Real-time track reconstruction with FPGAs in the LHCb Scintillating Fibre Tracker beyond Run 3

Wander Baldini[1], Giovanni Bassi[2,3], Andrea Contu[4], Riccardo Fantechi[2], Jibo He[5,6], Brij Kishor Jashal[7], Sofia Kotriakhova[1,8], Federico Lazzari[2,9], Maurizio Martinelli[10,11], Diego Mendoza[7], Michael J. Morello[2,3], Arantza De Oyanguren Campos[7], Lorenzo Pica[2,3], Giovanni Punzi[2,9], Qi Shi[5], Francesco Terzuoli[2,12], Giulia Tuci[13], Ao Xu[2], and Jiahui Zhuo[7]

[1]INFN Sezione di Ferrara, Ferrara, Italy
[2]INFN Sezione di Pisa, Pisa, Italy
[3]Scuola Normale Superiore, Pisa, Italy
[4]INFN Sezione di Cagliari, Monserrato, Italy
[5]University of Chinese Academy of Sciences, Beijing, China
[6]Hangzhou Institute for Advanced Study, UCAS, Hangzhou, China
[7]Instituto de Fisica Corpuscular, Centro Mixto Universidad de Valencia - CSIC, Valencia, Spain
[8]Università di Ferrara, Ferrara, Italy
[9]Università di Pisa, Pisa, Italy
[10]INFN Sezione di Milano-Bicocca, Milano, Italy
[11]Università di Milano Bicocca, Milano, Italy
[12]Università di Siena, Siena, Italy
[13]Physikalisches Institut, Ruprecht-Karls-Universitat Heidelberg, Heidelberg, Germany

E-mail: ao.xu@cern.ch

Abstract. Finding track segments downstream of the magnet is an important and computationally expensive task of the high-level trigger of the LHCb Upgrade I. These segments are utilised to form high-quality physics tracks with precision momentum measurement when combined with those reconstructed in the vertex track detector, and to reconstruct tracks belonging to long-lived particles decayed after the vertex track detector. LHCb is currently developing a new real-time tracking device based on distributed system of FPGAs, dedicated to the reconstruction of track-segment primitives in the forward Scintillating Fibre tracker detector at the full LHC collision rate. The aim is to significantly accelerate the online reconstruction in Run 4 extending the physics performance of the experiment, and to develop a new heterogeneous technology capable of affording even higher instantaneous luminosity conditions foreseen for Run 5 and beyond. In this report we discuss the first detailed study of the reconstruction performance expected from this device, based on an accurate simulation of its architecture at the bit level.

1 Introduction

The LHCb Upgrade I [1] is designed to run at an instantaneous luminosity of $2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ during Run 3 and 4 of LHC, with a bunch-crossing rate of 40 MHz and a mean number of proton-proton interactions per bunch crossing of six. The new data acquisition system consists of a single stage detector readout, followed by event-building on a local area network and by real time reconstruction and selection. The latter is a two-stage filtering by various computing elements, referred to as High Level Trigger (HLT), helped by a suitable intermediate storage system. In the ultimate version of the experiment, the LHCb Upgrade II [2], the detector will operate at a maximum instantaneous luminosity of $1.5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and a number of proton-proton interactions per bunch crossing of about 40. In these more demanding conditions, it is advantageous to perform early low-level reconstruction at the readout level, using custom processors to produce intermediate, more compact data structures, hereafter referred to as primitives. Using primitives rather than raw data as the starting point of the HLT reconstruction can save both bandwidth and computing resources for higher-level tasks in the selection of physics objects, which require the full flexibility of a high-level software environment. For such reconstructions to be effective and practical, it must be able to operate at an actual 30 MHz event rate without time-multiplexing, with a limited amount of buffering, and transparently at the readout level.

