Real-time analysis, alignment and calibration at LHCb in Run 3

Kate Richardson on behalf of the LHCb collaboration

Large Hadron Collider Physics Conference, Boston

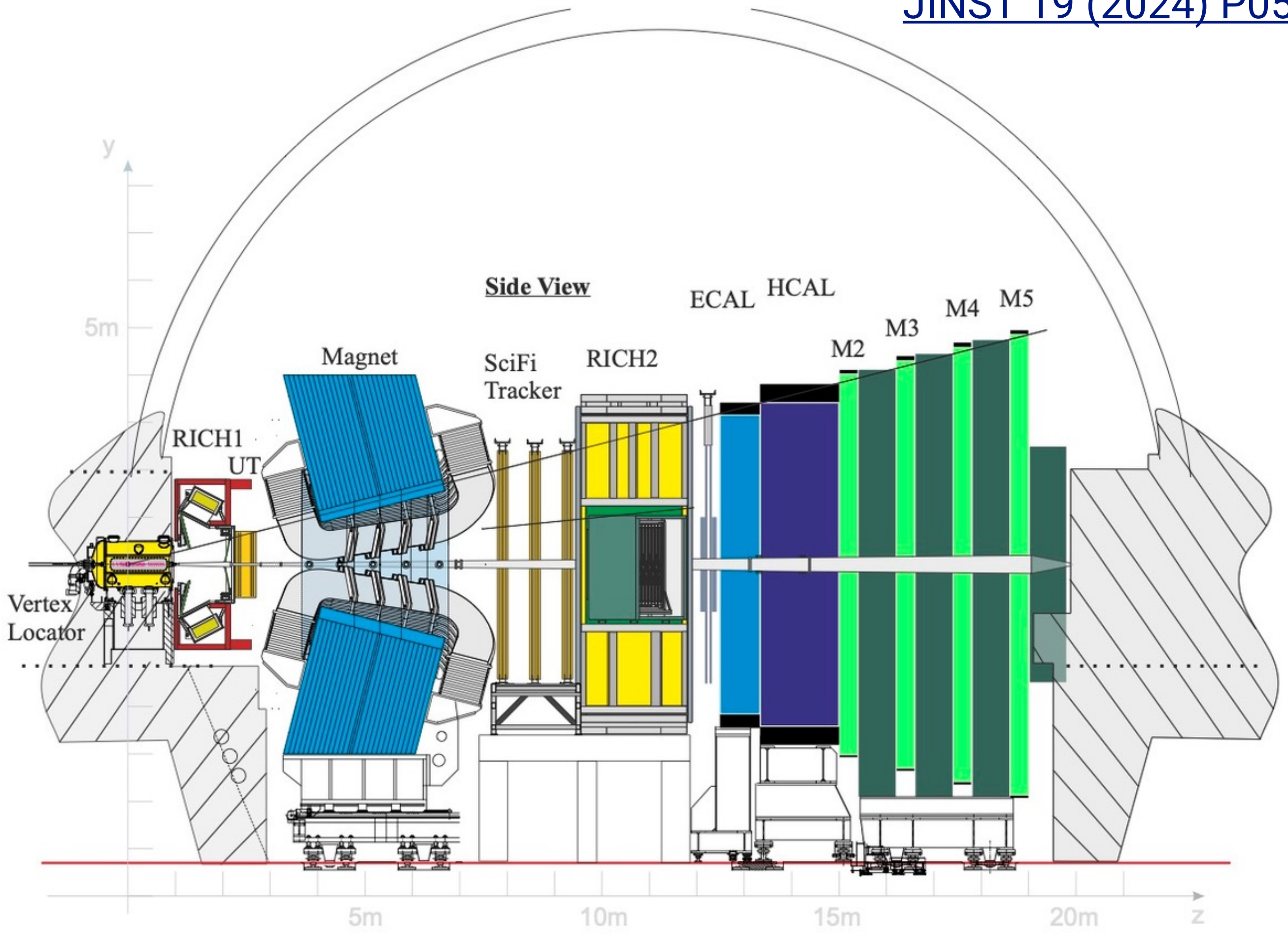
7 June 2024

Massachusetts Institute of Technology

The upgraded LHCb detector

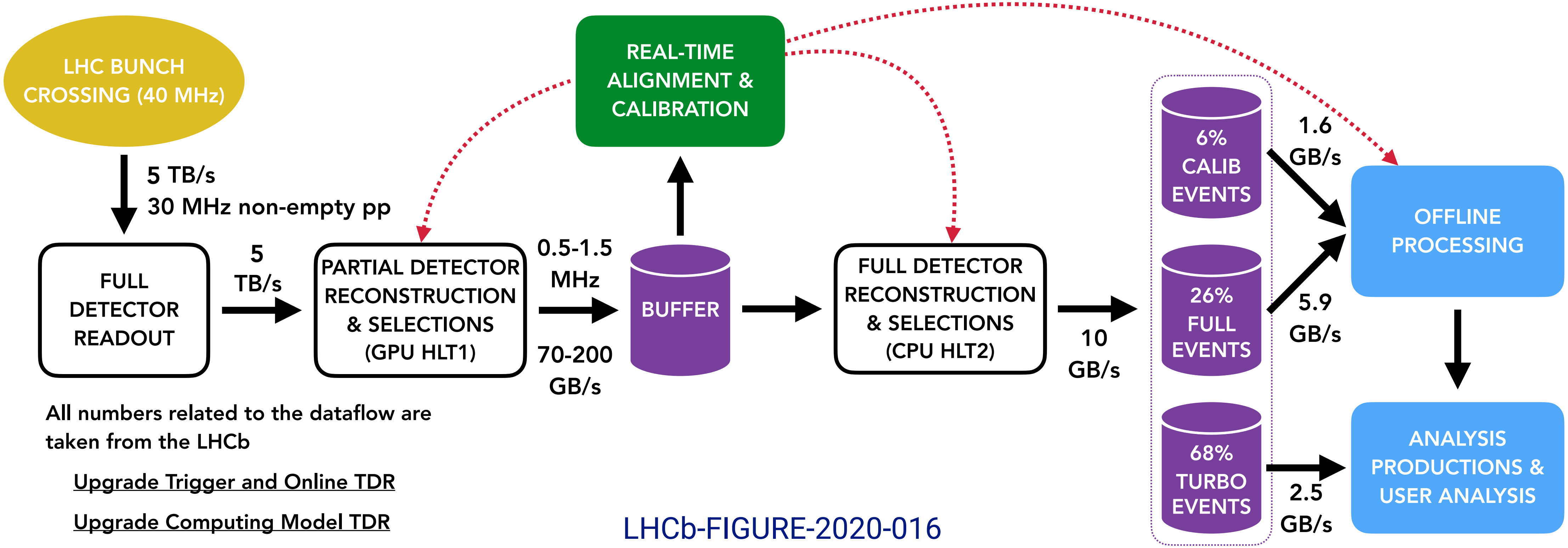
- Increased instantaneous luminosity $5\times$ from Run 2 to 2×10^{33}cm^{-2}s^{-1}
- New tracking detectors
- Improved readout electronics to meet rate requirements
- **No hardware trigger**

JINST 19 (2024) P05065



Dataflow in LHCb Run 3

- Software-only trigger allows more flexible selections



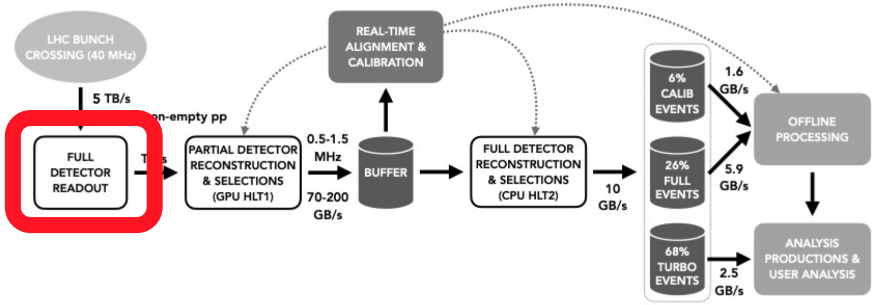
All numbers related to the dataflow are taken from the LHCb

