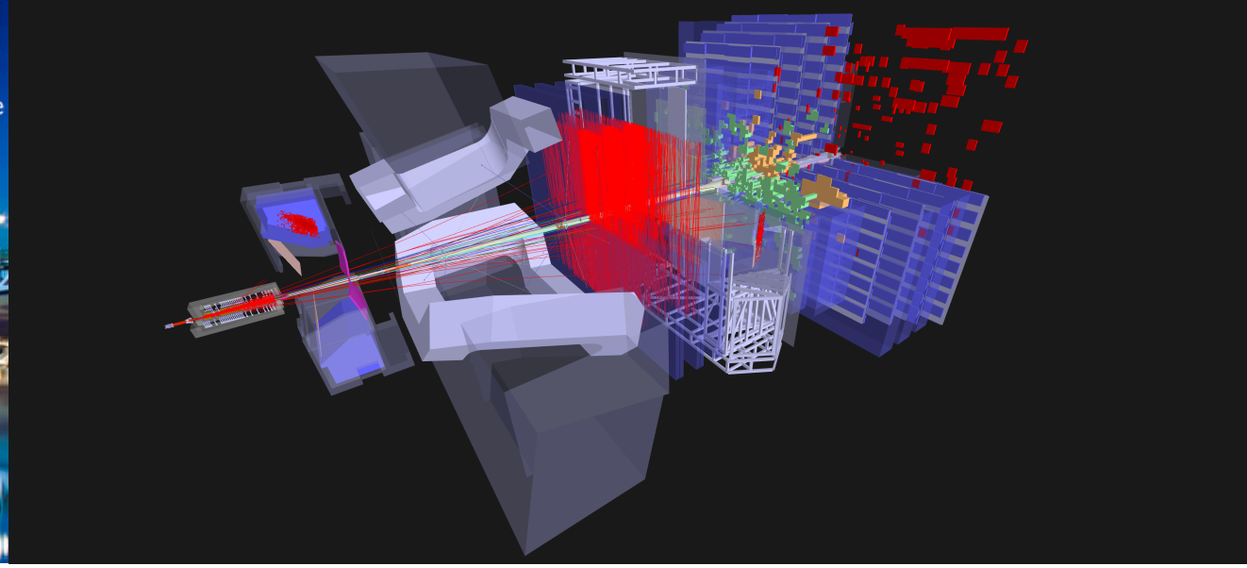



# FPGA-based architecture for a real-time track reconstruction in the LHCb Scintillating Fibre Tracker beyond Run 3

Michael J. Morello et al. [see in the backup]
on behalf of the LHCb Collaboration

8th International Connecting The Dots Workshop
*Mini-workshop on Real time tracking : triggering events with tracks*
October 10th-13th, 2023

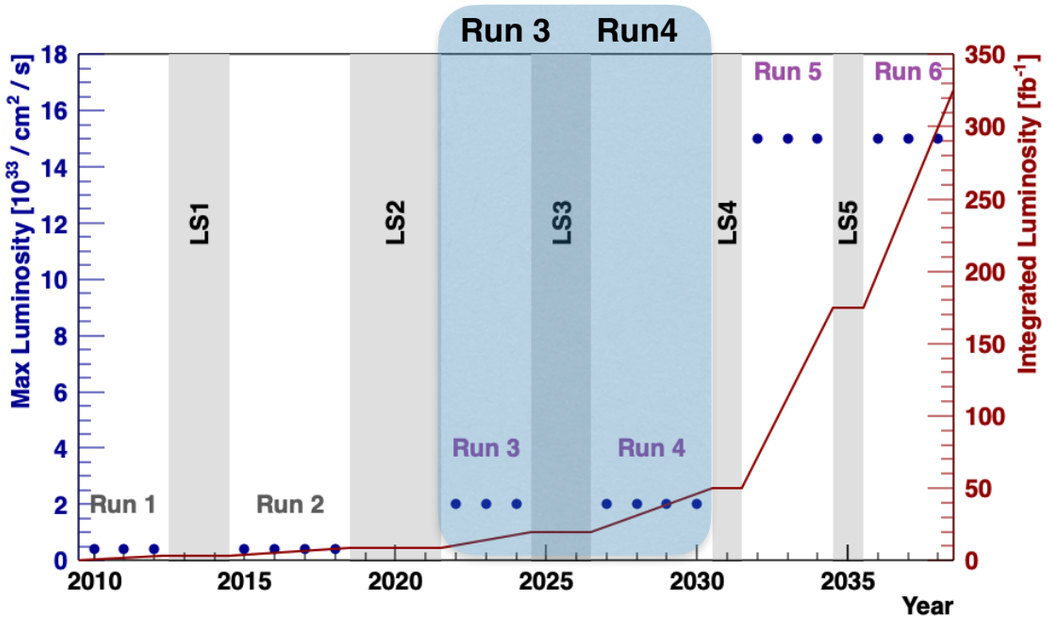
SCUOLA NORMALE SUPERIORE

INFN
Istituto Nazionale di Fisica Nucleare

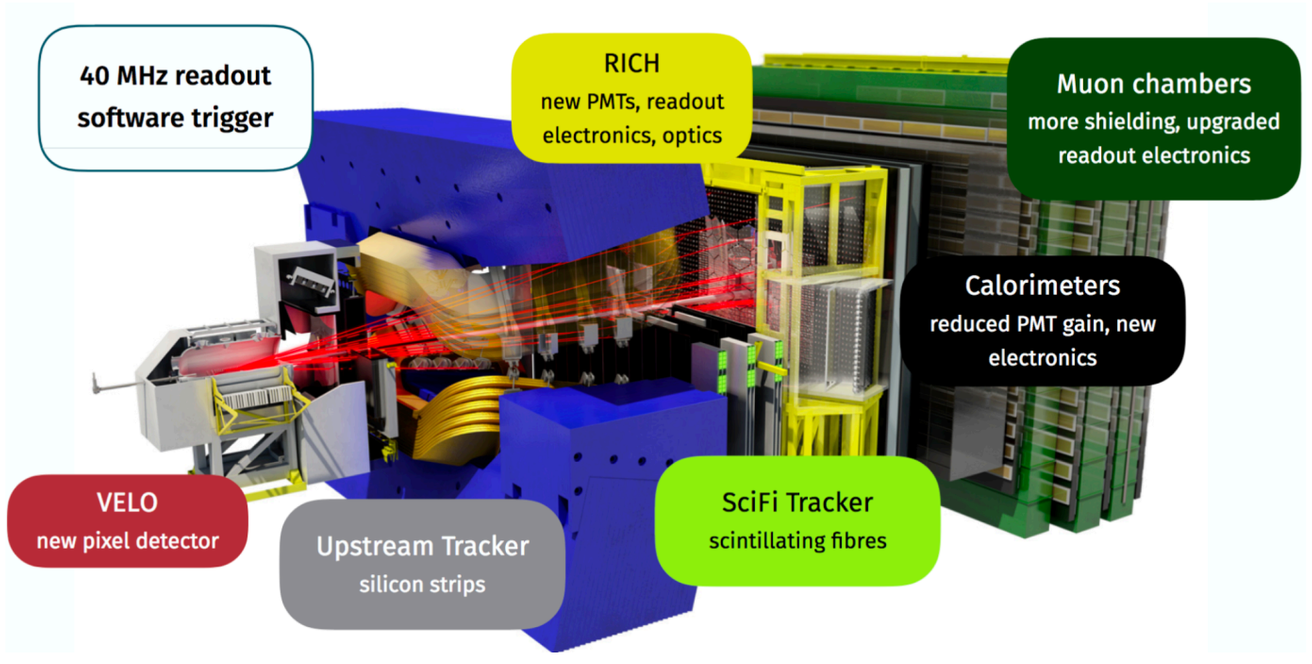
LHCb RTA

LHCb

# The LHCb Upgrade I

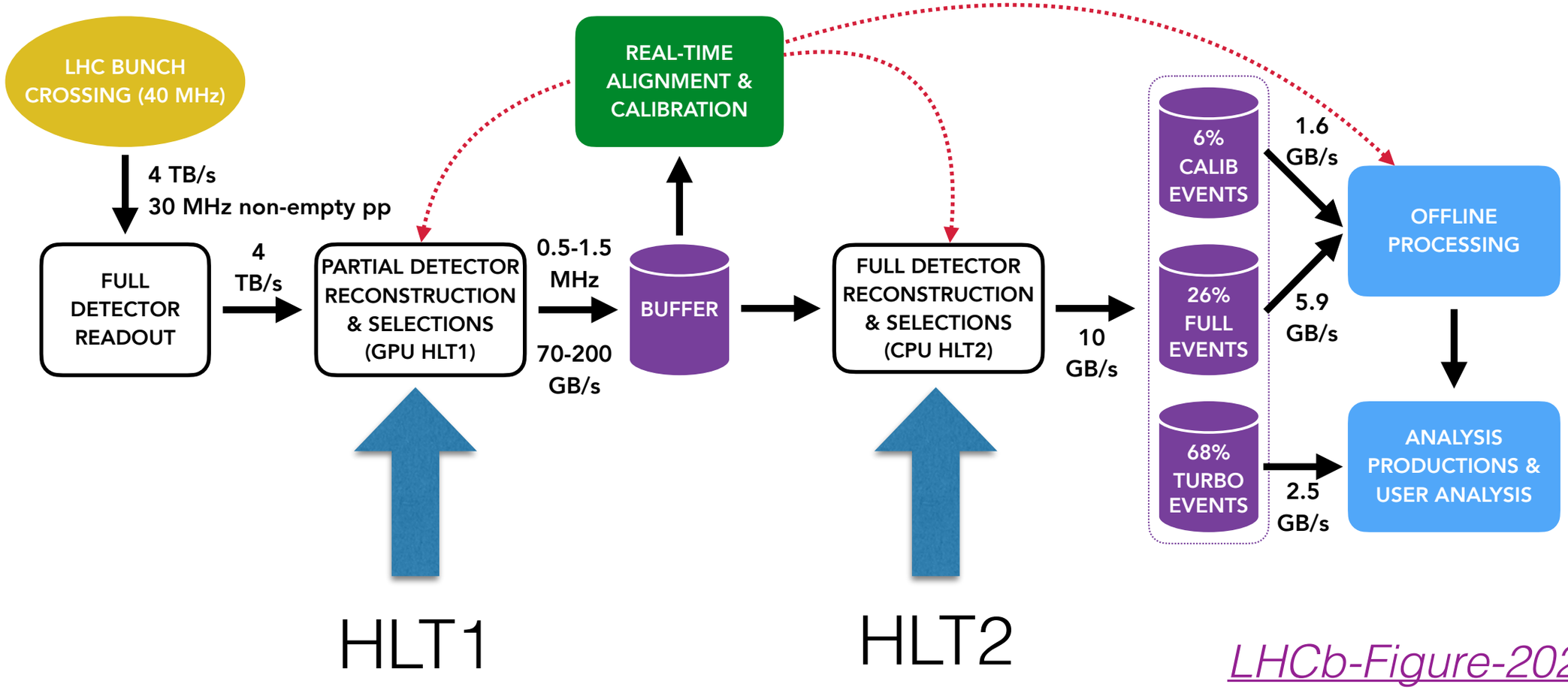



*The LHCb upgrade I [2305.10515]*
*LHCb upgrade I Seminar*




- Visible interactions (pile-up) 7.6 (1.1).
- $\sqrt{s}$ = 13.6 TeV (13 TeV).
- Luminosity: $2 \times 10^{33}$ cm$^{-2}$s$^{-1}$ (4 x 10$^{32}$ cm$^{-2}$s$^{-1}$).
- Expected integrated lumi: 50 fb$^{-1}$ (9 fb$^{-1}$).
- Brand-new tracking detectors → VELO, Upstream Tracker, Scintillating Fibre Tracker.
- Read-out of all sub-detectors at 40MHz → **Reconstructing all tracks in the High Level Trigger.**
- Goals wrt Run 2 → same efficiency/fb$^{-1}$ for muonic modes, x2 efficiency/fb$^{-1}$ for hadronic modes.

