# Allen: a GPU trigger for LHCb
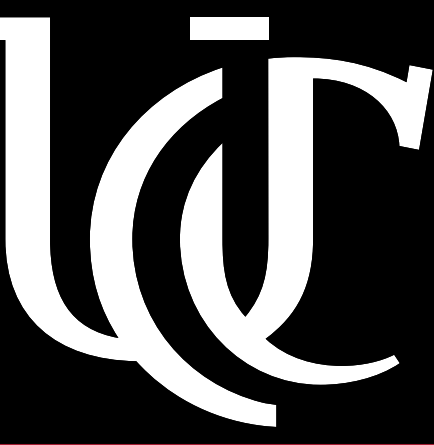
T. Boettcher[1]*, D. C. Craik[1], M. Sokoloff[2], M. Williams[1], on behalf of the LHCb Allen team
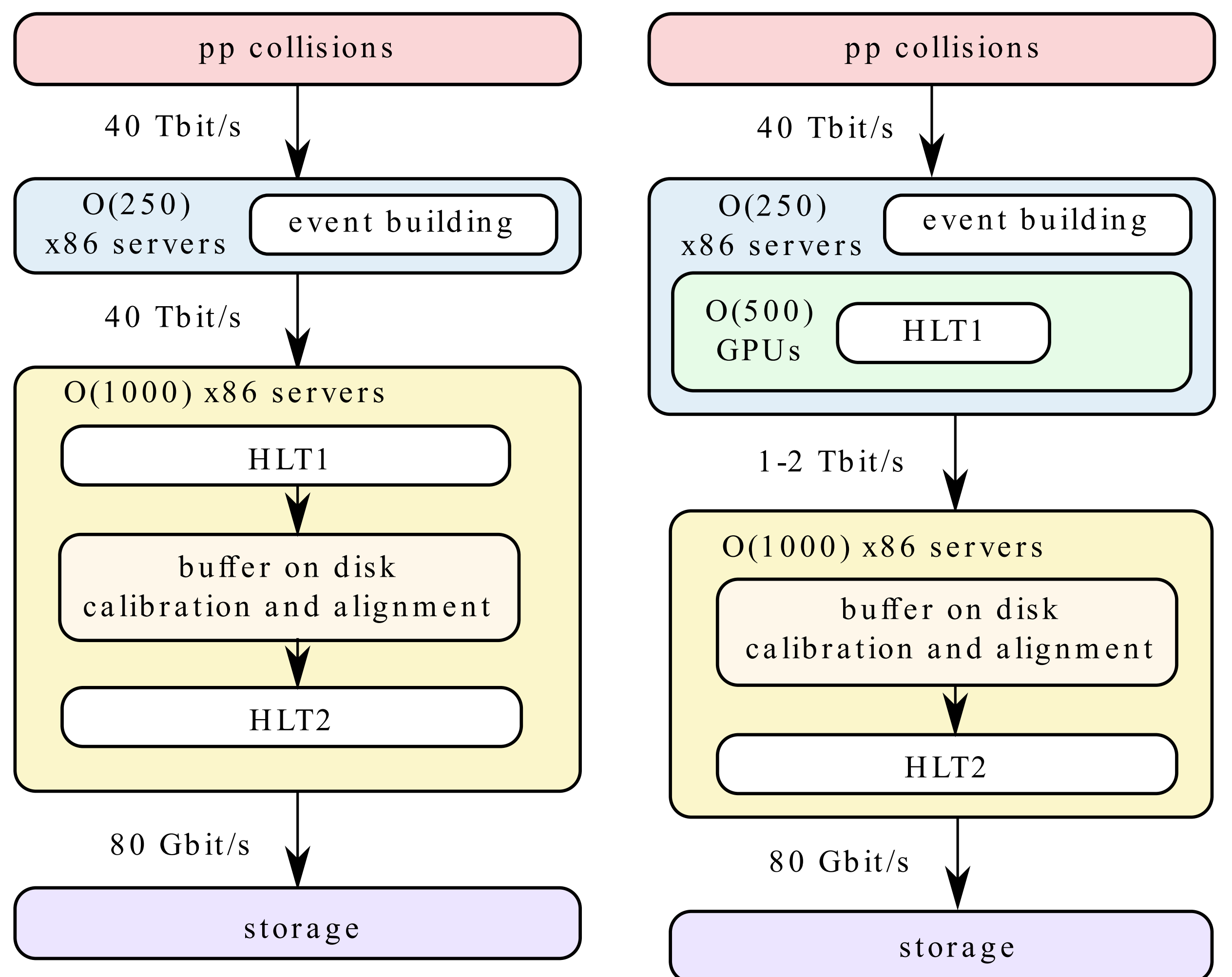
[1] MIT, [2] Cincinnati    * non-IRIS-HEP

