They do not, however, give a perfect reflection of the performance expected on the full farm. In particular, the effect of calibration and alignment tasks which run in parallel with the HLT1 and HLT2 tasks is neglected, as are the I/O issues associated with sending events from L0 to the HLT1 tasks, the buffering of HLT1 events and the action of reading them back into HLT2 tasks, and the overhead from sending the events accepted by HLT2 to the offline storage. For this reason it makes little sense to quote detailed performance numbers for the HLT from these tests.

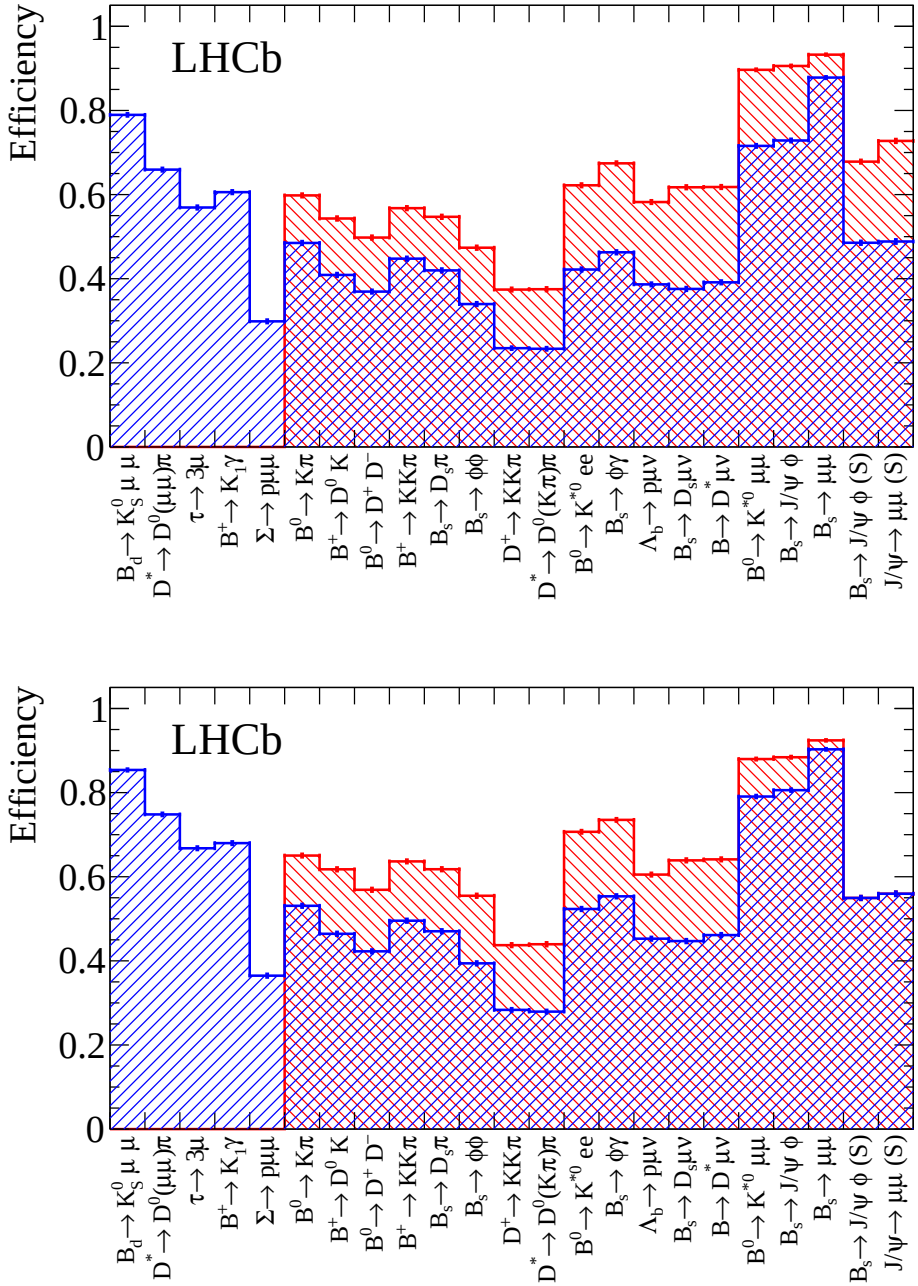

Figure 16: Efficiencies per signal mode for (top) 2016 and (bottom) 2017 data-taking periods measured in simulation. Red (left-slanted) hatched plots are when the entire L0 bandwidth is granted to this signal mode, whereas blue (right-slanted) hatched plots are following the bandwidth division. Signals which appear only in blue are used for performance validation and are not part of the optimization itself. Channels followed by "(S)" are selected in a kinematic and geometric volume which is particularly important for spectroscopy studies.

6.3 Optimization of the HLT timing and disk buffers

The timing budget of the HLT is defined as the average time available for each HLT task to process an event, when the processing farm is fully loaded.[1] With a traditional single-stage HLT, as was the case in Run 1, the timing budget is easily determined because the events must always be processed as they arrive during the collider runtime. Therefore, it is simply the number of HLT tasks divided by the input event rate, which amounted to about 50 ms for a farm with around 50000 logical cores as was available in 2015. The calculation is more complicated for the two-stage HLT used in Run 2. Since the second stage is deferred and can occur during LHC downtime there are two timing budgets, one for the HLT1 stage and one for the HLT2 stage. These budgets depend on the assumptions made about the length and distribution of the LHC runtime and downtime periods.

The LHC downtime is not uniformly distributed throughout the year. Most occurs during a winter shutdown, lasting several months, and "technical stops" lasting approximately two weeks each and distributed throughout the year. The runtime is consequently also concentrated, with a peak structure of repeated 10–15 hour-long collision periods with inter-fill gaps of 2–3 hours between them. The timing budget is determined by simulating the rate at which the disk buffer fills up and empties, using the processing speed measurements and the most recent LHC fill structure as a guide.[2] The objective is to ensure that the disk buffer will never exceed more than 80% capacity at any point throughout the remaining data taking period, and is evaluated every two weeks using the actual disk occupancy at the time as the starting point. The output rate of HLT1 is adjusted to keep the disk buffer usage within the desired limits. This output rate is controlled by switching between two HLT1 configurations, where the tighter configuration sacrifices some rate and efficiency for the inclusive general purpose trigger lines while protecting the trigger lines used for specific areas of the physics programme. The buffer usage is monitored throughout the year, and biweekly simulations are made using the present buffer capacity, HLT1 output rate and HLT2 throughput to determine the projected disk usage until the end of the year. Should a significant fraction of these simulations exceed the 80% usage threshold, the HLT1 configuration is tightened. An example simulation and the disk usage throughout 2017 are shown in Fig. 17.

6.4 Efficiency measurement method

All efficiencies are measured on background-subtracted data using the so-called TISTOS method described in the Run 1 performance paper [1] and briefly recapped here. Offline-selected signal events are divided into the following categories:

[1] The processing farm consists of processors with a certain number of physical cores, but a single task will not generally fully load a single physical core. For this reason the number of logical HLT tasks which are launched for each physical CPU core is optimized by measuring the overall throughput of events in the farm. For the 2015 HLT farm, this number is typically around 1.6, depending on the node in question.

[2] In 2015 this optimization used the 2012 fill information.

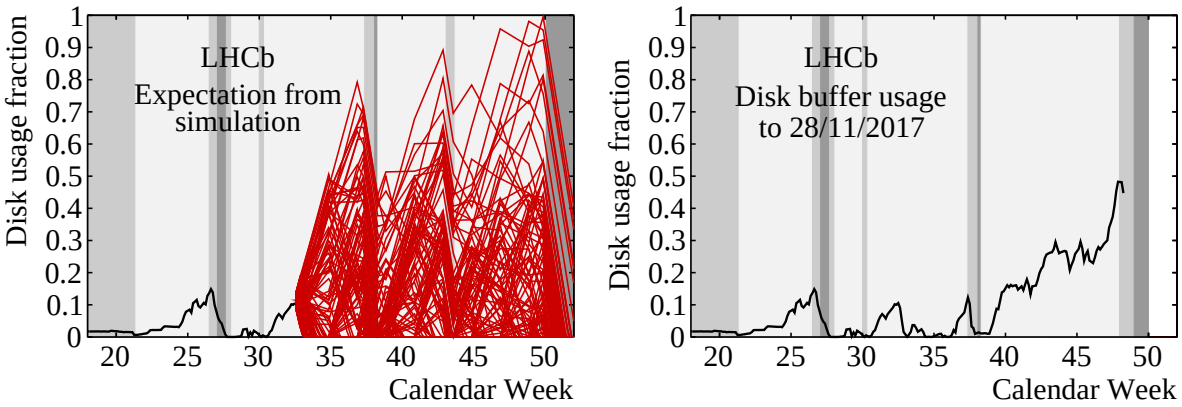

Figure 17: Disk buffer usage projections during (left) and at the end of (right) the 2017 data-taking period. During data taking, simulations (red, left) are used every two weeks to determine the probability of exceeding the 80% usage threshold. In 2017, the loose HLT1 configuration was used for the entire year leading to a maximum buffer capacity of 48% (black, right). LHC Technical Stops and Machine Development (MD) periods are shown in dark and light grey, respectively. The schedule changed between when this simulation was run in week 32 and the end of the year. An MD period was removed and the duration of the data taking was reduced.

- **TIS:** Events which are triggered independently of the presence of the signal decay. These are unbiased by the trigger selection except for correlations between the signal decay and the rest of the event. (For example when triggering on the "other" B in the event and subsequently looking at the momentum distribution of the "signal" B, the correlation in their momenta is caused by the fact that they both originate in the same fragmentation chain.)
- **TOS:** Events which are triggered on the signal decay independently of the presence of the rest of the event.