Upgrade Trigger and Online TDR

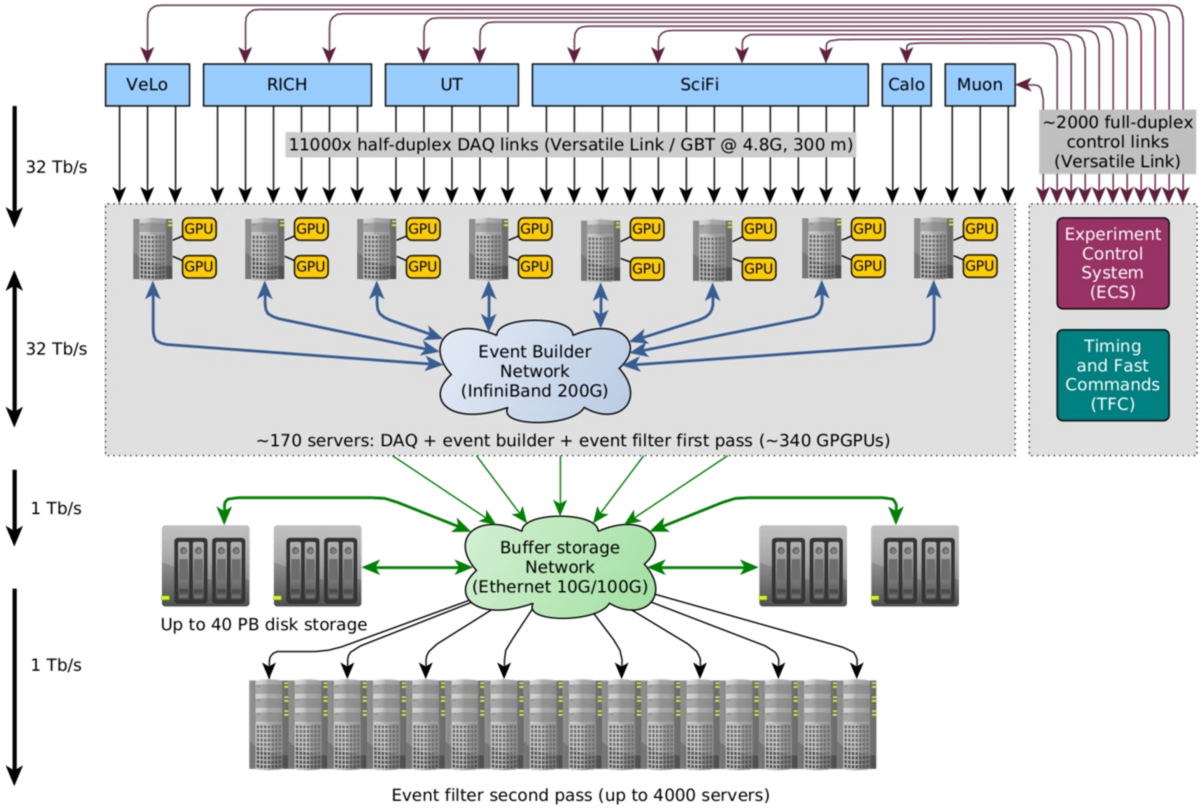
Upgrade Computing Model TDR

LHCb-FIGURE-2020-016

DAQ architecture

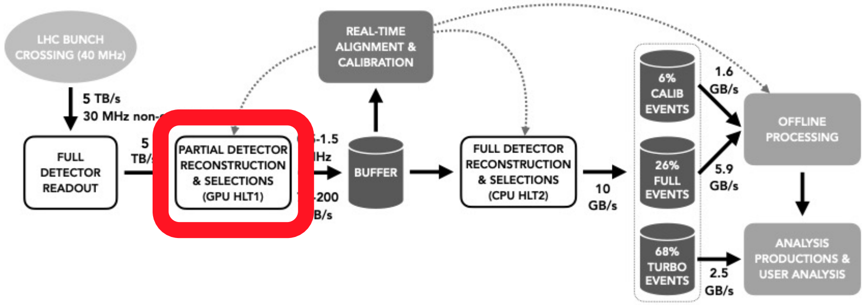


- $\mathcal{O}(500)$ FPGA readout boards receive data from subdetectors at 30 MHz
- Event builder units reorder raw data from front-end boards into event packets to be processed by HLT1
- Throughput of 5 TB/s
- Off-the-shelf components reduce cost

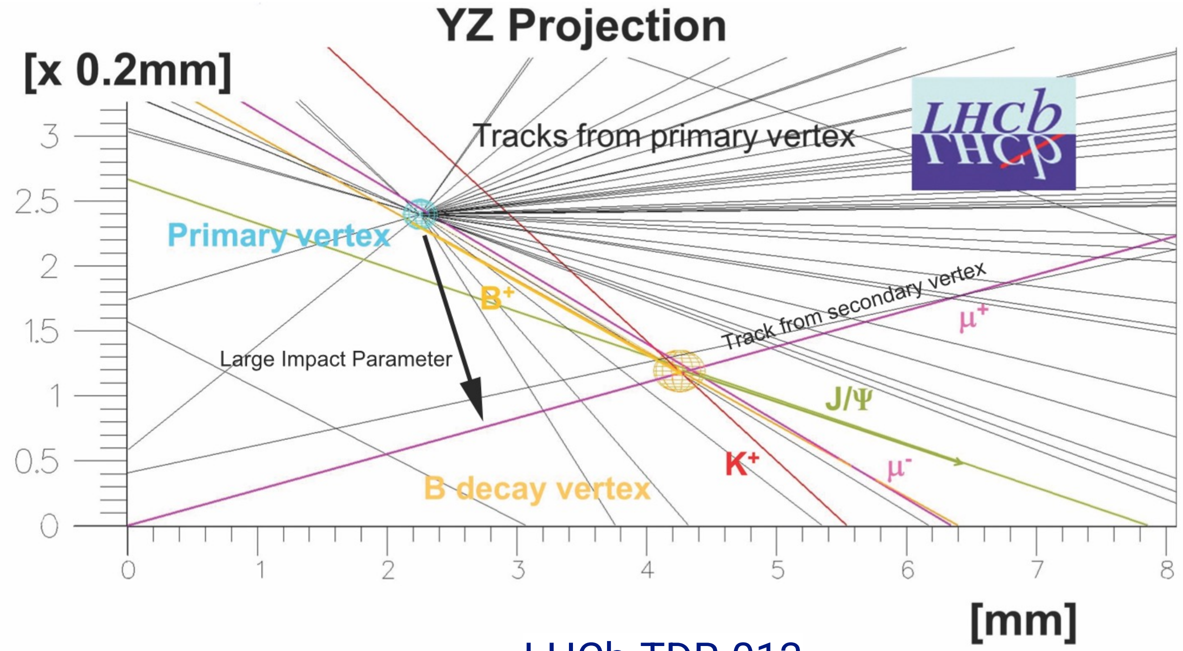


JINST 19 (2024) P05065

HLT1 requirements

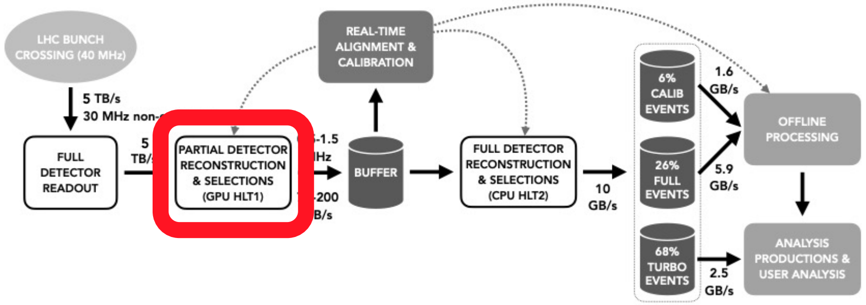


- To perform selections at HLT1 level, we need:
  - Subdetector reconstruction for VELO, UT, SciFi, ECAL, and MUON
  - Primary and secondary vertex reconstruction
  - Track fitting
  - Electron and muon PID
- All at a 30 MHz rate!