# DAQ and trigger in Upgrade I

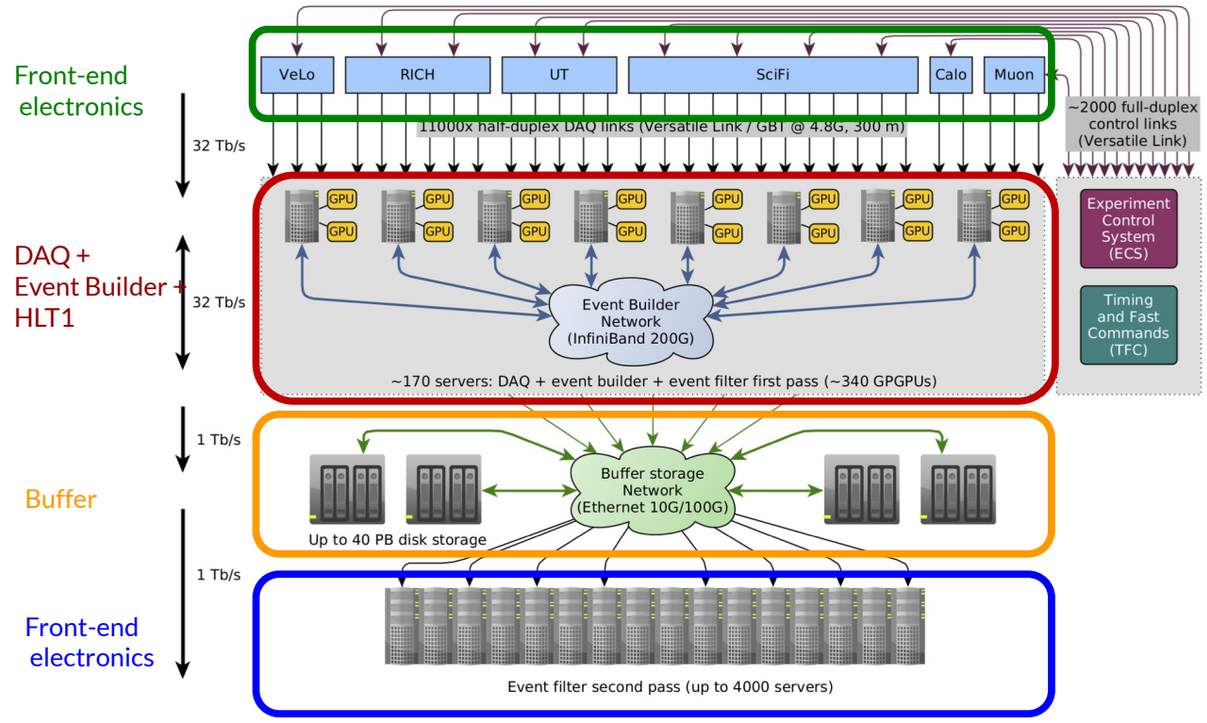



*LHCb-Figure-2020-016*

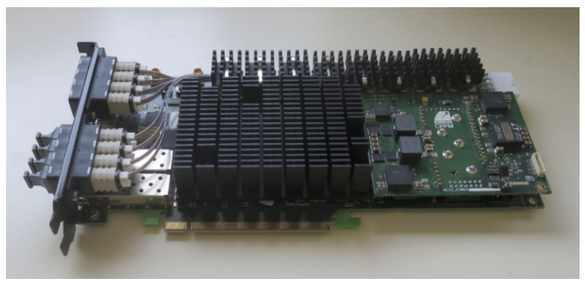
Triggering divided into two stages. HLT1 uses an array of GPU servers to perform a faster event reconstruction. HLT2, based on CPU servers, performs a complete reconstruction with an offline-level quality, permanently stored for subsequent analysis.

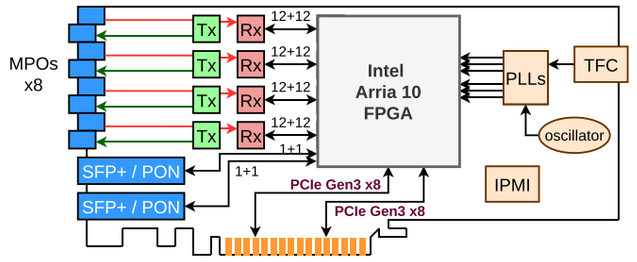
# DAQ and trigger in Upgrade I

- Custom PCIe40 cards to receive data from sub-detectors (aka TELL40), equipped with an Intel Arria 10 FPGA chip.
- A fast Event Builder (EB) network routes up to 32 Tb/s data rate.
- **170 EB servers with 2 GPU cards for HLT1**.
- 40 PB disk storage.
- HLT2 runs asynchronously on CPUs to take advantage of the LHC's dead time.
- **First example of complex reconstruction on FPGA at 30 MHz**: 2D VELO clustering already deployed in TELL40 (FPGA) [*10.1109/TNS.2023.3273600*].



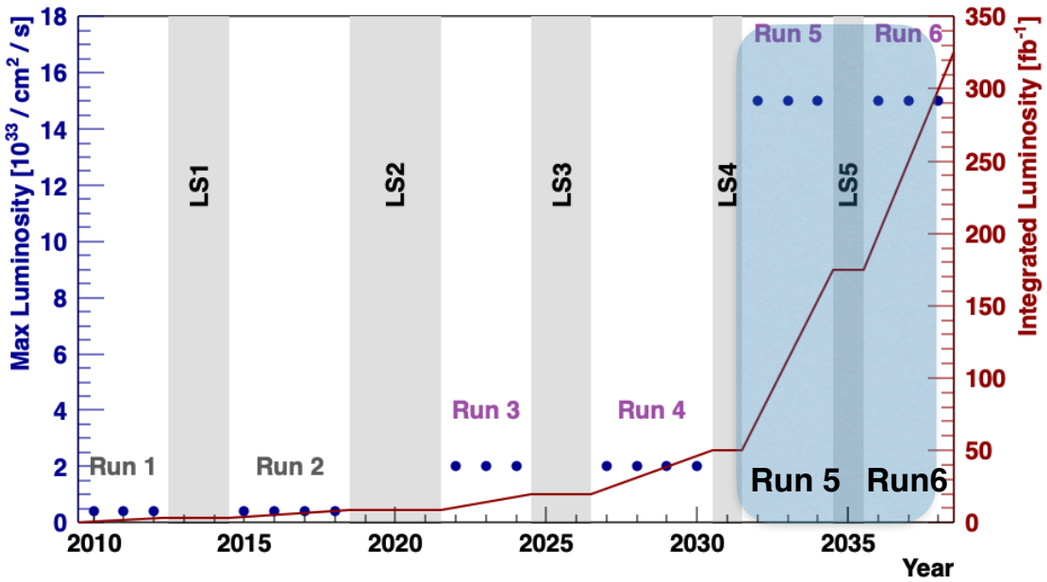

PCIe40 readout card





# LHCb in Run 5&6 (Phase II)

- Target instantaneous luminosity: ~ 1.5 x $10^{34}$ $cm^{-2}s^{-1}$
- Target integrated luminosity: ~300 $fb^{-1}$.
- Visible interactions (pile-up): ~40.
- Keep same performance in such more difficult conditions, timing will be required in some detectors.
- About 200 Tb/s data to be **reconstructed** in real time.

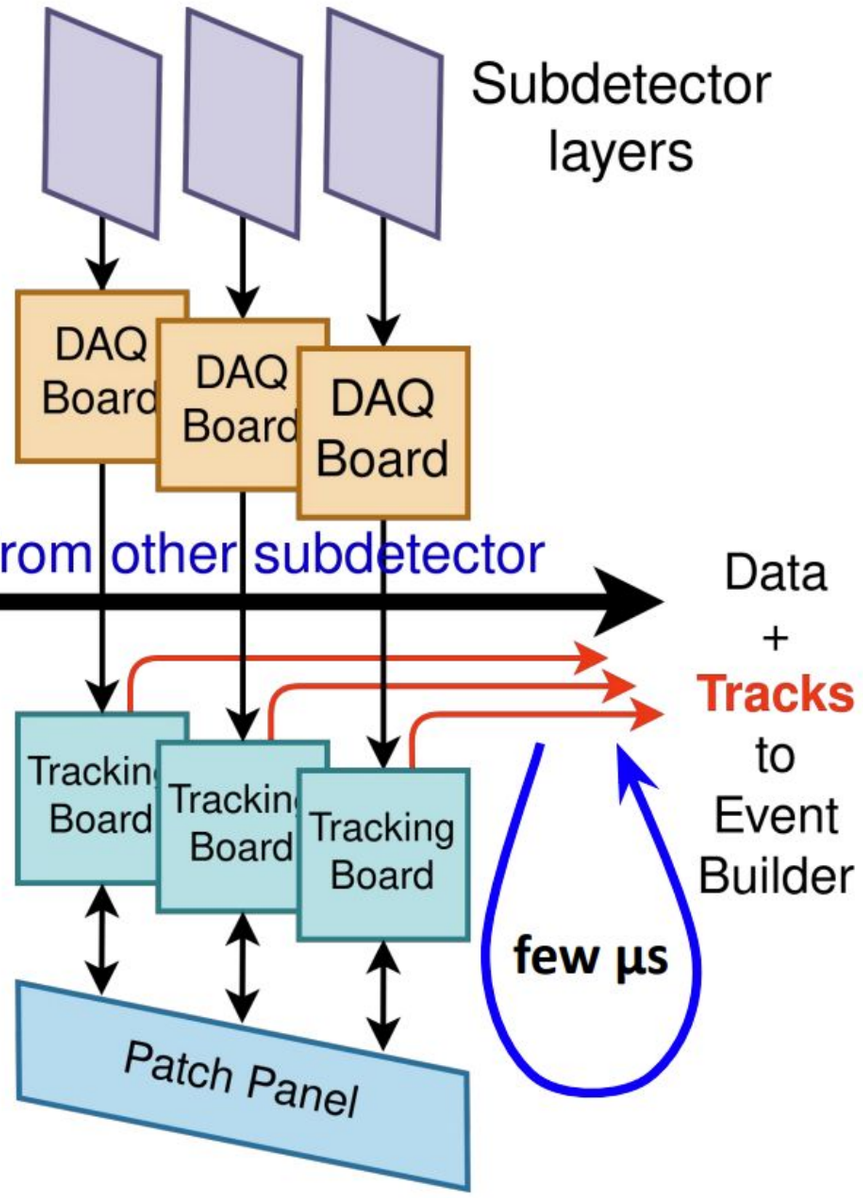


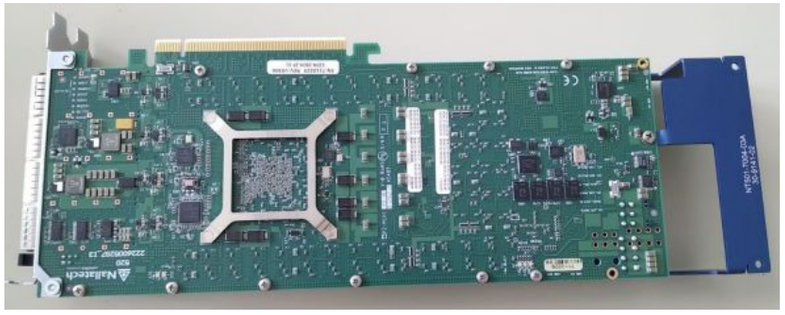

- More and more processing has to be performed earlier in the DAQ chain to efficiently reduce data size as soon as possible. **Costs could be an important limitation.** Greener solutions are needed.
- **Moving to a "heterogeneous-computing" paradigm is of paramount importance**.
- Take advantage of Run 4 (same conditions as Run 3) to develop novel approaches.


LHCb-TDR-023

# Real-time tracking with FPGAs

- Modern FPGAs can perform parallel data processing, integrated in the DAQ architecture, with high throughputs, low latencies and better energy efficiency than CPUs and GPUs (for certain tasks).
- So fast to allow "local" reconstruction **before event building**. Thus, also saving bandwidth immediately.
- This talk: specifically aimed at the realization of a real-time tracking unit to reconstruct standalone tracks downstream the magnet on FPGAs with the "**artificial retina**" algorithm.