All efficiencies quoted in this paper are TOS efficiencies, given by

$$\epsilon = \frac{N(\text{TOS and TIS})}{N(\text{TIS})}, \tag{4}$$

where $N(\text{TIS})$ is the number of signal TIS events in the sample, while $N(\text{TOS and TIS})$ is the number of signal events which are both TOS and TIS. The number of signal events passing and failing the TOS criterion is measured using a histogram sideband subtraction, as described in Ref. [36]. In order to reduce the correlations between TOS and TIS events, the efficiency is plotted as a function of the p_{T} and, where appropriate, decay time of the signal particle.

6.5 Samples used for performance measurements

The performance of the L0 and HLT trigger selections is evaluated using samples of trigger-unbiased signals collected during Run 2, shown in Figs. 18 and 19. The signal channels

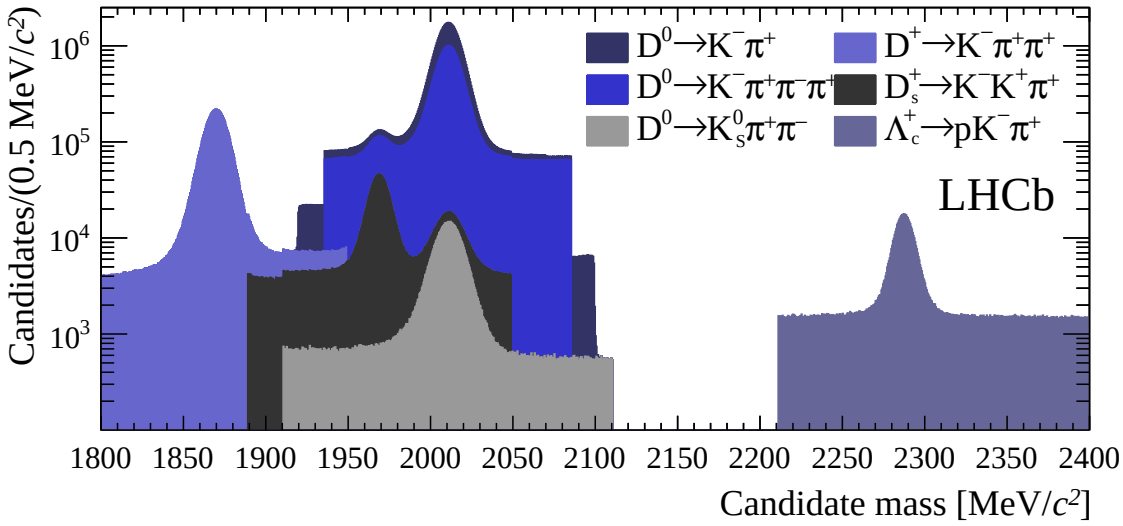

Figure 18: Charm candidates used for the evaluation of the trigger performance.

are representative of the LHCb physics programme, and are selected using relatively loose criteria on the kinematics and displacement of the b-hadron and the final-state particles. As the c-hadron signals are all fully selected by exclusive TURBO trigger lines, only their L0 and HLT1 efficiencies can be measured. The b-hadron signals are selected by offline selections without imposing any trigger requirements, and therefore, they can be used to measure the L0, HLT1, and HLT2 efficiencies. The exception is the $B^0 \to K^{*0}\gamma$ decay, where requiring TIS at HLT1 or HLT2 results in a signal yield and purity which are too small to be usable. Therefore, this mode is only used to study the L0 photon and electron trigger performance.

6.6 L0 performance

The efficiencies of the L0 trigger lines in Run 2 are shown in Fig. 20 for c-hadrons and Fig. 21 for b-hadrons, respectively, as a function of the p_{T} and the data-taking period. The L0 is optimized to fill the available ≈ 1 MHz bandwidth for a given set of LHC running conditions, and so the L0 efficiency evolves as a function of those conditions. In particular, if the LHC has to run with a reduced number of colliding bunches, the required rejection factor to reach 1 MHz output rate is smaller and the L0 criteria can be correspondingly loosened, which is the cause of the jumps visible in the bottom plots. In the case of the $B^0 \to K^{*0}\gamma$ decay, the low efficiency of the dedicated L0 photon trigger is due to the limited information available to separate electrons and photons within the L0 system. This identification relies on the amount of electromagnetic showering observed in the preshower detector, before the photons or electrons reach the ECAL, and whether there is a hit or not in the SPD detector. The chosen working point is such that the L0 photon trigger has a high purity but relatively low efficiency. However, many genuine photons are selected

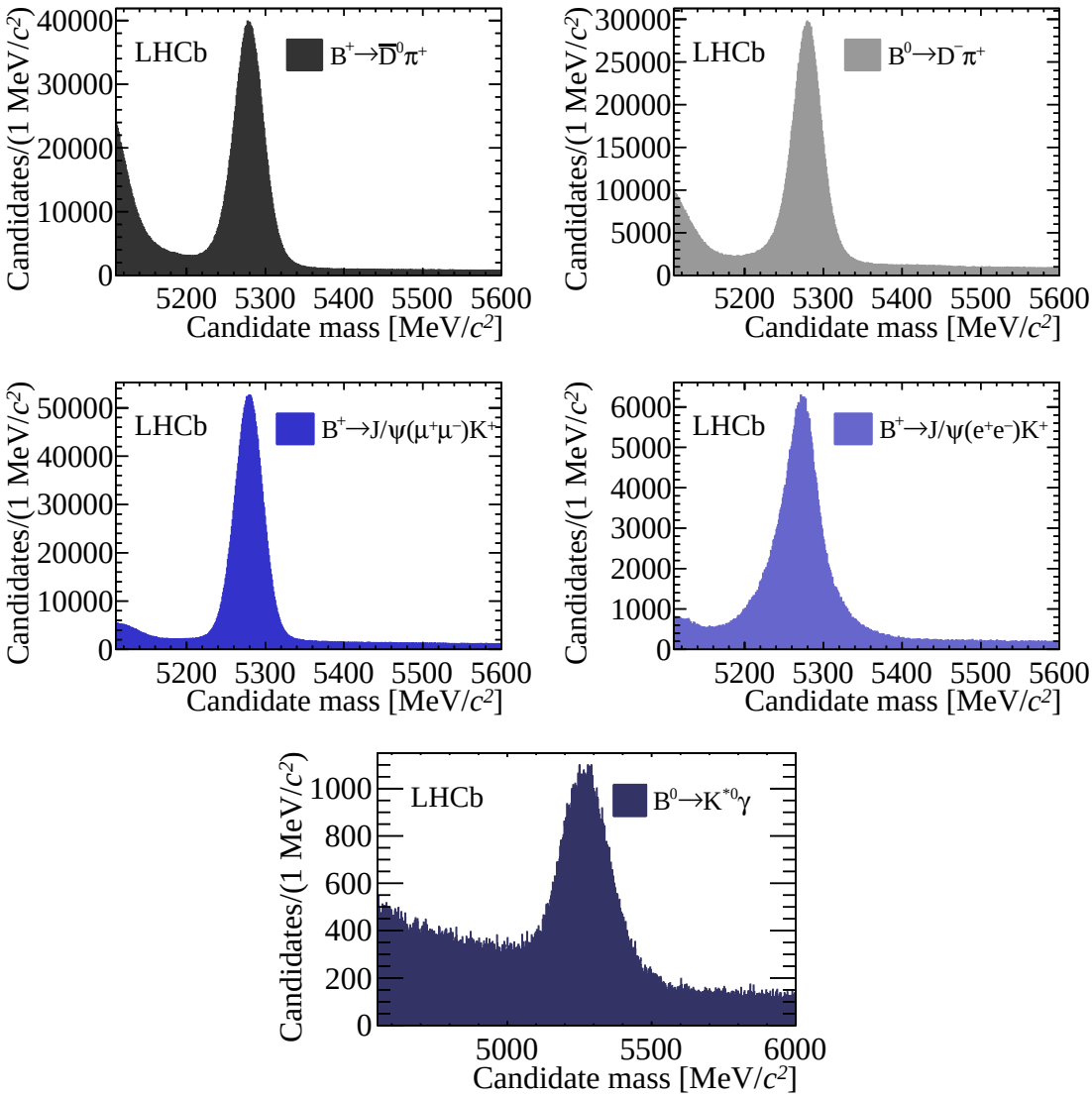

Figure 19: Beauty candidates used for the evaluation of the trigger performance.

by the L0 electron trigger, which is also efficient for photons. In addiction, a significant amount of photons convert in the detector material between the magnet and the SPD plane. These converted photons are reconstructed as neutral clusters offline, but leave a hit in the SPD detector and are therefore triggered as electrons. In practice the $B^0 \to K^{*0}\gamma$ signal is selected using both electron and photon L0 trigger lines to account for these effects.