LHCb-TDR-013

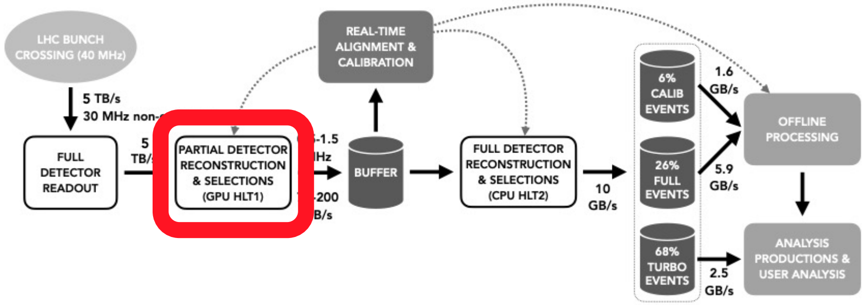
Why GPUs?



- HLT1 is an inherently parallelizable task, at multiple levels
  - We run multiple streams on each GPU, each with a slice of events
  - Within algorithms, threads are used to parallelize over objects (vertices, tracks, etc.)
- Limited I/O bandwidth is acceptable because small raw event data $\mathcal{O}$(100 kB/event) means thousands of events still fit in $\mathcal{O}$(10 GB) memory
- Cheaper and more scalable than CPU alternative
- Fit well into LHCb DAQ architecture
- Run with 323 Nvidia RTX A5000 GPUs

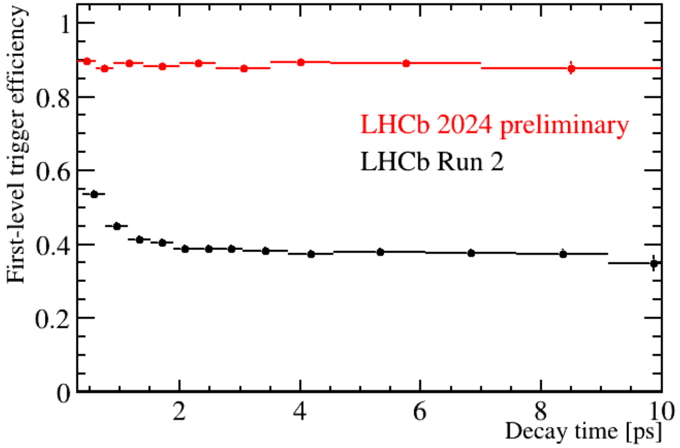
Comput. Softw. Big Sci. 4 (2020) 7

HLT1 performance



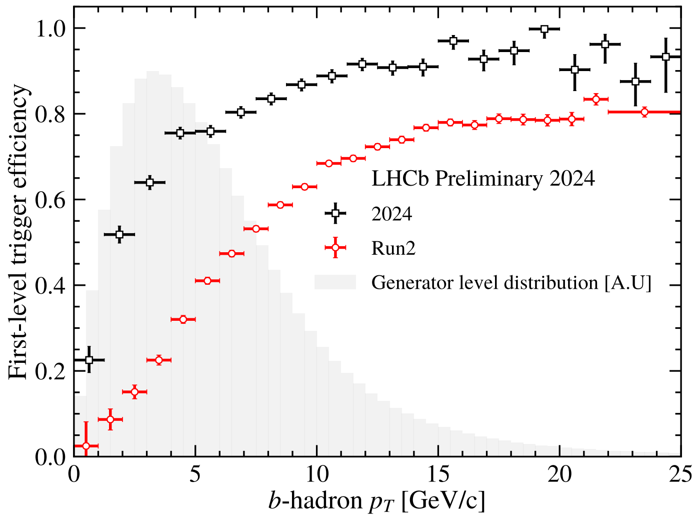
- Excellent track reconstruction efficiency, both with and without the UT
- Efficiencies equal or better compared to Run 2

