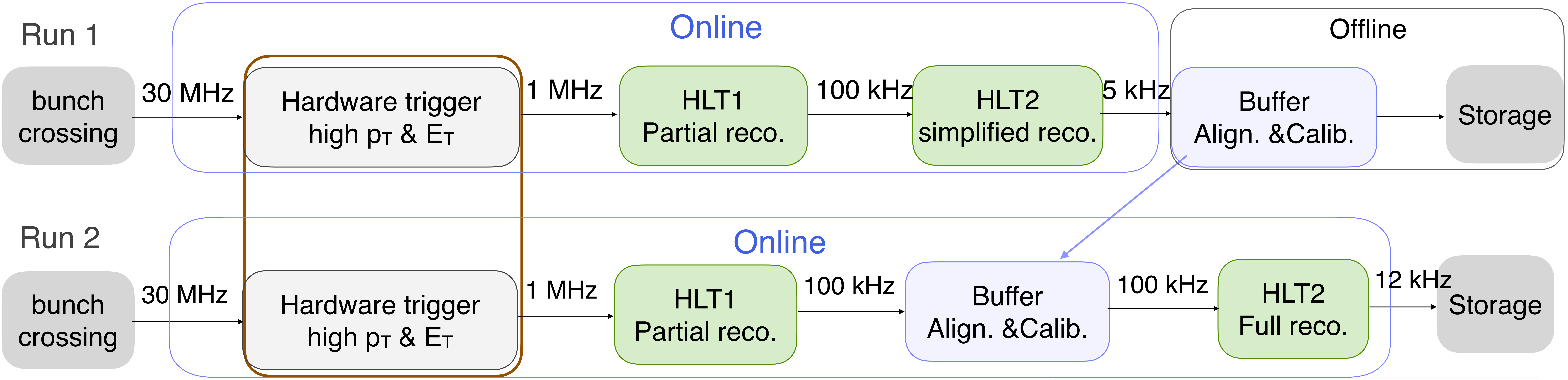
RTA for the LHCb Upgrade II

UII-Tracking Workshop

Peilian Li
(on behalf of RTA)

06.03.2024, Évian-les-Bains

Trigger at Runs 1 & 2


Run 1

Online

Offline

bunch crossing → 30 MHz → Hardware trigger high p_T & E_T → 1 MHz → HLT1 Partial reco. → 100 kHz → HLT2 simplified reco. → 5 kHz → Buffer Align. &Calib. → Storage

Run 2

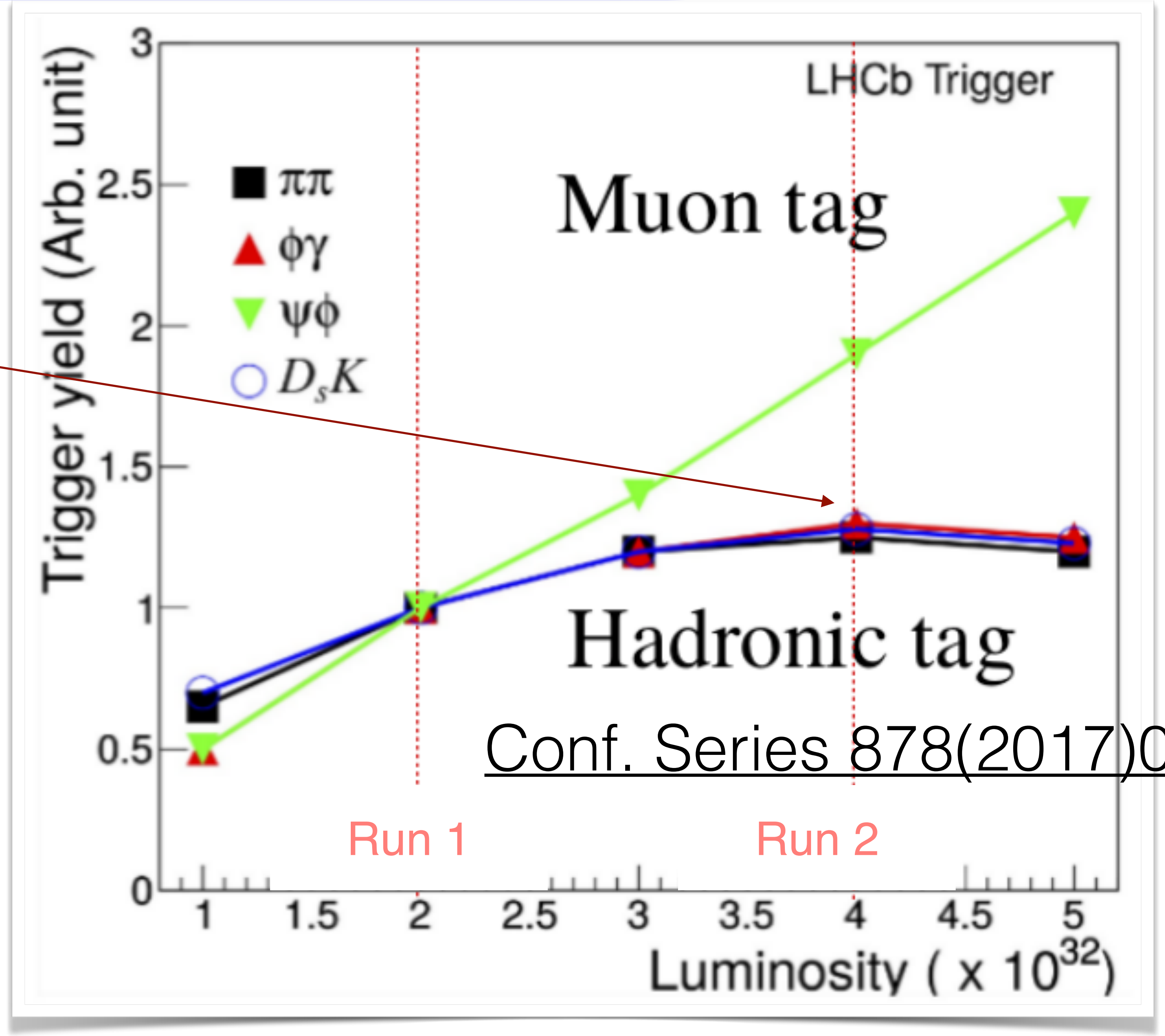
Online

bunch crossing → 30 MHz → Hardware trigger high p_T & E_T → 1 MHz → HLT1 Partial reco. → 100 kHz → Buffer Align. &Calib. → 100 kHz → HLT2 Full reco. → 12 kHz → Storage

Hardware trigger: 30→1 MHz read-out limits

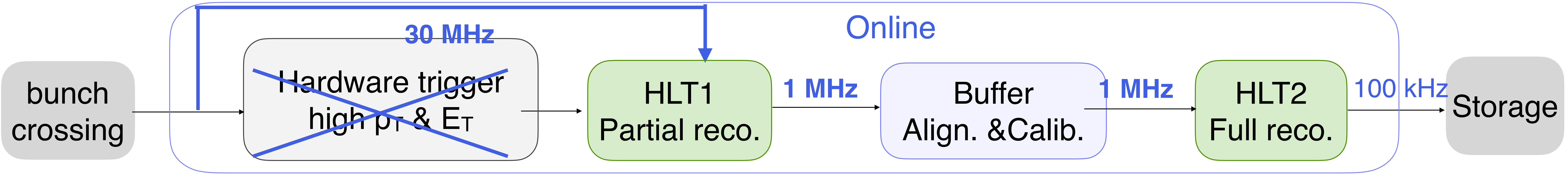
→ based on muon detector and calorimeters

- Hardware trigger is not an option for Run 3, as rate limit of 1 MHz saturates fully hadronic modes



Conf. Series 878(2017)012012

Trigger at Run 3



bunch crossing

Online

30 MHz

~~Hardware trigger high p_T & E_T~~

HLT1 Partial reco.

1 MHz

Buffer Align. &Calib.

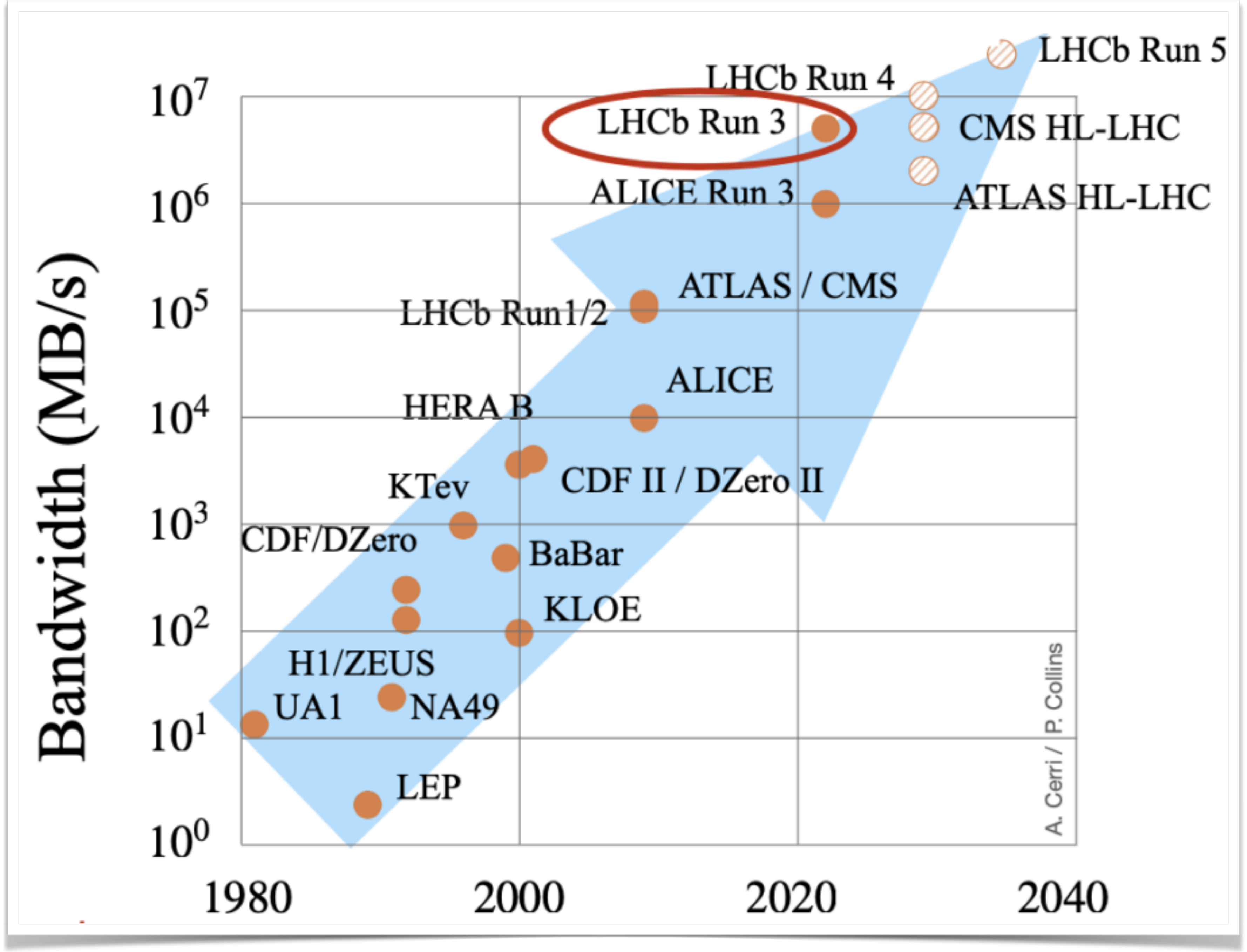
1 MHz

HLT2 Full reco.

100 kHz

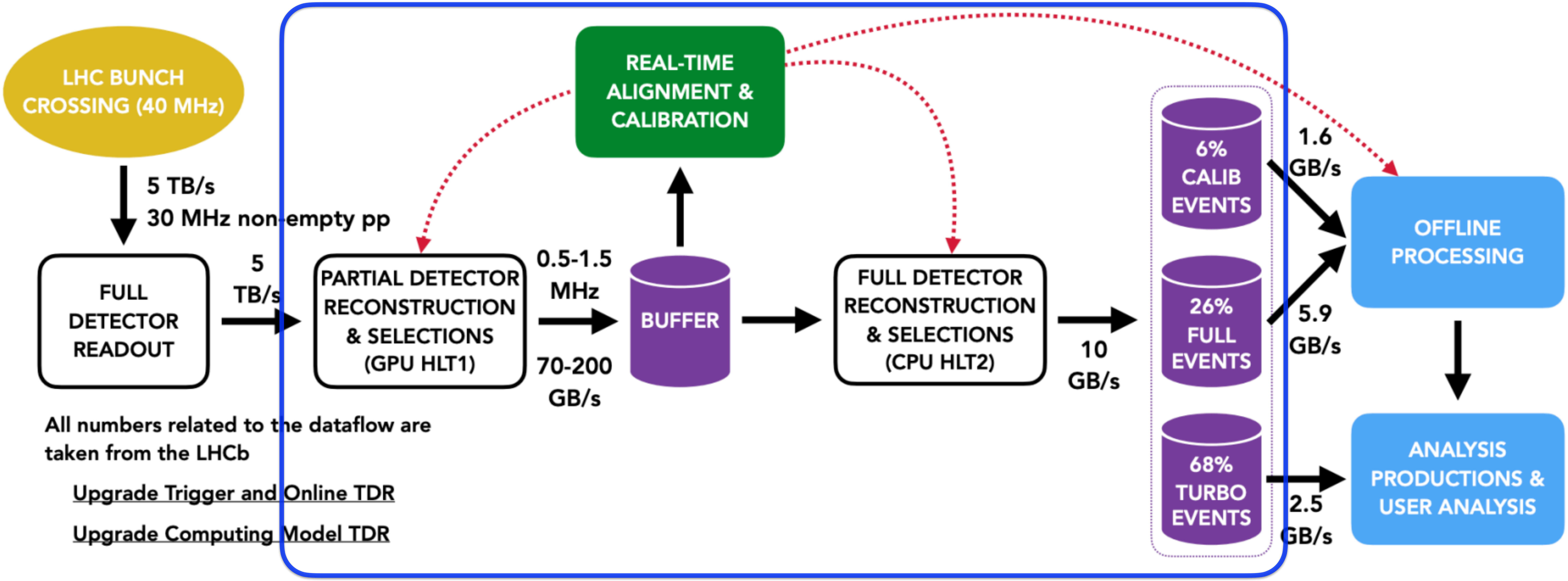
Storage

- Remove hardware trigger, fully software trigger
- Read out the full detector at 30 MHz in HLT1
- Real time alignment and calibration
- 10x higher data rate than Run 2 but with 3x larger disk buffer only
- Full offline-quality reconstruction in "real-time"
- Increase of hadronic trigger efficiency by 2~4x w.r.t. Run 2



Highest data processing rate of any HEP experiment!

