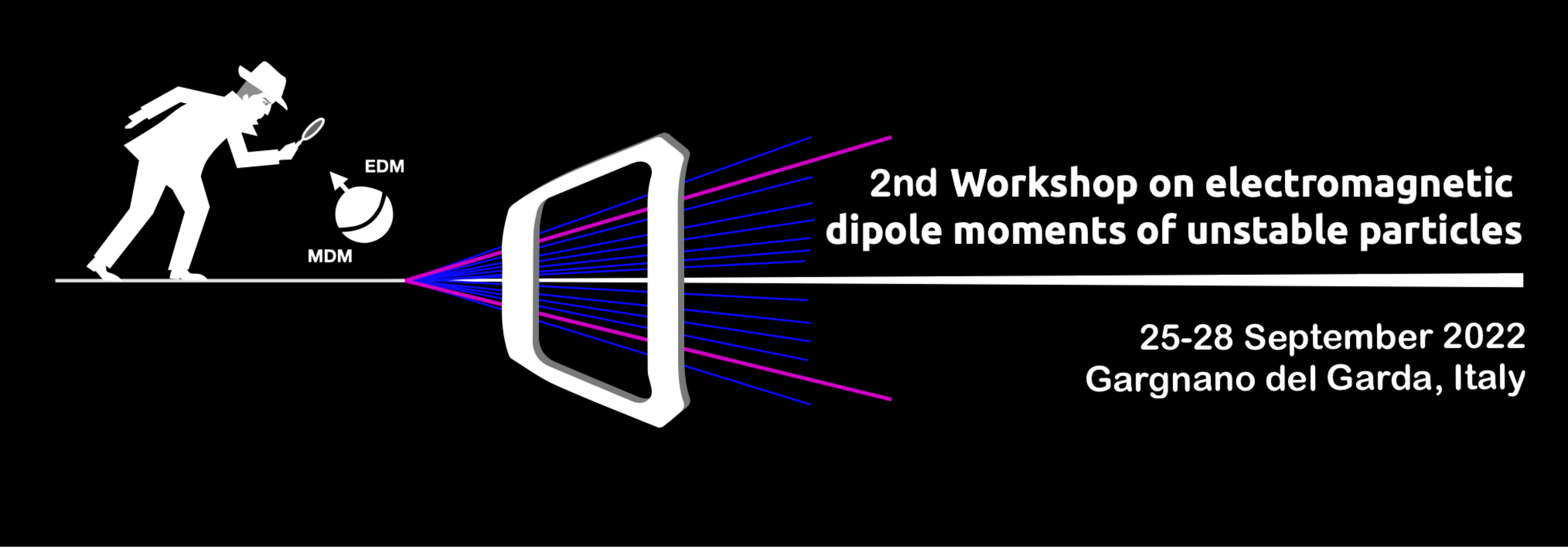
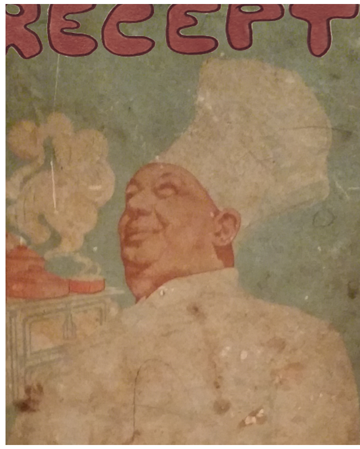


Trigger strategy for long lived particles

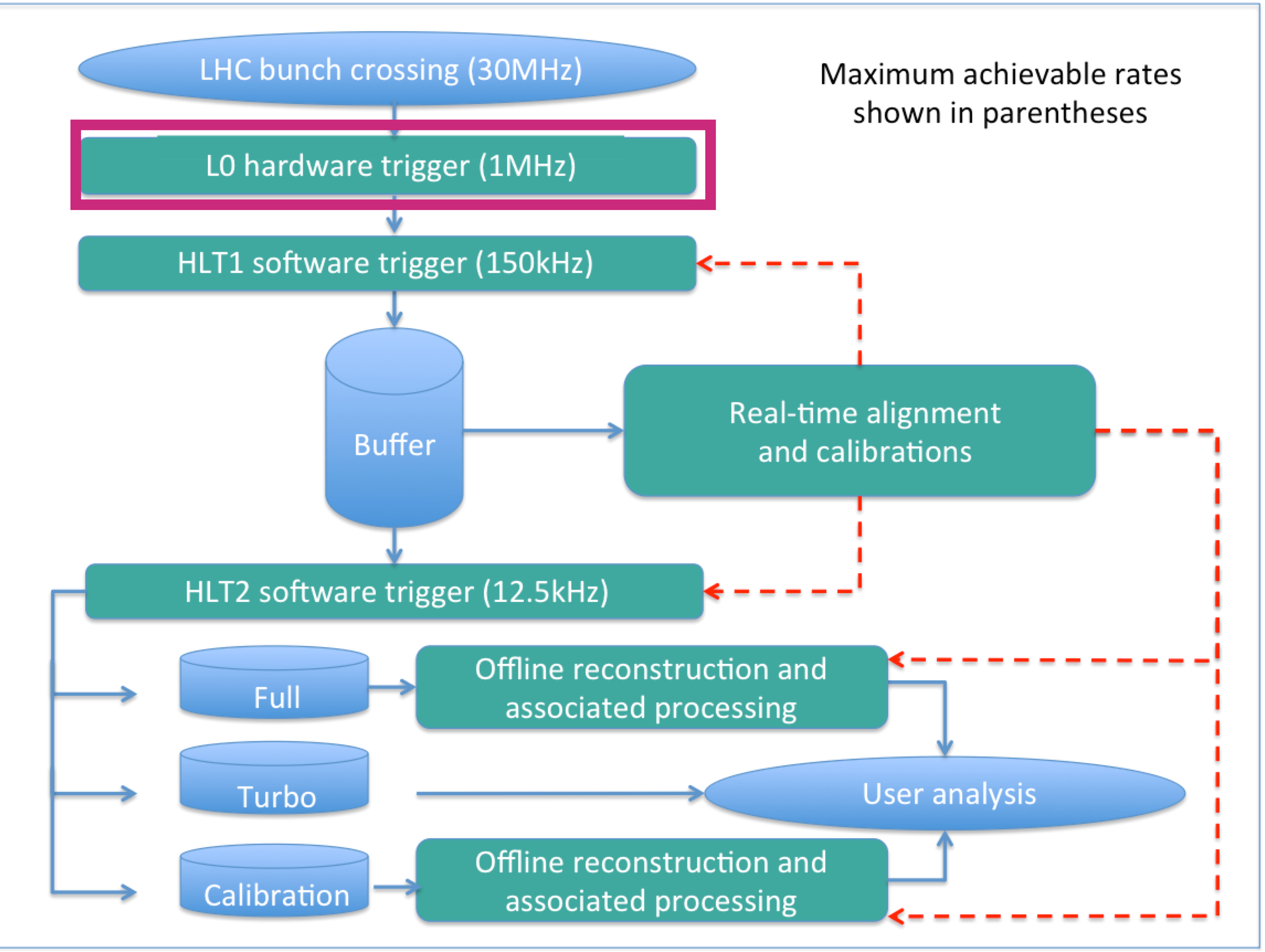
C. Agapopoulou (LPNHE), L. Calefice, V. Gligorov, A. Hennequin, L. Henry, B. Kishor Jashal, A. Oyanguren Campos, L. Pica, V. Svintozelskyi, J. Zhuo

LPNHE Paris
Sorbonne Université

European Research Council
Established by the European Commission



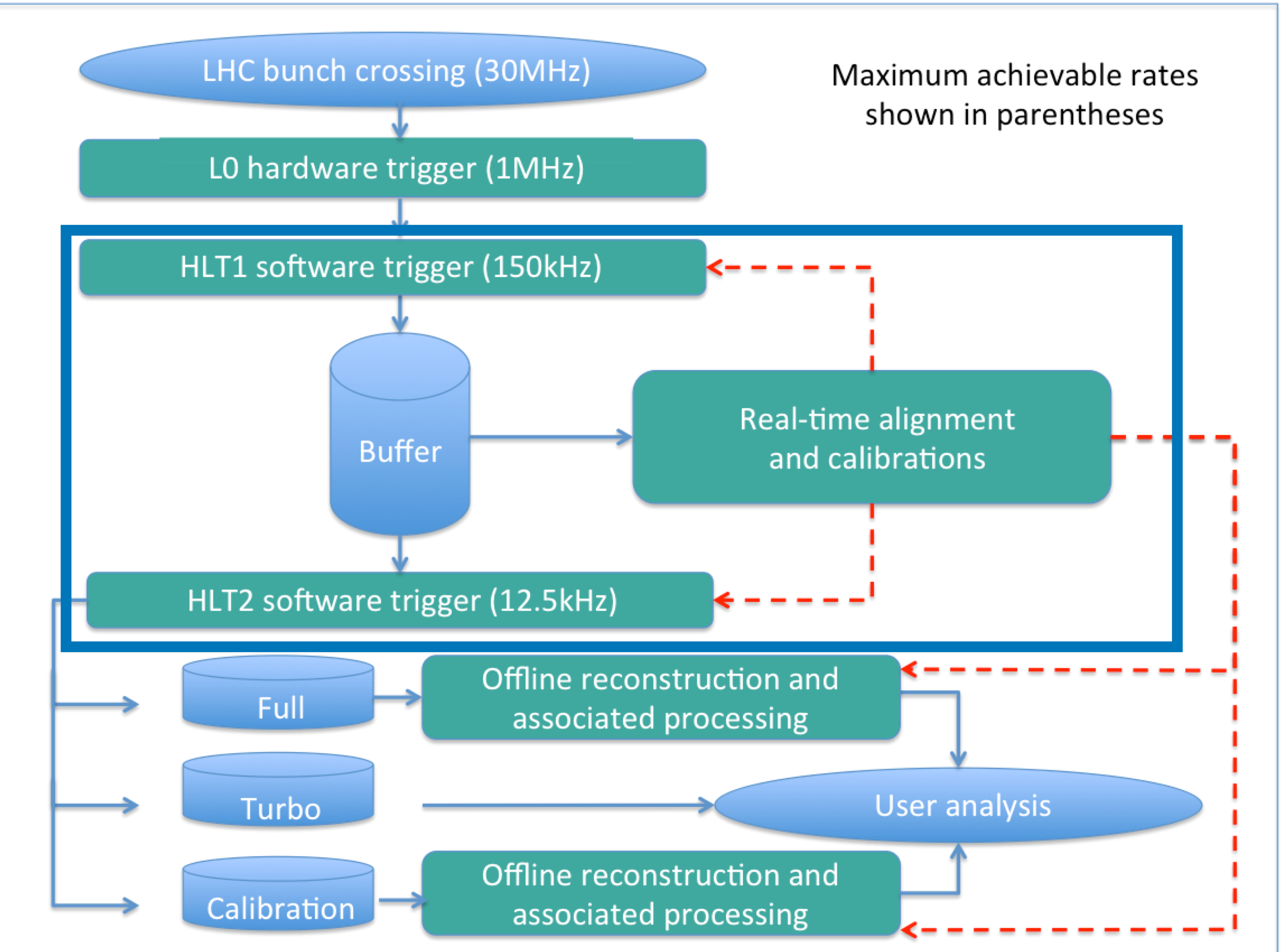
The LHCb Run1/2 trigger



L0 hardware trigger

- Calorimeter clusters and/or muon signals
- Reduced rate down to 1 MHz
- Made software trigger job easier but…
- Low efficiency for low-momentum signatures + saturation for fully hadronic channels!

The LHCb Run1/2 trigger

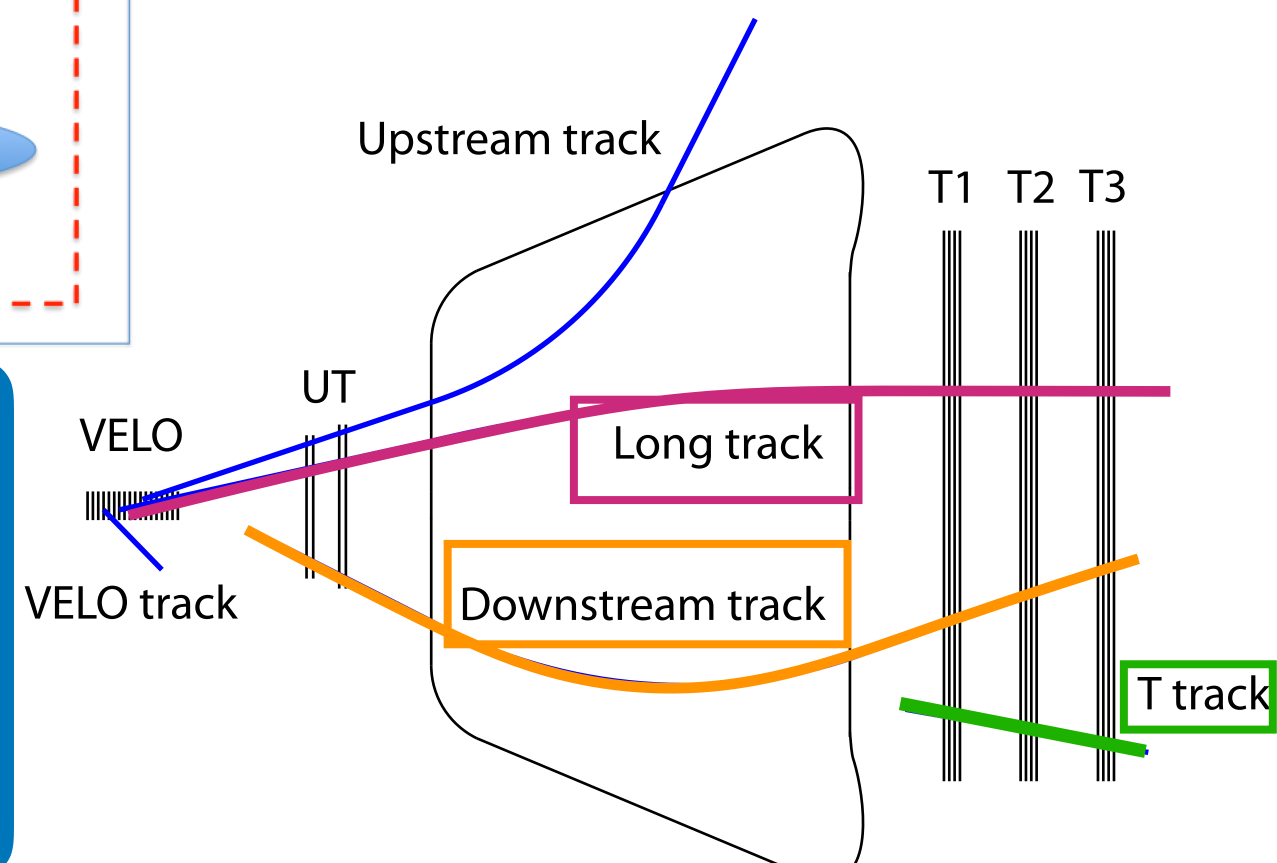


L0 hardware trigger

- Calorimeter clusters and/or muon signals
- Reduced rate down to 1 MHz
- Made software trigger job easier but…
- Low efficiency for low-momentum signatures + saturation for fully hadronic channels!

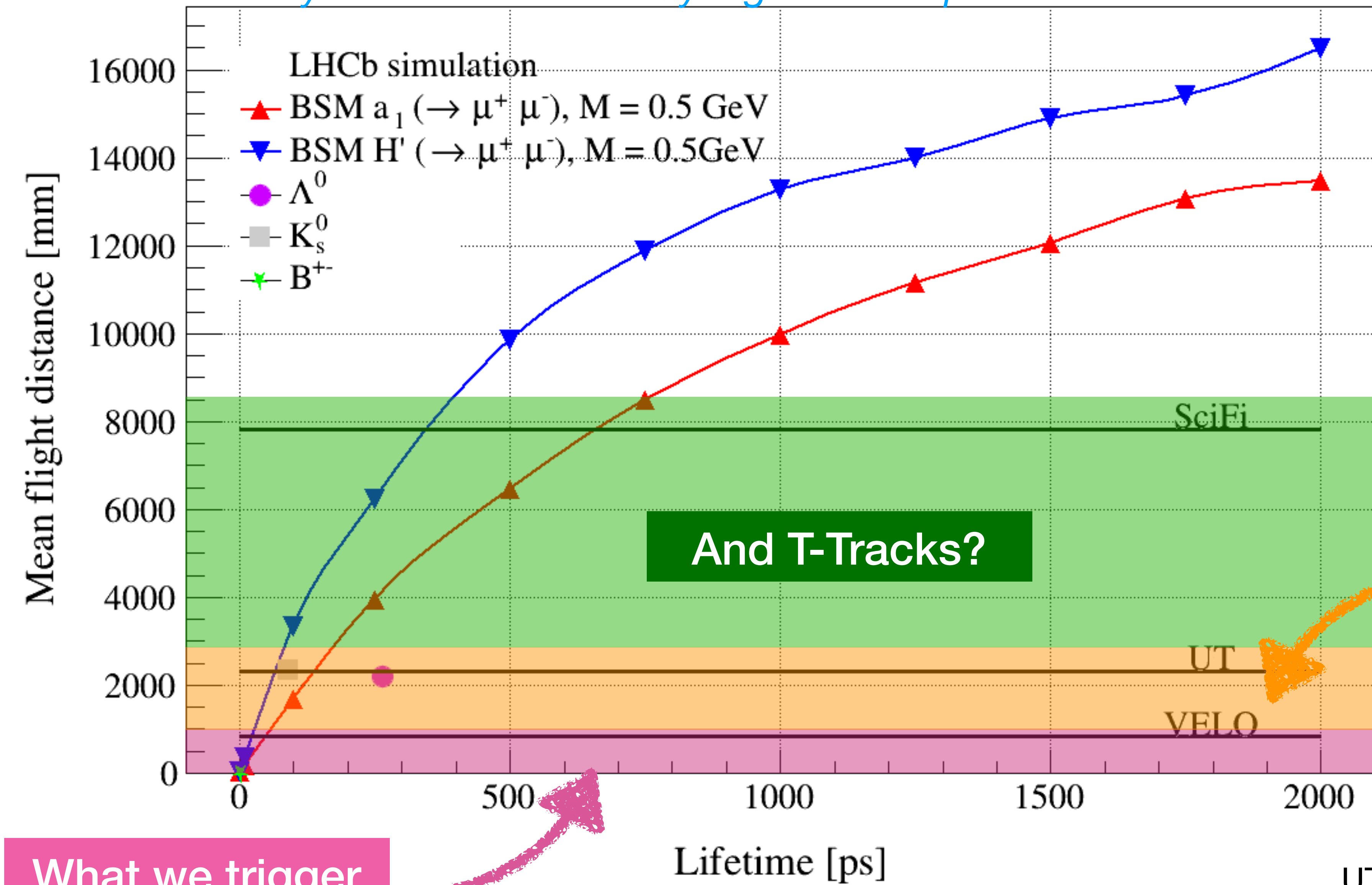
Software High Level Trigger (HLT)