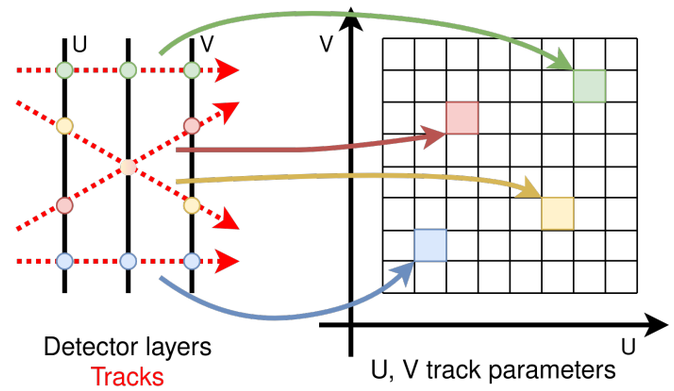
PCIe 16x board, 1 Intel Stratix 10 FPGA, 16 optical links

NIM A453 (2000) 425-429

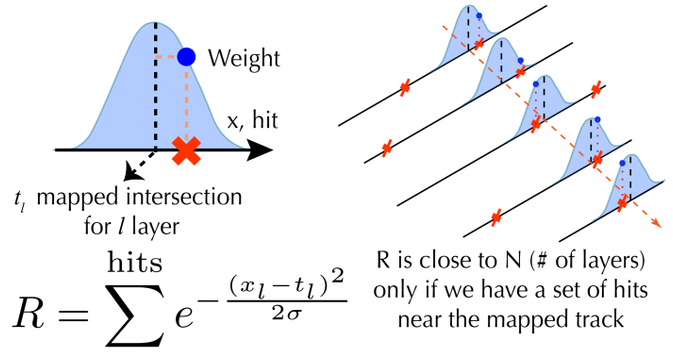
# The “artificial retina” architecture

- **Step 0 -** Discretize space of track parameters (pattern cells) and generate track intersections with detector planes (receptors) and connect them to cells (mapping).
- **Step 1 -** Detector hits are distributed (Switching Network) only to a reduced number of cells according the mapping of Step 0 (LUT).
- **Step 2 -** A logic unit (engine) for each cell accumulates a Gaussian weight proportional to the distance with the receptors.
- **Step 3 -** Tracks are identified as local maxima of accumulated weights, above a certain threshold, over the cells grid.

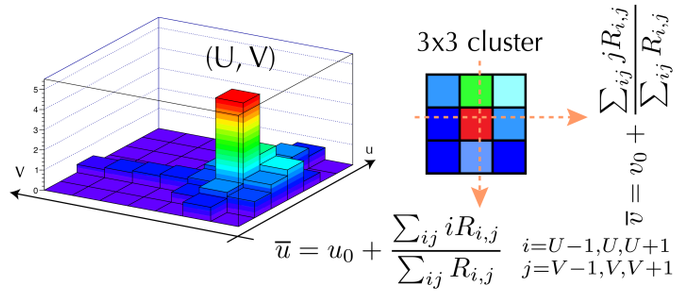
Conceived for parallelism (cells work in parallel): high-throughput and low-latencies. FPGA size limitations overcome by spreading cells over several chips (without increasing latency).




**Step 2: Accumulating weights (each cell)**

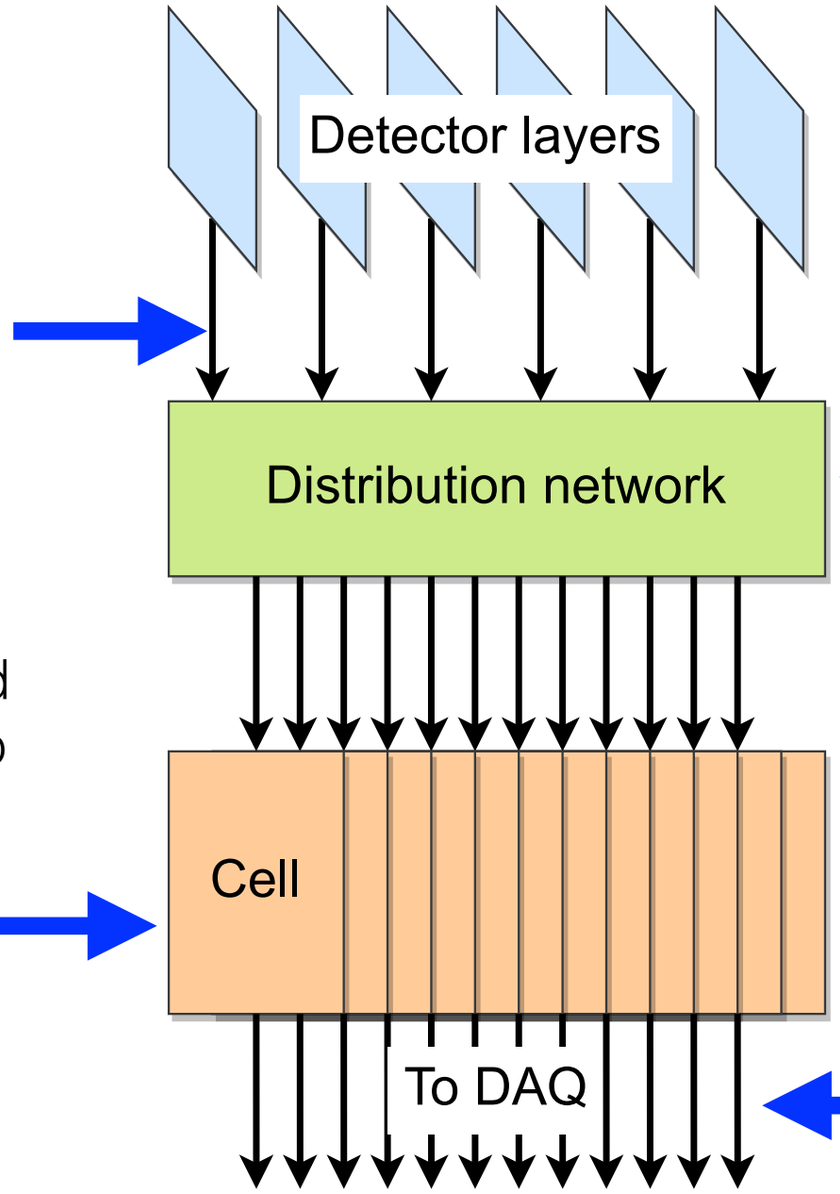



**Step 3: Find the local maxima and compute centroid**

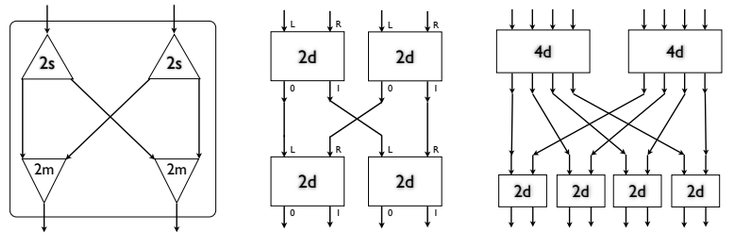


# The “artificial retina” architecture

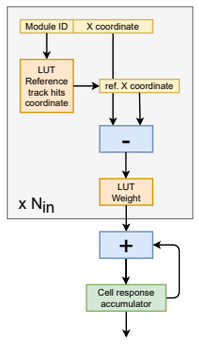
Input from detector and data preparation.

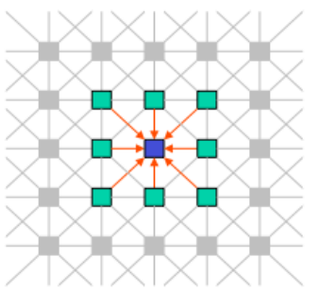



Distribution network of custom switches exchange hits among FPGAs and sends them to appropriate cells using lookup tables (allows large throughput).




Calculate cell weights and performs interpolations to extract track parameters. Engines work in parallel.



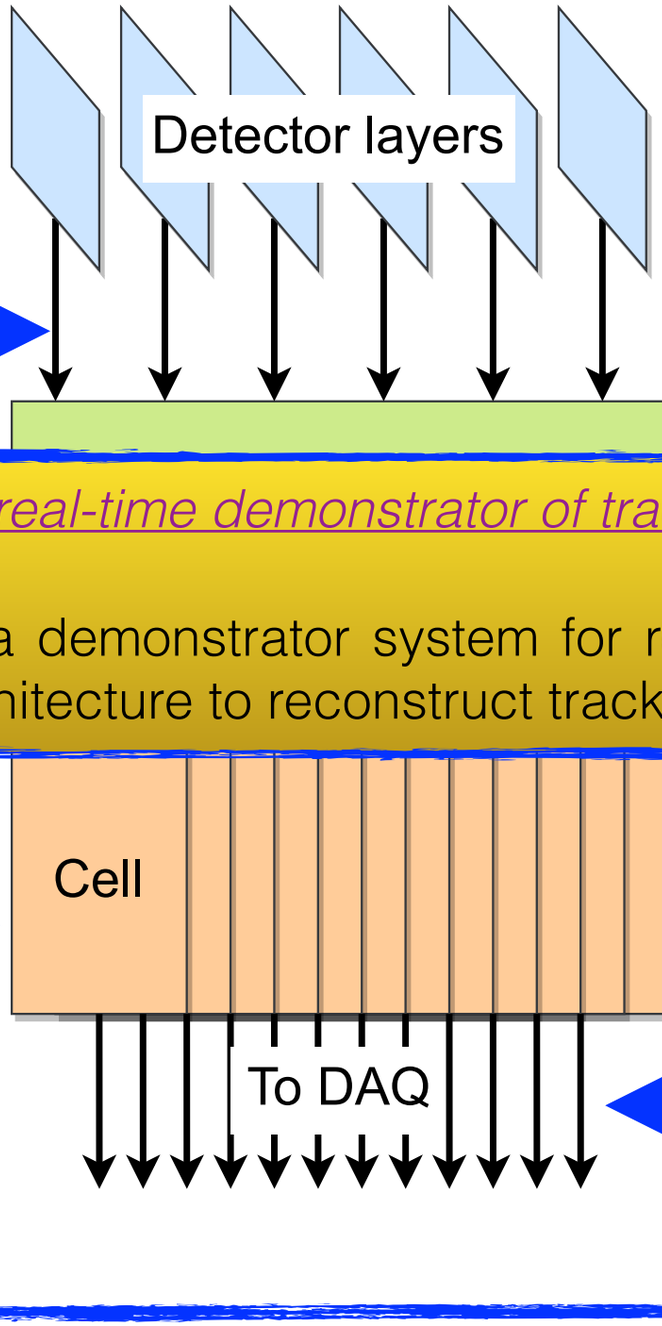



Primitive tracks are forwarded to the Event Builder, as an “embedded track-detector”, as sub-detector hits.

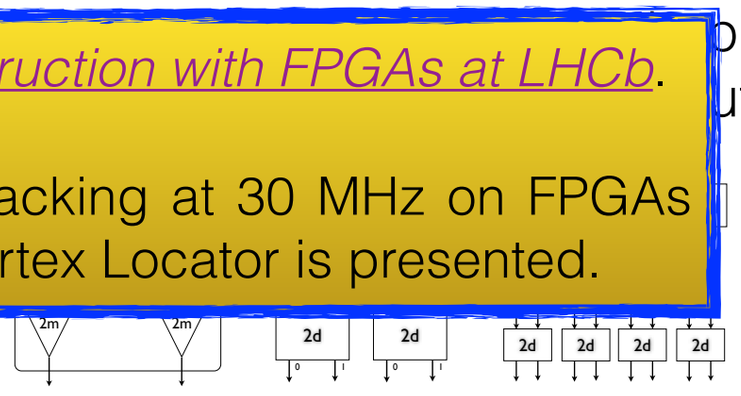
NIM A453 (2000) 425-429

# The “artificial retina” architecture

Input from detector and data preparation.

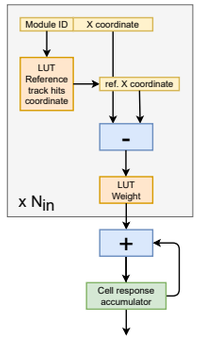


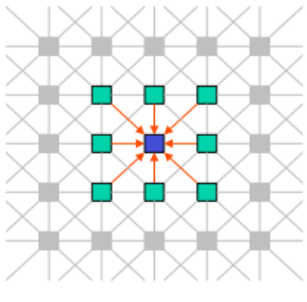

Distribution network of custom switches exchange hits among FPGAs and sends them to




ut).

Ca
pe
extract track parameters.
Engines work in parallel.



Primitive tracks are forwarded to the Event Builder, as an “embedded track-detector”, as sub-detector hits.