B_d



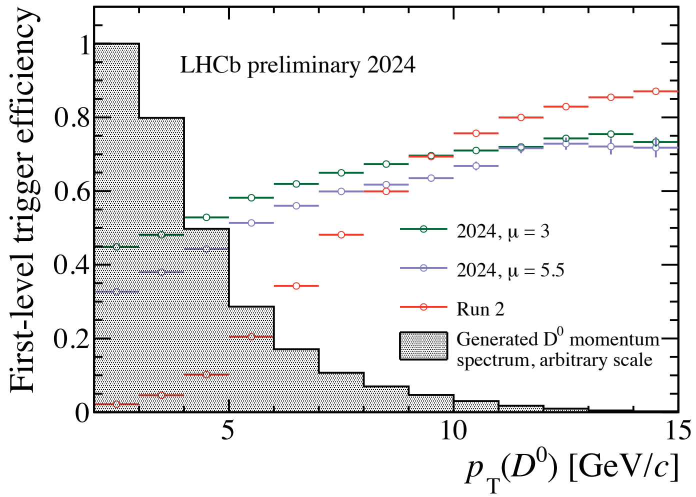
LHCb-FIGURE-2024-014

$B^{\pm} \to K^{\pm}e^{+}e^{-}$



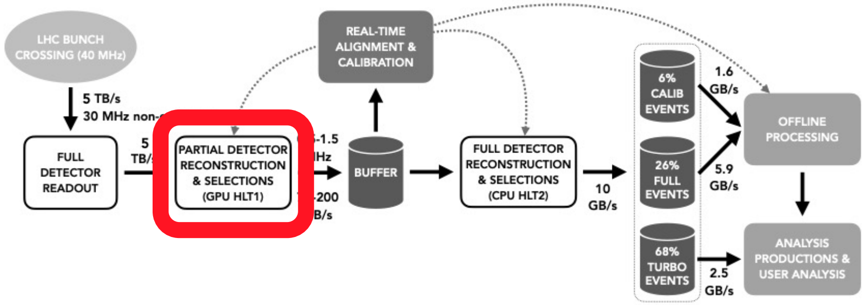
LHCb-FIGURE-2024-007

$D^0 \to K^{-}\pi^{+}$

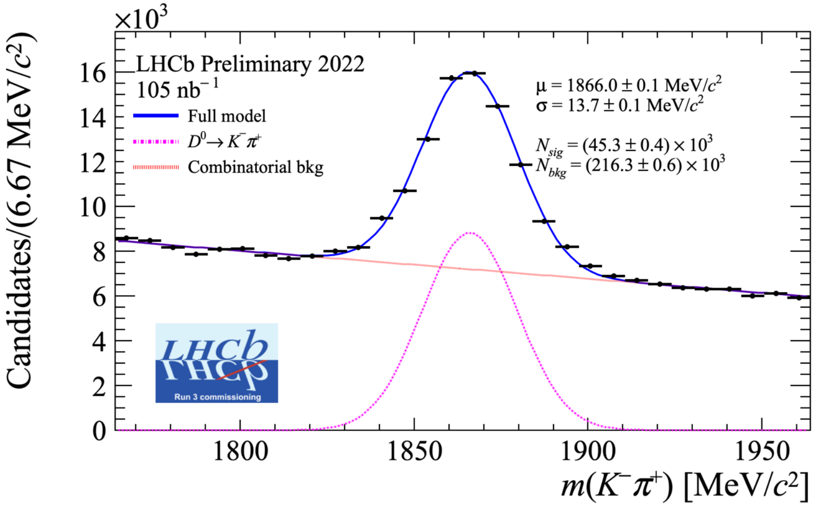


LHCb-FIGURE-2024-006

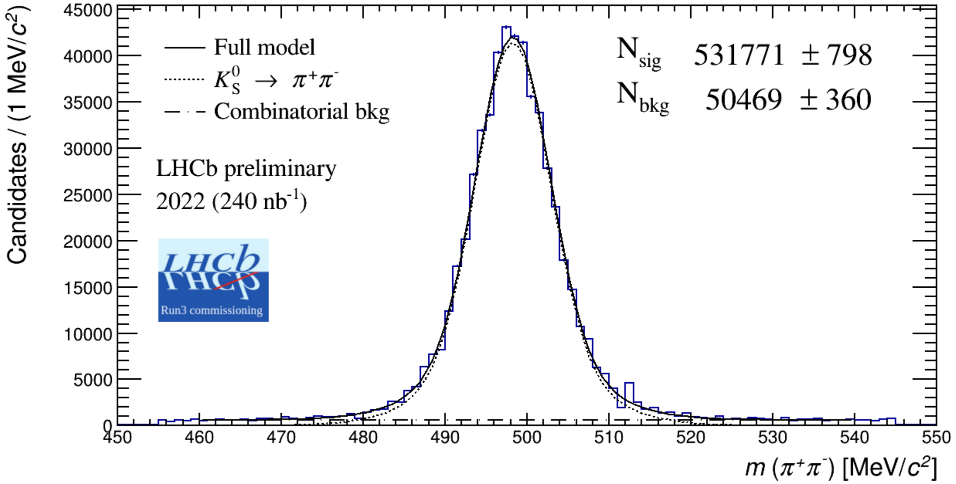
HLT1 online monitoring



- Monitoring is necessary to allow for real-time supervision of reconstruction and selections in HLT1 to find issues quickly
- LHCb has a monitoring infrastructure for aggregation and display but HLT1 monitoring buffers need to be periodically transferred to host
- Monitor all events at full 30 MHz rate → access to events discarded by HLT1

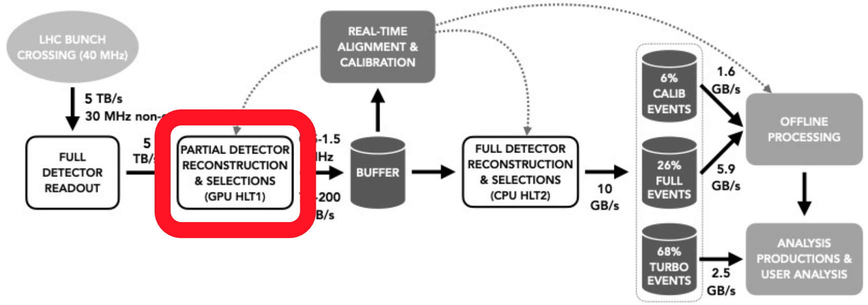


LHCb-FIGURE-2023-009



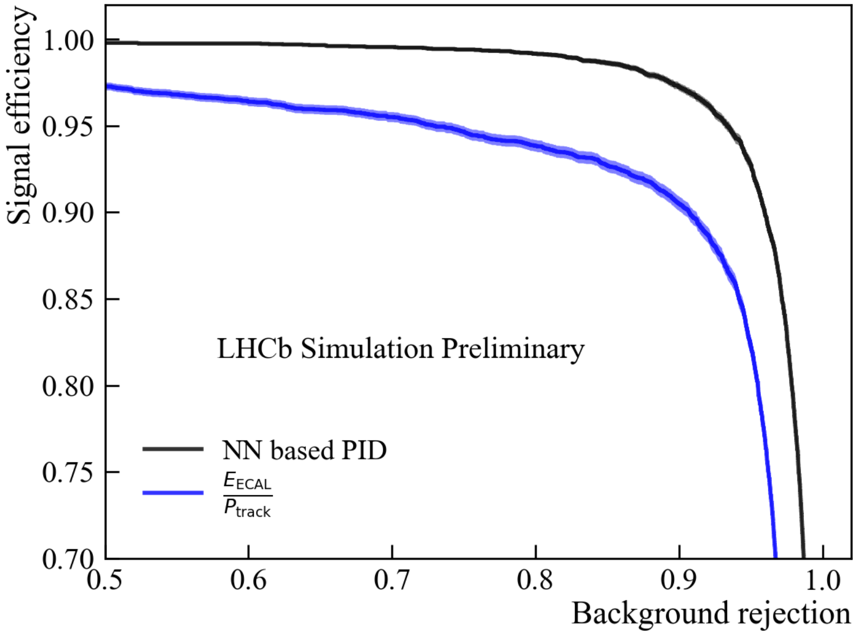
LHCb-FIGURE-2023-005

Machine learning in HLT1



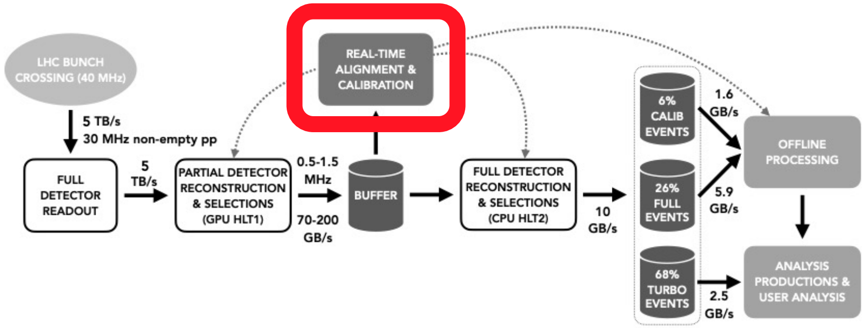
- Requirements of running in the HLT1 environment are that any model needs to be small and fast
- For physics we want our models to be robust and possibly monotonic
- Use Lipschitz neural networks to achieve this
- Currently in use for PID, selections, and more

Mach.Learn.Sci.Tech. 4 (2023) 3, 035020

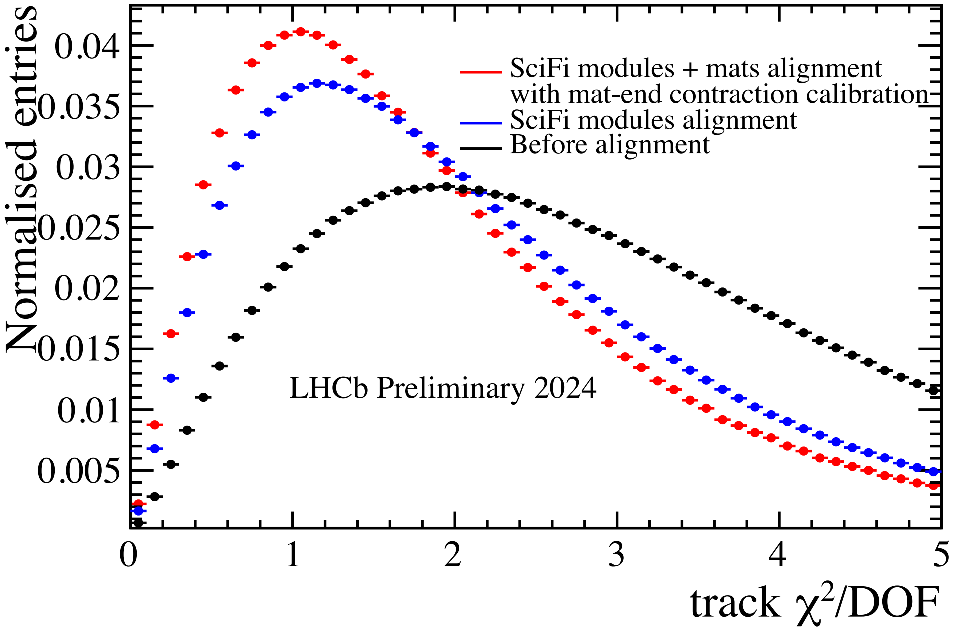


LHCb-FIGURE-2024-003

Alignment and calibration

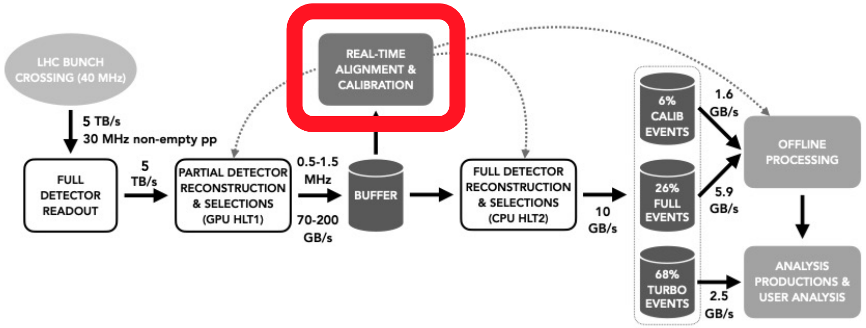


- Alignment for VELO, RICH mirrors, UT, SciFi, and MUON
- Calibration for RICH, ECAL, and HCAL
- Alignment process based on analyzer and iterator to obtain convergence
- Needs to be done in real time before HLT2 for best performance and turbo event model (more on this later)



