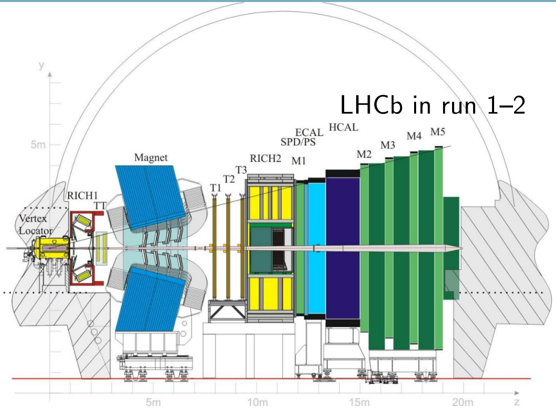
GPU-based software trigger for LHCb experiment

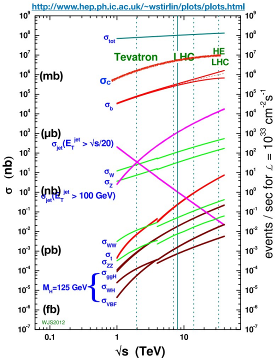
Anton Poluektov
on behalf of LHCb collaboration
RTA project

Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France

9 April 2024

LHCb experiment in 2010-2018

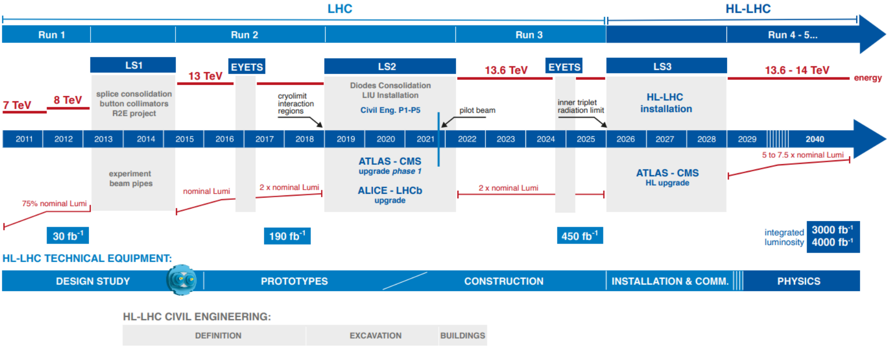




Forward spectrometer, optimised for b and c decays. $2 < \eta < 5$

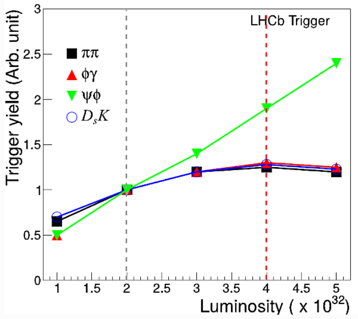
- Excellent vertex resolution (weak decays)
- High-precision tracking before and after the magnet
- PID in broad range of momenta $3 < p < 150\,\mathrm{GeV}$
- Efficient trigger, including fully-hadronic final states, $\sim$12 kHz output rate

LHC timeline



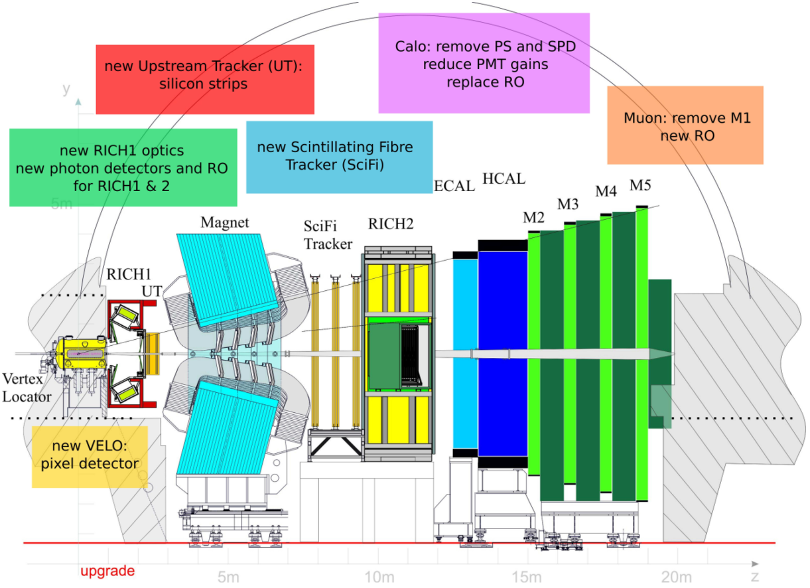
- LHC Run 2 finished in 2018
  - LHCb: $\int \mathcal{L}dt = 9\,\text{fb}^{-1}$ collected in 2010-2018
- Long shutdown until 2022: upgrade of the machine and detectors
  - LHCb Upgrade I: major upgrade/replacement of the subsystems and readout
- Run 3 until 2026 $\rightarrow$ HL-LHC upgrade $\rightarrow$ Run 4 ...
  - LHCb goal: $50\,\text{fb}^{-1}$ by the end of Run 4 $\rightarrow$ Upgrade II