- Split into two steps, HLT1 (coarse) & HLT2 (refined)
- Real time alignment & calibration
- Trigger heavily based on track & vertex reconstruction



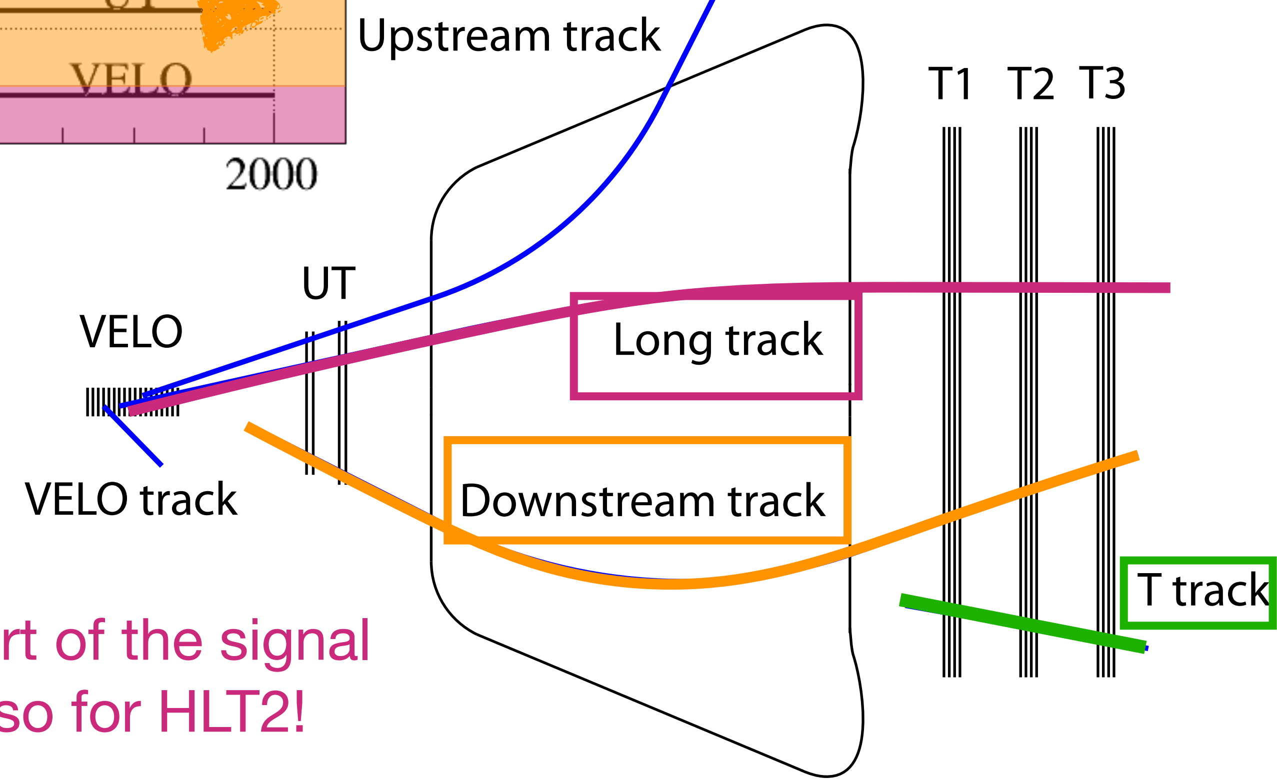
LLPs reach at LHCb

Plot by D. Mandoza & A. Oyanguren Campos



LHCb's unique rapidity coverage and PID capabilities extremely interesting for LLP searches

+ Downstream only in HLT2



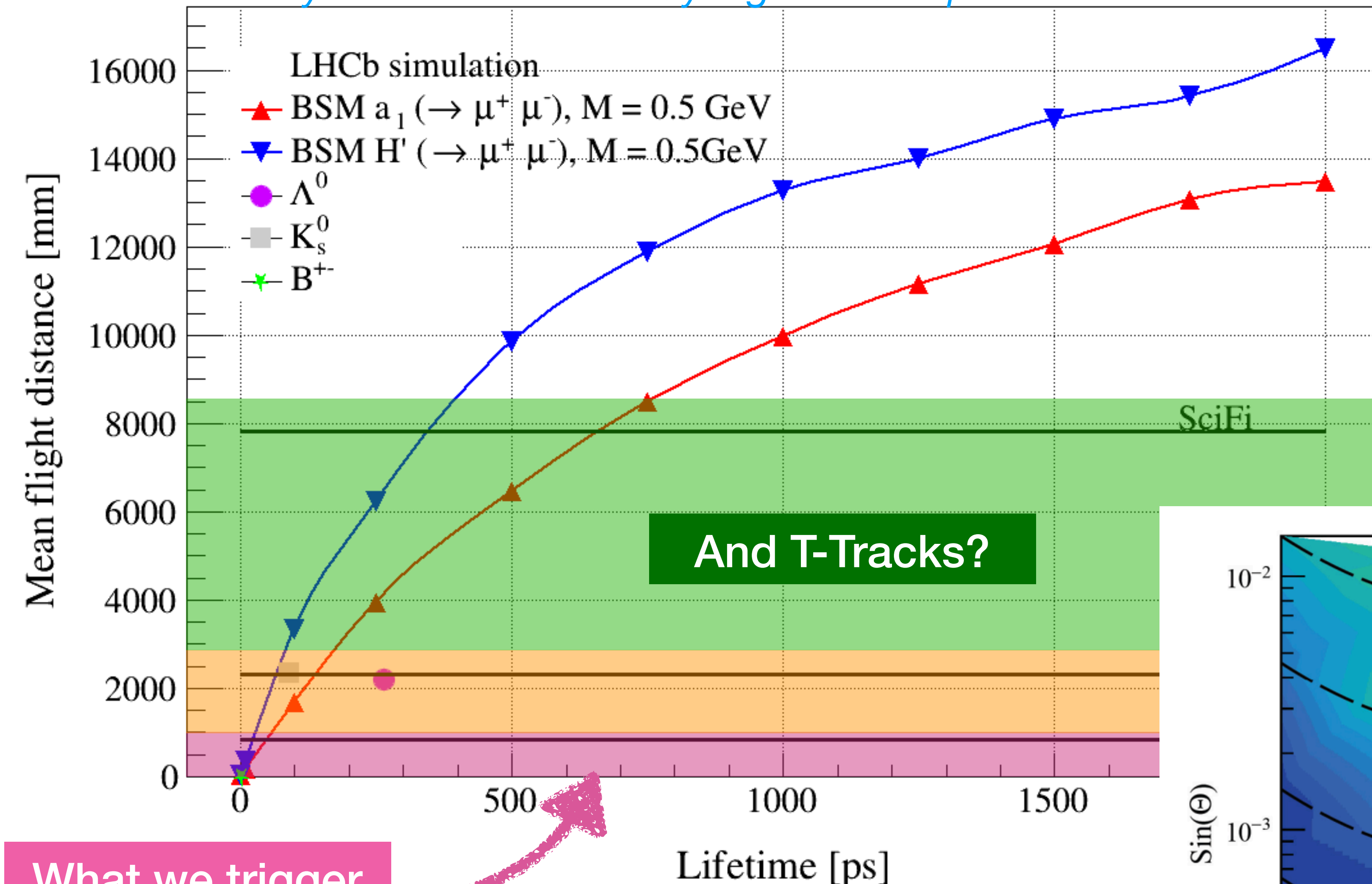
What we trigger on in HLT1

But currently not exploited to its full potential!

LLP analyses need to rely on long tracks that are not part of the signal to pass HLT1, and if the decay is after the magnet, also for HLT2!

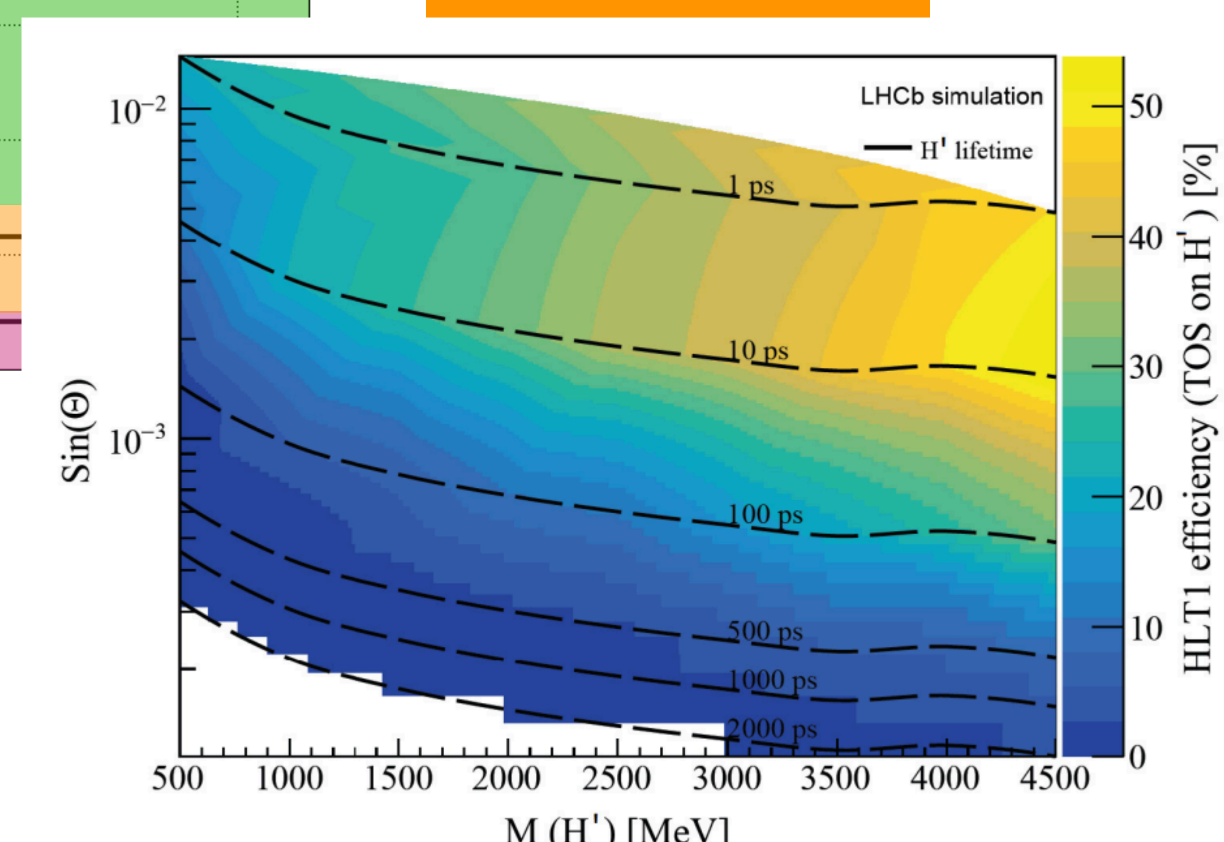
LLPs reach at LHCb

Plot by D. Mandoza & A. Oyanguren Campos



What we trigger on in HLT1

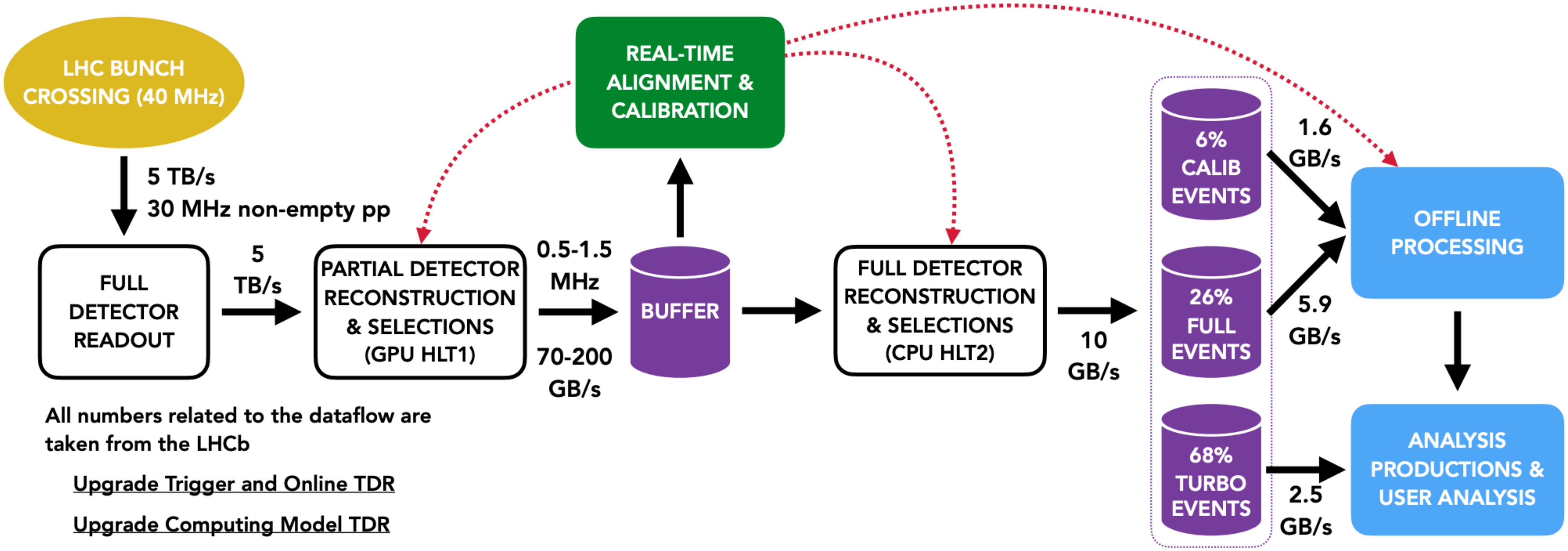
LHCb's unique rapidity coverage and PID capabilities extremely interesting for LLP searches



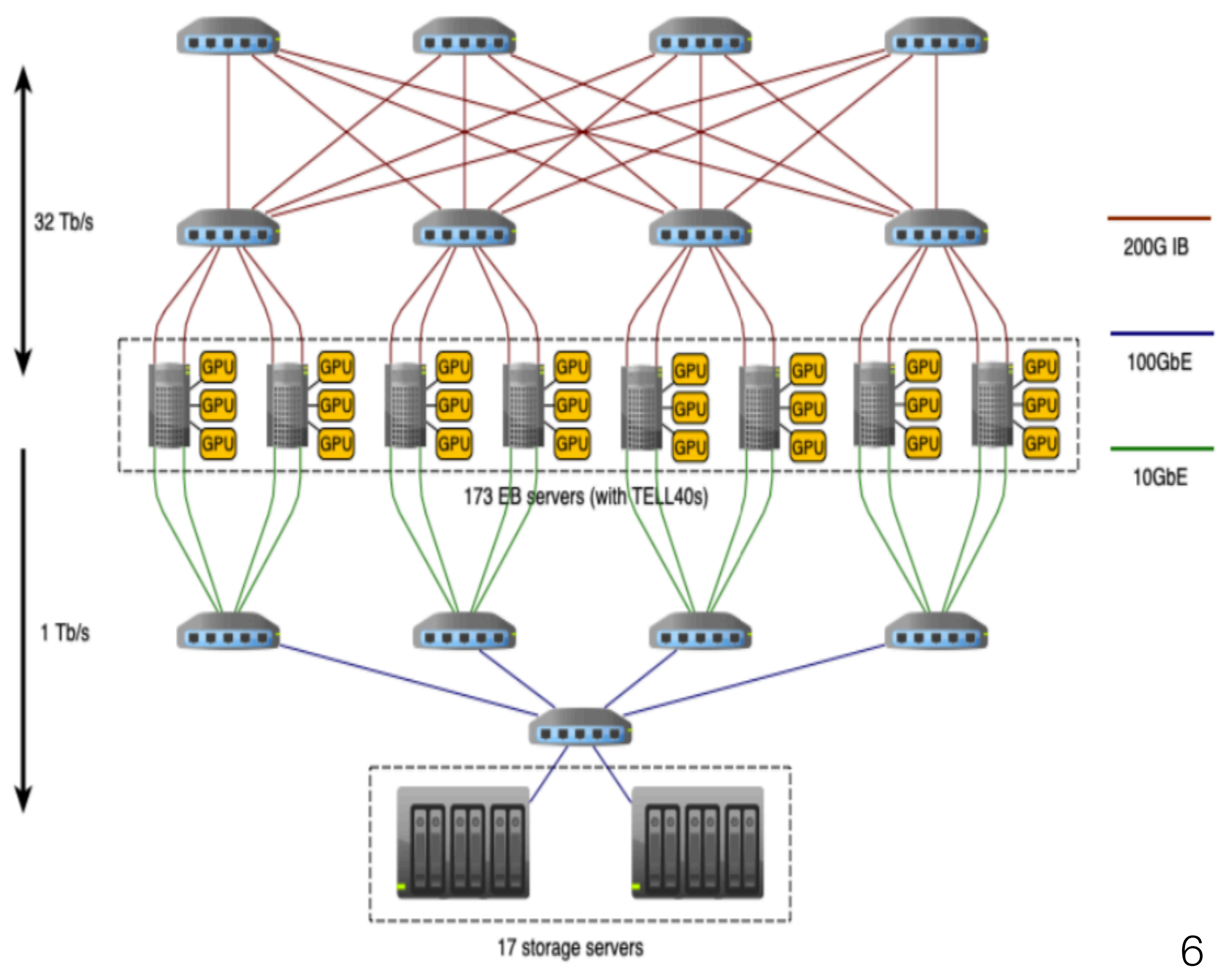
But currently not exploited to its full potential!

LLP analyses need to rely on long tracks that are not part of the signal to pass HLT1, and if the decay is after the magnet, also for HLT2!

