GPU-based Online Trigger at LHCb

Peilian Li
on behalf of the LHCb collaboration
(University of Chinese Academy of Sciences)

The International Workshop on CEPC
Hangzhou, 2024-10-24

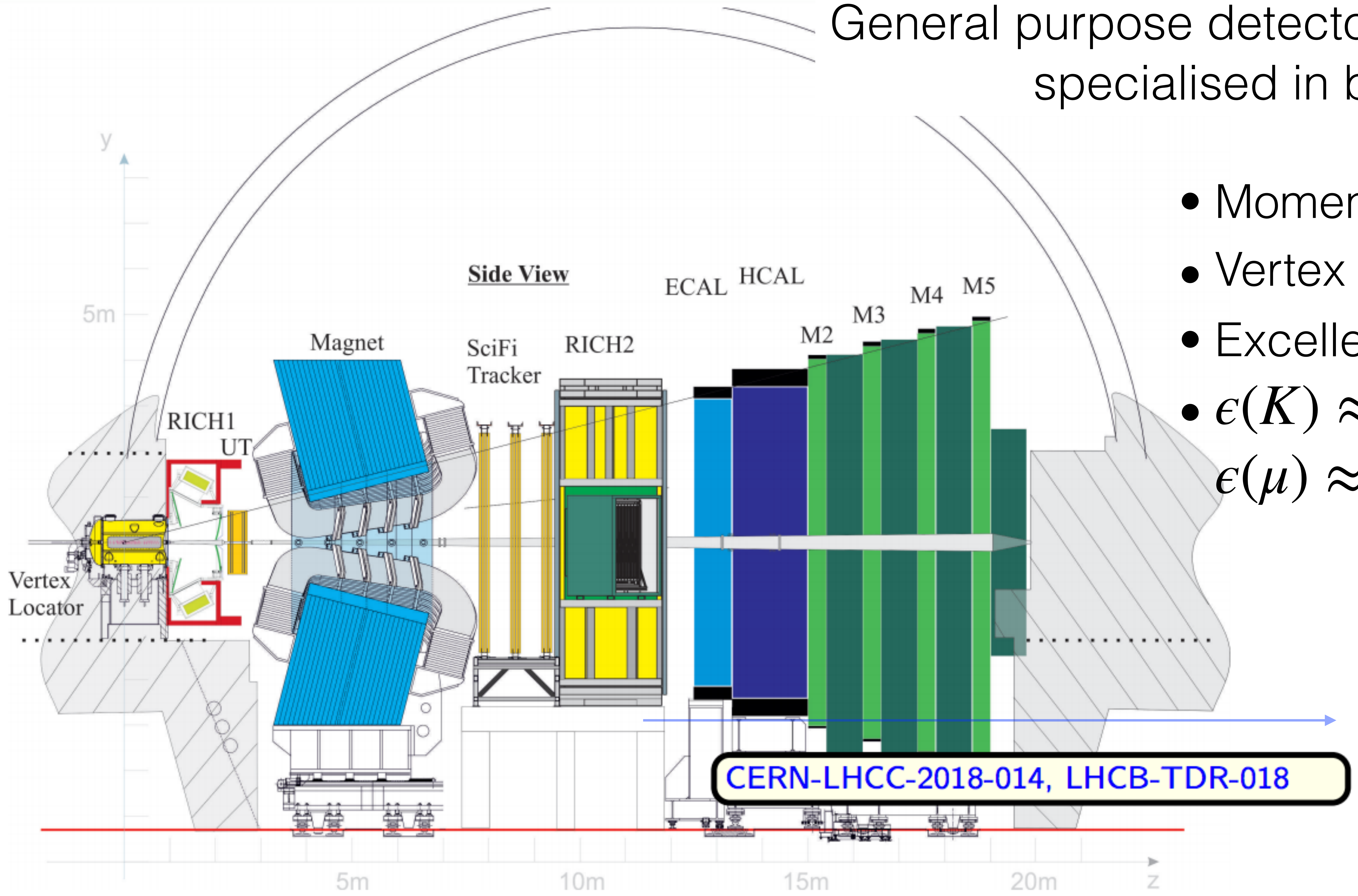
Outline

- LHCb detector for Run 3
- Trigger strategy
- Allen design
- Track reconstruction with GPU
- Performances

See Tracking with FPGA in the next Talk by Ao XU

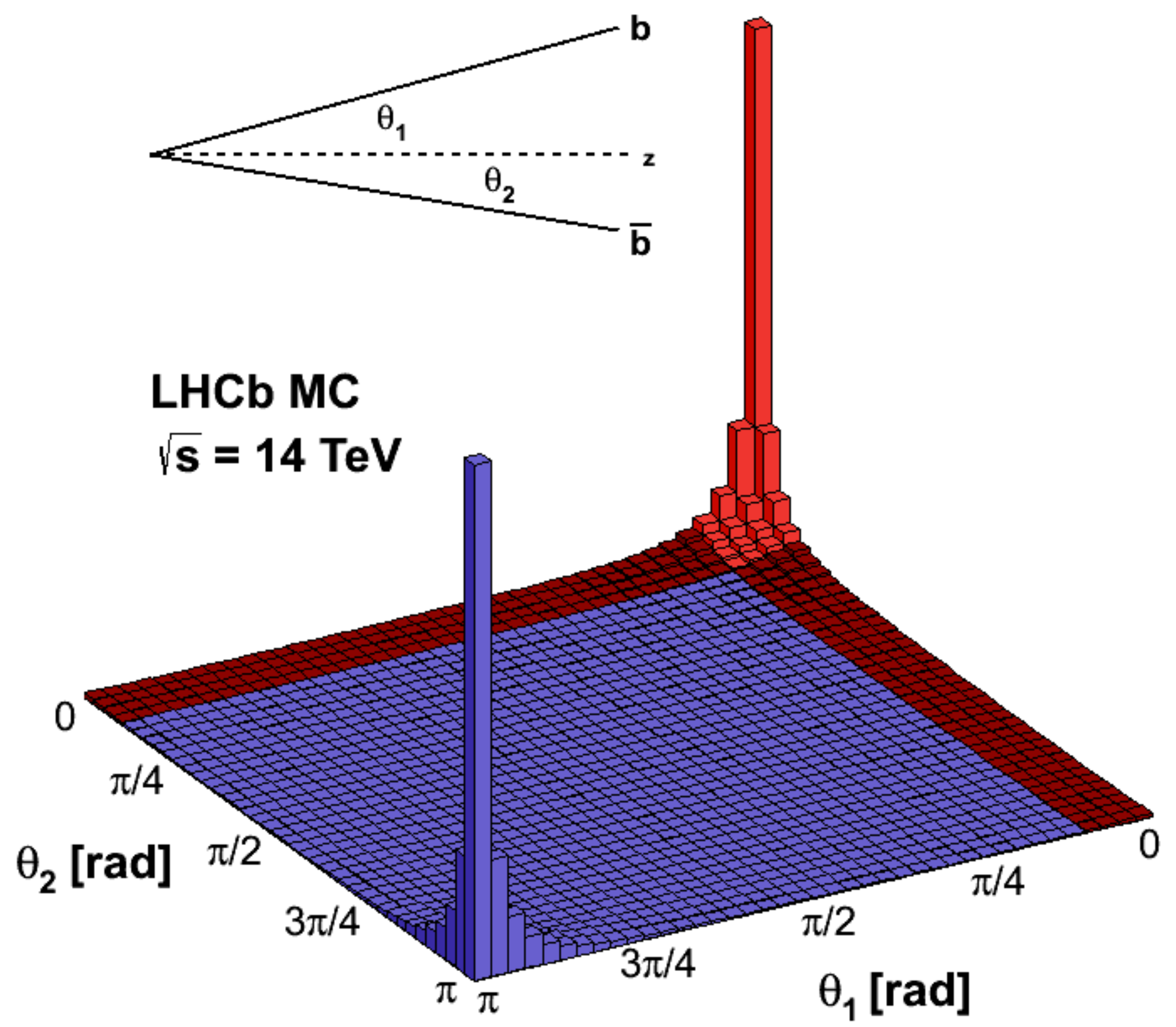
- Summary

LHCb Detector for Run 3



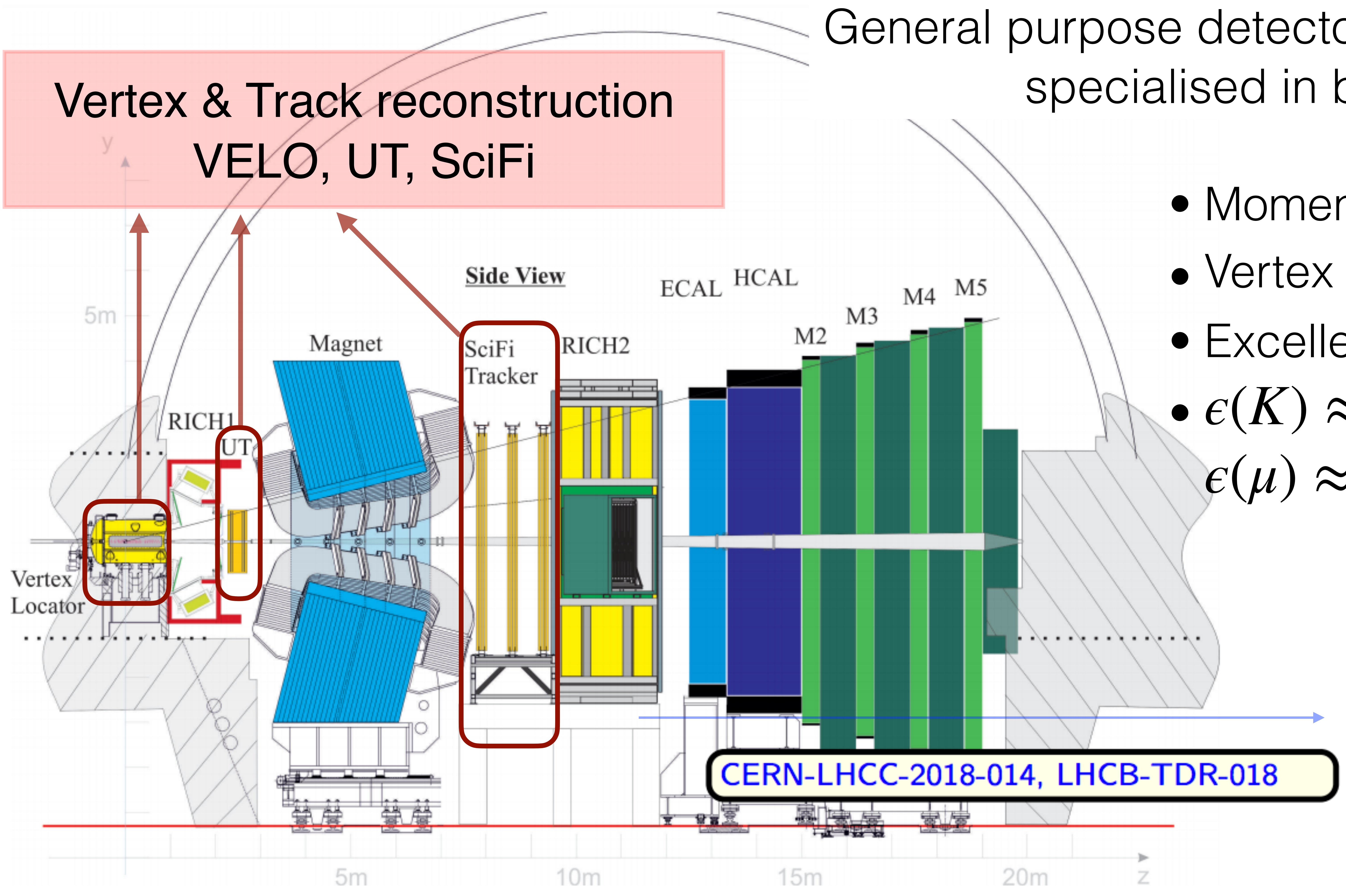
General purpose detector in the forward region ($2 < \eta < 5$)
specialised in beauty and charm physics

- Momentum resolution: 0.5%~1%
- Vertex resolution: $\sigma_{IP} \sim 35\mu m$
- Excellent particle identification
- $\epsilon(K) \approx 95\,\%$, misID $p(\pi \rightarrow K) \approx 5\,\%$
  $\epsilon(\mu) \approx 97\,\%$

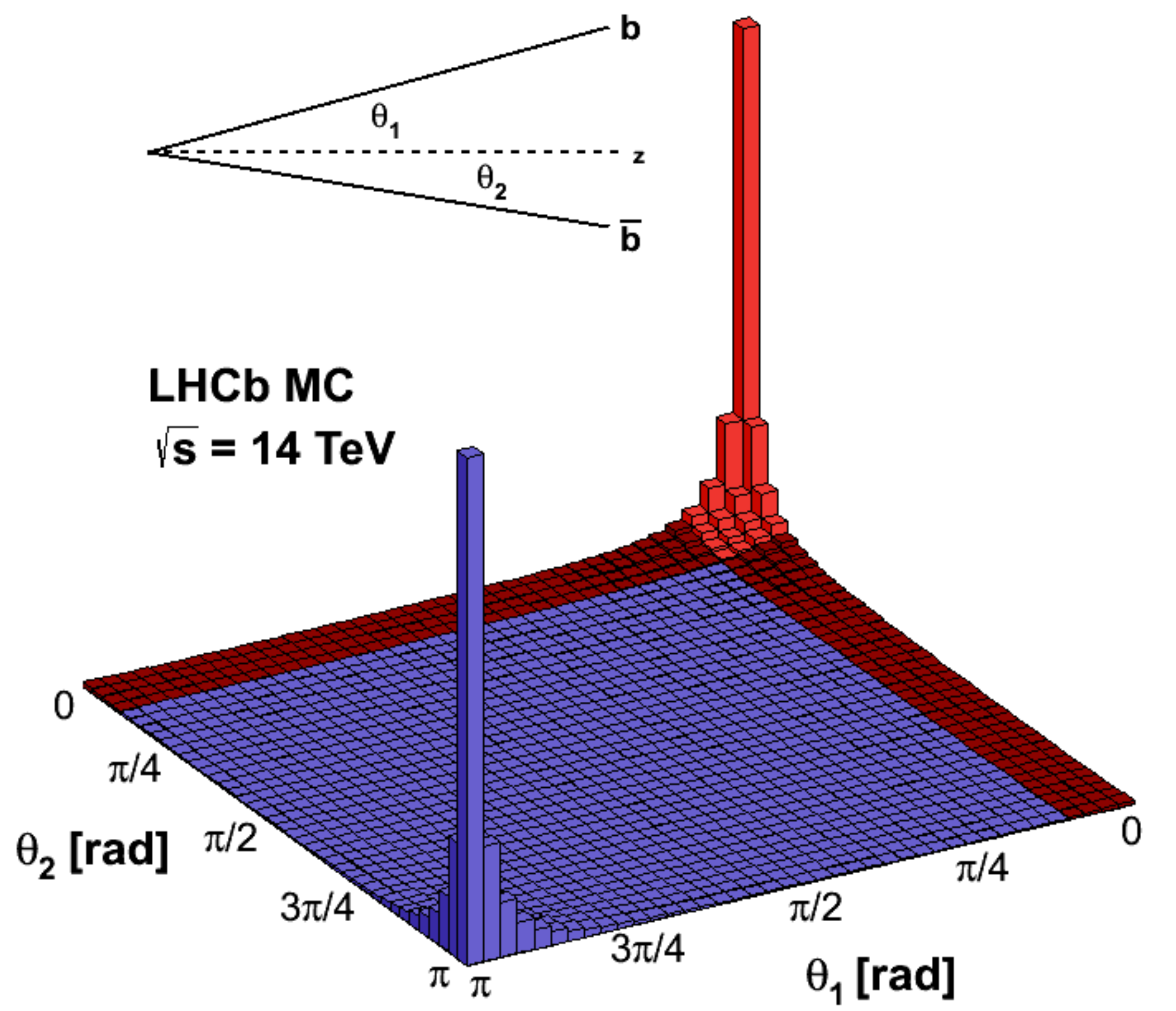


LHCb Detector for Run 3

General purpose detector in the forward region $(2 < \eta < 5)$
specialised in beauty and charm physics



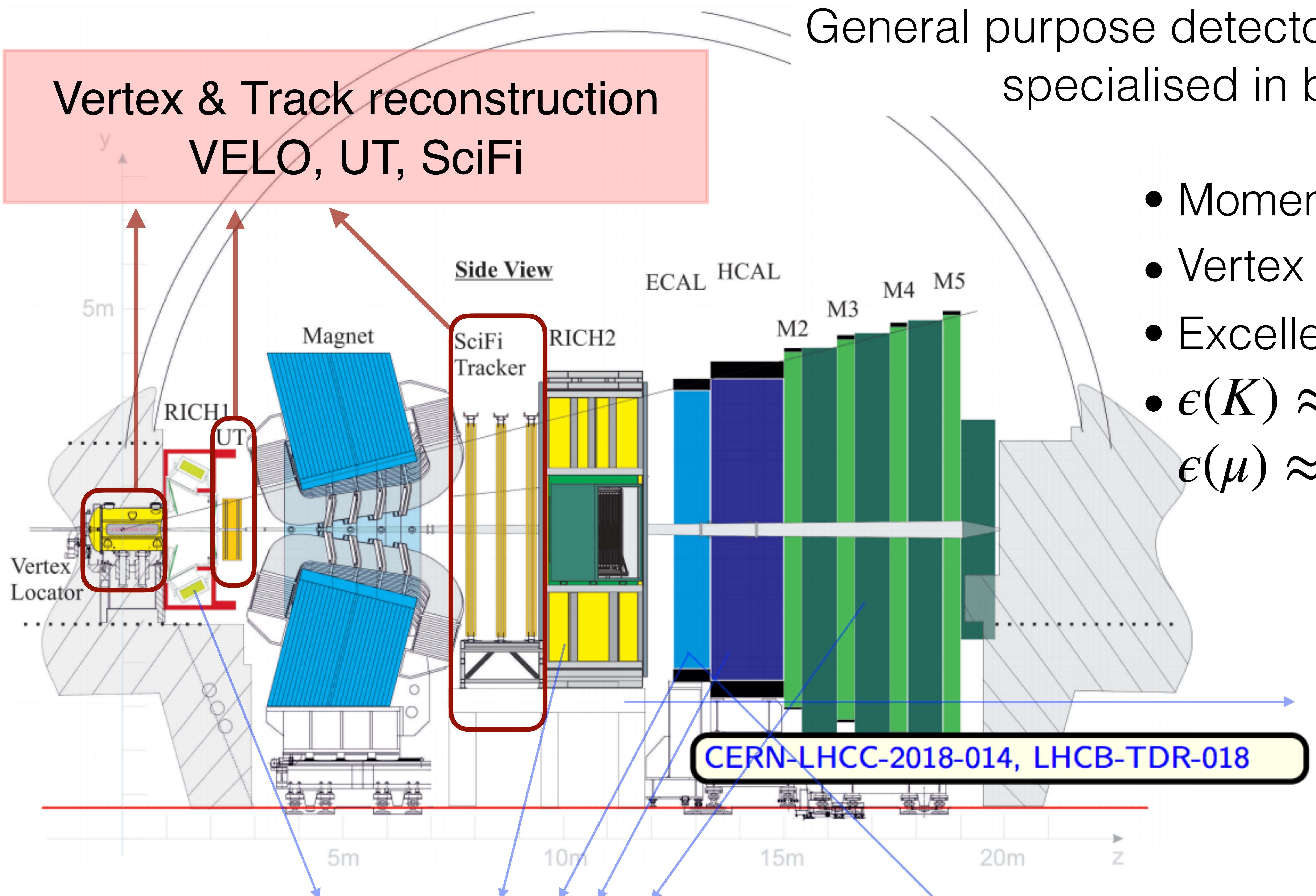
- Momentum resolution: 0.5%~1%
- Vertex resolution: $\sigma_{IP} \sim 35\mu m$
- Excellent particle identification
- $\epsilon(K) \approx 95\,\%$, misID $p(\pi \rightarrow K) \approx 5\,\%$
  $\epsilon(\mu) \approx 97\,\%$



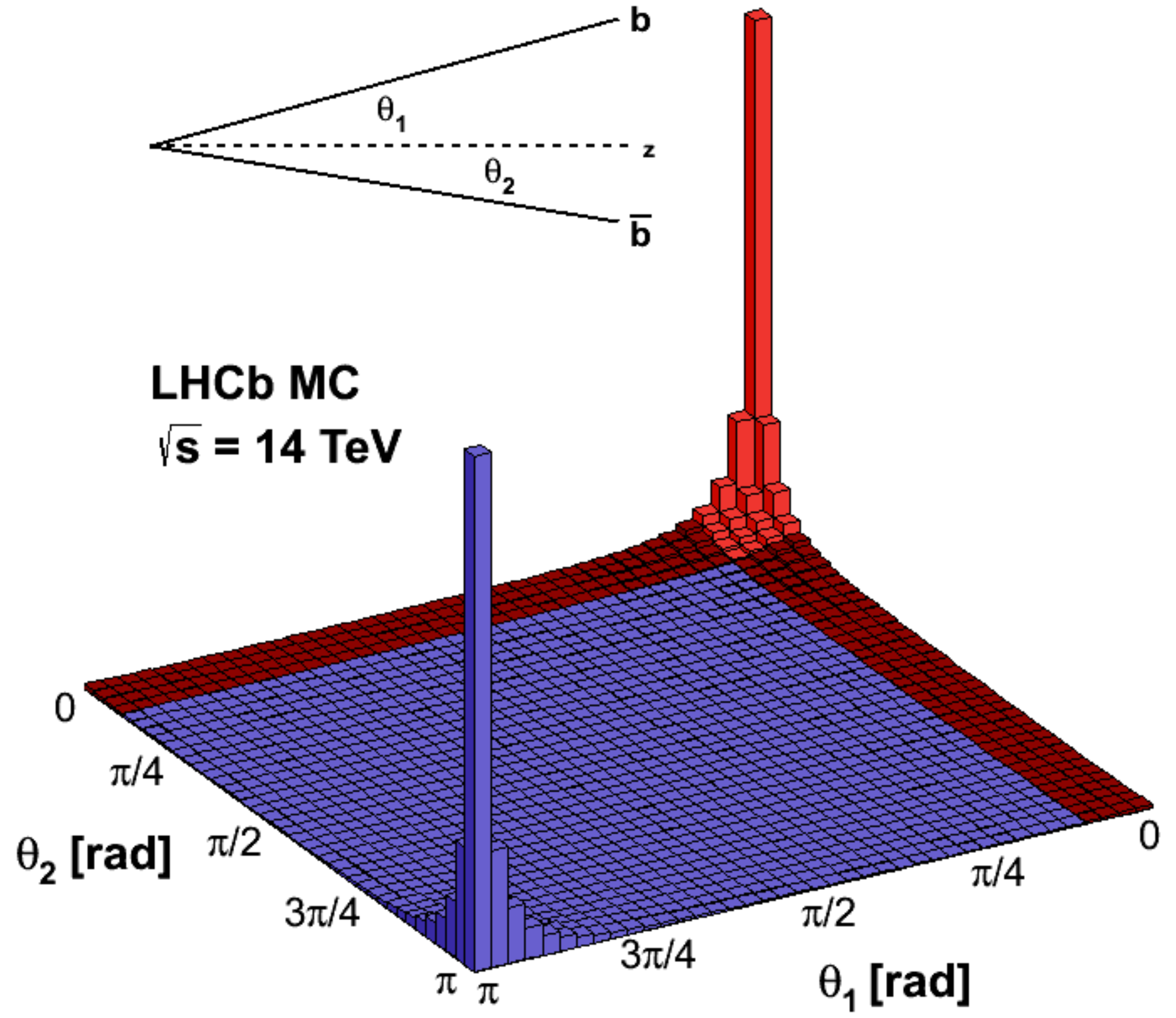
LHCb Detector for Run 3

General purpose detector in the forward region ($2 < \eta < 5$)
specialised in beauty and charm physics

Vertex & Track reconstruction
VELO, UT, SciFi



- Momentum resolution: 0.5%~1%
- Vertex resolution: $\sigma_{IP} \sim 35\mu m$
- Excellent particle identification
- $\epsilon(K) \approx 95\,\%$, misID $p(\pi \to K) \approx 5\,\%$
  $\epsilon(\mu) \approx 97\,\%$

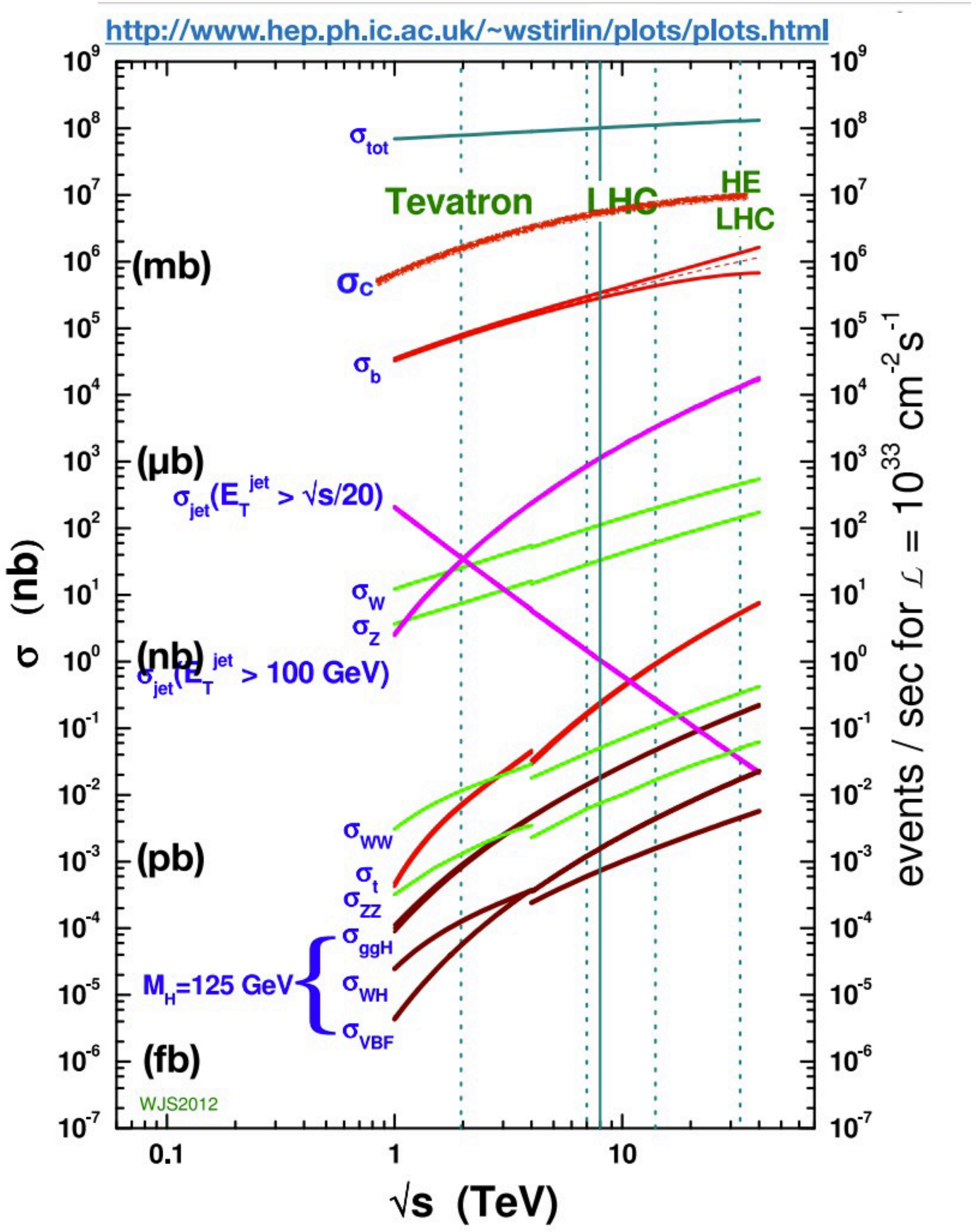


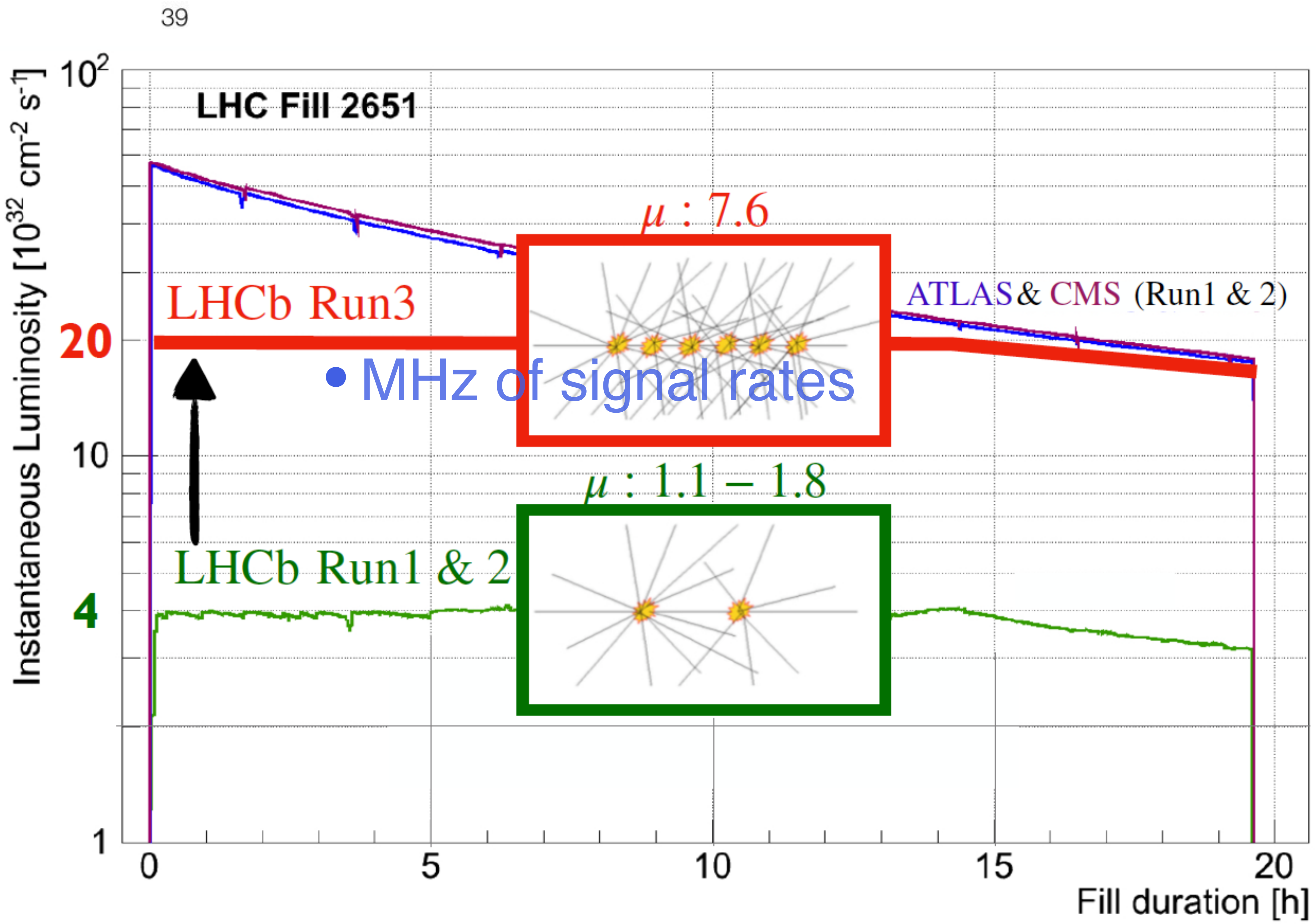
Particle identification:
RICH, MUON, ECAL

Neutral reconstruction:
ECAL

Challenges for LHCb Run 3

- Luminosity of 2×10^{33} $cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV, visible collisions per bunch $\mu \sim 5$