LHCb upgrade case



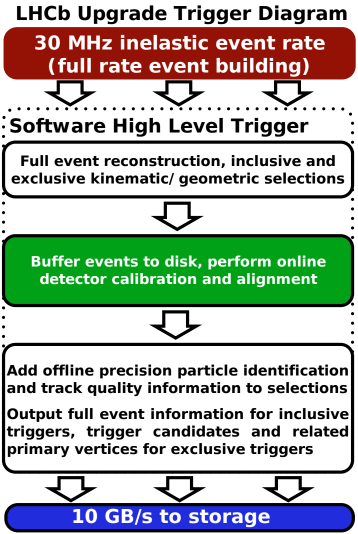
- Instantaneous luminosity:
  4×10^{32} (Run 2) $\rightarrow 2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
- Run 1–2 trigger:
  - First stage: hadrware L0 (40→1 MHz) using high p_T/E_T signatures
  - 1 MHz limit saturates hadronic modes already in Run 2
    (higher rate ⇒ higher thresholds)
- The only solution: read full event at bunch-crossing rate and apply track reconstruction/IP selections.
- Upgrade/replace subsystems:
  - Cope with higher occupancy.
  - Faster/higher precision tracking
- Fully replace DAQ and trigger.

LHCb upgrade



Complete replacement of DAQ, fully software trigger (HLT1 + HLT2)

Upgraded DAQ+trigger: functional diagram

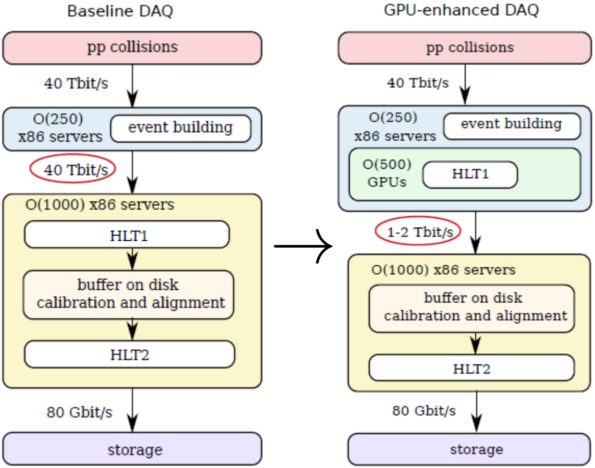


HLT1: [LHCb upgrade computing TDR]

- Subdetector reconstruction:
  - VELO: clustering, tracking, vertex reconstruction
  - UT, SciFi: tracking
  - Muon: Hit-track matching
- Global event reconstruction:
  - Track fit (Kalman filter)
  - Reconstruction of secondary vertices
- Selections: [LHCb-PUB-2019-013]
  - Single displaced tracks
  - Two-track displaced vertices
  - Single displaced muons
  - Low-mass displaced two-muon vertices
  - High-mass dimuons

Upgraded DAQ+trigger: hardware [Comput Softw Big Sci 4, 7 (2020)]

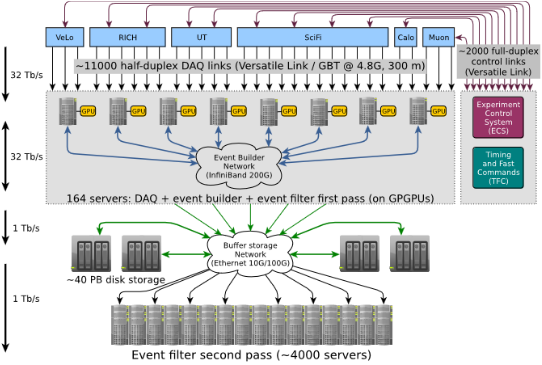
Baseline CPU-based design was replaced by GPU-accelerated one



- HLT1 runs on EB nodes
- Reduce network bandwidth between EB and filter farms
- Free up filter farm CPU for HLT2 only

