Proceedings of the CTD 2022
PROC-CTD2022-28
July 29, 2022

Tracking on GPU at LHCb's fully software trigger

Alessandro Scarabotto

LPNHE, Sorbonne Université, CNRS/IN2P3, Paris, France,
On behalf of the LHCb Collaboration.

ABSTRACT

The LHCb experiment will use a fully software-based trigger to collect data from 2022 on, at an event rate of 30 MHz. During the first stage of the High-Level Trigger (HLT1), a partial track reconstruction, using efficient parallelisation techniques on GPU cards, is performed. This stage will reduce the event rate by around a factor 30. Reconstructing tracks at 30 MHz represents a challenge which needs to be faced with very performant tracking algorithms. These proceedings focus on reconstruction and performance aspects of HLT1, giving particular attention to the algorithms reconstructing particles traversing the whole LHCb detector.

PRESENTED AT

Connecting the Dots Workshop (CTD 2022)
May 31 - June 2, 2022

1 Introduction

The LHCb detector is a single-arm forward spectrometer installed at the Large Hadron Collider (LHC) and specialized to study heavy flavour physics [1, 2]. During Run 1 and 2 of the LHC, the experiment collected 9 fb^{-1} of data. However, many crucial studies are still statistically limited and require an increased amount of data together with an improved detector performance, to achieve uncertainties comparable to the associated theory. LHCb is undergoing a major upgrade during the Long Shutdown 2 in order to collect 50 fb^{-1} by 2033 (Run 3 and 4) [3]. To achieve this, the instantaneous luminosity will be increased by a factor 5 compared to Run2, reaching $2 \cdot 10^{33} cm^{-2} s^{-1}$.

To sustain the experimental challenges of the increased luminosity, higher granularity and more radiation tolerant sub-detectors are needed, together with new front-end and readout electronics. A sketch of the upgraded LHCb detector is shown in Figure 1.

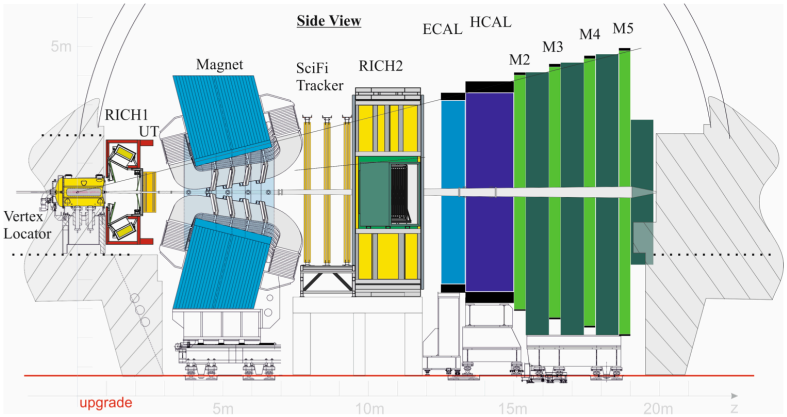


Figure 1: Side view of the LHCb upgraded detector.

In particular, all tracking sub-detectors are undergoing major upgrades:

- The Vertex Locator (VELO), a sub-system located around the proton-proton collision point and devoted to precisely measuring primary and secondary vertices, is composed of 26 tracking layers based on 50x50 μm^2 pixel technology.
- The Upstream Tracker (UT), a sub-detector before the dipole magnet designed to reconstruct long-lived particles decaying after the VELO, is composed of 4 layers of silicon strips.
- The Scintillating Fiber tracker (SciFi), located after the magnet to provide measurements of particle momenta, is constructed of 12 scintillating fiber layers divided into 3 stations.

The Run 2 triggering scheme is a major limiting factor to exploit the $2 \cdot 10^{33} cm^{-2} s^{-1}$ luminosity upgrade. As shown in Figure 2, the trigger yield for many hadronic channels was already saturating with Run 2 conditions ($4 \cdot 10^{33} cm^{-2} s^{-1}$).

This limitation is mainly caused by the LHCb's first level trigger (L0) implemented in hardware, which requires minimum energy depositions in the calorimeter to trigger on hadrons. To resolve this limitation, events will be selected by a fully software-based trigger during the upcoming data taking. Removing the L0 trigger will increase the hadronic channels' triggering efficiency by a factor of 2 for beauty hadron decays and typically larger factors for charm and light-hadron decays with respect to Run 2.

The software-based High Level Trigger (HLT) will perform the event reconstruction at the 30 MHz non-empty bunches collisions the LHC. The full LHCb data flow is presented in Figure 3.