The potential of the new Run 3 trigger



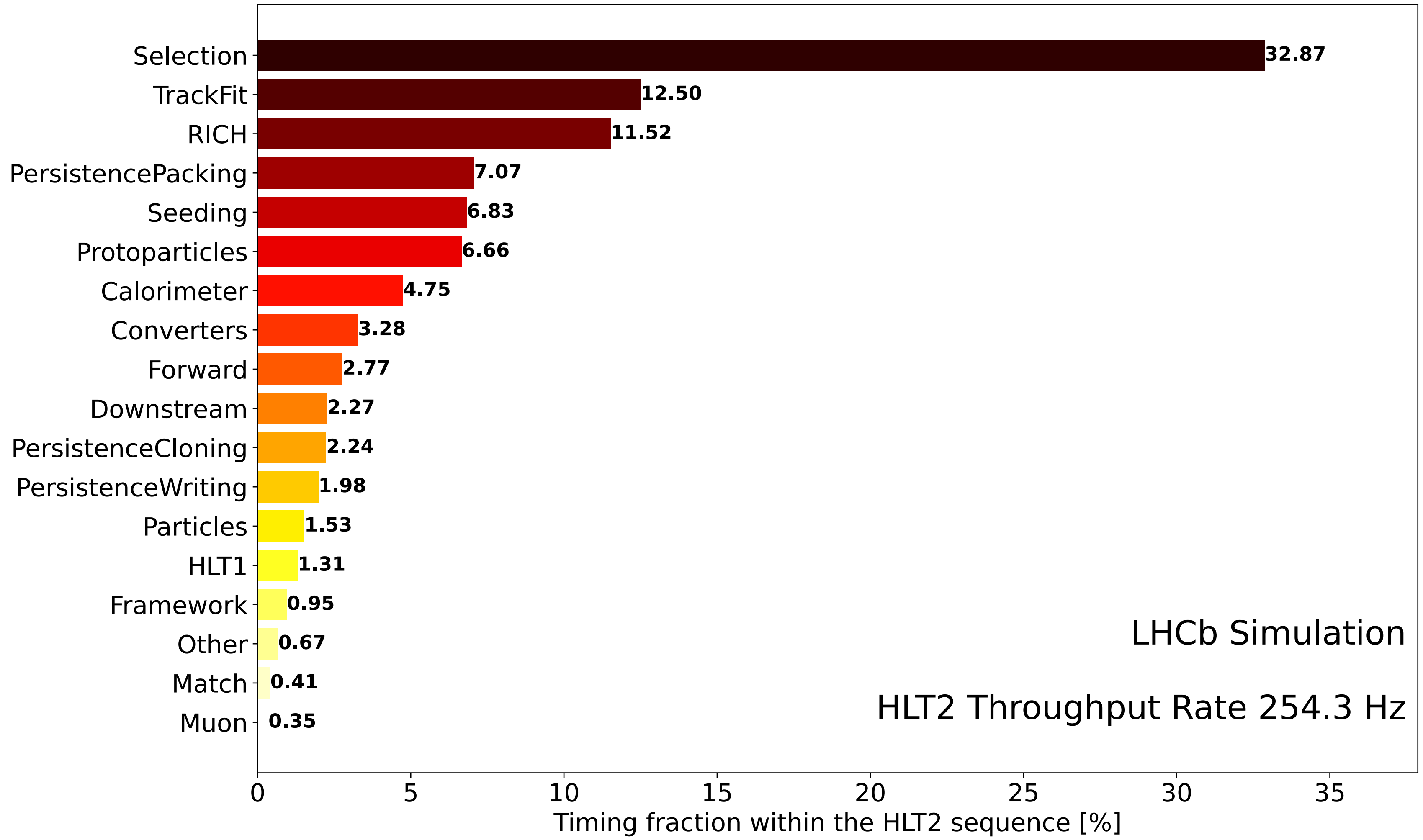
- In Run 3, complete removal of L0 trigger + reimplementation of **HLT1 entirely on GPUs!**
- Ideal for highly parrallelizable algorithms (such as tracking)
- **30 MHz** input to the Event Builder (EB), equipped with ~170 GPUs
- Reduce rate by x30 to be able to write out at **1 MHz**
- For start of Run 3, 1 GPU / server →2 remaining free slots per server, system can be upscaled in the future



LLP triggers in HLT2

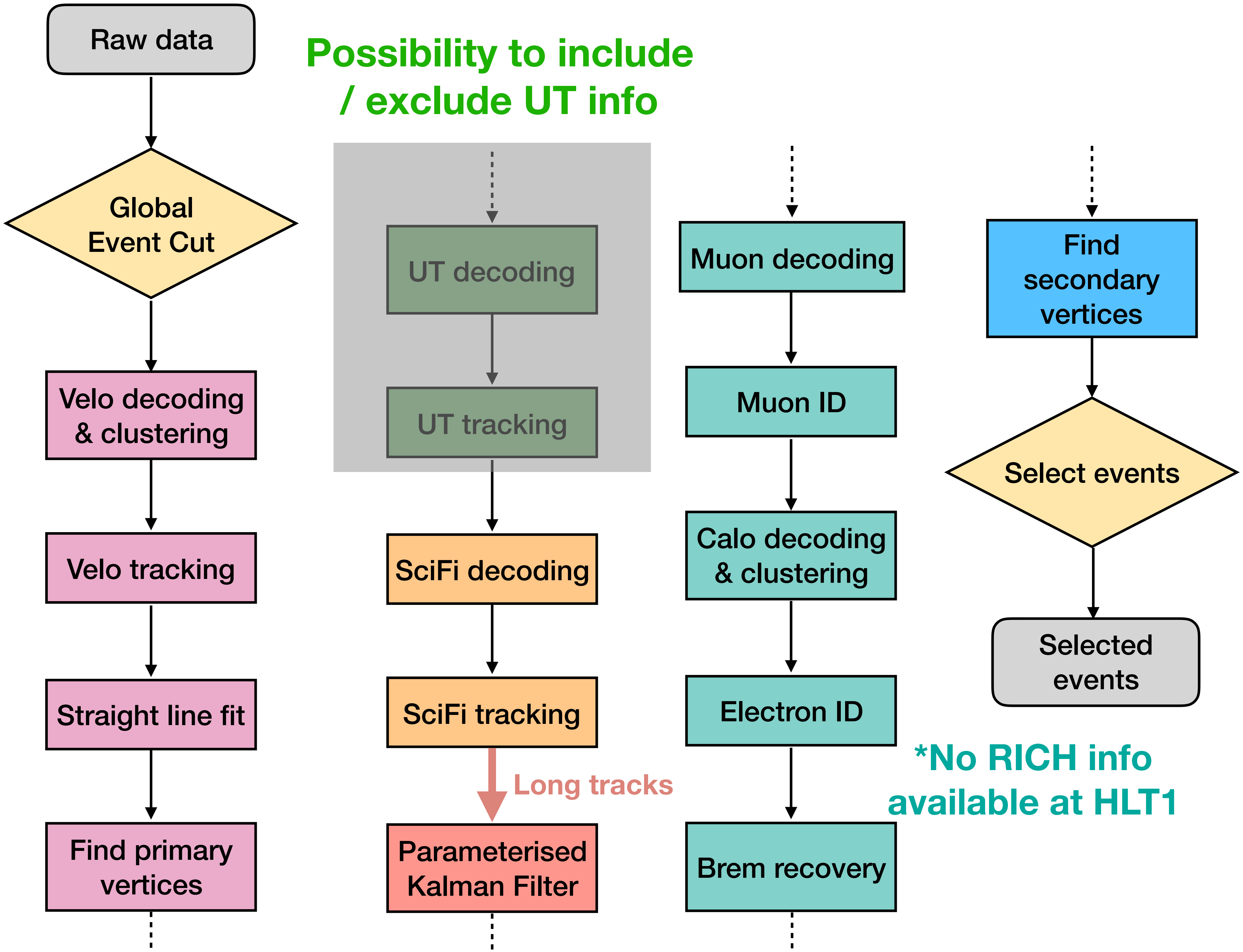
- HLT2 triggers rely on offline-quality reconstructed objects
  - T-track reconstruction from L. Henry, I. Sanderswood & M. Wang
  - LLP vertexing from G. Tonani

- Selections takes up 30% of total HLT2 throughput but
  - O(1000) lines envisioned, what's a few more?



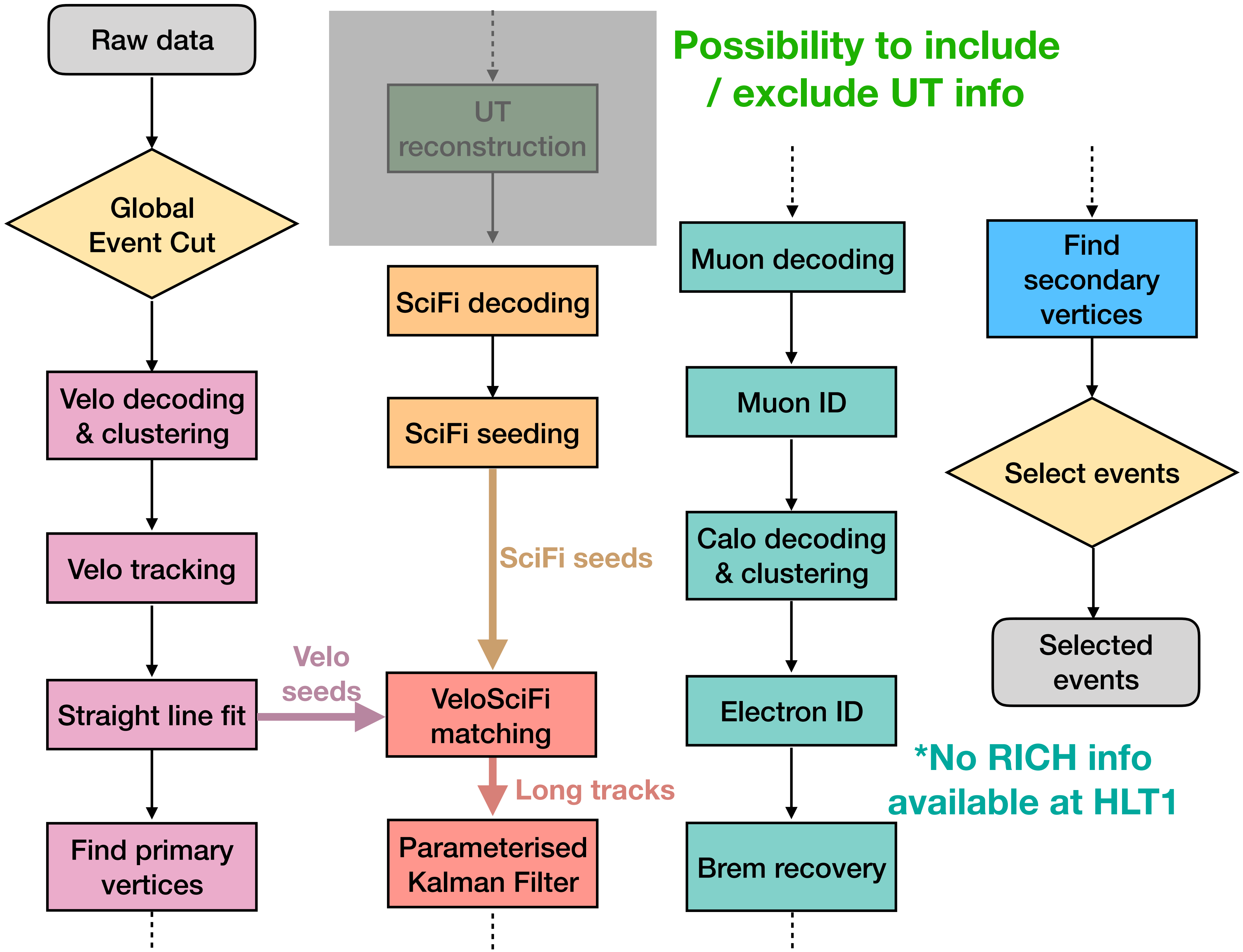
So what about HLT1?

Original HLT1 algorithm sequence



Reconstruction and selections heavily based on tracking… but only long tracks!

Seeding & matching at HLT1



Alternative long track reconstruction based on SciFi seeding - can enable Downstream reconstruction and T-track selections at the first trigger level!

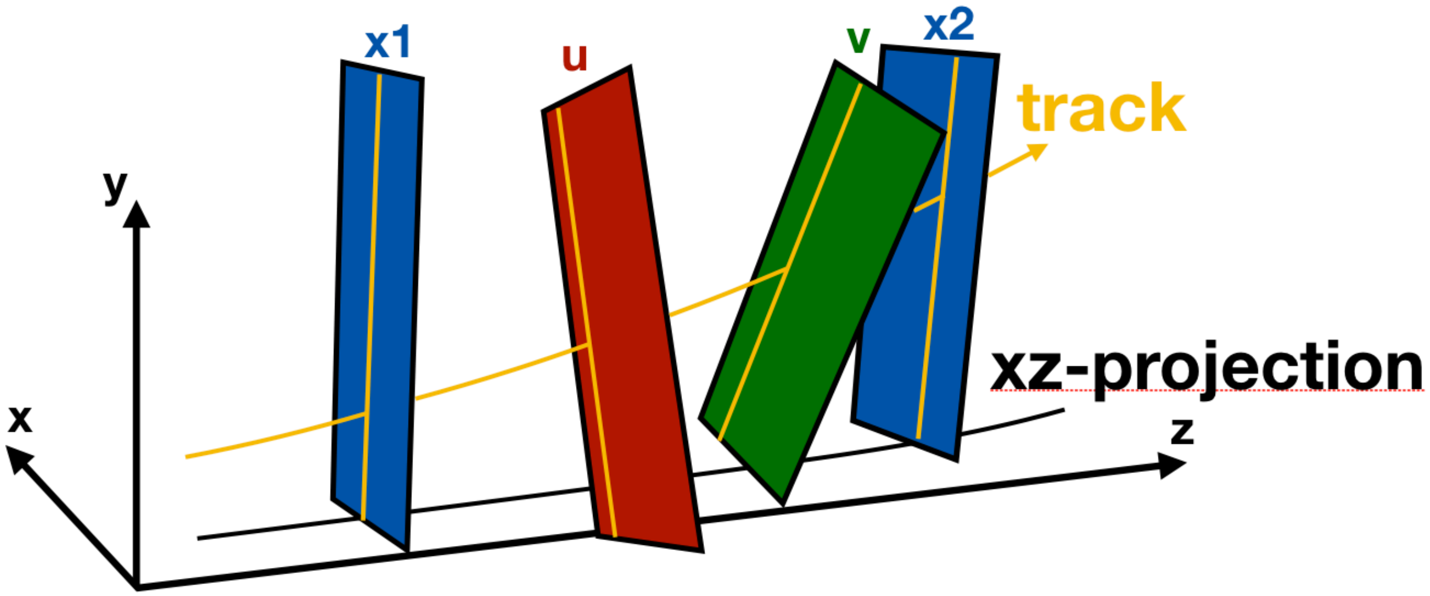
Seeding and matching algorithms

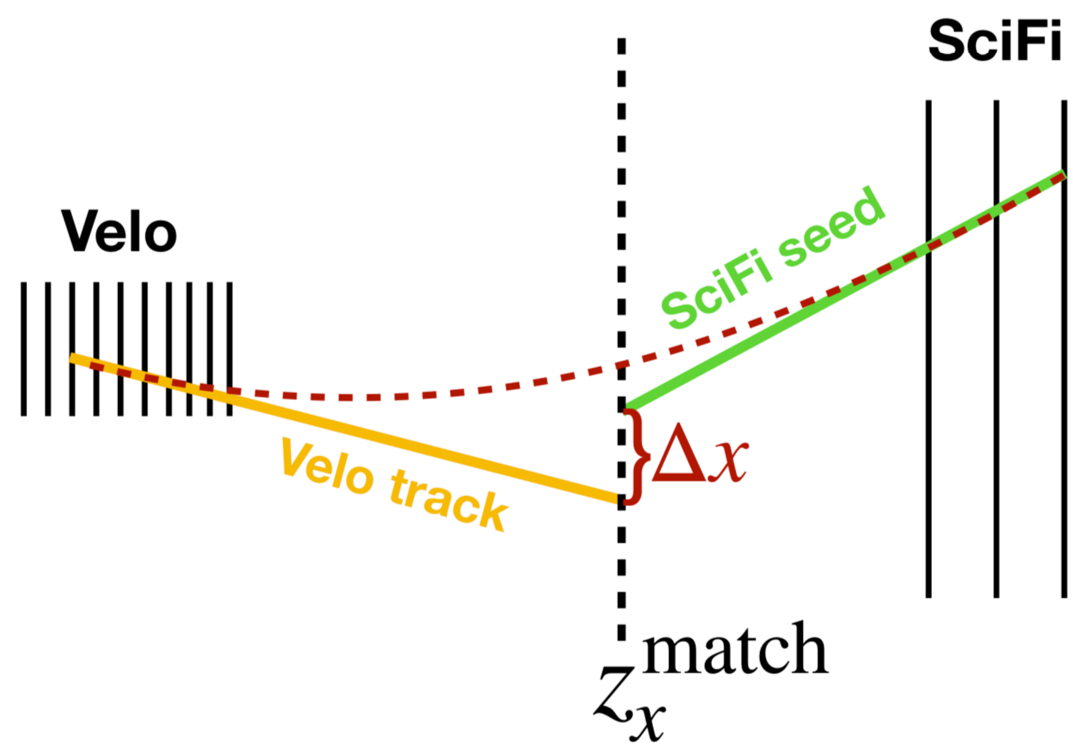
HLT1 Seeding:

- Adapted from HLT2 T-Track reconstruction
- Simplified version to increase speed (optimised for p > 3 GeV/c, tighter hit requirements)
- GPU-adapted: parallelisation across all hits, fewer cases, different UV-hit treatment
- Seeds with $dp/p \sim 10 - 30\,\%$

(Velo-SciFi) Matching:

- Input track seeds extrapolated to matching position as straight lines
- Parametrisation of magnetic field to calculate optimal matching position
- Matching to VELO brings dp/p down to 1-2%
- Can be adapted to UT-SciFi matching for downstream track reconstruction