The efficiency of each L0 trigger is measured with respect to events where the corresponding SPD criterion from Table 1 has already been applied. The distribution of SPD hits for $B^+ \to \overline{D}^0\pi^+$ signal candidates in Run 2 data is shown in Fig. 22, and is representative of typical heavy-flavour signals in LHCb. The efficiency of the SPD

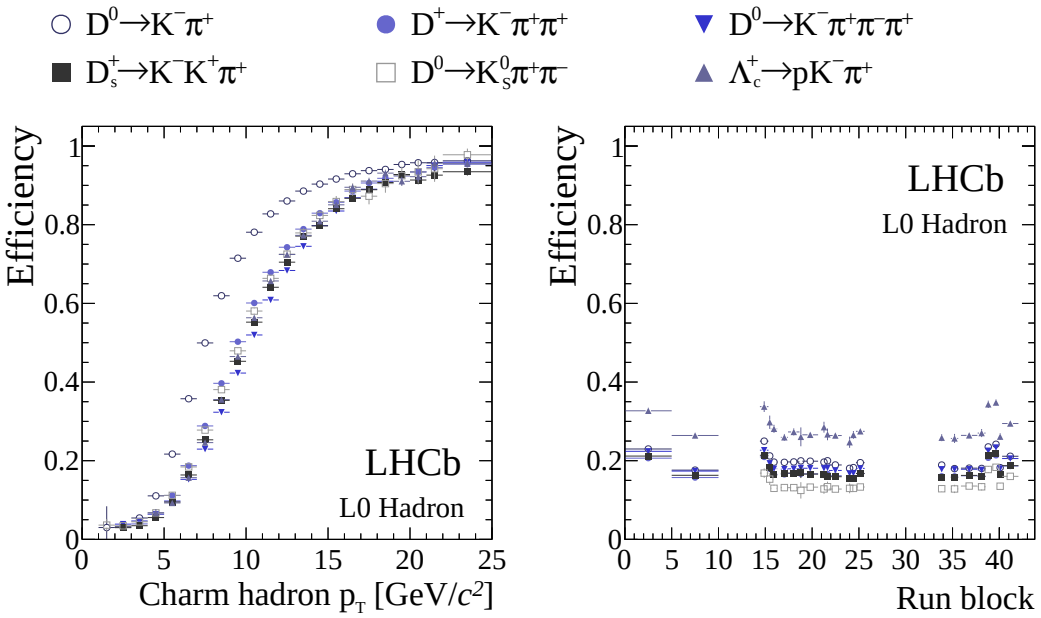

Figure 20: Efficiencies of the L0 trigger lines in Run 2 data for c-hadron decays. The left plot shows the efficiency as a function of the hadron p_{T}, while the right plot shows the evolution of the efficiency as a function of the different trigger configurations used during data taking. The three blocks visible in the plot, separated by vertical gaps, correspond to the three years of data taking (2015–2017). The L0 hadron efficiency is shown.

thresholds is generally around 90% for the L0DiMuon and 50% for the other heavy-flavour L0 trigger lines. The advantage of these SPD requirements is that they allow looser L0 kinematic thresholds.

The L0 trigger efficiencies as functions of the hadron p_{T} and η are shown in Fig. 23, except for the photon trigger where the signal yields are too small. The efficiency is relatively flat in η for any given p_{T} bin, although the calorimeter-based trigger lines do have a slightly better efficiency at high pseudorapidities.

6.7 HLT1 performance

The HLT1 trigger stage processes approximately 1 MHz of events that pass the L0 trigger, and reduces the event rate to around 110 kHz, which are further processed by HLT2. The HLT1 reconstruction sequence was described in Sec. 4, while this section describes the performance of the HLT1 trigger lines.

6.7.1 Inclusive lines

HLT1 has two inclusive trigger lines which select events containing a particle whose decay vertex is displaced from the PV: a line which selects a single displaced track with high p_{T}, and a line which selects a displaced two-track vertex with high p_{T}. The single track

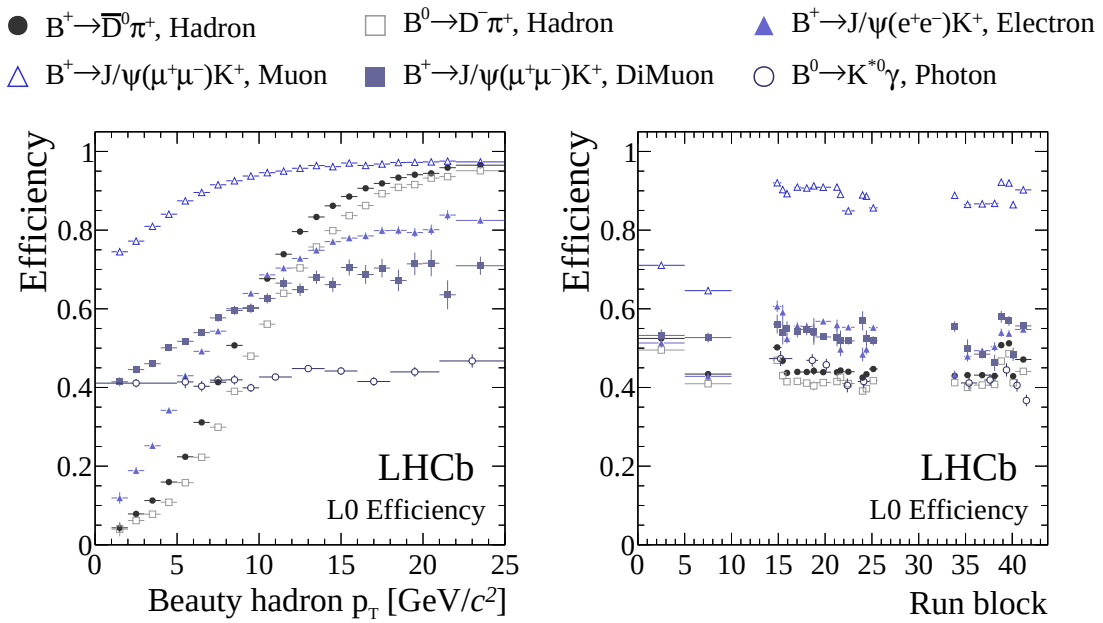

Figure 21: Efficiencies of the L0 trigger lines in Run 2 data for b-hadron decays. The left plot shows the efficiency as a function of the hadron p_{T}, while the right plot shows the evolution of the efficiency as a function of the different trigger configurations used during data taking. The three blocks visible in the plot, separated by vertical gaps, correspond to the three years of data taking (2015–2017). The plotted L0 efficiency for each b-hadron is described in the legend above the plots.

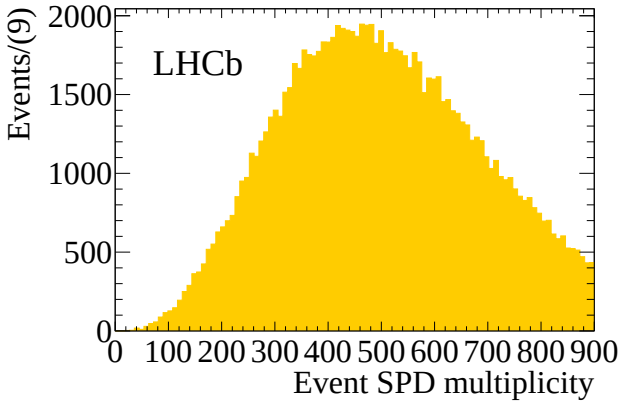

Figure 22: The SPD hit multiplicity of events containing $B^+ \to \overline{D}^0\pi^+$ candidates in Run 2 data.

line is a reoptimization of the Run 1 inclusive single track trigger [37], while the displaced two-track vertex trigger is a new development for Run 2. Both lines start by selecting good quality tracks that are inconsistent with originating from the PV. The single-track trigger then selects events based on a hyperbolic requirement in the 2D plane of the track

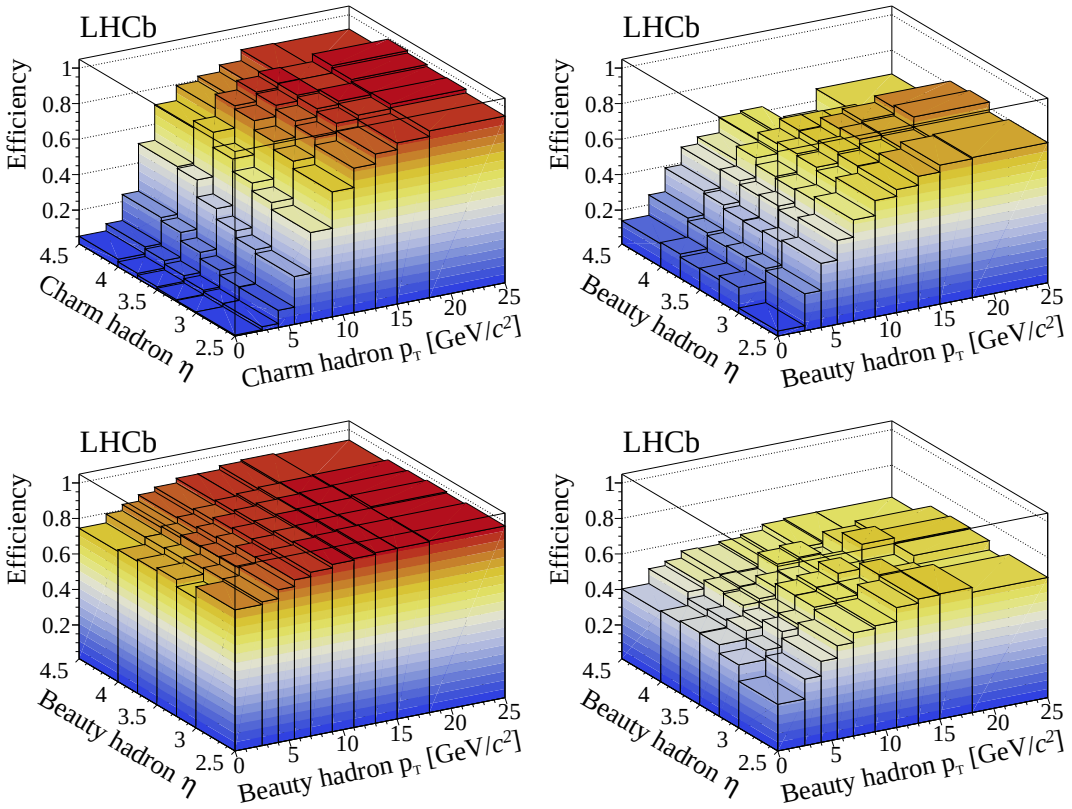

Figure 23: Two-dimensional efficiencies of the L0 trigger lines in Run 2 data: (top left) L0 hadron; (top right) L0 electron; (bottom left) L0 muon; and (bottom right) L0 dimuon. The L0 hadron efficiency is evaluated using $D^0 \to K^-\pi^+$ decays, whereas the others are evaluated using the relevant signals listed in Fig. 20 and Fig. 21.

displacement and $p_{\rm T}$.[3] The two-track displaced vertex trigger selects events based on a MatrixNet classifier [38] whose input variables are the vertex-fit quality, the vertex displacement, the scalar sum of the $p_{\rm T}$ of the two tracks, and the displacement of the tracks making up the vertex.