## The LHCb Detector



Side View

5m

Magnet · SciFi Tracker · RICH2 · ECAL · HCAL · M2 M3 M4 M5

RICH1 · UT

Vertex Locator

By

upgrade

5m · 10m · 15m · 20m · z
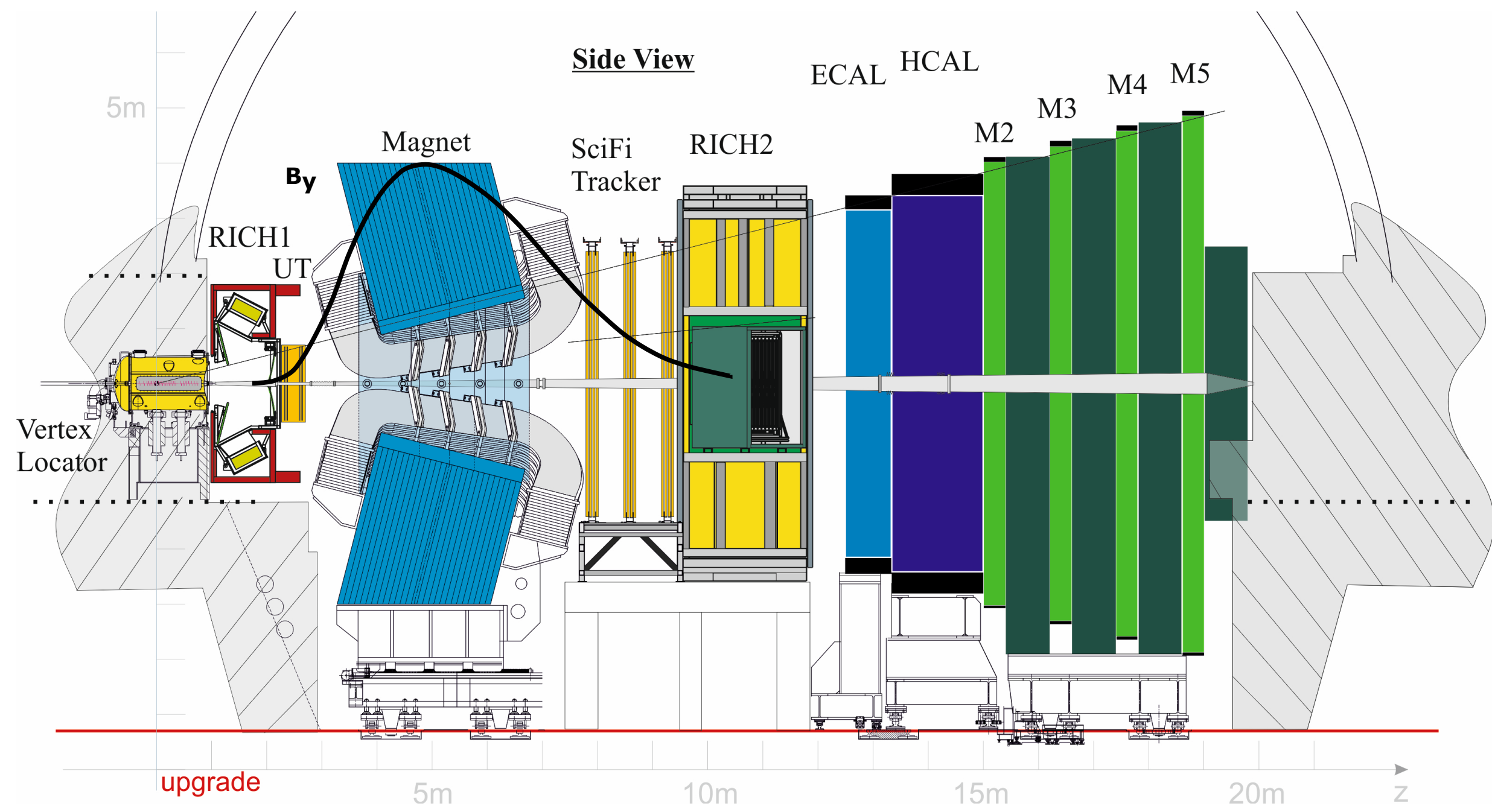
The upgraded LHCb detector
The y-component of the magnetic field is superposed to visualise in which parts of the detector trajectories are bent.