Warning: the exact numbers for BW, N(servers) have evolved since then

Upgraded LHCb DAQ: current implementation



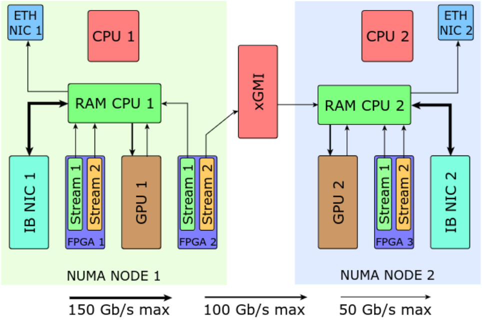
©CPPM

- Event rate: 30 MHz non-empty bunch crossing
- Event size: $\sim$ 100 kB
- Input bandwidth: $\sim$ 32 Tbit/s

- New PCIe40 readout boards
  - 24 optical inputs, PCIe interface
- Event builder network using commercial technology
  - 200 Gbit/s InfiniBand© network with remote direct memory access

Event builder hardware [IEEE Trans Nucl Sci, 70, 6 (2023)]

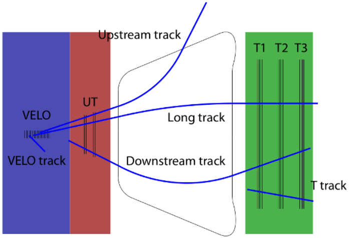
Current configuration: 164 2-CPU server nodes

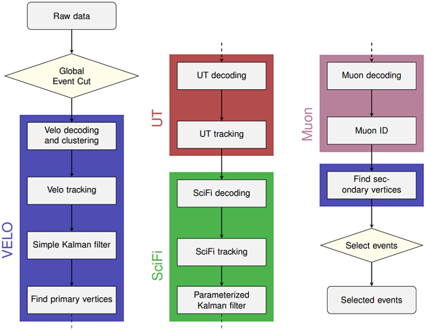


2-CPU server node hardware diagram

- **CPU:** ×2 AMD EPYC 7502, 32 cores
- **GPU:** ×2 NVIDIA RTX A5000
- **RAM:** 512 GB DDR4
- **Network:** ×2 NVIDIA ConnectX-6 HDR
- **Readout:** ×3 PCIe40 FPGA boards

- Framework for GPU-based execution of an algorithm sequence [GitLab repo], [Documentation]
- Cross-architecture compatibility: Runs on CPU, NVidia GPU (CUDA), AMD GPU (HIP)
- Algorithm sequences defined in python, generated at runtime
- Three levels of parallelism: Intra-collision (tracks, clusters), collisions, collision batches

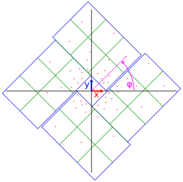




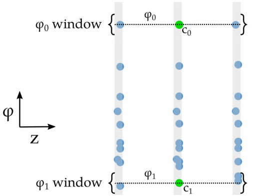
Allen project: parallel algorithms

Fast parallel algorithms developed for tracking, vertexing etc.

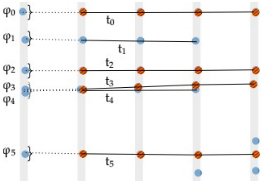
E.g. reconstruction of tracks in VELO:



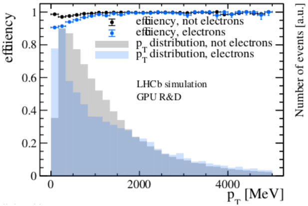
1. Sorting hits in ϕ



2. Triplet seeding



3. Triplet forwarding



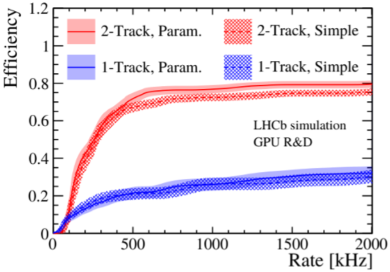
VELO tracking performance

[D. Campora, N. Neufeld, A. Riscos Nez, IPDPSW 2019]

Allen project: HLT1 performance [Comput Softw Big Sci 4, 7 (2020)]

Trigger	Rate [kHz]
1-Track	215 ± 18
2-Track	659 ± 31
High-p_T muon	5 ± 3
Displaced dimuon	74 ± 10
High-mass dimuon	134 ± 14
Total	999 ± 38