These trigger lines were primarily optimized for inclusively selecting the decays of b and c hadrons, and were trained using 26 different b- and c-hadron decays in order to make them as efficient as possible on the full spectrum of possible decay topologies. Care was taken, however, to make sure that these trigger lines would also be efficient for more exotic displaced signatures, for example hypothetical supersymmetric particles. The performance of these trigger lines is shown in Figs. 24 and 25.The two-track line is more efficienct at low $p_{\rm T}$, whereas the single track line performs best at high $p_{\rm T}$, such that combined they provide high efficiency over the full $p_{\rm T}$ range.

[3]More complicated multivariate selection criteria, for example boosted decision trees using track quality information in addition to the displacement and $p_{\rm T}$, were tried but gave no significant increase in performance.

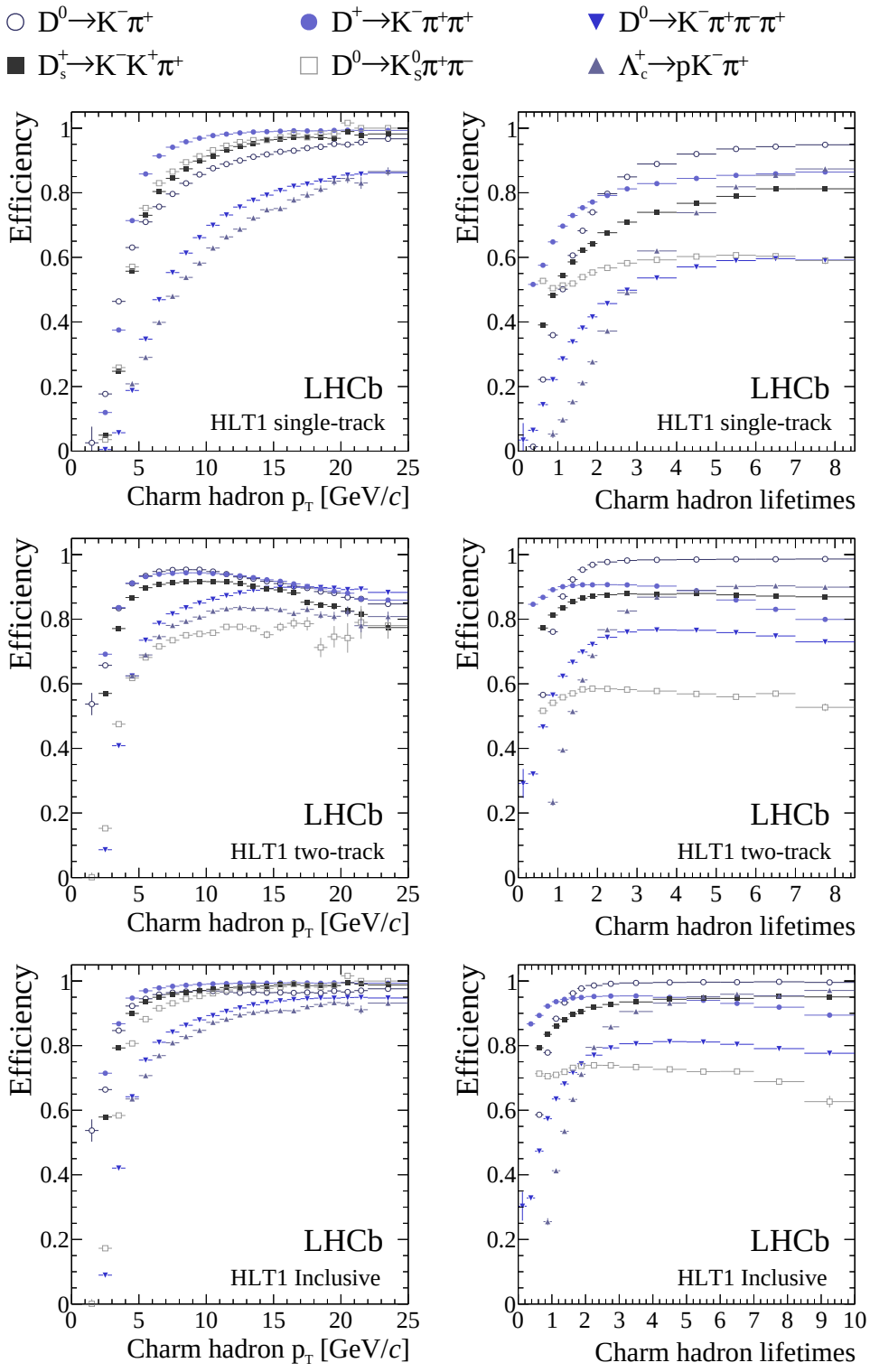

Figure 24: Efficiency of the HLT1 inclusive trigger lines as a function of (left) c-hadron p_{T} and (right) decay time. The decay time plots are drawn such that the x-axis is binned in units of the lifetime for each hadron in its rest frame. The plots in each column show, from top to bottom, the single-track, two-track, and combined HLT1 inclusive performance.

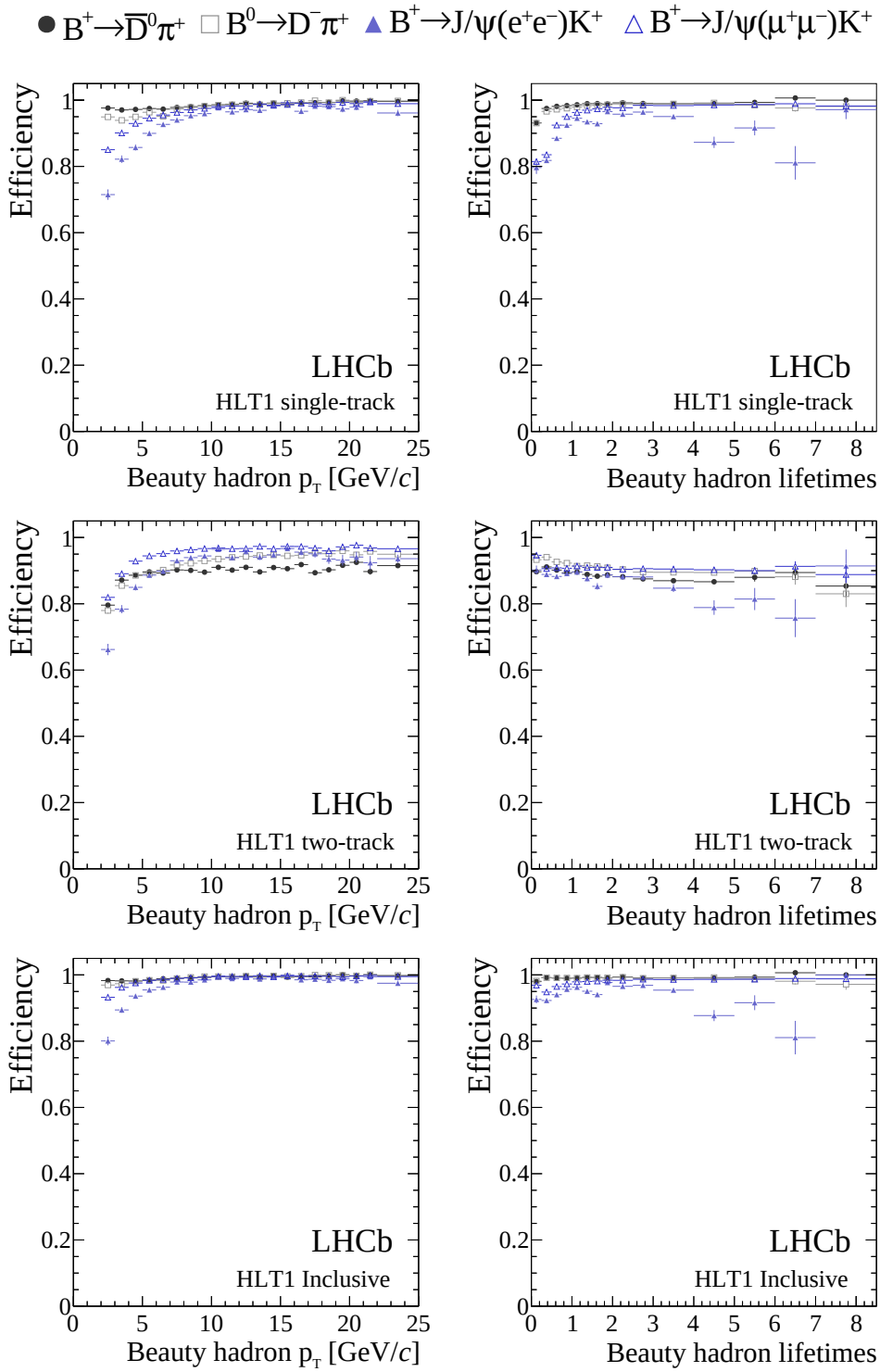

Figure 25: Efficiency of the HLT1 inclusive trigger lines as a function of (left) b-hadron p_{T} and (right) decay time. The decay time plots are drawn such that the x-axis is binned in units of the lifetime for each hadron in its rest frame. The plots in each column show, from top to bottom, the single-track, two-track, and combined HLT1 inclusive performance.