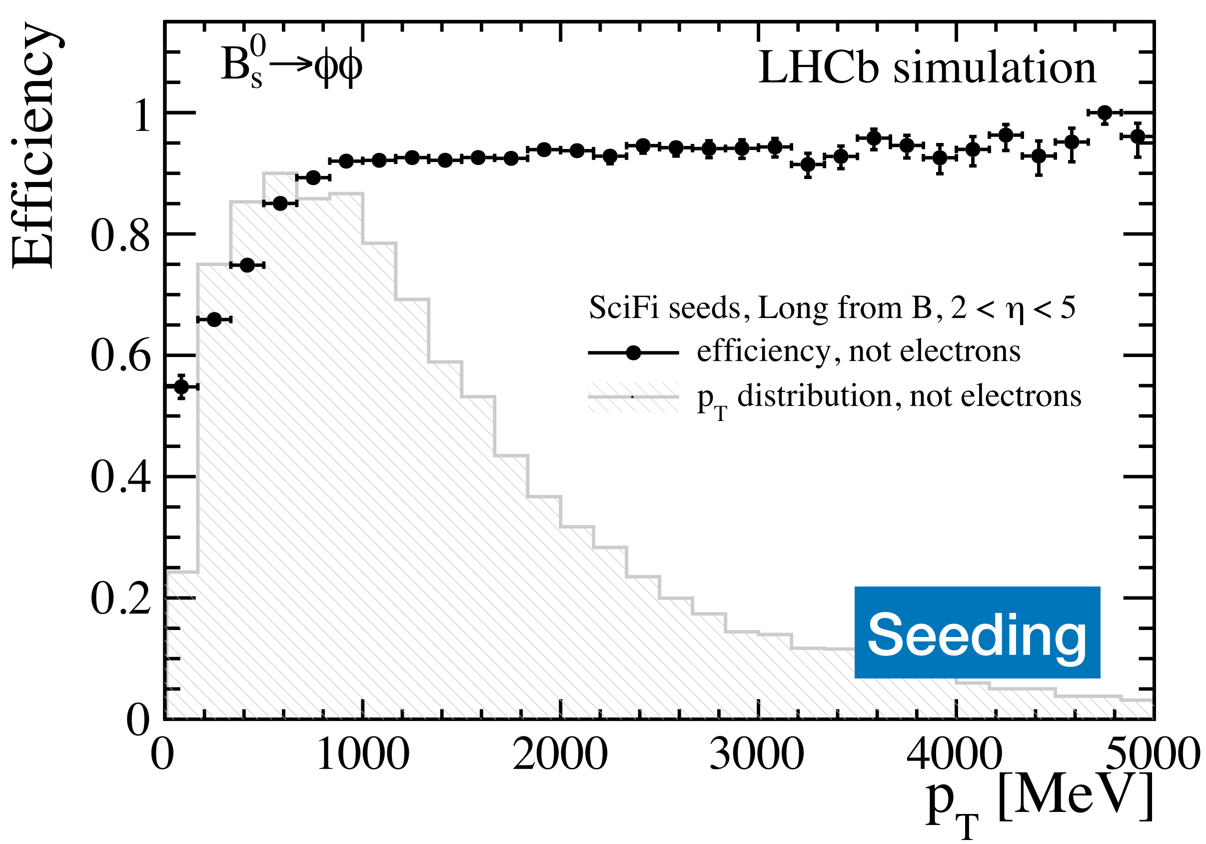


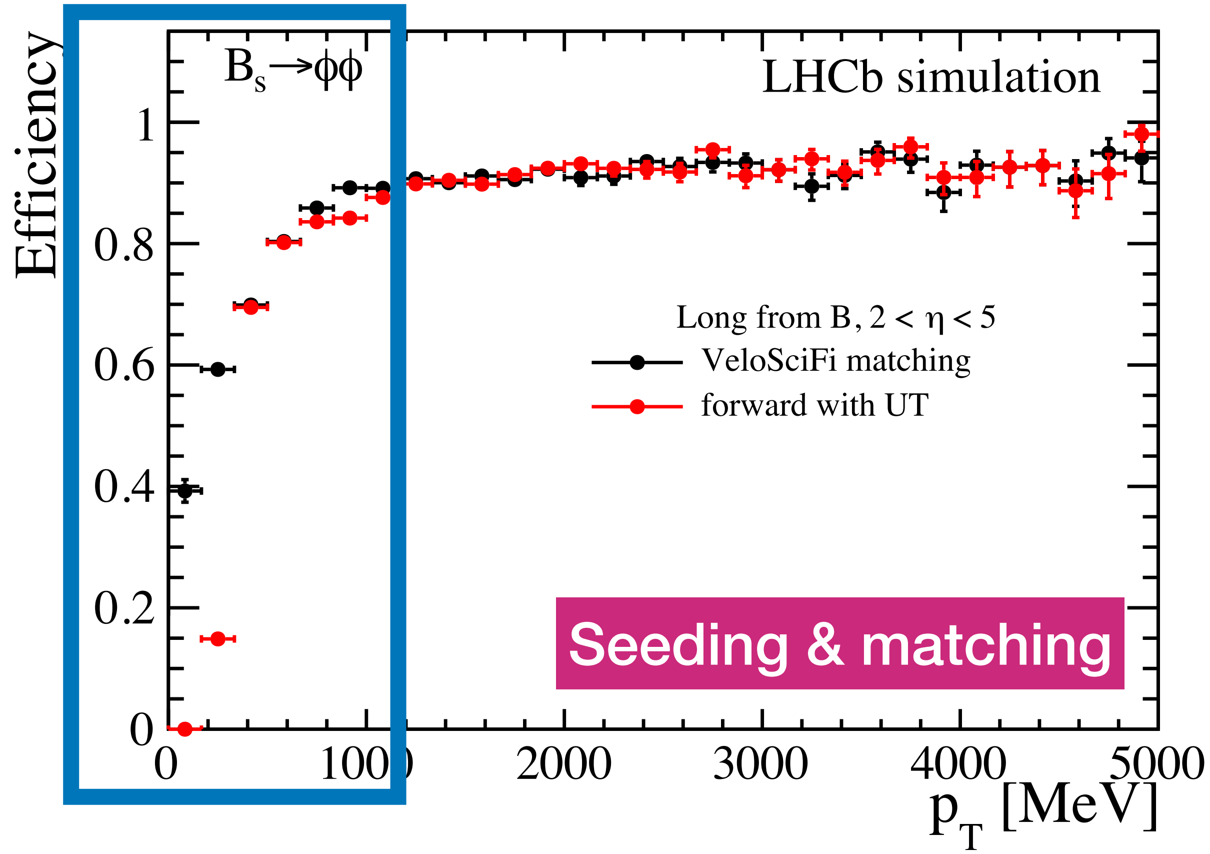
HLT1 seeding&matching performance

Efficiency	Seeding	Velo-SciFi matching	Forward (with UT)
Ghost rate	10 %	9 %	5 %
Long, p > 3 GeV	85 %	78 %	55 %
Long, p > 5 GeV	86 %	82 %	63 %
Long from B & p > 3 GeV	89 %	85 %	76 %
Long from B & P > 5 GeV	90 %	87 %	81 %
No VELO, UT+SciFi, p > 3 GeV	83 %	/	/
No VELO, UT+SciFi, p > 5 GeV	86 %	/	/
Long from B, electrons, P > 5 GeV	79 %	76 %	73 %

$B_s \to \phi\phi$

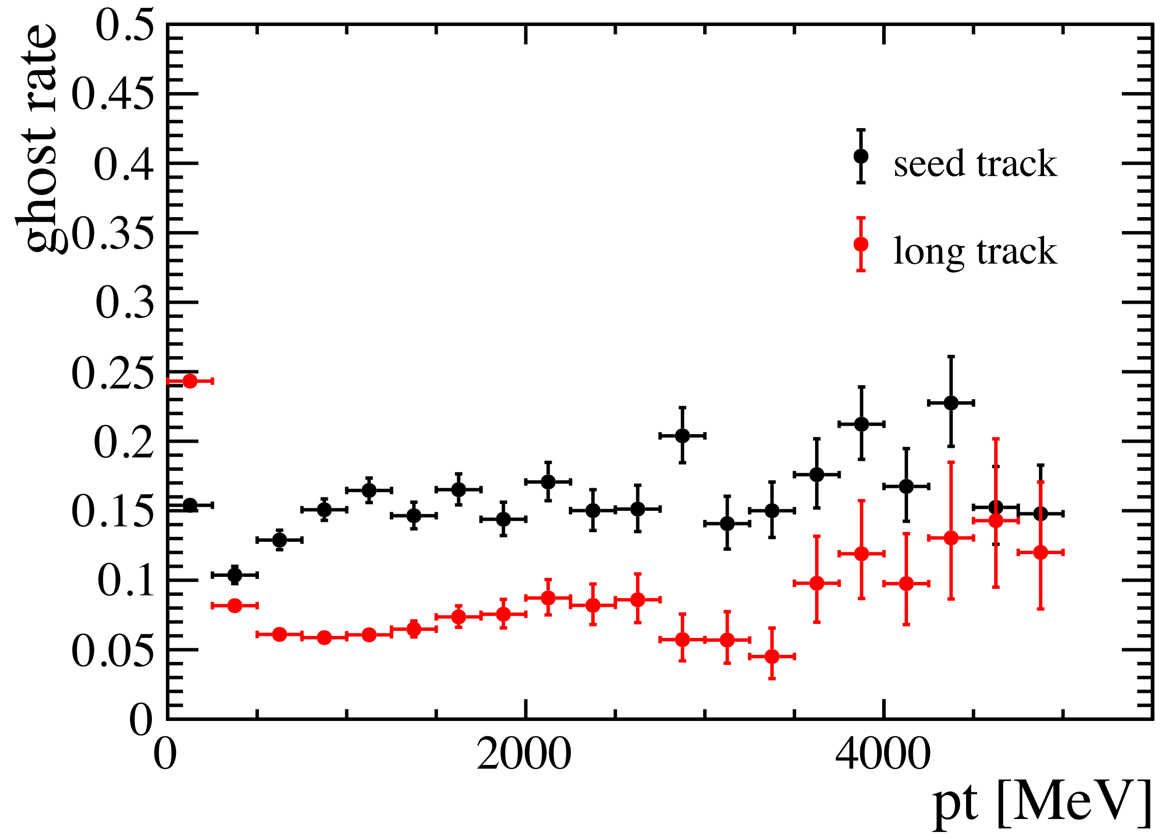
$B^0 \to K^{*0} e^+ e^-$

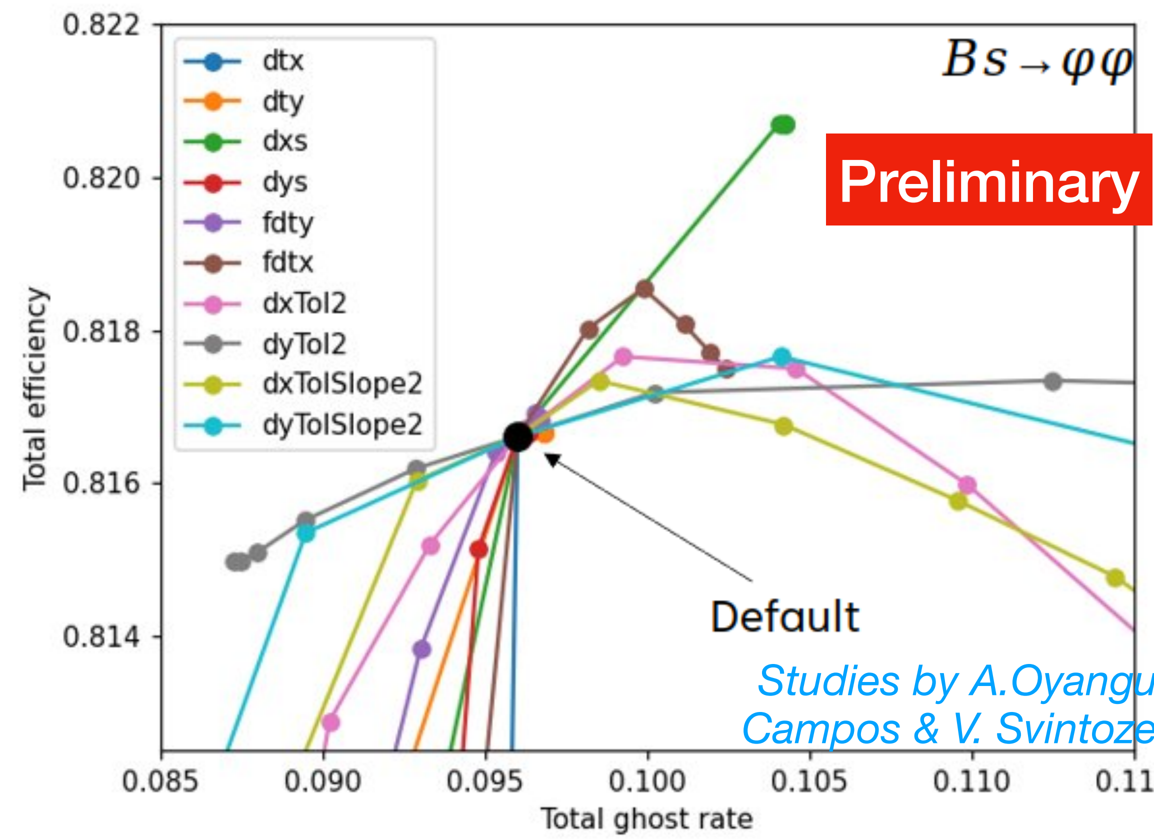




Ghost rate in HLT1

- Ghost rate ~ **10 %** for seeding, reduces to **8%** for long tracks
- Single-track lines are very sensitive to ghosts
- Long/Downstream tracks can profit from **UT information** to reduce the ghost rate (confirmed with the forward algorithm and simulation)
- However, **T-tracks will require particular care**
- Algorithm parameter scan:
  - Already performed for long tracks
  - **Ghost rate & equivalent tracking trigger rates brought down by 10% with less than 1% loss in efficiency**
  - Exercise to be performed on seeding algorithm
  - Exploring more advanced fake rejection techniques (MVA) - dedicated treatment required for the GPU architecture

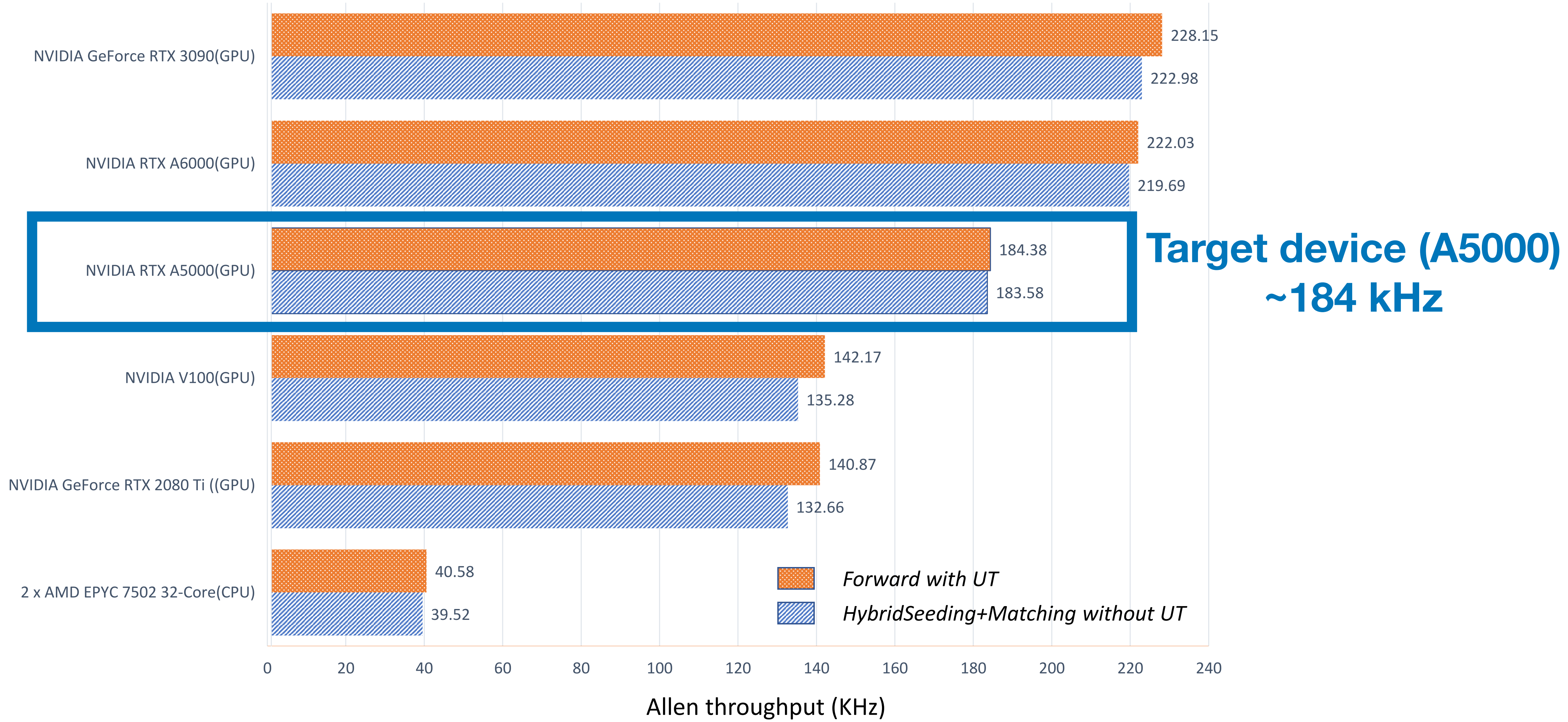




Studies by A.Oyanguren Campos & V. Svintozelskyi

Considerations for HLT1 LLP algorithms

Throughput



Reminder: HLT1 must run at **30 MHz**

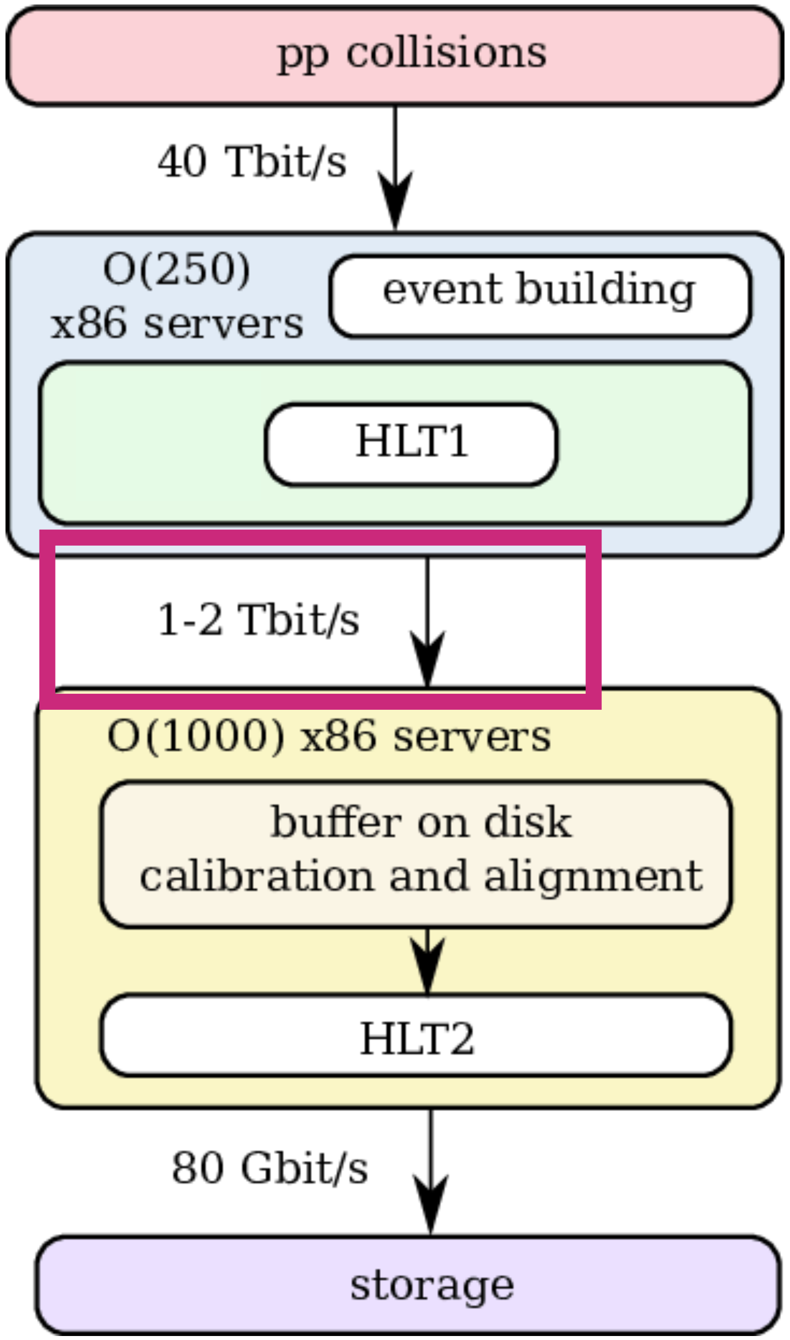
- With the currently installed ~200 GPUs, this translates to min **150 kHz / GPU**
- With only 20% margin, impact of any additional algorithms must be minimal… but