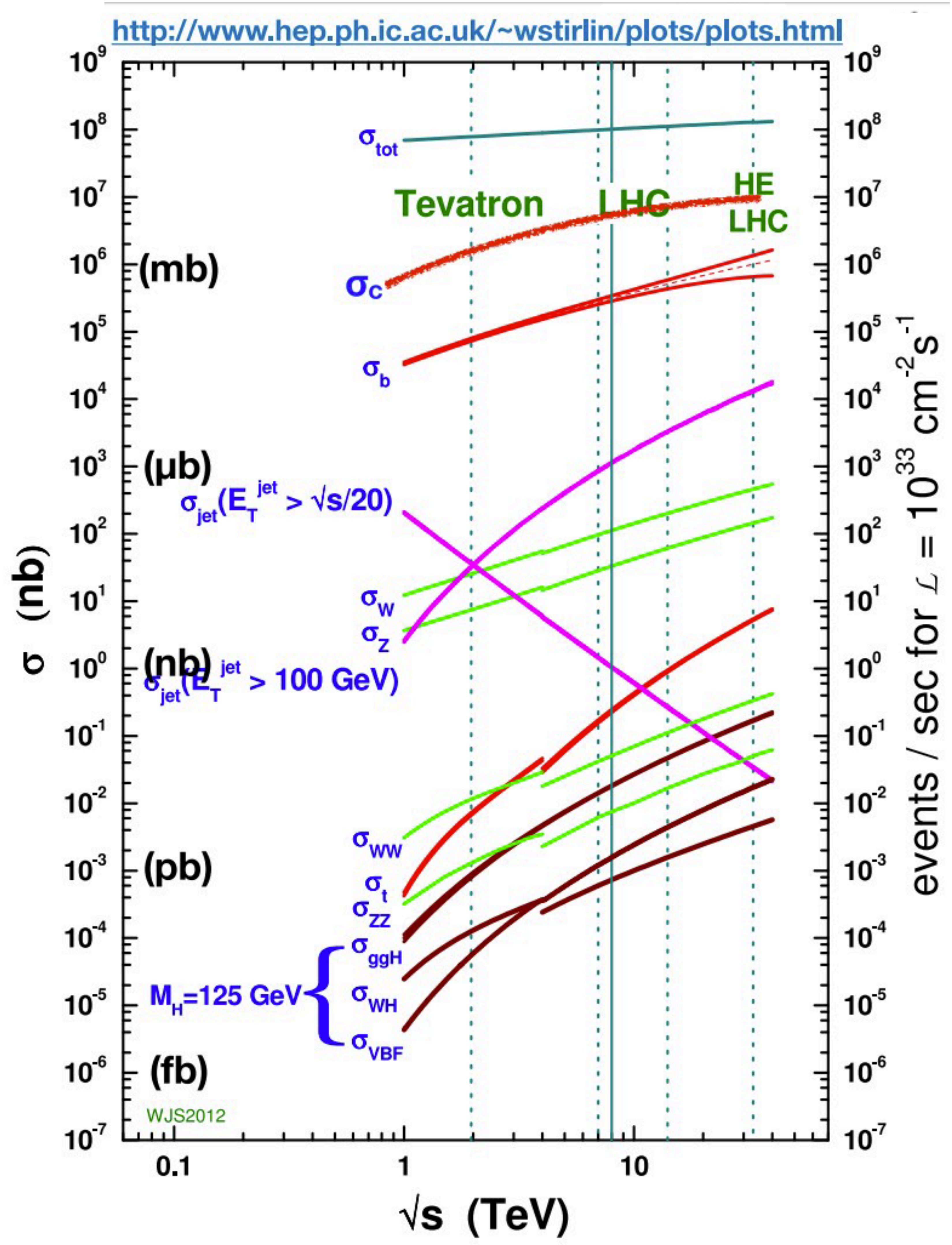




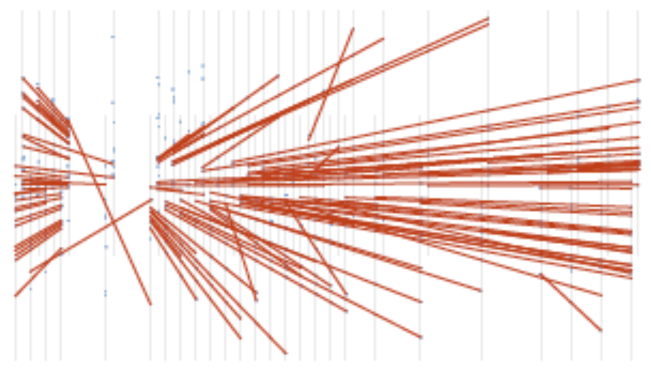
* μ is the average visible collisions /bunch

Challenges for LHCb Run 3

- Luminosity of $2 \times 10^{33}\ cm^{-2}s^{-1}$, $\sqrt{s}$ = 14 TeV, visible collisions per bunch $\mu \sim 5$



LHCb: Mainly beauty and charm physics

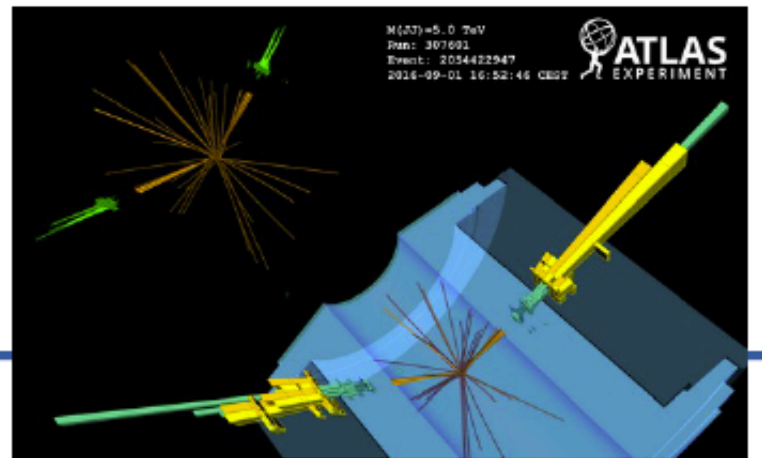


- Signal rates at MHz level
- Signal characteristics: Displaced vertices, momentum, particle type
- → No optimal local criteria for selection

- MHz of signal rates

ATLAS & CMS: Mainly Higgs properties, high p_T new phenomena

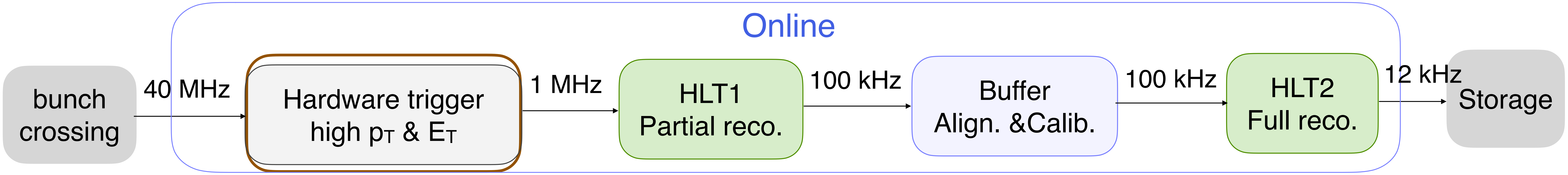
- Signal rates up to hundreds of kHz
- Signal characteristics: high pT / transverse energy
- → Local criteria for selection possible



Challenges for LHCb Run 3

Hardware trigger: 40→1 MHz read-out limits (fixed-latency trigger)

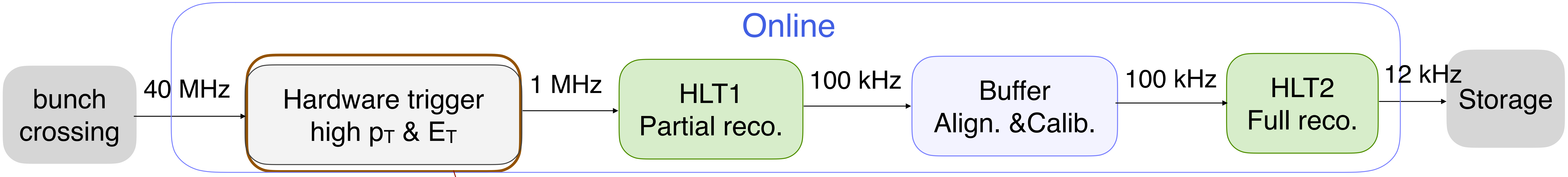
→ based on muon detector and calorimeters



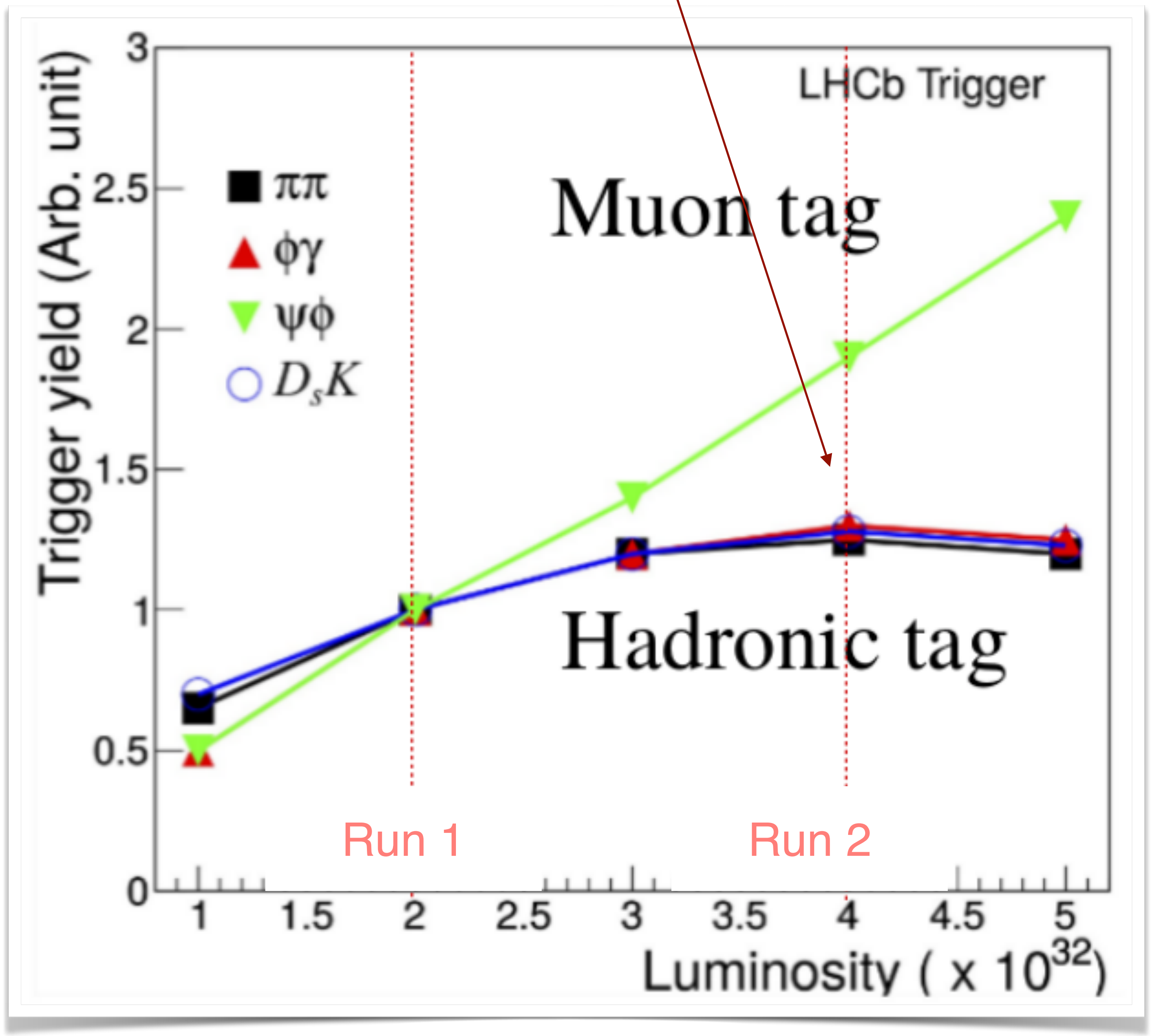
Challenges for LHCb Run 3

Hardware trigger: 40→1 MHz read-out limits (fixed-latency trigger)

→ based on muon detector and calorimeters

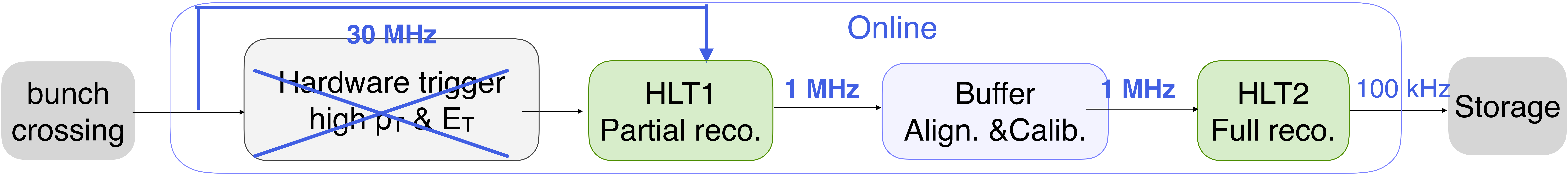


Conf. Series 878(2017)012012



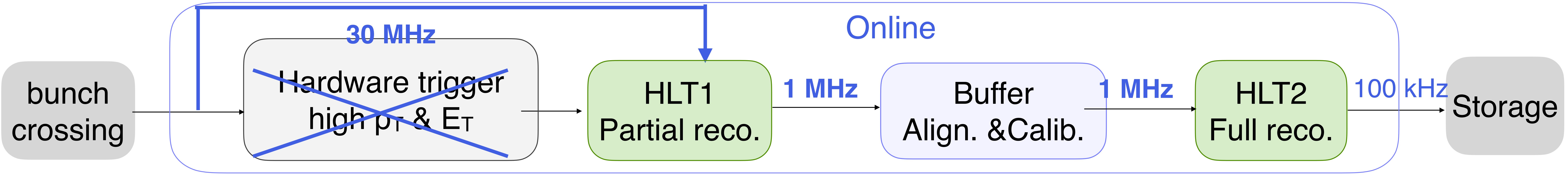
- Hardware trigger is not an option, as rate limit of 1 MHz saturates fully hadronic modes

LHCb Run 3 Trigger

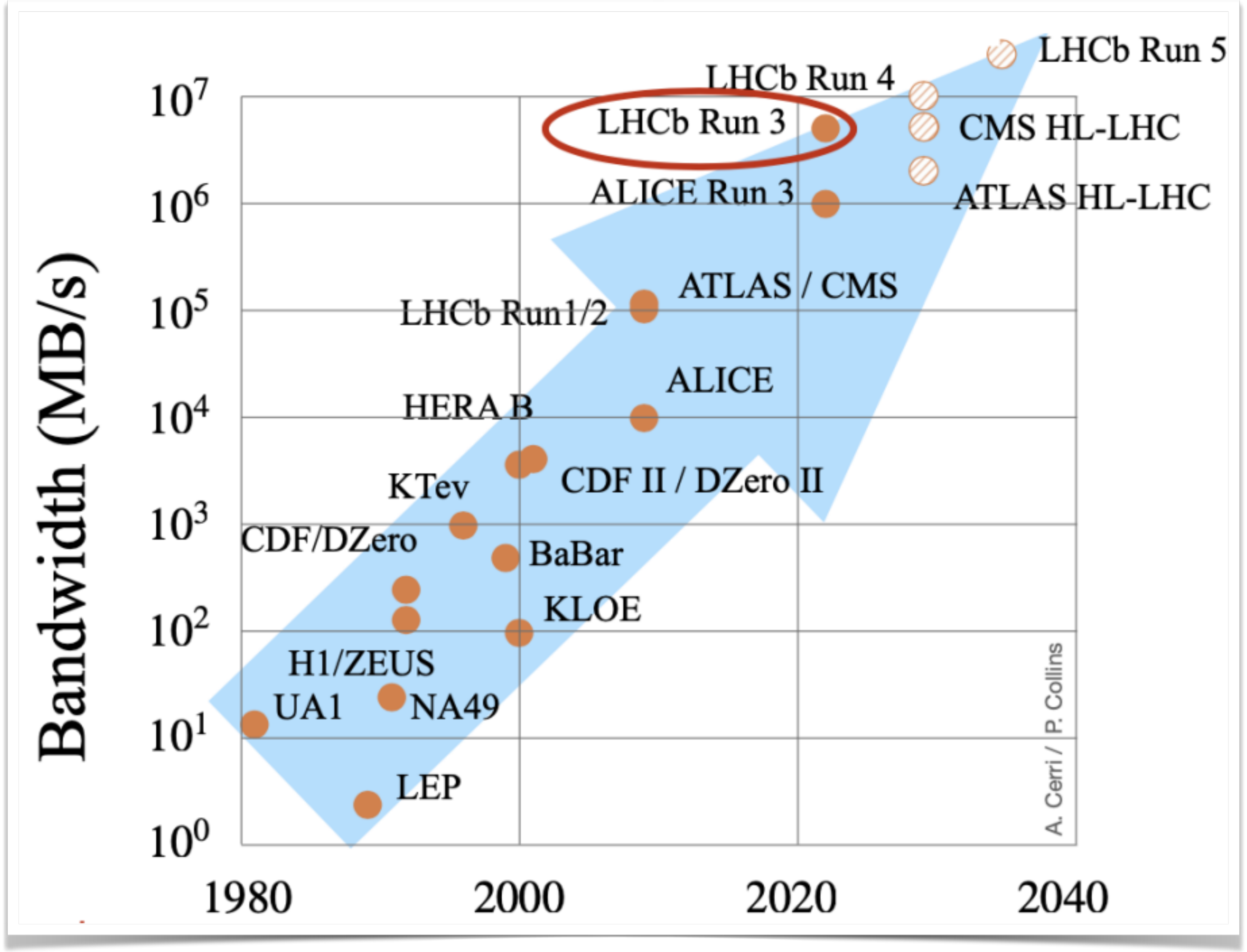


- Remove hardware trigger, fully software trigger
- Read out the full detector at 30 MHz in HLT1
- Real time alignment and calibration with 10x higher data rate than Run 2
- Full offline-quality reconstruction in “real-time”
- Increase of hadronic trigger efficiency by 2~4

LHCb Run 3 Trigger

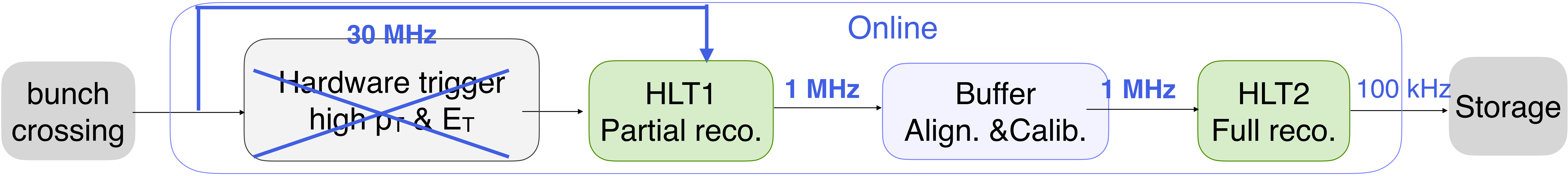


- Remove hardware trigger, fully software trigger
- Read out the full detector at 30 MHz in HLT1
- Real time alignment and calibration with 10x higher data rate than Run 2
- Full offline-quality reconstruction in "real-time"
- Increase of hadronic trigger efficiency by 2~4



Highest data processing rate of any HEP experiment!

LHCb Upgrade Trigger



Online - Real Time Analysis

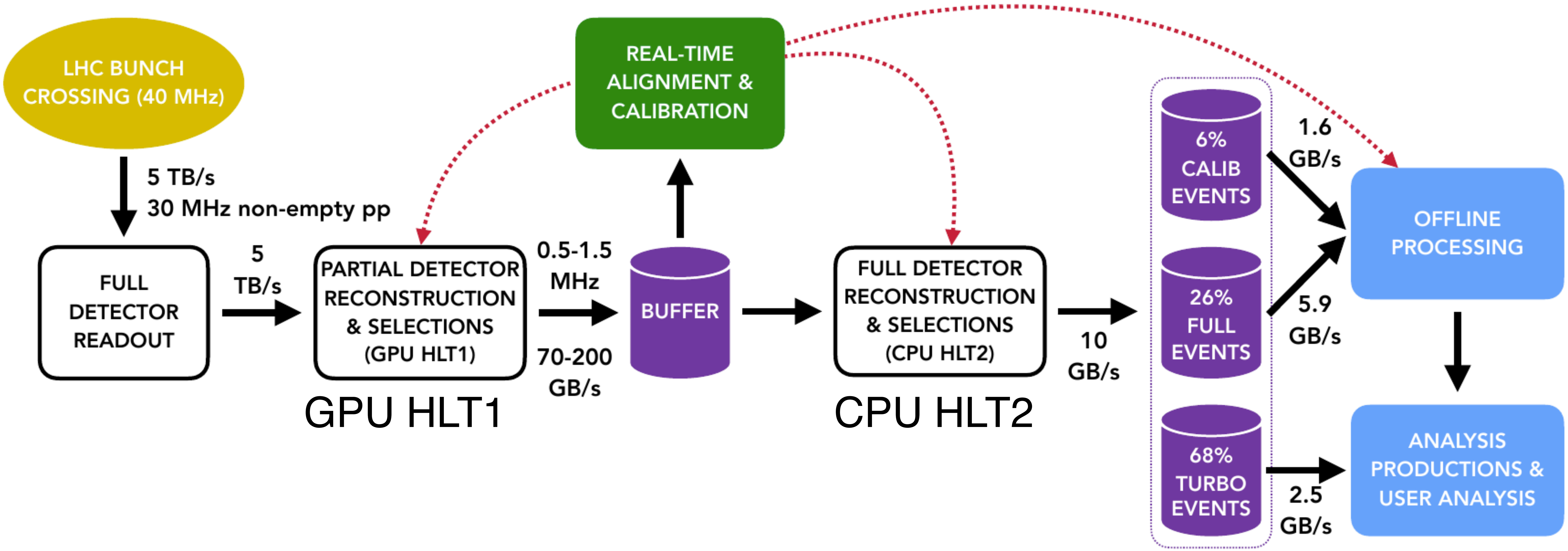
Run 1 & 2 trigger:
background rejection

Upgrade trigger:
background rejection & signals classification

LHCb Data Flow

All numbers related to the dataflow are taken from the LHCb
Upgrade Trigger and Online TDR
Upgrade Computing Model TDR

Alignment

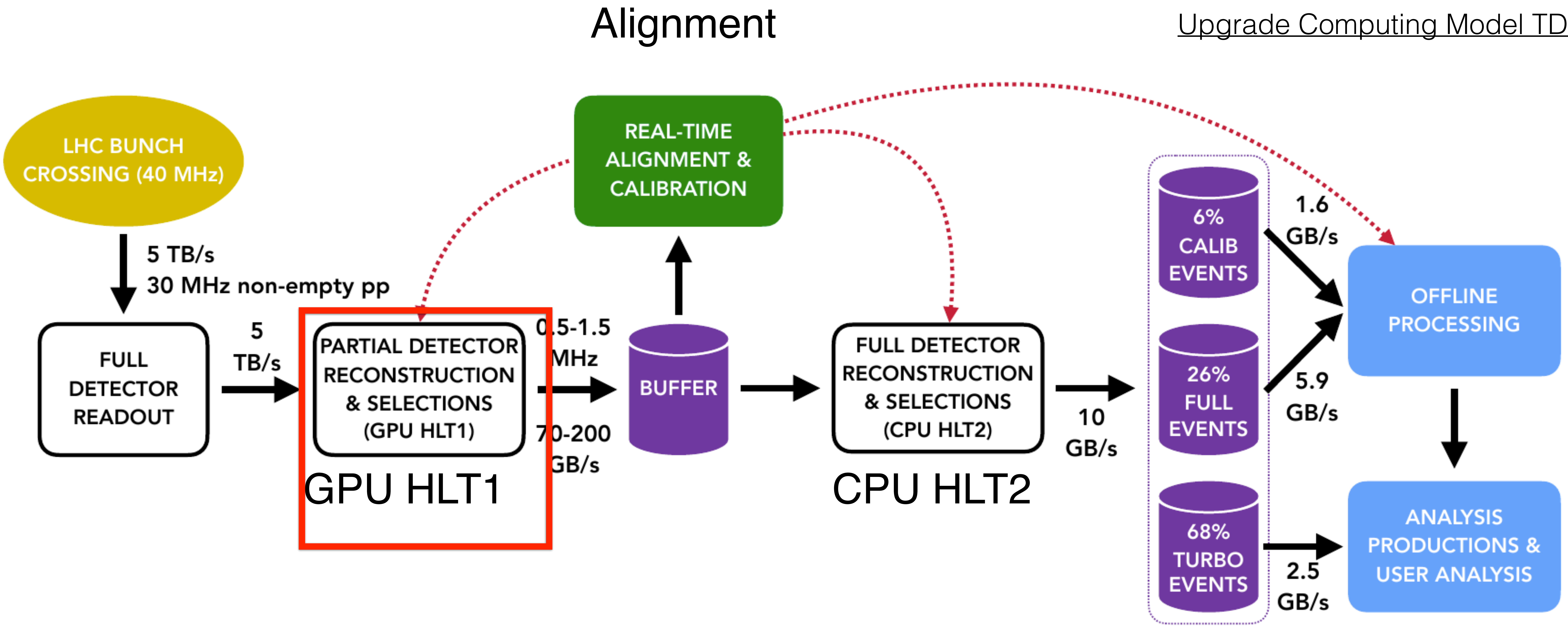


GPU HLT1

CPU HLT2

LHCb Data Flow

All numbers related to the dataflow are taken from the LHCb
Upgrade Trigger and Online TDR
Upgrade Computing Model TDR



First complete high-throughput GPU Trigger for a HEP experiment!

The Allen project (GPU HLT1)

- Named after Frances E. Allen
- Fully standalone software project: https://gitlab.cern.ch/lhcb/Allen
- Framework developed for processing LHCb’s HLT1 on GPUs

- Cross-architecture compatibility via macros & few coding guide lines
  - GPU code written in CUDA
  - runs on CPUs, Nvidia GPUs(CUDA), AMD GPUs (HIP)

Publications: Comput Softw Big Sci 4, 7 (2020), Technical Design Report (2020), Comput Softw Big Sci 6, 1(2022), EPJ Web of Conferences 251, 04009 (2021)