In this report, we describe the performance of a specialized firmware system to be deployed during the Long Shutdown 3 of LHC, running on an array of interconnected Field Programmable Gate Array (FPGA) boards. This system, referred to as DoWnstream Tracker (DWT), aims at reconstructing track primitives transparently during the readout of LHCb tracking stations downstream of the dipole magnet. The DWT computes track primitives to be used as seeds by the HLT reconstruction, using artificial retina algorithm [3]. The primitives reconstructed in each individual collision event are output in parallel by many physically separated boards, closely similar to the raw data output by any detector, making the online reconstruction faster, thus freeing computational resources to be used more efficiently for higher level tasks. Particularly, this will allow to extend the physics program of the experiment involving the long-lived particles [4], and will ensure the viability of this heterogeneous approach even at higher instantaneous luminosities [2, 5].

2 Tracking system of the LHCb Upgrade I

The tracking system of the LHCb Upgrade detector [1] consists of a hybrid pixel sensors vertex detector (VELO) surrounding the *pp* interaction region, a large-area silicon-strip detector (Upstream Tracker or UT) located upstream of a dipole magnet with a bending power of about 4 Tm, and three stations (T-stations) of scintillating fibres detectors (Scintillating Fibre Tracker or SciFi) placed downstream of the magnet. Each station of SciFi is composed of four detection layers with an x-u-v-x geometry, with vertical fibres in first and last layers, and tilted fibres by a stereo angle of $-5°$ and of $+5°$ in central layers. [1] Each layer has a hit resolution of 72 μm and a hit efficiency of 97.5%.

Reconstructed tracks fall into different categories depending on the subdetectors used in their reconstruction. Long tracks, which are reconstructed with hits in the VELO and the T-stations, are of most interest for physics analyses. They have excellent spatial resolution close to the primary interaction and precise momentum information. Tracks consisting of measurements only in the T-stations are referred to as T-tracks. They are not typically used in physics analyses, but instead are used as seeds to reconstruct the downstream tracks, which are reconstructed with measurements in the UT and the T-stations. Downstream tracks are essential for physics programmes that involve long-lived particles such as K^0_S mesons and $\mathit{\Lambda}$ baryons, a large fraction of which decay outside VELO.

The standalone T-track reconstruction is performed by solving the complex and heavy pattern recognition task, with the Seeding algorithm. It is implemented both in the second stage of HLT based on CPUs, and recently in the first stage of HLT based on GPUs [6]. A first implementation of downstream tracking in the first stage of HLT (HLT1) also exists and is to be included during the data-taking of Run 3 [7]. In the current configuration, the standalone T-track reconstruction in HLT1 consumes a substantial fraction of the available GPU power, about 30-40%. As a proof-of-concept, if the pattern recognition of x-z track projection is replaced with T-track primitives of x-z projection from DWT, the throughput of HLT1 sequence increases by about 20%. If the pattern recognition is replaced with 3D T-track primitives from DWT, the throughput increase is about 30% [4].

3 Artificial retina architecture

The technology in the core of the proposed system is a highly-parallel architecture for pattern recognition known as artificial retina [3]. The Retina architecture is an arrangement of parallel computing units fed by a custom switching network programmed to perform a computation resembling the Hough transform.

[1] LHCb uses a right-handed coordinate system with the z-coordinate along the beam axis, and the vertical y-coordinate coinciding with the direction of the magnetic field.