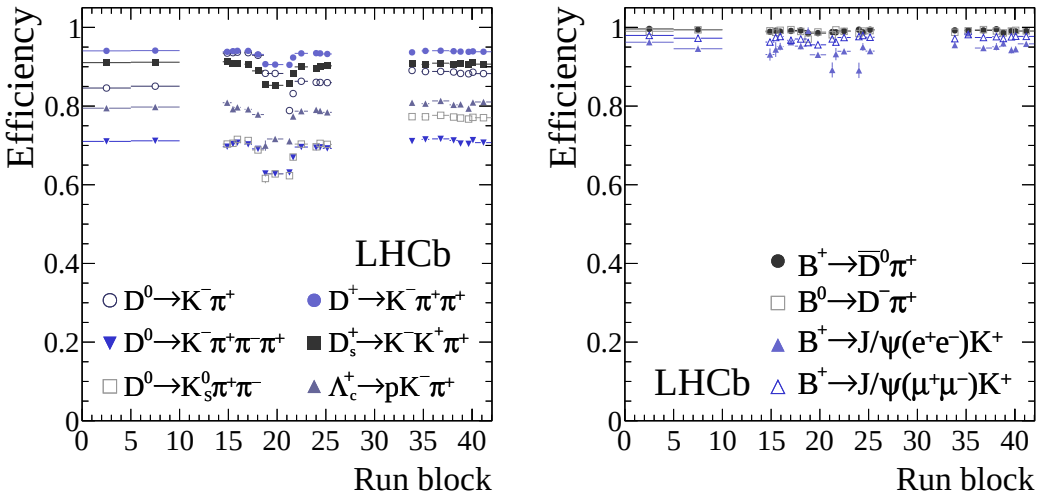

Figure 26: The HLT1 efficiency as a function of the different trigger configurations used during data taking for (left) c-hadrons and (right) b-hadrons. The three blocks visible in the plot, separated by vertical gaps, correspond to the three years of data taking (2015–2017).

Unlike the L0 trigger configurations, which changed frequently in response to varying LHC conditions, the HLT1 trigger configuration was kept largely stable, with some updates at the end of each data-taking year. The variation in the total HLT1 efficiency as a function of the data-taking period is shown in Fig. 26. The b-hadron efficiencies have been stable throughout the Run 2 data taking. The c-hadron efficiencies decreased midway through 2016, when a tighter HLT1 configuration was used to prevent the disk buffer from overflowing due to unexpectedly high LHC efficiency and availability. The improvements seen for some of the c-hadron channels in 2017 with respect to 2016 are caused by changes in the corresponding reference offline selections leading to a different average p_{T} and displacement of the c-hadron, not any intrinsic variation in the HLT1 performance or thresholds.

6.7.2 Muon lines

The HLT1 muon lines select muonic decays of b and c hadrons, as well muons originating from decays of W and Z bosons. As muons have an intrinsically cleaner signature than hadrons, the muon lines make use of simple rectangular selection criteria as opposed to the multivariate inclusive lines. There are four primary lines: one line that selects a single displaced muon with high p_{T} for flavour physics; a second single muon line that selects very high p_{T} muons, without displacement criteria, for electroweak physics; a third line that selects a dimuon pair compatible with originating from the decay of a charmonium or bottomonium resonance, or from Drell-Yan production; and a fourth line that selects displaced dimuons with no requirement on the dimuon mass. The efficiencies of the lines relevant for b-hadron decays are shown in Fig. 27 as obtained from data with the TISTOS method. Note that because these HLT1 muon trigger lines only run on events selected

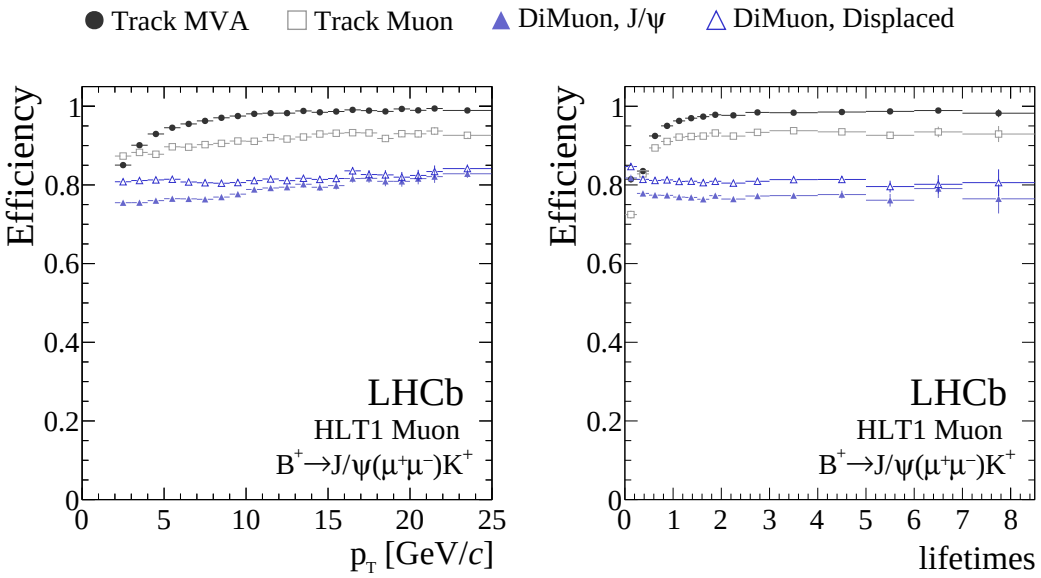

Figure 27: The efficiency of the HLT1 muon trigger lines as a function of the (left) b-hadron p_{T} and (right) units of the average B^+ decay time. The decay time plot is drawn such that the x-axis is binned in units of the B^+ lifetime in its rest frame. The efficiency of the inclusive single-track HLT1 trigger is plotted for reference.

by `L0Muon` and `L0DiMuon` trigger lines, their absolute efficiency is lower than that of the inclusive single-track HLT1 trigger, which runs on all `L0`-selected events. In addition to these lines, for Run 2 a new line dedicated to lower-p_{T} dimuons has been developed which has tighter criteria on the displacement of the dimuon but runs on all `L0`-selected events, rather than just the muon ones. This line is particularly important for selecting rare decays of strange hadrons, that are not triggered by the `L0` muon lines, increasing their HLT1 efficiency up to a factor three [39].

6.7.3 Calibration trigger lines

HLT1 contains two primary types of calibration trigger lines: a line which selects $D^0 \to K^-\pi^+$ candidates with significant displacement from the PV, and a line which selects $J/\psi \to \mu^+\mu^-$ candidates. The former is used for providing a pure sample of good tracks (the D^0 decay products) for the alignment of the tracking system, while the latter is used to provide a pure sample of muons for the alignment of the muon chambers. In addition, other trigger lines select events enriched in off-axis VELO tracks or tracks which populate the lower-occupancy regions of the RICH detectors, for use in the VELO and RICH alignment, respectively. The purity and yield of the calibration trigger lines is illustrated in Fig. 28, which shows the D^0 and J/ψ candidates reconstructed online in their respective lines for a specific fill corresponding to approximately 18.5 pb^{-1} of integrated luminosity.

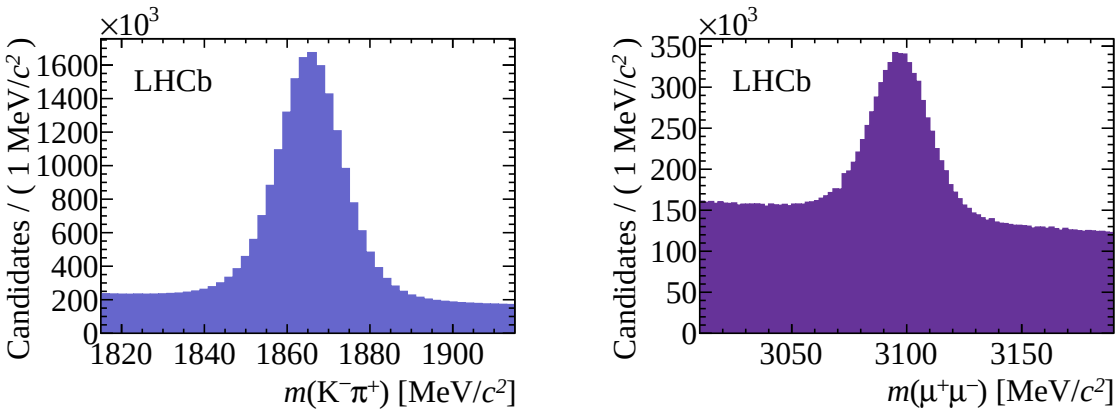

Figure 28: The D^0 (left) and J/ψ (right) candidates selected by the HLT1 calibration lines. Both plots show candidates reconstructed online.

6.7.4 Low multiplicity event and exclusive trigger lines

Special trigger lines for low-multiplicity events are needed to enable the study of central exclusive production (CEP). This kind of process takes place by colourless, low-p_{T} t-channel exchange between protons and can result in particle production in the central rapidity region. The protons remain intact and are deflected only slightly, so such production is typically accompanied by large ranges of rapidity with little detector activity, known as "rapidity gaps". The trigger development initially focussed on acquiring large samples of exclusively-produced dimuon candidates, but evolved to cover final states involving hadrons and calorimeter objects.