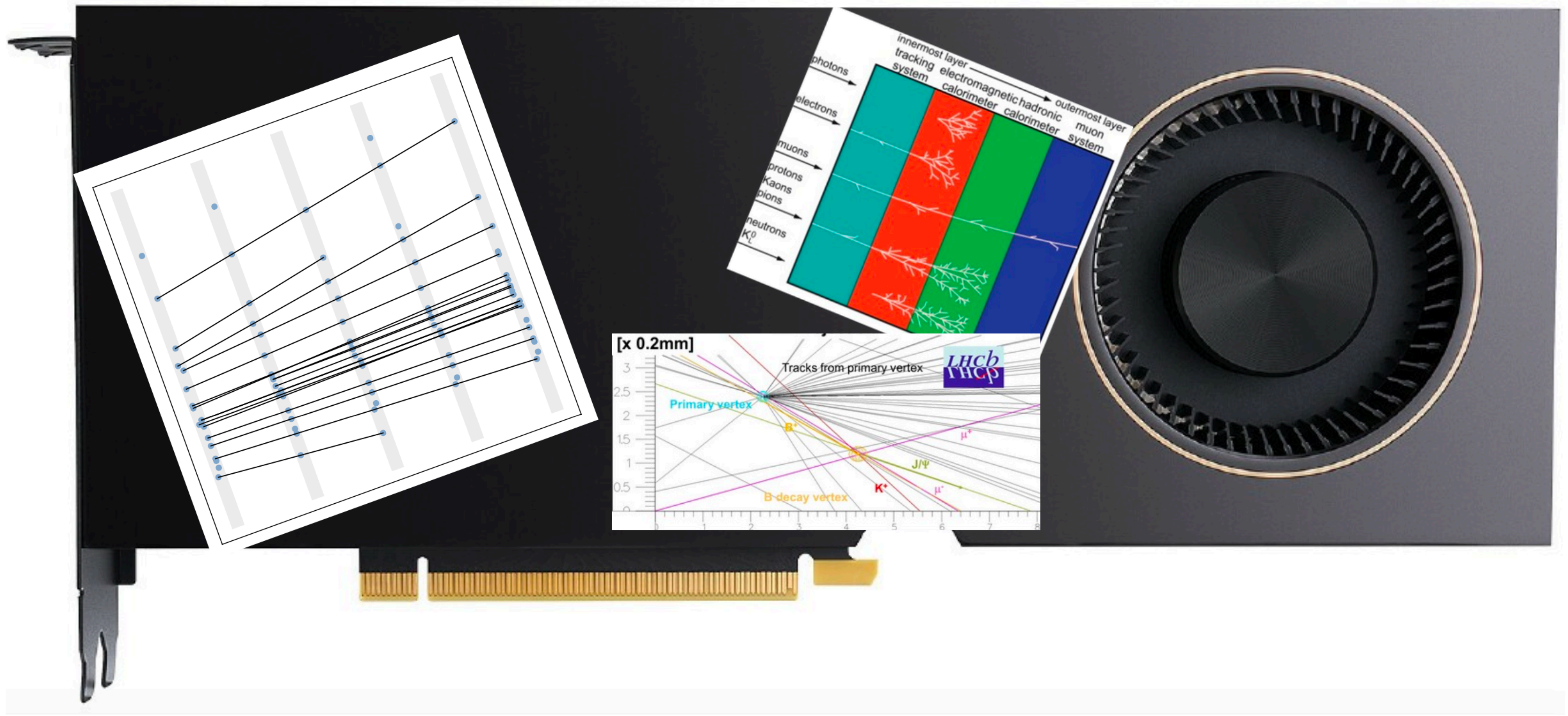
Minimise copies to / from GPU

Server

Raw data

Selection decisions

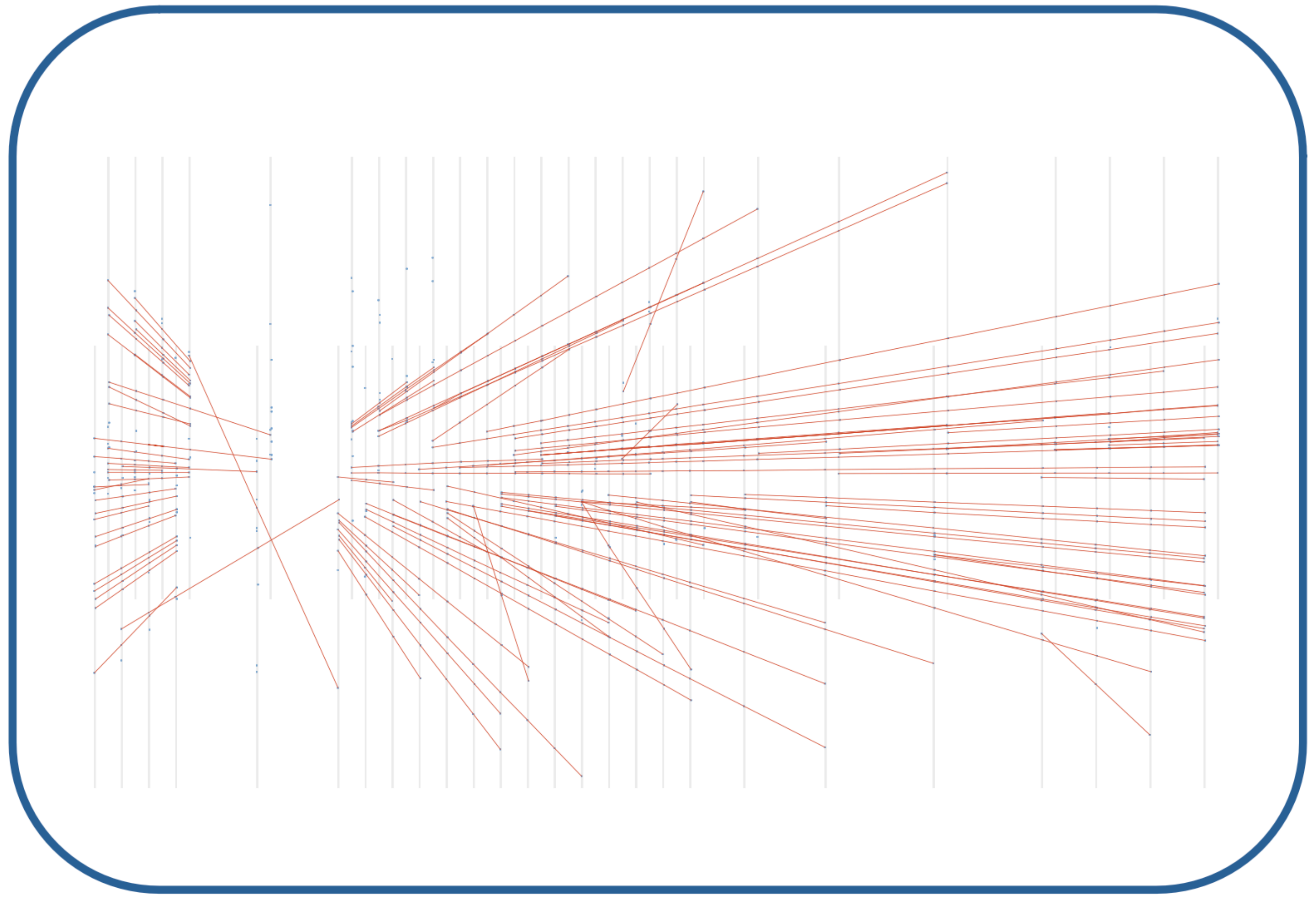
GPU



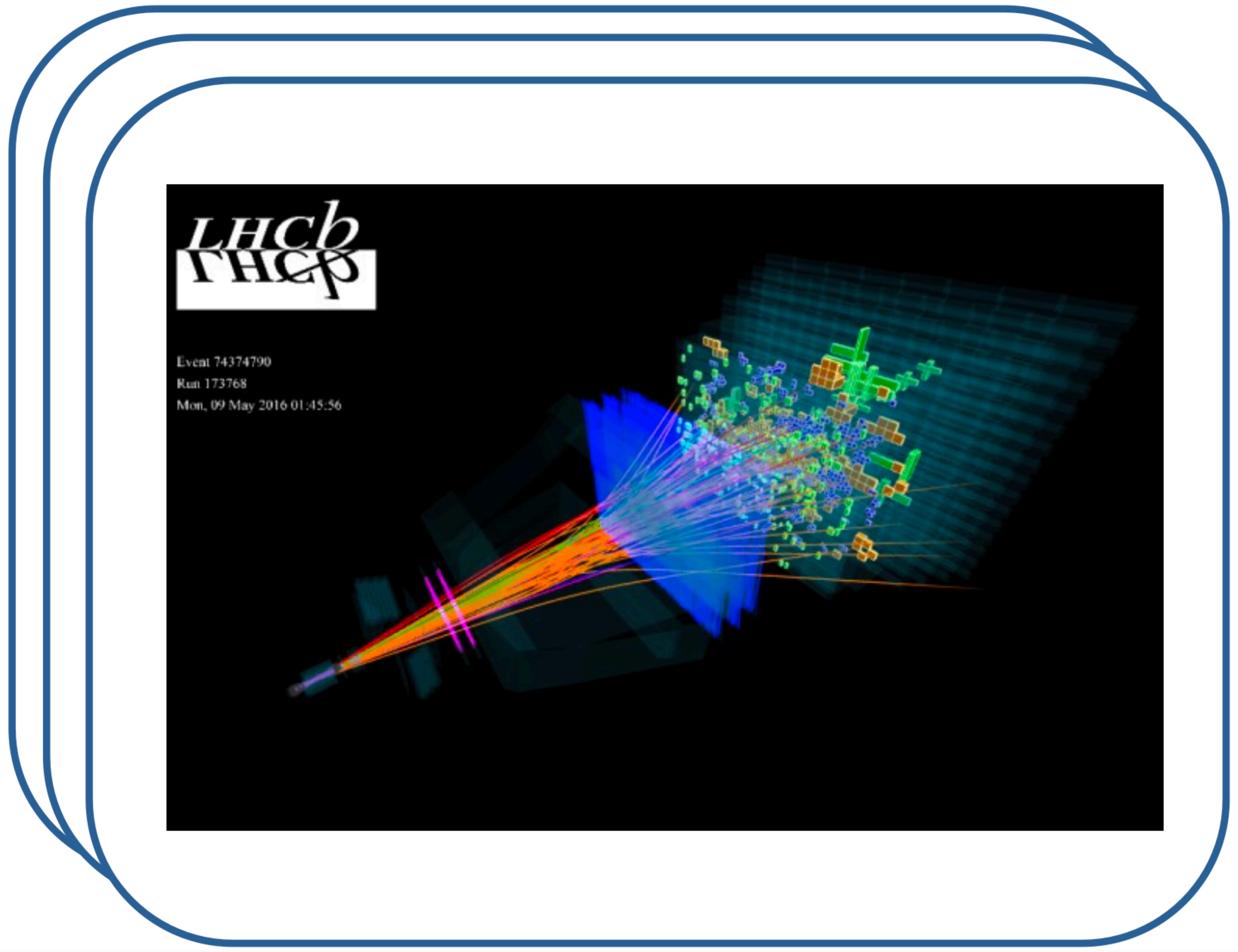
from Dorothea's slides

Three levels of parallelisation

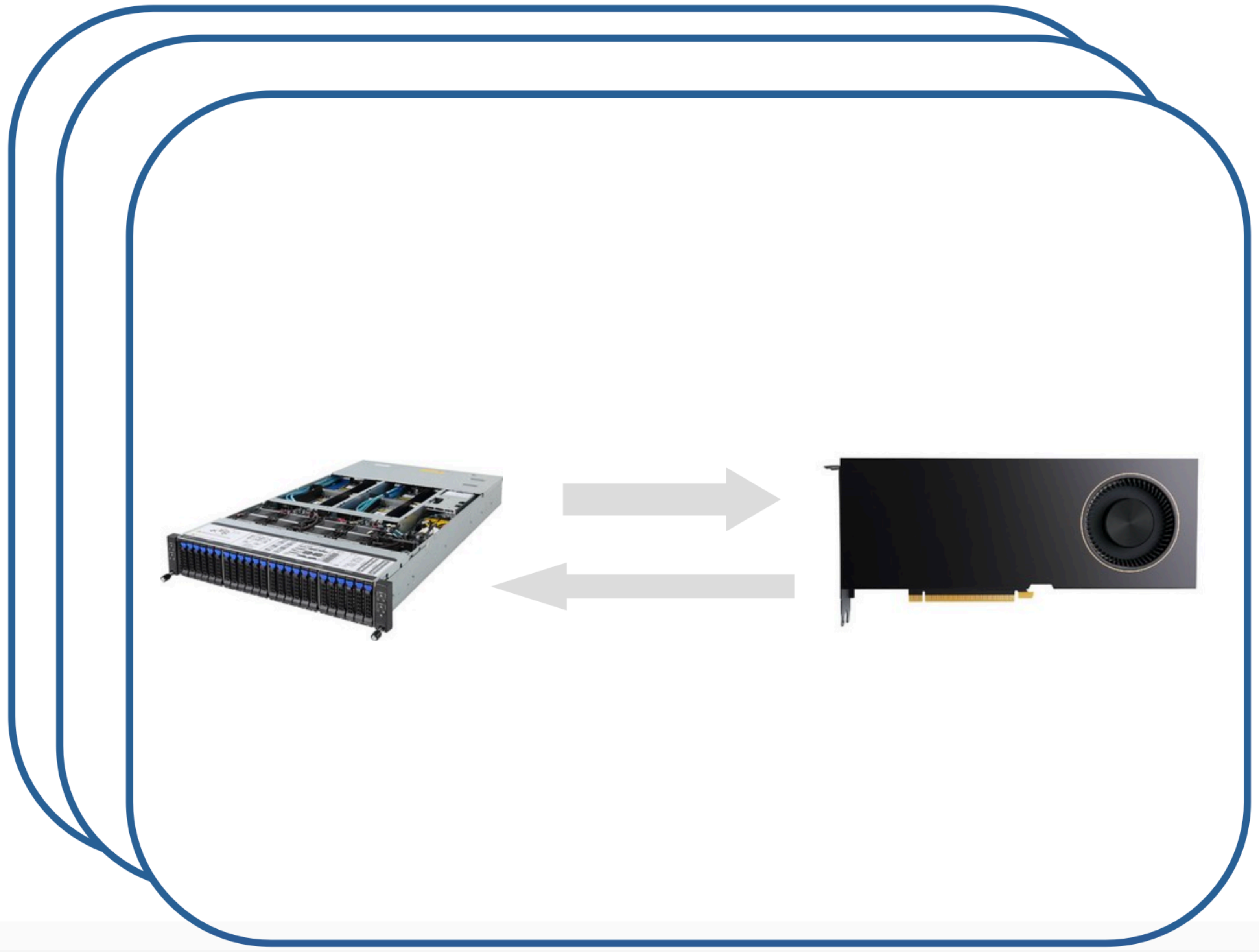
Intra-collision: Tracks, vertices, …



Proton collisions

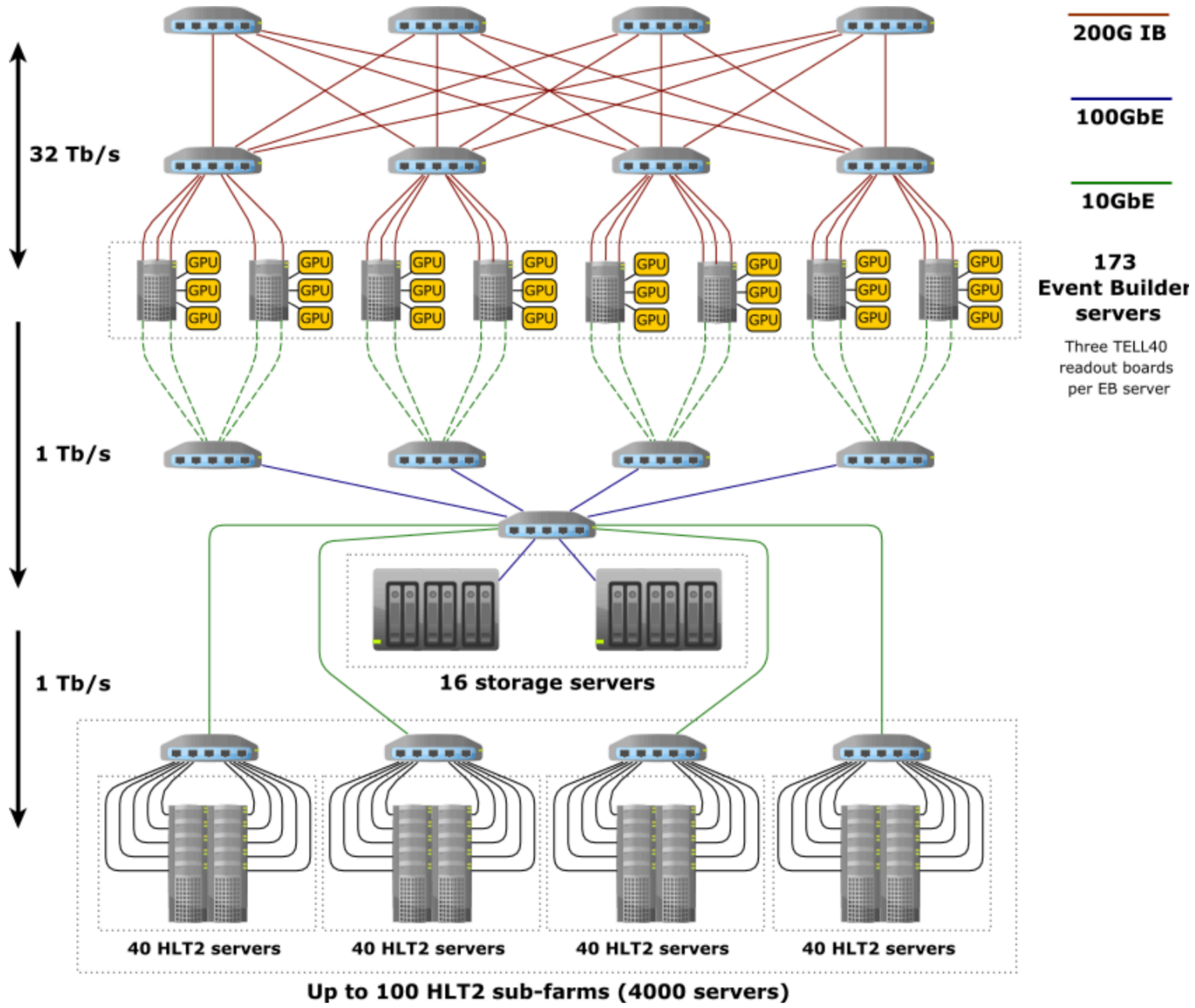


Collision batches



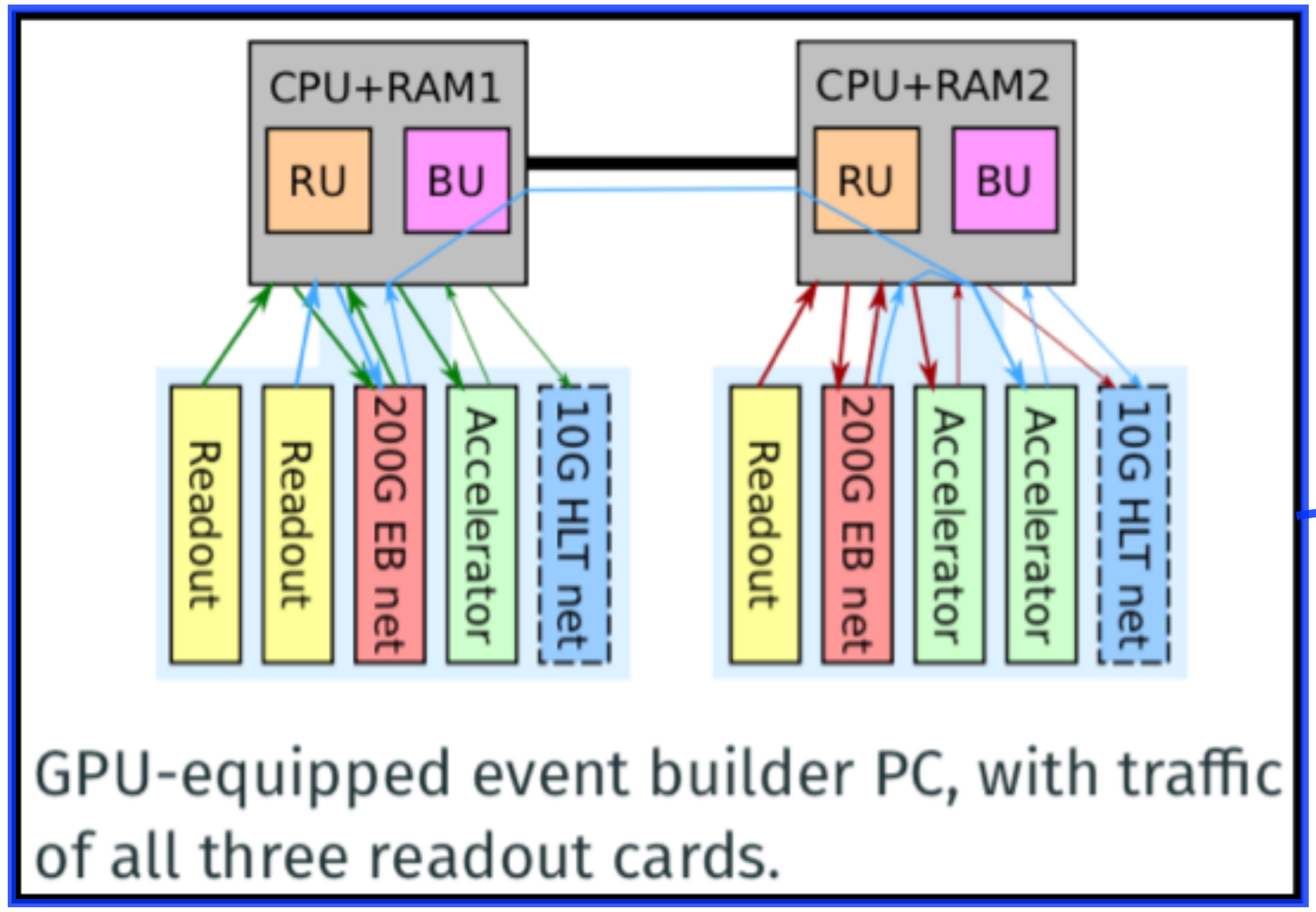
from Dorothea's slides

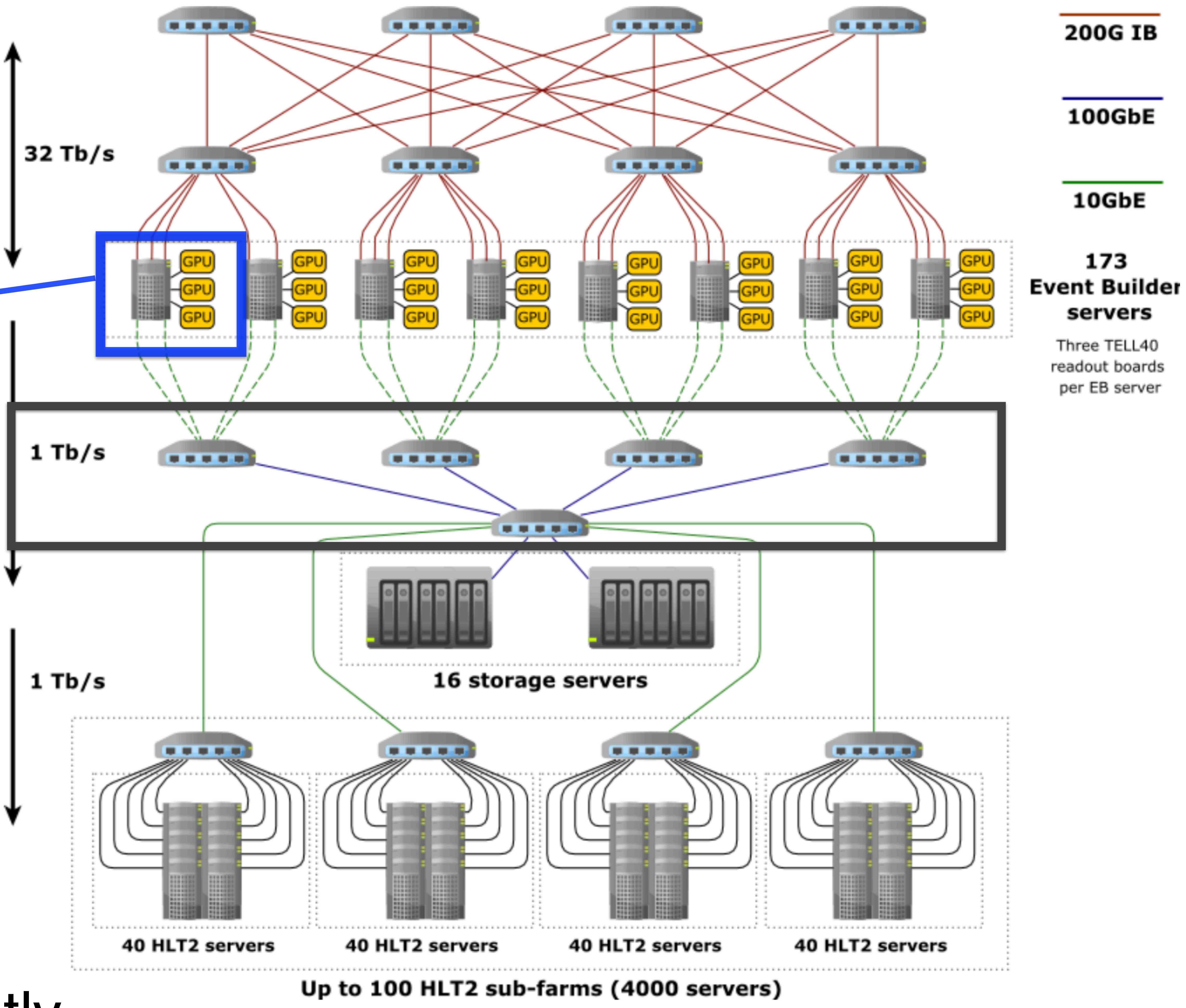
Allen design



Computing and Software for Big Science(2020)4:7

Allen design

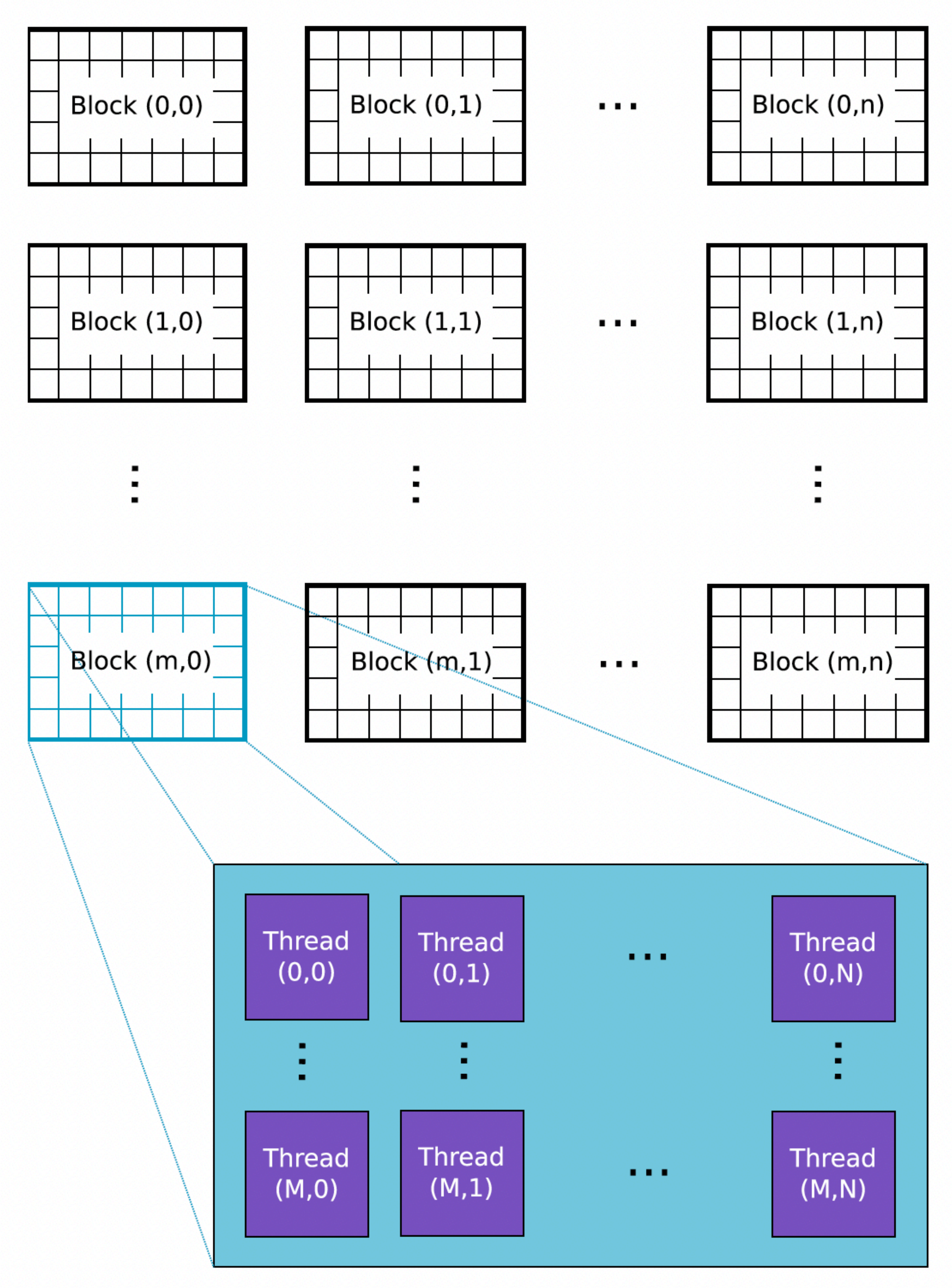




- Take as input LHCb raw data (5TB/s) at 30 MHz
- Each Event-builder (EB) holds 3 GPU cards via PCIe slots
- ~500 NVIDIA RTX 5000 GPUs installed
- Reduce data volume by a factor 30-60, significantly reducing the networking from EB to CPU farms

Computing and Software for Big Science(2020)4:7

Allen design

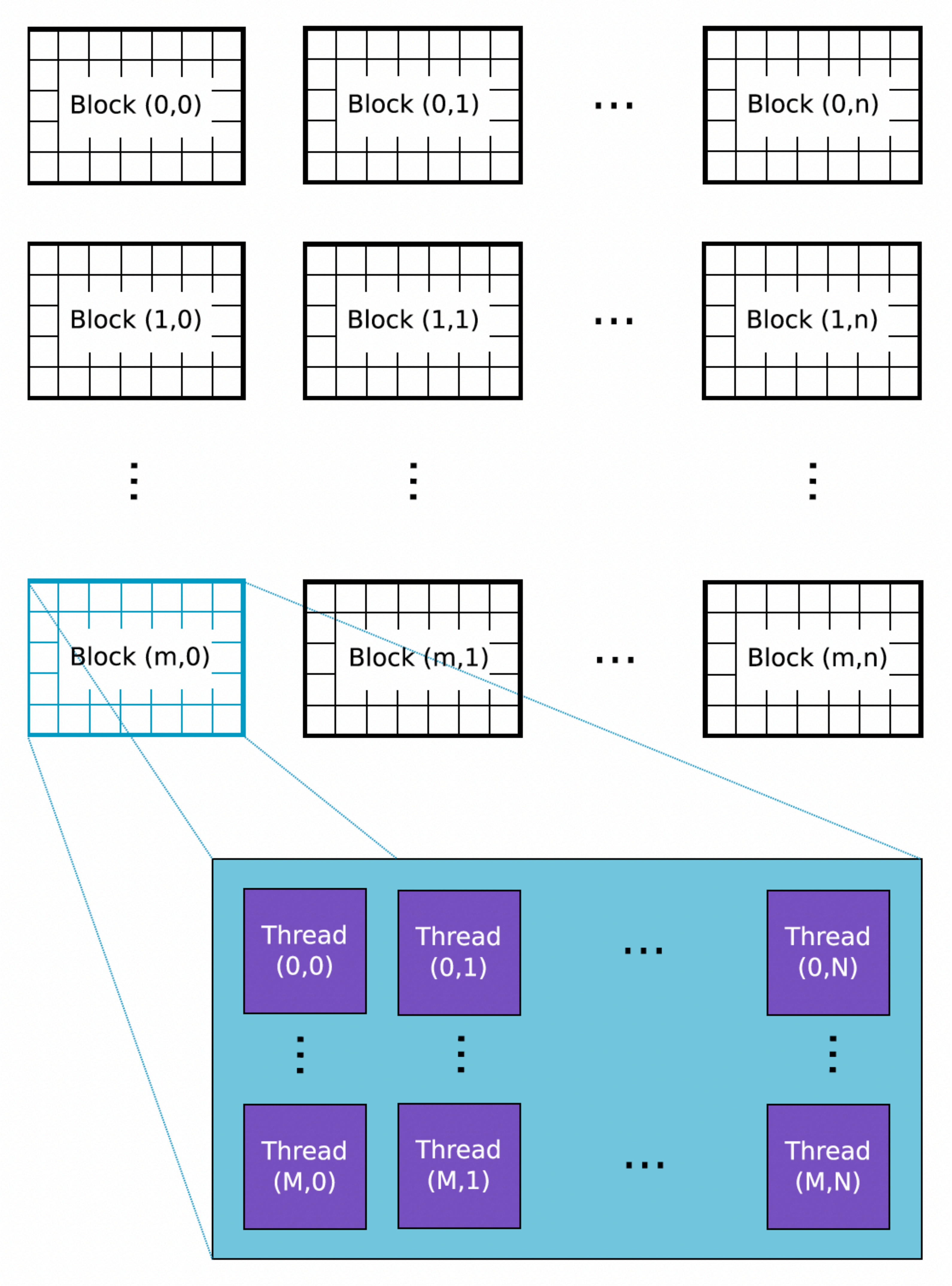


sharing a common memory

- Threads grouped into blocks, forming a grid to execute one kernel on the GPU
- Every GPU receives complete events from an EB unit and processing several thousand events at once

Computing and Software for Big Science(2020)4:7

Allen design

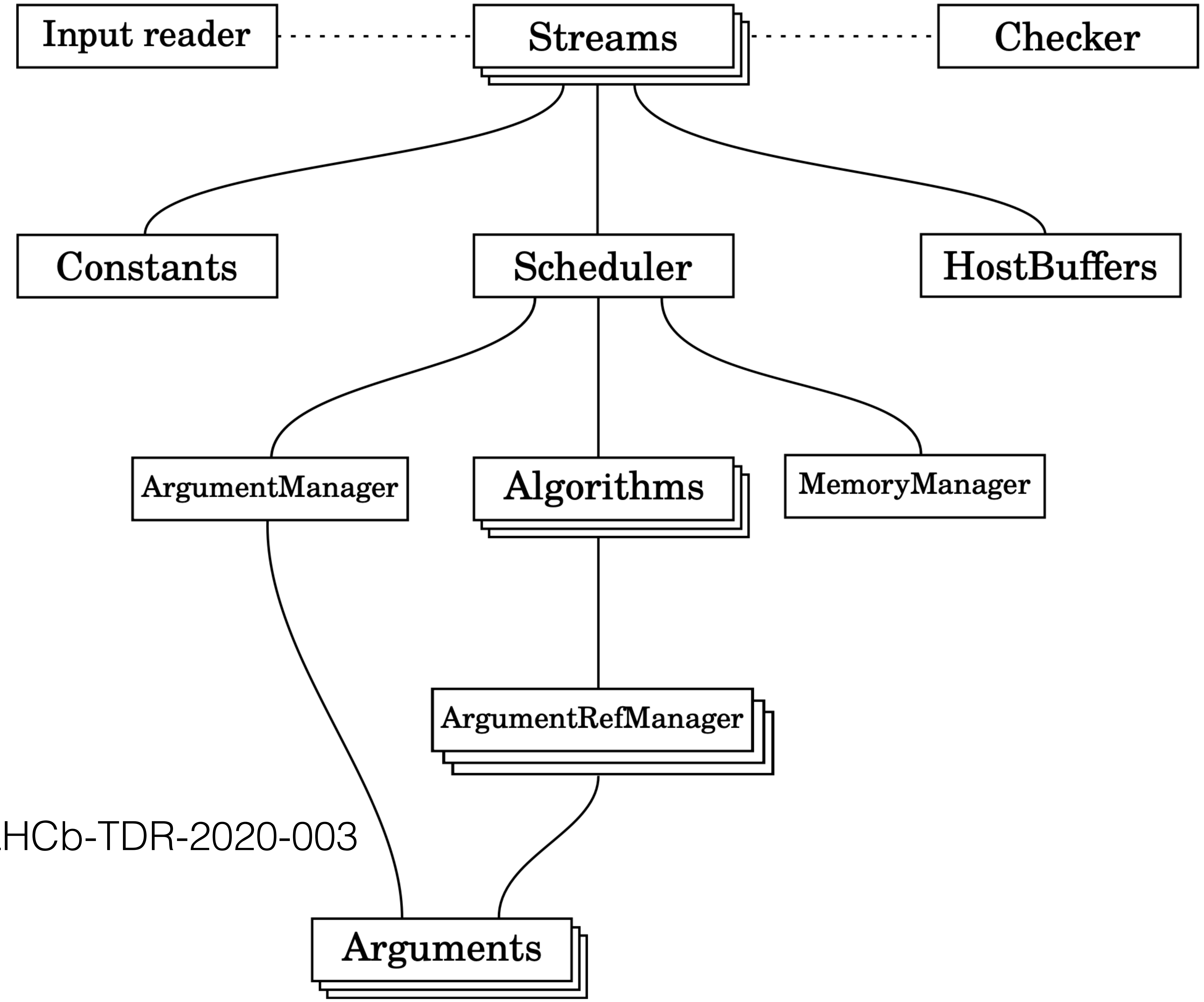


sharing a common memory

- Threads grouped into blocks, forming a grid to execute one kernel on the GPU
- Every GPU receives complete events from an EB unit and processing several thousand events at once
- Raw detector data copied to GPU, processed with the full HLT1 sequence
  - LHCb raw events ~ 100 kB
  - no limitation in PCIe connection between the CPU and GPU
  - only selected events copied from GPU to CPU (a reduction of a factor 30-60)
  - no Intra-GPU communication as each event is independent