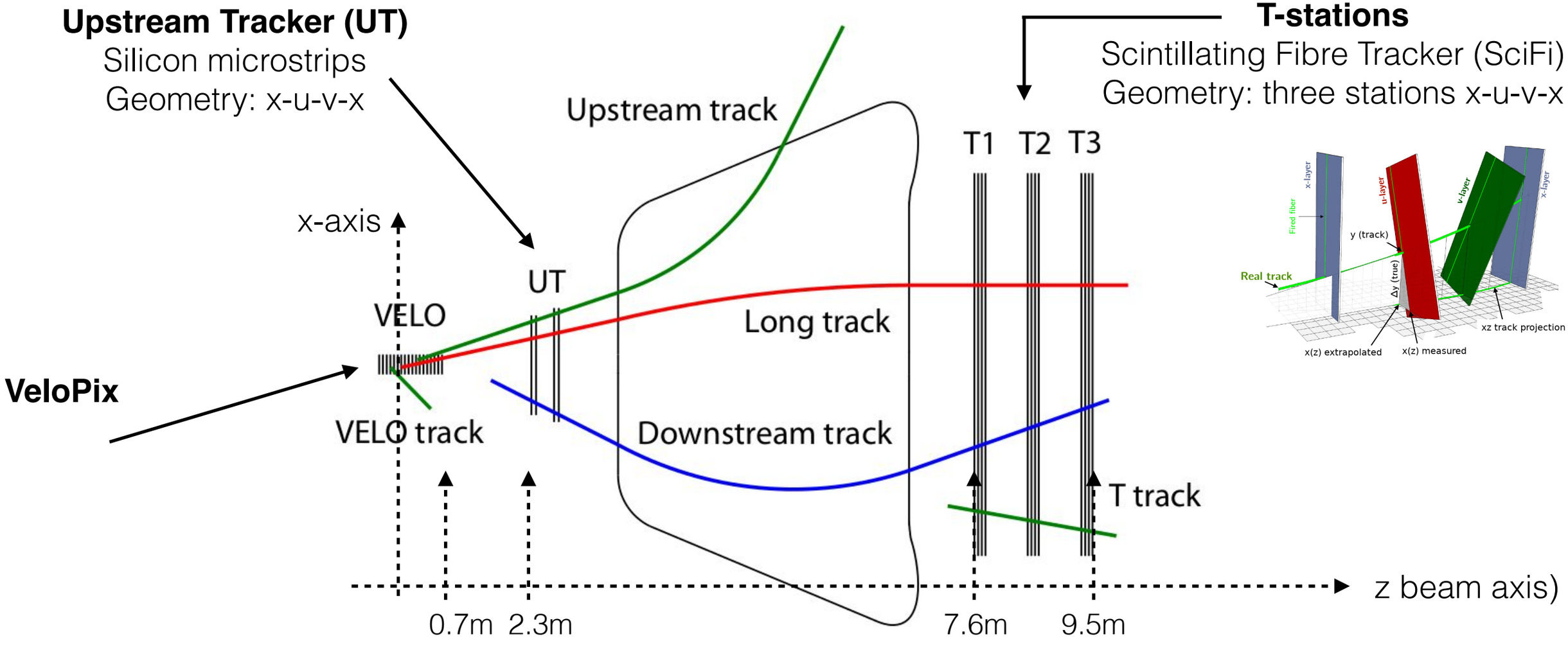
See next talk → F. Terzuoli, *A real-time demonstrator of track reconstruction with FPGAs at LHCb*.

The experience gained with a demonstrator system for real-time tracking at 30 MHz on FPGAs with the **“artificial retina”** architecture to reconstruct tracks in the Vertex Locator is presented.

# Tracking at LHCb-Upgrade I

**Upstream Tracker (UT)**
Silicon microstrips
Geometry: x-u-v-x

**T-stations**
Scintillating Fibre Tracker (SciFi)
Geometry: three stations x-u-v-x



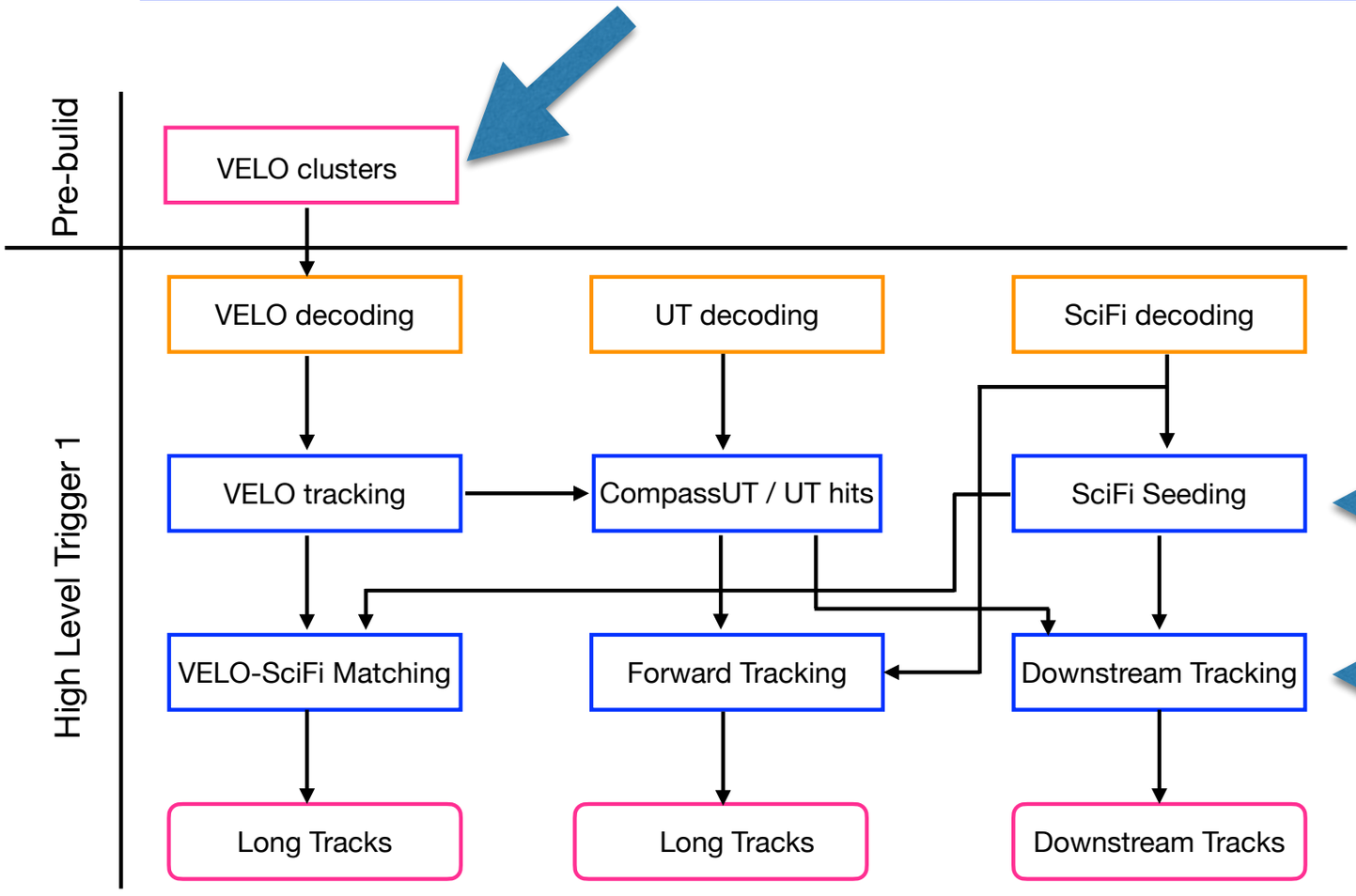




- **Long track**: reconstructible using VELO + UT + **T stations**. → beauty and charm core physics
- **Downstream track**: reconstructible using UT + **T stations**. → crucially to reconstruct for LLPs.
- **T-track**: reconstructible using **T stations**. → first stage of downstream tracks reconstruction.

# Tracking at HLT1 today

2D VELO clustering first example of RETINA pre-processing at the pre-build stage. HLT1 throughput improved by a factor >11%, with a bandwidth reduction of 14%. [*10.1109/TNS.2023.3273600*]

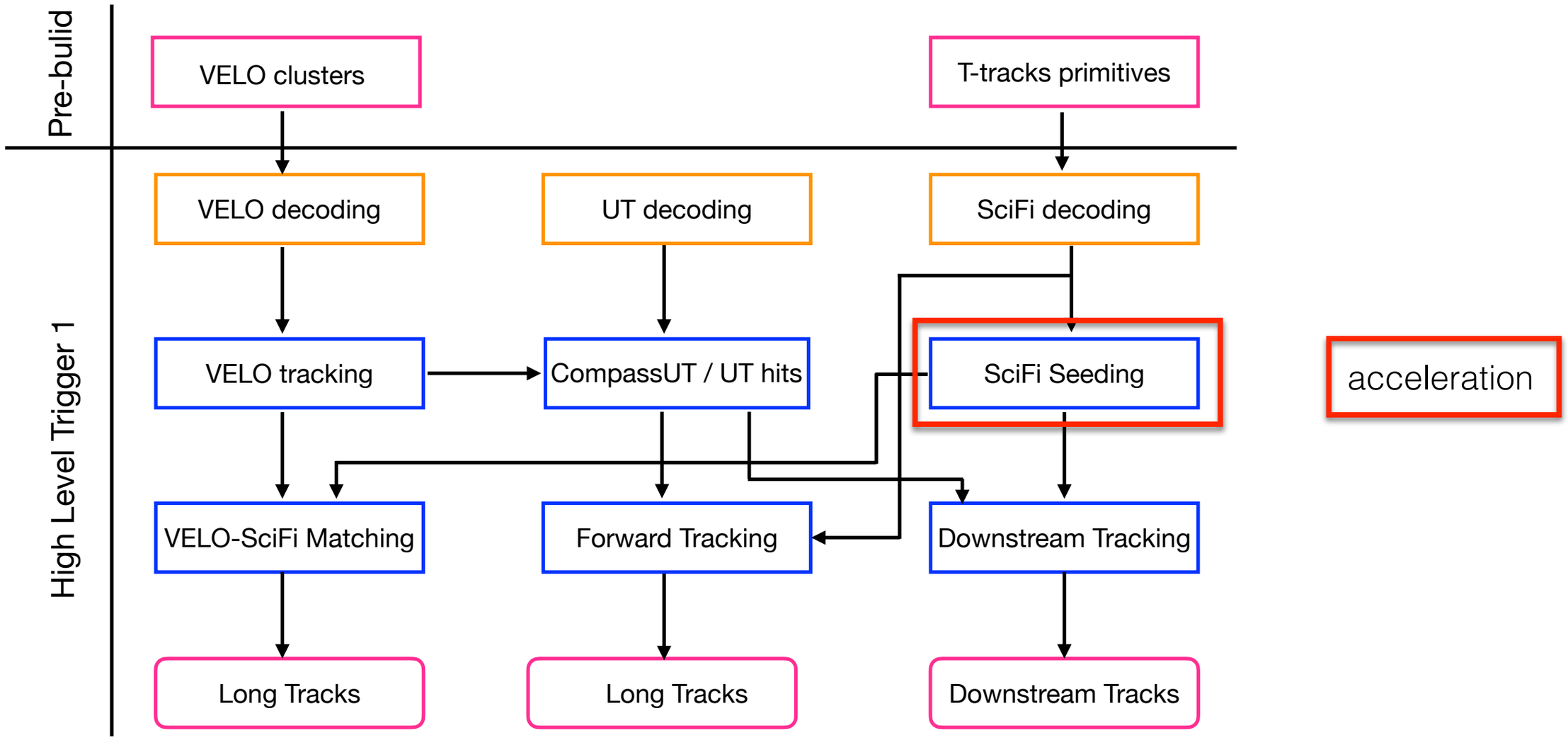



Algorithm used for reconstructing standalone T-tracks in the SciFi sub-detector. *See Louis Henry's talk.* About 35% of the entire sequence, corresponding to 2.8$\mu$s out of 8$\mu$s (125KHz) on NVIDIA RTX A5000 [*FIGURE-2022-010*].

Adding UT hits to form a "downstream track" requires additional GPU-time. Not yet in the baseline. *See Arantza De Oyanguren Campos's talk*.

→ Seeding is a very intensive pattern recognition task (high occupancy in SciFi).
→ T-track primitives essential in both in-out (**forward**) and out-in (**matching**) tracking algorithms.

