Track Reconstruction with Graph Neural Networks on Heterogeneous Architectures

Fotis Giasemis on behalf of the SMARTHEP/ANN4EUROPE team
Biennale LPNHE 2024

LHCb Detector and Dataflow

Velo
Vertex **Lo**cator
With silicon pixels
No magnetic field

UT
Upstream **T**racker
With silicon strips

SciFi
With **Sci**ntillating **Fi**bres

J. Phys.: Conf. Ser., vol. 878, p. 012012, 2017

All numbers related to the dataflow are taken from the LHCb

Upgrade Trigger and Online TDR
Upgrade Computing Model TDR
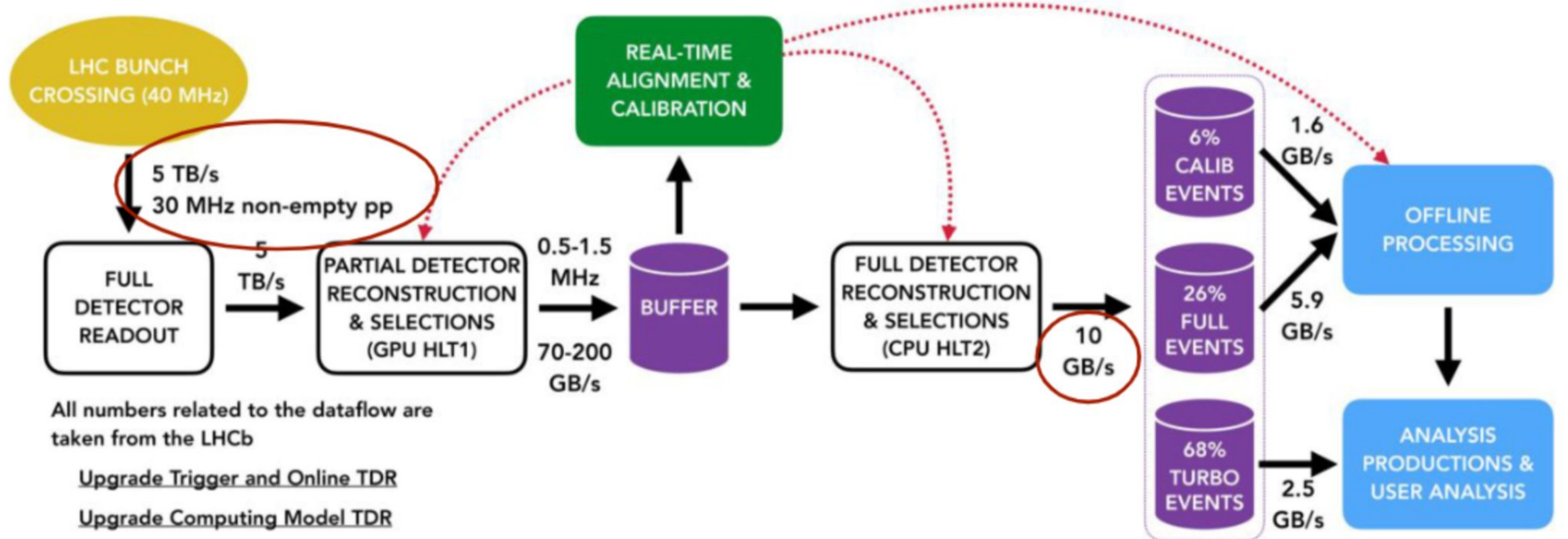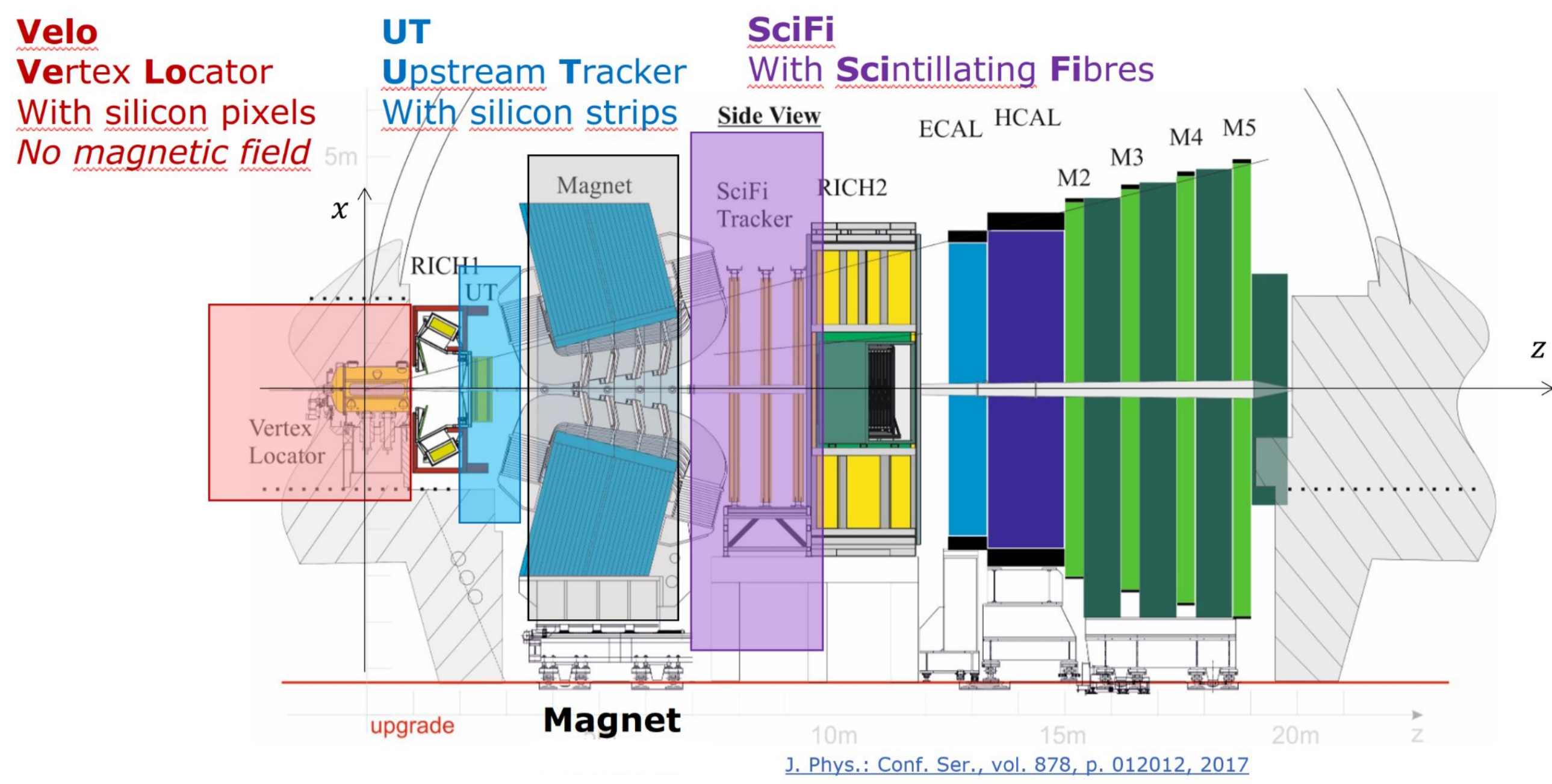
First Level Trigger (HLT1)

LHCb moved the first level trigger from CPU to GPU. This is the first time in high-energy physics that GPU processors have been used for a first-level trigger system.

Compared to the CPU baseline the GPU trigger allowed numerous additional physics functionality to be implemented: calorimenter reconstruction, low-momentum tracking, as well as finding tracks produced outside the LHCb vertex detector.