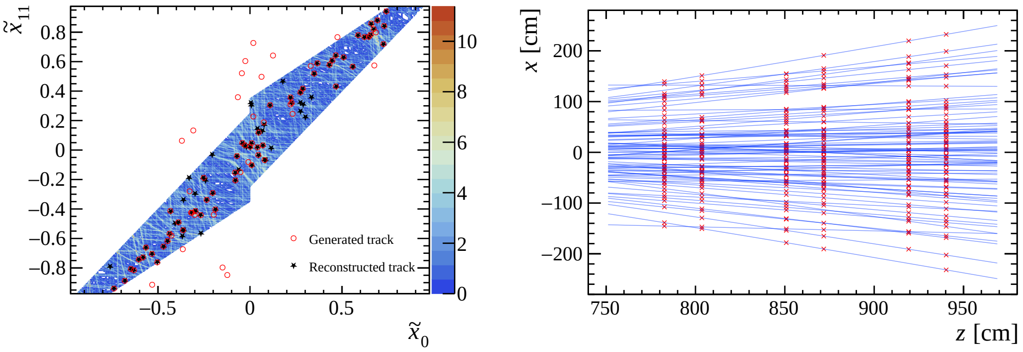

Figure 1: Excitation level of the axial retina from a single fully simulated event (left), where true tracks are indicated with red circles while reconstructed track candidates with black stars. Physical representation of reconstructed axial tracks in the SciFi detector (right).

The purpose is to find patterns of hits in its input that are compatible with a set of precalculated reference tracks, referred to as receptors. The recognised patterns are output in parallel over a large number of parallel lines to avoid any bottleneck due to serialisation. The implementation is realised in practice by vertically segmenting both the distribution network and the logic cell into smaller blocks that can be allocated to individual FPGAs of an array, while preserving a certain amount of necessary horizontal connectivity. The whole system, together with its implementation details, has been developed to its current form, prototyped, and tested, over the course of a decade [4]. A hardware demonstrator has been installed and tested with live LHCb Run 3 data, by reconstructing a quadrant of the VELO detector [4].

The DWT performs real-time T-track reconstruction in two stages. Details, including a description of all developments, can be found in Ref. [4] and are outlined here. The first stage is to find x-z (or axial) track projection. Tracks are approximated, to a good extent, as straight lines in this initial reconstruction stage, disregarding the presence of a small component of fringe field in the SciFi region. The axial track projection is therefore parameterised in a two-dimensional space (x_0, x_{11}), where x_0 and x_{11} are the x-coordinates of the intersections of the track with two virtual planes located just before the first layer of the SciFi and just after the last one, respectively. To ensure the efficient use of FPGAs resources, a transformed space, $(\tilde{x}_0, \tilde{x}_{11})$, is used and both coordinates are normalised to the range $[-1, 1]$ [8]. The transformation ensures that hits from axial layers are uniformly distributed. This space is subdivided into a 648×648 square grid, chosen to balance the grid size and the cell occupancy. The active region of cells is determined to have an acceptance of 100% for all tracks with $p > 5\,\text{GeV}/c$, which corresponds to about 98% for tracks with $p > 3\,\text{GeV}/c$. This amounts to about 73k cells for both the top and bottom half of SciFi. Local maxima in the parametric space exceeding a certain excitation threshold is considered to be axial track candidates. The excitation threshold has an efficiency above 95% for generic particles reconstructible in the SciFi subdetector, but also with a fake-track rate of 95%. To reject fake tracks, track fits are performed over all hit combinations formed from the two hits closest to the cell receptor on each SciFi layer. These are linear fits with three parameters, using the same modified parabola model as in the Seeding algorithm to account for the fringe magnetic field. A threshold is set on the $\chi^2_{\rm A}$ of the fit for accepting the track candidate. An event display of axial track candidates is shown in Fig. 1.

The second stage is to associate u/v (or stereo) hits, which aims at a further reduction of the ghost rate at the prebuild level, and at the maximum achievable acceleration of the HLT tracking sequences. First studies on the subject are reported in Refs. [8, 9] and showed a promising performance. Due to the negligible effect of the fringe field in the y-z plane, the vertical projection of track trajectories (stereo tracks) can be approximated as straight lines originating from the nominal interaction point. The stereo track is, therefore, described with one single parameter in the first stage of stereo hits association. The parameter is chosen to be the y coordinate of the intersection of the track with a virtual plane located in the middle of the SciFi. The one-dimensional space of the y coordinate is then discretised with a limited number of bins, Δy, referred to as stereo retina. In the current study, 45 bins are used. For each candidate axial track, to find compatible u/v hits, the u/v coordinates are transformed into y coordinates according to the geometric relation $y = \frac{x_{{\rm pred},u/v} - x_{{\rm meas},u/v}}{\tan\alpha}$, where $x_{{\rm pred},u/v}$ is the predicted x coordinate of the track trajectory (obtained from the axial track fit), $x_{{\rm meas},u/v}$ is the measured u/v coordinate and $\tan\alpha$ is the tangent of the small stereo angle of the SciFi u/v layers. The hits are sent and accumulated with unit weight in the corresponding Δy bin. A dedicated Switching Network, as that of the axial