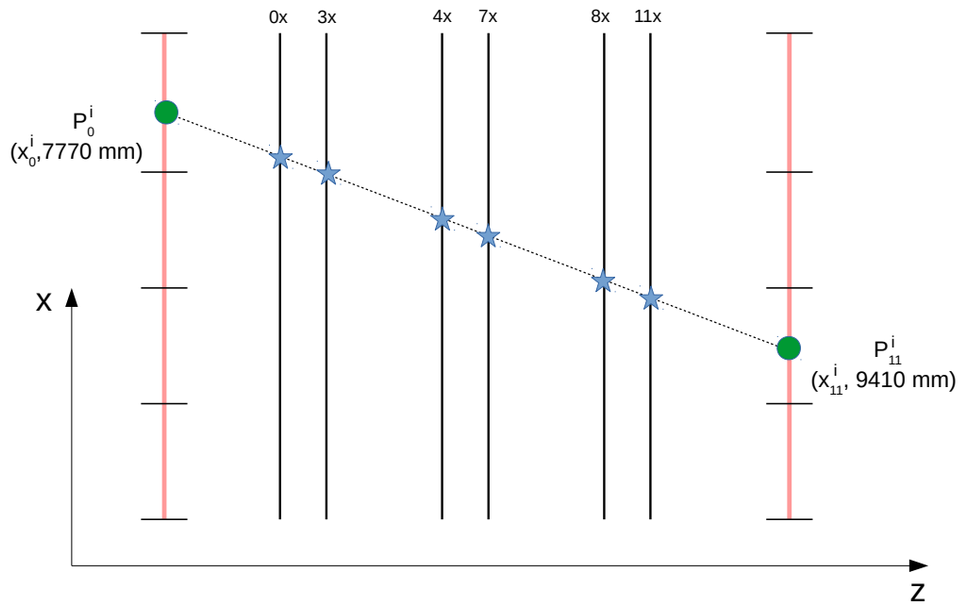
# Downstream Tracker in Run 4

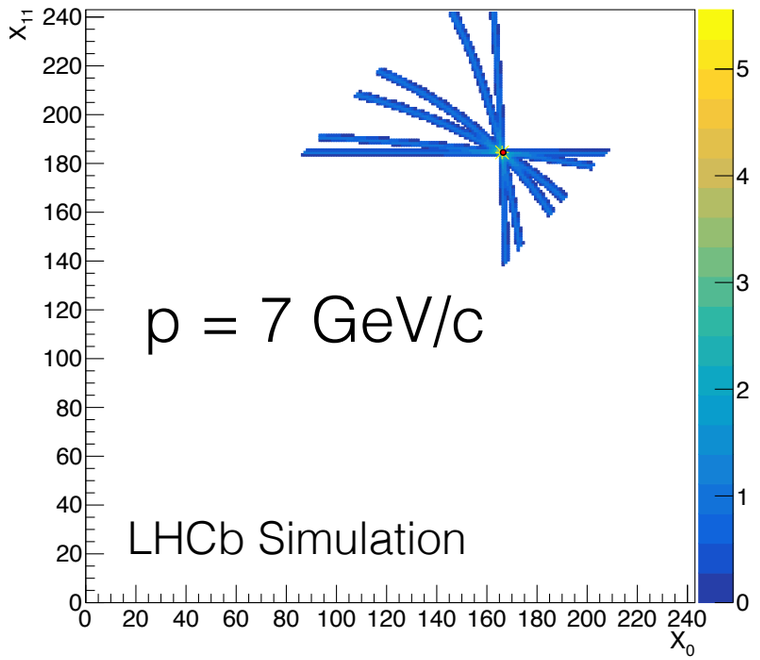



Proposal for a Downstream Tracker (DWT) RETINA-like tracking in Run4 under scrutiny at LHCb. RETINA will provide on-the-fly T-track primitives, at pre-build level, to greatly accelerate SciFi seeding tracking algorithm, while saving GPU resources for higher-level tasks. **The DWT significantly extends LHCb's physics reach for long lived particles (and not only).**

# Tracking 6-layers of SciFi

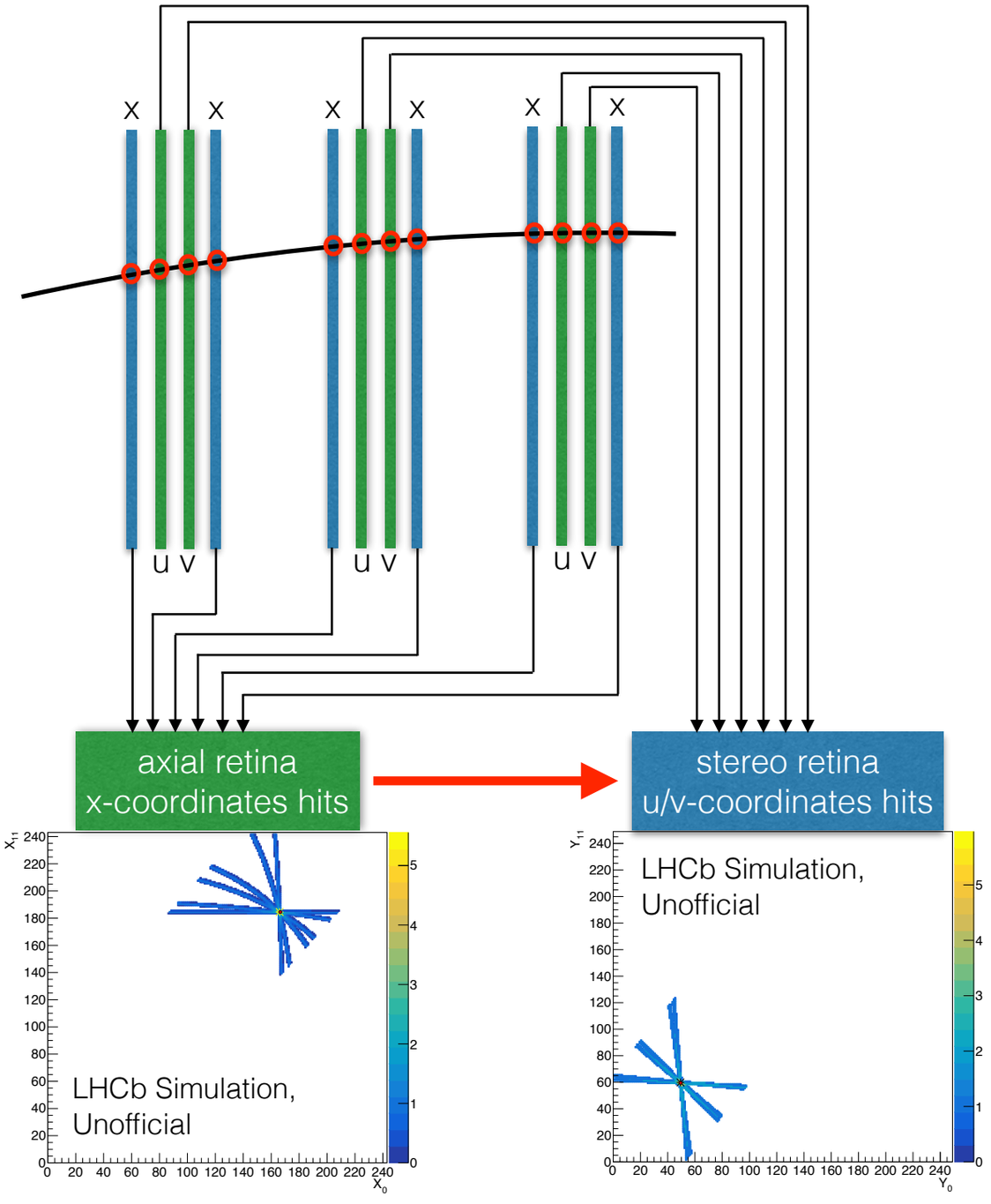
- SciFi tracks are parameterized as straight lines (Retina must be 2D):
  - $x_0$ $(x_{11})$ : x-coordinates of the intersection between the first (last) axial layer;
  - similar for the stereo association $y_0$ $(y_{11})$.
- Interesting physics tracks (Long and Downstream) distributed over the diagonal region, being $x_0 \approx x_{11}$ (and $y_1 \approx y_{11}$ )
- Typical size of a Retina system is needed, about 150k cells for the whole SciFi sub-detector (axial layers).



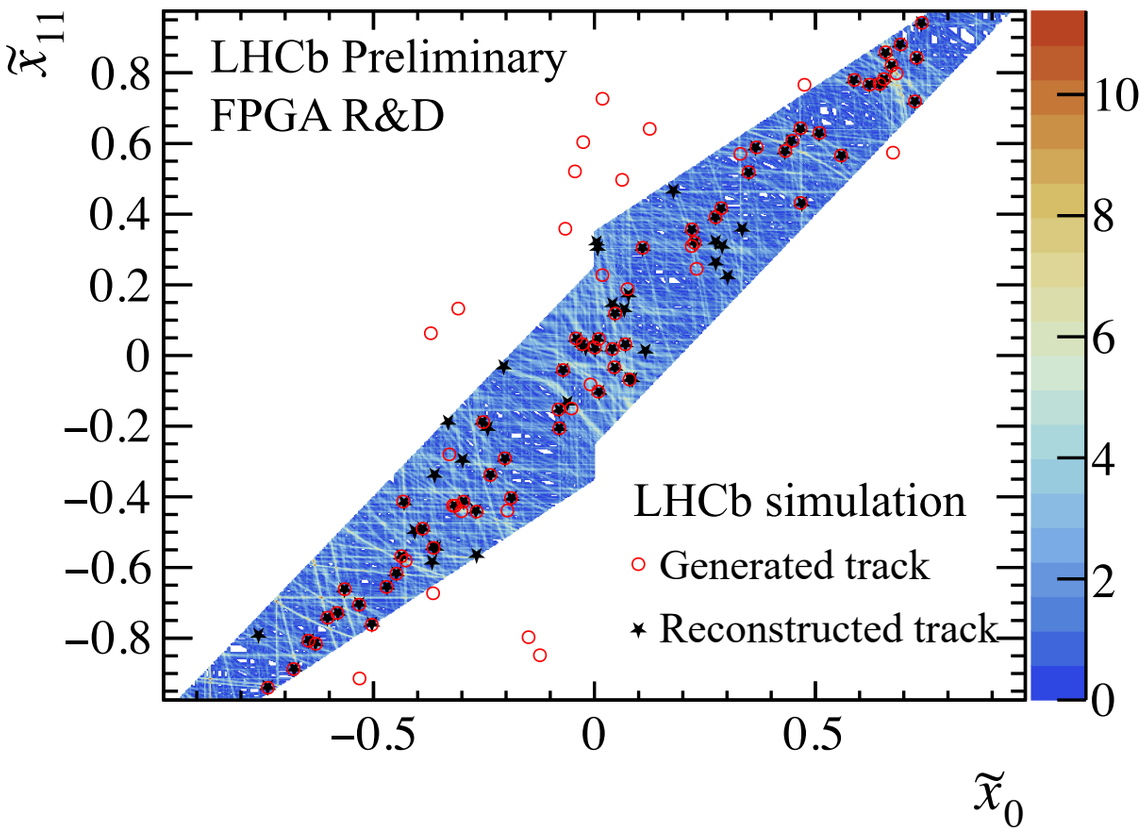
# Reconstruction of 3D T-tracks

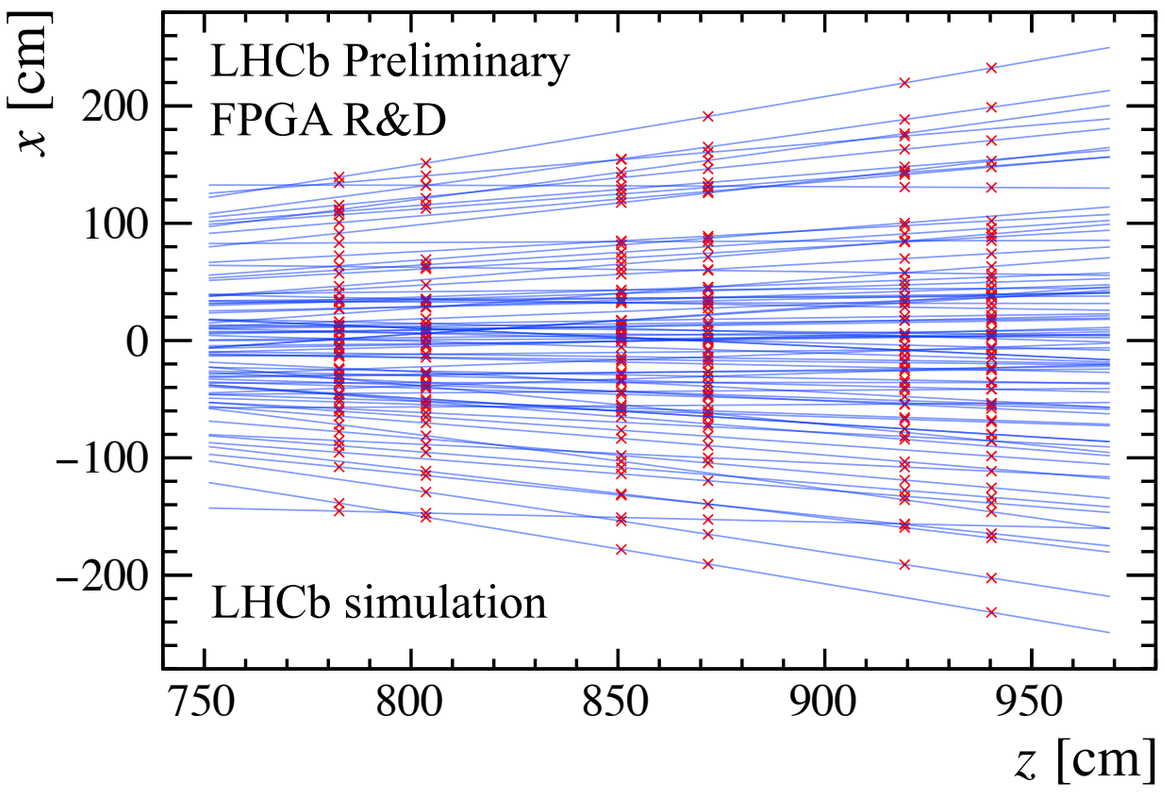
- Reconstruction of T-tracks factorized in two stages.
- **Pattern recognition:** find the x-z track projection using only axial layers.
  - tracks approximated as straight lines (2D Retina).
  - for each local maximum found, a linear $\chi^2$ fit to a parabola is executed (on DSP blocks of FPGAs) in order to kill ghost tracks and evaluate parabola parameters.
- **Stereo association**: x-z projection of track candidate is used as "seed" to extract y-coordinates from u/v layers and associate y-z track projection. **Still in progress (not presented here).**



# An example of axial retina

*LHCB-FIGURE-2023-027*




This is just a top half of the SciFi corresponding to 73k cells of granularity.