Run 3 Data flow


LHC BUNCH CROSSING (40 MHz)
5 TB/s
30 MHz non-empty pp
FULL DETECTOR READOUT
5 TB/s
PARTIAL DETECTOR RECONSTRUCTION & SELECTIONS (GPU HLT1)
0.5-1.5 MHz
70-200 GB/s
BUFFER
REAL-TIME ALIGNMENT & CALIBRATION
FULL DETECTOR RECONSTRUCTION & SELECTIONS (CPU HLT2)
10 GB/s
6% CALIB EVENTS
26% FULL EVENTS
68% TURBO EVENTS
1.6 GB/s
5.9 GB/s
2.5 GB/s
OFFLINE PROCESSING
ANALYSIS PRODUCTIONS & USER ANALYSIS

All numbers related to the dataflow are taken from the LHCb

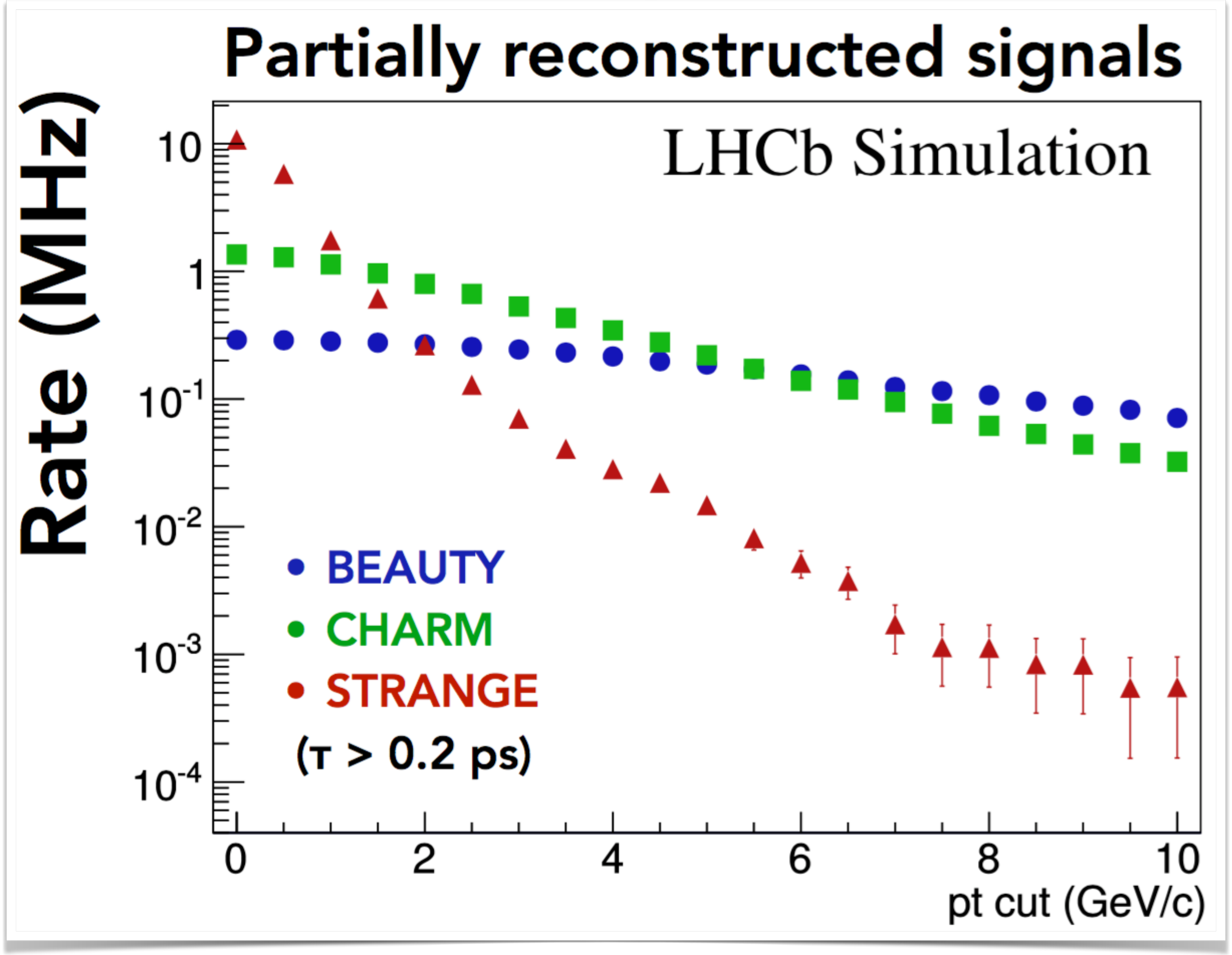
Upgrade Trigger and Online TDR

Upgrade Computing Model TDR

- From single HLT framework → two separate ones:
  - GPU (Allen) for HLT1, can be run from Gaudi too
  - CPU (Gaudi/Moore) for HLT2

What will it be for UII?

- 5x - 7.5x higher luminosity
- ~40 primary vertices
- “triggerable” decay in every event
- linear increase of **output rate** with luminosity
- **Larger event size** (more PVs + timing info)
- HLT2 computing and storage needs **scale quadratically**



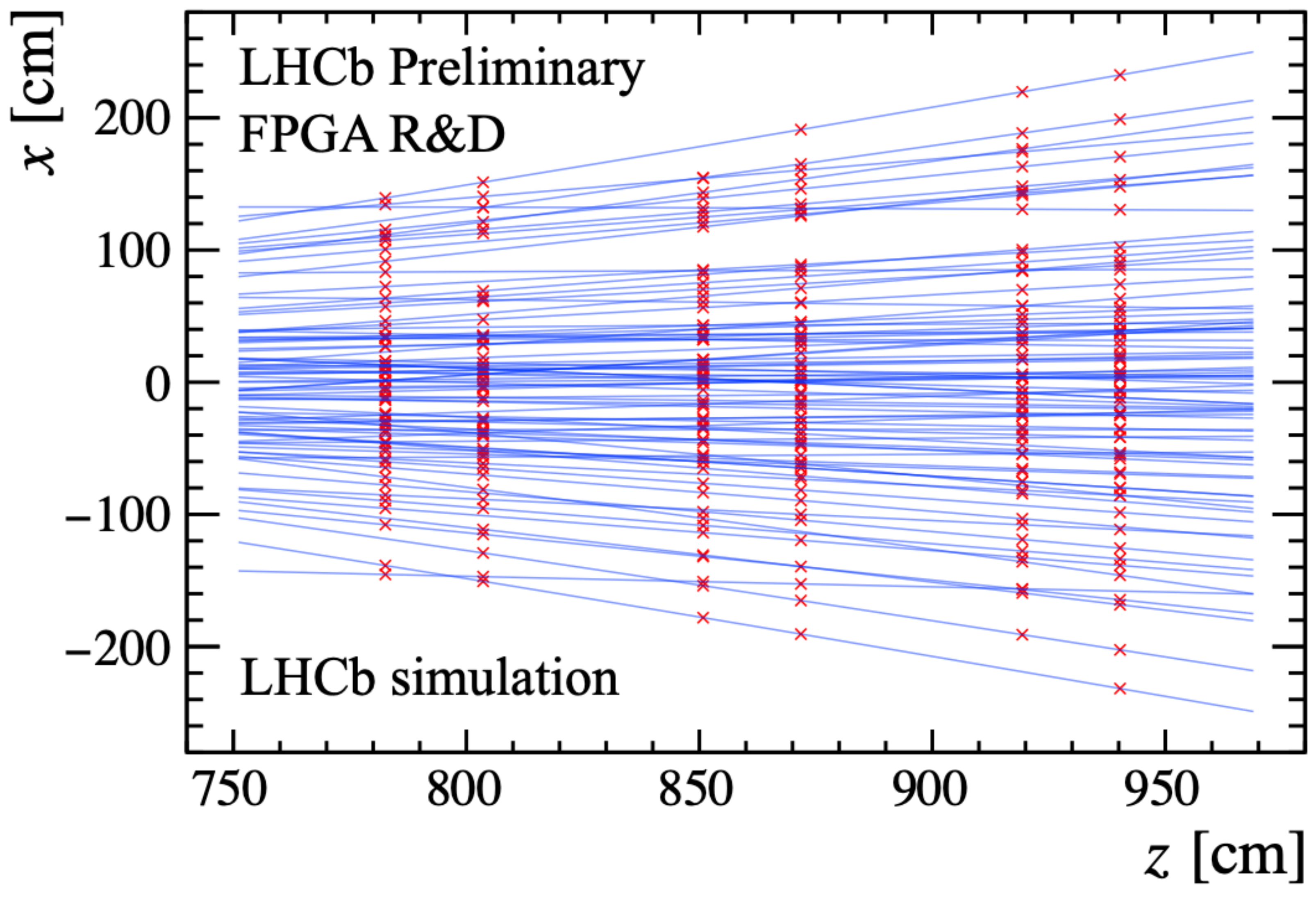
Baseline from FTDR

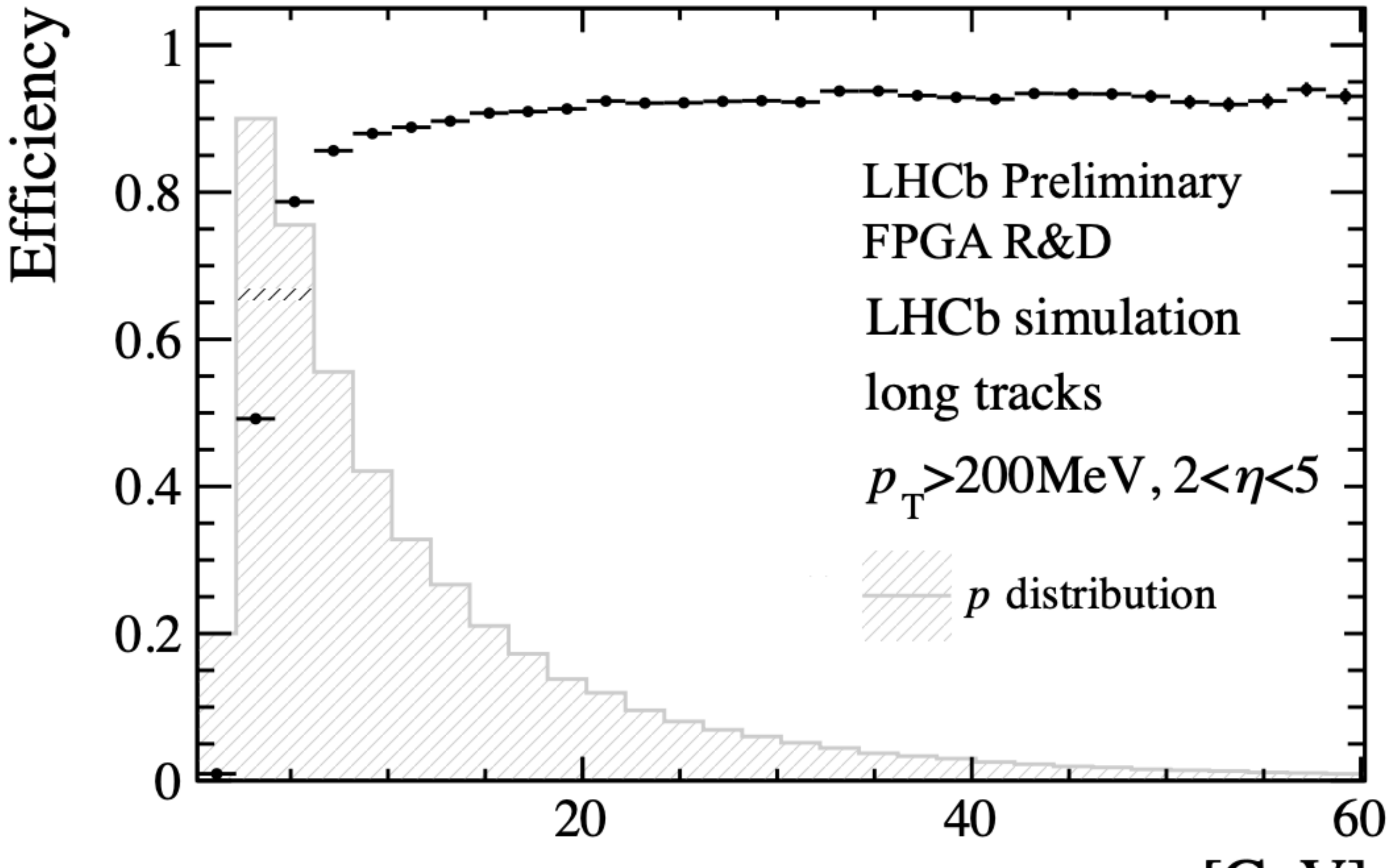
- Signal dominated, two-step trigger as in Run 3, both HLT1 and HLT2 reconstruction with GPUs
  - Luminosity scale factor of 7.5 w.r.t Run 3, HLT1 output rate scaling factor ~7.5
  - Event size scaling factor of 4: estimate based on the scaling provided by the sub-detectors
  - HLT1 performs **partial event reconstruction & inclusive selection at 30 MHz**
  - HLT2 performs **full reconstruction and inclusive + exclusive** selections

- Major changes / questions:
  - Pileup mitigation using timing information
  - Explore more exclusive selection & partial persistency
  - HLT2 ported to GPUs to handle increased complexity in limited resources
  - When/where to apply Calibration & alignment?