Computing and Software for Big Science(2020)4:7

Allen design



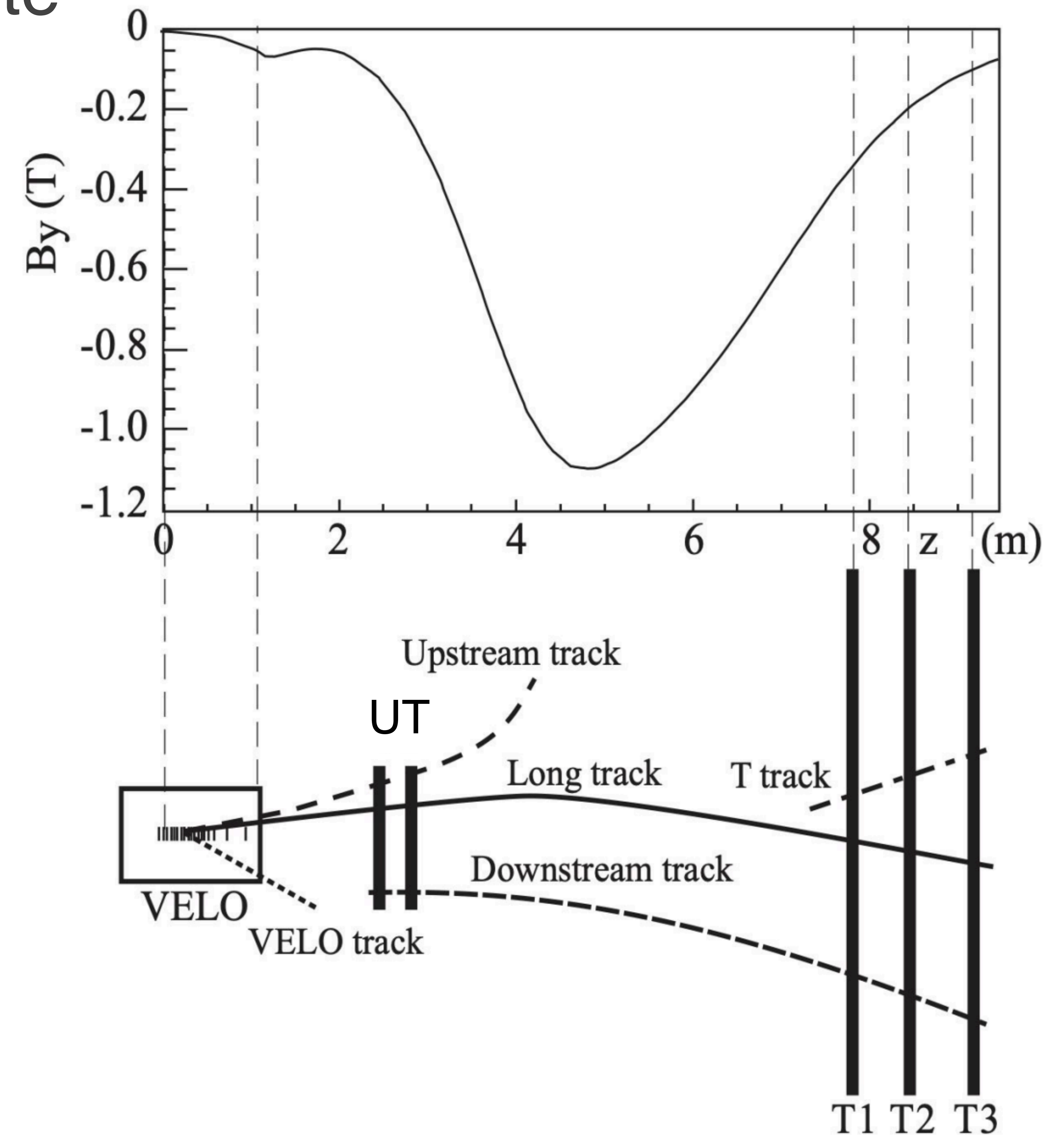
LHCb-TDR-2020-003

- Algorithms sequences defined in python and generated at run-time
- Multi-event processing with dedicated scheduler
- Memory manager allocates large chunk of GPU memory at start-up
- Reconstruction algorithms re-designed for parallelism (SOA) and low memory usage: O(MB) per core

Track reconstruction with GPU

- Filters the 30 MHz pp collision to 1 MHz
- Partial reconstruction using hits from VELO, (UT), SciFi & Muon
  - → High momentum long charged track reconstruction & muon identification
  - → Few inclusive single and two-track selections to reduce rate

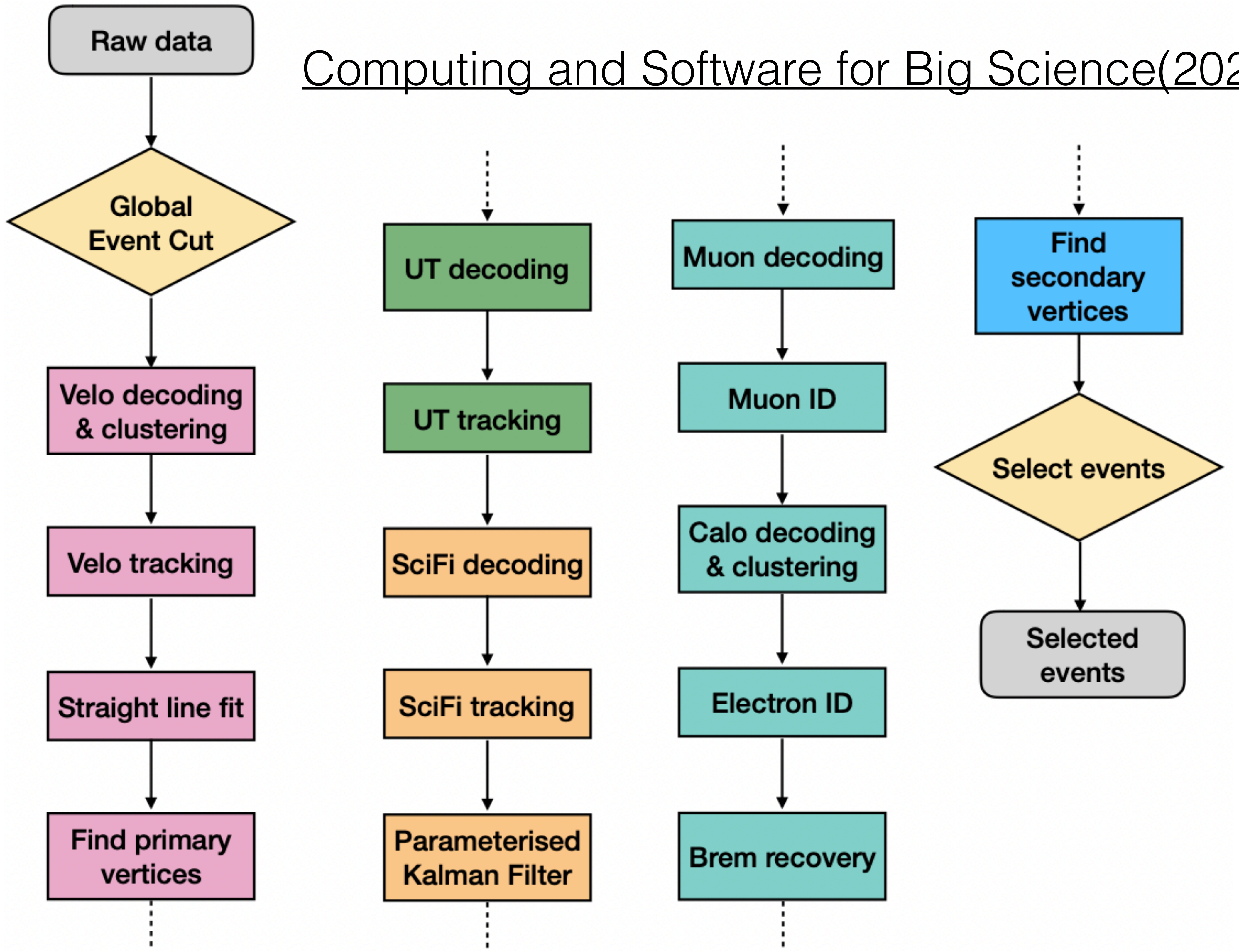
Computing and Software for Big Science(2020)4:7

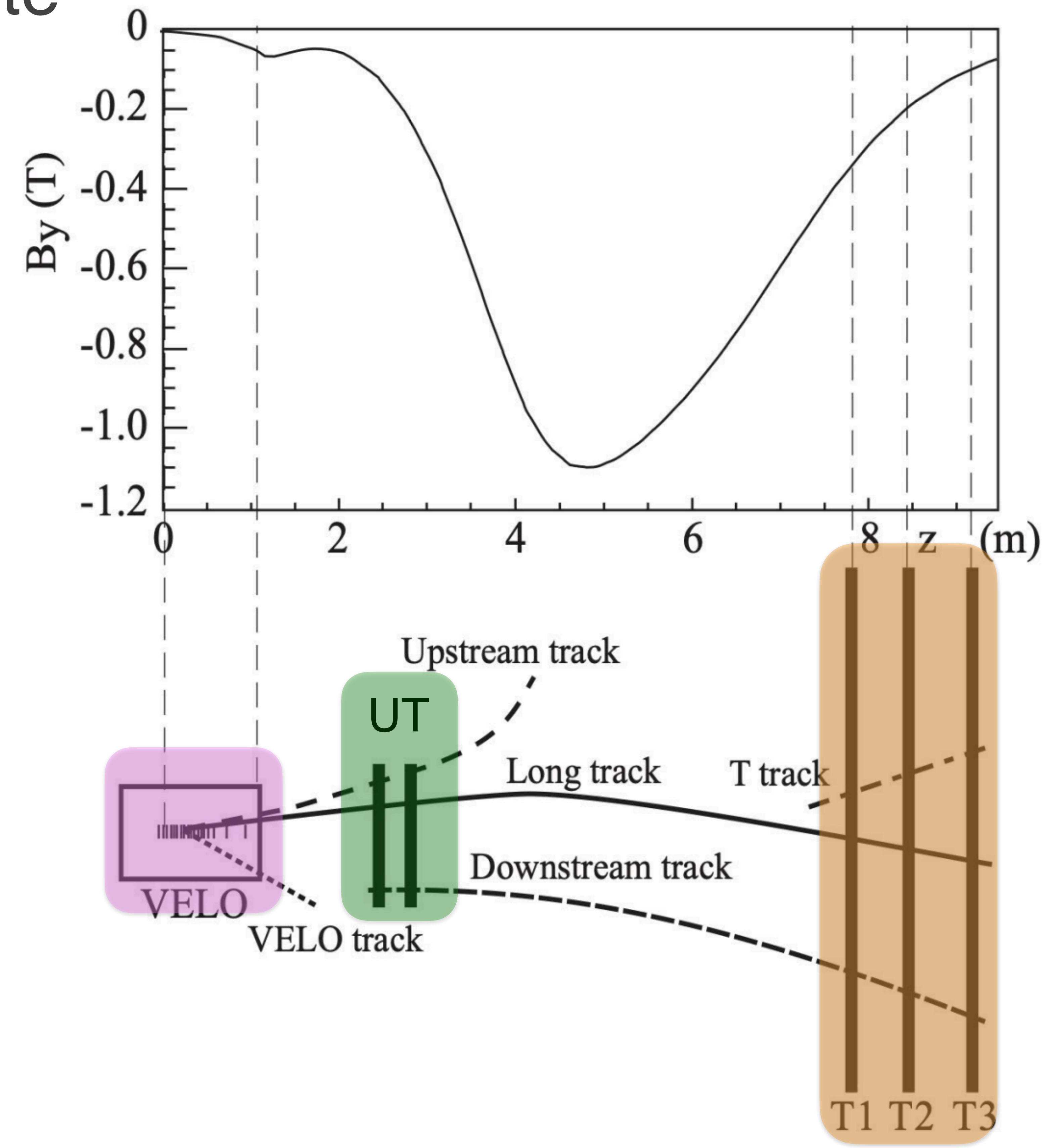


Track reconstruction with GPU

- Filters the 30 MHz pp collision to 1 MHz
- Partial reconstruction using hits from VELO, (UT), SciFi & Muon
  - → High momentum long charged track reconstruction & muon identification
  - → Few inclusive single and two-track selections to reduce rate

Computing and Software for Big Science(2020)4:7

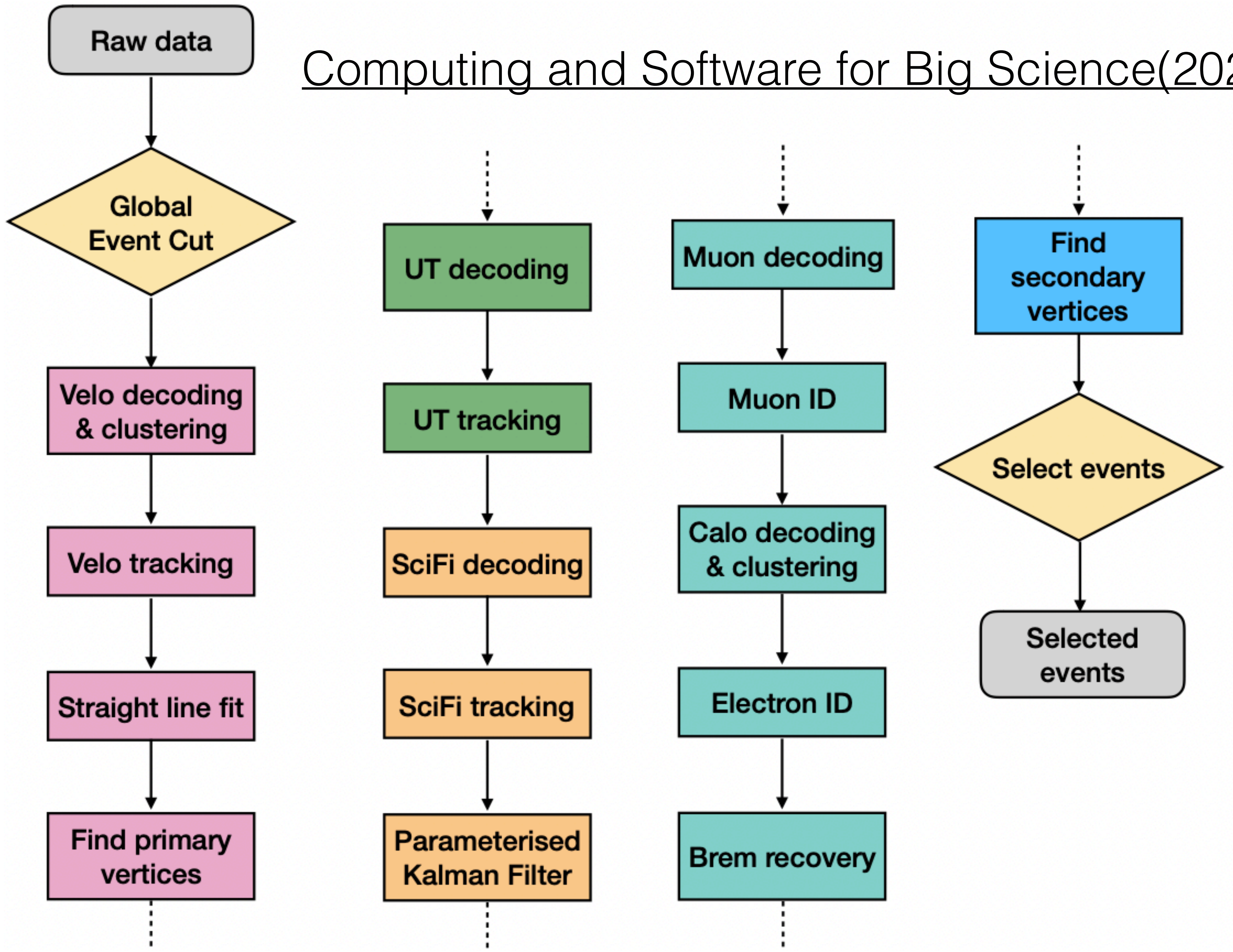


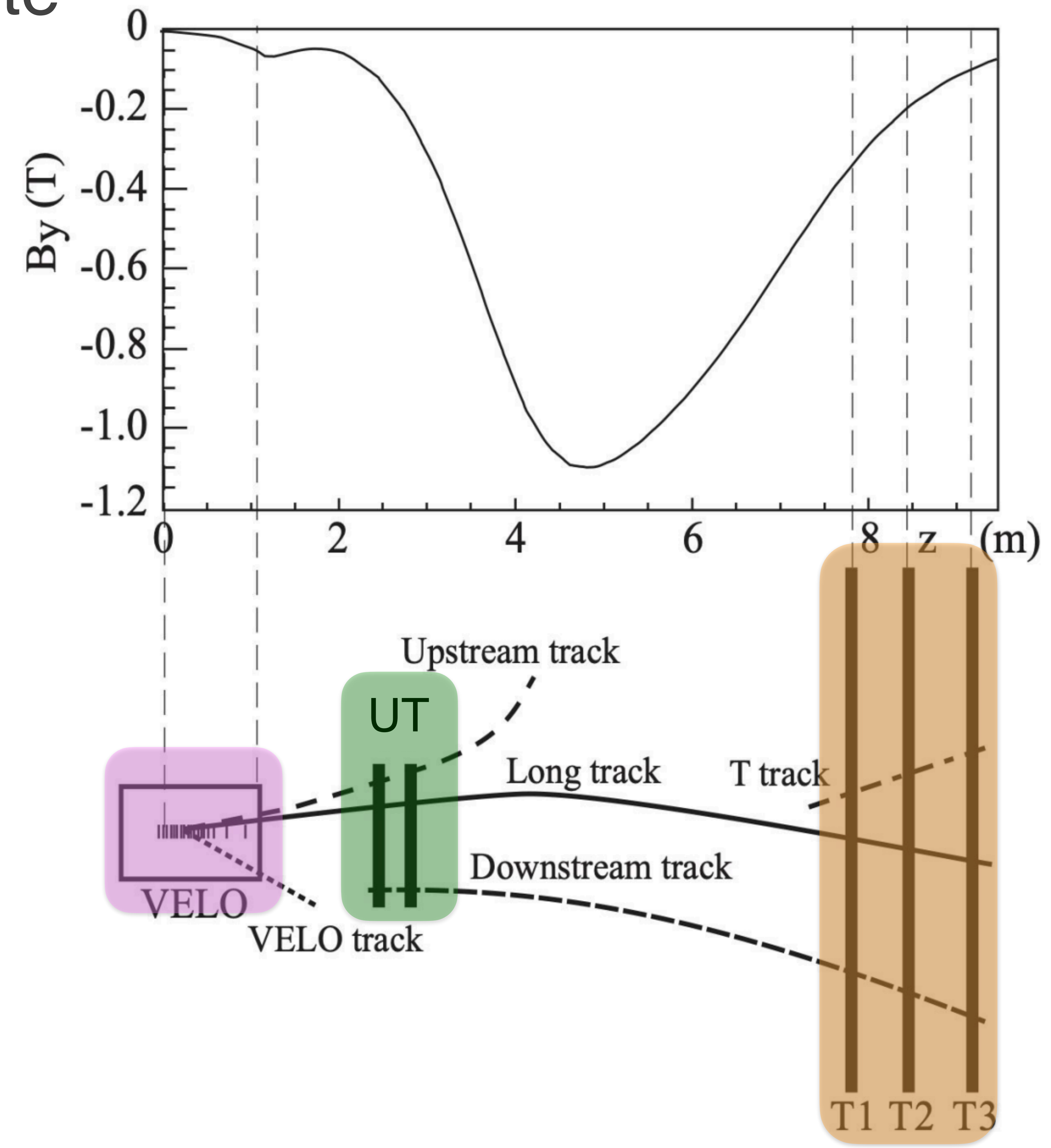


Track reconstruction with GPU

- Filters the 30 MHz pp collision to 1 MHz
- Partial reconstruction using hits from VELO, (UT), SciFi & Muon
  - → High momentum long charged track reconstruction & muon identification
  - → Few inclusive single and two-track selections to reduce rate

Computing and Software for Big Science(2020)4:7

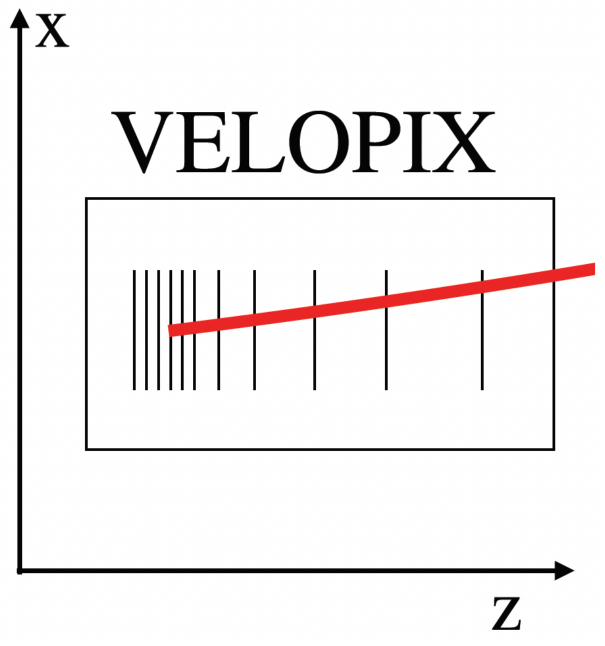




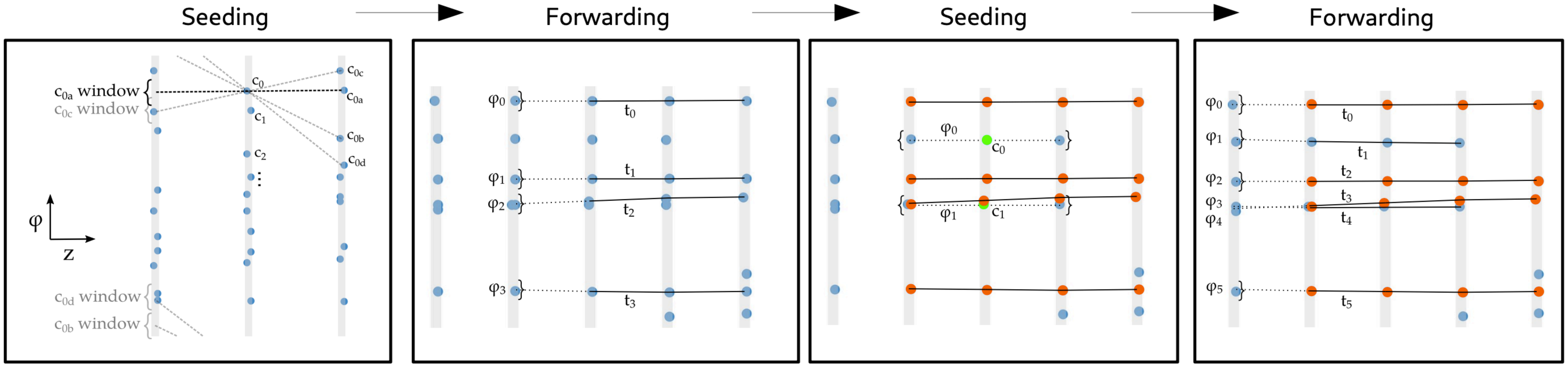
- Velo clustering with FPGA is used in data taking

More details in the next talk by Ao

VELO: Tracking



- 26 layers of silicon pixels detector

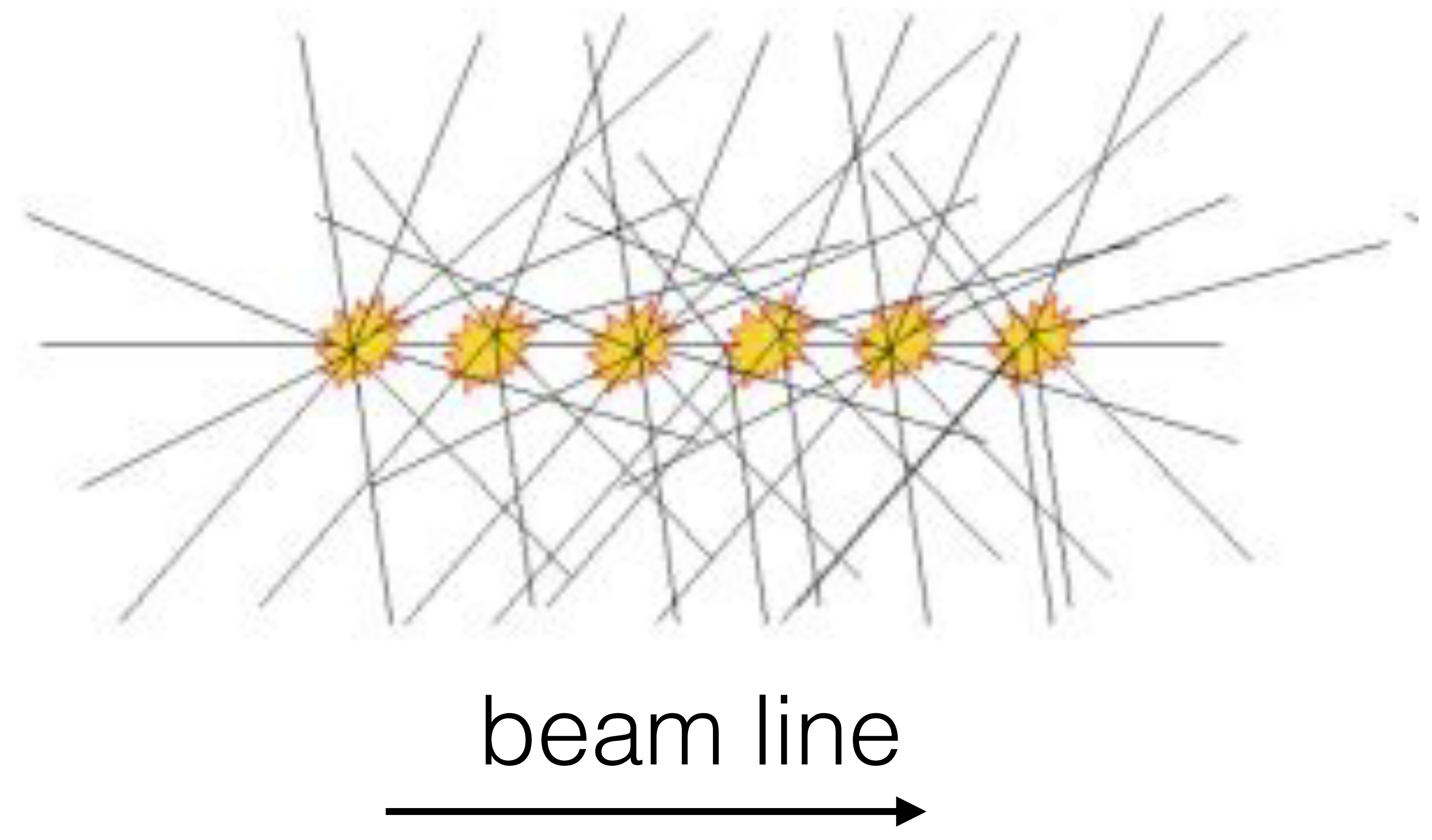


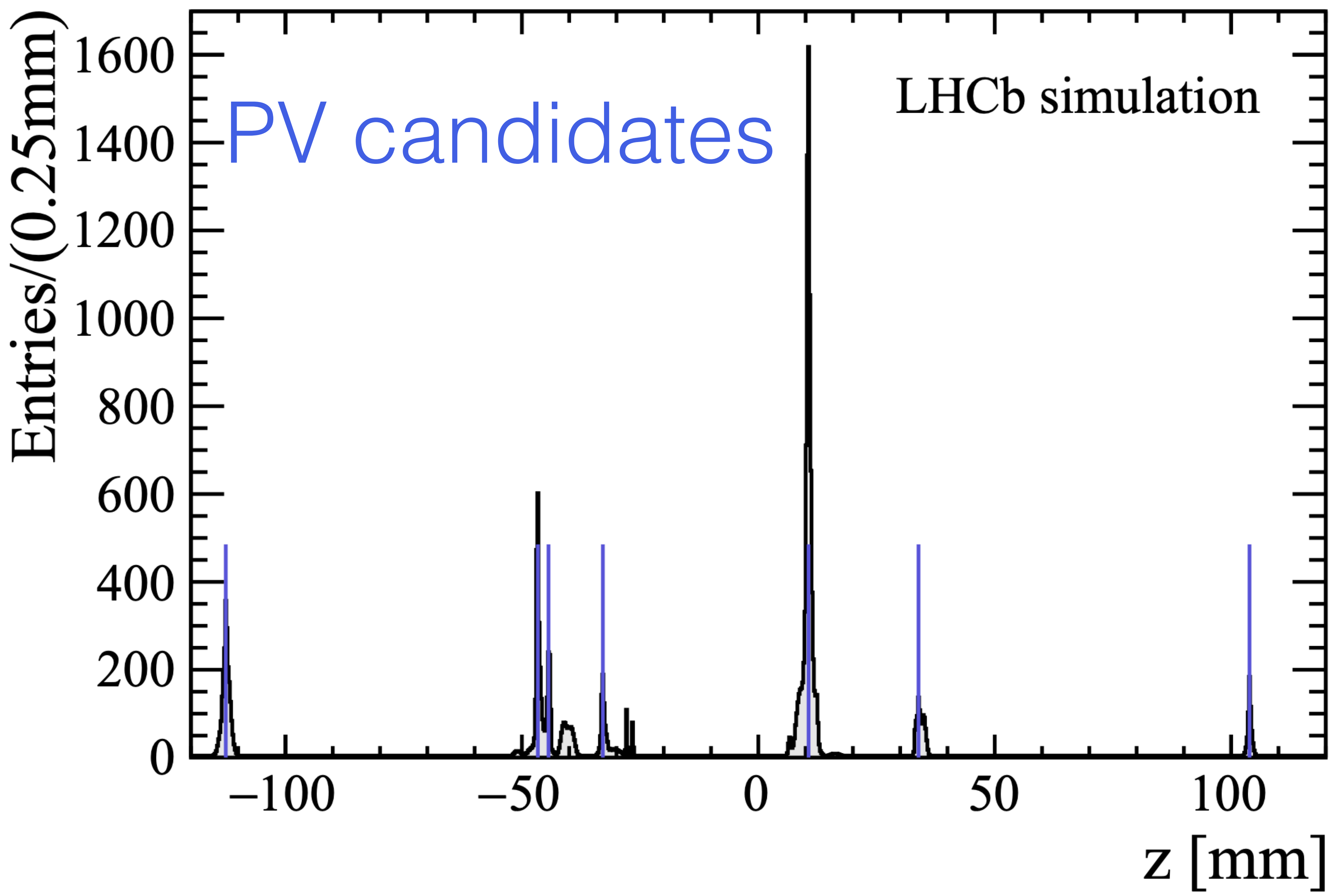
- Build “triplets” of three hits on consecutive layers → parallelisation
- Choose them based on alignment in phi
- Hits sorted by phi →memory accesses as contiguous as possible: data locality
- Extend triplets to next layer → parallelisation

D D. Campora, N. Neufeld, A. Riscos Núñeez: “A fast local algorithm for track reconstruction on parallel architectures”, IPDPSW 2019