Raw detectors' information is sent to a custom data processing center, where FPGA cards receive data at an average of 5 TB/s. Then, the Event Builder (EB) CPU servers use this information to produce packets of events that can be processed directly by the HLT. The first High Level Trigger (HLT1) reduces the 30 MHz

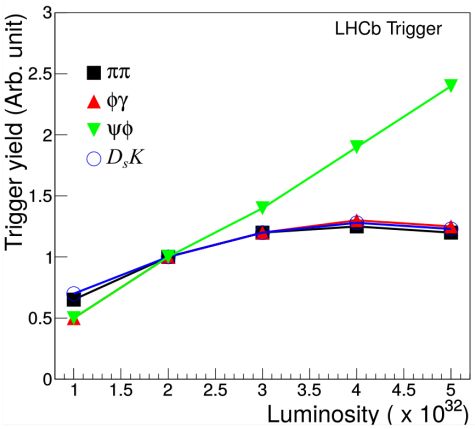


Figure 2: Run2 trigger yield dependence on instantaneous luminosity. [5] Various decays are identified by different colors. Hadronic trigger yields saturate when increasing the luminosity.

input rate to 1 MHz selecting interesting events by partially reconstructing them [4]. Graphical Processing Units (GPUs) are ideal for this purpose, as the reconstruction can be parallelised at the event level and further at the track level. Moreover GPUs can be added to each EB server with zero extra overhead costs.

A real-time alignement and calibration is then performed inside a buffer before processing events with the second High Level Trigger (HLT2). HLT2 performs a full event reconstruction on CPUs, reducing the rate from 1 MHz down to around 50 kHz. HLT2 selected events will be then processed offline to create final samples to be analysed.

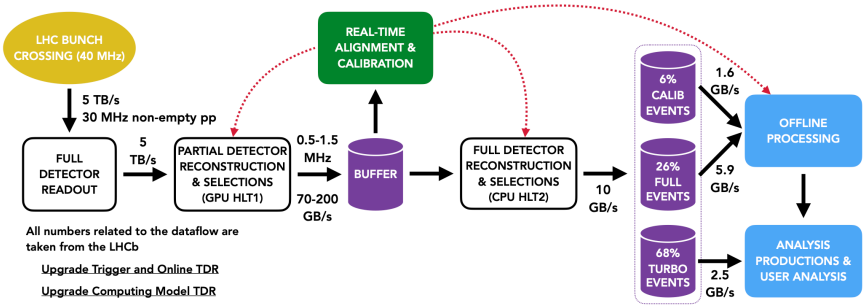


Figure 3: LHCb upgrade dataflow focusing on the real-time aspects. [6]

2 Reconstruction on GPUs

The Allen project takes care of the HLT1 reconstruction on GPUs at LHCb. Charged particle reconstruction is performed down to a momentum threshold of 3 GeV. For LHCb's physics purposes, different types of tracks must be reconstructed (see Figure 4):

- VELO tracks leaving hits only in the VELO subdetector,
- Upstream tracks which are bent out of the magnet after the UT,
- Long tracks crossing all tracking subdetectors. These tracks are very important for LHCb as they provide the best particle momentum measurement,

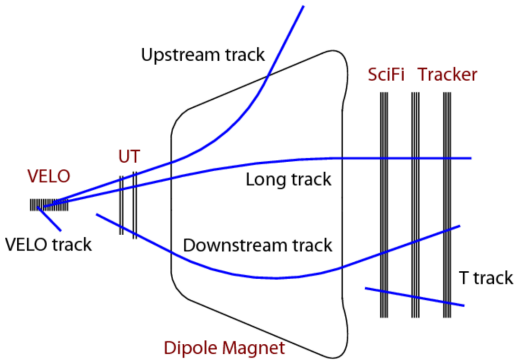


Figure 4: Types of tracks reconstructed by the LHCb experiment.

Allen features various algorithms to perform the event reconstruction, using input raw data from the LHCb's subdetectors. The default sequence is presented in Figure 5 and composed of:

- A global event cut, which rejects the 10% of highest occupancy minimum bias events prior any processing, based on UT and SciFi raw data information,
- VELO reconstruction using the "Search by Triplet" algorithm, performing an optimized straight line fit [9],
- Primary vertex finding and fitting, extrapolating VELO tracks as straight lines to the beamline,
- VELO-UT tracking with the "CompassUT" algorithm [11], providing the first charge and momentum measurement and having therefore the first meaningful objects for physics selections,
- Forward reconstruction, extrapolating VeloUT tracks to the SciFi region using a B field's parametrized trajectory,
- Parametrized Kalman filter to improve the estimate of certain physics variables, such as the impact parameter (IP),
- Muon identification, extrapolating the reconstructed forward tracks to find matching hits in the muon system,
- Calorimeter reconstruction to correctly identify electrons from corresponding hits in the calorimeter and recover their lost radiated energy with a Bremsstrahlung recovery algorithm,
- Secondary vertex finding from long-lived particles decay,
- Event selection with trigger lines to cover most of the LHCb physics programme.