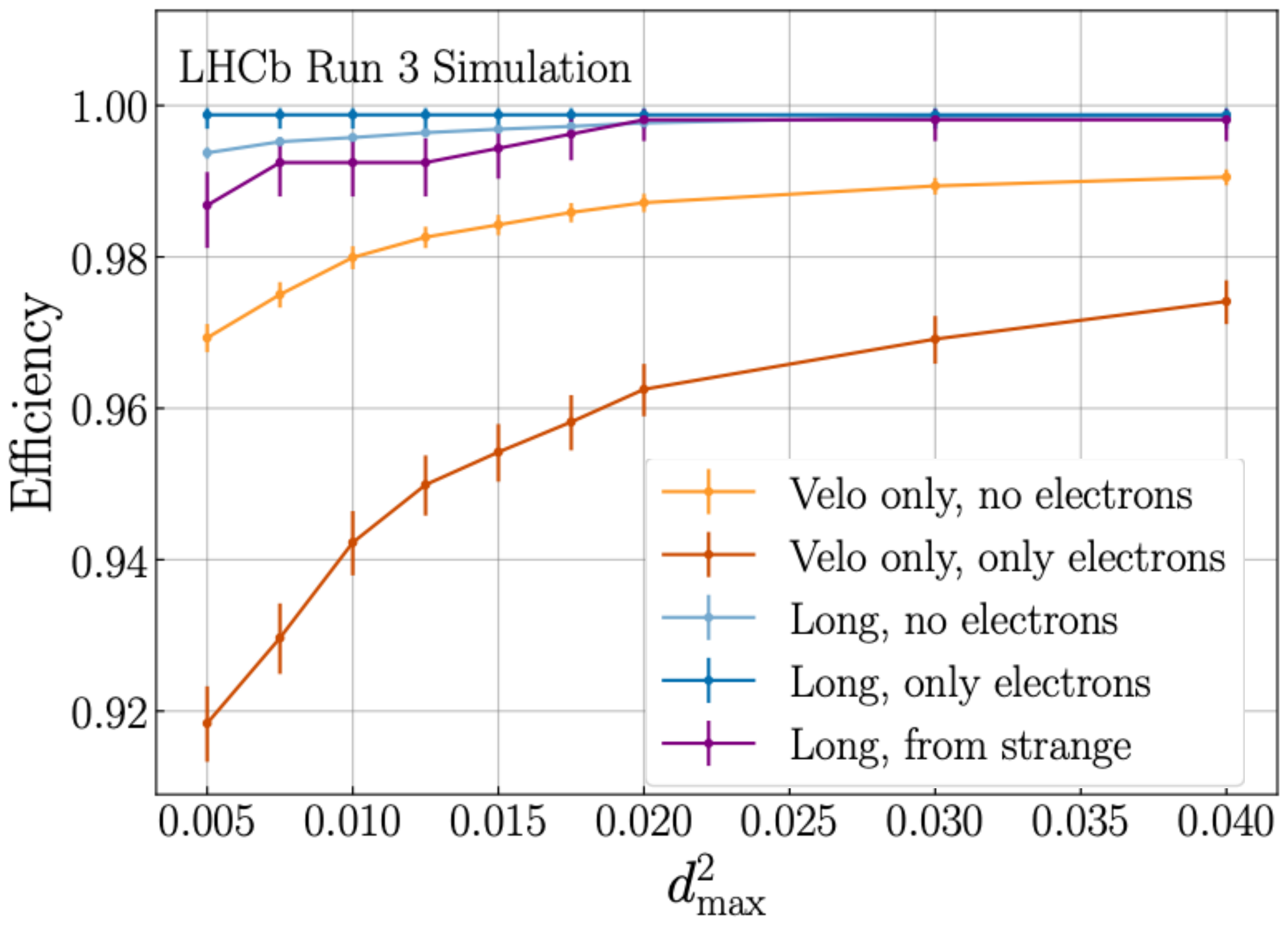
Architectures

- GPUs for both HLT1 & HLT2 reconstruction: driven by the current cost of hardware
  - Professional GPUs used in Run 3
- Alternative R&Ds: Clustering & Reconstruction in FPGAs → G. Punzi's talk tomorrow
  - Retina clustering applied in Run 3 successfully
  - **SciFi seed tracks in FPGAs proposal for Run 4** (approved by LHCb and in review by LHCC) will provide **excellent demonstrator**
- Testbed for other accelerators: IPUs, Machine learning applications (etx4velo)





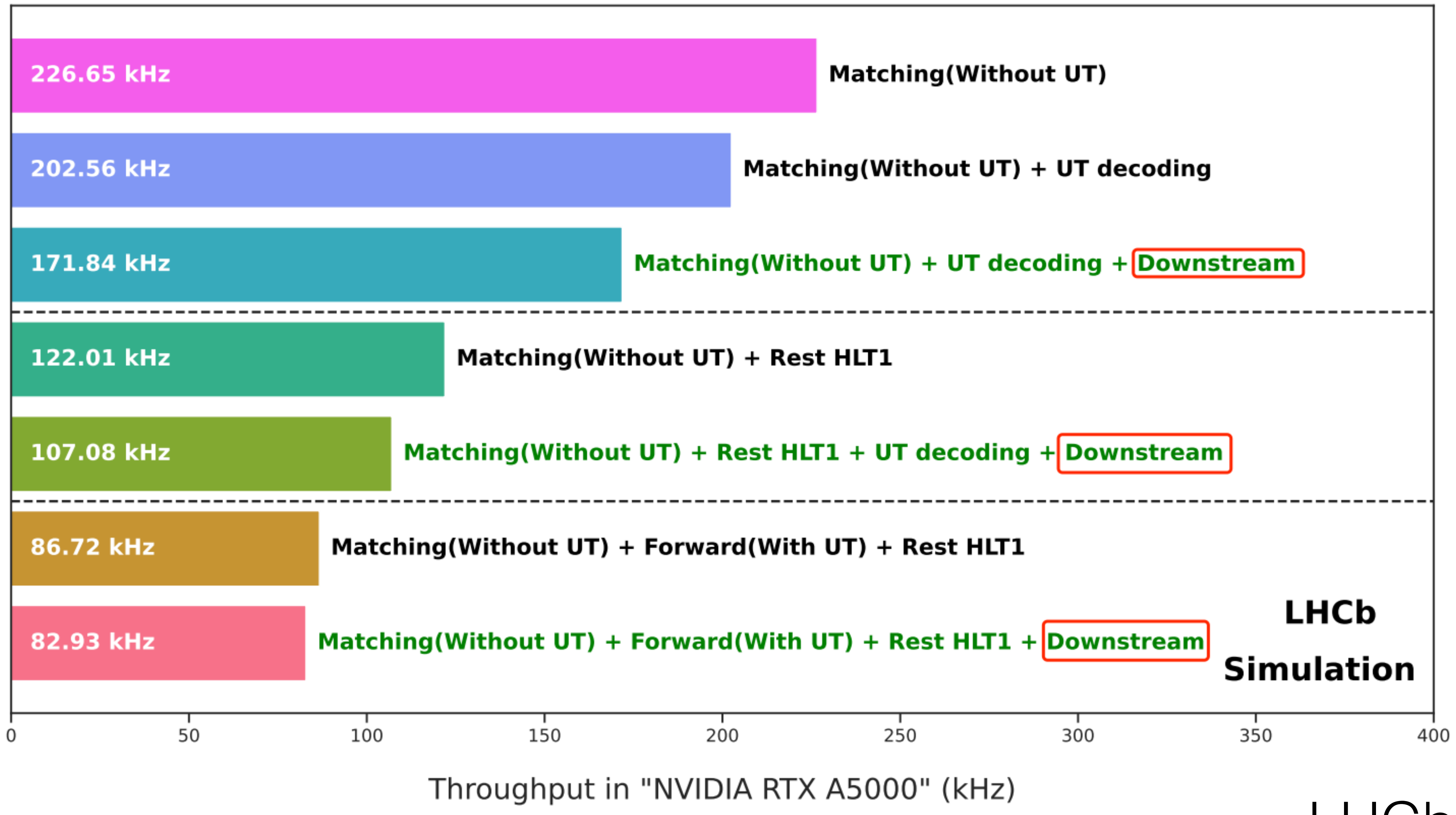
LHCb-FIGURE-2023-027

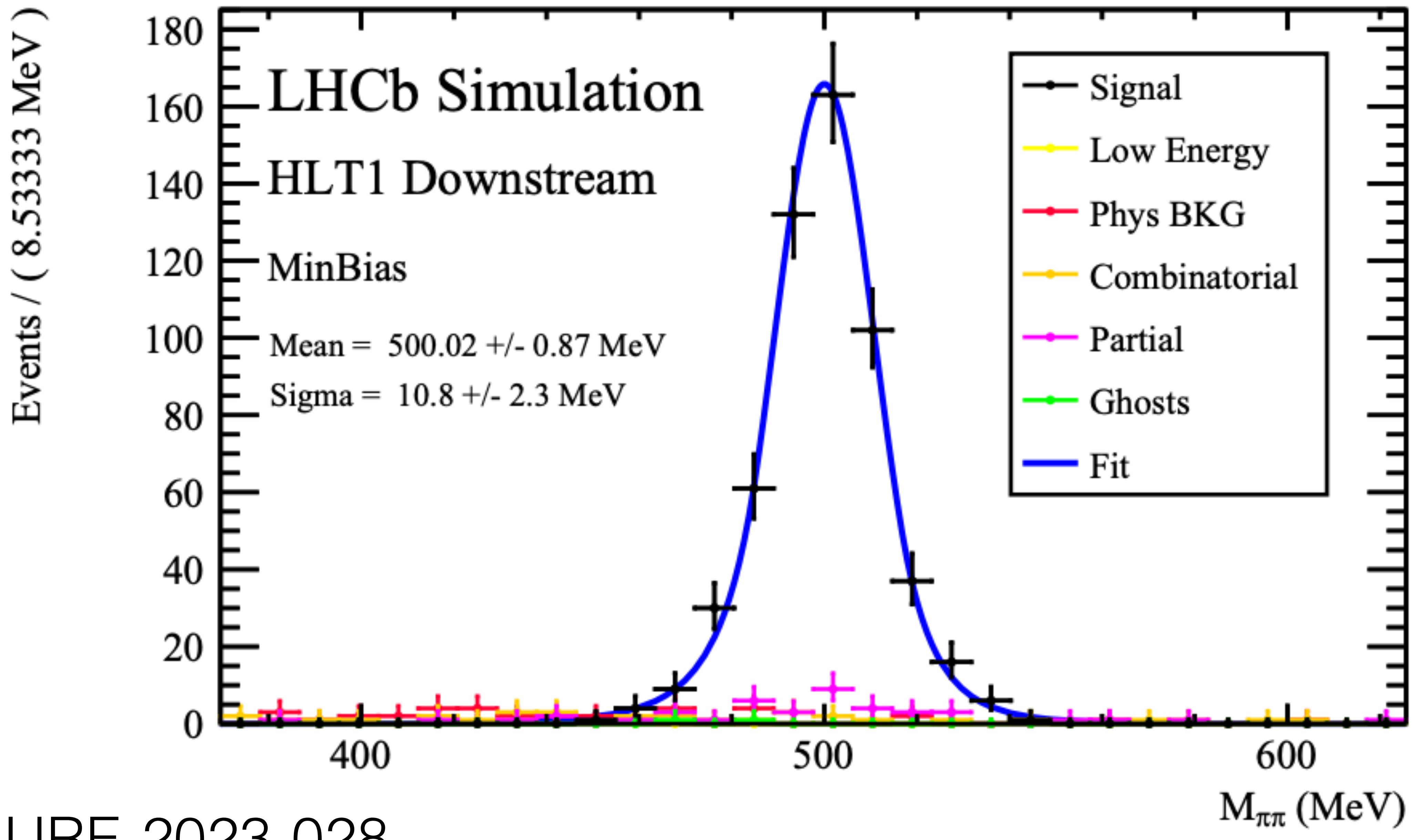


LHCb-FIGURE-2023-024

Lessons learnt from Run 3 HLT1

- Allen is our **first heterogeneous platform**
  - Support cross-architecture (CPU/GPU) programming, multi-event scheduling…
- Lots of **additional achievements than planned** as good demonstrators for UII
  - Two Long tracking methods, **Downstream tracking, ECAL reconstruction**
  - **RICH decoding** ready, reconstruction in progress
  - More **exclusive selections**, luminosity, monitoring and more