- Standalone package
- Run configurable sequence of algorithms
- Process O(1000) events in parallel
- Handles host/device memory allocation and transfers between host and device
- Sequence configurable at compile time
- Algorithms configurable at runtime
- Run tracking and vertex reconstruction

## GPU-enhanced Trigger



pp collisions · 40 Tbit/s · O(250) x86 servers — event building · 40 Tbit/s · O(1000) x86 servers — HLT1 — buffer on disk calibration and alignment — HLT2 · 80 Gbit/s · storage

pp collisions · 40 Tbit/s · O(250) x86 servers — event building · O(500) GPUs — HLT1 · 1-2 Tbit/s · O(1000) x86 servers — buffer on disk calibration and alignment — HLT2 · 80 Gbit/s · storage
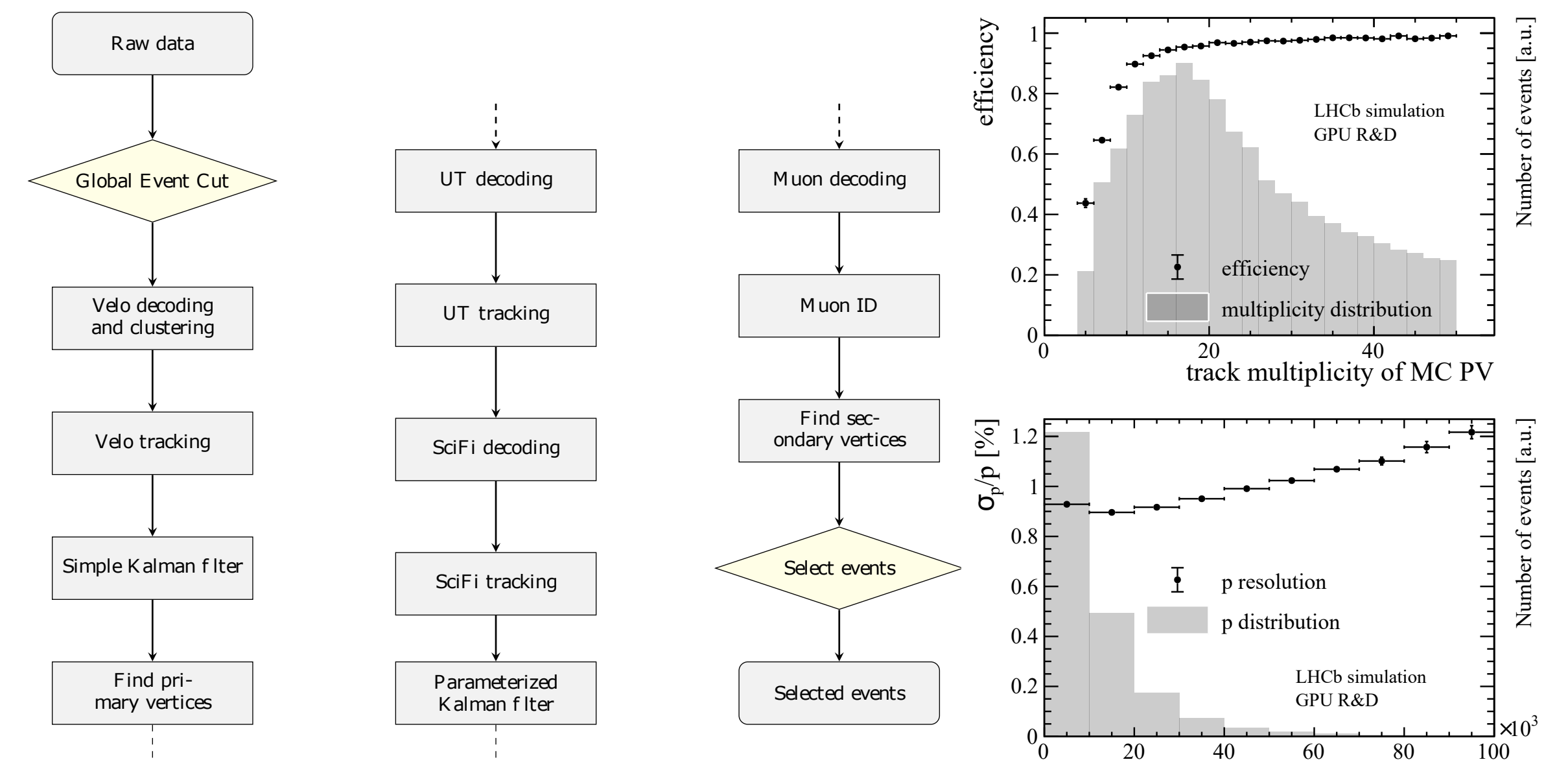
Overview of the LHCb trigger for Run III: (left) baseline and (right) GPU-enhanced. A GPU-based first-stage (HLT1) trigger would reduce the data rate between the two server farms by a factor of 30-60. Cost saving on networking offsets the cost of GPUs.
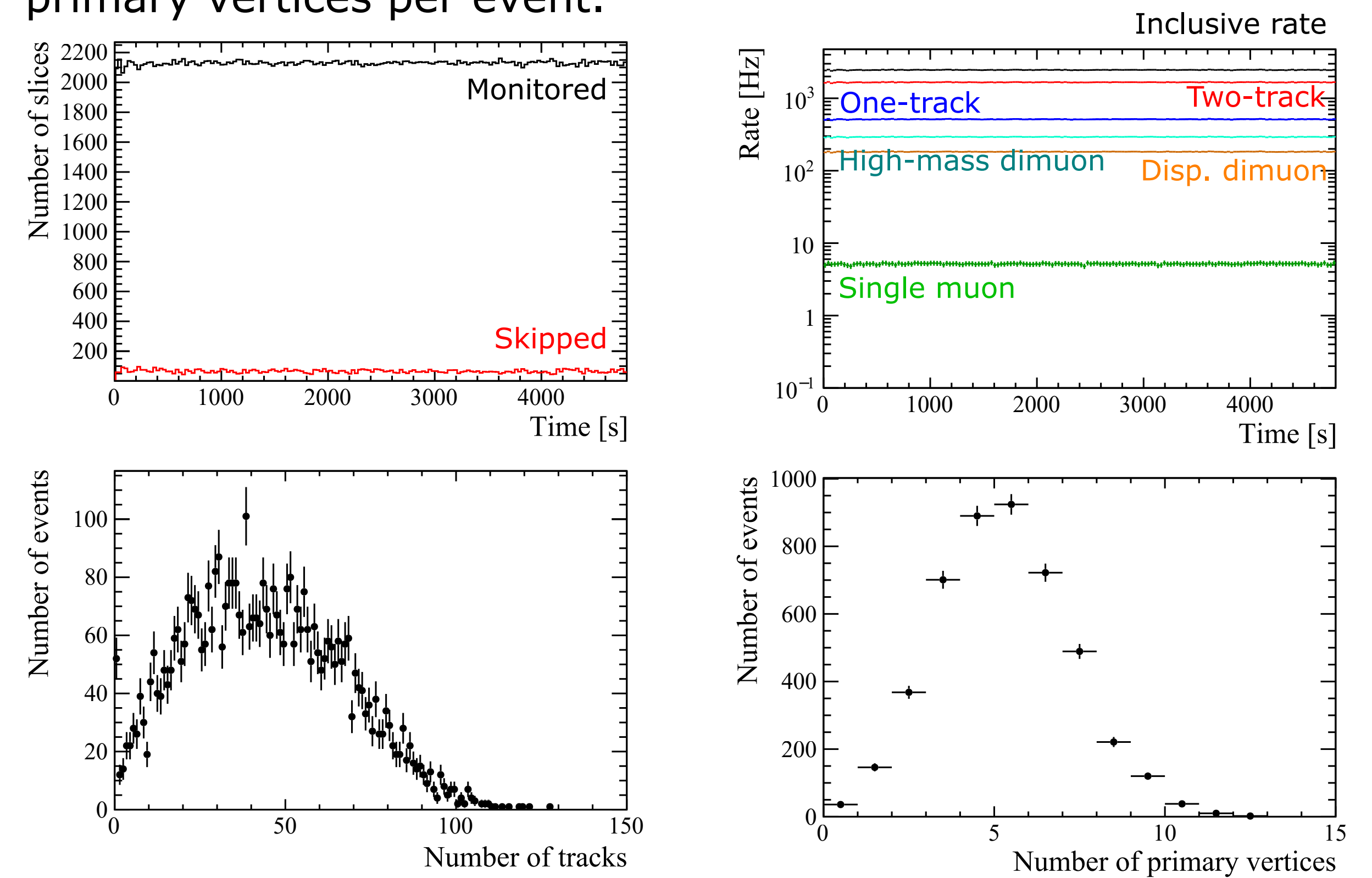
## Reconstruction