2.1 Velo tracking

Reconstruction at LHCb's HLT1 level starts from tracks leaving hits in the detector closest to the proton-proton collision point: the VELO. This tracking sub-detector is designed to reconstruct particles in the LHCb forward acceptance (10-300 mrad) and it is composed of 26 modules on each side of the beam pipe. In this first region ($\sim$ 1 meter), the effect of LHCb's magnet field is negligible. Tracks are therefore expected to be straight lines. The "Search by Triplet" algorithm [9] achieves optimal computational performance with a double parallelization, at both the event and track level.

As trajectories are expected to start close to the XY plane origin, tracks can be reconstructed selecting hits by ϕ (angle in the XY plane) and z. The algorithm can then be explained in 3 steps:

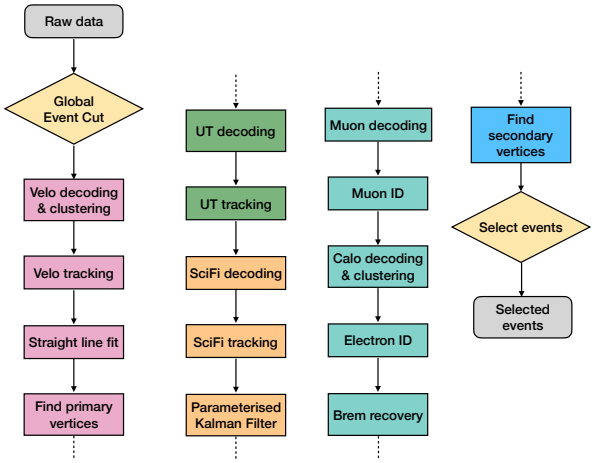


Figure 5: Schematic of HLT1 reconstruction flow at LHCb.

1. Track seeding operating on 3 consecutive modules, looking for hits inside ϕ windows,
2. Forwarding the search to next modules to complete the track,
3. Repeat the seeding process on the leftover hits.

As shown in Figure 6, the track reconstruction efficiency is about 99 % in all ranges of momenta and transverse momenta.

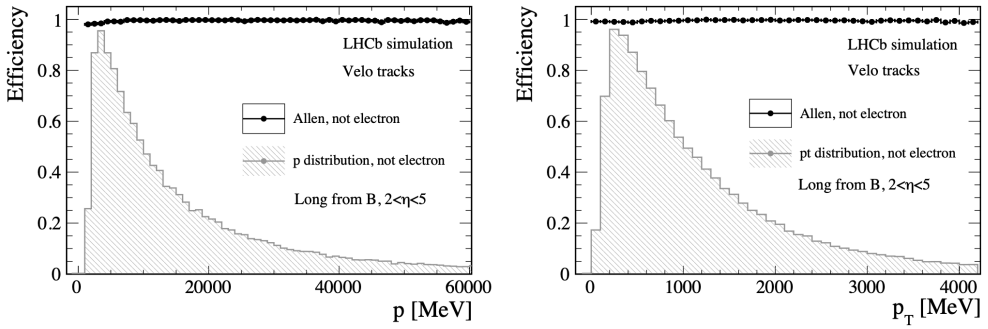
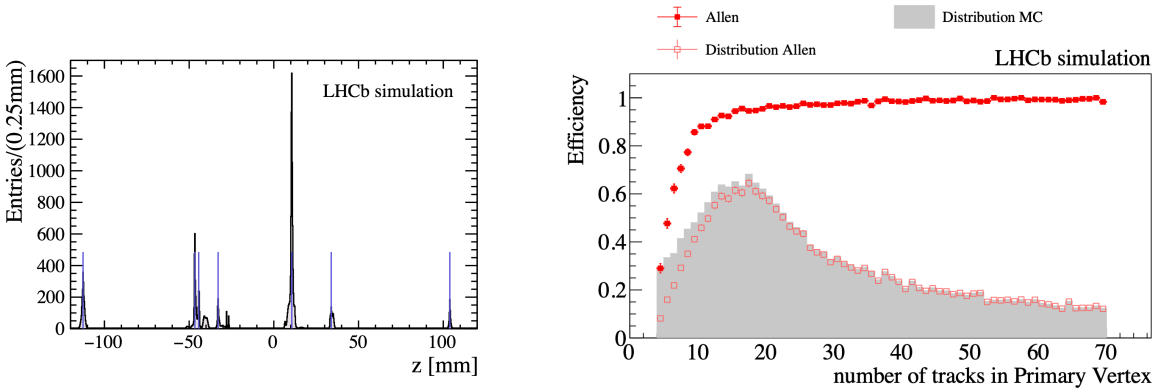