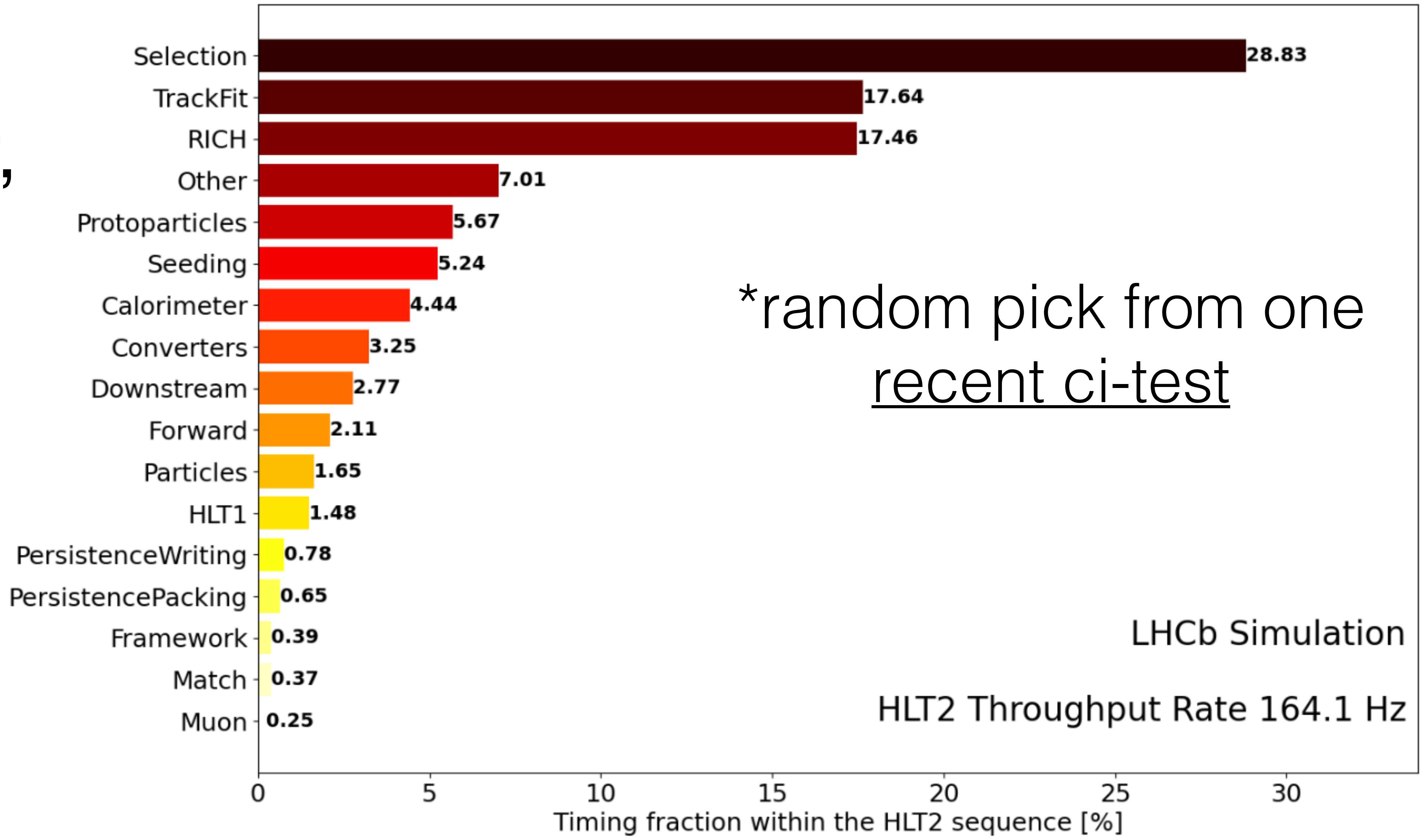
LHCb-FIGURE-2023-028

Lessons learnt from Run 3

- Run 3 experience shows us **possibility to migrate HLT2 reconstruction to GPU**
  - **Consistent reconstruction** between HLT1 & HLT2 → as in Run 2 which was beneficial in many aspects
  - How difficult to port **Track fit → most time consumption in Run 3 HLT2 reconstruction**

- **40% time of Track fit due to extrapolations in magnetic field**
- Still using well know **Runge-Kutta extrapolator, not vectorisable** horizontally
  - **Can look very different on GPUs**

- About **40% throughput from Selection +** ProtoParticles + Persistency in HLT2
- Smarter in selections & partial persistency?



*random pick from one recent ci-test

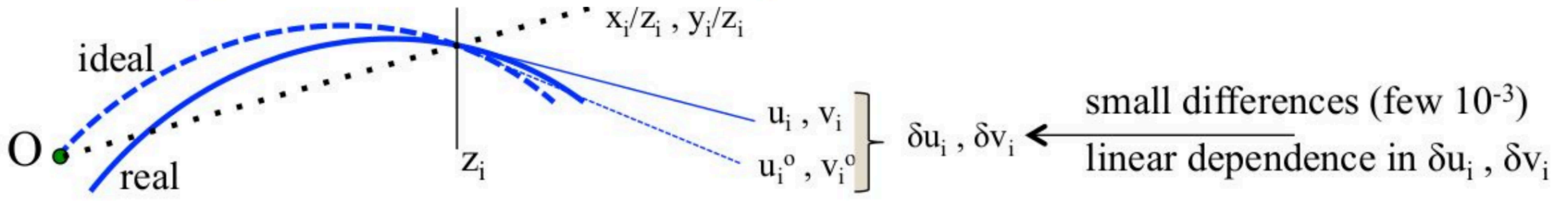
Beyond Runge-Kutta: Paramaterisations / ML

- Very nice math/ML master project with K. Spenlo - PhD student at LHCb/Belle2 at Ljubljana working together with A. Usachov

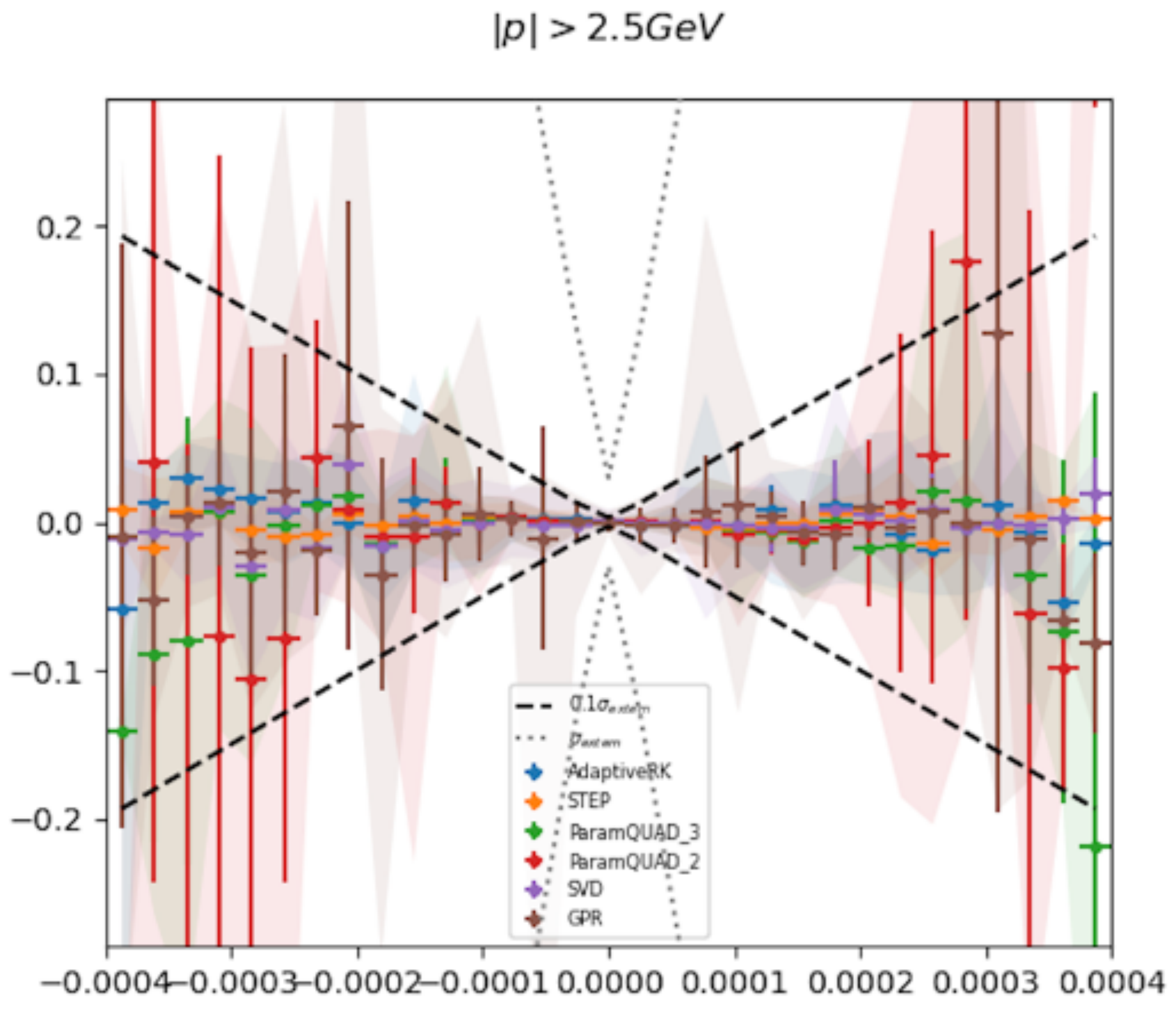
Comput. Phys. Commun. 265,108026 (2021)

- **Started with famous parameterisation** with help from Pierre, later with a bunch of ML approaches

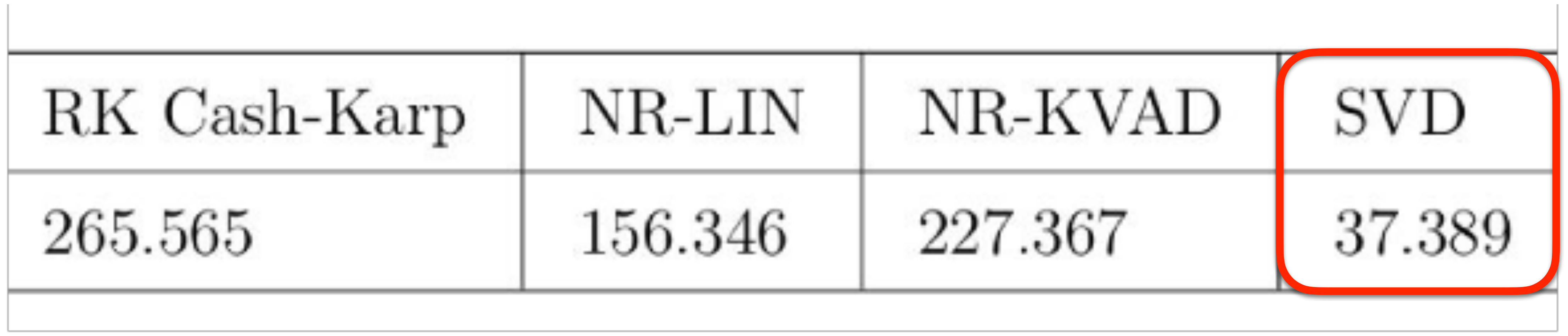
$$\mathbf{f}(\mathbf{x}_i) = \sum_{k=1}^{K_1} \mathbf{A}_k(x_i, y_i) \left(\frac{q}{p}\right)^k + \sum_{k=1}^{K_2} (\mathbf{B}_k(x_i, y_i)\, \delta u + \mathbf{C}_k(x_i, y_i)\, \delta v) \left(\frac{q}{p}\right)^k$$



ideal
O
real
z_i
x_i/z_i , y_i/z_i
u_i , v_i
u_i^o , v_i^o
δu_i , δv_i
small differences (few 10^{-3})
linear dependence in δu_i , δv_i



- Standalone python project with many extrapolators: **RK, parameterised, SVD..**
- Estimate of timing (Backward extrapolations)
- **Promising results** with Singular Value Decomposition **SVD extrapolator**