# First look at physics performance

- LHCb GEANT-based SciFi simulation, with simulated realistic pp collisions at the LHCb-Upgrade I conditions ($\nu$ = 7.6).
- **At this stage working point chosen to have 90% of efficiency for generic long tracks with p > 5GeV.**
- Efficiencies ‘comparable' with the CPU-HLT2 Hybrid Seeding [*ref*] and GPU-HLT1 Standalone Seeding [*ref*].
- Ghost rate is about 50% (about 1 fake track for each real track), to be compared with 22% of Allen-HLT1 (axial-only).
- Stereo information from u- and v-coordinate hits not yet included. Performance will benefit by using that. **Trade-off with GPUs under investigation.**

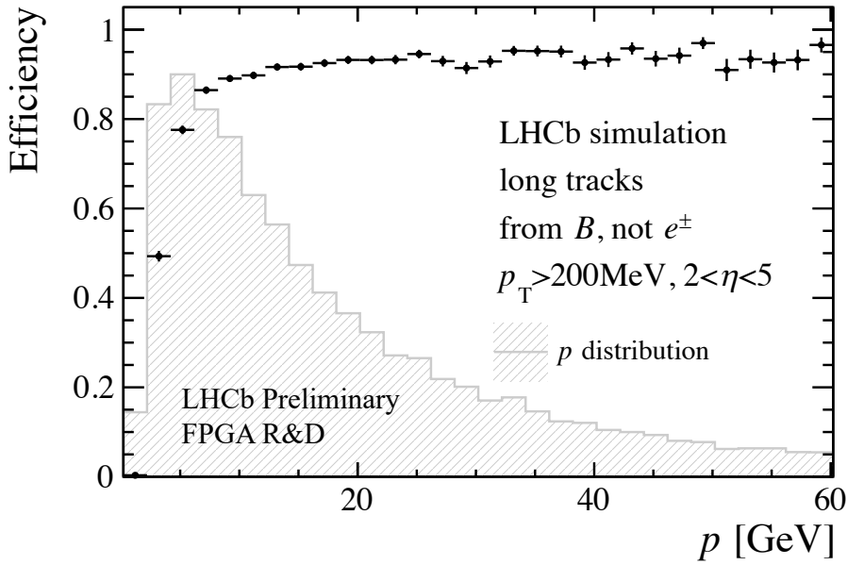
Table 1: Axial reconstruction efficiencies for different simulated samples and different track categories. The ghost rate is also shown. Event-averaged values are shown in brackets. The physics fiducial requirements $p_{\mathrm{T}} > 200\,\mathrm{MeV}/c$ and $2 < \eta < 5$ are applied.

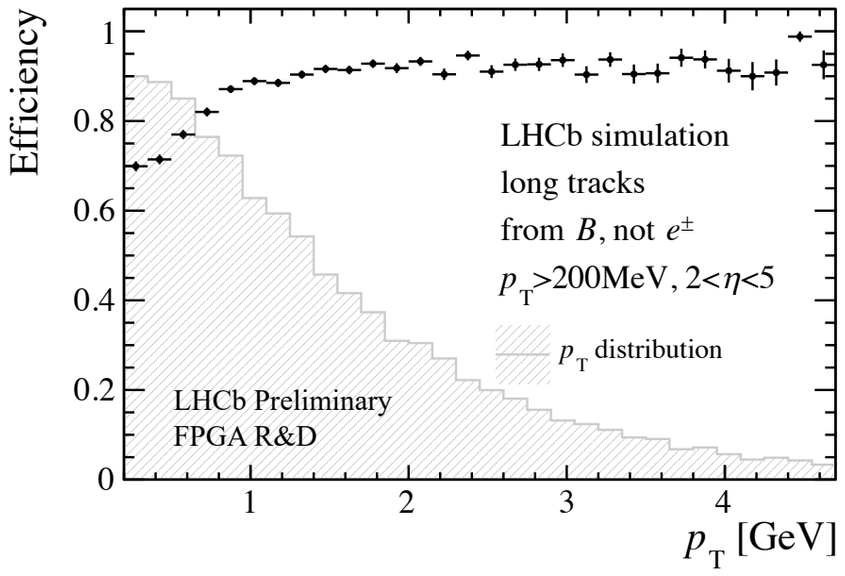
| Track type | MinBias [%] | $D^0 \to K^0_{\mathrm{S}}\pi^+\pi^-$ [%] | $B^0_s \to \phi\phi$ [%] |
|---|---|---|---|
| T-track | 71 (72) | 70 (71) | 70 (71) |
| T-track, $p > 3\,\mathrm{GeV}$ | 83 (84) | 81 (82) | 82 (83) |
| T-track, $p > 5\,\mathrm{GeV}$ | 89 (90) | 88 (89) | 88 (88) |
| Long | 77 (79) | 76 (77) | 76 (77) |
| Long, $p > 3\,\mathrm{GeV}$ | 85 (86) | 83 (84) | 84 (84) |
| Long, $p > 5\,\mathrm{GeV}$ | 90 (91) | 89 (90) | 88 (89) |
| Long from $B$ not $e^\pm$, $p > 3\,\mathrm{GeV}$ | - | - | 88 (87) |
| Long from $B$ not $e^\pm$, $p > 5\,\mathrm{GeV}$ | - | - | 91 (90) |
| Down | 75 (76) | 74 (75) | 75 (75) |
| Down, $p > 3\,\mathrm{GeV}$ | 84 (85) | 82 (83) | 83 (84) |
| Down, $p > 5\,\mathrm{GeV}$ | 89 (91) | 88 (89) | 88 (89) |
| Down from strange not $e^\pm$, $p > 3\,\mathrm{GeV}$ | - | 82 (82) | - |
| Down from strange not $e^\pm$, $p > 5\,\mathrm{GeV}$ | - | 88 (89) | - |
| Down from strange not long not $e^\pm$, $p > 3\,\mathrm{GeV}$ | - | 82 (82) | - |
| Down from strange not long not $e^\pm$, $p > 5\,\mathrm{GeV}$ | - | 87 (88) | - |
| ghost rate [%] | 49 (40) | 52 (47) | 53 (47) |
| ghost rate / (1 - ghost rate) | 0.9 (0.7) | 1.1 (0.9) | 1.1 (0.9) |

*LHCB-FIGURE-2023-027*

# FPGA primitives into GPU-HLT1

- **Main DWT purpose is to provide good quality primitives to Allen-HLT1 and speed up the standalone T-track reconstruction. HLT2 will also accelerate, since primitives can be re-used there.**
- DWT already has a very good physics performance for being a co-processor.
- Ghost rate is under control. It will be absorbed by the GPU track refitting and processing at full precision.
- Optimal working point (and many other features of the algorithm) to be chosen when T-track primitives are plugged into Allen and used as ‘seeds’ in the GPU-seeding algorithm.

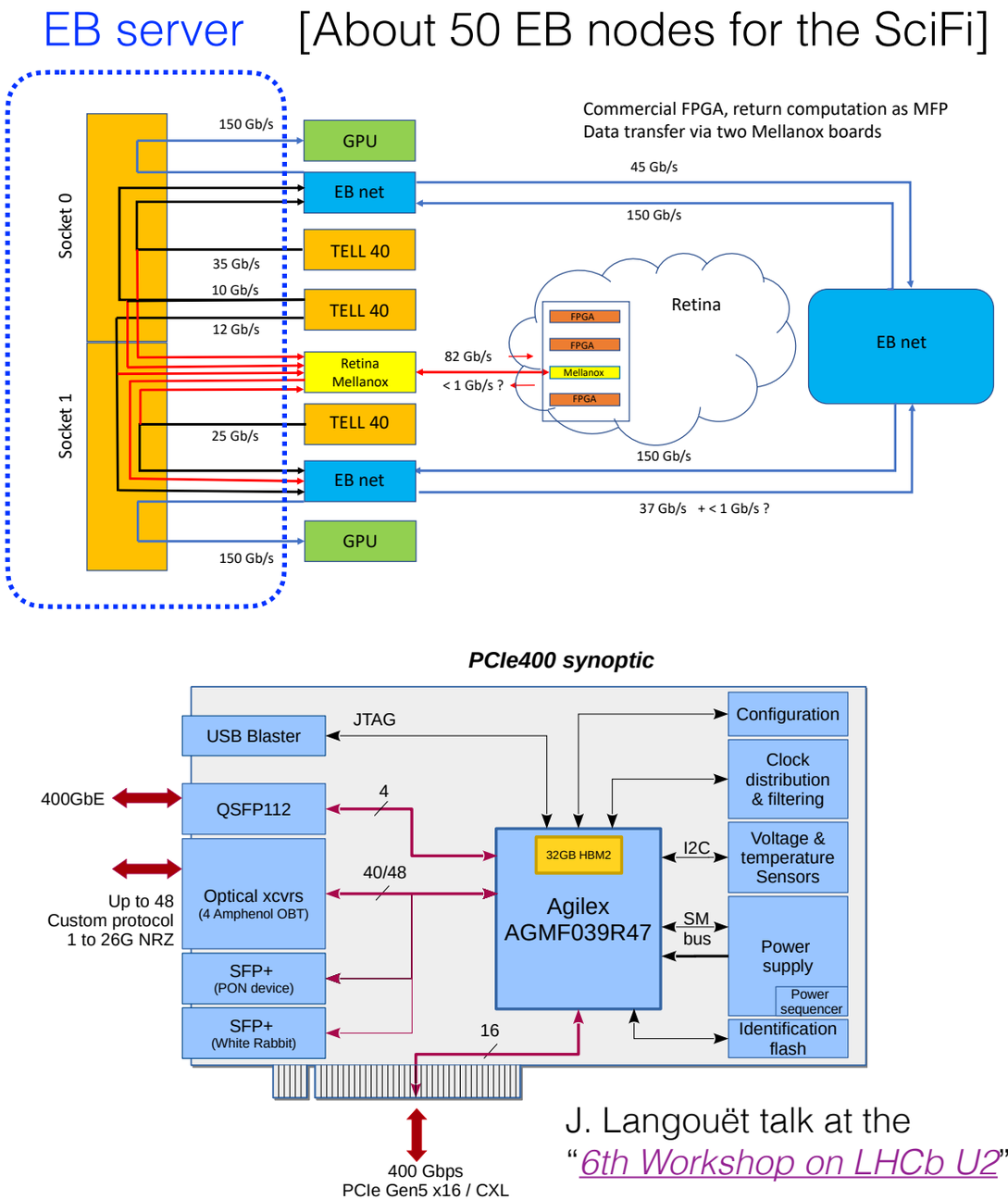



LHCB-FIGURE-2023-027

# Resource usage and integration

- DAQ integration is crucial, strong constraints from available servers/PCIe slots/bandwidth, several options under discussion to minimize impact on operations and maximize performance.
- A cell-engine for processing a 1D SciFi hit with all functionalities, requires no more than 1000 LEs. [cell for stereo association simpler < 500LEs].
- DWT is modular and can be implemented on 64 FPGA boards for the axial reconstruction and 32 boards for the stereo association.
- **About 2.8 × $10^6$ LEs for each tracking board.**
- Commercial FPGAs chips with similar or larger number of LEs are already available on the shelves, such as the Intel Agilex 7 AGM039 (3.85 x$10^6$ LEs) envisioned for the PCIe400 boards.