VELO: Vertex reconstruction

LHCb-Figure-2020-005

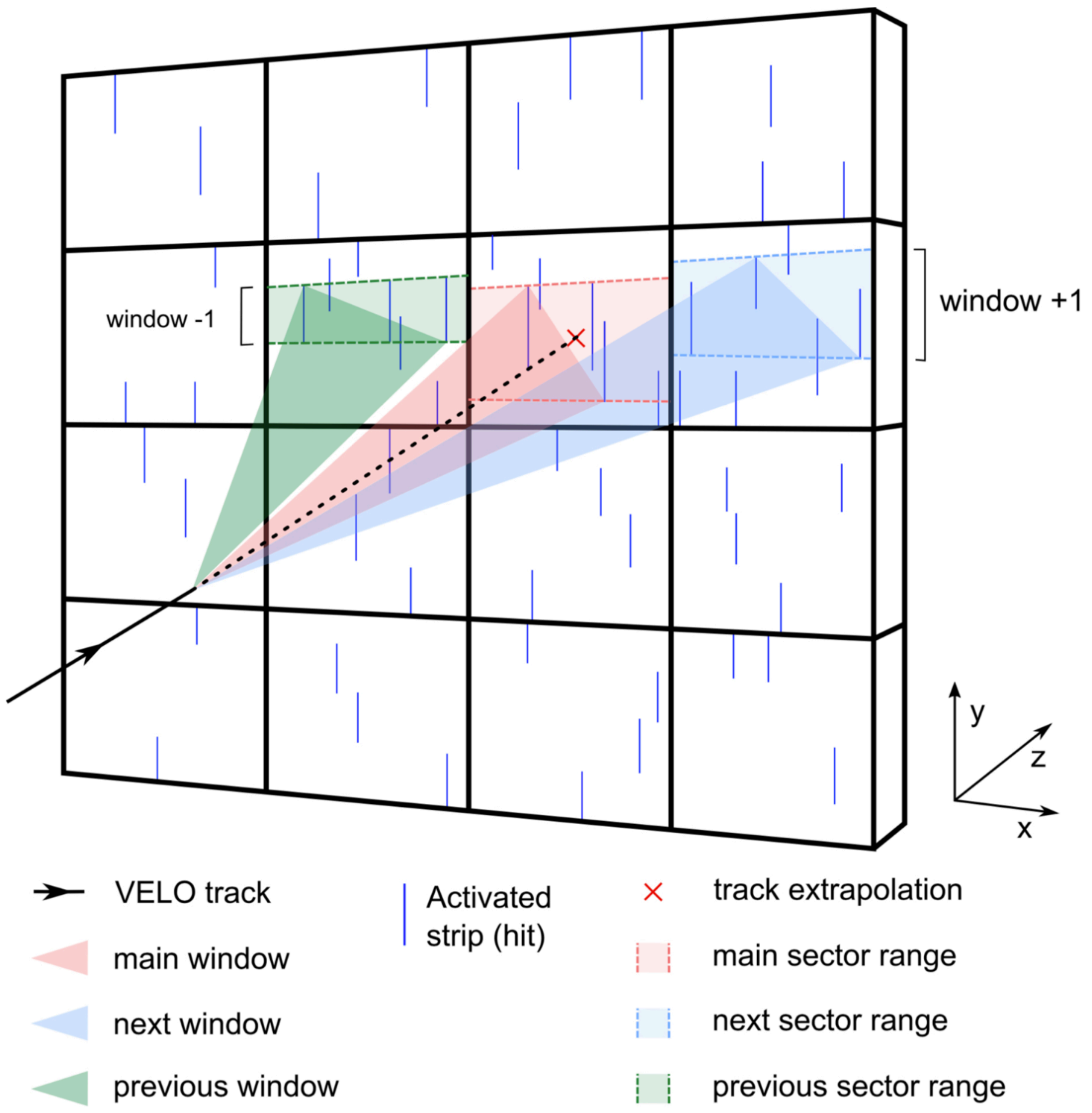




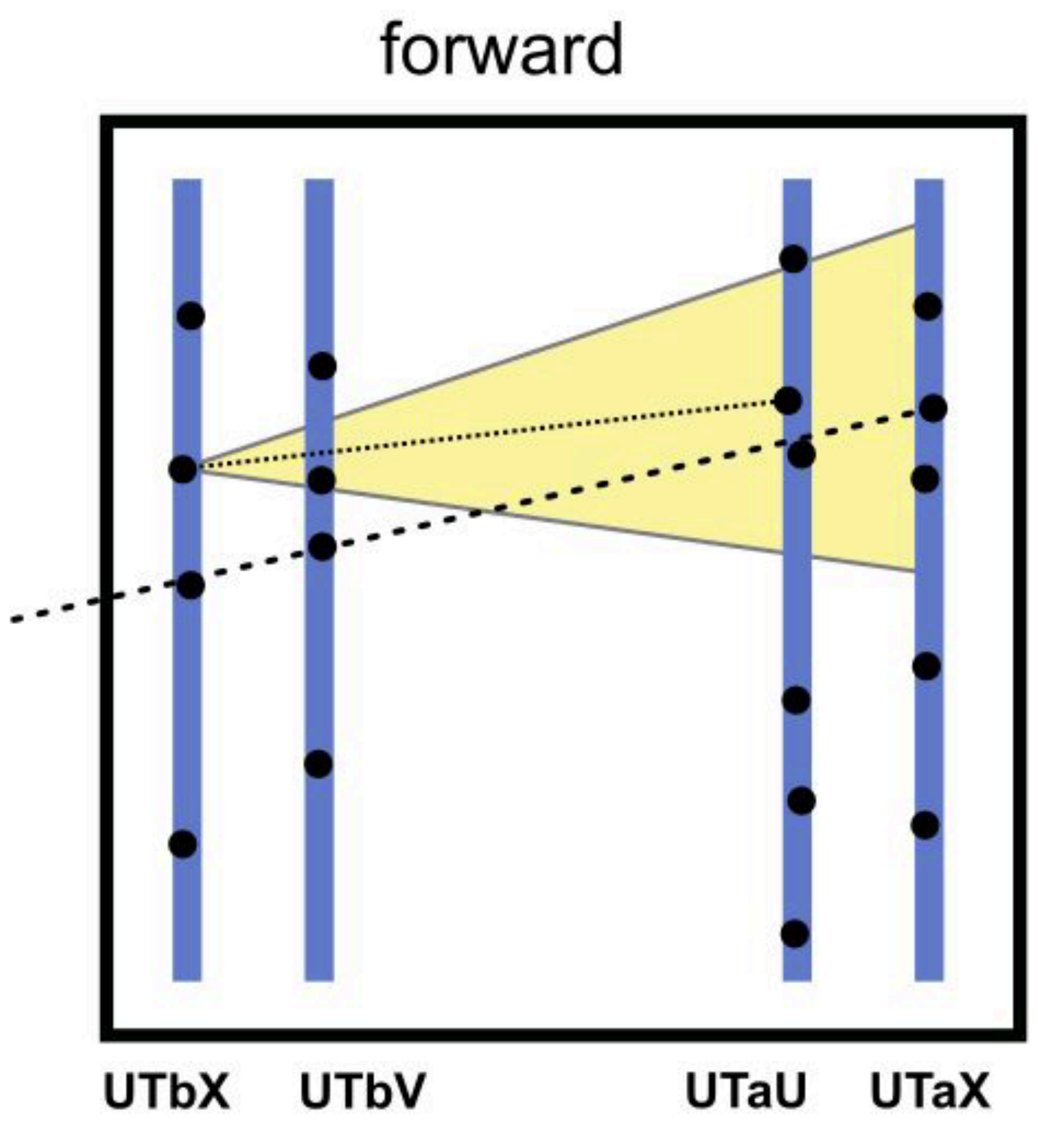
- Primary vertices (PVs) are extended along the beam direction (z-axis)
- Histogram the tracks' z position closest to the beam line
- Every track contributes to every PV candidate with a weight
- No inter-dependence between PV candidates, as every track contributes to every PV
- PV fitting can be done in parallel for every candidate

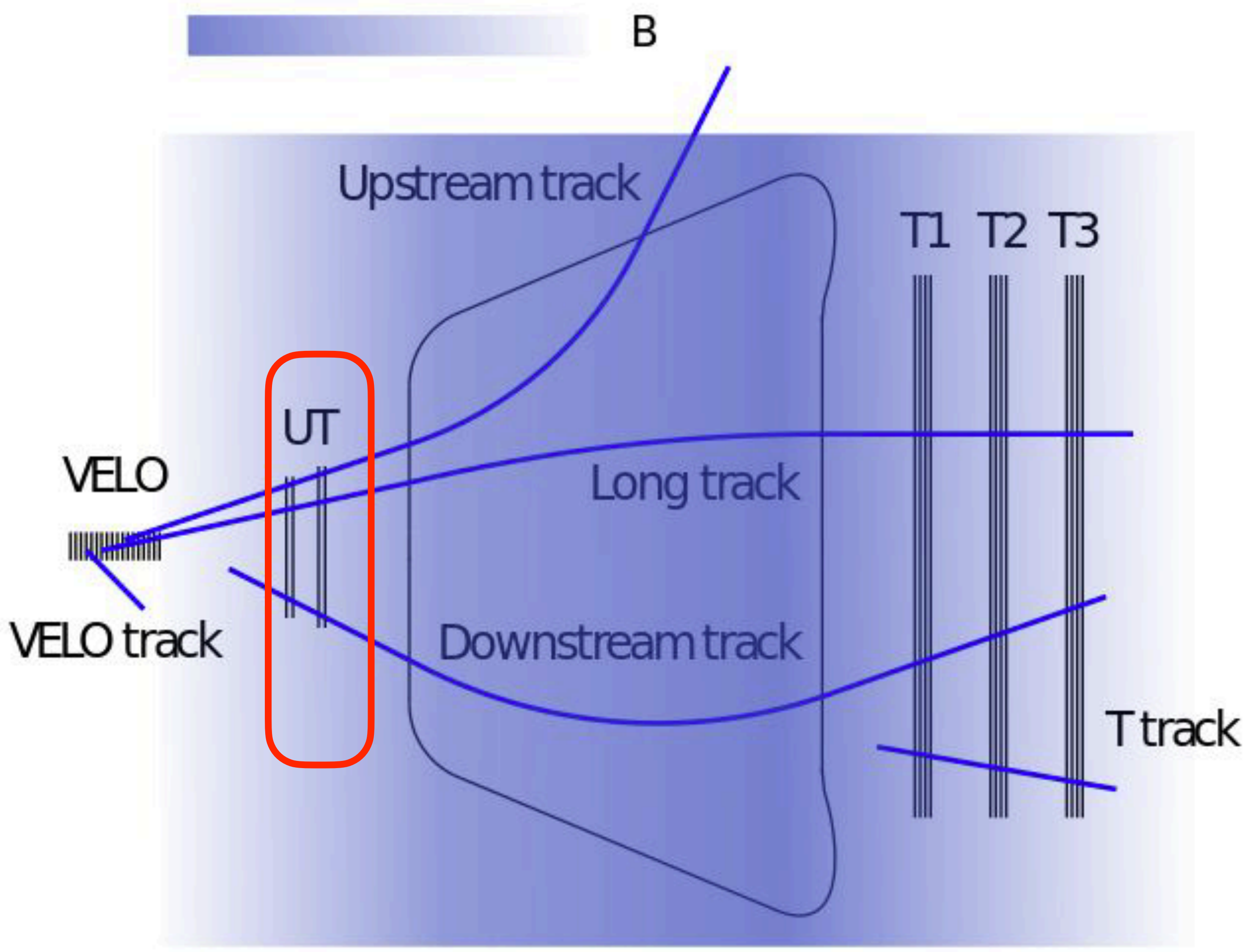
UT: Tracking

- Four layers of silicon strip detector

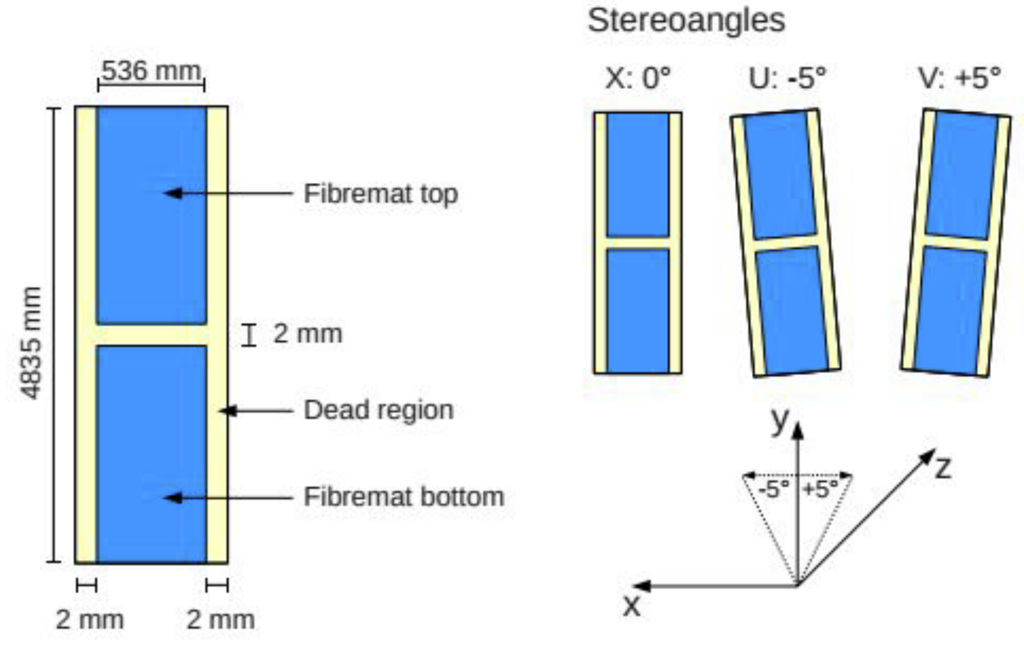


- Extrapolate VELO tracks to the UT planes based on lookup table for minimum momentum requirement
- Define search regions in each UT plane → parallelisation
- Trackless finding inside windows from 4 layers building combinatorics → parallelisation

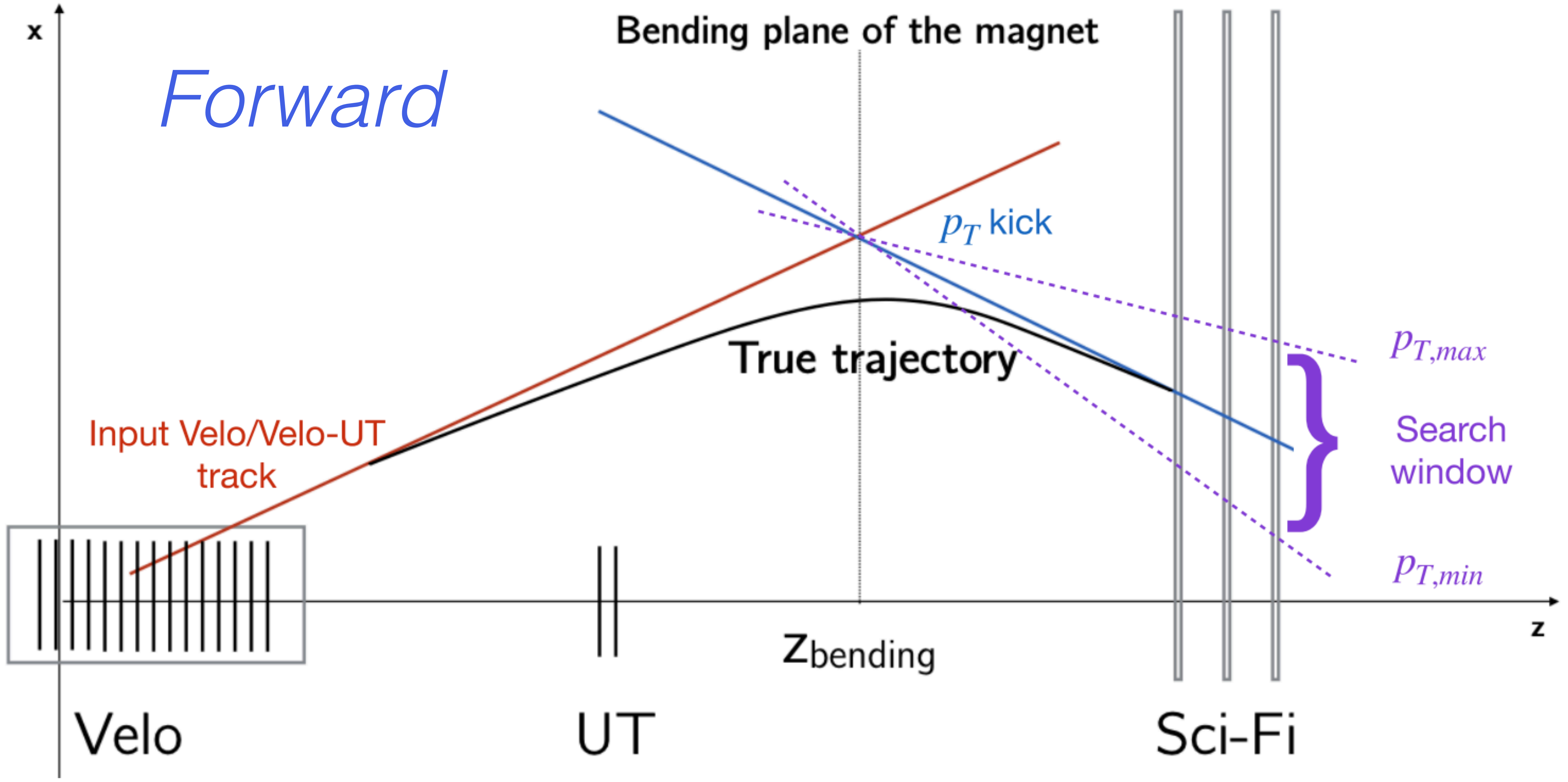


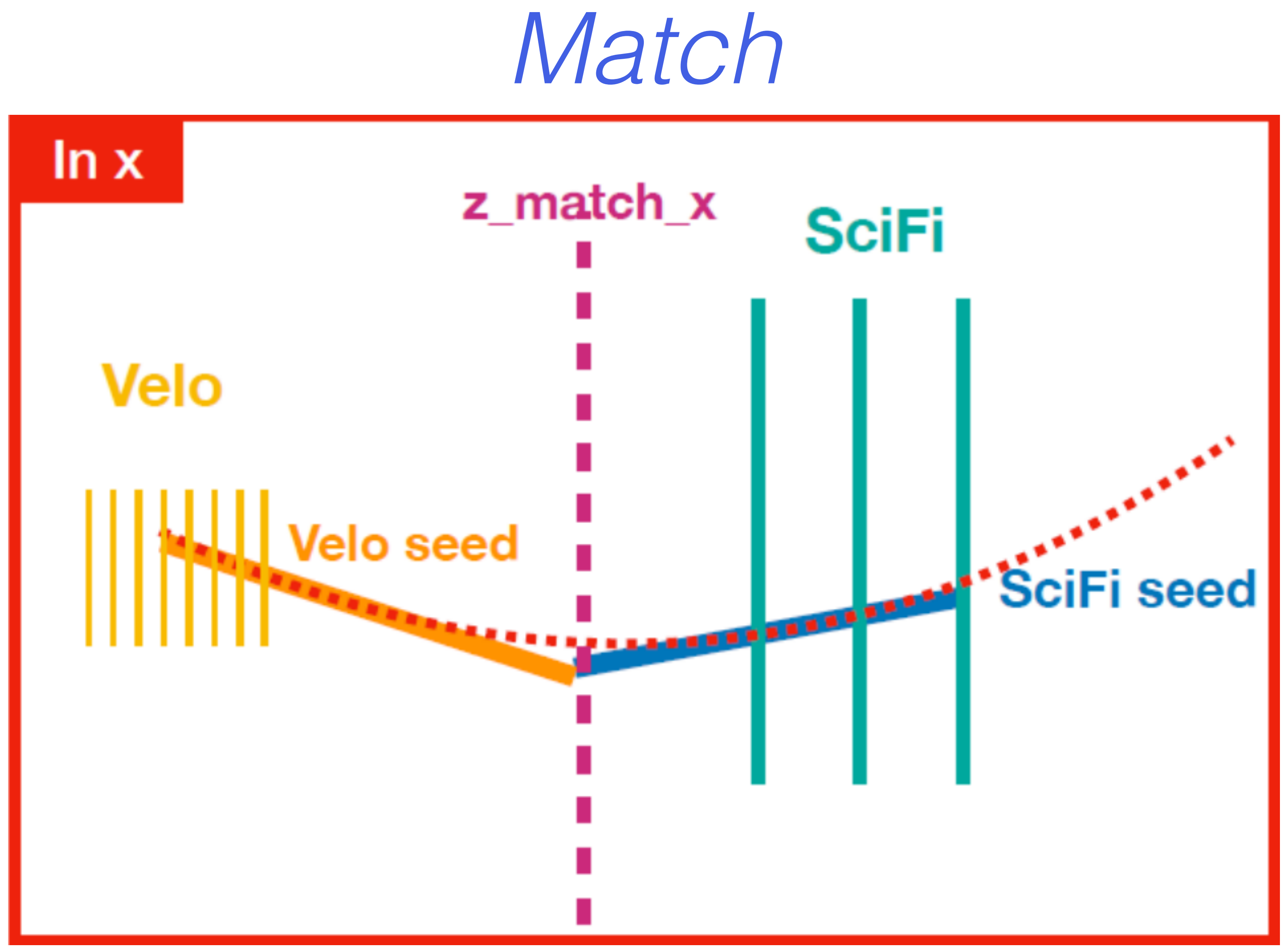


SciFi: Long track reconstruction

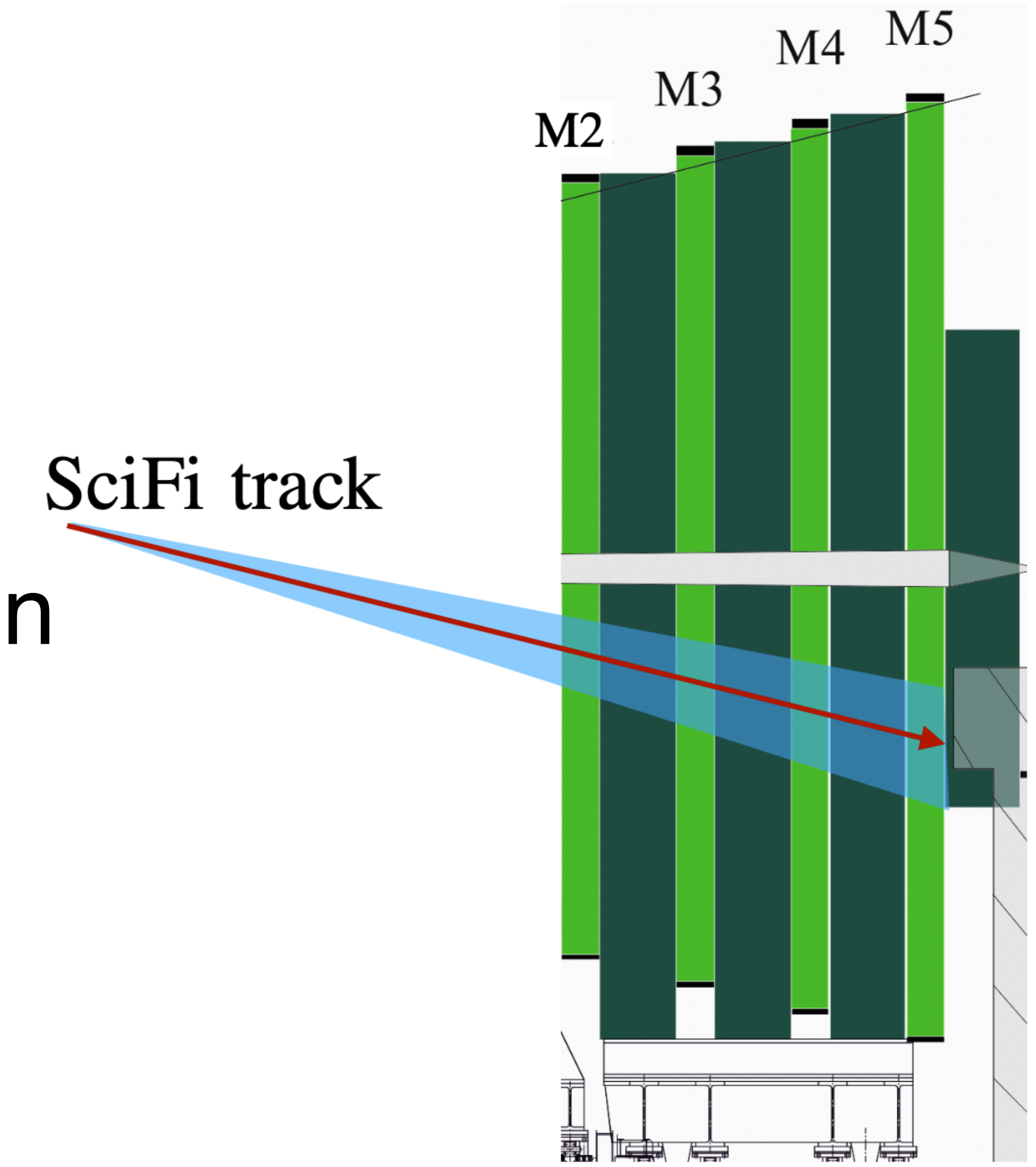


- 3 stations with 4 layers scintillating fibres each (*xuvx* configuration)
  - Extrapolate each Upstream track in the 12 layers of the SciFi
  - Build triplets combinations using T1/2/3, Best triplets selected according to local parameterisation of magnetic field
  - Forward all triplets to remaining layers with an extra parameterised corrections in the non-bending plane





Muon identification & track fit

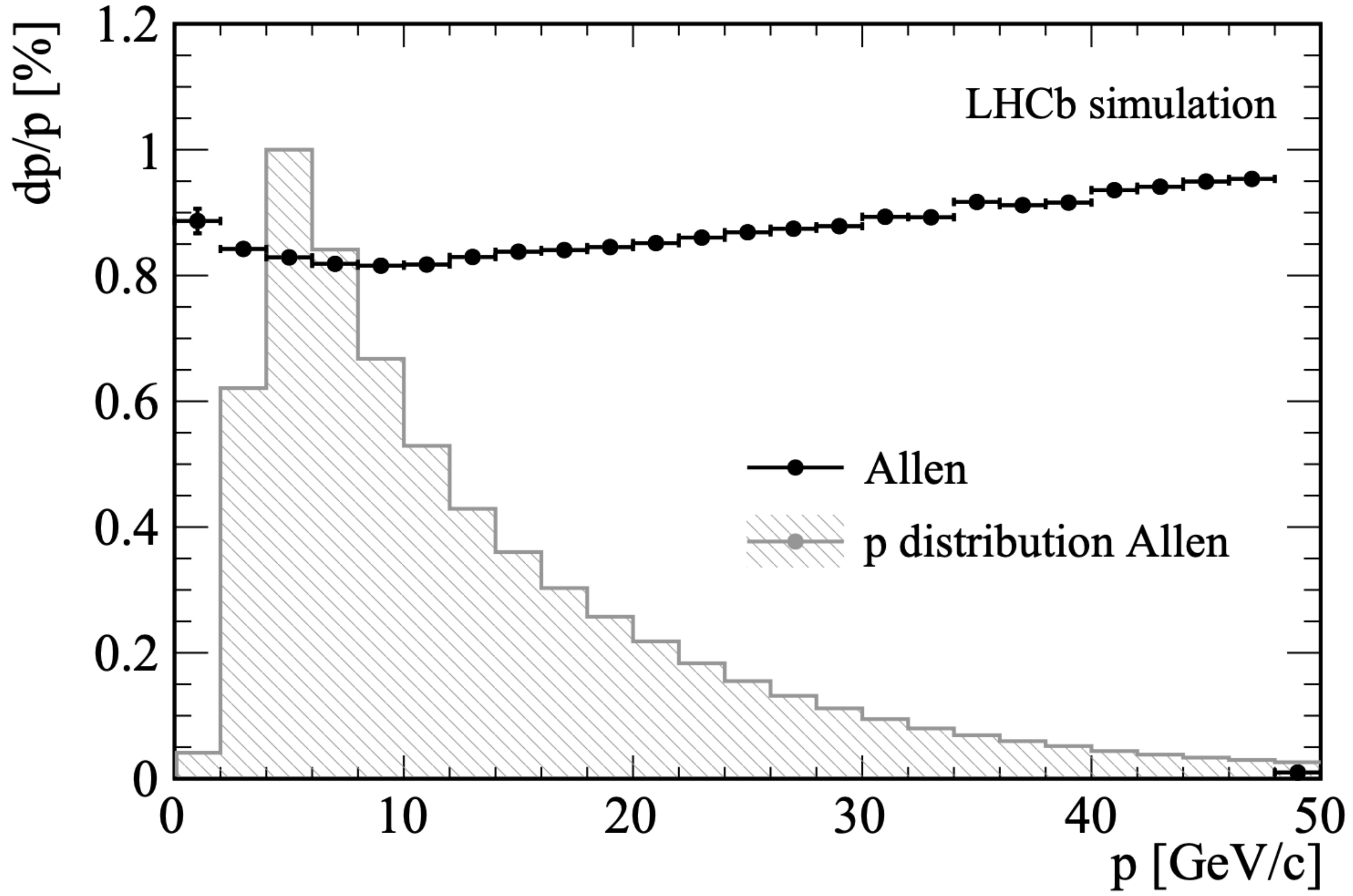


Muon identification

- Project Long tracks to 4 layers of Multi-wire proportional Muon chambers
- Find hits in side the FoI for μ identification
- Parallelise across tracks and muon chambers

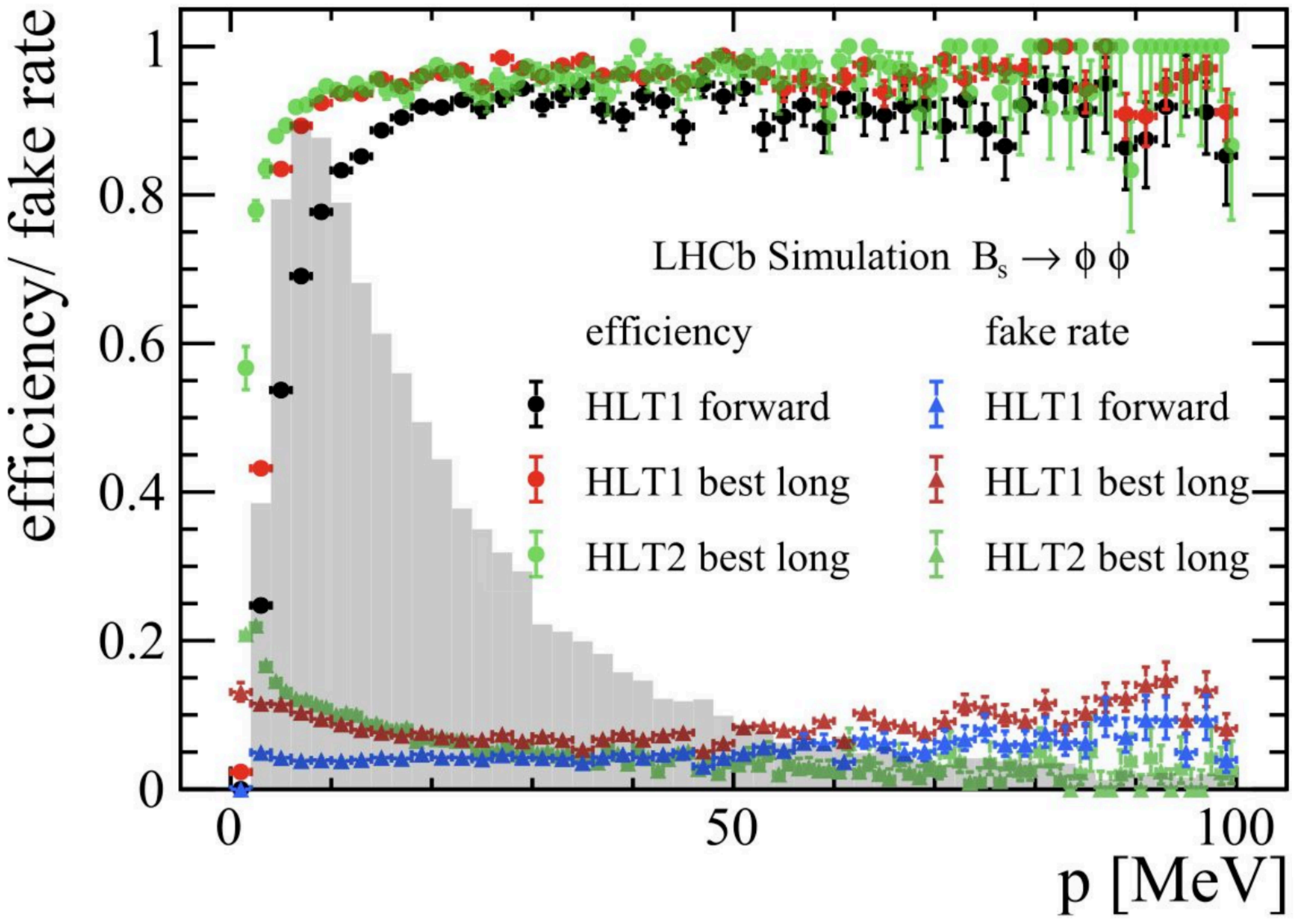
Track fit

- Goal: improve track description close to the beam line for precise determination of the impact parameter
- Only fit part of the track within the Velo detector
- Parameterized Kalman filter → no need for magnetic field map and detector material description