Figure 6: Velo tracking efficiency as a function of momentum and transverse momentum (p_T). [8]

2.2 Primary Vertex finding

One of the physics requirement in HLT1 is to reconstruct all primary vertices (PVs) from proton-proton collisions. VELO tracks can be extrapolated as straight lines to the beamline, and PVs can be identified by looking at the z position of the intersection between the track and the beamline. Figure 7 (a) shows a histogram of all VELO tracks' point of closest approach inside a single event. Peaks can be then isolated and fitted using Gaussian density distributions. These density distributions take the uncertainty on the track states into account. The primary vertex reconstruction efficiency as a function of VELO tracks associated to each PV is presented in Figure 7 (b).



(a) (b)

Figure 7: Primary vertex clustering (a) and efficiency as a function of number of associated tracks (b). [10]

2.3 Velo-UT tracking

The UT subdetector is composed of 4 silicon strip module planes. The first and last plane, labeled as X, are composed of vertical modules while the two middle planes, labeled as U and V, contains modules tilted by $+5°$ and $-5°$ respectively. By combining the information from the U/V and X layers, it is possible to determine the Y coordinates of the hits. The UT subdetector allows to reconstruct charged particles which decay after the VELO and low momentum tracks bending out of the magnetic field region, not reaching the SciFi. Moreover, it allows a first momentum and charge estimate. VELO-UT tracks, because of their still poor momentum resolution, are not mainly used for physics selections, but are rather inputs to the Forward algorithm.

VELO tracks are extrapolated to the UT planes taking into account the small magnetic field present in the region. Hits are searched in the planes assuming a minimum momentum which defines the search windows size. To perform the UT tracking, a 3-hit search is carried out following the VELO track extrapolation and allowing for one missing hit [11]. If more than one tracklet per VELO input track is found, a χ^2 fit is performed and the tracklet with the lowest χ^2 is selected. The algorithm is parellalized over input VELO tracks and also over tracklet candidates per input track. The reconstruction efficiency as a function of momentum and transverse momentum is shown in Figure 8.

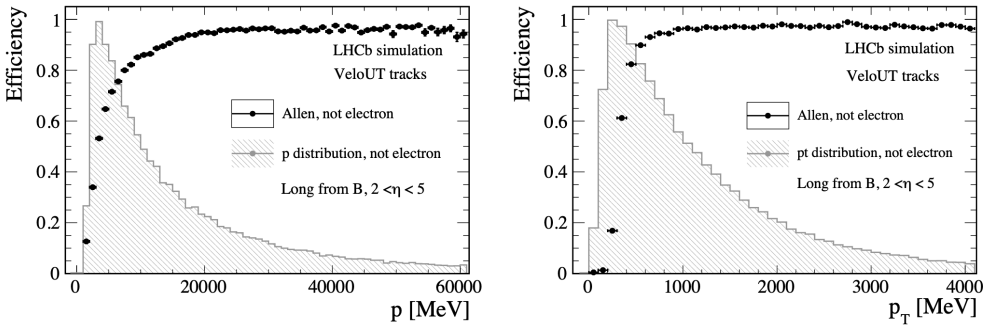


Figure 8: Velo-UT tracking efficiency as a function of momentum and transverse momentum (p_T).