J. Langouët talk at the "*6th Workshop on LHCb U2*"

# Conclusions

- The Retina DWT is a co-processor running at the pre-build stage at 30 MHz with no time-multiplexing.
- It aims at providing T-track primitives of ‘very good’ quality to Allen-HLT1 (and HLT2) and accelerate Seeding and Downstream tracking.
- Projected physics performance, for a FPGA coprocessor of reasonable size, is already excellent. Work is ongoing to integrate primitives into Allen-HLT1 and precisely quantify all benefits.
- The DWT is the first real-life test bed for this new architecture and will pave the way for an even better integrated heterogeneous system (with much greater acceleration) in view of the challenges of Upgrade II.

# BACKUP

# The LHCb-RETINA team

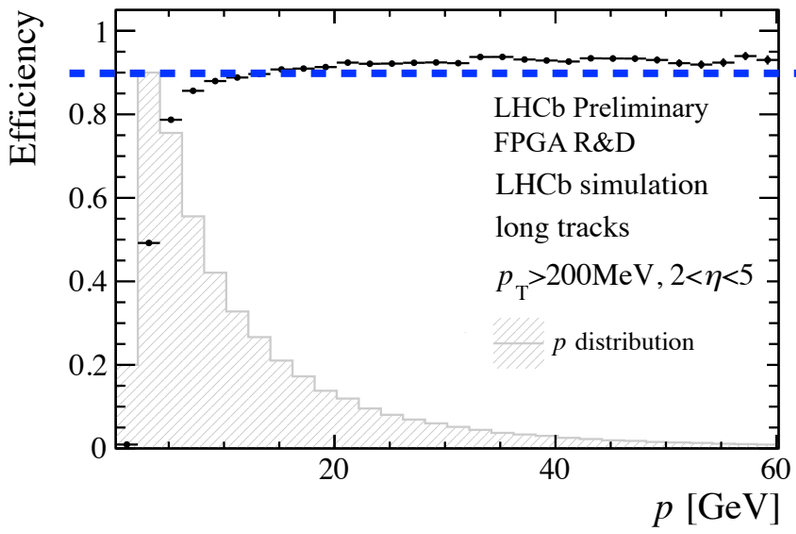
Wander Baldini[1,2], Giovanni Bassi[3,4], Andrea Contu[5], Riccardo Fantechi[3], Jibo He[6,7], Brij Kishor Jashal[8], Sofia Kotriakhova[1,9], Federico Lazzari[3,10], Maurizio Martinelli[11,12], Diego Mendoza[8], Michael J. Morello[3,4], Arantza De Oyanguren Campos[8], Lorenzo Pica[3,4], Giovanni Punzi[3,10], Qi Shi[6], Francesco Terzuoli[3,13], Giulia Tuci[14], Ao Xu[3], Jiahui Zhuo[8]

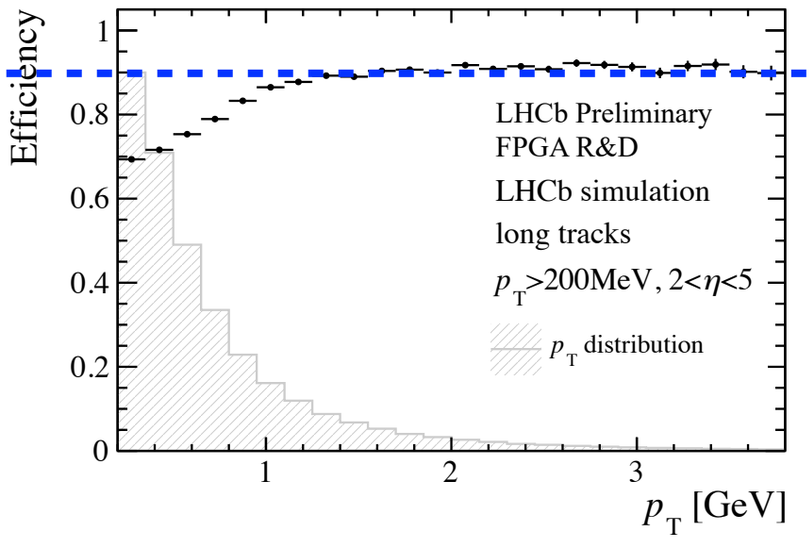
[1] *INFN Sezione di Ferrara, Ferrara, Italy*
[2] *European Organization for Nuclear Research (CERN), Geneva, Switzerland*
[3] *INFN Sezione di Pisa, Pisa, Italy*
[4] *Scuola Normale Superiore, Pisa, Italy*
[5] *INFN Sezione di Cagliari, Monserrato, Italy*
[6] *University of Chinese Academy of Sciences, Beijing, China*
[7] *Hangzhou Institute for Advanced Study, UCAS, Hangzhou, China*
[8] *Instituto de Fisica Corpuscular, Centro Mixto Universidad de Valencia - CSIC, Valencia, Spain*
[9] *Università di Ferrara, Ferrara, Italy*
[10] *Università di Pisa, Pisa, Italy*
[11] *INFN Sezione di Milano-Bicocca, Milano, Italy*
[12] *Università di Milano Bicocca, Milano, Italy*
[13] *Università di Siena, Siena, Italy*
[14] *Physikalisches Institut, Ruprecht-Karls-Universitat Heidelberg, Heidelberg, Germany*

# Abstract

Finding track segments downstream of the magnet is some of the most important and computationally expensive task of the first stage of the new GPU-based software trigger of the LHCb Upgrade I, that has started operation in Run 3. These segments are essential to form all good physics tracks with a very high precision momentum measurement, when combined with those reconstructed in the vertex track detector, and to reconstruct long-lived particles, such as Kshort and strange baryons, decaying after the vertex track detector, largely boosting the physics reach of the experiment. In this talk, we discuss the collaboration plans to install a real-time tracking device based on distributed system of FPGAs, dedicated to the reconstruction of particles trajectories in the forward Scintillating Fibre tracker detector, with the aim to preserve the full physics potential of the experiment in Run 4, and in view of Run 5 (Upgrade II) at higher instantaneous luminosity. This system will enhance the DAQ system of the experiment, and will run in real time during physics data taking, reconstructing tracks on-the-fly at the LHC collision rate, before the software trigger processing begins. The design and the expected performance of this device, with the capability of processing events at the full LHC collision rate of 30 MHz is discussed. Proposed by Michael J. Morello <michael.joseph.morello@cern.ch>