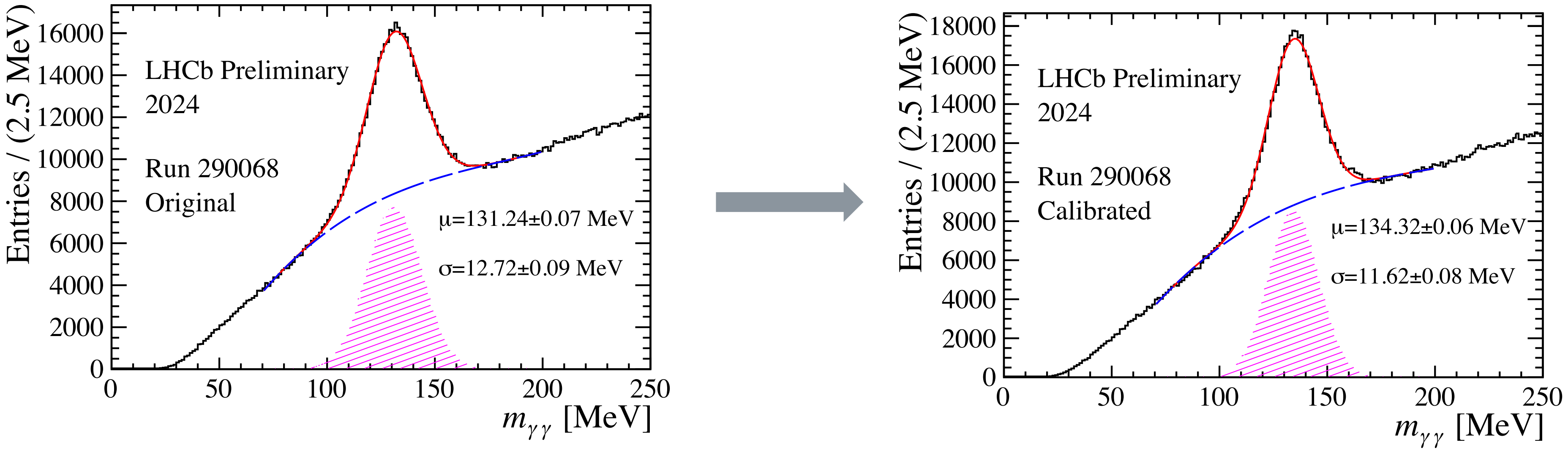
LHCb-FIGURE-2024-009

A&C monitoring

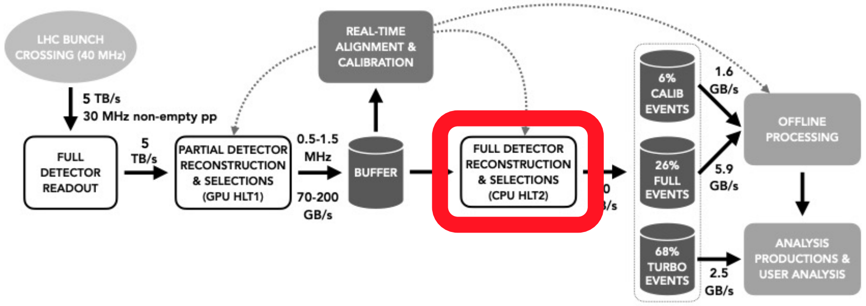


- Monitor alignment and calibration quantities in real-time to catch issues as fast as possible
- Perform HLT2 level reconstruction on subset of events to compare with HLT1 in real-time

LHCb-FIGURE-2024-009



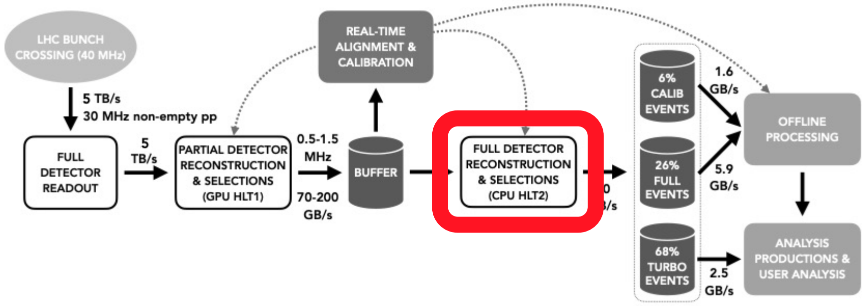
HLT2 requirements



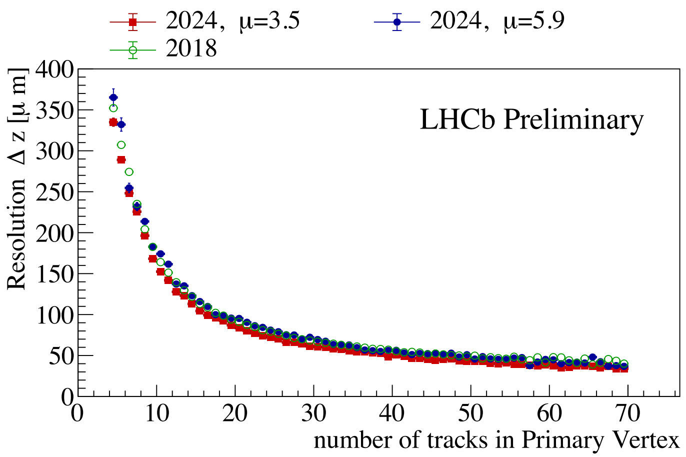
- Full, offline-quality reconstruction at 500 kHz
  - Complete detector decoding, track fit, and PID
- $\mathcal{O}(2700)$ selections written by analysts cover entire LHCb physics program
- To meet rate requirements several improvements were made for Run 3:
  - Use structure of arrays memory to vectorize tracking algorithms
  - Functors (function objects) are designed to be agnostic to input and output type
  - Compile a functor cache instead of just-in-time compilation
  - Event scheduler handles data dependencies and composite nodes (AND, OR, etc.)