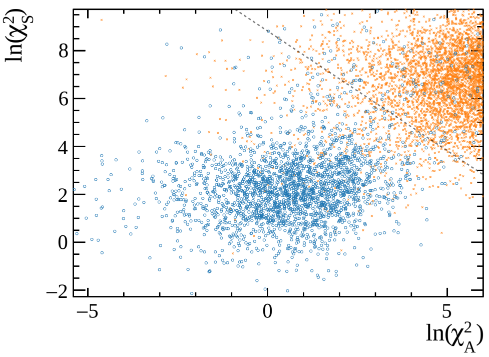

Figure 2: Two-dimensional distribution of log-sized $(\chi^2_{\rm A}, \chi^2_{\rm S})$. Truth-matched track candidates are displayed in blue while truth-unmatched in orange. The working point is indicated with the dashed black line.

reconstruction, is needed to distribute stereo hits to relevant bins. In order to save bandwidth, only axial track candidates passing the requirement of $\chi^2_{\rm A} < 400$ are considered for stereo association. The average (maximum) number of u/v hits filling each stereo retina is expected to be of the order of 120 (340).

For each stereo retina, bins with no less than 5 hits and with at least 1 hit on 5 different SciFi u/v layers are considered to be stereo track candidates. Then a χ^2 fit to a linear model $y = a_y + b_y \times z$ is performed on all combinations of 5 hits from different SciFi u/v layers. Finally a stereo track candidate is found if the resultant minimum $\chi^2_{\rm S}$ over all fit combinations passes a pre-determined quality requirement on the two-dimensional $(\chi^2_{\rm A}, \chi^2_{\rm S})$ space. Figure 2 shows the two-dimensional distribution of $(\chi^2_{\rm A}, \chi^2_{\rm S})$ for truth-matched an truth-unmatched reconstructed tracks. The average number of combinations to be run in each stereo retina is about 150 and is highly correlated to the number of bins of the stereo retinas.

4 Performance of T-track reconstruction with DWT

To estimate the performance of track primitives reconstructed with DWT, three simulated samples with the nominal LHCb Run 4 conditions are used, which differ in event topology. The first is a sample of generic inelastic events (the so-called Minimum Bias sample), while the other two are filtered samples containing a hard collision in each event, producing a $D^{*+} \to D^0\pi^+ \to [K^0_{\rm S}\pi^+\pi^-]\pi^+$ decay or a $B^0_s \to \phi\phi \to [K^+K^-][K^+K^-]$ decay, respectively. The beam energy is set to 7 TeV, and the average number of pp interactions per bunch crossing is set to 7.6. The dedicated DWT emulator, a C++ software which uses integers to emulate the firmware implementation at bit-level, is developed to simulate the behaviour of the DWT. A Monte Carlo generated particle is deemed to be reconstructible in the SciFi subdetector if it produced at least one x-coordinate hit and one u/v-coordinate hit in each of the three T-stations. The minimum number of hits for a track to be reconstructible in the SciFi subdetector is therefore equal to six. In our samples, the average number of reconstructible particles per event is around 150, with a significant tail extending to values up to 400 (see Ref. [8] for more details). The most crowded events (about 10% of the total) are discarded by the Global Event Cut (GEC) requirement [10], before entering the High Level Trigger. We will therefore quote all reconstruction performance parameters for events passing this requirement.