Rates of HLT1 lines on minimum bias events

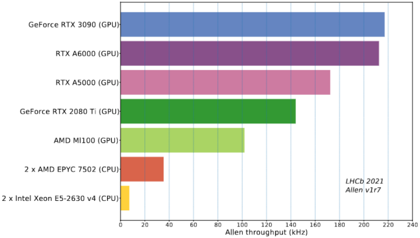


Efficiency of 1-Track and 2-Track selections with $B_s^0 \to \phi\phi$ MC

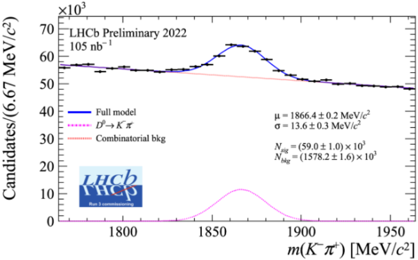
Signal	GEC	TIS -OR- TOS	TOS	GEC × TOS
$B^0 \to K^{*0}\mu^+\mu^-$	89 ± 2	91 ± 2	89 ± 2	79 ± 3
$B^0 \to K^{*0}e^+e^-$	84 ± 3	69 ± 4	62 ± 4	52 ± 4
$B_s^0 \to \phi\phi$	83 ± 3	76 ± 3	69 ± 3	57 ± 3
$D_s^+ \to K^+K^-\pi^+$	82 ± 4	59 ± 5	43 ± 5	35 ± 4
$Z \to \mu^+\mu^-$	78 ± 1	99 ± 0	99 ± 0	77 ± 1

Efficiencies of HLT1 selection for benchmark signals

Allen project: HLT1 performance

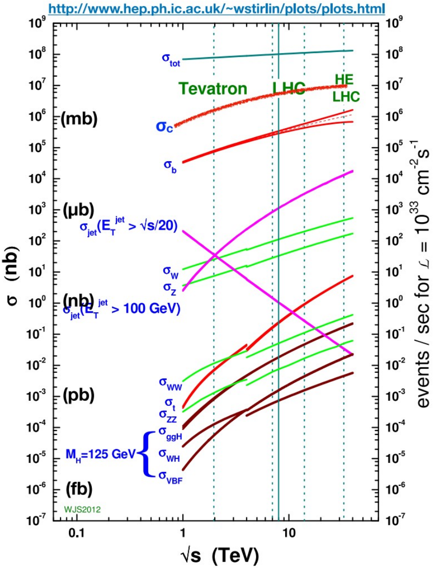


HLT1 throughput for various GPU cards.
[LHCB-FIGURE-2020-014]



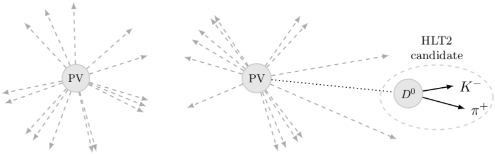
$D^0 \to K^-\pi^+$ peak directly from HLT1 (2022)
[LHCB-FIGURE-2023-009]

HLT2 signal rates



- Signal rates at $\mathcal{L} = 2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$:
  O(10) MHz charm
  O(1) MHz beauty
- Output bandwidth limited to 10 GB/ s. Up to 100 kHz with full event size of 100 kB.
- Need to reduce the event size for higher rate

Selective persistency: write out only the "interesting" part of the event.

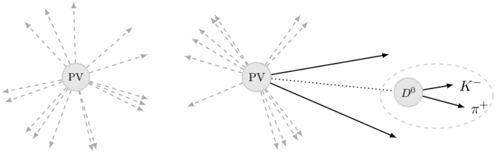


- Turbo stream:
  - Minimum output: only HLT2 signal candidates

Limitations: cannot refit tracks and PVs offline, rerun flavour tagging etc.
Advantage: Event size $O(10)$ smaller than RAW

Persistency model [LHCb upgrade computing TDR]

Selective persistency: write out only the “interesting” part of the event.



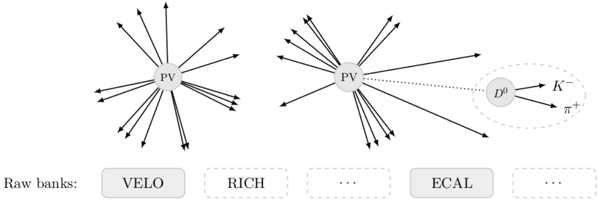
- `Turbo` stream:
  - Minimum output: only HLT2 signal candidates
  - Optionally: (parts of) *pp* vertex (*e.g.* "cone" around candidate for spectroscopy searches)

Limitations: cannot refit tracks and PVs offline, rerun flavour tagging etc.
Advantage: Event size $O(10)$ smaller than RAW

Persistency model

[LHCb upgrade computing TDR]

Selective persistency: write out only the "interesting" part of the event.