CERN-THESIS-2020-331 J. Phys.: Conf. Ser. 1525 (2020) 012052

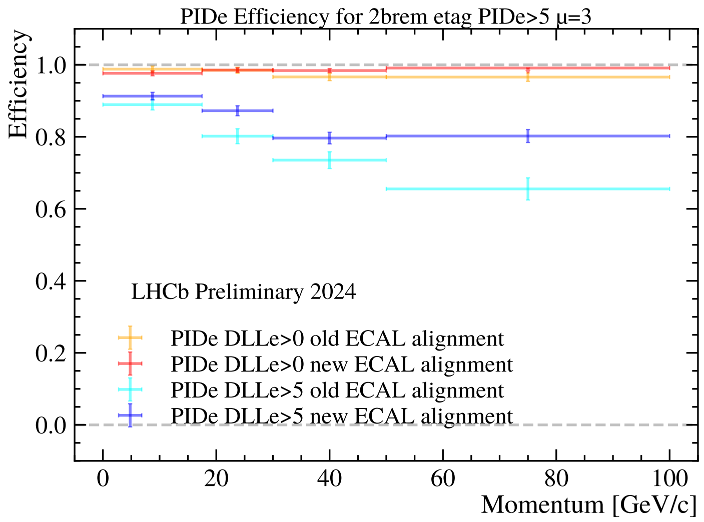
HLT2 performance

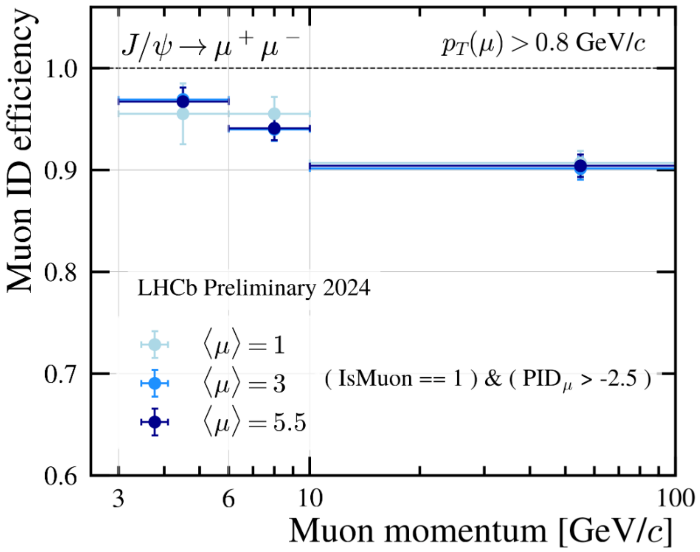


- Tracking algorithm has excellent momentum resolution
- Efficient PID, neutral reconstruction, and selections



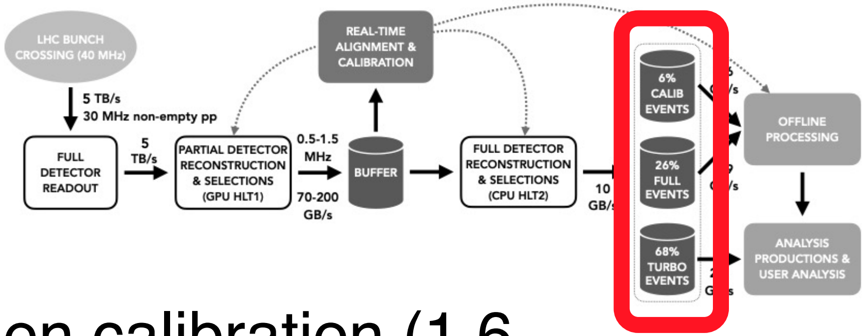
LHCb-FIGURE-2024-011





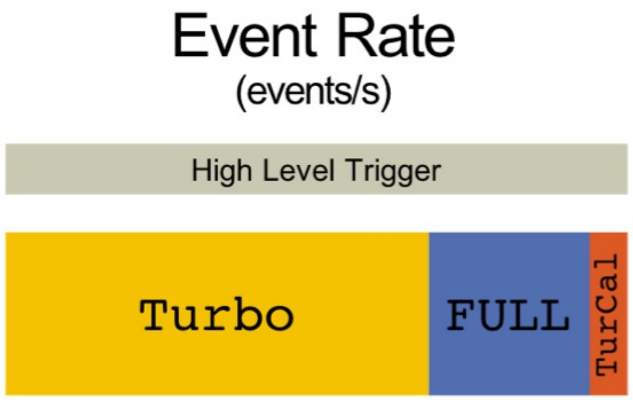
LHCb-FIGURE-2024-010

HLT2 persistency model



- Maximum output bandwidth of 10 GB/s split between calibration (1.6 GB/s), full (5.8 GB/s), and turbo (2.5 GB/s)
- Turbo model allows full flexibility in what objects are persisted
  - Reduces average event size leading to increase in number of events able to be selected (bandwidth [GB/s] = average event size [GB] x rate [Hz])
  - Baseline for Run 3 – approximately 70% of events are turbo

JINST 14 (2019) P04006



LHCb-TDR-018

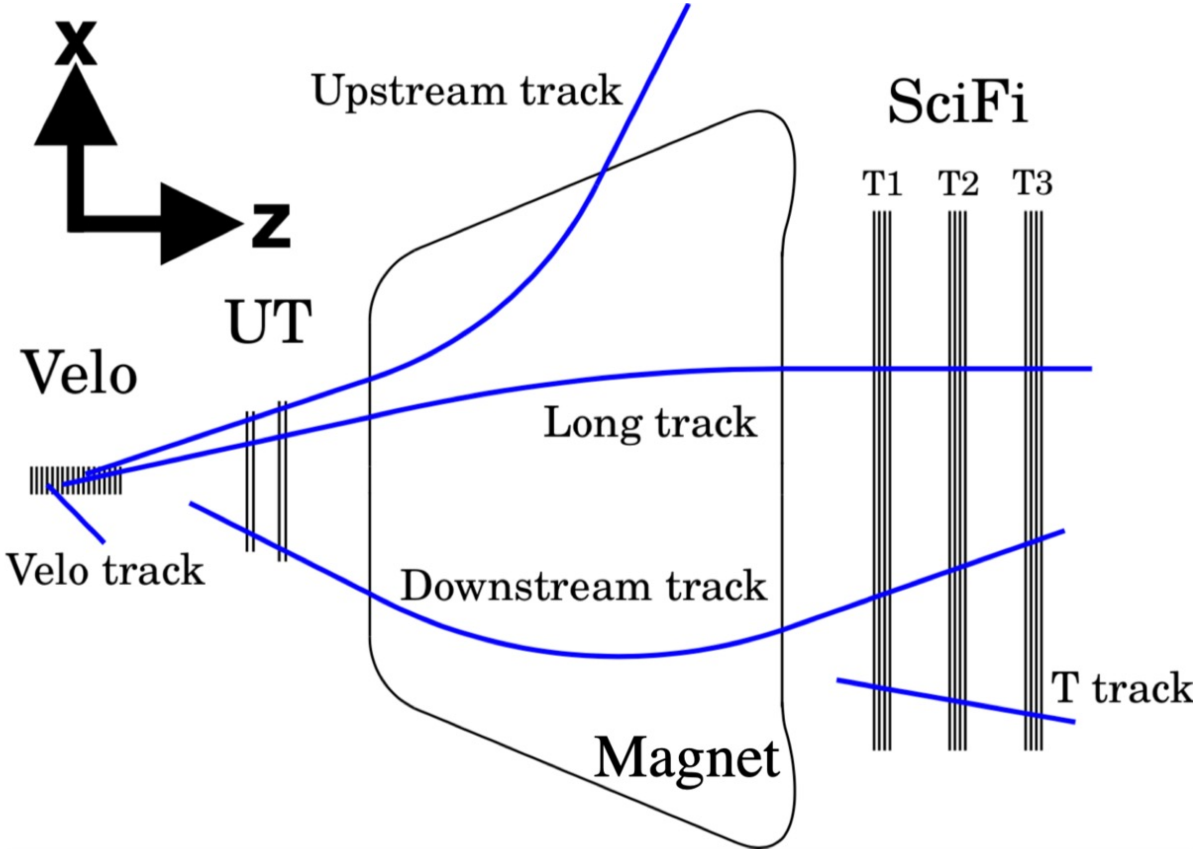
Summary

- LHCb is running a completely software-based trigger at 30 MHz (5 TB/s)
- HLT1 completes a partial reconstruction and selection at the full event rate by utilizing parallelization on GPUs
- Real-time alignment and calibration allows HLT2 to complete an offline-quality reconstruction to meet rate and turbo requirements
- HLT2 developments have met the rate challenge and the turbo model allows for higher statistics within a set bandwidth

Thank you for listening!