(Left) sequence of algorithms; (top right) efficiency of vertex reconstruction; (bottom right) momentum resolution of track reconstruction.



Raw data → Global Event Cut → Velo decoding and clustering → Velo tracking → Simple Kalman filter → Find primary vertices

UT decoding → UT tracking → SciFi decoding → SciFi tracking → Parameterized Kalman filter

Muon decoding → Muon ID → Find secondary vertices → Select events → Selected events

efficiency — track multiplicity of MC PV — LHCb simulation GPU R&D — efficiency — multiplicity distribution

$\sigma_p/p$ [%] — p [MeV/c] — p resolution — p distribution — LHCb simulation GPU R&D

## Monitoring

Integrated monitoring of selection rates and reconstructed objects: (top left) number of data slices monitored; (top right) event selection rates; (bottom left) number of reconstructed tracks per event; (bottom right) number of reconstructed primary vertices per event.



Number of slices — Time [s] — Monitored — Skipped

Rate [Hz] — Time [s] — Inclusive rate — One-track — Two-track — High-mass dimuon — Disp. dimuon — Single muon

Number of events — Number of tracks

Number of events — Number of primary vertices

## Throughput

- Throughput rate allows running on O(500) GPUs
- Throughput still scales with theoretical GPU performance
- New GPU performance expected to improve further by start of Run III

Options to do even more in HLT1:
- Improve low-momentum tracking efficiency
- Decode calorimeter information
- Remove global event cut
- Downstream track reconstruction



Throughput [kHz] — Theoretical 32 bit TFLOPS — LHCb simulation GPU R&D — Tesla V100 32GB — Quadro RTX 6000 — GeForce RTX 2080 Ti — Tesla T4 — GeForce GTX 1080 Ti — GeForce GTX TITAN X — GeForce GTX 1060 6GB — GeForce GTX 680 — GeForce GTX 670

Throughput on Tesla V100 16GB [kHz] — Number of events [a.u.] — SciFi raw data volume [kB] — LHCb simulation GPU R&D — Data volume distribution — Throughput