The ghost rate is defined as the fraction of reconstructed tracks not associated to an actual Monte Carlo particle (truth-matching), relative to the total number of reconstructed track candidates. All efficiencies and ghost rates are determined by applying the fiducial physics requirements $p_{\rm T} > 200\,{\rm MeV}/c$ and $2 < \eta < 5$, in addition to the standard LHCb definition of reconstructible tracks for the different categories listed.

In order to evaluate the physics performance, it is important to determine a reasonable working point for the requirement on both the $\chi^2_{\rm A}$ and $\chi^2_{\rm S}$ values to accept high-quality reconstructed 3D T-track primitives. We choose to use threshold values in the plane $(\chi^2_{\rm A}$,$\chi^2_{\rm S})$ with a linear boundary, as indicated in Fig. 2, which return a reconstruction efficiency of 90% for generic long tracks having a momentum threshold of $p > 5\,{\rm GeV}/c$. The resulting efficiencies, including acceptance, in finding 3D track using the chosen working point are shown in Tab. 1 for all track categories and simulated samples. Generic downstream tracks display a reconstruction efficiency of about 88(83)% with a momentum requirement of $p > 5(3)\,{\rm GeV}/c$. A reconstruction efficiency of about 90% is found for Long tracks from B decays (not electrons and $p > 5\,{\rm GeV}/c$) and an efficiency of 88(83)% for Down tracks from strange (not electrons and $p > 5(3)\,{\rm GeV}/c$). With this selection the ghost rate reduces to about 17%, which corresponds to 0.2 fake

Table 1: Averaged reconstruction efficiencies for different simulated samples and different track categories. The ghost rate is also shown. Event-averaged values are shown in brackets. The physics fiducial requirements $p_{\mathrm{T}} > 200\,\mathrm{MeV}/c$ and $2 < \eta < 5$ are also applied to determine all efficiencies.

Track type	MinBias	$D^0 \to K^0_{\mathrm{S}}\pi^+\pi^-$	$B^0_s \to \phi\phi$
Long, $p > 3\,\mathrm{GeV}/c$	85 (86)	83 (84)	84 (85)
Long, $p > 5\,\mathrm{GeV}/c$	90 (91)	89 (90)	89 (89)
Long from B not $e^{\pm}$, $p > 3\,\mathrm{GeV}/c$	-	-	88 (87)
Long from B not $e^{\pm}$, $p > 5\,\mathrm{GeV}/c$	-	-	90 (90)
Down, $p > 3\,\mathrm{GeV}/c$	84 (85)	83 (84)	83 (84)
Down, $p > 5\,\mathrm{GeV}/c$	89 (91)	88 (89)	88 (89)
Down from strange not $e^{\pm}$, $p > 3\,\mathrm{GeV}/c$	-	83 (83)	-
Down from strange not $e^{\pm}$, $p > 5\,\mathrm{GeV}/c$	-	88 (88)	-
Down from strange not long not $e^{\pm}$, $p > 3\,\mathrm{GeV}/c$	-	83 (83)	-
Down from strange not long not $e^{\pm}$, $p > 5\,\mathrm{GeV}/c$	-	88 (89)	-
ghost rate	16 (10)	17 (12)	17 (13)
ghost rate / (1 - ghost rate)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)