RK Cash-Karp	NR-LIN	NR-KVAD	SVD
265.565	156.346	227.367	37.389

- **Promising for the track fit on GPUs**

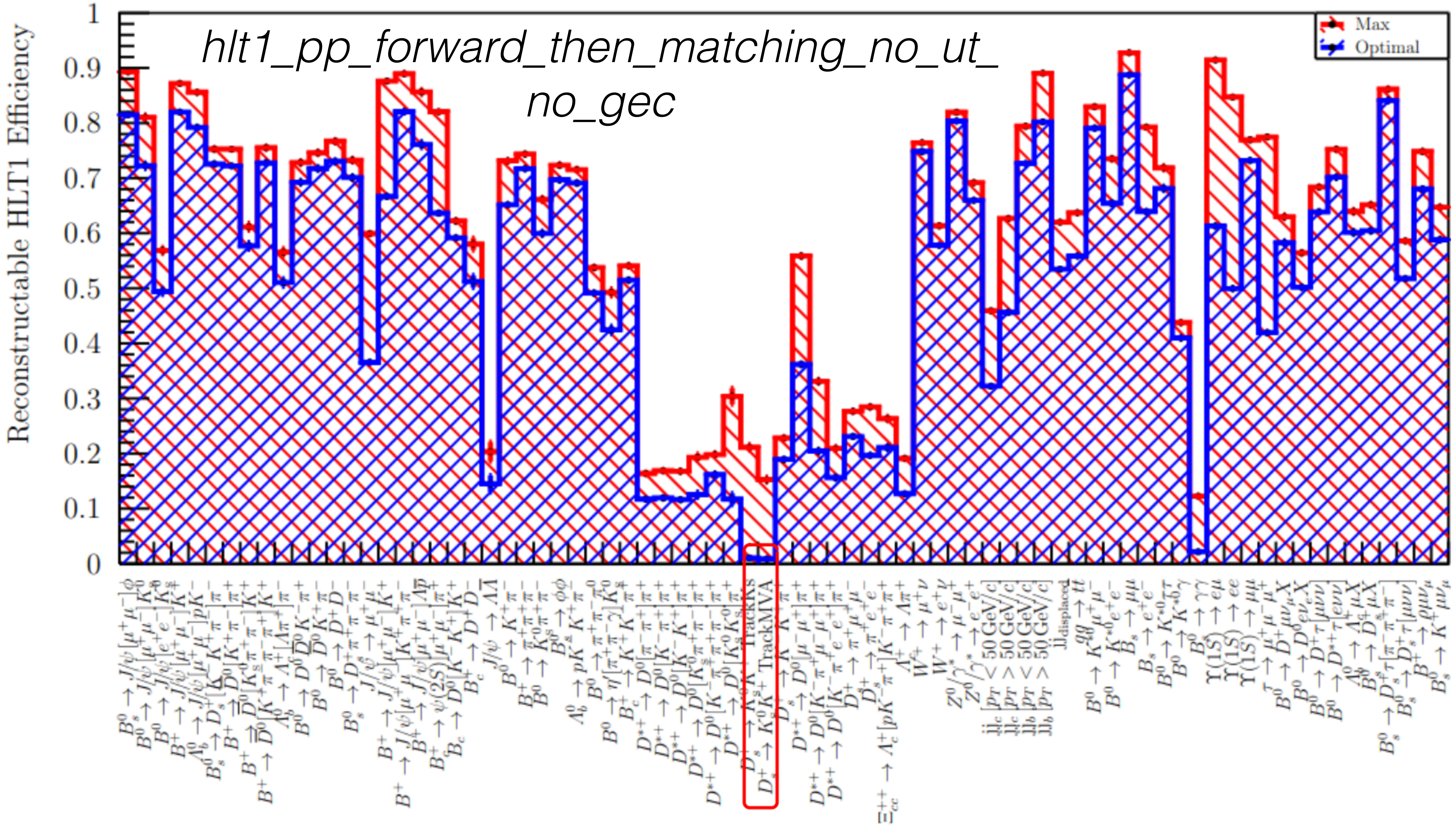
Lessons learnt from Run 3

- Interplay between detector and trigger performance very important
  - VELO PV & IP resolution, PV reconstruction efficiency
  - Momentum resolution / PID → signal / background ratio
  - Tracking efficiencies
  - Ghost rate

⇓

Joint efforts between sub-detectors and software would be essential

Tim’talk at RTA-WP3



TrackMVA p_T threshold goes to 10 GeV/c ..

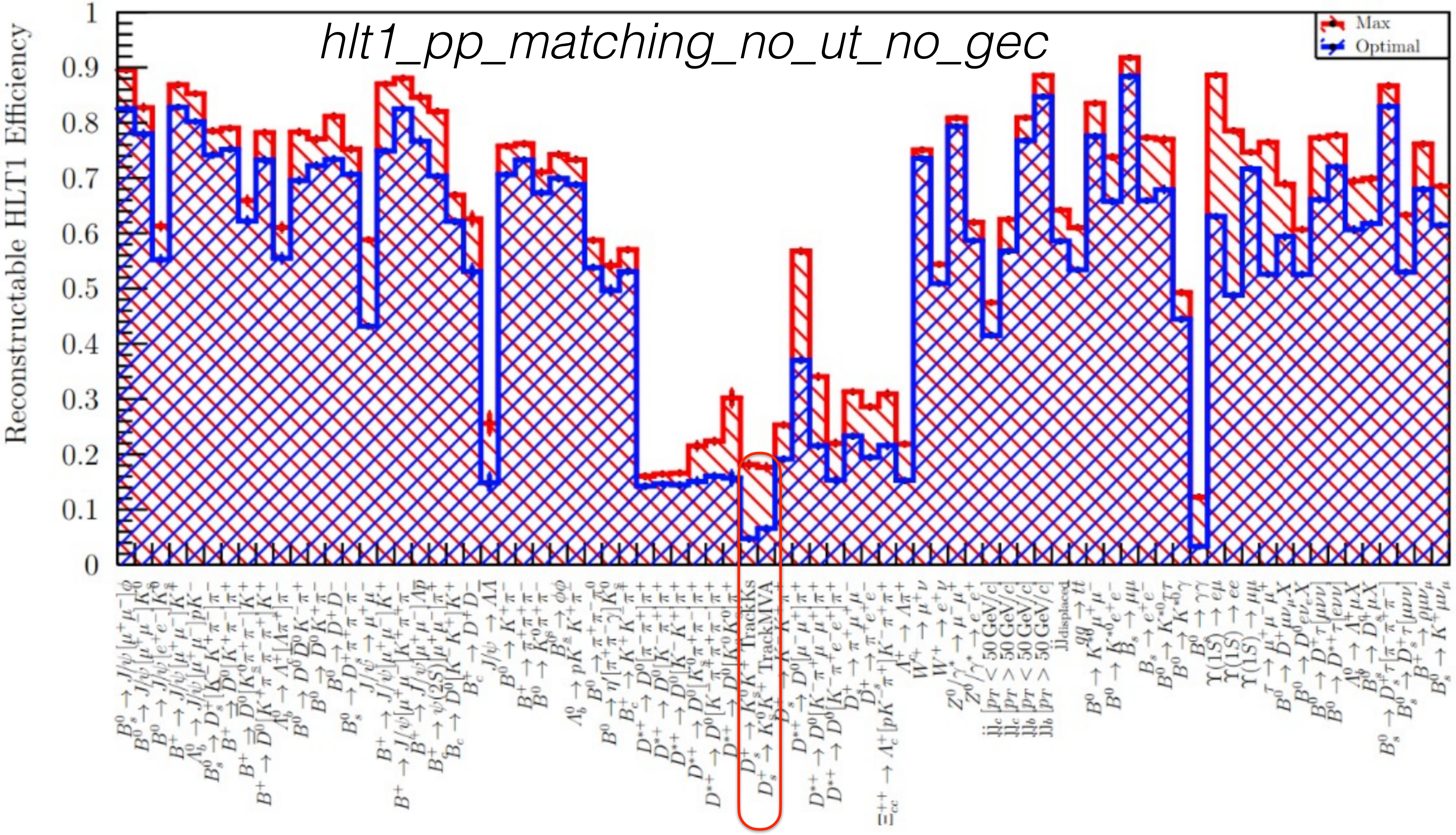
Lessons learnt from Run 3

- Interplay between detector and trigger performance very important
  - VELO PV & IP resolution, PV reconstruction efficiency
  - Momentum resolution / PID → signal / background ratio
  - Tracking efficiencies
  - Ghost rate

⇓

Joint efforts between sub-detectors and software would be essential

Alessandro's talk at LHCb



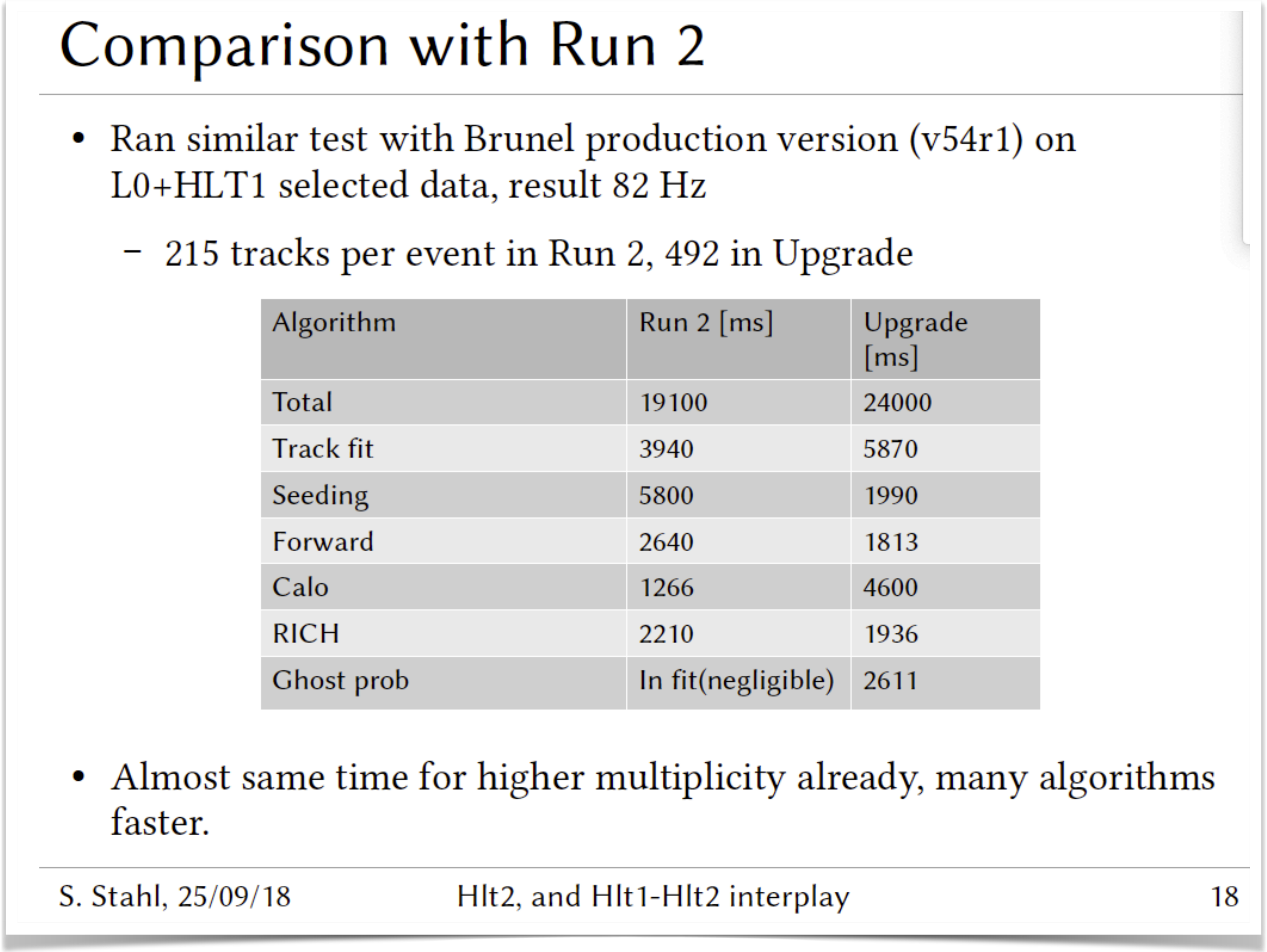
Lessons learnt even earlier

- Detector performance important for the reconstruction and trigger process

  - New detector + software optimisation → process a Run 3 event in the same time as a Run 2 event

  - More pixel trackers in UII expected to speed up reconstruction → **to be studied** with simulation & reconstruction

**Thanks Sascha for the interesting lesson! :)*



Comparison with Run 2

- Ran similar test with Brunel production version (v54r1) on L0+HLT1 selected data, result 82 Hz
  - 215 tracks per event in Run 2, 492 in Upgrade

Algorithm	Run 2 [ms]	Upgrade [ms]
Total	19100	24000
Track fit	3940	5870
Seeding	5800	1990
Forward	2640	1813
Calo	1266	4600
RICH	2210	1936
Ghost prob	In fit(negligible)	2611