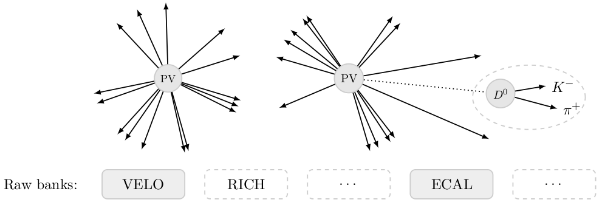


- `Turbo` stream:
  - Minimum output: only HLT2 signal candidates
  - Optionally: (parts of) *pp* vertex (*e.g.* "cone" around candidate for spectroscopy searches)

  Limitations: cannot refit tracks and PVs offline, rerun flavour tagging etc.
  Advantage: Event size $O(10)$ smaller than RAW
- `FULL` stream: all reconstructed objects in the event
  - + selected RAW banks

Persistency model

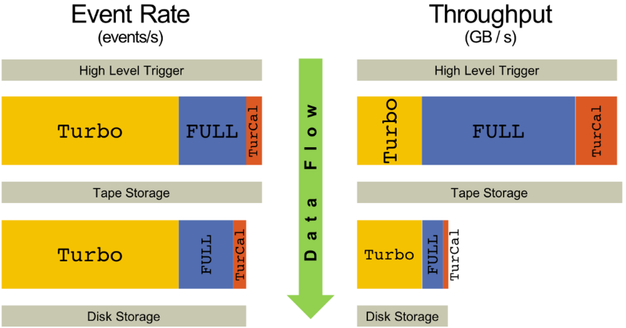
Selective persistency: write out only the "interesting" part of the event.



- `Turbo` stream:
  - Minimum output: only HLT2 signal candidates
  - Optionally: (parts of) *pp* vertex (*e.g.* "cone" around candidate for spectroscopy searches)

  Limitations: cannot refit tracks and PVs offline, rerun flavour tagging etc.
  Advantage: Event size $O(10)$ smaller than RAW
- `FULL` stream: all reconstructed objects in the event
  - + selected RAW banks
- `TurCal` stream: HLT2 candidates and selected RAW banks
  Used for offline calibration and performance measurement

HLT2 output rates



Rate and bandwidth to tape

stream	rate fraction	throughput (GB/s)	bandwidth fraction
FULL	26%	5.9	59%
Turbo	68%	2.5	25%
TurCal	6%	1.6	16%
total	100%	10.0	100%

Disk bandwidth

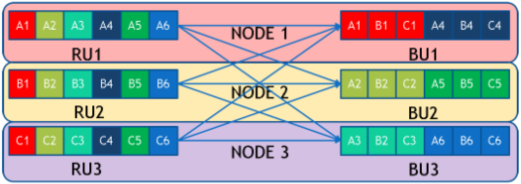
stream	throughput (GB/s)	bandwidth fraction
FULL	0.8	22%
Turbo	2.5	72%
TurCal	0.2	6%
total	3.5	100%

Summary

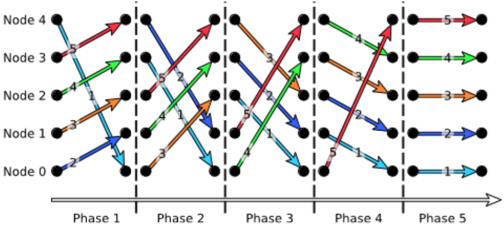
- LHCb started taking data after upgrade in 2022 (Run 3)
  - Commissioning with LHC in 2022
  - 2023: run with open VELO after incident; UT commissioning
  - Plan to run with maximum performance in 2024–2025
- Aim to increase instantaneous luminosity to $2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ (5 times pre-upgrade).
- Major redesign of readout and trigger compared to Run 2
- Remove hardware L0 stage, read out full detector at $30\,\mathrm{MHz}$ non-empty bunch crossing rate
  - Need to cope with 32 Tbit/s input bandwidth: *highest in any physics experiment to date*
- HLT filtering farm:
  - Architecture of split trigger with disk buffer, alignment and calibration ⇒ offline-quality output.
  - GPU-based HLT1 stage in the event builder farm
  - CPU-based HLT2
  - Increase physics output by moving most of signal rate to `Turbo` stream (reduced size, no RAW information).
  - 10 GB/s output bandwidth to tape for further analysis

Backup

Event builder



Event building process



Linear shifting scheduling