Since low levels of activity are anticipated for CEP, events with more than 30 hits in the SPD are rejected at the hardware level. Lower-bounds are also placed on relevant detector activity measurements as appropriate for each final state. These criteria indirectly favour the selection of events with exactly one pp interaction, as opposed to either multiple- or zero-interaction events. At the HLT1 stage, the low-multiplicity events containing muons or electromagnetic calorimeter objects occur at a low enough rate that can be selected with no additional requirements. but low-multiplicity events containing hadrons are required to have at least two tracks reconstructed in the VELO.

In addition, the low p_{T} thresholds implemented in the Run 2 HLT1 tracking allowed several special exclusive HLT1 trigger selections to be implemented for the first time, notably trigger lines that select two-body beauty and charm hadron decays without biasing their decay times [40]. In 2018 the HeRSCheL detector [41] is employed in the `L0` selection of CEP events, allowing for a reduction of the p_{T} thresholds.

6.7.5 HLT1 bandwidth division

The HLT1 bandwidth is preferentially allocated to the inclusive and muon trigger lines which, by selecting b- and c- hadron decays, cover most of the LHCb physics programme.

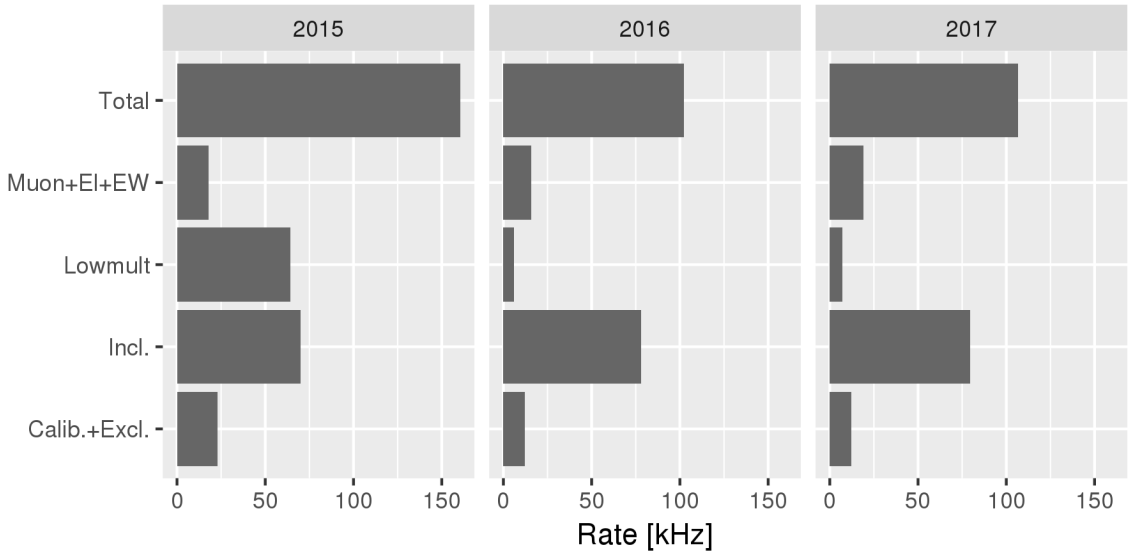

Figure 29: Rates of the main groups of HLT1 trigger lines and the total HLT1 rate as a function of the year of data taking, shown for the trigger configuration used to take most of the luminosity in each year.

The large disk buffer available in Run 2 also makes it possible to allow generous rates for other trigger lines, however, with a total HLT1 output rate of 150 kHz which is around two times the Run 1 average. The HLT1 rates and the overlaps in the events selected by the different HLT1 trigger lines are shown in Fig. 29.

6.8 HLT2 performance

The HLT2 trigger stage reduces the event rate to around 12.5 kHz, at which point the remaining events are saved to permanent storage for further analysis. The HLT2 reconstruction sequence was described in Sec. 5, while this section describes the performance of a representative set of HLT2 trigger lines.

6.8.1 Inclusive b-hadron trigger lines

The HLT2 inclusive b-hadron trigger lines look for a two-, three-, or four-track vertex with sizable p_{T}, significant displacement from the PV, and a topology compatible with the decay of a b hadron. As in Run 1 [42, 43], these "topological" trigger lines rely on a multivariate selection of the displaced vertex. This selection is implemented in a MatrixNet classifier whose inputs have been discretized [43] in order to minimize the variation in selection performance with varying detector conditions and speed up the evaluation time. The efficiency of the topological trigger lines is increased for decays involving muons by relaxing the requirement on the multivariate discriminant whenever one or more of the tracks associated with the topological vertex is positively identified as a muon or electron.

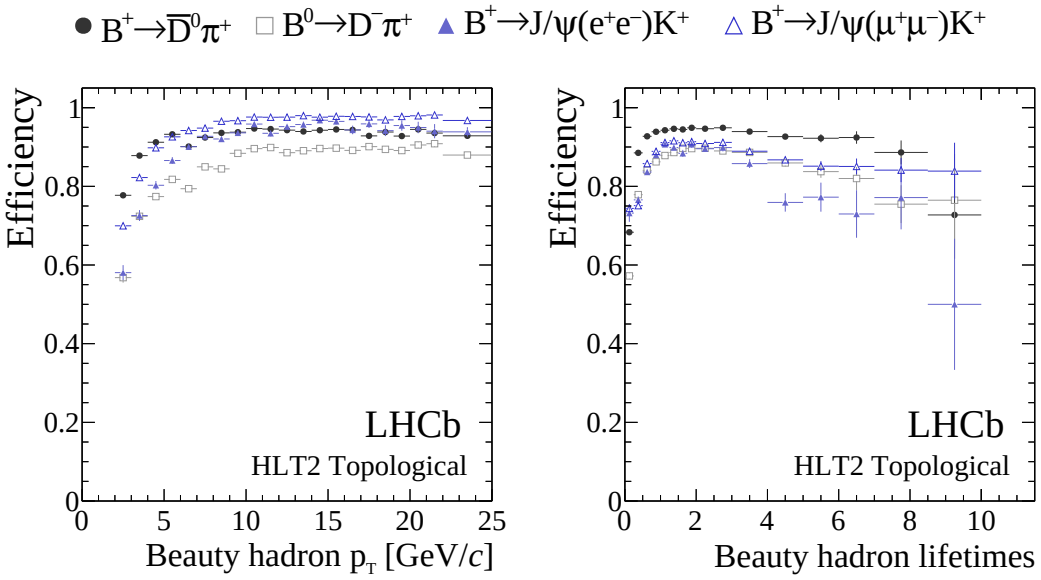

Figure 30: Efficiency of the HLT2 topological trigger lines as a function of the (left) b-hadron p_{T} and (right) in units of the average b-hadron decay time. The decay time plots are drawn such that the x-axis is binned in units of the lifetime for each hadron in its rest frame. The plots show the combined efficiency of the topological trigger lines for each b-hadron decay mode.

The topological trigger lines are trained to separate signal b-hadron decays which can be fully reconstructed inside the detector acceptance from those which cannot, as well as from displaced vertices formed from the decays of c hadrons originating from the PV. The displaced vertices from c hadrons are the most numerous background. Harder to discriminate against, however, are the backgrounds from b-hadron decays that are only partially contained in the detector acceptance, or b-hadron decays in which much of the energy is taken by neutral particles. The selection has been reoptimized [44] for Run 2, taking advantage of the full offline reconstruction now available in HLT2 to loosen the selection criteria when building vertex candidates, and fully relying on the multivariate algorithm to discriminate between signal and background. The resulting efficiencies are shown in Fig. 30 and Fig. 31 for a specific decay mode, while the evolution of the efficiency as a function of the data-taking conditions is shown in Fig. 32.

6.8.2 Muon and dimuon trigger lines

The HLT2 muon and dimuon trigger lines select a wide spectrum of signals: low-mass Drell–Yan dimuons for electroweak physics, dimuons originating from the PV for production measurements, dimuons with displacement from the PV for the study of b-, c-, and s-hadron decays and heavy dimuons for exotic particle searches and electroweak physics. As mentioned in Sec. 4.2, in Run 2 the HLT2 and offline muon-identification procedures are identical. Owing to this improvement and because muons provide a relatively rare and clean event signature, the dimuon trigger lines generally have a high efficiency which is

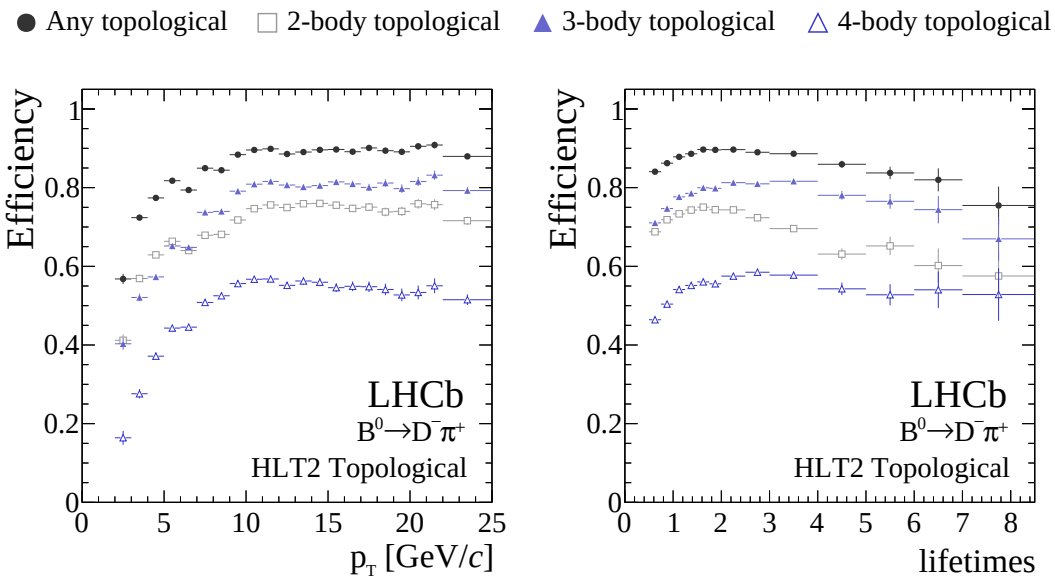

Figure 31: Efficiency of the HLT2 topological trigger lines as a function of the (left) b-hadron p_{T} and (right) in units of the average b-hadron decay time. The decay time plots are drawn such that the x-axis is binned in units of the lifetime for each hadron in its rest frame. The plots show the different contributions of the 2-, 3-, and 4-body topological trigger lines to a specific decay.