Reminder, we expect to double n. GPUs, this should allow for many more functionalities

Output rate limitations

- **Output rate limited to 1 MHz (~ 1 Tb/s)**
- Limitation from the writing speed to the HLT1-HLT2 buffer (online alignment & calibration)
- Current HLT1 rates way above limit → rescaling campaign planned when all lines are in place
- Largest fraction taken up by two-track / low p_T lines
- Seeding & matching access lower p_T tracks → higher rates
- Special care on rate must be taken for any new lines!

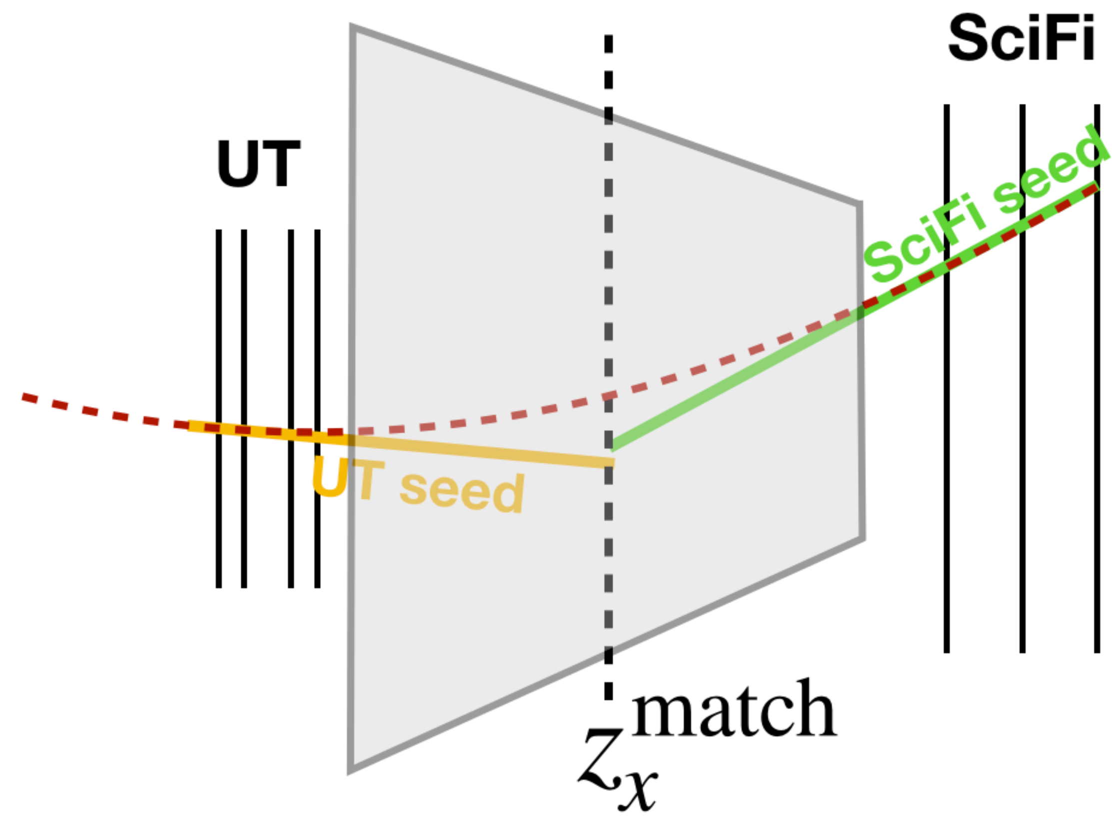
	HLT1-Matching (kHz)	HLT1-Forward (kHz)
TwoTrackMVACharm	5340	4689
TwoTrackMVA	858	744
SingleTrackMVA	708	432
KsToPiPi	1071	645
Hlt1LowPtMuon	2034	2490
Hlt1LowPtDiMuon	1782	408
Total	**11300**	**7800**



Adapting to Downstream Tracking

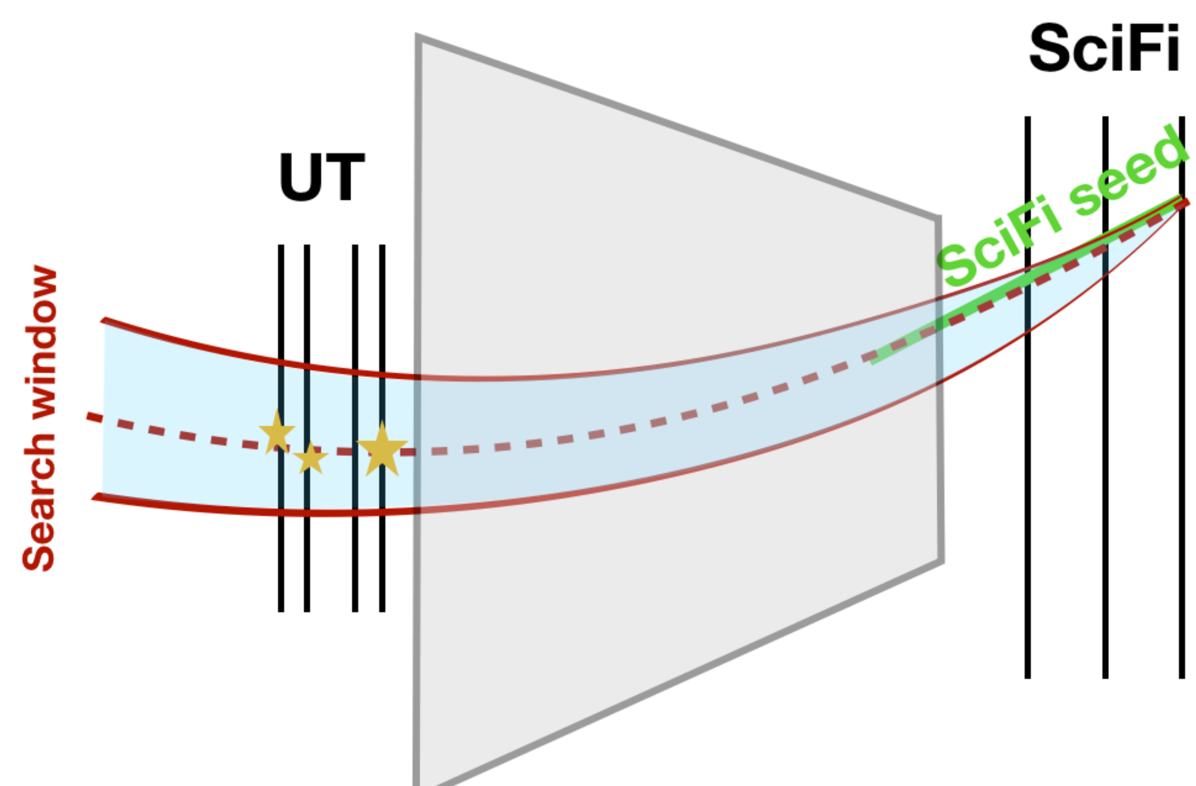
UT seeding:

- Standalone reconstruction of UT segments & matching with SciFi seeds
- Preliminary tests with entire UT give encouraging tracking efficiency, but far from throughput target
- Using only remaining UT hits (unused for long reco) quite promising to speed up algorithm
- Very well adapted for migration to FPGA tracking



Downstream tracking (ala HLT2):

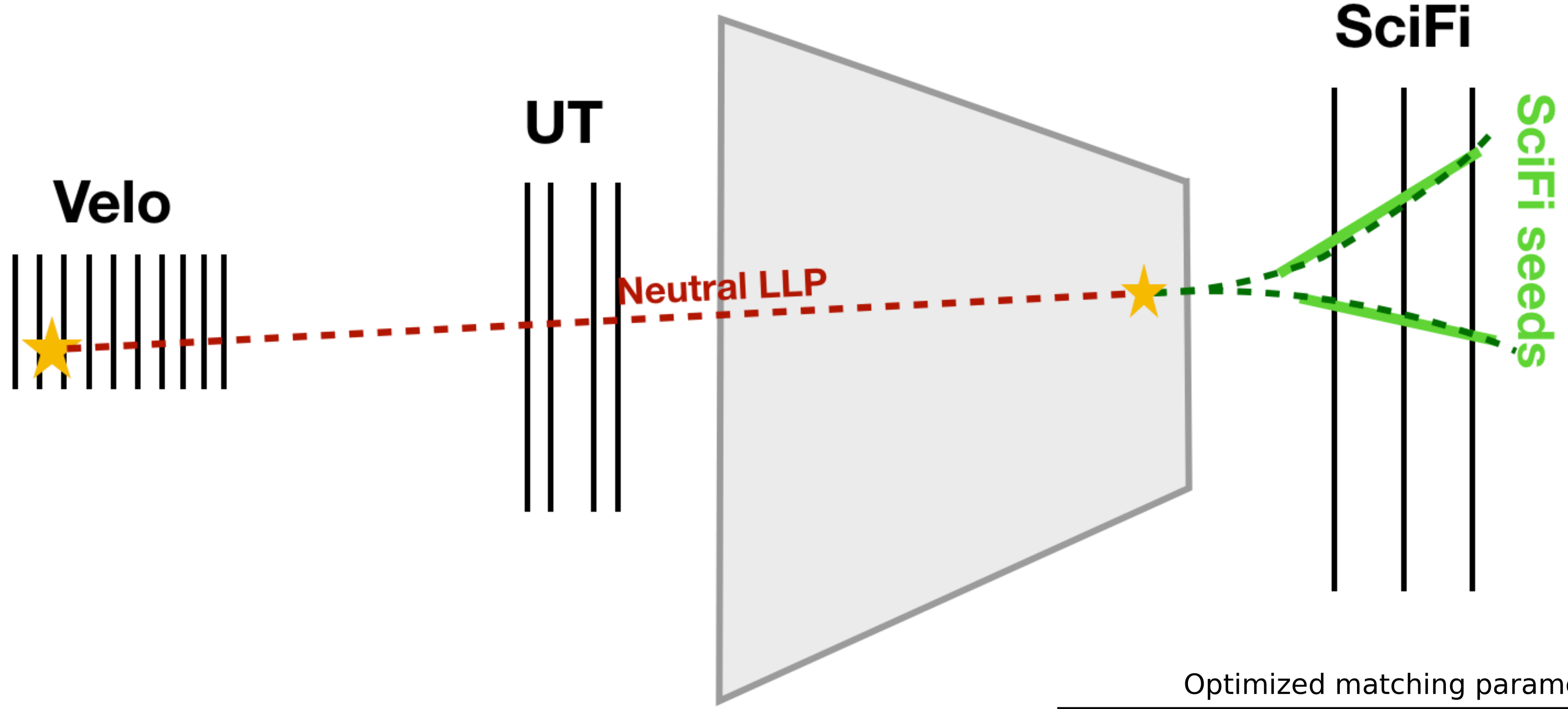
- Propagate SciFi seeds upstream of magnet and look for unused UT hits (hit flagging)
- Algorithm speed dependent on window size → potentially stronger limits on momentum range
- Probably the most realistic solution in terms of performance and throughput for early Run 3



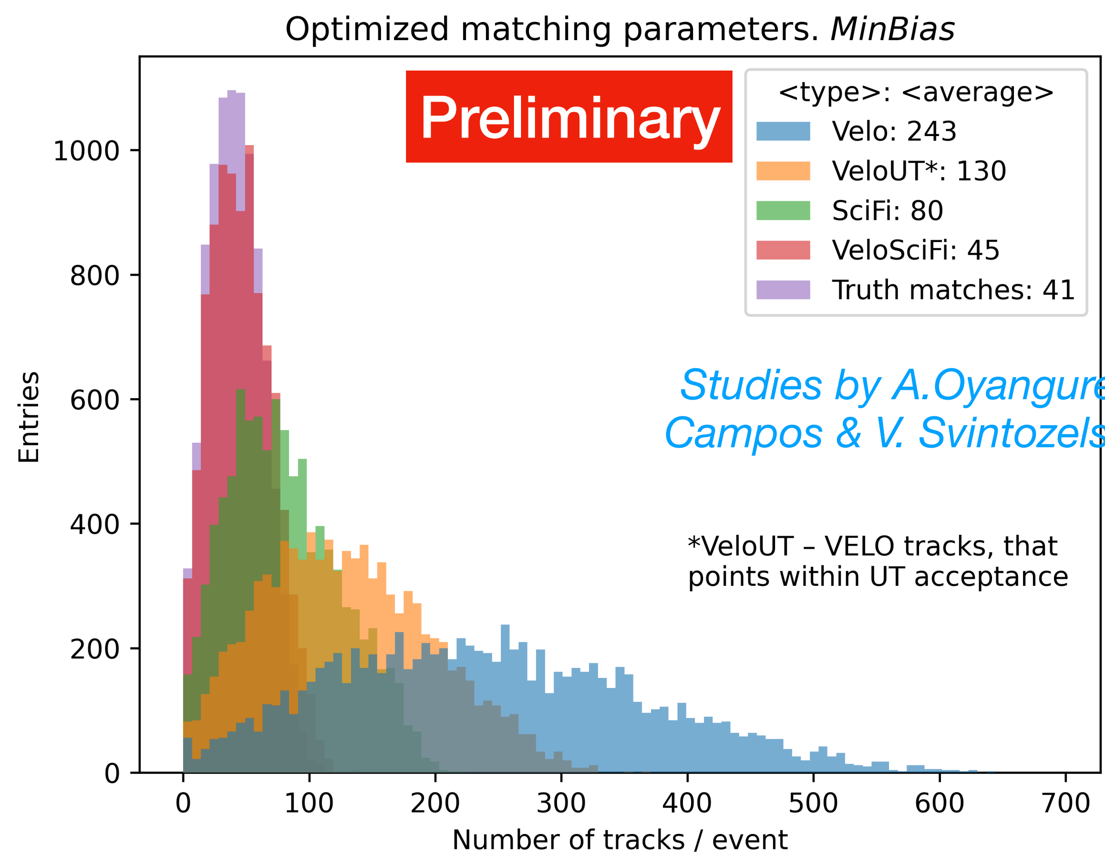
Both options under investigation

Adapting to ultra-LLPs

Can we trigger directly on the LLP decay after the UT?