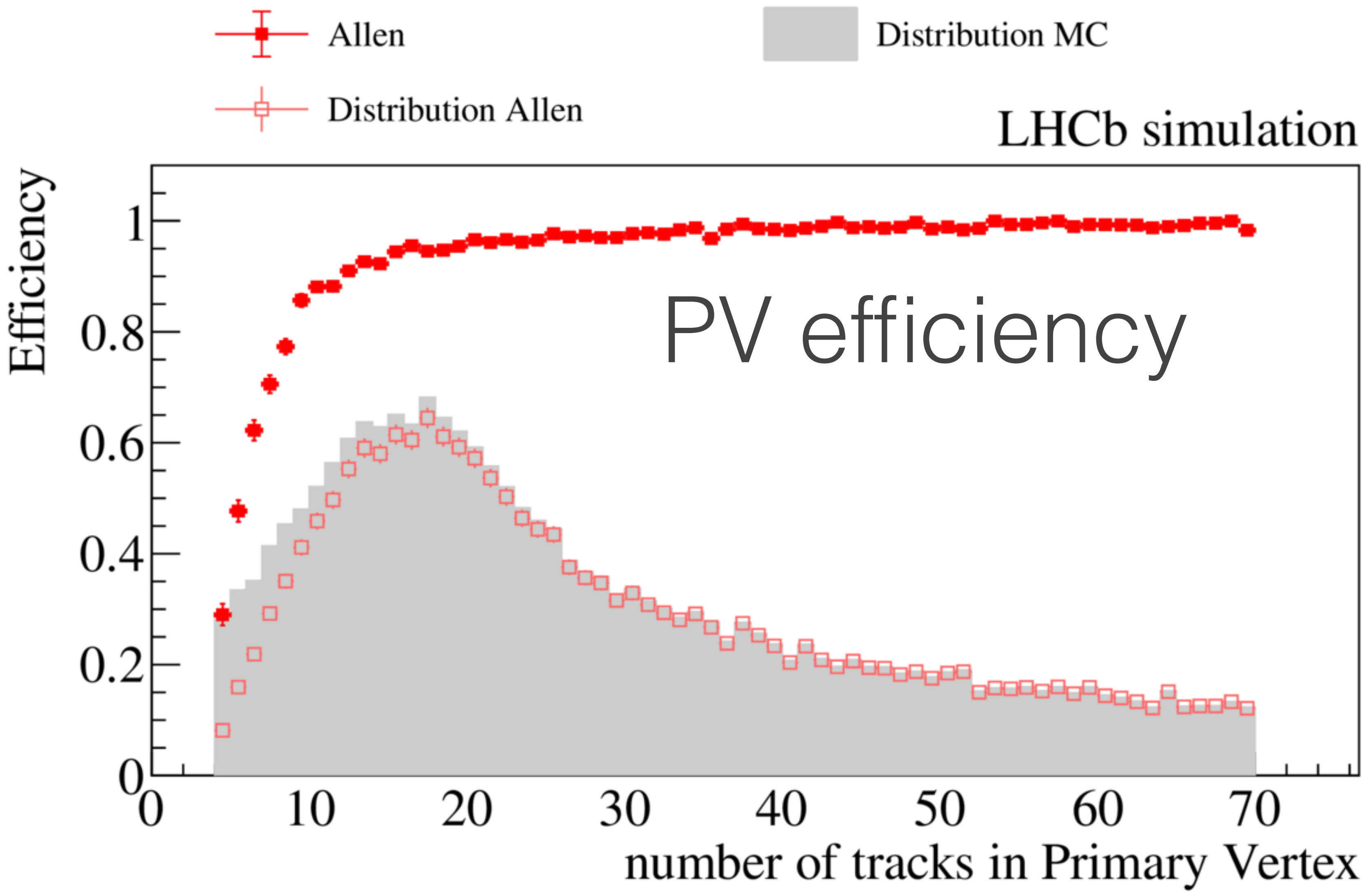


HLT1 Performance

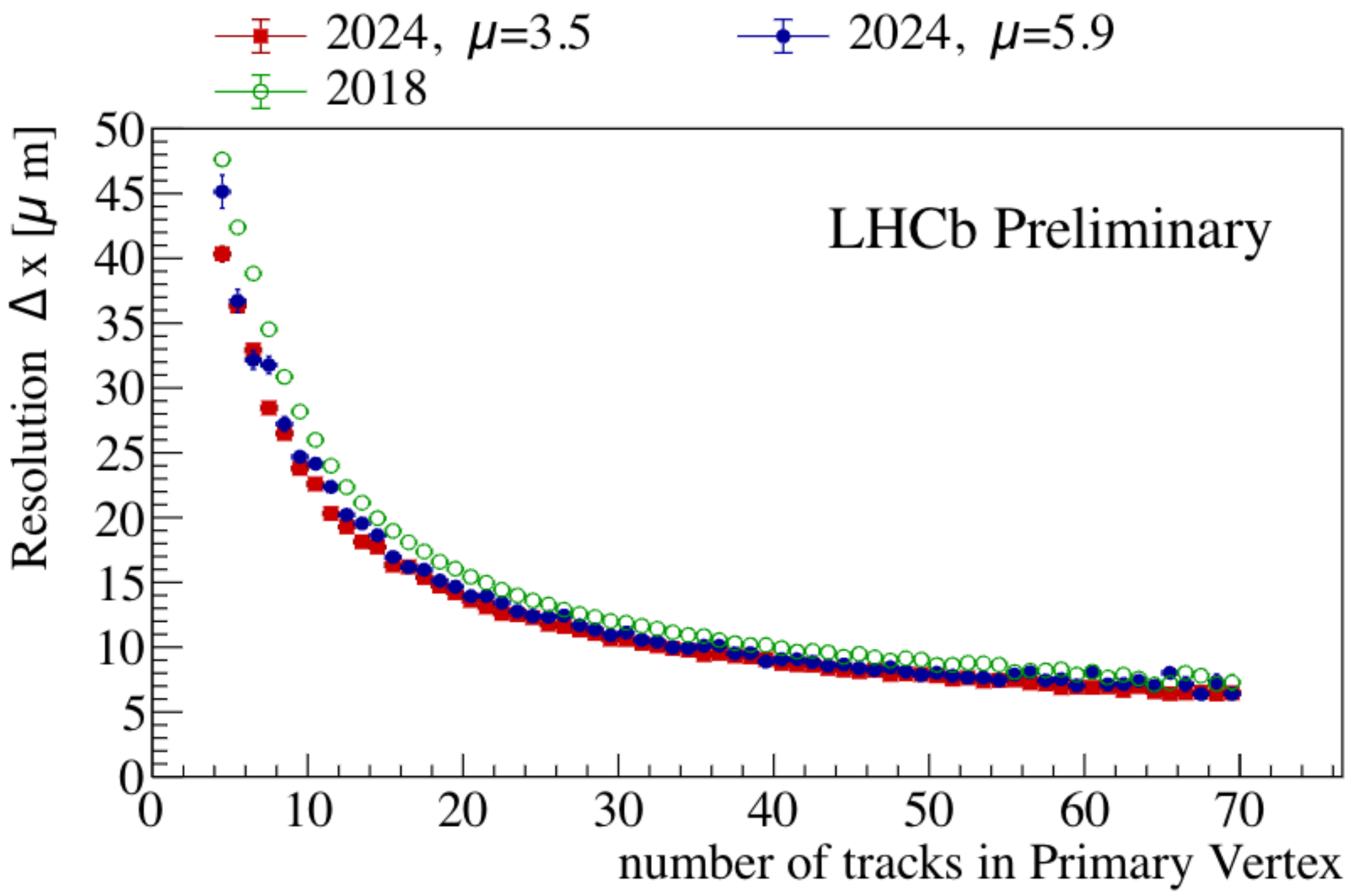
IEEE Access, vol. 12, 2024



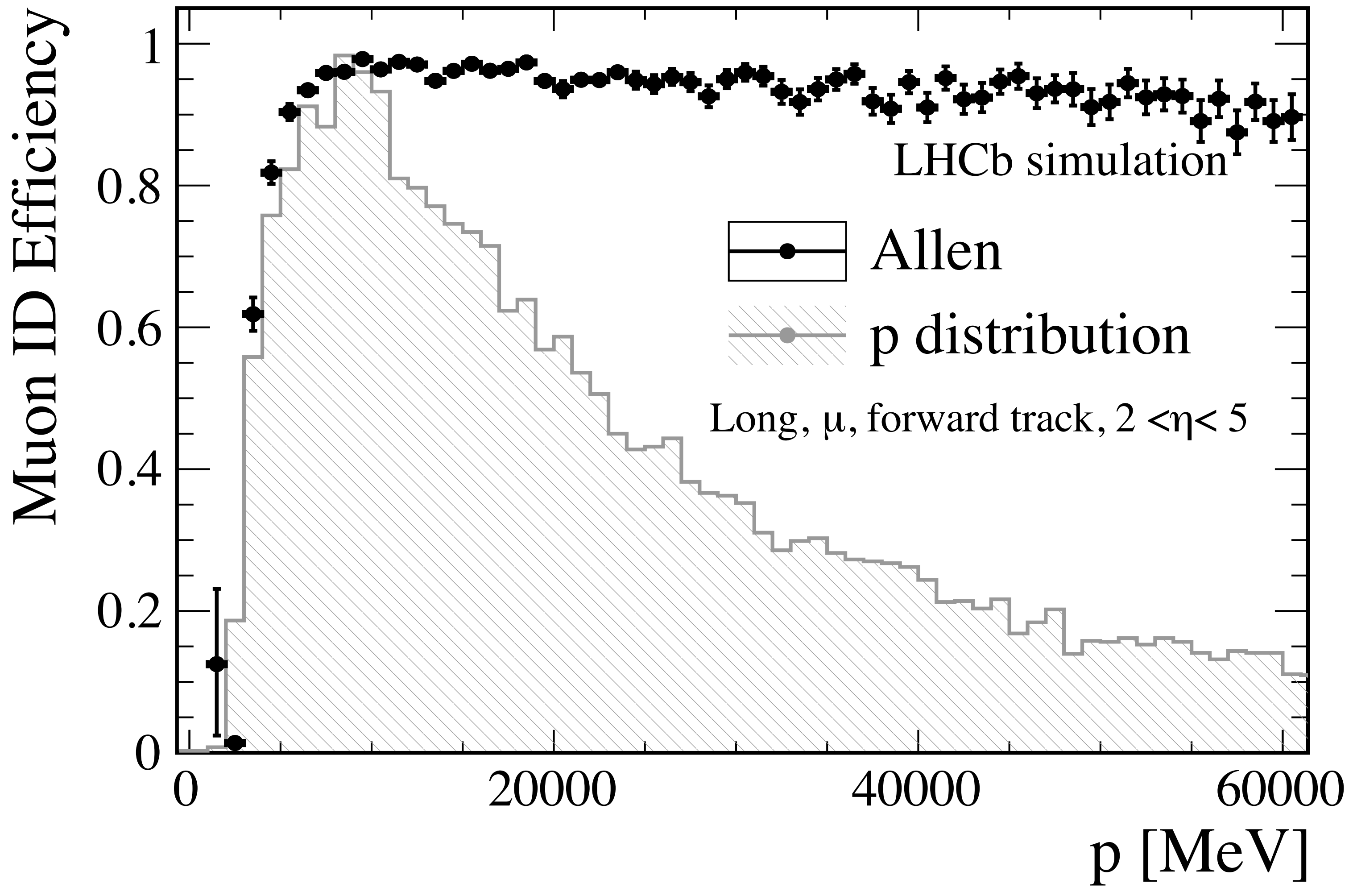
LHCb-Figure-2020-014



LHCb-FIGURE-2024-011



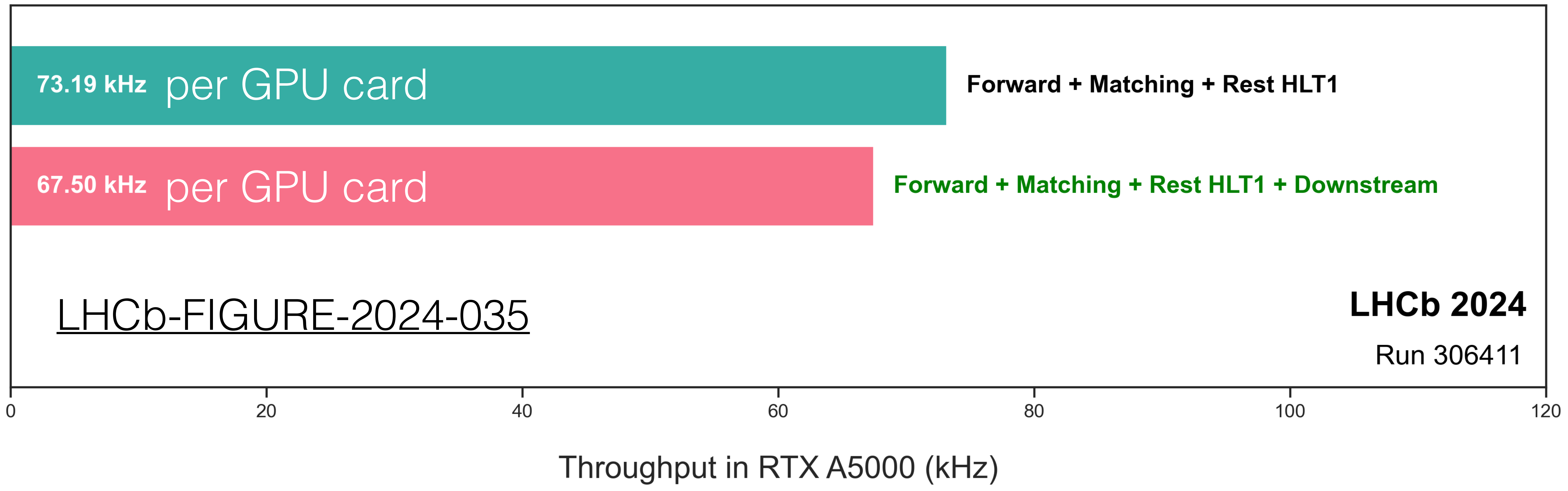
LHCb-Figure-2020-014



- About 95% long track efficiency about p>5 GeV
- More than 90% PV reconstruction efficiency with number of tracks larger than 10
- More than 95% Muon identification
- About 2-3% $\pi \to \mu$ misidentification when momentum > 20 GeV

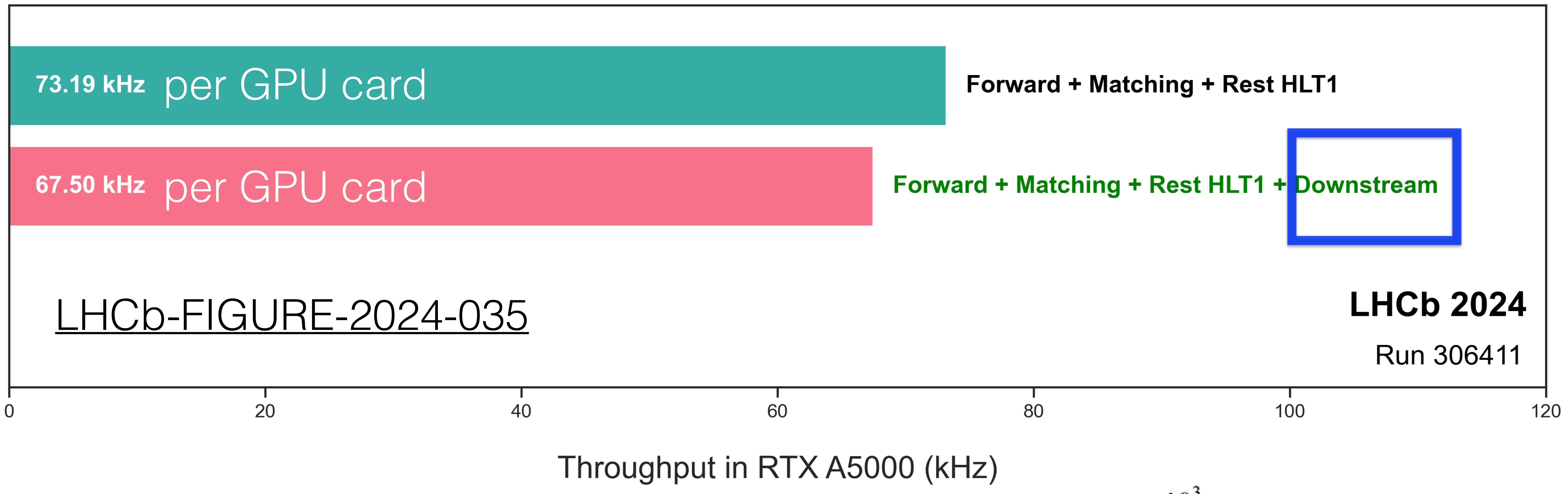
HLT1 Throughput Performance

- O(500) Nvidia RTX A5000 GPUs implemented
- O(300) enough for designed HLT1 trigger on 30MHz ⇒ plenty of space for more physics

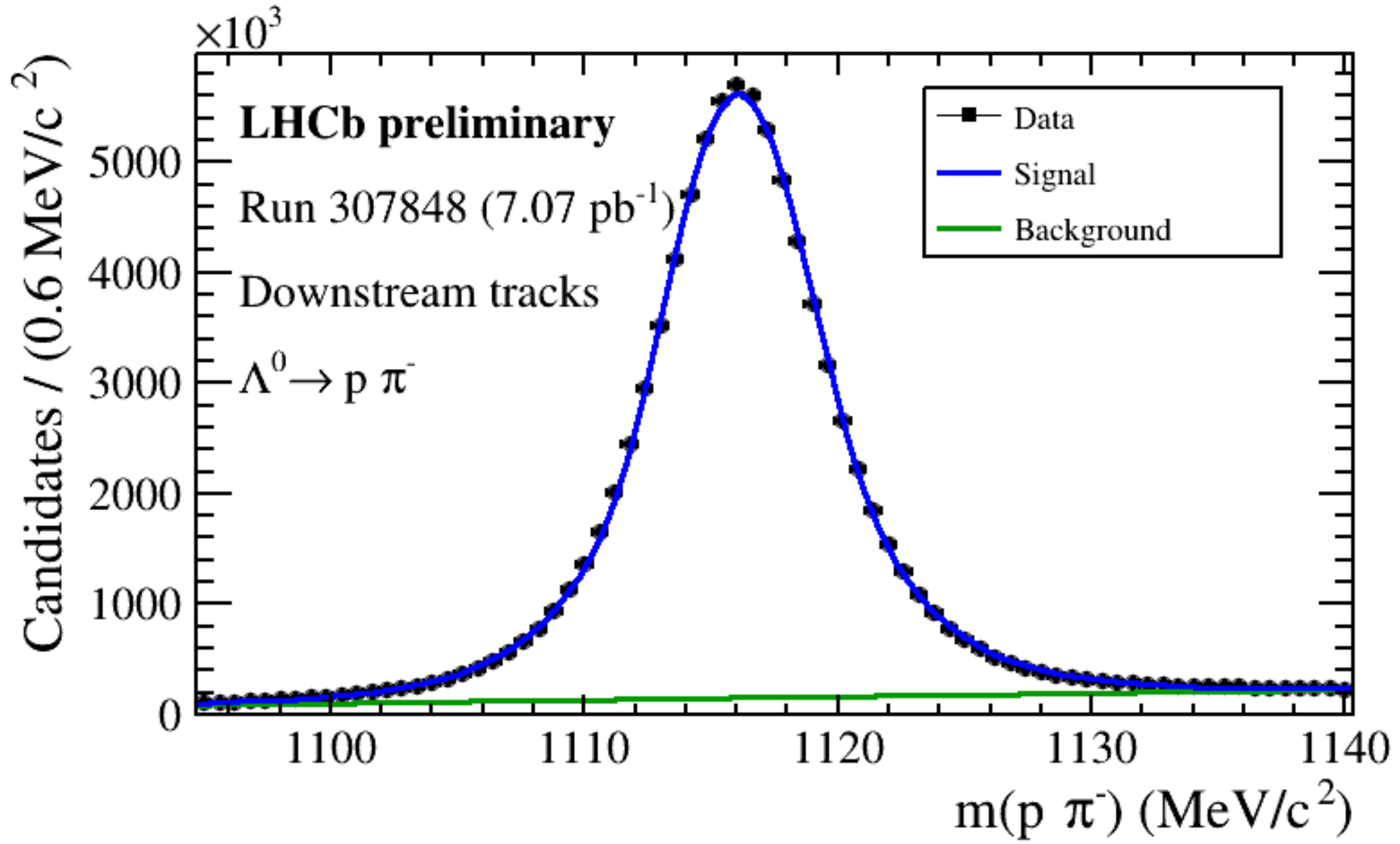


HLT1 Throughput Performance

- O(500) Nvidia RTX A5000 GPUs implemented
- O(300) enough for designed HLT1 trigger on 30MHz ⇒ plenty of space for more physics



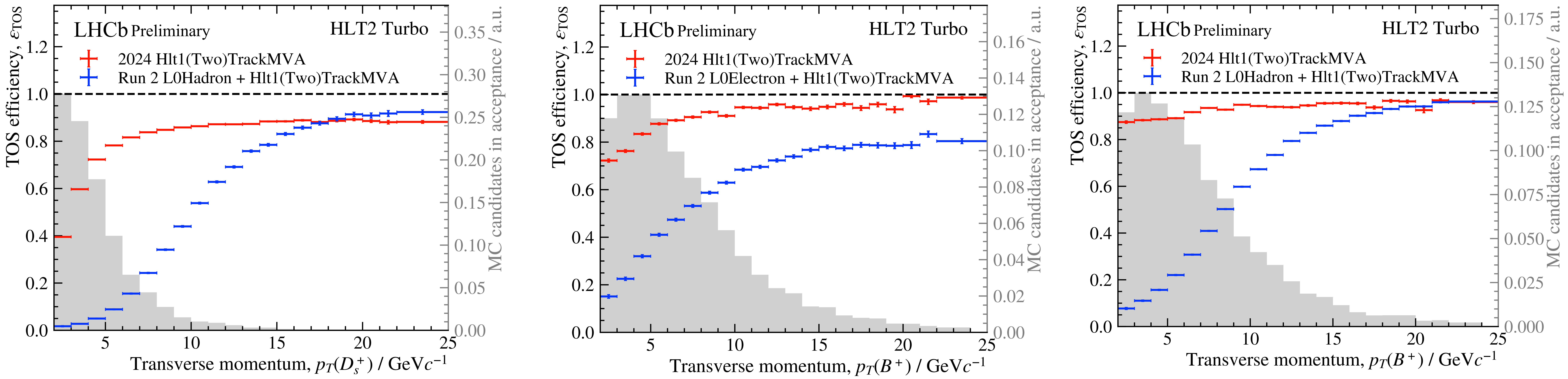
- Downstream tracks reconstructed on GPU, extending the potentials for the physics for long-lived particles
- ECAL & RICH reconstruction on GPU in progress



HLT1 Trigger Performance with 2024 data

- The real-time analysis philosophy proved to be valid
- Removal of hardware trigger results in significant improvement in the trigger efficiency for dielectrons, hadronic c and b decay channels
  - Huge gain at low p_T region, beneficial for the charm and strange physics programme

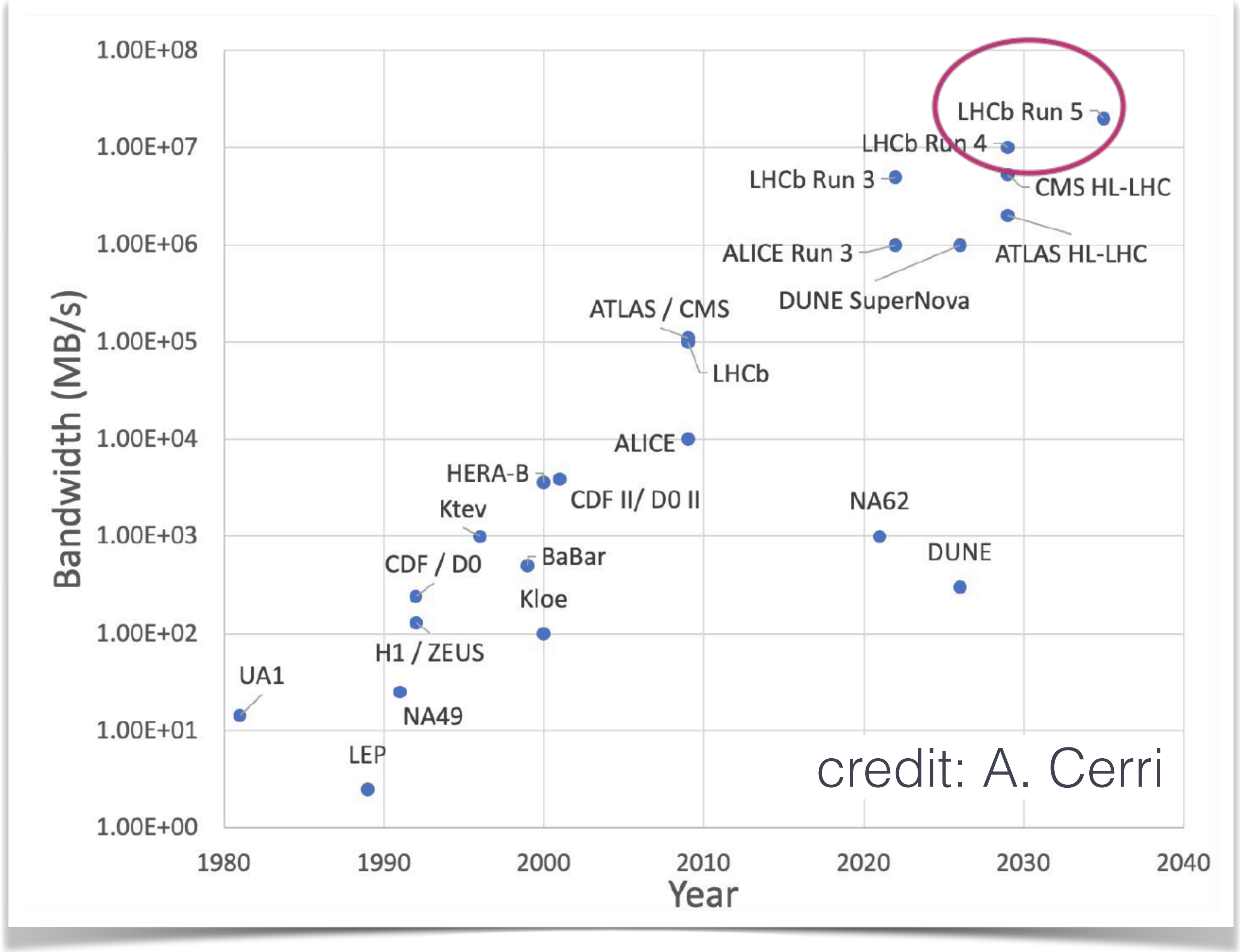
LHCb-Figure-2024-030



$D_s^+ \to K^+K^-\pi^+$ $\quad$ $B^+ \to J/\psi(e^+e^-)K^+$ $\quad$ $B^+ \to D^0(K^-pi^+)\pi^+$

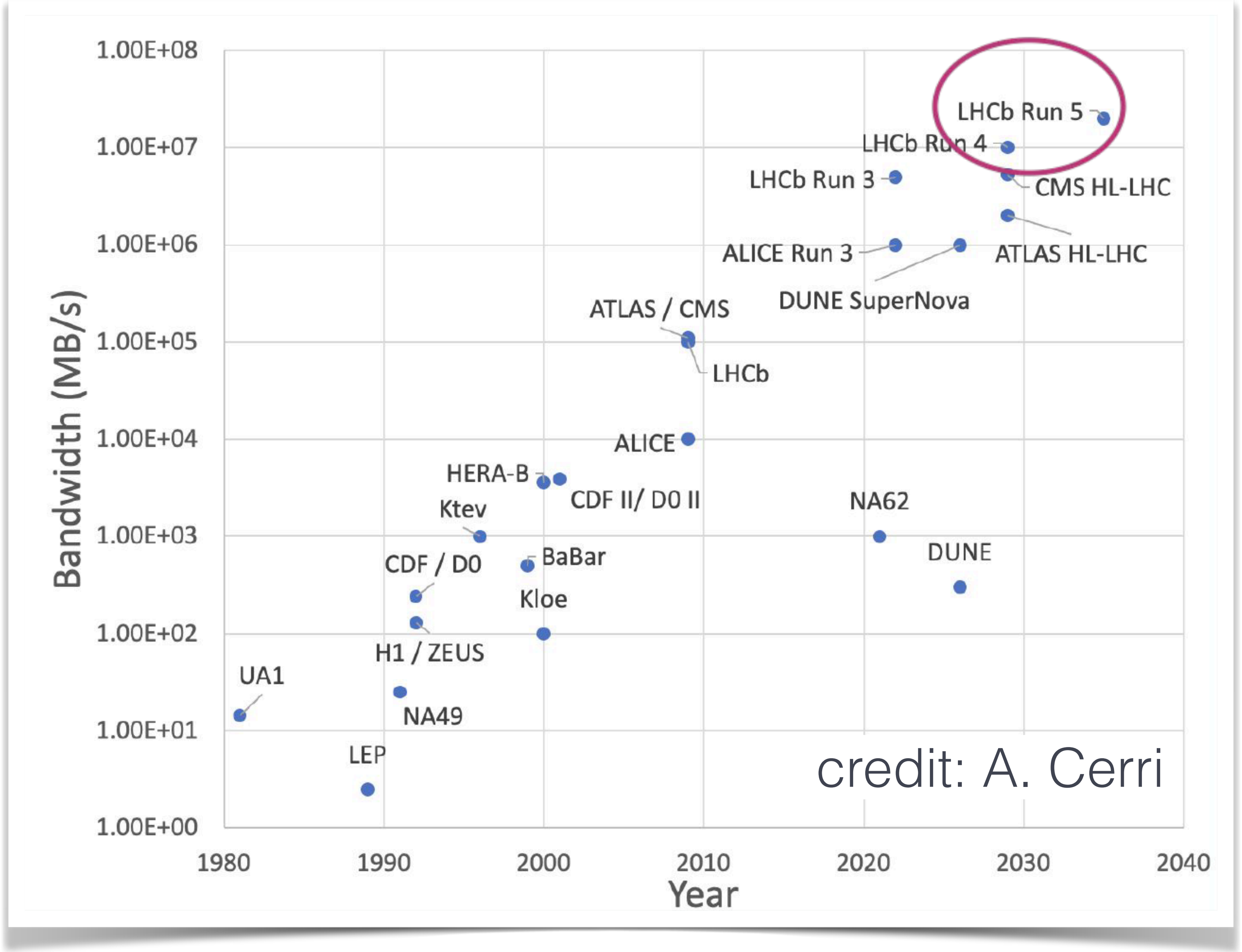
Toward the future upgrade

- LHCb planning Upgrade II for Long Shutdown 4
  - FTDR approved in 2022 and Scoping document in review
  - Luminosity: $2 \times 10^{33} \rightarrow 1.5 \times 10^{34}\ cm^{-1} \cdot s^{-1}$
  - Pile up: 5 → 40
  - Exciting challenges in trigger and DAQ:
    ⇒ 200 Tb/s of data to be processed in real time, 4D reconstruction with time…



Toward the future upgrade

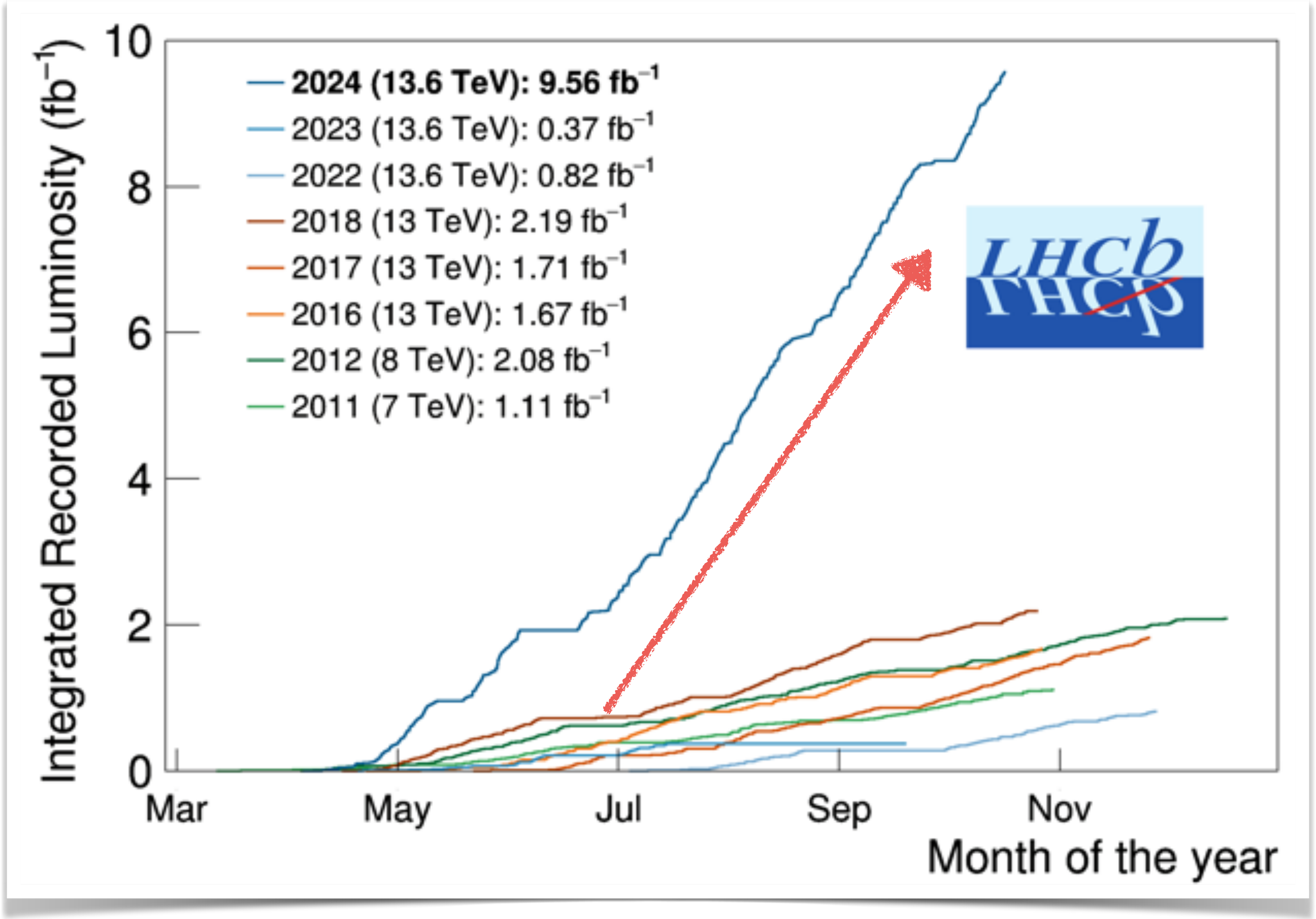
- LHCb planning Upgrade II for Long Shutdown 4
  - FTDR approved in 2022 and Scoping document in review
  - Luminosity: $2 \times 10^{33} \rightarrow 1.5 \times 10^{34}\ cm^{-1} \cdot s^{-1}$
  - Pile up: 5 → 40
  - Exciting challenges in trigger and DAQ:
    ⇒ 200 Tb/s of data to be processed in real time, 4D reconstruction with time…