2.4 Forward tracking

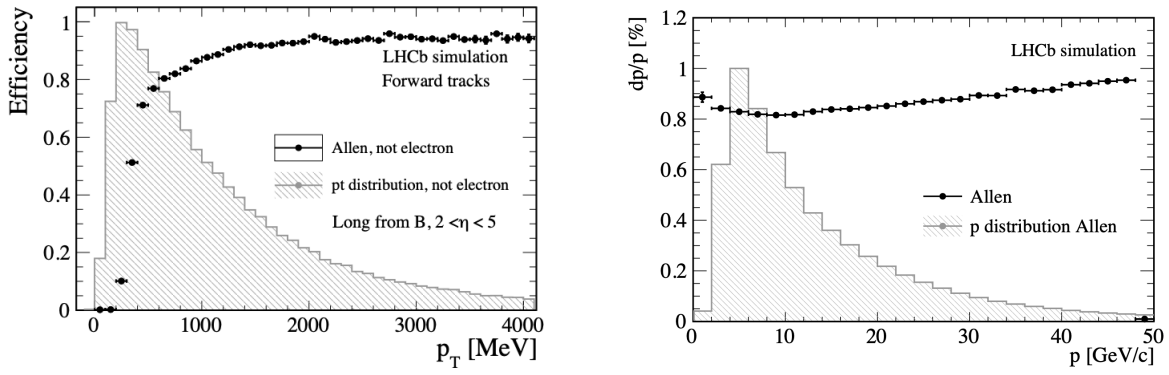
For the LHCb's physics purposes, HLT1 needs to be able to reconstruct long tracks with 500 MeV p_T requirement and a 3 GeV momentum requirement, and evaluate their momenta with a resolution better than 1 %. The SciFi tracker, located after the dipole magnet, allows the reconstruction of these long tracks thanks to its 12 scintillating fibers layers. They are divided into 3 T stations with 4 layers each, labeled as X-U-V-X. X layers are vertical while U and V layers are tilted by $\pm 5°$, respectively, providing in combination also Y information to the hits.

The forward tracking algorithm takes as input previously reconstructed VELO-UT tracks and parallelizes their extrapolation inside the GPU. The input tracks have already a first momentum and charge estimate from the VELO-UT algorithm. A parametrization of the magnetic field is used to estimate their trajectories extrapolating them to each of the 12 layers of the SciFi. The size of the search windows is decided based on the first momentum estimate from the VELO-UT algorithm.

A selected collection of hits is saved for each SciFi layer and for each input track. Triplets of hits are searched for, either looking at the first X layer in each T-station (first seed) or the last X layer in each T-station (second seed). Performing combinations in only 2 seeds is found to be the optimal, balancing between physics and speed performances of the algorithm. Triplets are computed first combining all hits in the first and last T-station search window (doublet), then adding a hit in the central T station estimated from the corresponding doublet.

To achieve a better throughput performance, the triplet search is performed using a maximum of 32 central hits inside the search window. A second level of parellelization is set on the triplet candidates. Theχ^2 of trajectories can be evaluated at this stage to avoid processing triplets that are unambiguously wrong (with very high χ^2). Matching a VELO-UT track with 3 hits in the SciFi provides a well defined trajectory which can be determined by a 4 parameter equation [13]. The triplet candidates are then extended over the other 3 X layers and 6 U/V layers (using Y information from the VELO-UT slope) requiring a minimum of 9 hits.

As a final step, all remaining tracklets in combination with their respective VELO-UT input track are fitted and the candidate with the lowest χ^2 is selected. The reconstruction efficiency is shown in Figure 9 (a) and plateaus at 90 % for tracks with $p_T > 1$ GeV. The final candidate fit provides also the momentum estimate, shown in Figure 9 (b), which has a resolution below 1 %.



(a) (b)

Figure 9: Forward tracking efficiency (a) as a function of transverse momentum (p_T) and track momentum resolution as a function of momentum (b).

2.5 Forward tracking without UT

For the beginning of the 2022 data taking, the Upstream Tracker is not yet available. Therefore, the forward tracking should be adapted to reconstruct tracks using only VELO and SciFi information. As mentioned in the previous sections, the UT provides a first momentum estimate and a charge measurement of particles.

This is very important to reduce the search windows size and, as a consequence, the number of triplet combinations. To perform the forward tracking using VELO tracks as input only, a minimum momentum is chosen to limit the search window size. For this algorithm, the focus is on long tracks with momentum greater than 5 GeV and transverse momentum greater than 1 GeV. Both charge assumptions are tested opening double-sided search windows in the X direction, as charged particles are bent by the dipole magnet only in the XZ plane (see Figure 10).

VELO input tracks are extrapolated as a straight lines to the SciFi layers to find a central point from which two search windows are opened, representing the two charge assumptions. A small 1 cm overlap between the two windows avoids to lose efficiency for very high momentum tracks.