- Same challenges as in HLT2: mom. resolution, fake rejection, extrapolation position…
- But also for HLT1: **output rate and throughput**
- After long reconstruction, O(50) un-matched SciFi seeds
  - **manageable combinatorics** (current long-based PV and SV algorithms < 5% throughput)
- Need to evaluate if output rate can be managed:
  - **High rejection of random combinations**
  - **+ strategy for different physics channels**



Studies by A.Oyanguren Campos & V. Svintozelskyi

Conclusions

- The new software-only trigger is a unique opportunity to increase LHCb's acceptance for Long Lived Particles

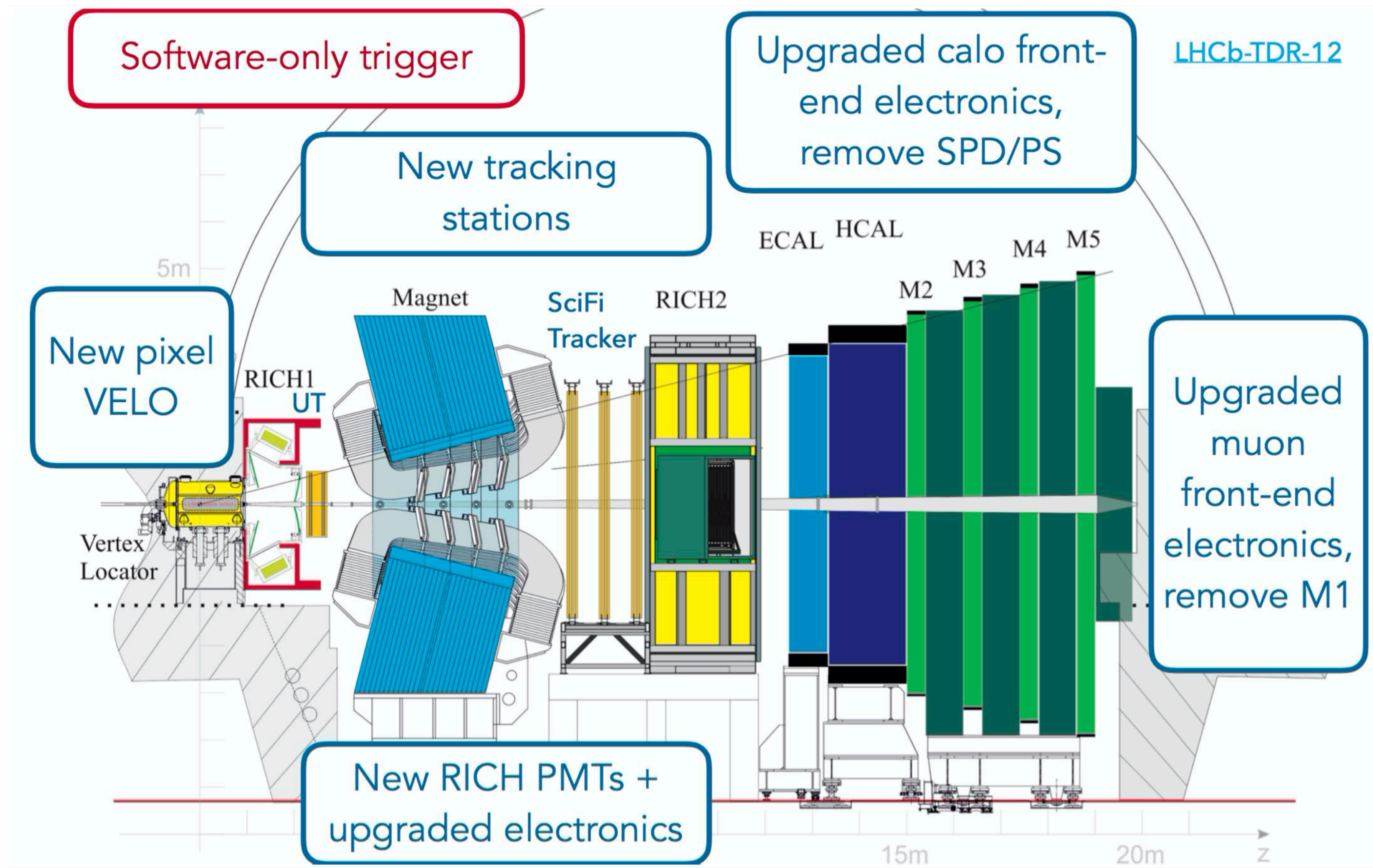
- For the second stage trigger (HLT2), offline-quality reconstruction is key:
  - Track and vertex reconstruction algorithms relying on displaced tracks advancing well

- For the first stage trigger (HLT1), the new GPU architecture is opening up reconstruction & trigger possibilities:
  - SciFi seeding recently implemented with compatible performance to the default (Forward tracking) - for the first time in HLT1
  - Exploring downstream tracking and triggering on T-tracks
  - Particular care must be given to throughput and output rate limitations

Stay tuned for more updates and thank you for your attention!

Backup

The LHCb U1 upgrade



Will start in 2022!

The LHCb detector at CERN:

- Single-arm forward spectrometer for high-precision flavour physics
- High precision tracking and vertexing
- Complemented with excellent PID

The U1 upgrade

- Instantaneous luminosity will increase by x5
- Major upgrade in all sub-detectors to handle increased rates
- Software-only trigger!

Are GPUs a good fit?

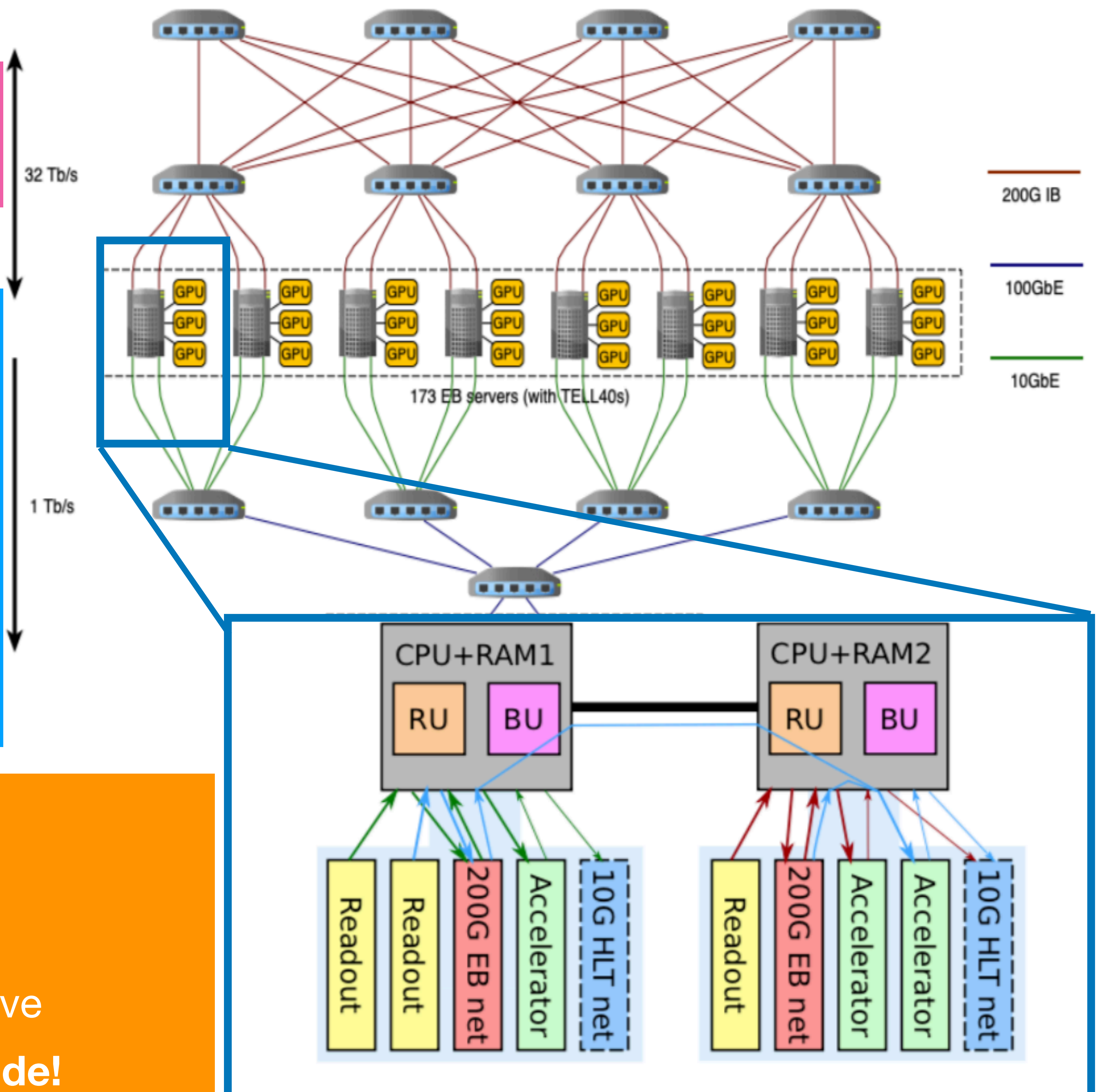
- Event builder farm equipped with 173 servers

- Each server has 3 free PCIe slots
  - Can be used to host GPUs
  - Sufficient cooling & power
  - Advantageous to have GPUs as self-contained processors
  - Sending data to GPU is like sending data to network card

- GPUs map well into LHCb DAQ architecture
- HLT1 tasks inherently parallelizable
- Smaller network between EB & CPU HLT
- Cheaper & more scalable than CPU alternative

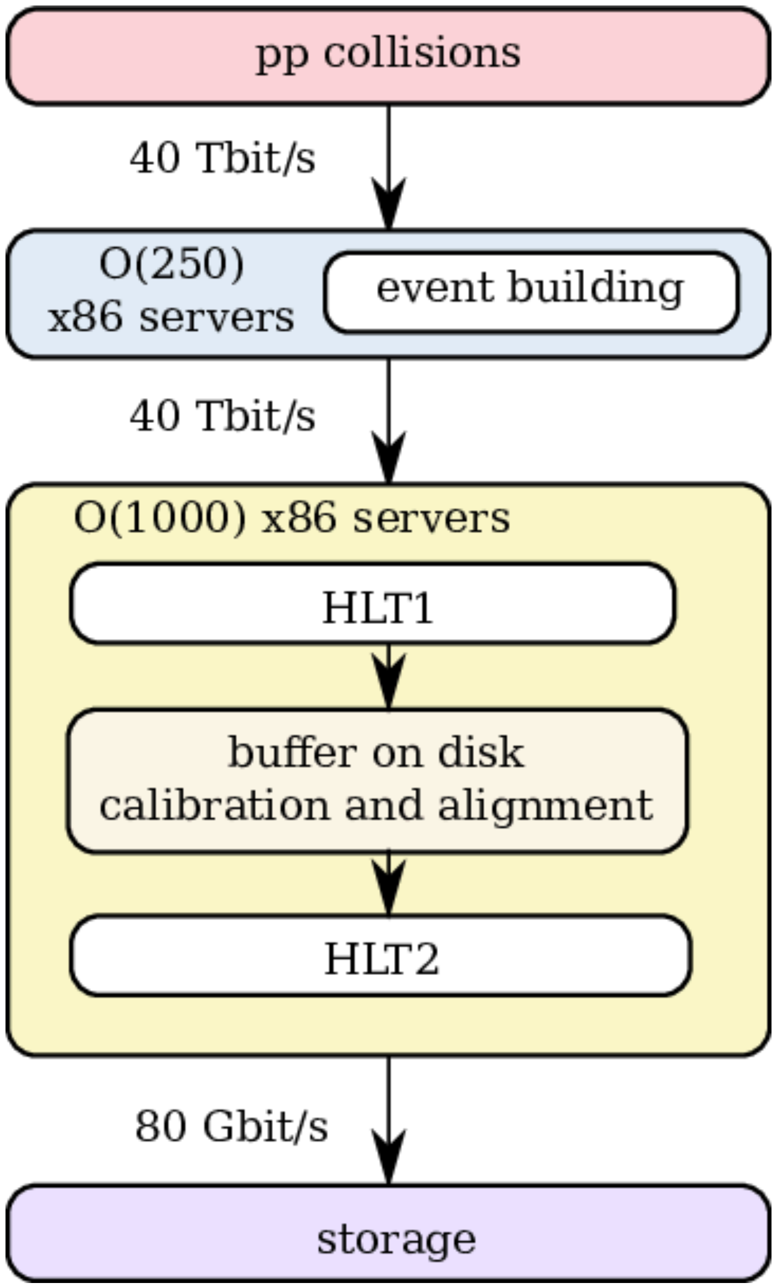
➡**Was chosen as the baseline for the upgrade!**

➡**Is implemented with O(200) Nvidia RTX A5000 GPUs**



GPU-equipped event builder PC, with traffic of all three readout cards.

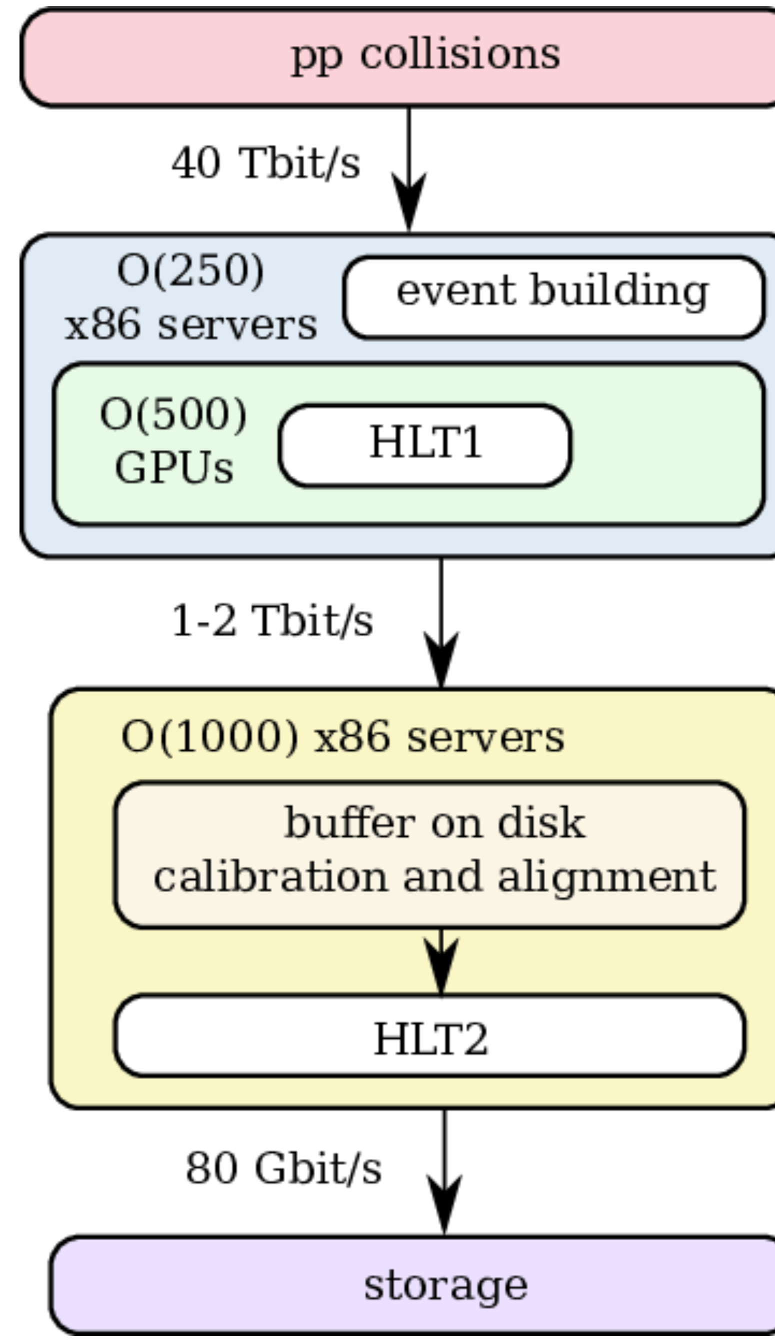
Architecture upgrade options



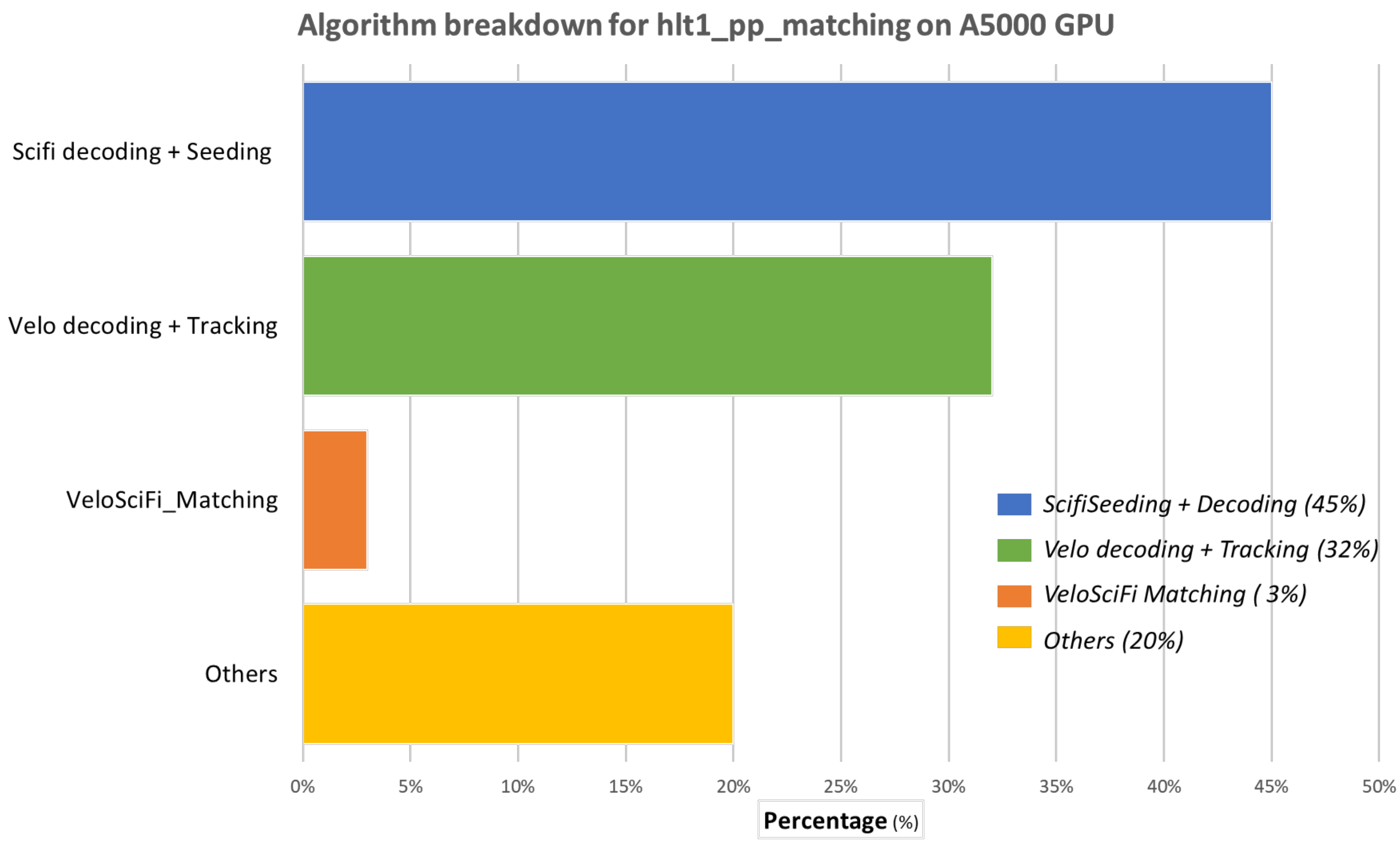
Detector data received by O(500) FPGAs and built into events in the EB servers

Two options:

1. Send full 40 Tb/s to a CPU processing server → extra network needed
2. Fill extra EB slots with GPUs → reduce rate locally to 1 Tb/s before full processing



Seeding&matching breakdown



Allen: a GPU HLT1 trigger platform

- Public software project: gitlab repo
- Supports three modes:
  - Standalone
  - Compiling within the LHCb framework for data acquisition
  - Compiling within the LHCb framework for simulation and offline studies
- Runs on CPU, Nvidia GPU (CUDA, CUDACLANG), AMD GPUs (HIP)
- GPU code written in CUDA
- Cross-architecture compatibility (HIP, CPU) via macros

Allen

pipeline passed

Welcome to Allen, a project providing a full HLT1 realization on GPU.