- Fully software trigger strategy, partial and full detector reconstruction both on GPUs
- Complementary R&D activities focusing on primitives reconstruction on FPGAs, IPU exploration

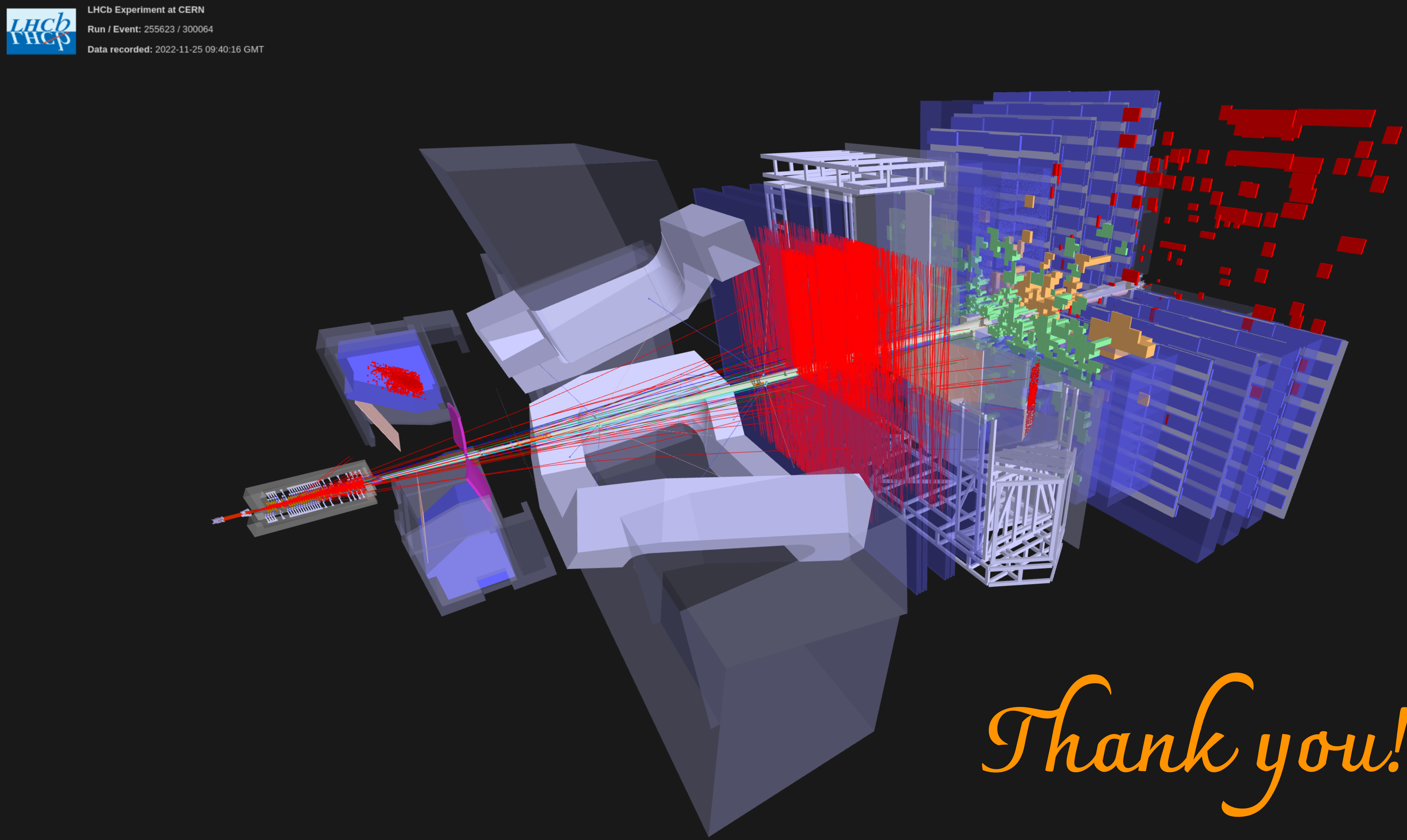
Summary

- LHCb Run 3 changes the trigger paradigm with software only data processing successfully
  - ✓ GPU-based HLT1 reduce data rate from 30MHz → 1 MHz
  - ✓ Great performance achieved in 2024 data



- Hybrid architecture (GPU+ FPGA +CPU) in Run 3 paves the way for the future upgrade
- R&D studies on optimal use of hybrid architectures (GPU/CPU/FPGA) for LHCb Upgrade II


LHCb Experiment at CERN
Run / Event: 255623 / 300064
Data recorded: 2022-11-25 09:40:16 GMT

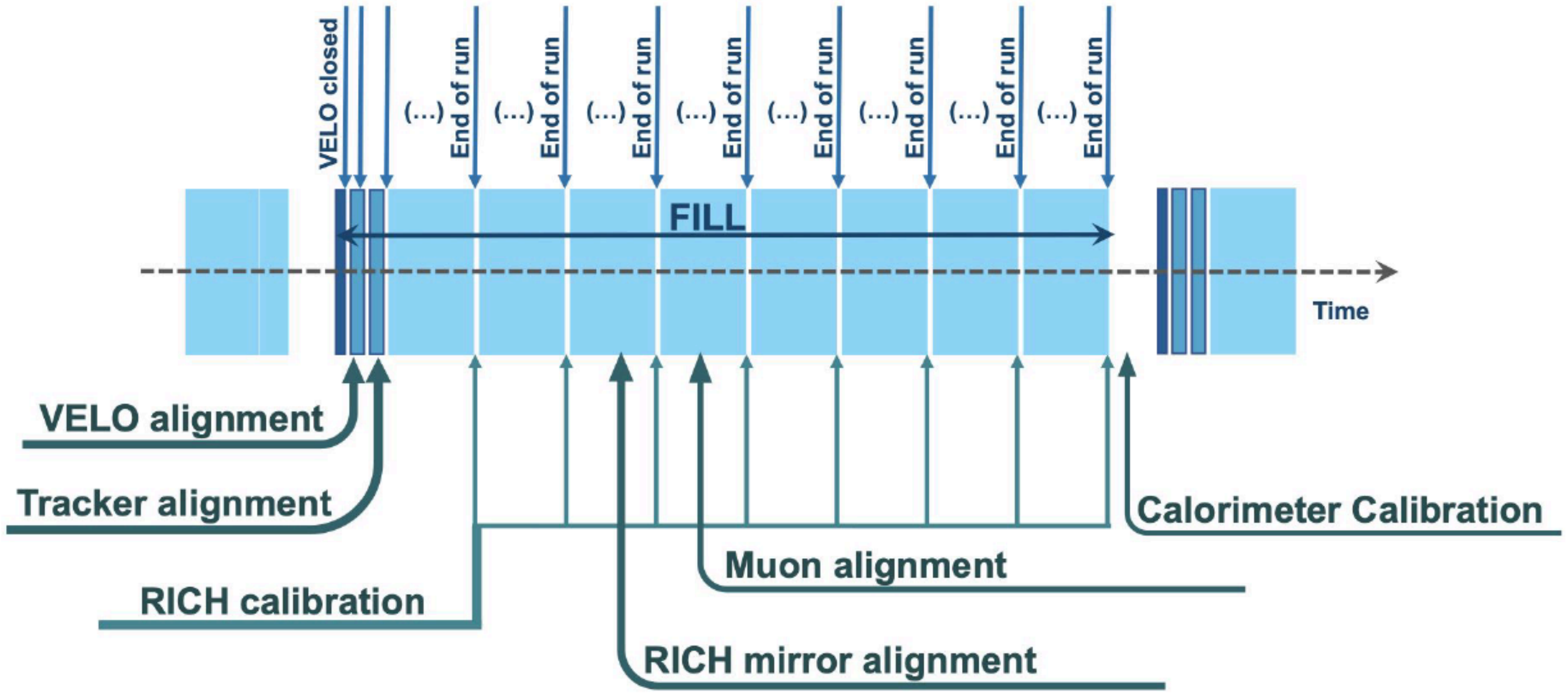

LHCb Experiment at CERN
Run / Event: 255623 / 300064
Data recorded: 2022-11-25 09:40:16 GMT

Thank you!

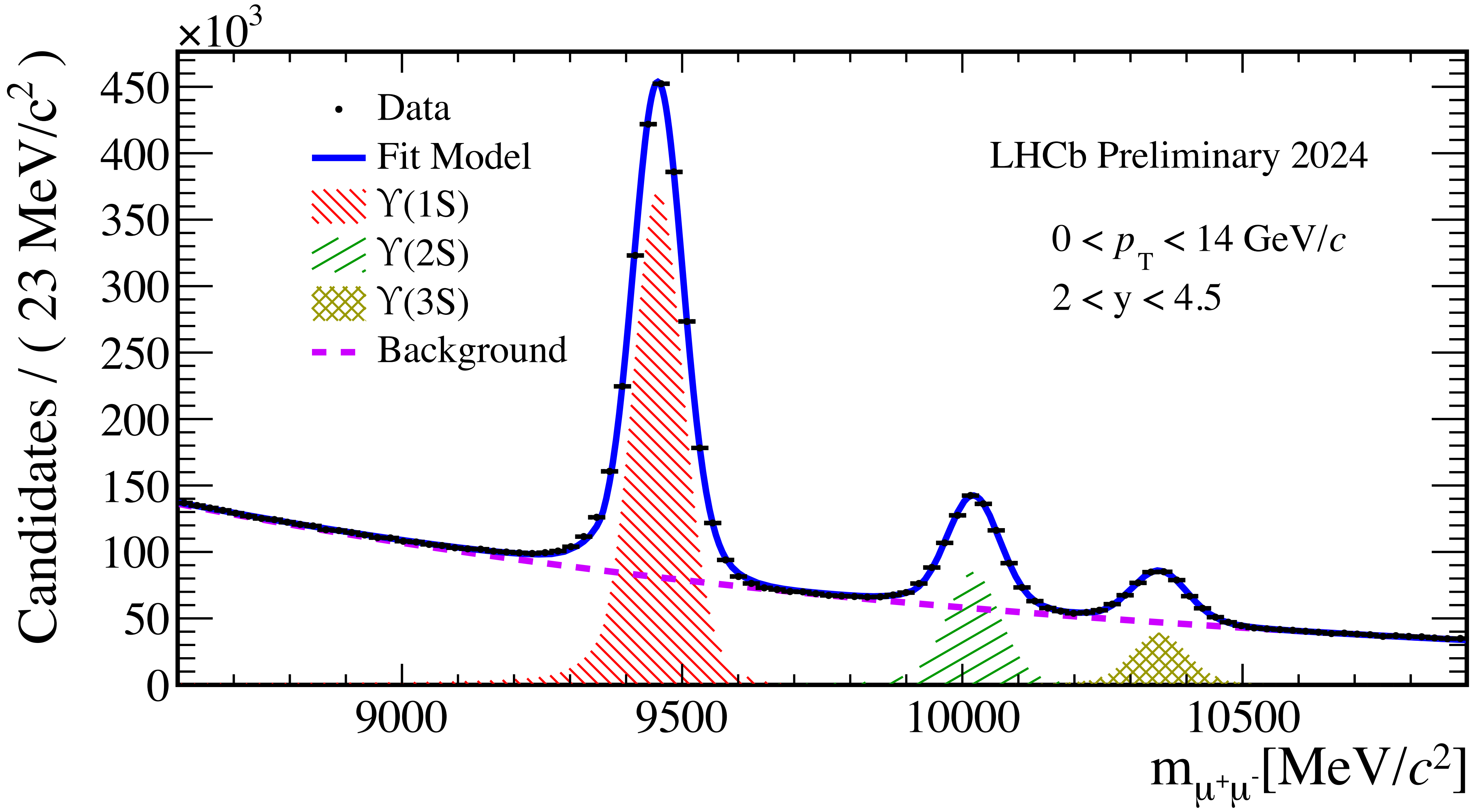
Alignment & Calibration

- Data passing HLT1 trigger stored in a buffer of O(30PB) for real-time alignment and calibration
- Crucial for efficient and pure selections require offline-quality reconstruction at the HLT2 level

Several minutes in Trackers & several hours for RICH & MUON



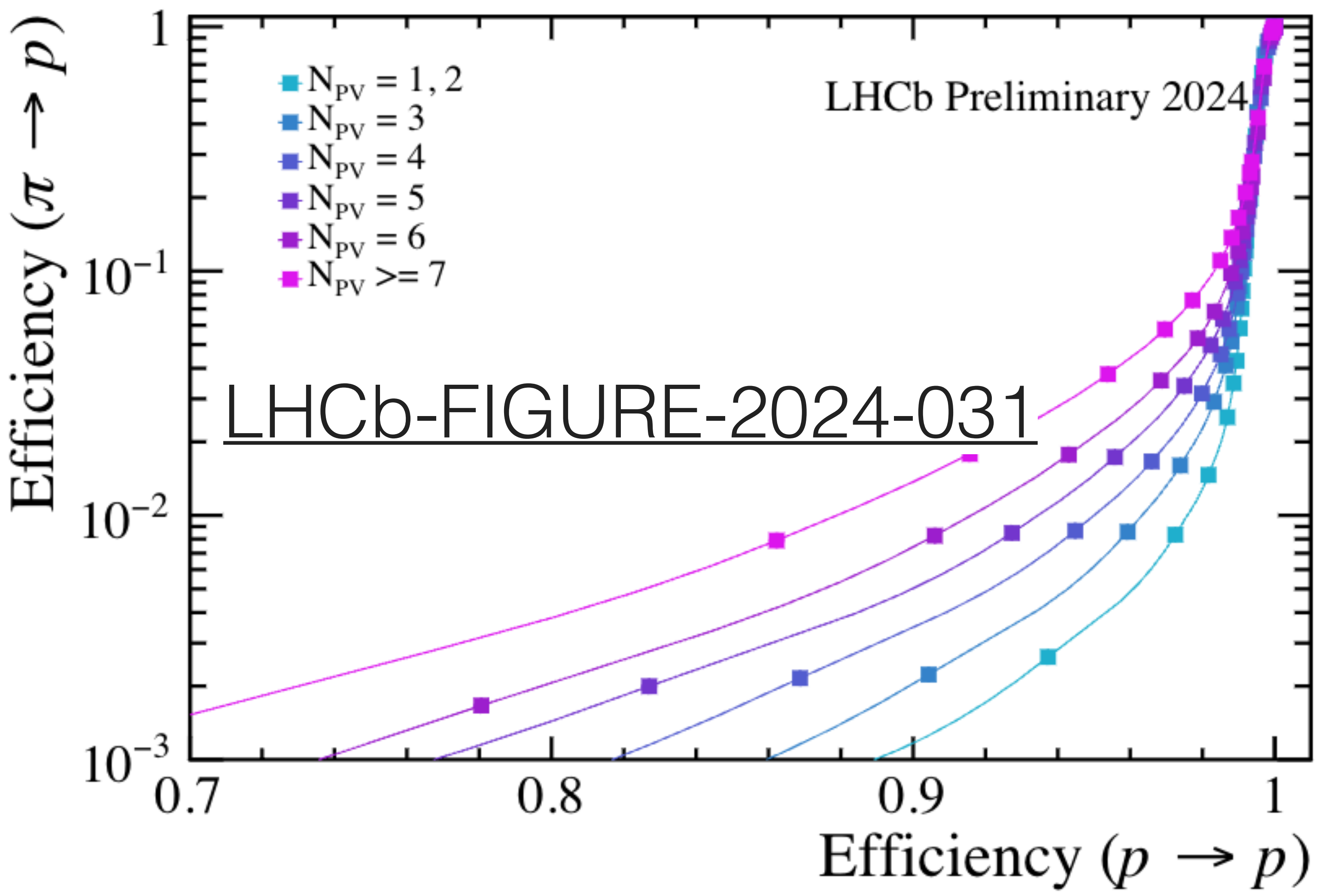
LHCb-FIGURE-2024-025

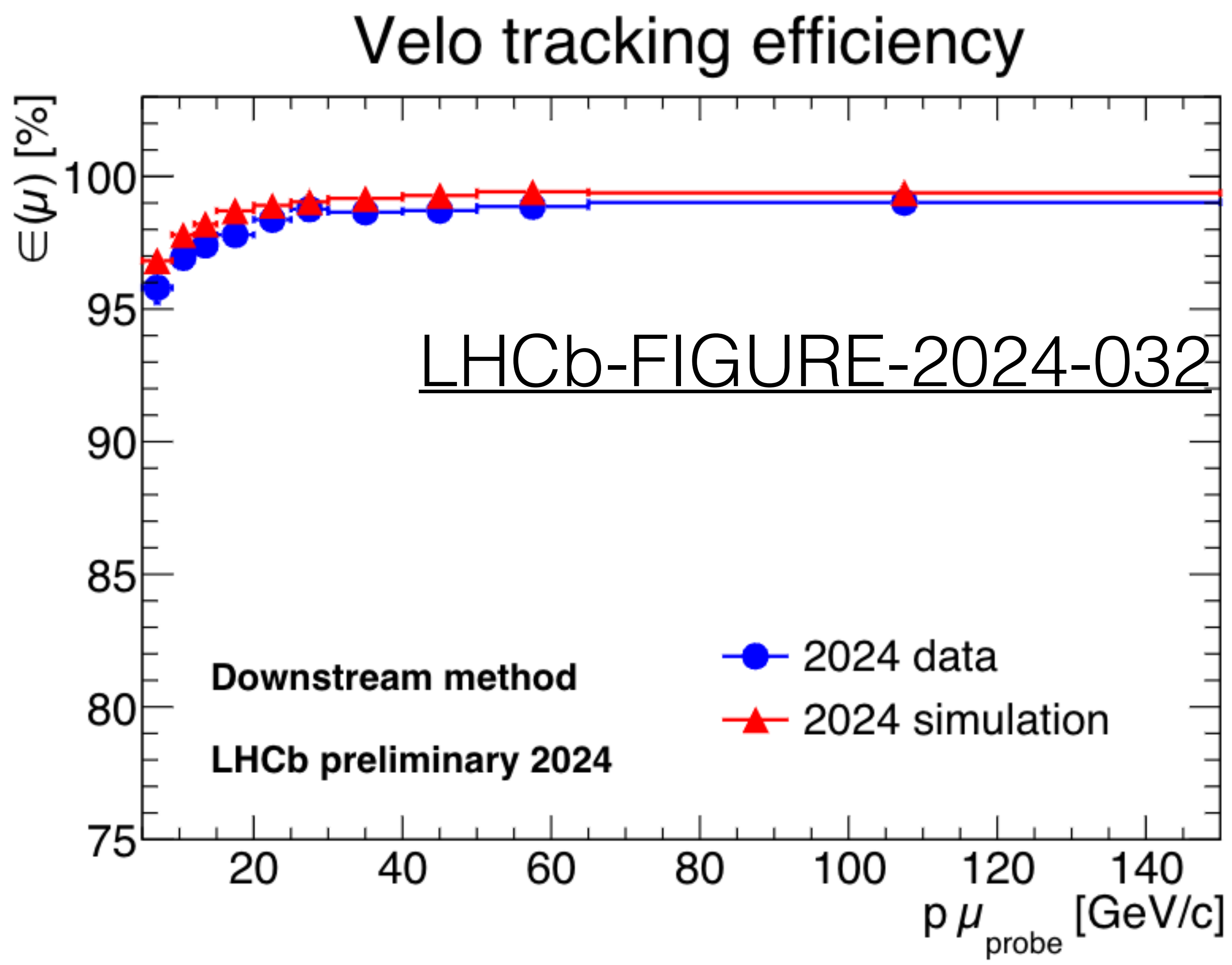


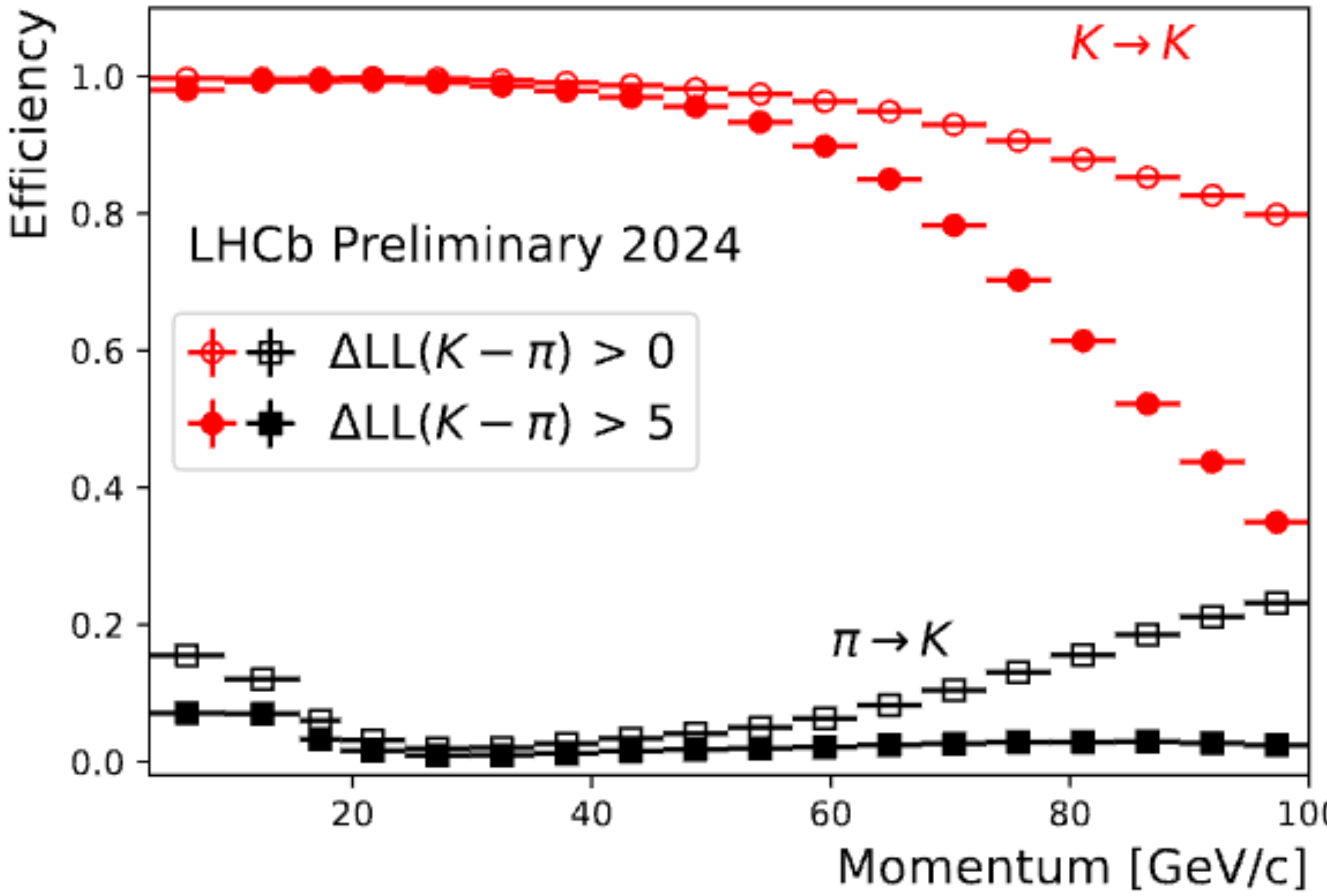
HLT2 Performance with 2024 data

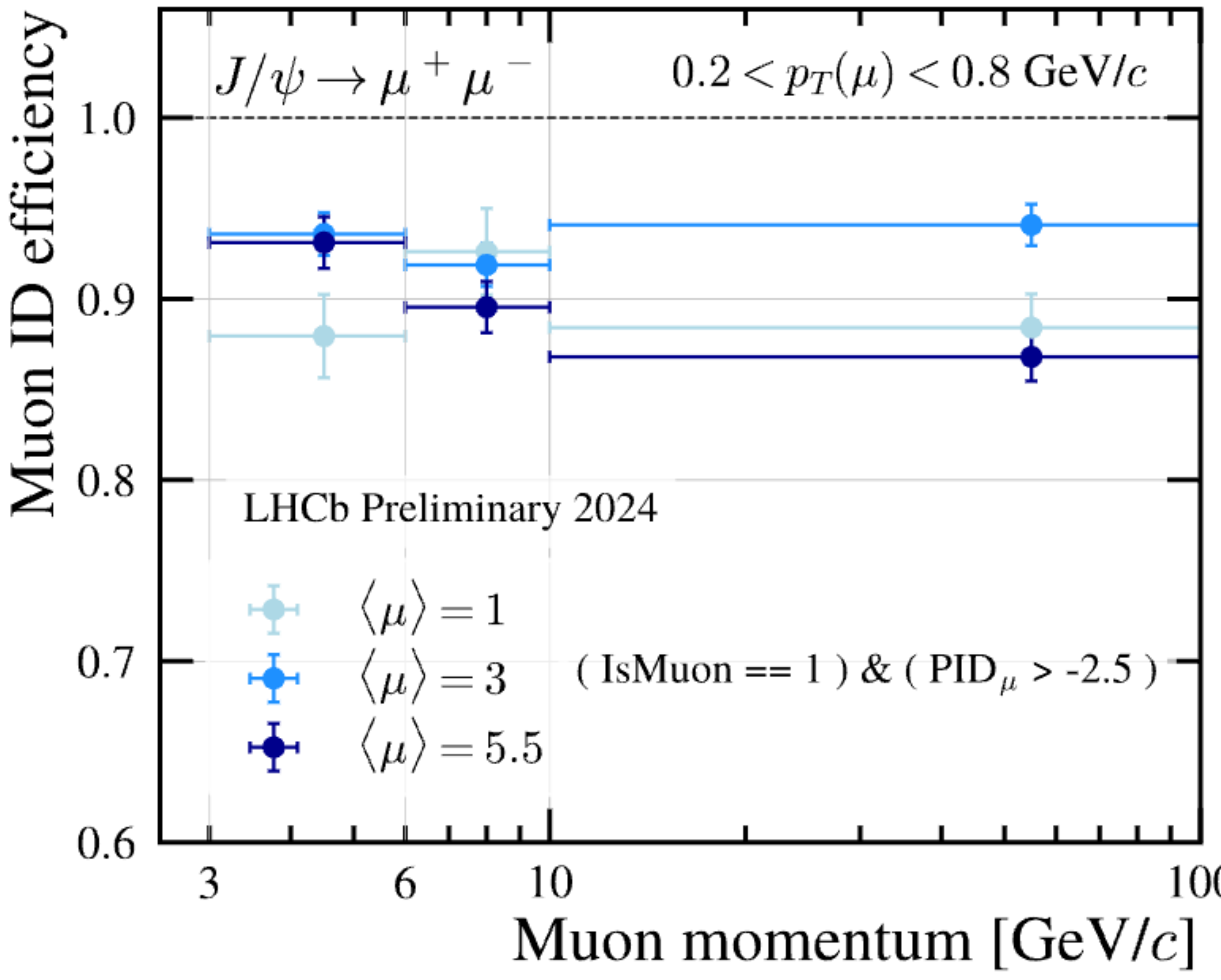
LHCb-FIGURE-2024-010

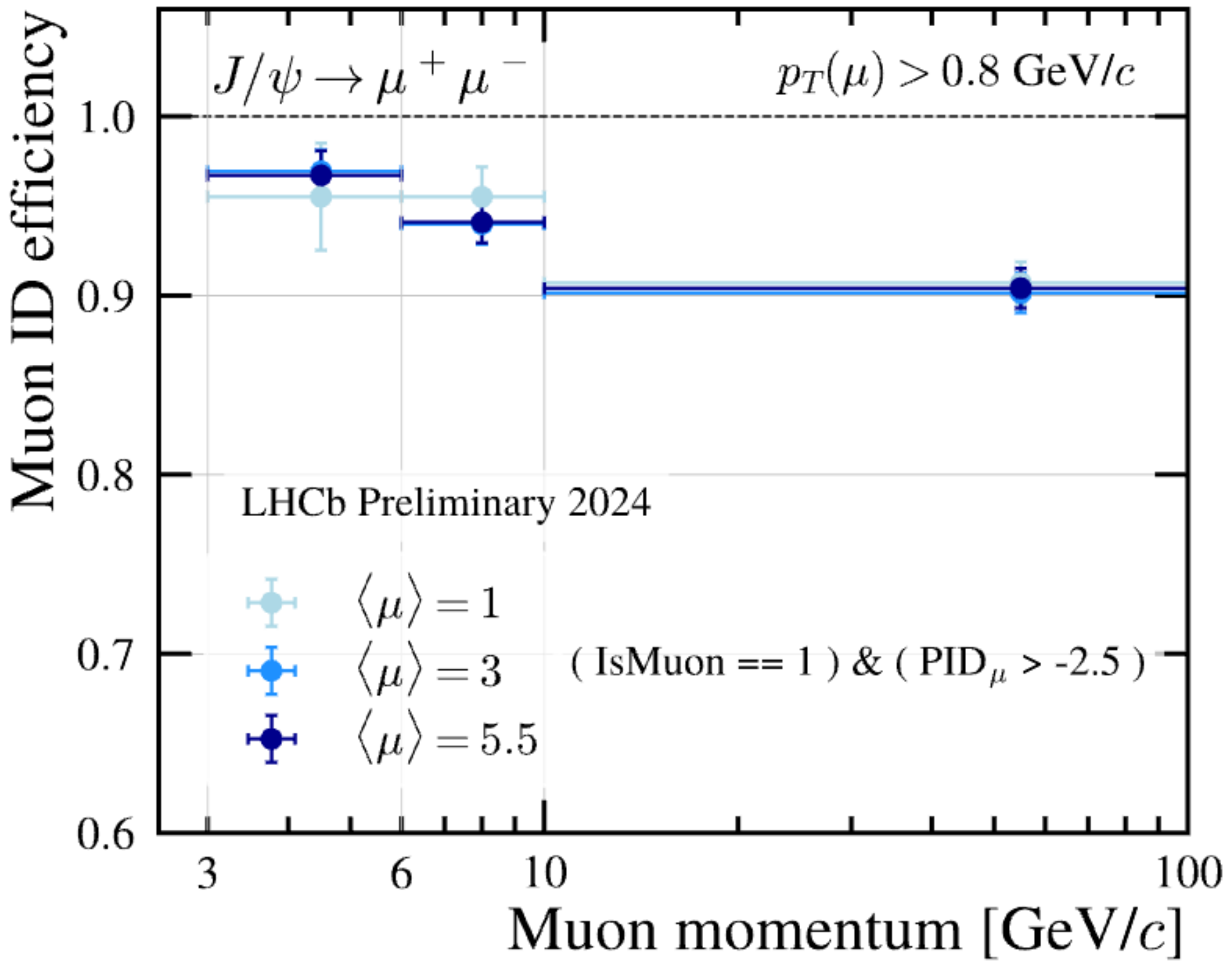
- Achieving TDR performance for tracking and Particle identification