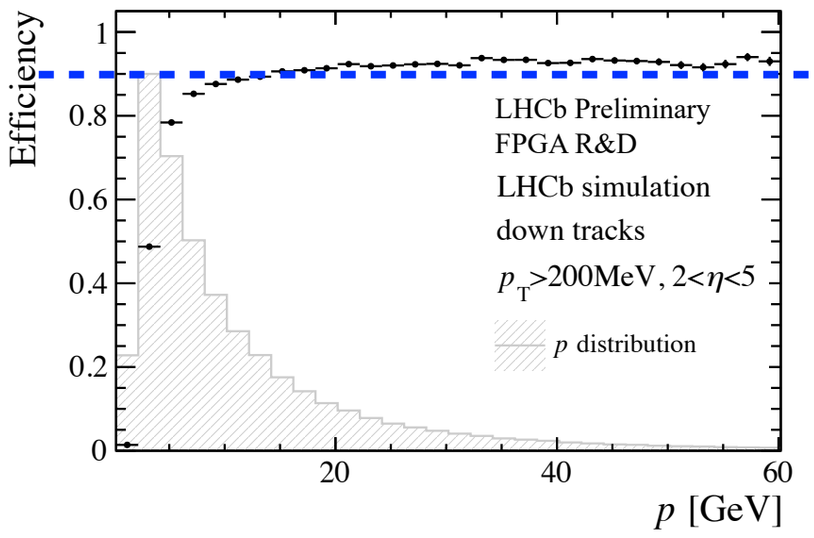
# DWT: physics performance

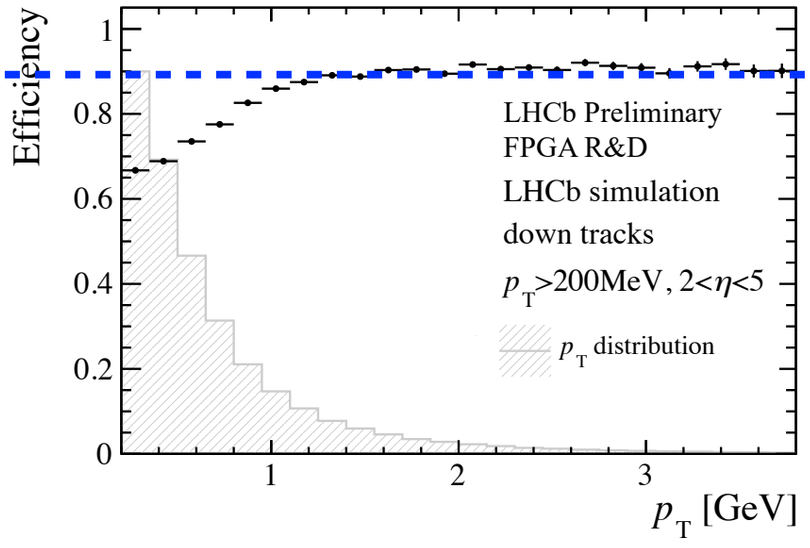



90%

T-tracks reconstructible as long track



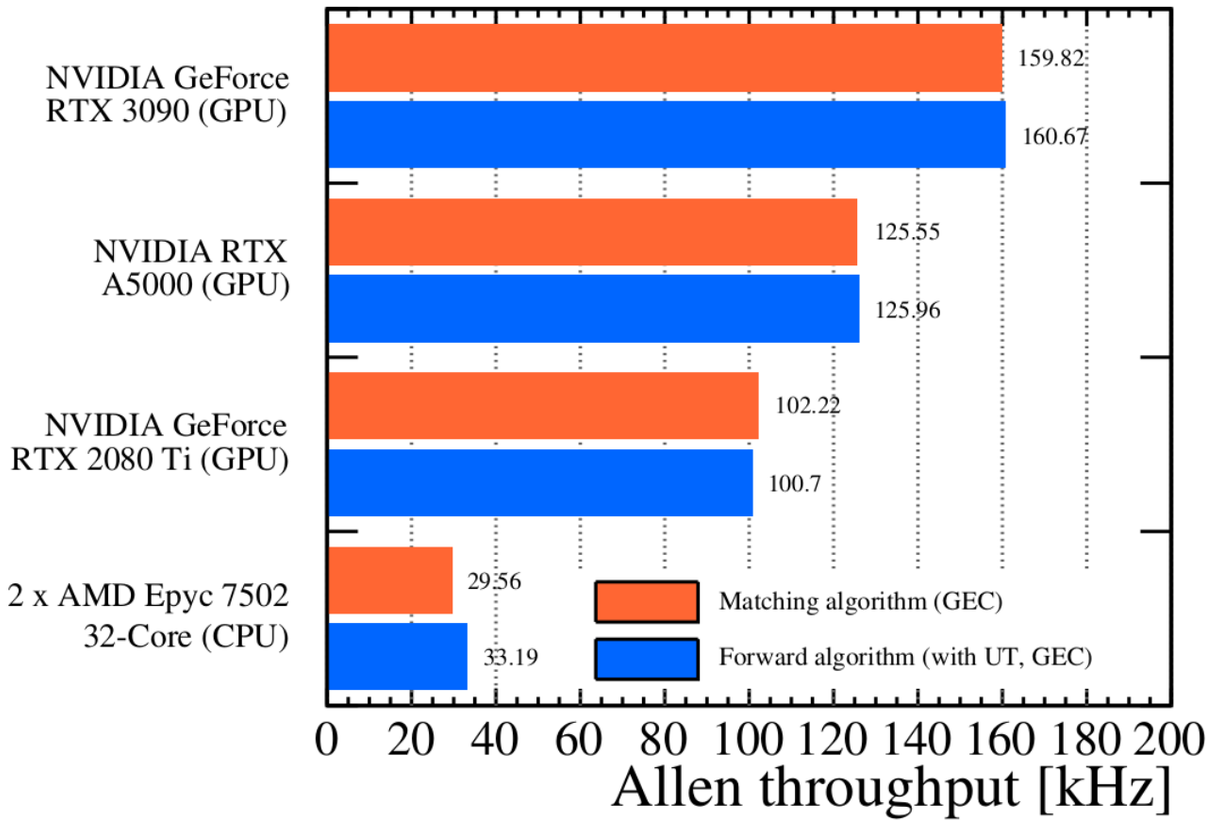



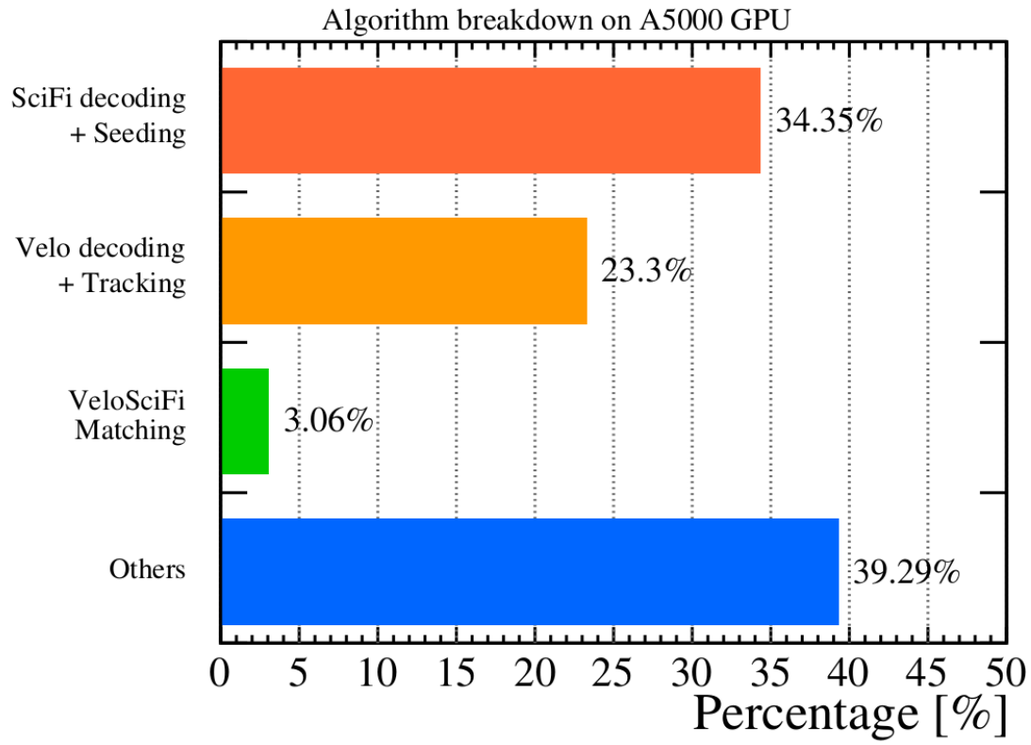

90%

T-tracks reconstructible as downstream track

# Seeding HLT1

See *L. Henry's talk at CTDs 2023*



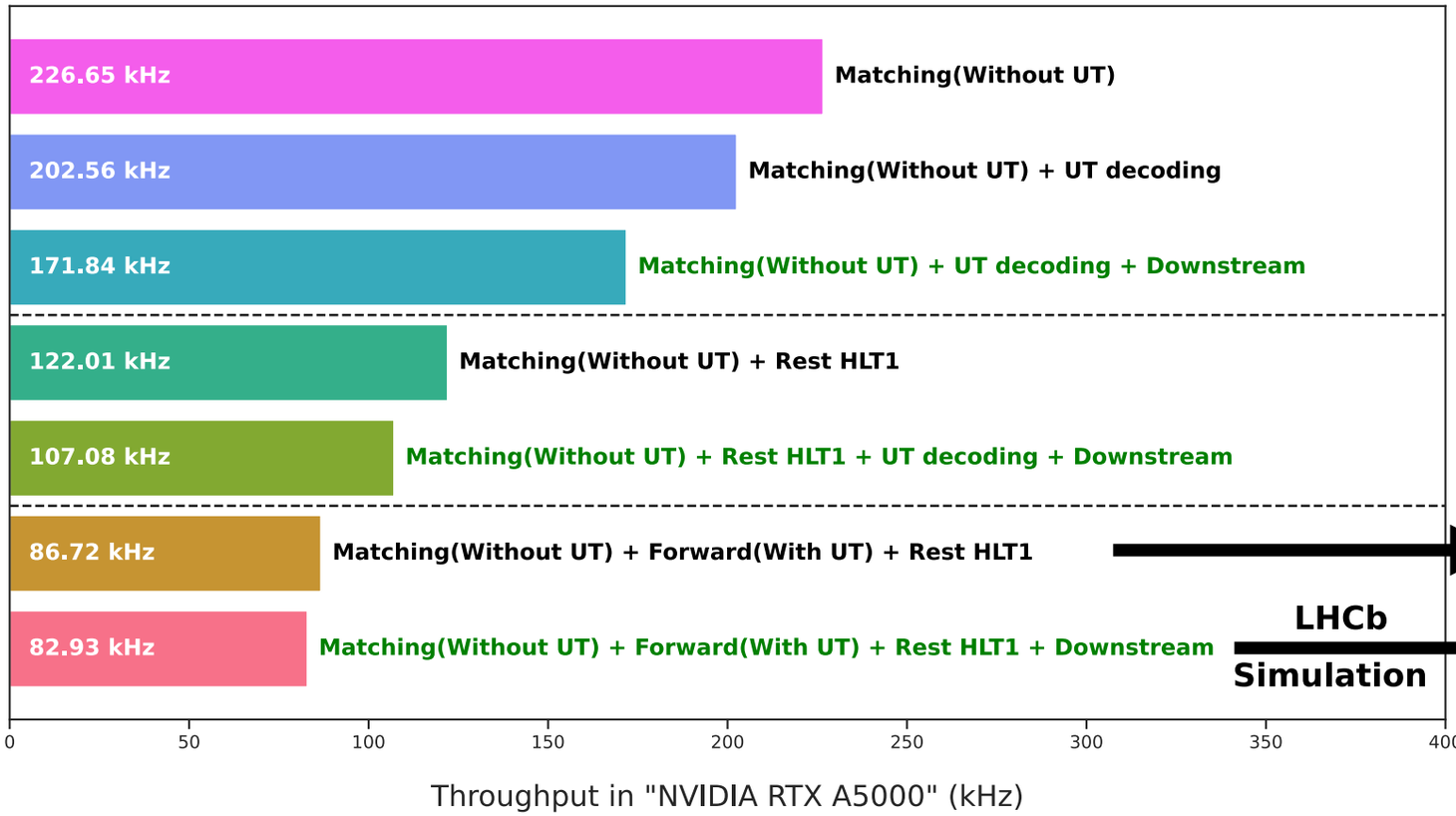




On NVIDIA RTX A5000 the entire HLT1 sequence takes ~8$\mu$s (on simulation).

SciFi decoding and Seeding takes 35% → ~2.8$\mu$s (VELO tracking takes ~1.8$\mu$s where ~0.6$\mu$s already removed thanks to the FPGA 2D clustering)

# Throughput Allen-HLT1

*LHCb-FIGURE-2023-028*




~2x170 = 340 GPUs
installed in the EB nodes

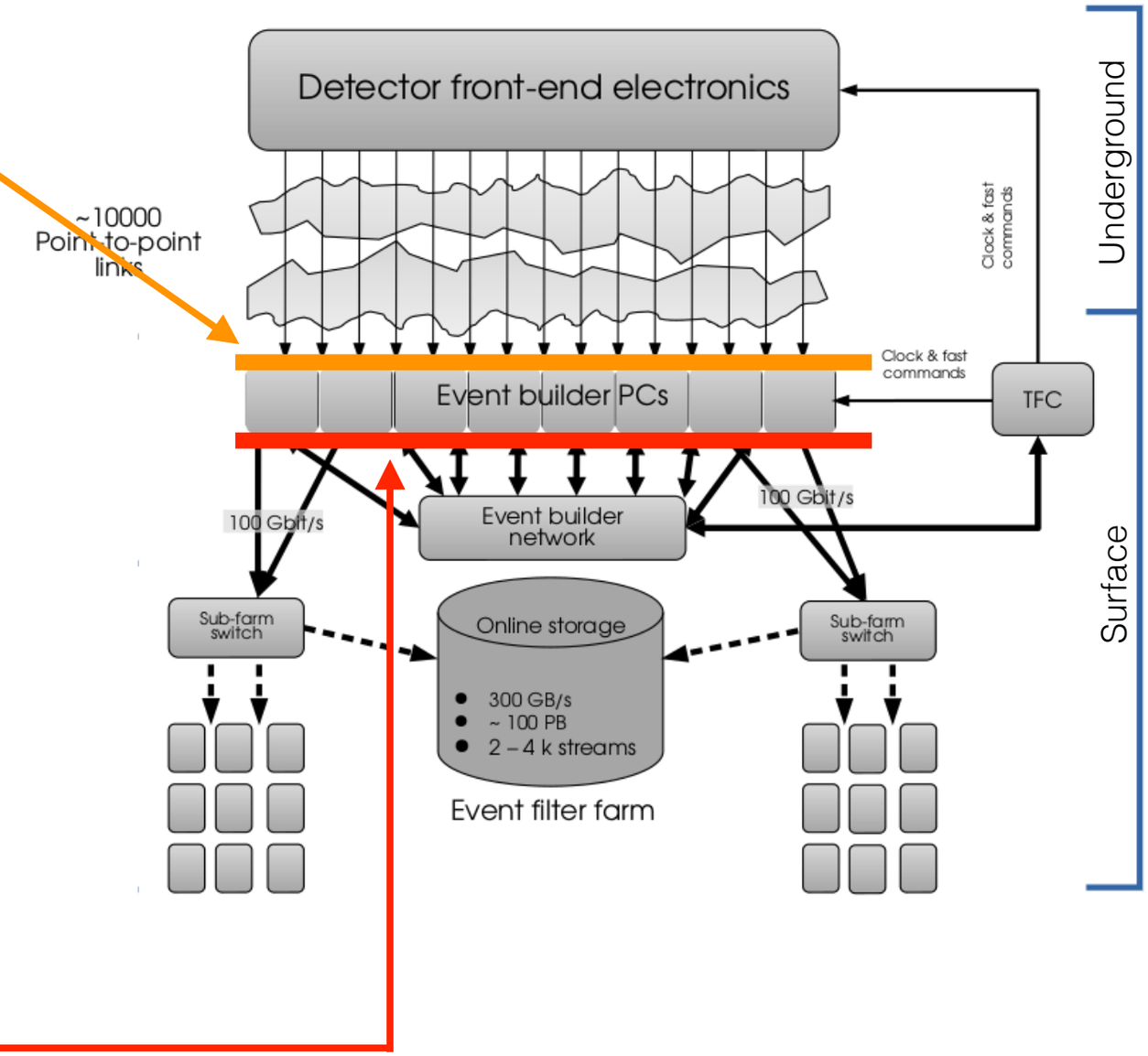
86.7kHz x 340= 29.4MHz

82.9kHz x 340= 28.1MHz

Figure 1: Through of the Allen sequence including the `Downstream` algorithm, executed on a RTA A5000 card using simulated MinBias sample with nominal Run3 conditions. The global effect of the inclusion of this new algorithm is 3 kHz. This can be shown by comparison of the last two boxes in the figure.

# DAQ integration (pre-build)

- Plan is to integrate at **Pre-Build** level. (effectively part of the readout)
- Many advantages: modularity, not necessary unpack the event, can reduce the data flow into the EB, appears as "virtual detector" producing ready-made tracks,…
- HLT1-Allen (~ 2x173 GPUs) is now running on GPUs installed in the EB, operating in the **Post-Build** level.



# Scintillating Fibre Tracker (SciFi)

- 3 tracking stations (T1,T2,T3) of scintillating fibre.
- 4 layers per station (x-u-v-x)
  - u/v layers tilted by a stereo angle of +5°/-5°.
  - Electronic readout at 40MHz.
  - Hit spatial resolution: ~100 μm.
  - High occupancy: an average of about 300 hits per layer, up to a maximum of 800 hits per layer.
- A small component of magnetic field (fringe field) is present in the SciFi region. Tracks are well approximated as parabola in x-z view, and as straight lines in y-z view.
- For this study SciFi divided in 4 independent quadrants. Negligible loss of efficiency because of tracks crossing different quadrants.