Documentation can be found here.

Mattermost discussion channels

- Allen developers - Channel for any Allen algorithm development discussion.
- Allen core - Discussion of Allen core features.
- AllenPR throughput - Throughput reports from nightlies and MRs.

Performance monitoring

- Allen throughput evolution over time in grafana
- Allen dashboard with physics performance over time

Documentation

Edit on GitLab

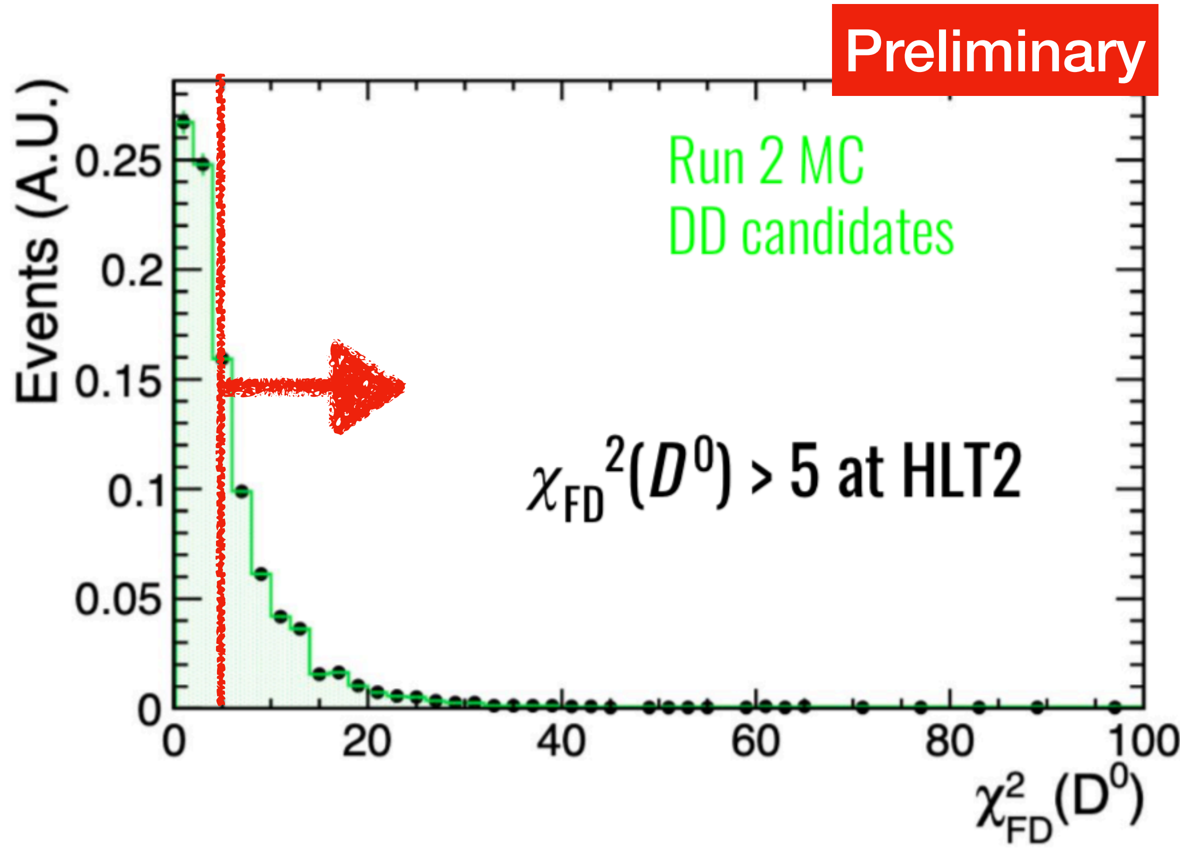
Welcome to Allen's documentation!

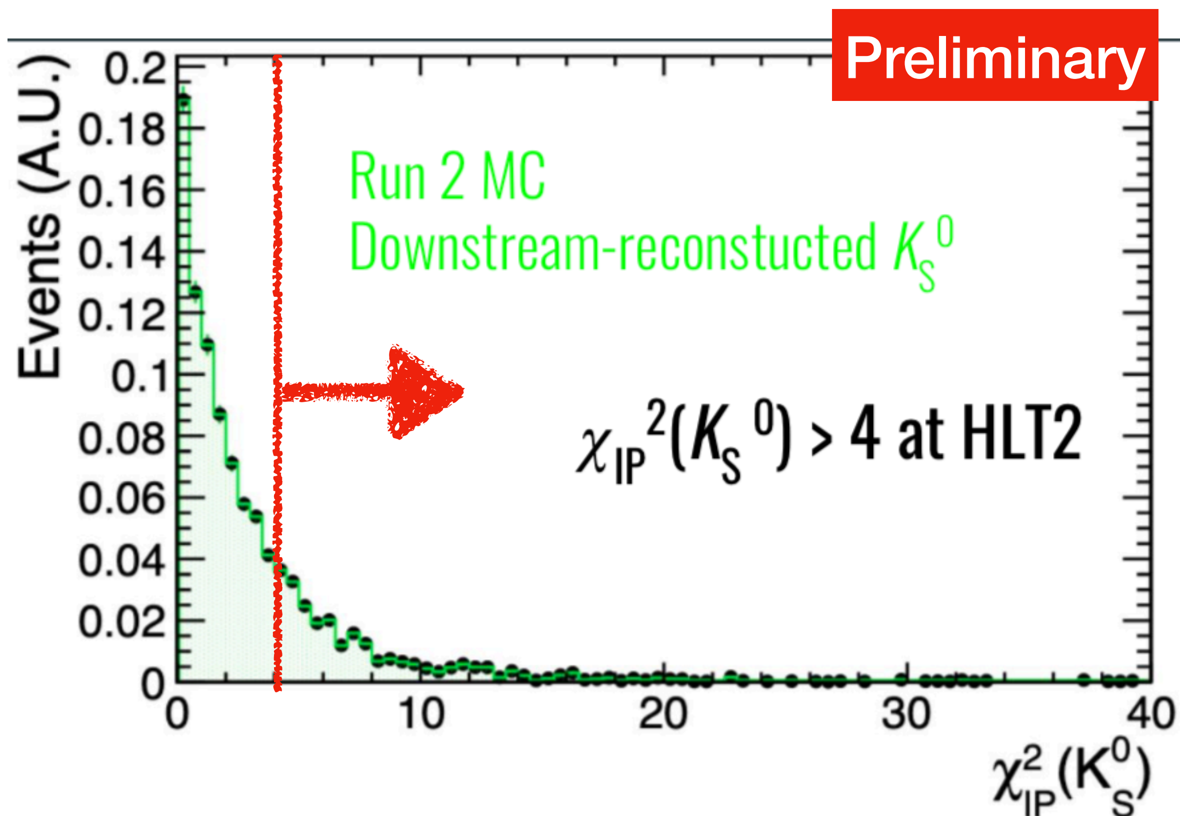
Allen is the LHCb high-level trigger 1 (HLT1) application on graphics processing units (GPUs). It is responsible for filtering an input rate of 30 million collisions per second down to an output rate of around 1-2 MHz. It does this by performing fast track reconstruction and selecting pp collision events based on one- and two-track objects entirely on GPUs.

This site documents various aspects of Allen.

Run 1/2 limitations

- Run1/2 L0 triggers only on calo clusters and muons + HLT1 triggered only on **Long Tracks** -> Sub-optimal for LLP channels
- Mitigation possible when part of the decay contained **Long Tracks** and/or leptons:
  - Ex $\Lambda_b \to \Lambda J/\psi(\to \mu\mu)$ triggering on the muons
  - Not possible for channels such as $D^0 \to K_S^0 K_S^0$ and $\Lambda_b \to \Lambda\gamma$
- $D^0 \to K_S^0 K_S^0$ Run 2 studies by L. Pica, G. Punzi, G. Tuci:
  - Extremely interesting for CPV in charm
  - Run 2 analysis performed in **LL**, **LD** and **DD**
  - Simulation: **~86%** of decays with at least one **Downstream** K_S^0,
  - but only **25%** in data
  - HLT1 only triggers on **Long Tracks** + HLT2 selections are inefficient for **Downstream**

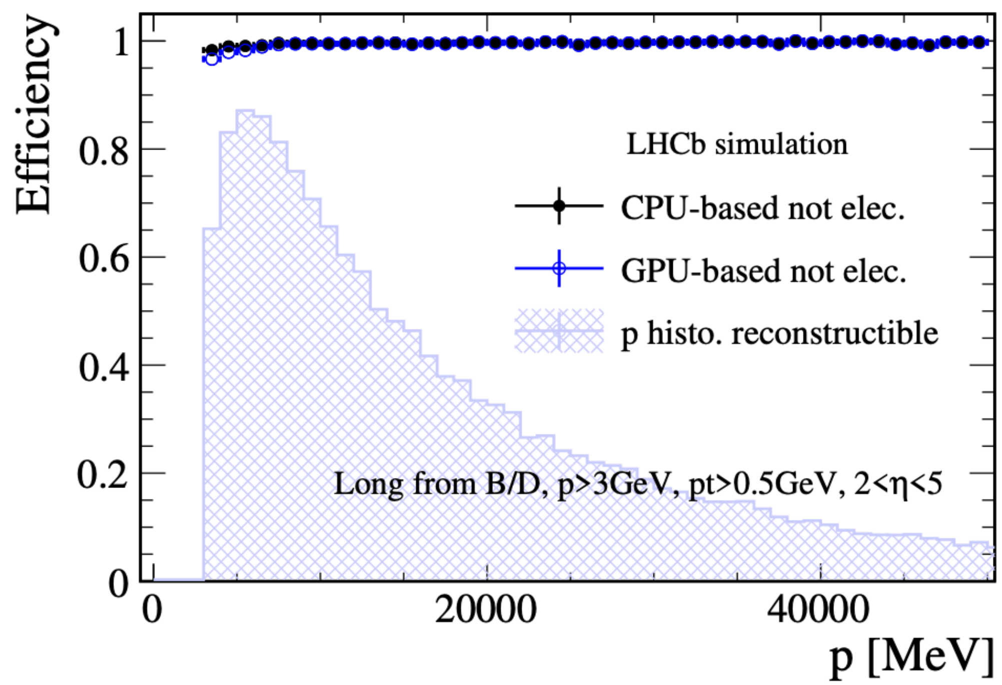


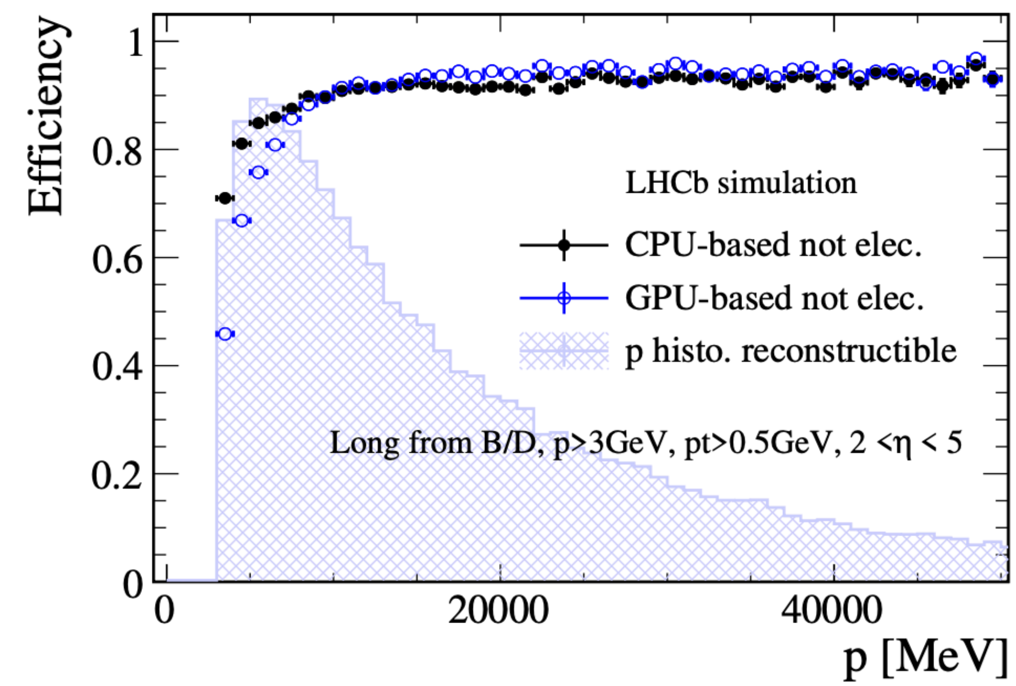


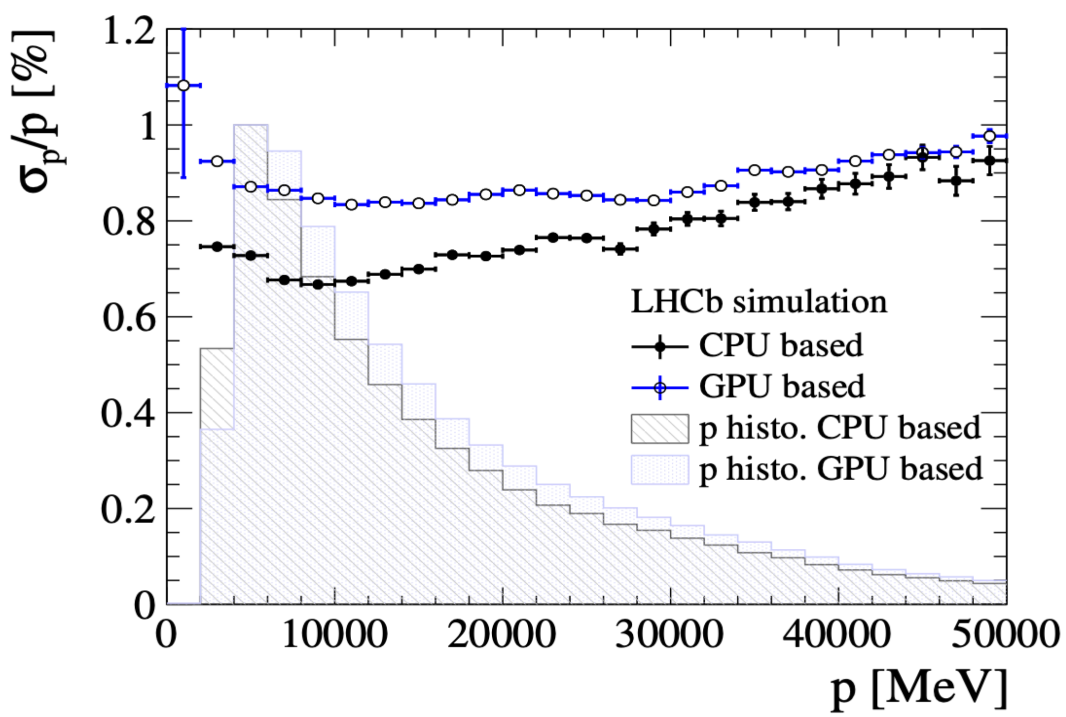
A lot of room for improvement on the trigger side!

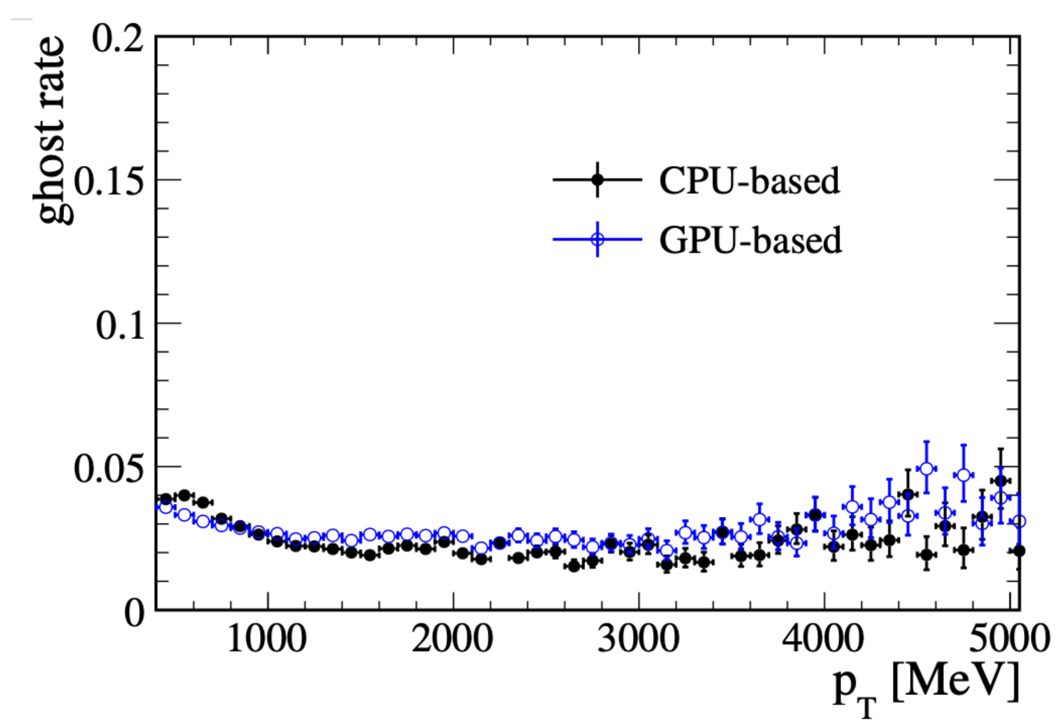
HLT1 CPU/GPU tracking performance

Compatible performance between CPU and GPU!