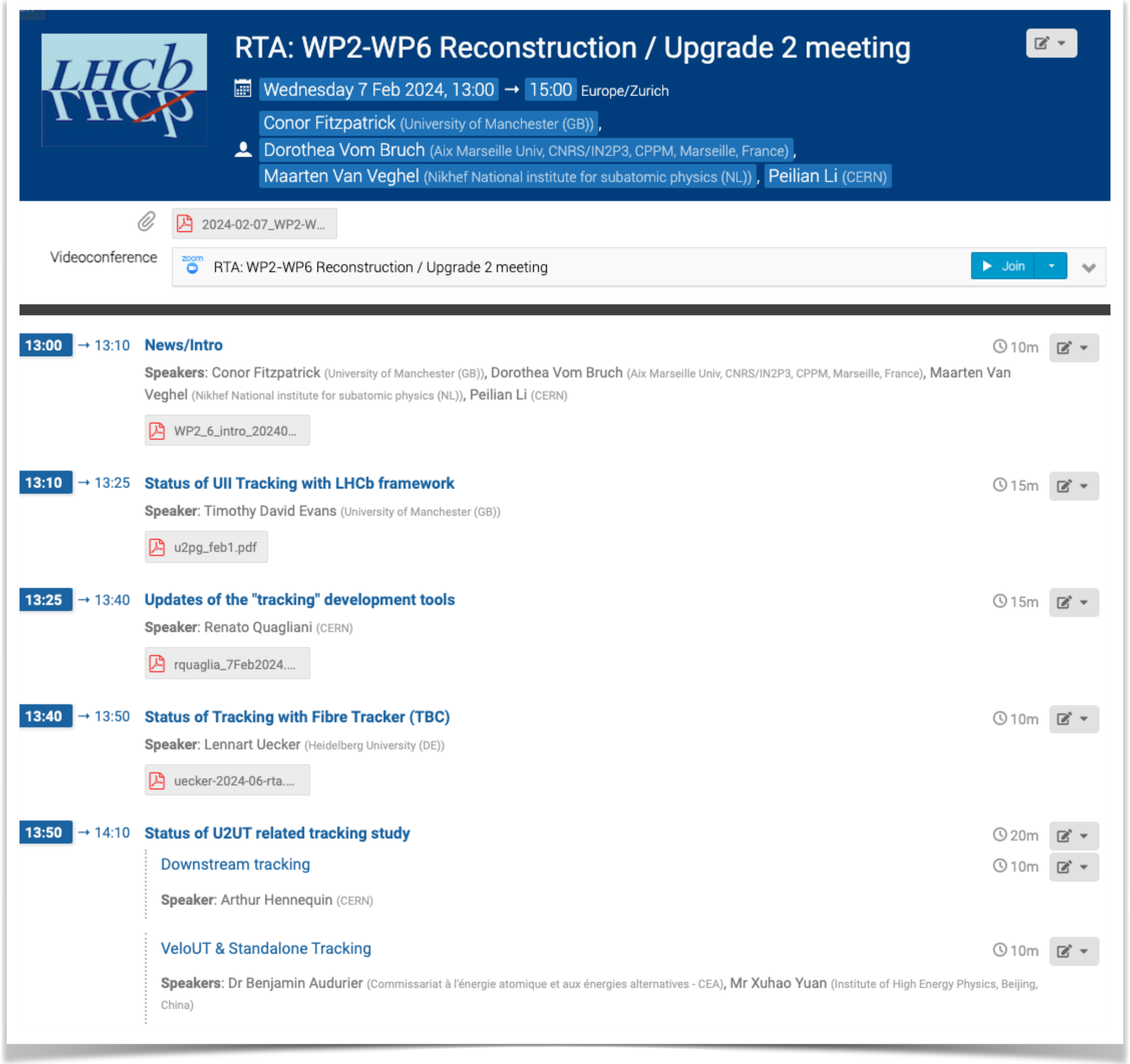
- Almost same time for higher multiplicity already, many algorithms faster.

S. Stahl, 25/09/18 Hlt2, and Hlt1-Hlt2 interplay 18

Preparation of Run 5 reconstruction

- Joint efforts between sub-detector & software needed to move forward
- Overview meeting organised in joint WP2 & WP6 meeting

  - Share common tools and framework
  - Converge on common performance metrics
  - Very productive discussion on the action points to **make long track reconstruction possible**



RTA: WP2-WP6 Reconstruction / Upgrade 2 meeting

Wednesday 7 Feb 2024, 13:00 → 15:00 Europe/Zurich

Conor Fitzpatrick (University of Manchester (GB)), Dorothea Vom Bruch (Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France), Maarten Van Veghel (Nikhef National institute for subatomic physics (NL)), Peilian Li (CERN)

2024-02-07_WP2-W...

Videoconference: RTA: WP2-WP6 Reconstruction / Upgrade 2 meeting

13:00 → 13:10 **News/Intro** 10m

Speakers: Conor Fitzpatrick (University of Manchester (GB)), Dorothea Vom Bruch (Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France), Maarten Van Veghel (Nikhef National institute for subatomic physics (NL)), Peilian Li (CERN)

WP2_6_intro_20240...

13:10 → 13:25 **Status of UII Tracking with LHCb framework** 15m

Speaker: Timothy David Evans (University of Manchester (GB))

u2pg_feb1.pdf

13:25 → 13:40 **Updates of the "tracking" development tools** 15m

Speaker: Renato Quagliani (CERN)

rquaglia_7Feb2024....

13:40 → 13:50 **Status of Tracking with Fibre Tracker (TBC)** 10m

Speaker: Lennart Uecker (Heidelberg University (DE))

uecker-2024-06-rta....

13:50 → 14:10 **Status of U2UT related tracking study** 20m

Downstream tracking 10m

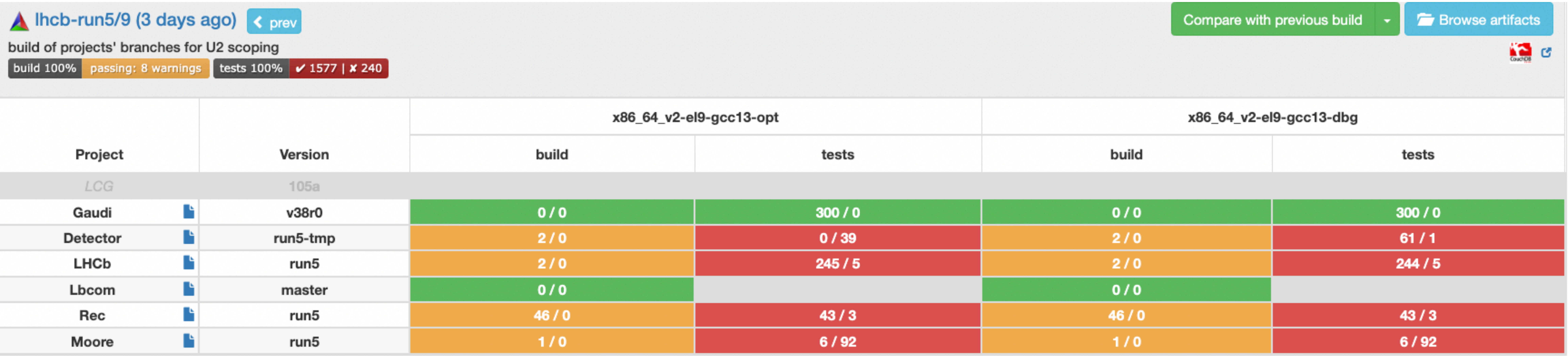
Speaker: Arthur Hennequin (CERN)

VeloUT & Standalone Tracking 10m

Speakers: Dr Benjamin Audurier (Commissariat à l'énergie atomique et aux énergies alternatives - CEA), Mr Xuhao Yuan (Institute of High Energy Physics, Beijing, China)

LHCb framework for Run 5

- **Nightly build set** up with *run5* branches in Detector, LHCb, Rec, Moore (Many thanks to Ben, Marco Cl., Tim & Renato!)
- Instruction to build with *lb-dev* or the *full stack* & Mattermost channel for discussion
- **Long tracks reconstruction** with fake clusterings in TV, UP, MP possible
- More details about the reconstruction performance in talks tomorrow afternoon

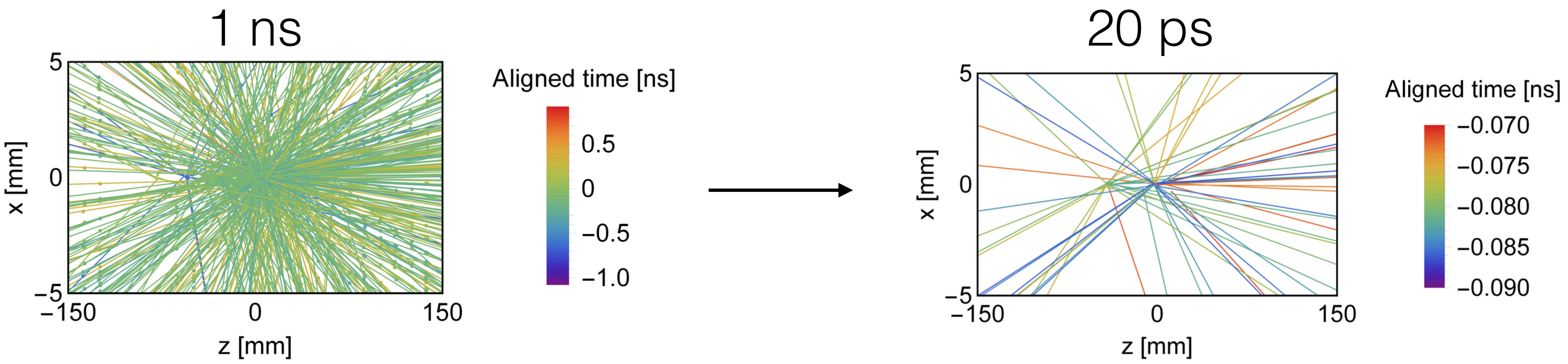


build of projects' branches for U2 scoping

build 100% passing: 8 warnings tests 100% ✔ 1577 | ✘ 240

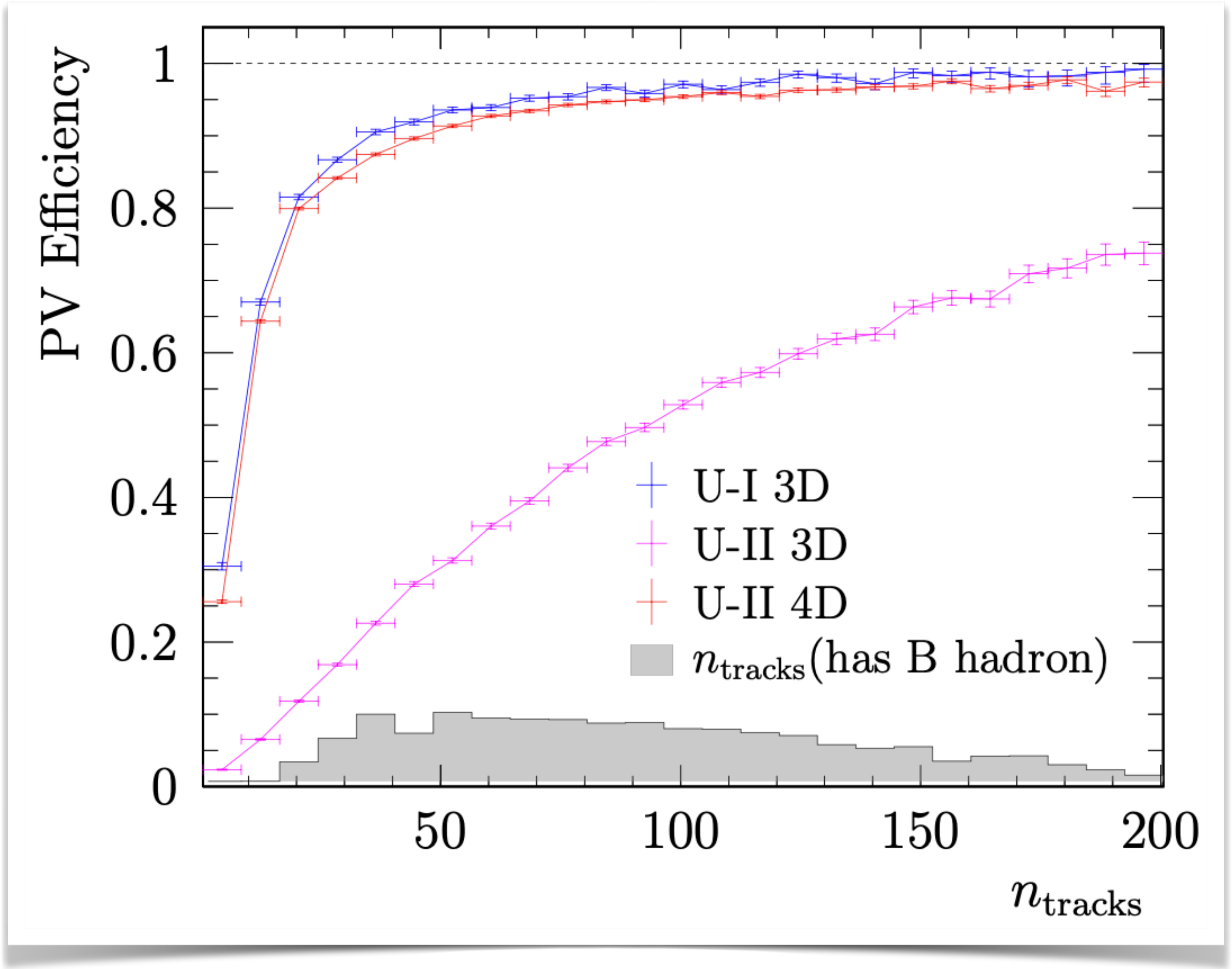
Project	Version	x86_64_v2-el9-gcc13-opt		x86_64_v2-el9-gcc13-dbg	
		build	tests	build	tests
LCG	105a				
Gaudi	v38r0	0 / 0	300 / 0	0 / 0	300 / 0
Detector	run5-tmp	2 / 0	0 / 39	2 / 0	61 / 1
LHCb	run5	2 / 0	245 / 5	2 / 0	244 / 5
Lbcom	master	0 / 0		0 / 0	
Rec	run5	46 / 0	43 / 3	46 / 0	43 / 3
Moore	run5	1 / 0	6 / 92	1 / 0	6 / 92

Pileup mitigation using timing



Magical solution to all our problem?