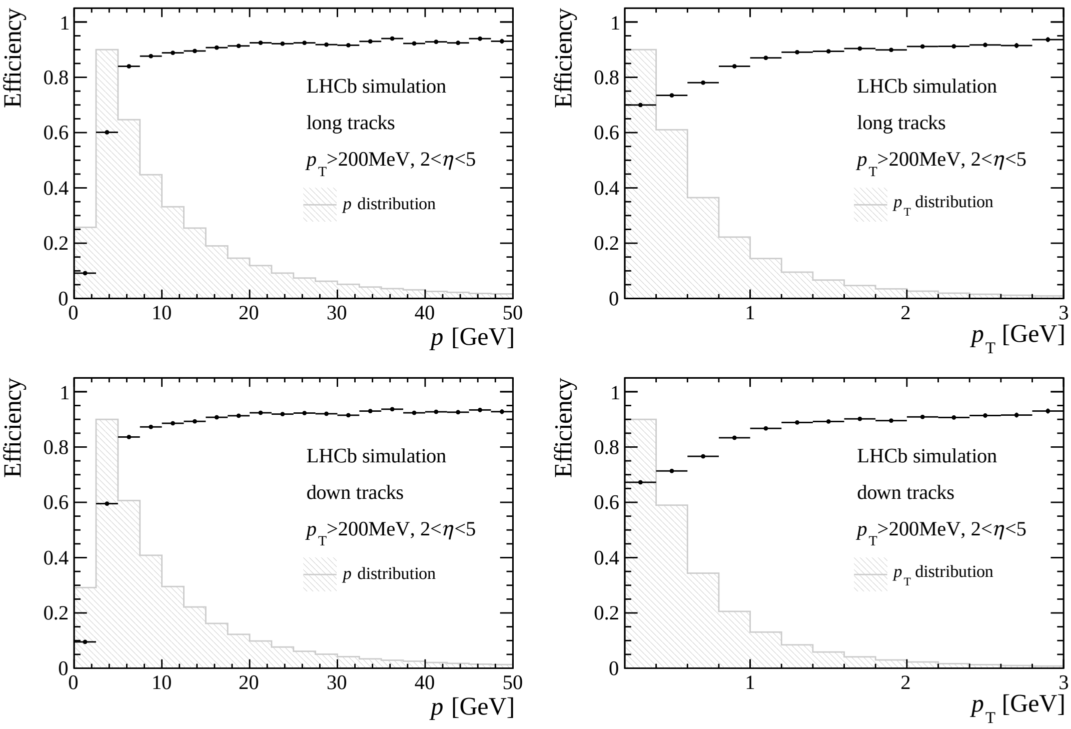

Figure 3: Differential 3D reconstruction efficiency as a function of p and p_{T} for (top) Long tracks and (bottom) Downstream tracks. The $B^0_s \to \phi\phi$ sample is used. The physics fiducial requirements $p_{\mathrm{T}} > 200\,\mathrm{MeV}/c$ and $2 < \eta < 5$ are also applied to determine efficiencies.

track for each truth-matched reconstructed track. The ghost rate of Seeding algorithm implemented in HLT1 is about 11% (5% with the clone killing [11]), which can be compared with the value found with the DWT. Efficiencies higher than those obtained with the current configurations of the system can be obtained by relaxing the chosen quality requirements, at the price to have an increase of the number of reconstructed ghost tracks. This is not expected to be an issue, due to the strong ability to remove ghost tracks in downstream tracking algorithms. These results indicate that the number of Retina cells planned for this device is adequate for obtaining a good performance.

For illustrative purpose, the differential 3D reconstruction efficiencies as a function of p and p_{T}, for generic long and downstream tracks are shown in Fig. 3. The differential efficiency and ghost rate for