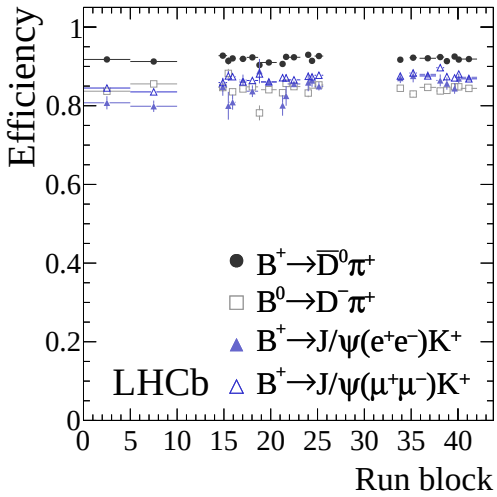

Figure 32: Evolution of the HLT2 efficiency as a function of the different trigger configurations used during data taking.

only limited in some cases by the rate of the selected signal, most notably for production measurements. This is illustrated in Fig. 33 where the efficiency of the HLT2 muon trigger lines is shown for $B^+ \to J/\psi K^+$ decays. Note that the muon topological trigger lines have a lower absolute efficiency compared to the hadron topological trigger lines because they

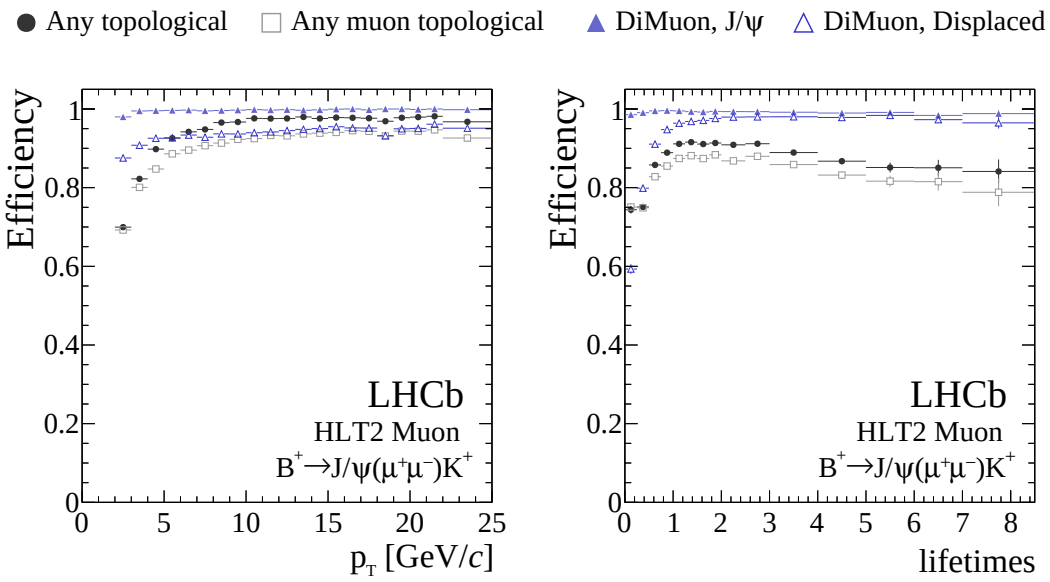

Figure 33: The TOS efficiency of the HLT2 muon trigger lines as a function of the (left) B^+ p_{T} and (right) units of the average B^+ decay time. The decay time plot is drawn such that the x-axis is binned in units of the B^+ lifetime in its rest frame. The efficiency of the inclusive topological ("any topological") trigger lines and topological trigger lines requiring one track identified as a muon ("any muon topological") are plotted for reference.

only process events passing the HLT1 single-muon selection. In addition to the standard inclusive muon lines used in Run 1, for Run 2 new lines have been developed in particular for dimuons with lower p_{T} for exotic particle searches (*e.g.* dark photons) and for rare strange-hadron decays [39].

6.8.3 Exclusive and calibration trigger lines

In addition to the inclusive trigger lines, the full offline reconstruction performed at the start of HLT2 means that it is possible to fully reconstruct certain decays of interest and select them using dedicated trigger lines without any loss in efficiency compared to the offline analysis. This is especially important for c-hadron trigger lines because around 10% of all 13 TeV proton-proton collisions produce a $c\bar{c}$ pair, and it is not possible to write all c-hadron signals to offline storage. In order to reduce the necessary disk space, the LHCb exclusive c-hadron trigger lines make extensive use of the TURBO stream. All events selected by those trigger lines, except those containing neutral particles, are sent to the TURBO stream. The selection criteria of these trigger lines are usually a slightly looser version of those used in the offline analysis, enabling the candidates saved in the TURBO stream to be directly used by the analysts. In total, over 200 different exclusive trigger lines which select the decays of c hadrons are deployed in Run 2. They are generally tuned to have S/B ratios well in excess of 1 already at the output of the trigger, with the final selection performed offline using information reconstructed in the trigger and

tuned to minimize systematic uncertainties. The purity achievable using the trigger-level information has already been illustrated in Fig. 18 for a representative sample of c-hadron decays.

In addition, HLT2 contains a suite of calibration trigger lines, which are used to measure the performance of the track-finding and particle-identification algorithms in a data-driven way. These trigger lines select high-yield charm, charmonium, and K^0_S decays using a tag-and-probe approach, where the probe particle is kept unbiased with respect to either the tracking or particle-identification information. There are around 50 such lines in total, and they select around 500 Hz of calibration signals.

6.8.4 Low multiplicity event trigger lines

At the HLT2 stage there are dedicated selections for each relevant final state with a low track multiplicity. There are 32 lines: two to select exclusive dimuon production, three to select exclusive production of photons or electrons, and the remainder to select various hadronic final states, dominated by lines that select low-p_T hadrons.

The HLT2 trigger efficiencies have been determined in data and are shown in Fig. 34 for two channels of particular interest: dimuon and dihadron. The dimuon HLT2 trigger efficiency is determined using a sample of independently triggered candidates reconstructed in events containing exactly two muon tracks inside the detector acceptance. The dimuon candidate is required to have satisfied the relevant low-multiplicity `L0` trigger. The efficiency is shown as a function of dimuon mass, where the rise at $800\,\text{MeV}/c^2$ results from the $400\,\text{MeV}/c$ p_T requirement for each muon. In the case of exclusive production, where the candidate is expected to be produced with low p_T, this leads to an implicit lower bound on the mass of the exclusively-produced object at $m(\mu^+\mu^-) \approx 800\,\text{MeV}/c^2$. The non-zero efficiency for candidates with $m(\mu^+\mu^-) \lesssim 800\,\text{MeV}/c^2$ arises from candidates with higher p_T.

The dihadron HLT2 trigger efficiency, which includes the effect of a 50% prescale, is determined using $\phi(1020) \to K^+K^-$ candidates reconstructed in low-multiplicity events and triggered independently of the signal candidate. The $\phi(1020)$ candidate is required to pass the relevant low-multiplicity `L0` and HLT1 trigger lines, and the background from misidentified pions is reduced using information from the RICH sub-detectors. The efficiency is shown as a function of the p_T of the $\phi(1020)$ meson.

6.8.5 HLT2 bandwidth division

The HLT2 bandwidth is divided into the full stream, containing inclusive trigger lines, and the TURBO stream, which contains exclusive trigger lines that fully reconstruct relevant decays. Most of the full stream rate is taken up by the topological b-hadron, inclusive c-hadron, and dimuon trigger lines, while the TURBO stream rate is divided among several hundred exclusive c-hadron trigger lines. As the TURBO stream trigger lines perform a full selection of high-purity signals, their rates are generally proportional to the signal abundance. The HLT2 rates and the overlaps in the events selected by the different HLT2

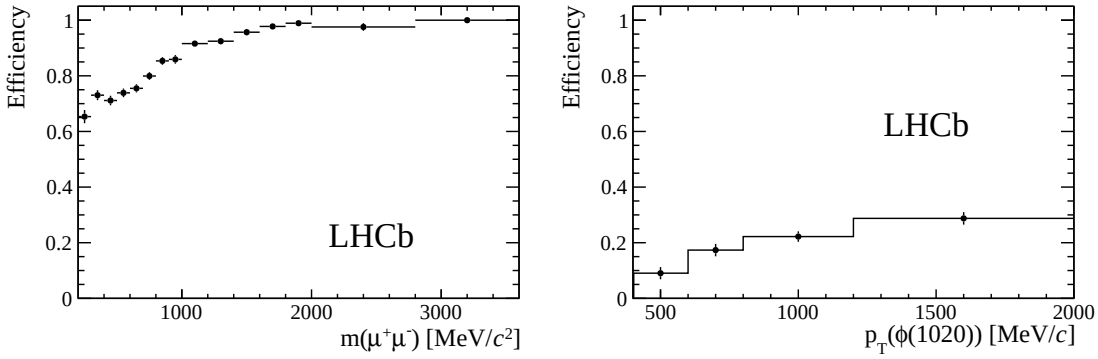

Figure 34: HLT2 trigger efficiencies of the dedicated selections for low-multiplicity events: (left) for dimuon candidates as a function of dimuon mass, and (right) for $\phi(1020)$ candidates as a function of candidate p_{T}.