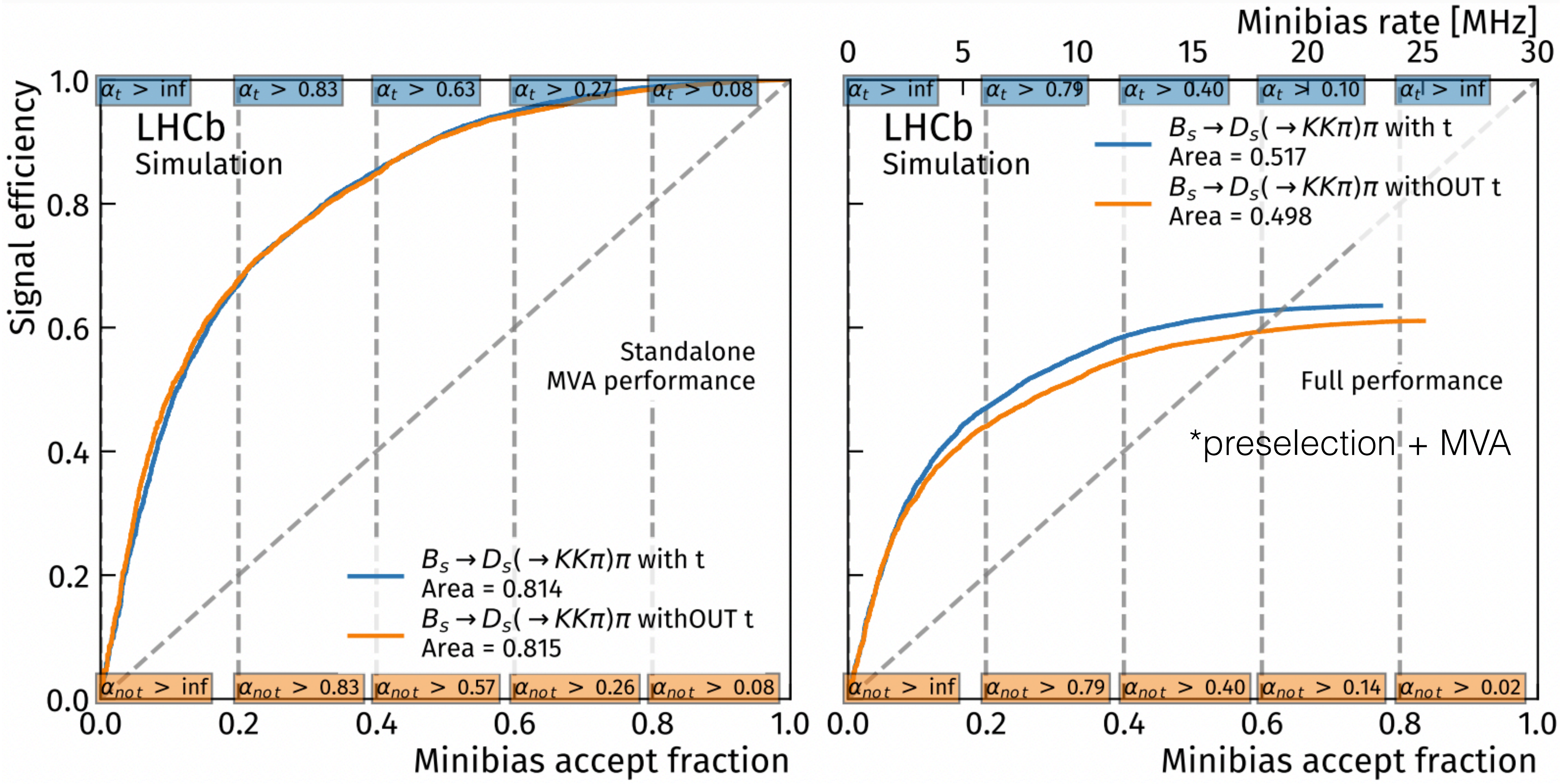
- First studies on VELO & ECAL show **timing is crucial for track-PV association + background rejection** of photon signatures
- How helpful for the selection and reduction of data rate?



Timing effect on HLT1 trigger

- Very preliminary study in a summer student project by F. Harz using *UII VELO simulation*
- Follow similar selection in 2-track MVA, w./w.o. time information used in χ^2_{IP}, χ^2_{vtx} and χ^2_{flight}

  - **4% increase in signal TOS efficiency**
  - **8% reduction of minibias rate**
  - Exploring further usage of time information
  - Exclusive selections in HLT1?



*preselection + MVA

**signal efficiency is TOS efficiency on $B_s^0 \to D_s^- \pi^+$*

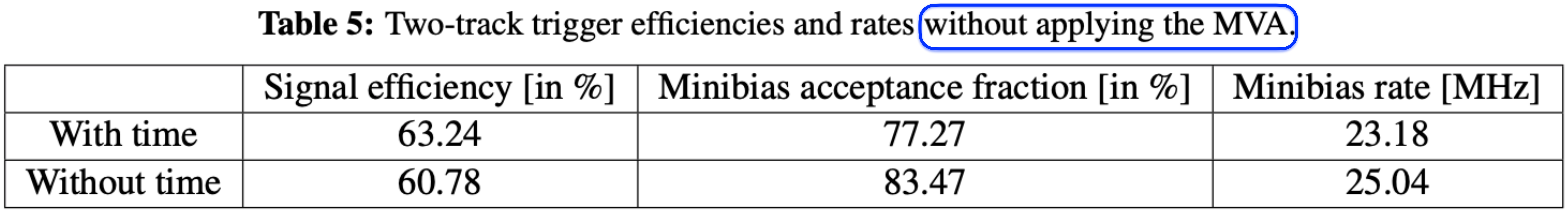


Table 5: Two-track trigger efficiencies and rates without applying the MVA.

	Signal efficiency [in %]	Minibias acceptance fraction [in %]	Minibias rate [MHz]
With time	63.24	77.27	23.18
Without time	60.78	83.47	25.04

Questions asked by U2PG

Relevance of magnetic field in UT (+ Velo) region?

- Matching method (Velo + SciFi tracks, not require magnetic field in UT region) works in both HLT1 & HLT2 Run 3 → Caveat: have to estimate how it scales in higher lumi
- Very few physics need Upstream tracks, where 10% uncertainty might be fine
- Pattern recognition of Velo-UT-MS tracks might be challenging without momentum estimate of Velo-UT part

Implications of modules with "low" design hit efficiency?

- Could mask to the designed efficiency randomly to estimate the tracking efficiency
- In the VELO layout optimisation, this is considered by requiring more stations

Impact of not levelling at the start of the fill on reconstruction systematics?

- In Run 3, Velo tracking efficiency goes down a bit with higher mu, as well as PID
- Deeper study requires simulations and reconstruction
- In terms of systematics, as long as MC can reflect the levelling, would be able to evaluate properly (significant work, not a showerstopper nor priority)

Summary

- LHCb UII will increase the instantaneous luminosity by another factor of 5~7
- Major challenges and great opportunities:
  - Keep software-only trigger at 30 MHz with much higher complexity of events
  - Adapt trigger strategy to deal with signal decays in every event
  - Both HLT1 and HLT2 reconstruction with GPUs
- Many proposals: timing information, heterogeneous systems, ML, FPGA clustering/tracking …
- Joint efforts between sub-detectors and software important to develop demonstrators of the detector and physics performance
  - First nightly build with TV, UP, MP included enable the reconstruction of long tracks
  - Early look at 1-track/2-track trigger with UII samples…

Thank you!

Back up

- Very preliminary study in a summer student project by F. Harz using UII VELO simulation
  - 1-track MVA

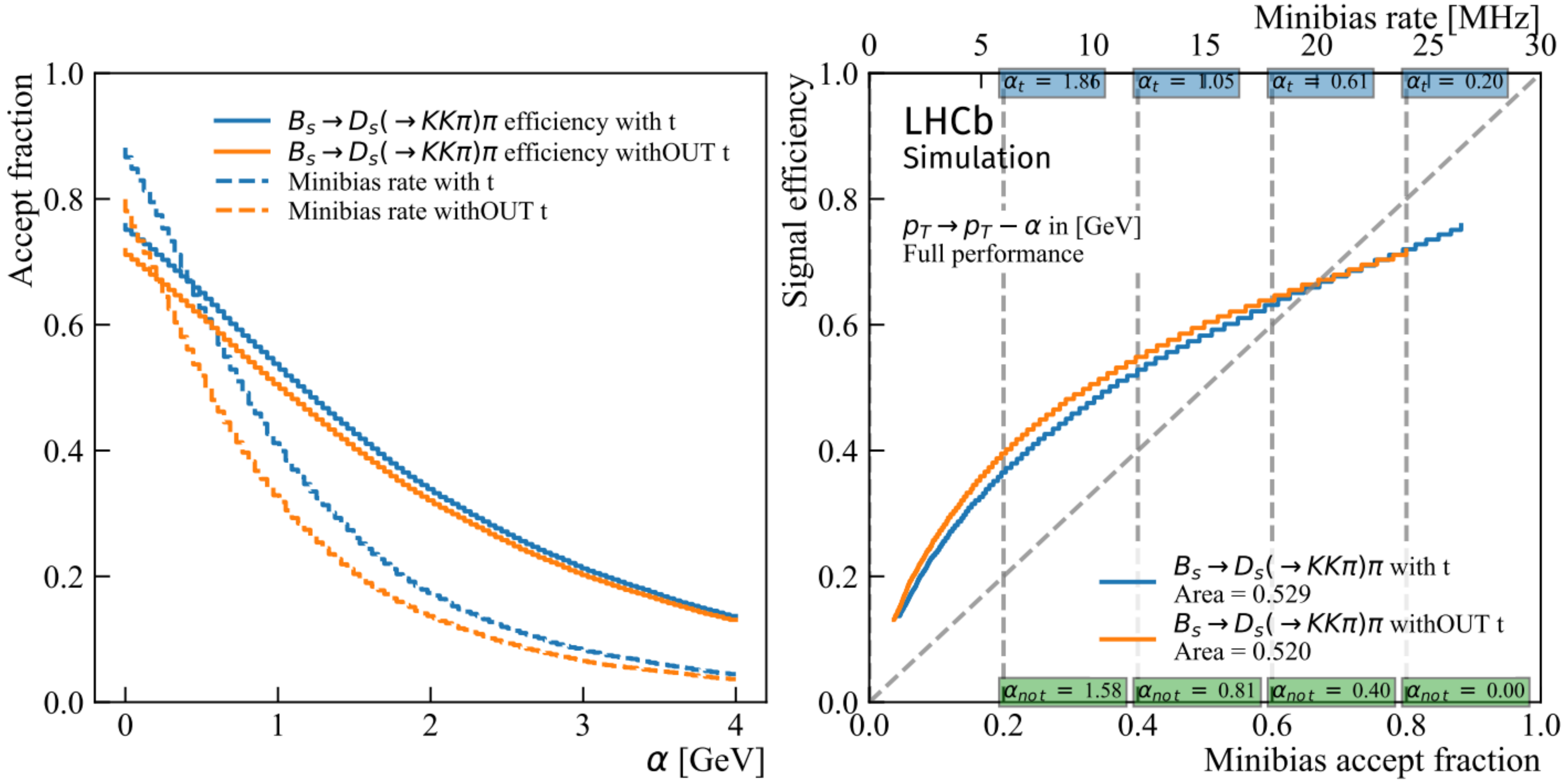


Figure 4: Effect of the p_T-re-tuning ($p_T \to p_T - \alpha$) on the signal efficiency and minibias rate for the one-track trigger while applying a 50 ps time window compared to not applying this time window. The left side shows the acceptance fraction of the signal and minibias for different values of α. The right side shows the same information by drawing the signal efficiency against t