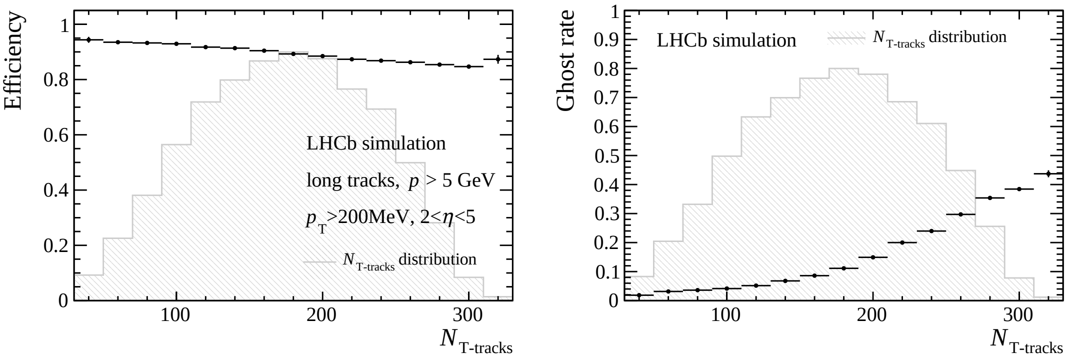

Figure 4: Differential (left) 3D reconstruction efficiency for generic Long tracks having a momentum threshold of $p > 5\,\text{GeV}/c$ and (right) ghost rate as a function of the number of T-tracks. The $B_s^0 \to \phi\phi$ sample is used. The physics fiducial requirements $p_{\text{T}} > 200\,\text{MeV}/c$ and $2 < \eta < 5$ are also required to determine efficiencies.

generic long tracks having a momentum threshold of $p > 5\,\text{GeV}/c$ as a function of the number of T-tracks are also shown in Fig. 4.

5 Conclusions

In this report we present the first comprehensive study of the performance of T-track primitives reconstructed in real time with the artificial retina architecture. The study utilises fully simulated events under LHCb Upgrade I conditions and accurate bit-level emulation of FPGA firmware implementation. This allows for a reliable estimate of the amount of hardware to be deployed for the purpose of the project and paves the way for a global optimisation of hyper-parameters both in the proposed system and along with the downstream reconstruction algorithms in the LHCb high level trigger.

References

[1] LHCb, R. Aaij *et al.*, *The LHCb Upgrade I*, JINST **19** (2024) P05065, arXiv:2305.10515.

[2] LHCb Collaboration, *Framework TDR for the LHCb Upgrade II - Opportunities in flavour physics, and beyond, in the HL-LHC era*, tech. rep., CERN, Geneva, 2021. CERN-LHCC-2021-012.

[3] L. Ristori, *An artificial retina for fast track finding*, Nucl. Instrum. and Meth. **A453** (2000) 425.

[4] M. J. Morello *et al.*, *Proposal for FPGA-based tracking in the LHCb downstream region*, tech. rep., CERN, Geneva, 2024. LHCb-PUB-2024-001.

[5] LHCb Collaboration, *LHCb Data Acquisition Enhancement TDR*, tech. rep., CERN, Geneva, 2024. CERN-LHCC-2024-001.

[6] LHCb Collaboration, *Standalone track reconstruction and matching algorithms for the GPU-based High Level Trigger at LHCb*, , LHCB-FIGURE-2022-010.

[7] A. De Oyanguren Campos, B. K. Jashal, and Z. Jiahui, *Downstream track reconstruction algorithms for GPU-based High level trigger at LHCb*, , Talk at the Trigger and readout at LHC experiments for Run3 and beyond, Valencia, March 2023.

[8] A. Di Luca, *Real-time reconstruction of tracks in the Scintillatin Fibre Tracker of the LHCb Upgrade*, Master's thesis, Università di Pisa, Pisa, 2018. CERN-THESIS-2018-272.

[9] M. J. Morello *et al.*, *Real-time reconstruction of long-lived particles at LHCb using FPGAs*, J. Phys. Conf. Ser. **1525** (2020) 012101, arXiv:2006.11067.

[10] LHCb, R. Aaij *et al.*, *A Comparison of CPU and GPU Implementations for the LHCb Experiment Run 3 Trigger*, Comput. Softw. Big Sci. **6** (2022) 1, arXiv:2105.04031.

[11] L. Calefice, *Standalone track reconstruction on GPUs in the first stage of the upgraded LHCb trigger system & Preparations for measurements with strange hadrons in Run 3*, PhD thesis, TU Dortmund and Sorbonne Université/LPNHE, 2022, CERN-THESIS-2022-343.