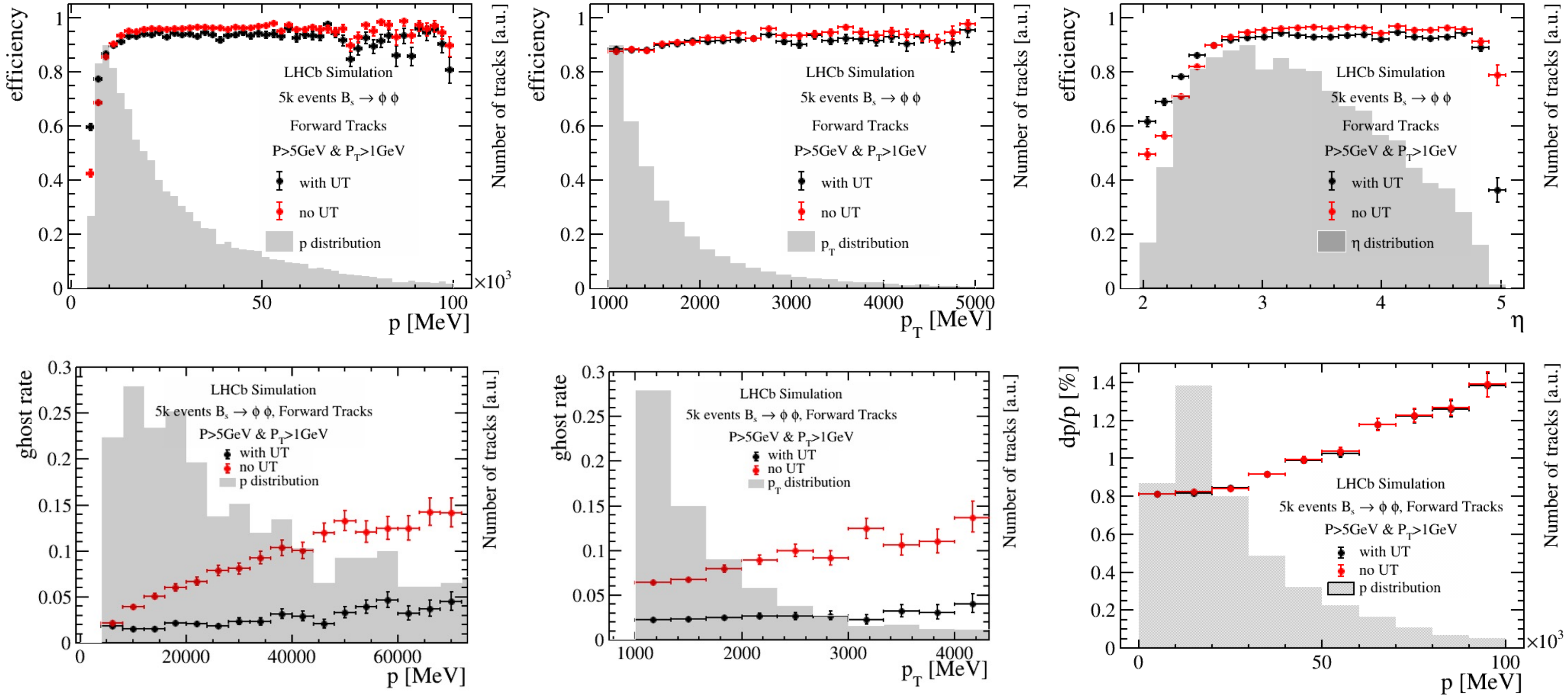
Backup

Types of tracks in LHCb



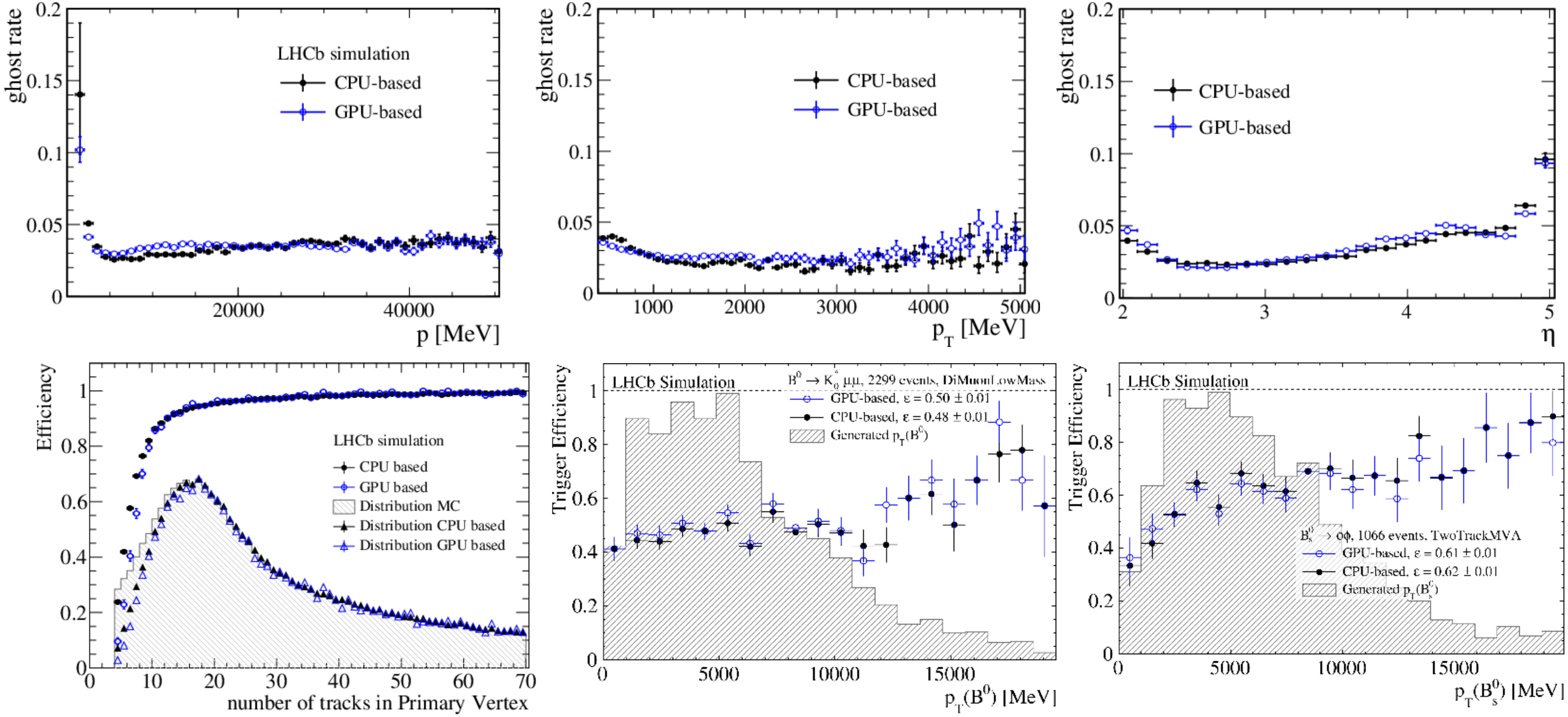
LHCb-DP-2022-002

HLT1 tracking efficiencies



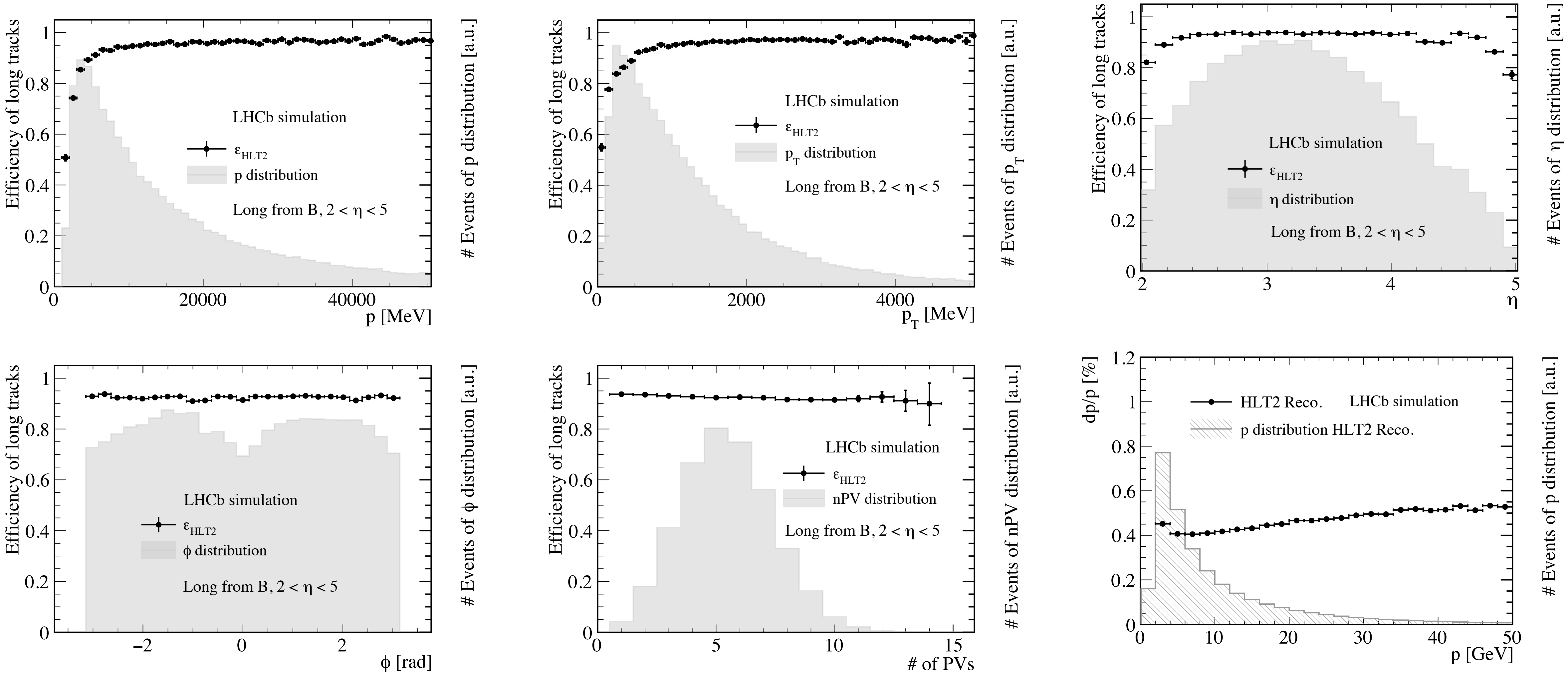
LHCb-FIGURE-2022-007

HLT1 CPU vs GPU performance



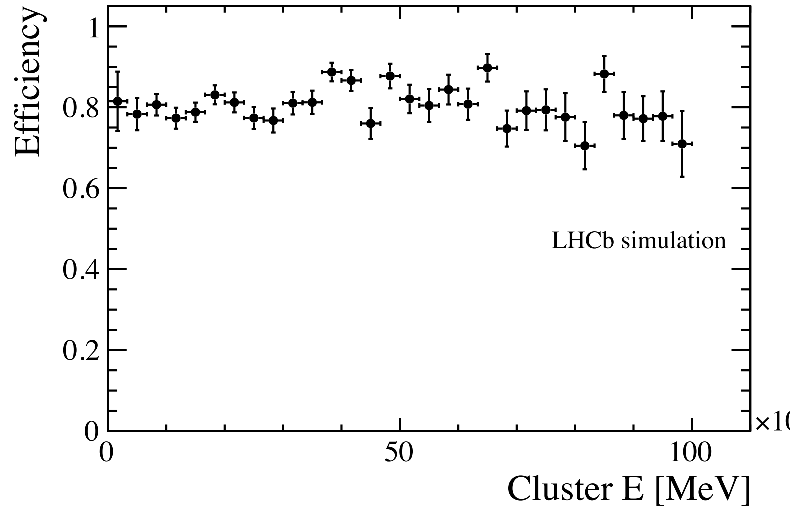