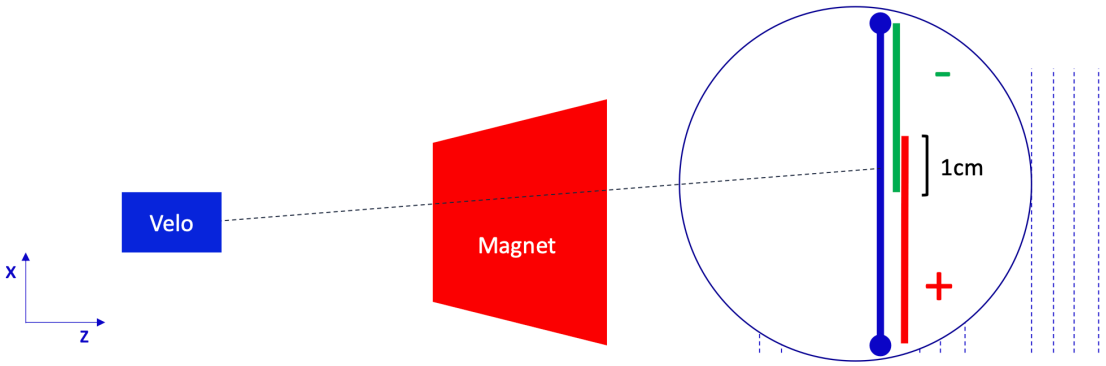
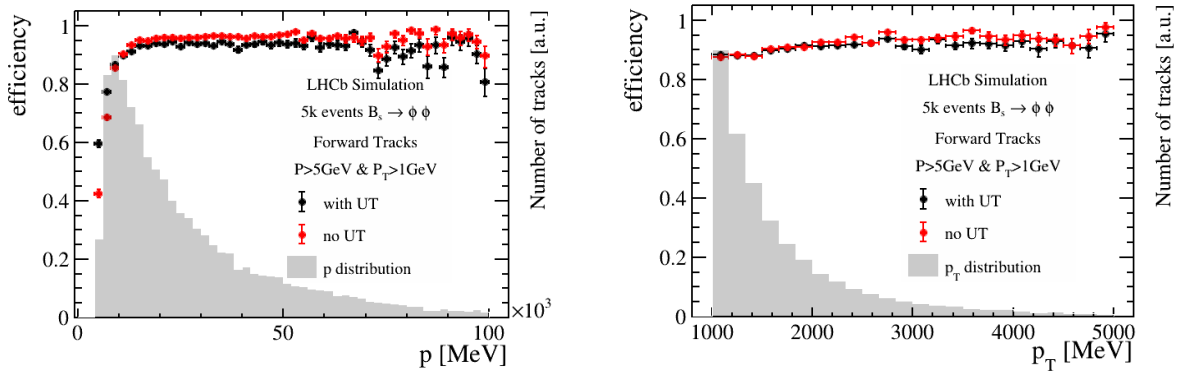


Figure 10: Forward tracking without UT scheme. The double-sided search window strategy allows to test both charge particle assumptions.

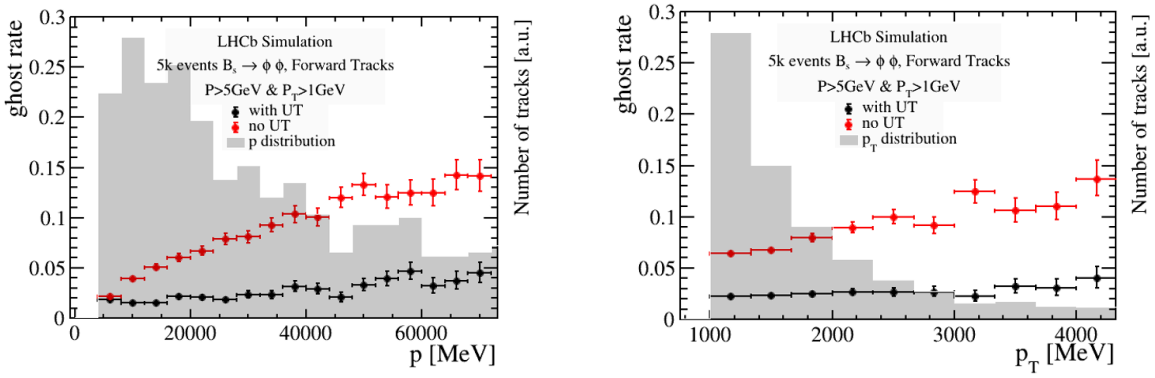
Triplets are then formed with the same technique explained in section 2.4, combining hits in search windows with the same charge and seed assumptions. The next steps of the algorithm remain unchanged.

The goal of the algorithm is to keep physics efficiency and throughput comparable to the forward tracking "with UT", considering the smaller subset of tracks of interest (P > 5 GeV and p_T > 1 GeV). This comparison of reconstruction efficiencies as a function of momentum and transverse momentum is shown in Figure 11. The throughput is going to be discussed in the next section 3. The greater number of triplet combinations causes an increase of the ghost rate compared to the forward tracking version with UT (see Figure 12).



(a) (b)

Figure 11: Forward tracking efficiency (a) as a function of momentum and (b) transverse momentum (p_T).