total rate because the preselection (for α

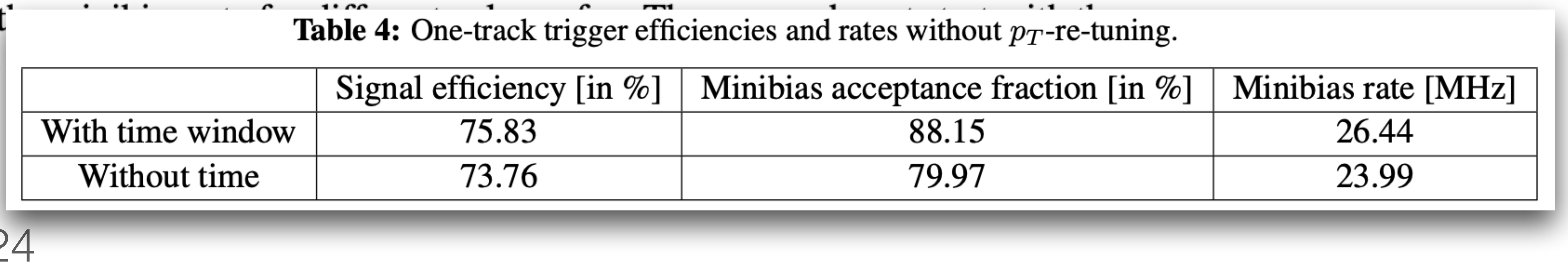
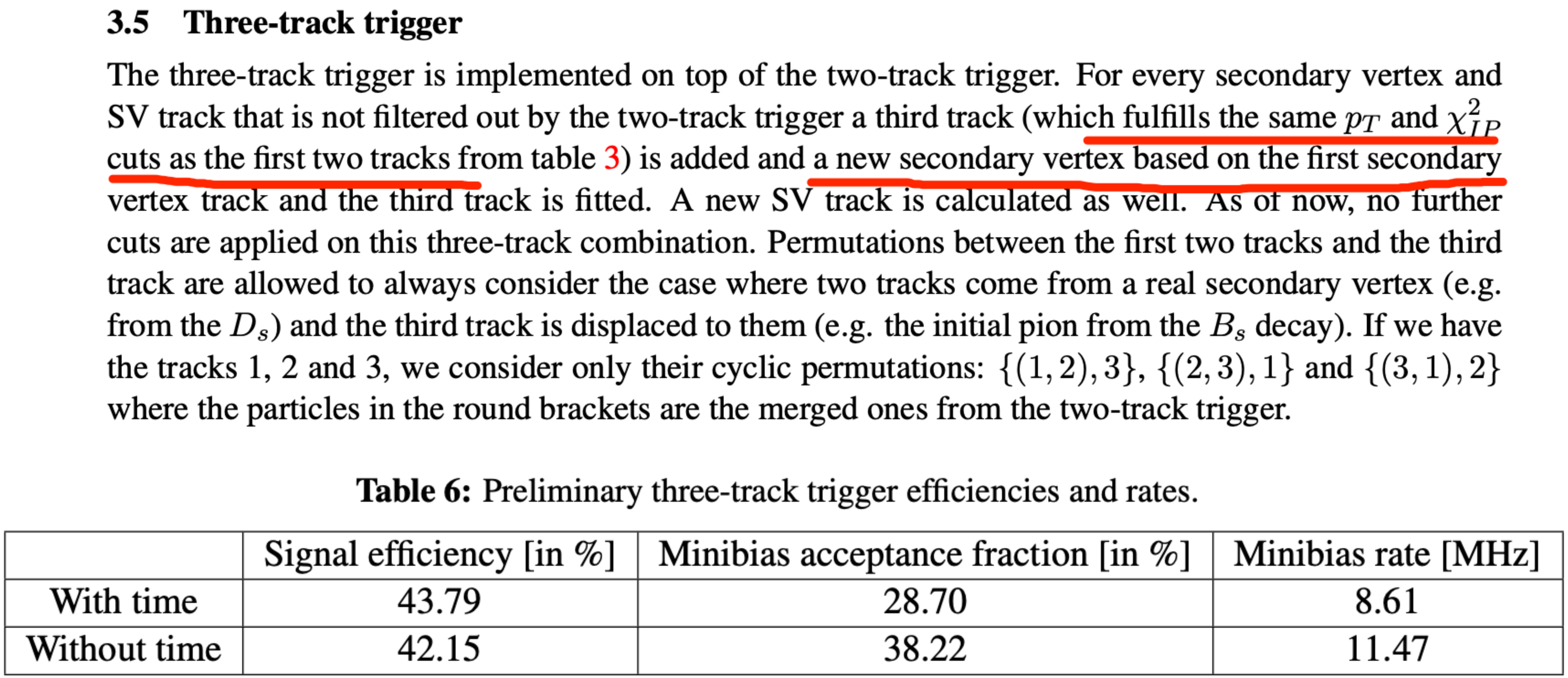


Table 4: One-track trigger efficiencies and rates without p_T-re-tuning.

	Signal efficiency [in %]	Minibias acceptance fraction [in %]	Minibias rate [MHz]
With time window	75.83	88.15	26.44
Without time	73.76	79.97	23.99

Back up

- Very preliminary study in a summer student project by F. Harz using UII VELO simulation
  - 3-track MVA, on top of the 2-track MVA with MVA cut at a fixed value



3.5 Three-track trigger

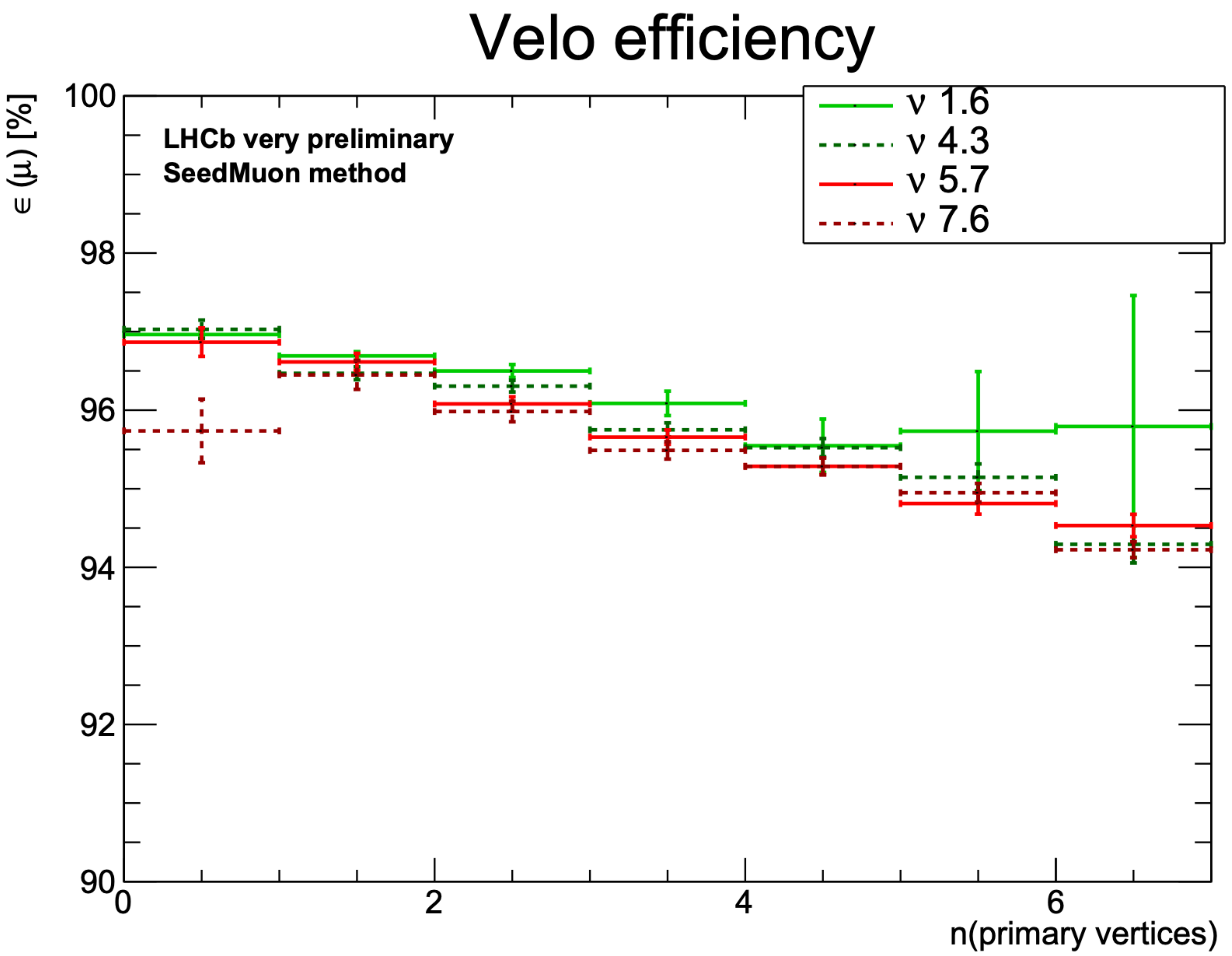
The three-track trigger is implemented on top of the two-track trigger. For every secondary vertex and SV track that is not filtered out by the two-track trigger a third track (which fulfills the same p_T and χ^2_{IP} cuts as the first two tracks from table 3) is added and a new secondary vertex based on the first secondary vertex track and the third track is fitted. A new SV track is calculated as well. As of now, no further cuts are applied on this three-track combination. Permutations between the first two tracks and the third track are allowed to always consider the case where two tracks come from a real secondary vertex (e.g. from the D_s) and the third track is displaced to them (e.g. the initial pion from the B_s decay). If we have the tracks 1, 2 and 3, we consider only their cyclic permutations: $\{(1,2),3\}$, $\{(2,3),1\}$ and $\{(3,1),2\}$ where the particles in the round brackets are the merged ones from the two-track trigger.

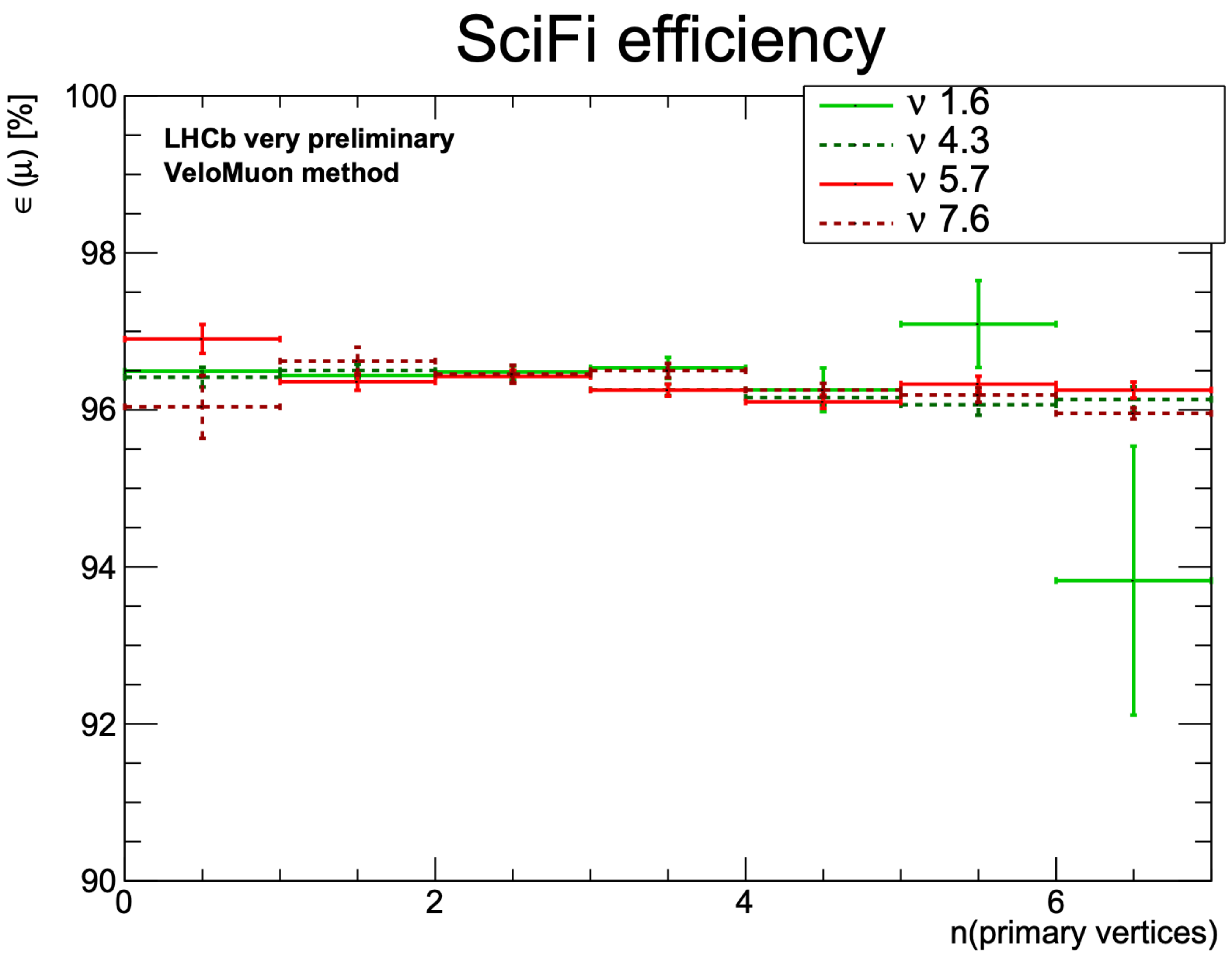
Table 6: Preliminary three-track trigger efficiencies and rates.

	Signal efficiency [in %]	Minibias acceptance fraction [in %]	Minibias rate [MHz]
With time	43.79	28.70	8.61
Without time	42.15	38.22	11.47

Back up

- Tracking efficiency from expected-24 MC samples with different mu (Rowina's report at RTA-WP4)
  - decrease a bit when mu goes up


Velo efficiency
ϵ (μ) [%]
n(primary vertices)
LHCb very preliminary
SeedMuon method
ν 1.6
ν 4.3
ν 5.7
ν 7.6


SciFi efficiency
ϵ (μ) [%]
n(primary vertices)
LHCb very preliminary
VeloMuon method
ν 1.6
ν 4.3
ν 5.7
ν 7.6