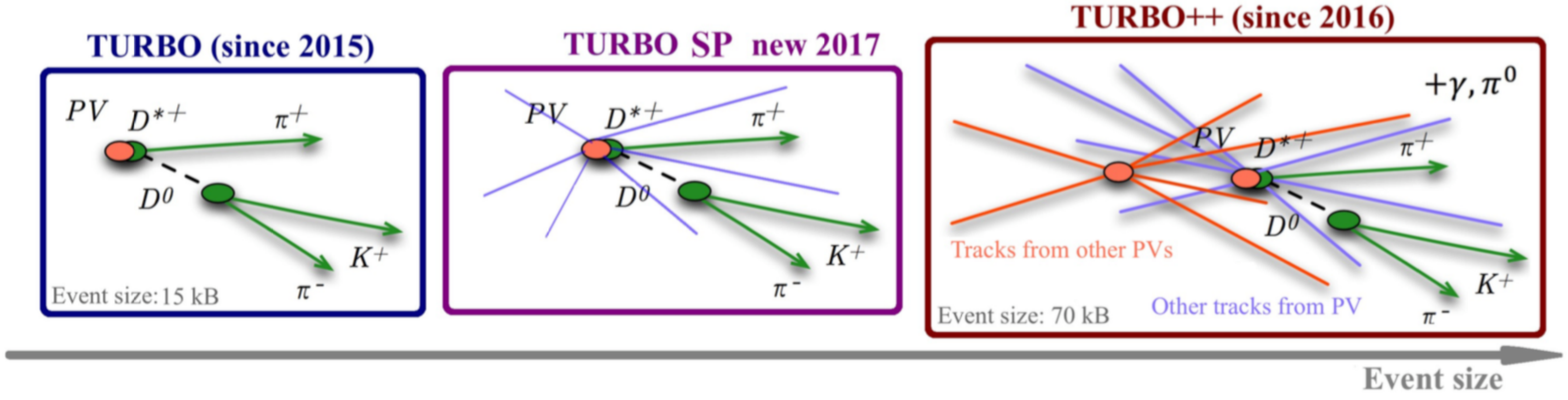


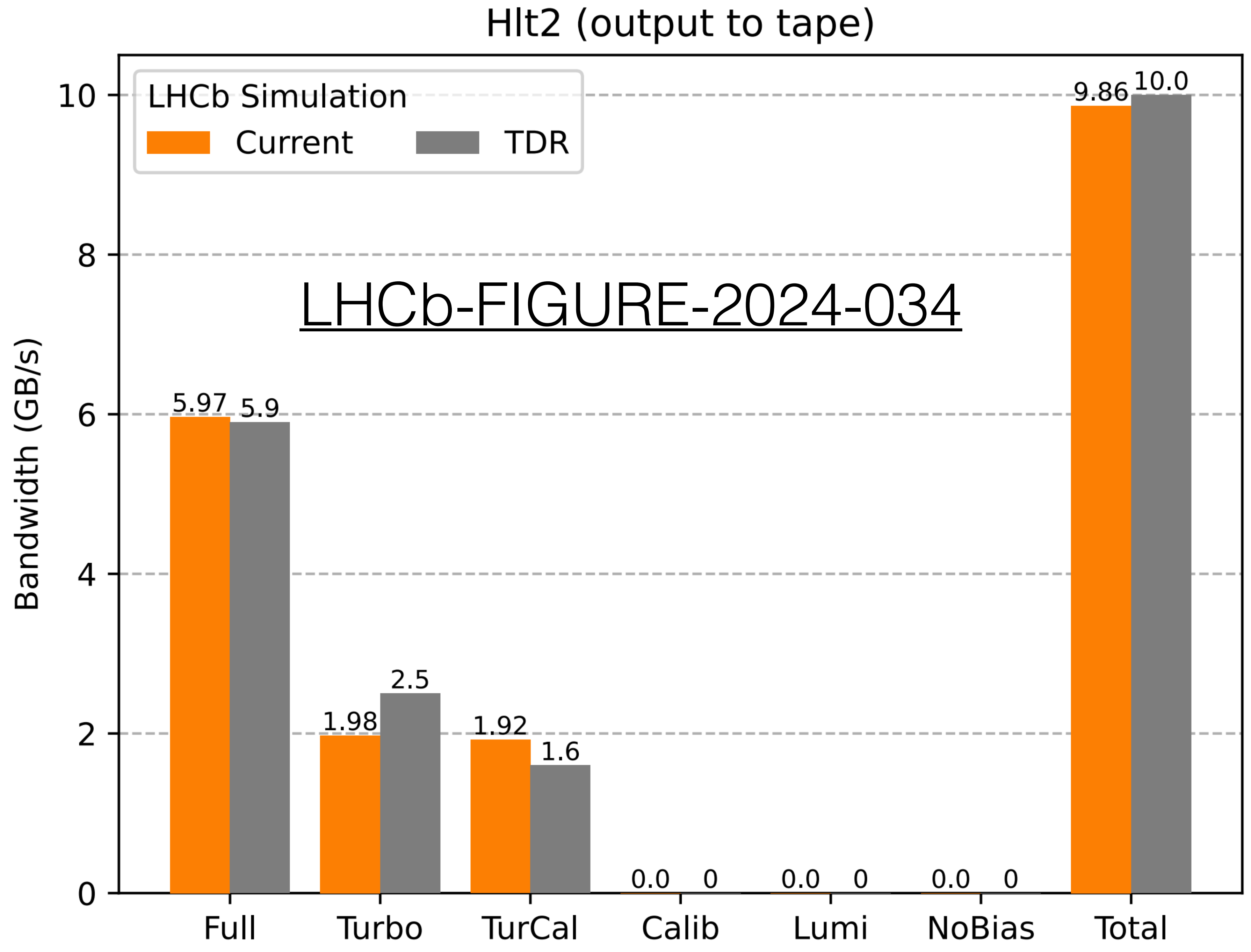




HLT2 Trigger Performance

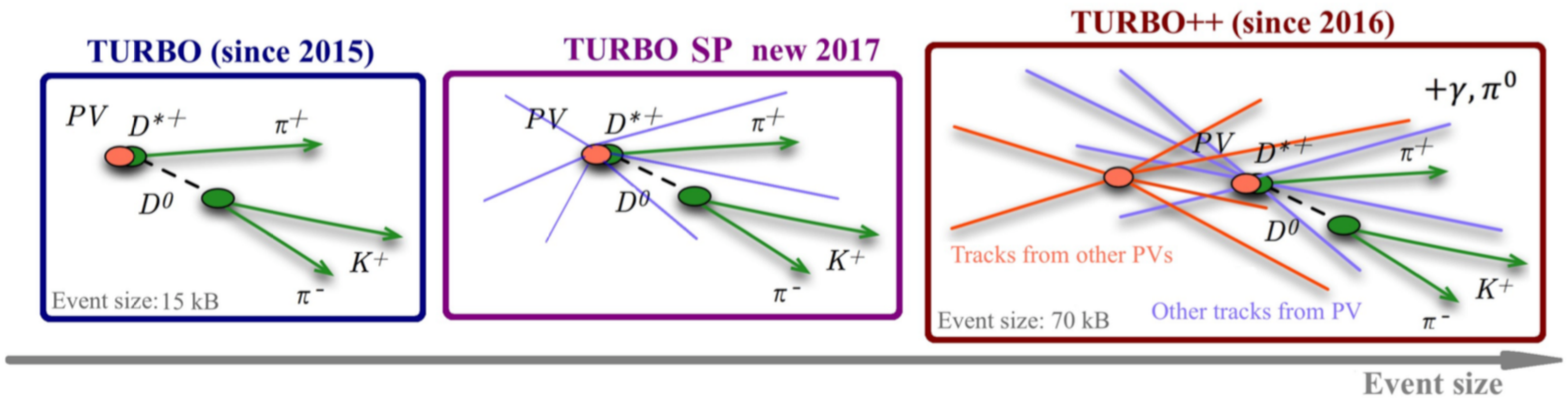
- Fixed output bandwidth of 10 GB/s
- Bandwidth [MB/s] ~ Trigger output rate [kHz] × average event size [kB]





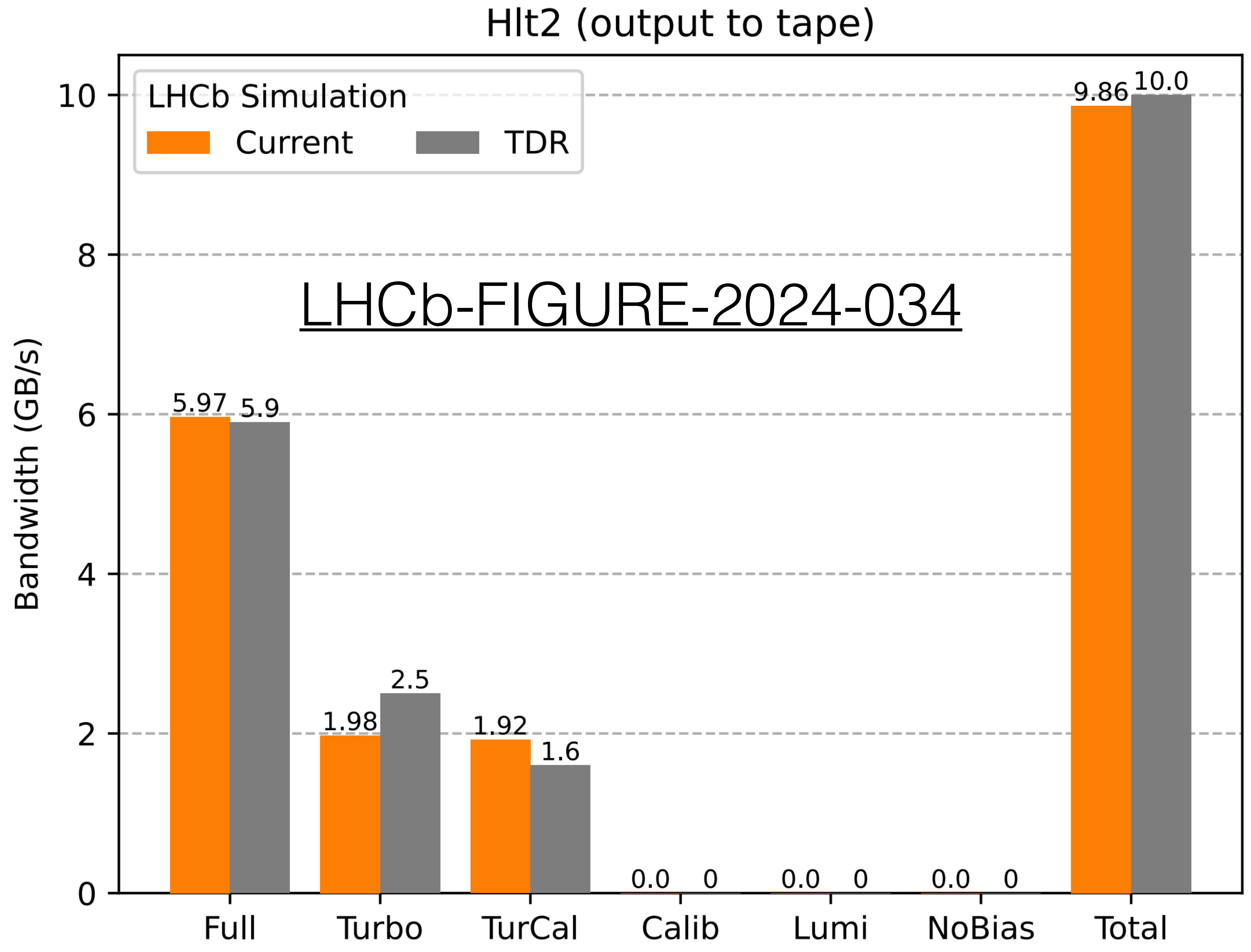
HLT2 Trigger Performance

- Fixed output bandwidth of 10 GB/s
- Bandwidth [MB/s] ~ Trigger output rate [kHz] × average event size [kB]



- Reduced event format: **throw away the raw event info & reduce event size** by saving only objects needed for physics analyses

$\Rightarrow$ *allowing more physics for the same bandwidth*



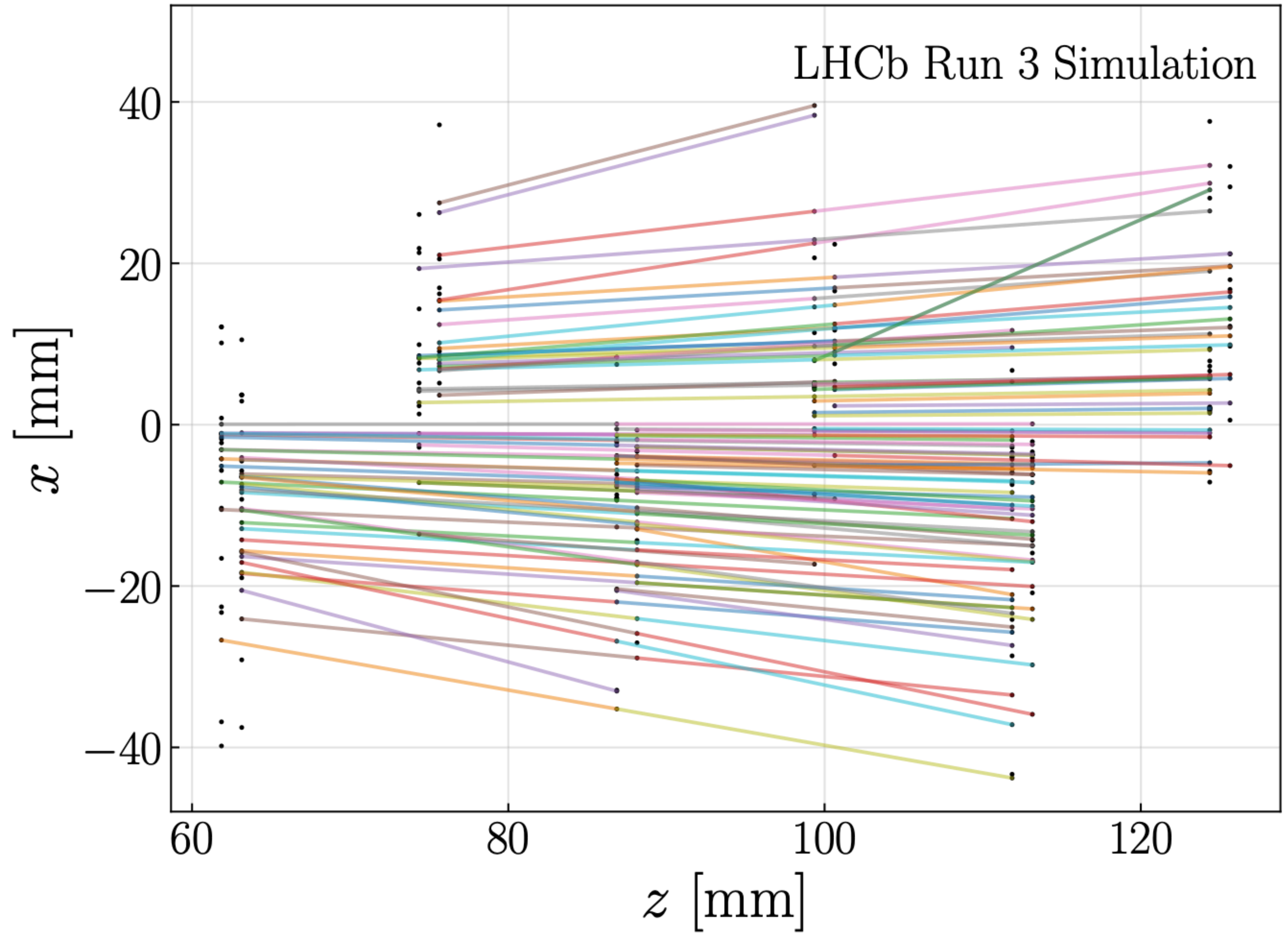
LHCb Performance Definitions

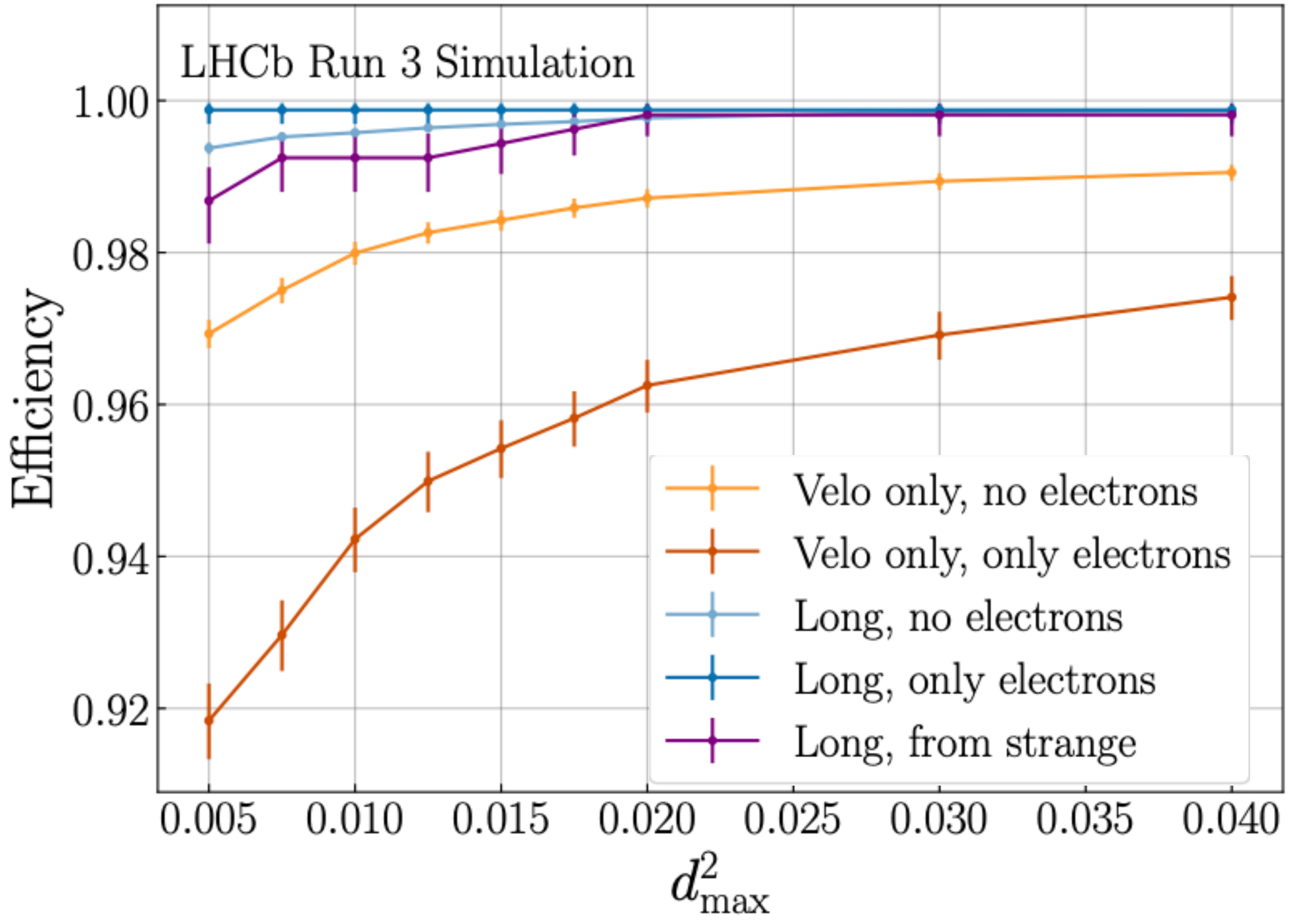
LHCb-Figure-2020-014

- A **track** is **matched** to a simulated particle if **at least 70% of the hits** come from the same simulated particle
- **Efficiency:** number of matched reconstructed tracks divided by number of reconstructible particles
- **Reconstructible particles** have a minimum number of hits in the sub-detectors for which the efficiency is being determined
- A **PV** is matched to a simulated PV if the **distance along the z-axis is less than five times the uncertainty** of the reconstructed PV
- **Muon identification efficiency** is determined with respect to all tracks matched to a simulated track
- **Computational performance (throughput)** measured with events representative of the Run 3 conditions on several GPU cards

GNN based track finding in VELO

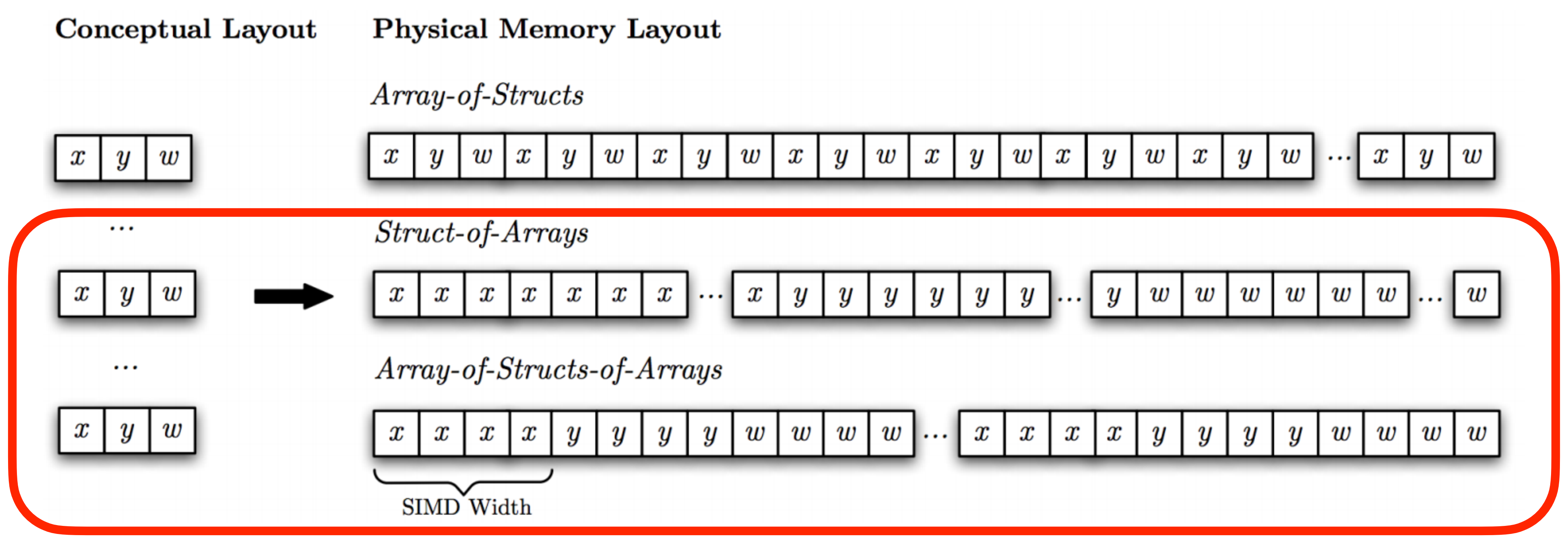
LHCb-FIGURE-2023-024

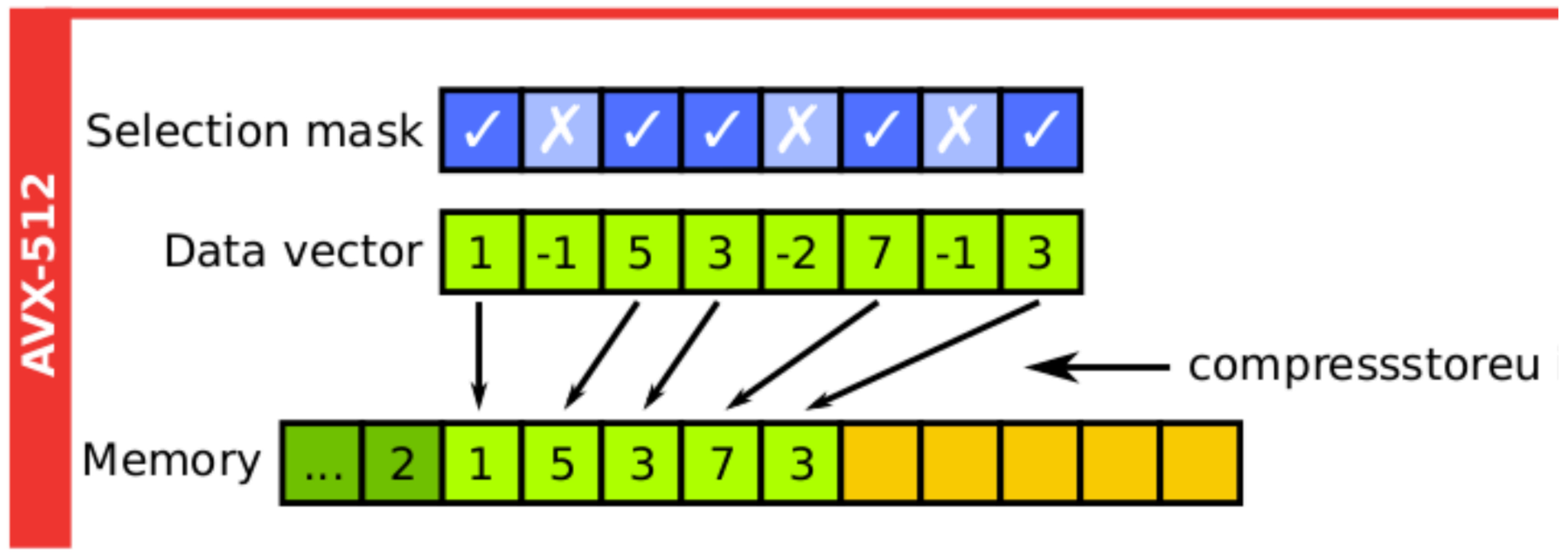


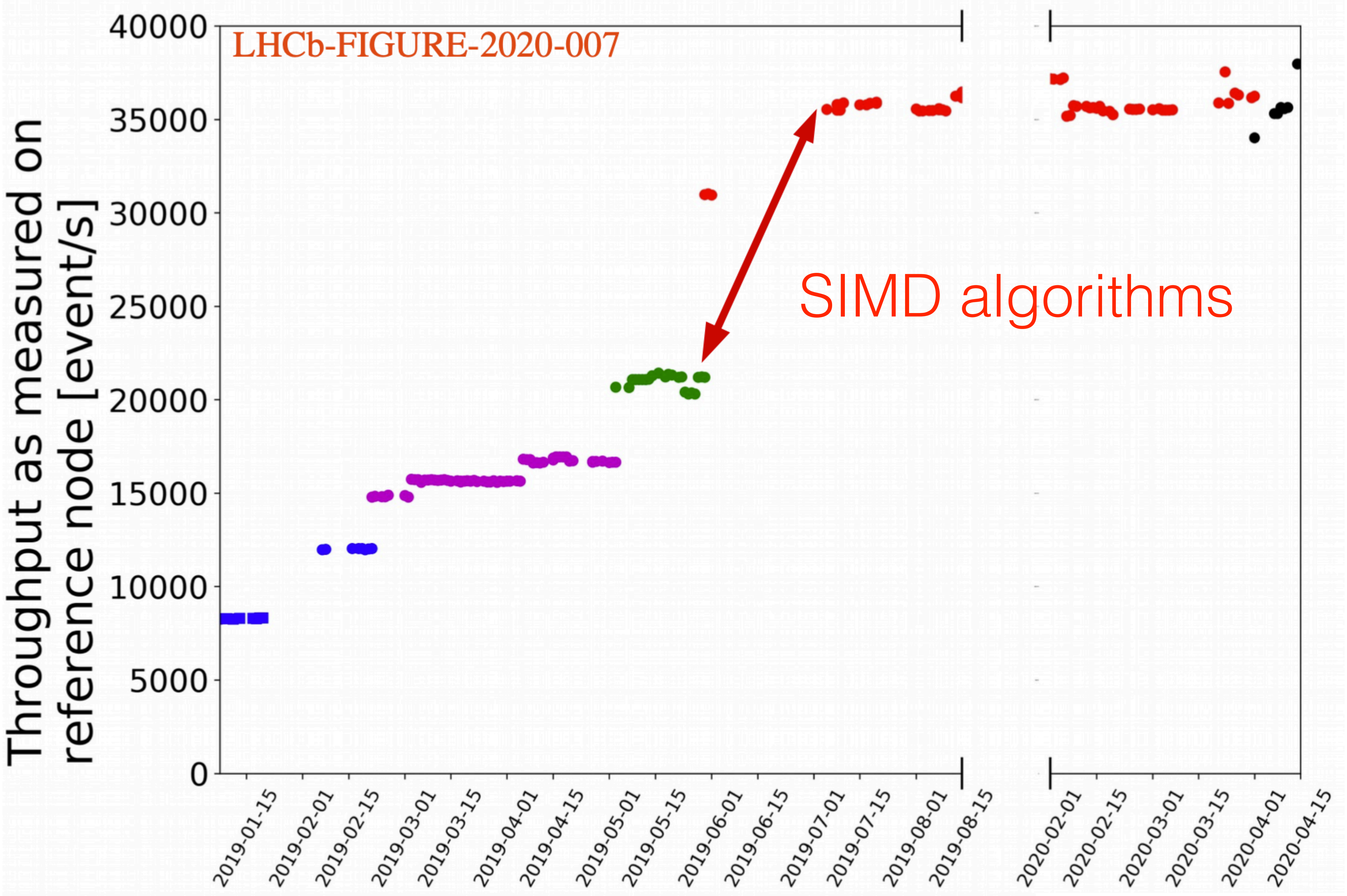


Parallelisation

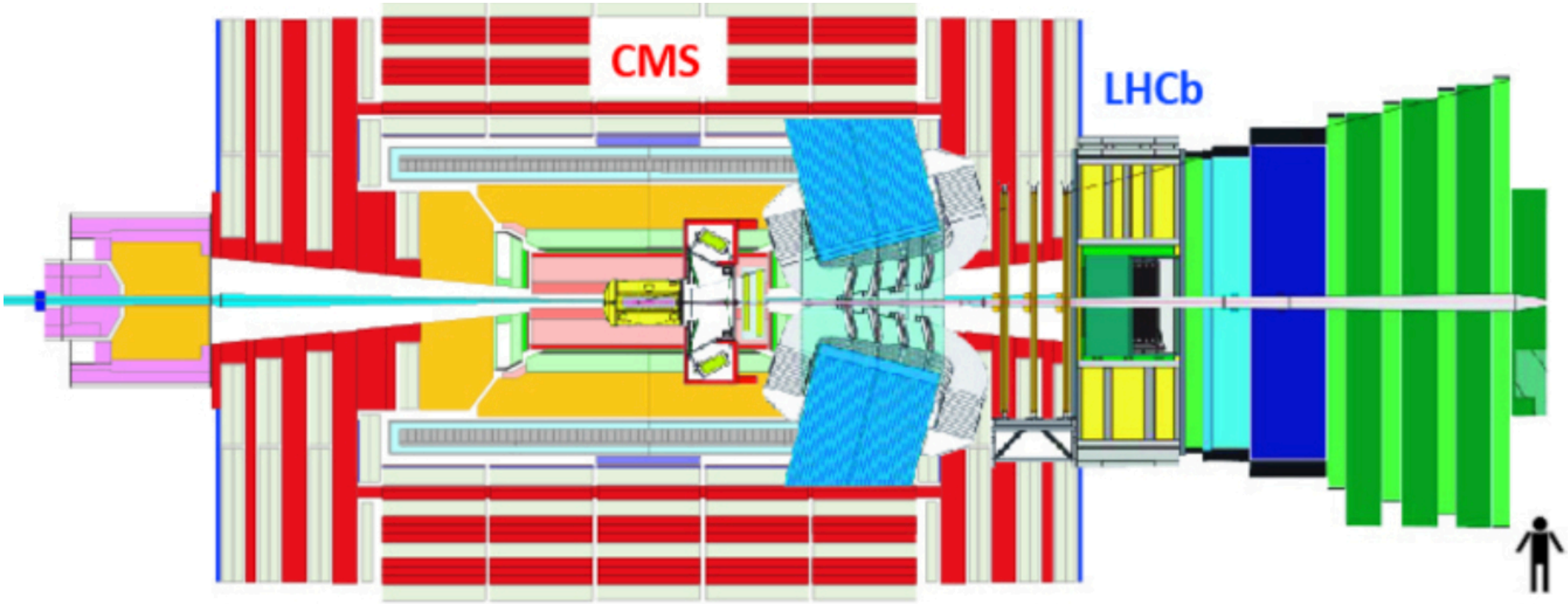
- Common intra-event parallelisation techniques as in GPU
- Significantly speed up the reconstruction







- Rewrote all reconstruction algorithms with SOA structure
- Developed custom SIMD wrappers to support all the backends (SSE, AVX2..)


CMS
LHCb