(a) (b)

Figure 12: Forward tracking ghost rate (a) as a function of momentum and (b) transverse momentum (p_T).

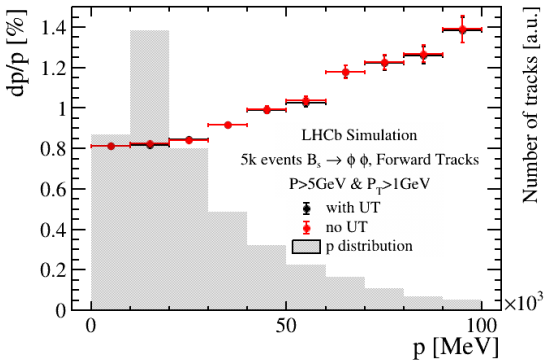


Figure 13: Forward tracking momentum resolution as a function of momentum.

2.6 Kalman fit

Performing a Kalman fit on the reconstructed tracks improves the estimate of important physical parameters, as momentum and track's impact parameter. LHCb's nominal Kalman filter uses a Runge-Kutta extrapolator with detailed detector description to precisely estimate noise due to multiple scattering and energy loss. To increase throughput and limit memory overhead, these calculations can be replaced by parametrizations. The default HLT1 Kalman fit features a VELO-only parametrization, performing a backward VELO track fit using the momentum estimate from the forward tracking. The impact parameter resolution and χ^2 for various implementations of the Kalman fit are shown in Figure 14.

3 Throughput

Throughput is a crucial topic when discussing the HLT1 performance. The first high level trigger at LHCb needs to take care of an 30 MHz input rate, which must be reduced by a factor 30 to proceed to HLT2. The Allen throughput per different CPU or GPU cards is presented in Figure 15. The chart shows performances with the default "with UT" sequence and a sequence where no UT information is used (i.e. no VELO-UT tracking and Forward tracking with no UT). The NVIDIA RTX A5000 was chosen as the default GPU card for data taking. The results are comparable between the "with UT" and "no UT" sequences, keeping in mind that the "no UT" sequence only reconstruct a subset of all long tracks (P $>$ 5 GeV and $p_T > 1$ GeV). These results are very promising as, considering a single GPU throughput of around 170 kHz for the A5000

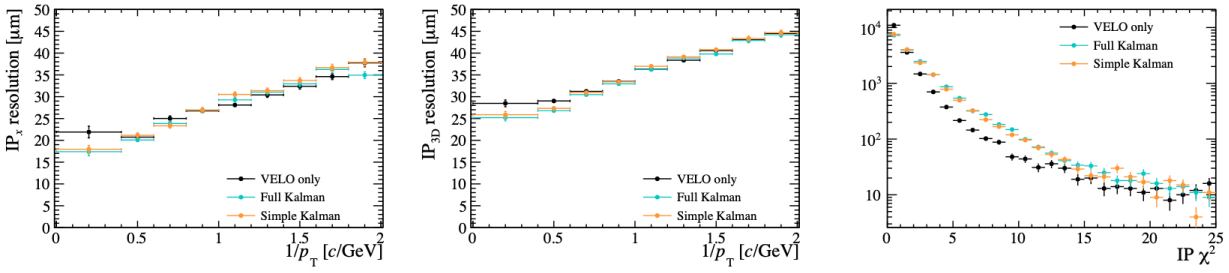


Figure 14: Impact parameter resolution and χ^2 for different implementations of the Kalman fit at HLT1. The default solution is shown in black and features a Velo-only Kalman fit taking the momentum estimate from the forward tracking. [4]

card, about 175 GPUs are needed to process the 30 MHz input rate. LHCb has 500 GPU slots available, which leaves a good amount of space for future improvements of the whole system. Allen in the first days of Run 3 is presented in Ref. [14].