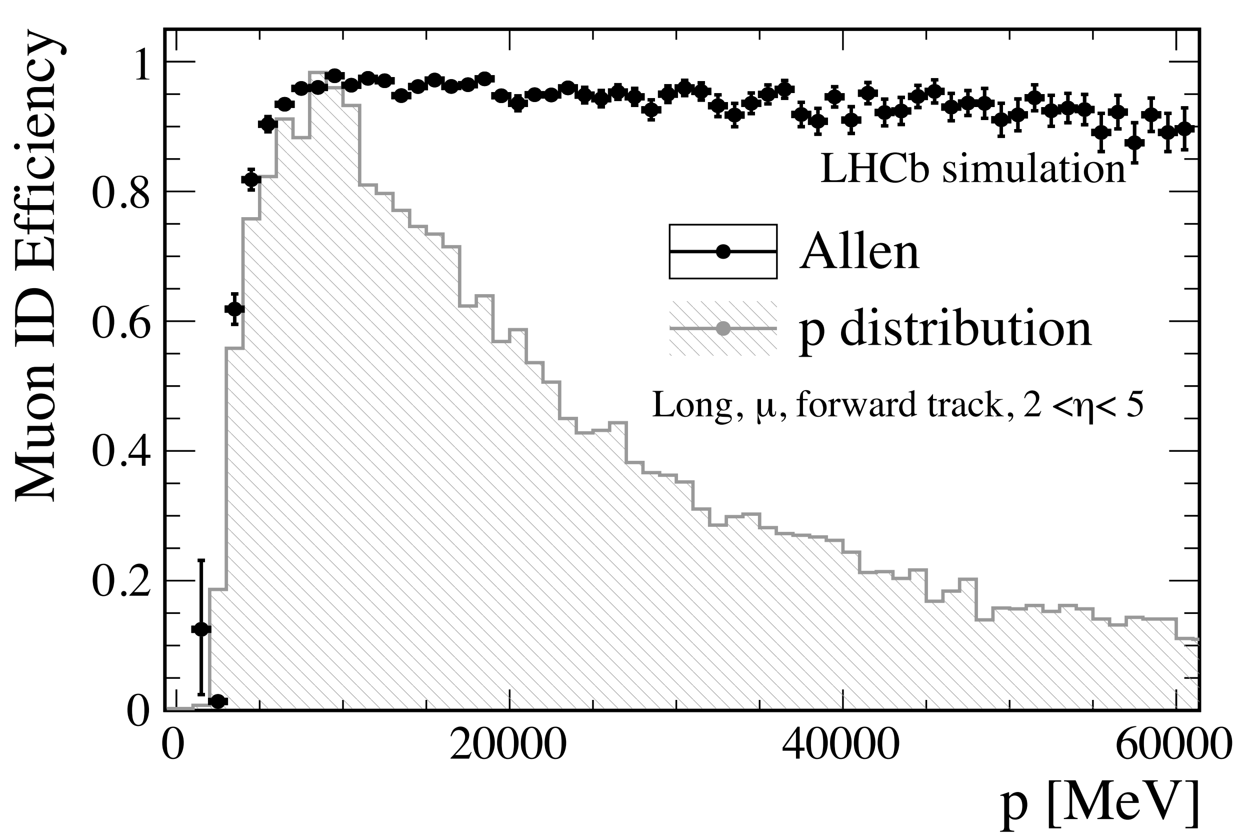


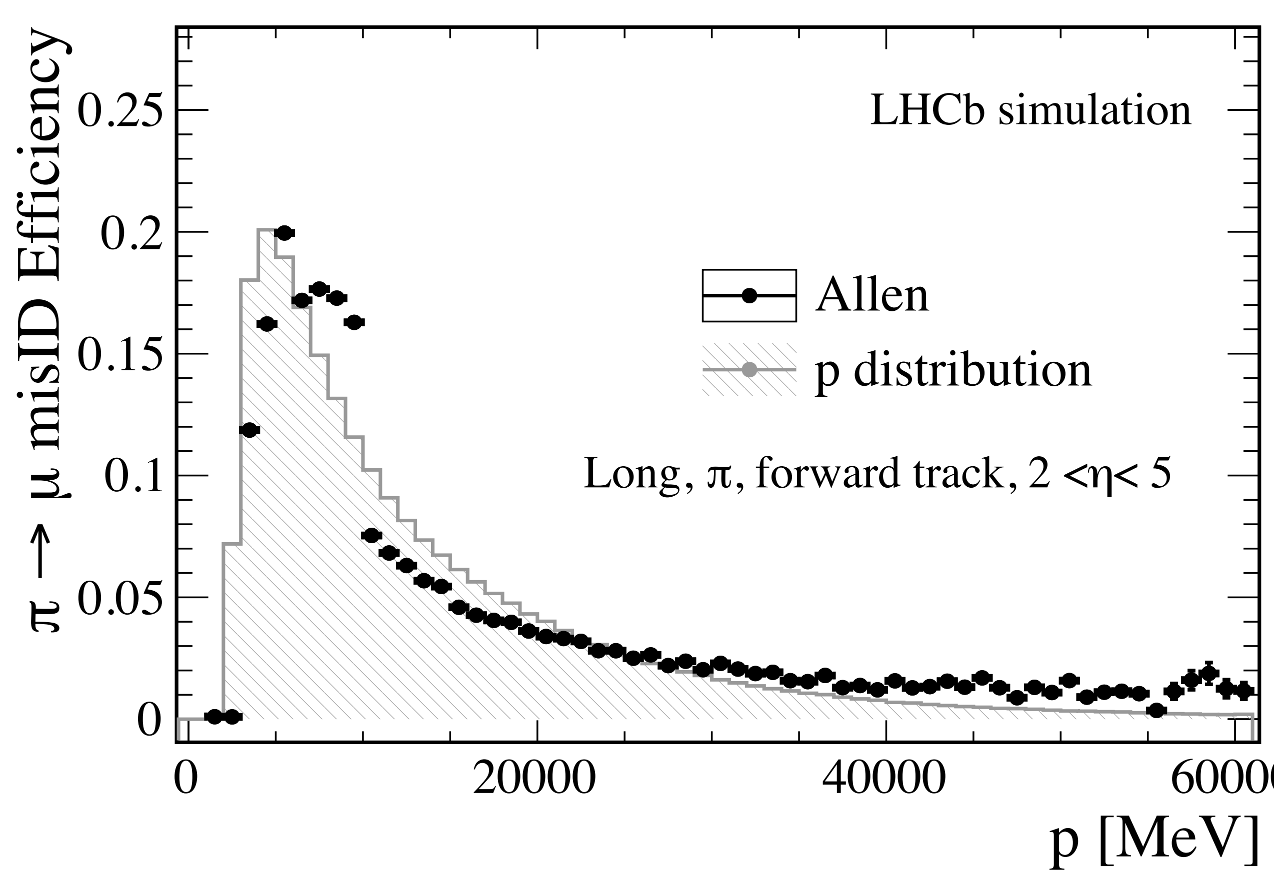


Comput. Softw. Big Sci. 6 (2022) no.1, 1

HLT1 muonID performance

- Excellent muon identification and misID background rejection



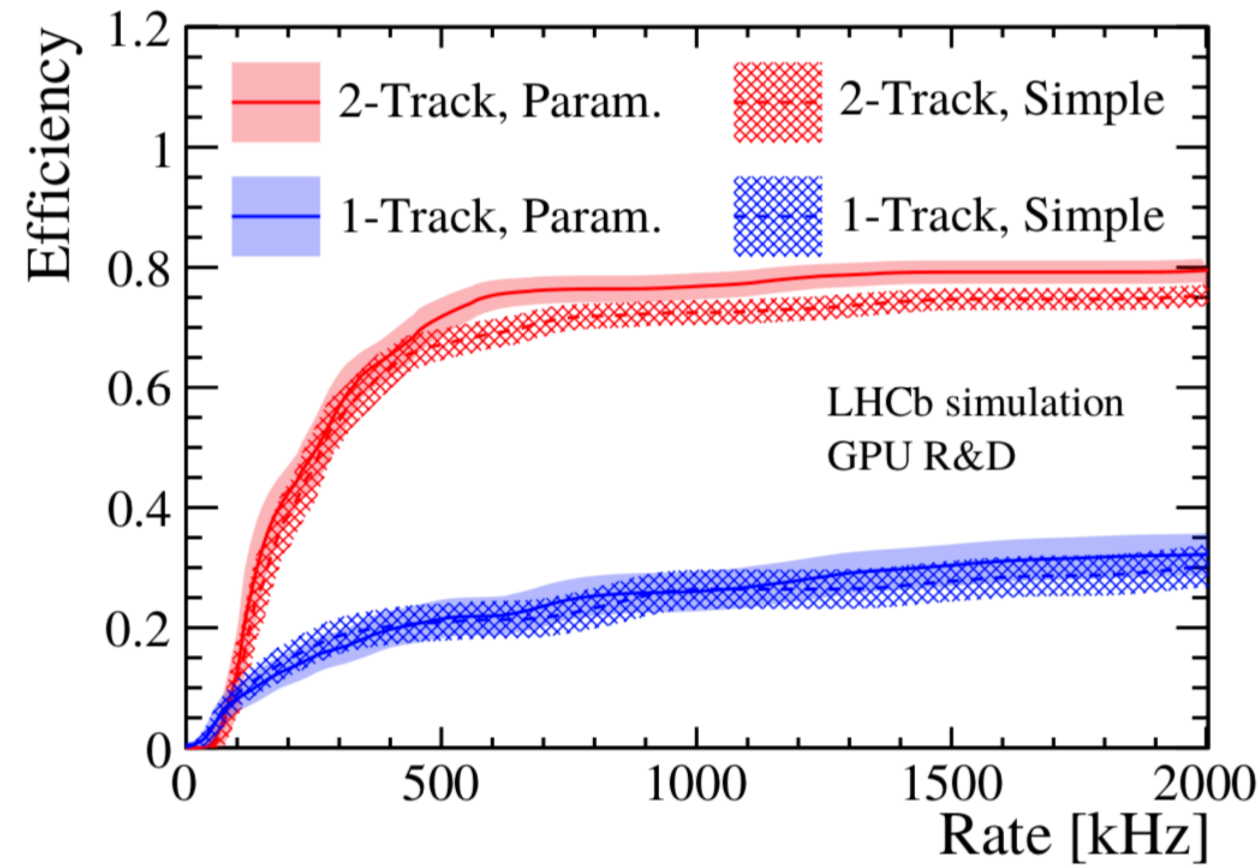


 LHCb-FIGURE-2020-014

Kalman filter

- Improve Impact Parameter (IP) resolution and reduce ghosts
- Nominal LHCb Kalman filter uses Runge - Kutta extrapolator + detailed detector description

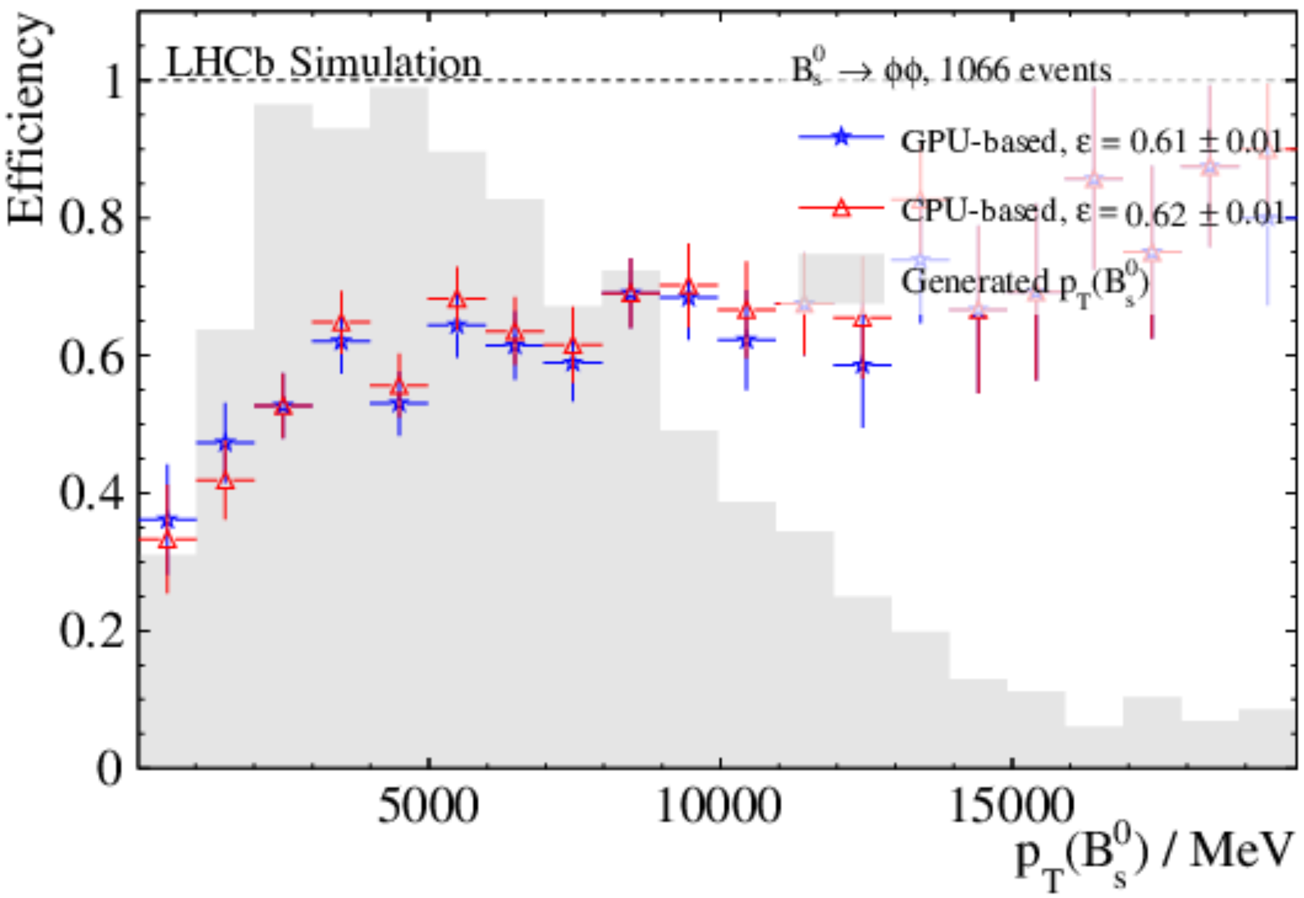
- In HLT1, for performance reason two alternatives based on parametrizations:
    - Full detector Parametrized Kalman Filter
    - Velo-Only Kalman Filter (fits only Velo segment, momentum estimate from full track)
    - IP resolution mostly impacted by Velo measurement -> Velo-Only option chosen, which significantly improves throughput



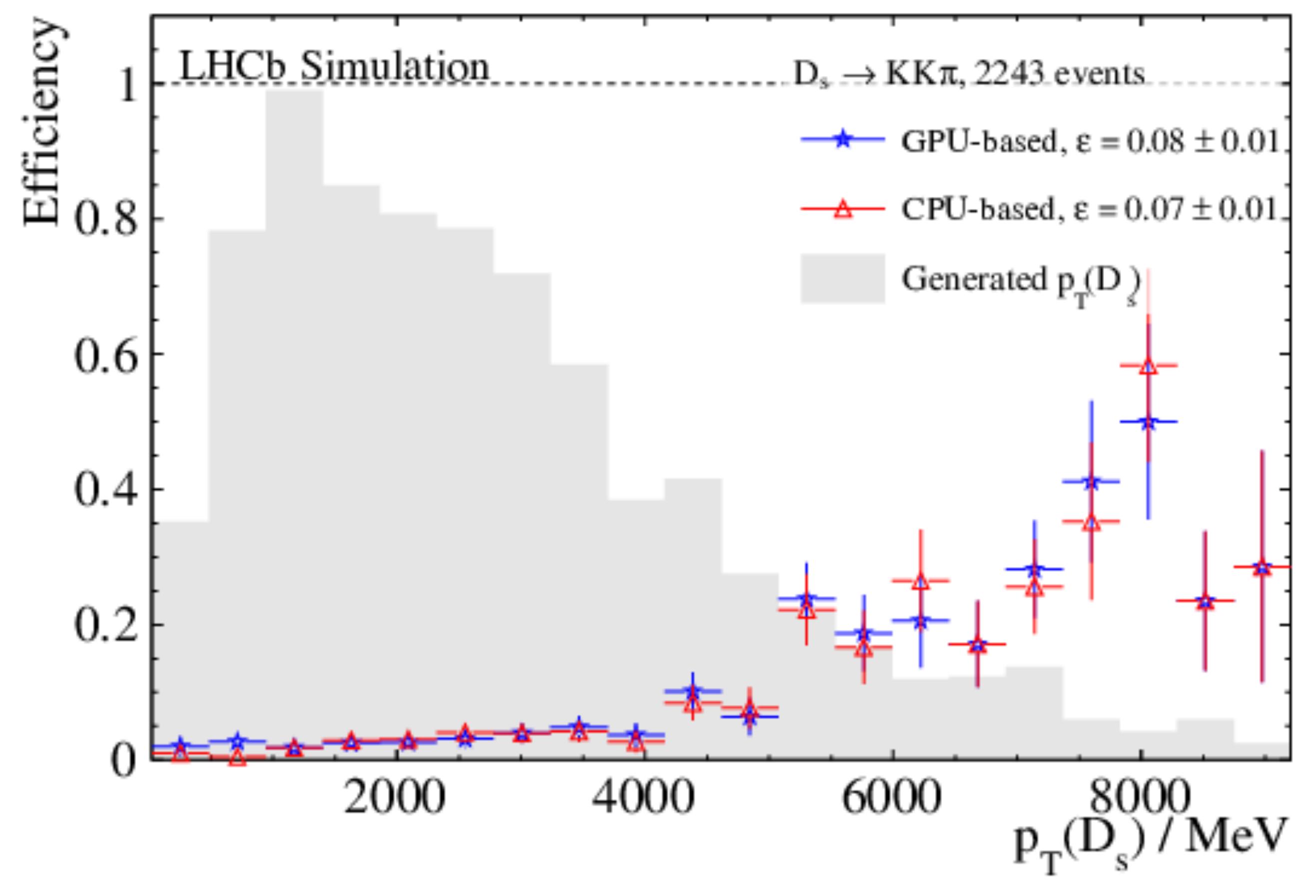
HLT1 selection performance

- Inclusive rate for the main HLT1 lines ~ 1 MHz
- Compatible performance between CPU and GPU

TwoTrackMVA line for $B_s^0 \to \phi\phi$



TrackMVA line for $D_s \to KK\pi$



[2105.04031]