 Comput. Softw. Big Sci. 6, 1 (2022) pp.1

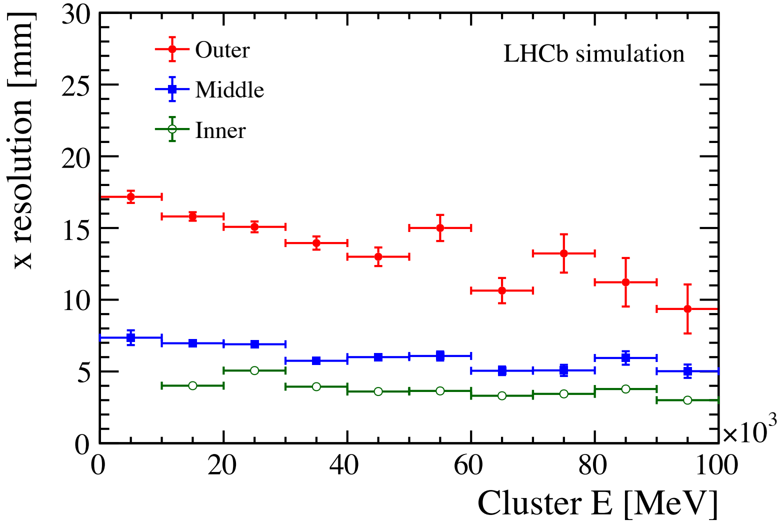
HLT2 tracking efficiencies

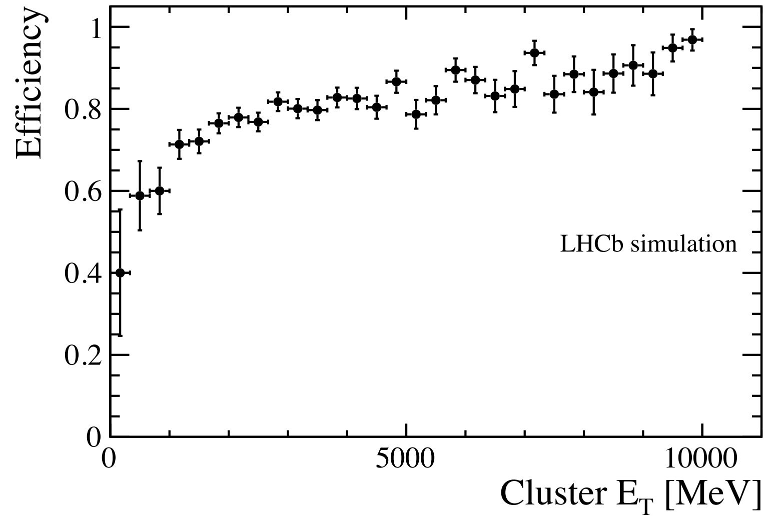


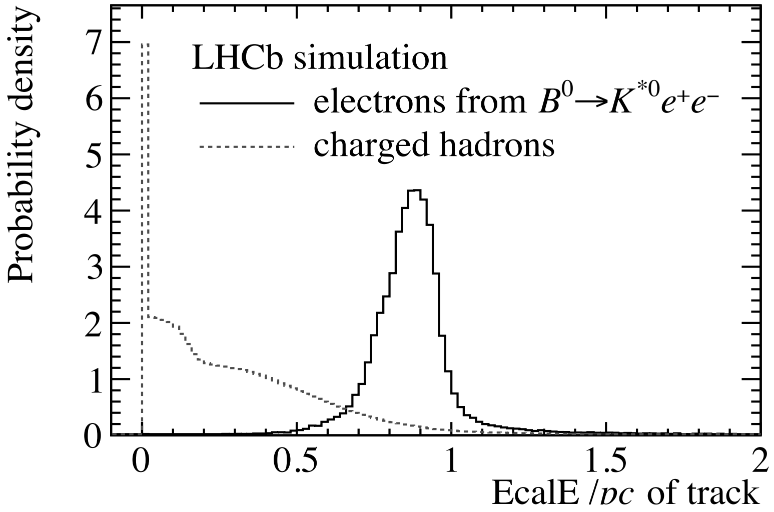
LHCb-FIGURE-2021-003

HLT2 ECAL efficiencies









LHCb-FIGURE-2021-003

HLT2 ghost rates

LHCb-FIGURE-2021-003