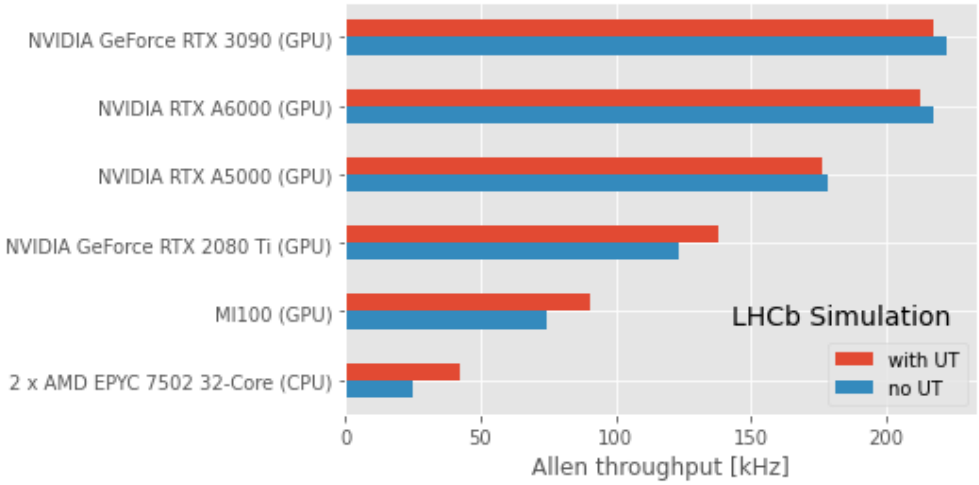


Figure 15: HLT1 throughput comparison between full sequence and sequence without UT information.

4 Conclusions

The goal of LHCb's HLT1 is to reconstruct events at the LHC inelastic rate of 30 MHz. GPU technology helps performing this task by parallelizing the reconstruction over single independent events and over tracks. These proceedings focused on presenting the HLT1 tracking algorithms together with their performances. A customized algorithm presented for the first time in this document, allows performing forward tracking without UT information. Around 175 GPUs are needed to take care of the LHC input collision rate and reduce it by a factor 30 for the next steps of the data flow. All the aforementioned characteristics allow HLT1 to be ready to face the next LHCb's data taking with good expected physics perfomance.

ACKNOWLEDGEMENTS

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 724777 "RECEPT". The author would like to acknowledge the support of the LHCb collaboration by thanking the LHCb computing, Online, RTA and simulation teams for their support and for producing the simulated LHCb samples used to develop and benchmark our algorithms.

References

[1] R. Aaij *et al.* [LHCb], "LHCb Detector Performance," Int. J. Mod. Phys. A **30** (2015) no.07, 1530022 doi:10.1142/S0217751X15300227 [arXiv:1412.6352 [hep-ex]].

[2] A. A. Alves, Jr. *et al.* [LHCb], "The LHCb Detector at the LHC," JINST **3** (2008), S08005 doi:10.1088/1748-0221/3/08/S08005

[3] I. Bediaga *et al.* [LHCb], "Framework TDR for the LHCb Upgrade: Technical Design Report," CERN-LHCC-2012-007.

[4] LHCb Collaboration, "LHCb Upgrade GPU High Level Trigger Technical Design Report," (2020) doi:10.17181/CERN.QDVA.5PIR

[5] A. Piucci, "The LHCb Upgrade," J. Phys. Conf. Ser. **878** (2017) no.1, 012012 doi:10.1088/1742-6596/878/1/012012

[6] LHCb Collaboration, "RTA and DPA dataflow diagrams for Run 1, Run 2, and the upgraded LHCb detector," (2020), https://cds.cern.ch/record/2730181

[7] R. Aaij *et al.* [LHCb], "A Comparison of CPU and GPU Implementations for the LHCb Experiment Run 3 Trigger," Comput. Softw. Big Sci. **6** (2022) no.1, 1 doi:10.1007/s41781-021-00070-2 [arXiv:2105.04031 [physics.ins-det]].

[8] LHCb Collaboration, "Performance of the GPU HLT1 Allen," (2020), https://cds.cern.ch/record/2722327

[9] Cámpora Pérez, Daniel Hugo and Neufeld, Niko and Riscos Nuñez, Agustin, "Search by triplet: An efficient local track reconstruction algorithm for parallel architectures," Journal of Computational Science **54** (2021) doi:10.1016/j.jocs.2021.101422

[10] LHCb Collaboration, "Run3 Primary Vertex reconstruction description and performance," (2020), https://cds.cern.ch/record/2714851

[11] Fernandez Declara, Placido and Cámpora Pérez, Daniel Hugo and Garcia-Blas, Javier and Vom Bruch, Dorothea and Daniel García, J. and Neufeld, Niko, "A Parallel-Computing Algorithm for High-Energy Physics Particle Tracking and Decoding Using GPU Architectures," IEEE Access **7** (2019), 91612-91626 doi:10.1109/ACCESS.2019.2927261

[12] LHCb Collaboration, "HLT1 forward tracking performance," (2022) https://cds.cern.ch/record/2810803

[13] Quagliani, Renato, "Study of double charm B decays with the LHCb experiment at CERN and track reconstruction for the LHCb upgrade," (2017) https://cds.cern.ch/record/2296404

[14] T. Boettcher, "Allen in the first days of Run 3," PROC-CTD2022-33.