Figure 35: Rates of the main categories of HLT2 trigger lines and the total HLT2 rate for each year of data taking, shown for the trigger configuration used to take most of the luminosity in the given year. TURBO, CALIBRATION, and FULL refer to different output data streams as discussed in Ref. [2].

trigger lines are shown in Fig. 35, where the exclusive trigger lines are counted as one item for brevity.

7 Conclusions

The design and performance of the LHCb Run 2 reconstruction and High Level Trigger have been presented. The use of real-time alignment and calibration and improvements in

the reconstruction software allows for events to be fully reconstructed in the High Level Trigger with equivalent quality to the Run 1 offline performance, and enables signals to be selected with a purity close to that achievable offline. This in turn enables physics analysis to be performed directly with the output of the reconstruction in the trigger. To this end, a significant fraction of triggered events is saved in a reduced "real-time analysis" format, saving only higher-level reconstructed objects relevant to physics analysis and not the full raw detector data. The successful deployment of this full real-time reconstruction and analysis during Run 2 is a critical stepping stone towards the LHCb upgrade, whose software trigger will have to deal with roughly 100 times greater data rates while maintaining a high acceptance over the same broad range of physics channels.

Acknowledgements

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).

References

[1] R. Aaij *et al.*, *The LHCb trigger and its performance in 2011*, JINST **8** (2013) P04022, arXiv:1211.3055.

[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] R. Aaij *et al.*, *A comprehensive real-time analysis model at the LHCb experiment*, arXiv:1903.01360.

[4] C. M. LHCb Collaboration, *Computing Model of the Upgrade LHCb experiment*, Tech. Rep. CERN-LHCC-2018-014. LHCB-TDR-018, CERN, Geneva, May, 2018.

[5] LHCb collaboration, A. A. Alves Jr. *et al.*, *The LHCb detector at the LHC*, JINST **3** (2008) S08005.

[6] LHCb collaboration, R. Aaij *et al.*, *LHCb detector performance*, Int. J. Mod. Phys. **A30** (2015) 1530022, arXiv:1412.6352.

[7] R. Aaij *et al.*, *Performance of the LHCb Vertex Locator*, JINST **9** (2014) P09007, arXiv:1405.7808.

[8] R. Arink *et al.*, *Performance of the LHCb Outer Tracker*, JINST **9** (2014) P01002, arXiv:1311.3893.

[9] M. Adinolfi *et al.*, *Performance of the LHCb RICH detector at the LHC*, Eur. Phys. J. **C73** (2013) 2431, arXiv:1211.6759.

[10] A. A. Alves Jr. *et al.*, *Performance of the LHCb muon system*, JINST **8** (2013) P02022, arXiv:1211.1346.

[11] T. Sjöstrand, S. Mrenna, and P. Skands, *PYTHIA 6.4 physics and manual*, JHEP **05** (2006) 026, arXiv:hep-ph/0603175; T. Sjöstrand, S. Mrenna, and P. Skands, *A brief introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, arXiv:0710.3820.

[12] I. Belyaev *et al.*, *Handling of the generation of primary events in Gauss, the LHCb simulation framework*, J. Phys. Conf. Ser. **331** (2011) 032047.

[13] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. **A462** (2001) 152.

[14] P. Golonka and Z. Was, *PHOTOS Monte Carlo: A precision tool for QED corrections in Z and W decays*, Eur. Phys. J. **C45** (2006) 97, arXiv:hep-ph/0506026.

[15] Geant4 collaboration, J. Allison *et al.*, *Geant4 developments and applications*, IEEE Trans. Nucl. Sci. **53** (2006) 270; Geant4 collaboration, S. Agostinelli *et al.*, *Geant4: A simulation toolkit*, Nucl. Instrum. Meth. **A506** (2003) 250.

[16] M. Clemencic *et al.*, *The LHCb simulation application, Gauss: Design, evolution and experience*, J. Phys. Conf. Ser. **331** (2011) 032023.

[17] P. d'Argent *et al.*, *Improved performance of the LHCb Outer Tracker in LHC Run 2*, JINST **9** (2017) P11016, arXiv:1708.00819.

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

[19] LHCb, S. Borghi, *Novel real-time alignment and calibration of the LHCb detector and its performance*, Nucl. Instrum. Meth. **A845** (2017) 560.

[20] O. Callot, *FastVelo, a fast and efficient pattern recognition package for the Velo*, LHCb-PUB-2011-001. CERN-LHCb-PUB-2011-001, LHCb.

[21] E. E. Bowen, B. Storaci, and M. Tresch, *VeloTT tracking for LHCb Run II*, LHCb-PUB-2015-024. CERN-LHCb-PUB-2015-024. LHCb-INT-2014-040.

[22] O. Callot and S. Hansmann-Menzemer, *The Forward Tracking: Algorithm and Performance Studies*, LHCb-2007-015. CERN-LHCb-2007-015.

[23] LHCb Collaboration, M. Stahl, *Machine learning and parallelism in the reconstruction of LHCb and its upgrade. Machine learning and parallelism in the reconstruction of LHCb and its upgrade*, J. Phys. : Conf. Ser. **898** (2017) 042042. 8 p.

[24] A. Dziurda, T. Lesiak, and V. Gligorov, *Studies of time-dependent CP violation in charm decays of B_s^0 mesons*, Apr, 2015. Presented 19 Jun 2015.

[25] R. Aaij *et al.*, *Optimization of the muon reconstruction algorithms for LHCb Run 2*, LHCb-PUB-2017-007. CERN-LHCb-PUB-2017-007.

[26] O. Callot and M. Schiller, *PatSeeding: a standalone track reconstruction algorithm*, LHCb-2008-042. CERN-LHCb-2008-042.

[27] M. Needham and J. Van Tilburg, *Performance of the track matching*, LHCb-2007-020. CERN-LHCb-2007-020.

[28] M. Needham, *Performance of the Track Matching*, LHCb-2007-129. CERN-LHCb-2007-129.

[29] A. Davis, M. De Cian, A. M. Dendek, and T. Szumlak, *PatLongLivedTracking: A tracking algorithm for the reconstruction of the daughters of long-lived particles in LHCb*, LHCb-PUB-2017-001. CERN-LHCb-PUB-2017-001.

[30] A. Hoecker *et al.*, *TMVA 4 — Toolkit for Multivariate Data Analysis. Users Guide.*, `arXiv:physics/0703039`.

[31] H. Voss, A. Hoecker, J. Stelzer, and F. Tegenfeldt, *TMVA - Toolkit for Multivariate Data Analysis*, PoS **ACAT** (2007) 040.

[32] M. De Cian, S. Farry, P. Seyfert, and S. Stahl, *Fast neural-net based fake track rejection in the LHCb reconstruction*, LHCb-PUB-2017-011. CERN-LHCb-PUB-2017-011.

[33] LHCb collaboration, R. Aaij *et al.*, *Measurement of the track reconstruction efficiency at LHCb*, JINST **10** (2015) P02007, arXiv:1408.1251.

[34] T. Skwarnicki, *A study of the radiative cascade transitions between the Upsilon-prime and Upsilon resonances*, PhD thesis, Institute of Nuclear Physics, Krakow, 1986, DESY-F31-86-02.

[35] V. Breton, N. Brun, and P. Perret, *A clustering algorithm for the LHCb electromagnetic calorimeter using a cellular automaton*, LHCb-2001-123.

[36] F. Archilli *et al.*, *Performance of the muon identification at LHCb*, JINST **8** (2013) P10020, arXiv:1306.0249.

[37] V. V. Gligorov, *A single track HLT1 trigger*, LHCb-PUB-2011-003.

[38] A. Gulin, I. Kuralenok, and D. Pavlov, *Winning the transfer learning track of Yahoo's learning to rank challenge with YetiRank*, in *Proceedings of the Learning to Rank Challenge* (O. Chapelle, Y. Chang, and T.-Y. Liu, eds.), vol. 14 of *Proceedings of Machine Learning Research*, (Haifa, Israel), pp. 63–76, PMLR, 25 Jun, 2011.

[39] F. Dettori, D. Martinez Santos, and J. Prisciandaro, *Low-p_T dimuon triggers at LHCb in Run 2*, LHCb-PUB-2017-023.

[40] M. W. Kenzie and V. Gligorov, *Lifetime unbiased beauty and charm triggers at LHCb*, LHCb-PUB-2015-026. CERN-LHCb-PUB-2015-026.

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

[42] V. V. Gligorov, C. Thomas, and M. Williams, *The HLT inclusive B triggers*, LHCb-PUB-2011-016. CERN-LHCb-PUB-2011-016. LHCb-INT-2011-030, LHCb-INT-2011-030.

[43] V. V. Gligorov and M. Williams, *Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree*, JINST **8** (2013) P02013, arXiv:1210.6861.

[44] T. Likhomanenko *et al.*, *LHCb topological trigger reoptimization*, J. Phys. Conf. Ser. **664** (2015) 082025.