Implementing this trigger on 340 GPU processors also resulted in very substantial cost savings with respect to the CPU baseline system.
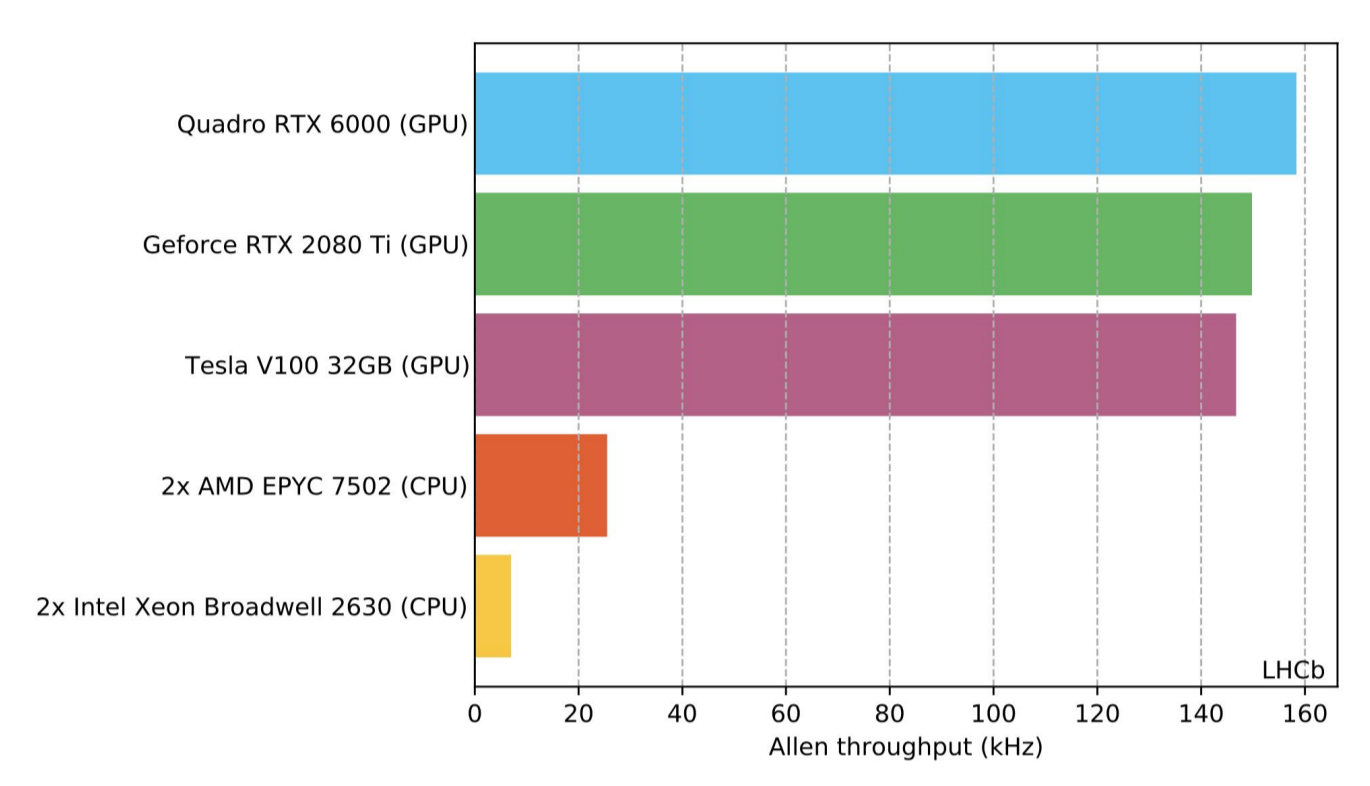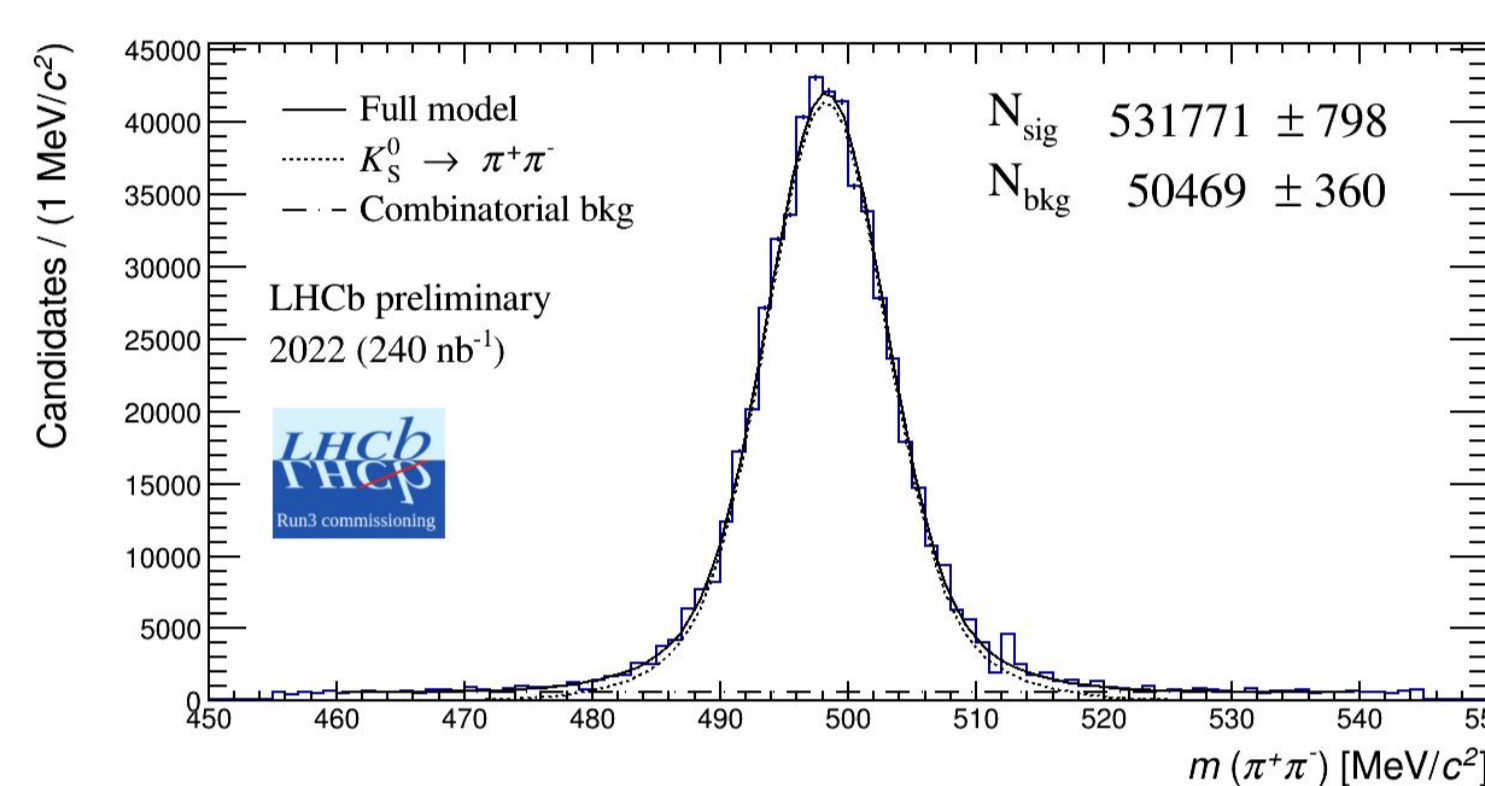
## ETX4VELO: Graph Neural Network-Based Pipeline for Track Finding in the Velo at LHCb

The focus is to evaluate deep-learning algorithms performance for EFFICIENCY and THROUGHPUT, and estimate how these models scale up with the increase of data rate.

For this purpose we developed the ETX4VELO pipeline which focuses on developing Graph Neural Networks (GNNs) algorithms for track reconstruction.
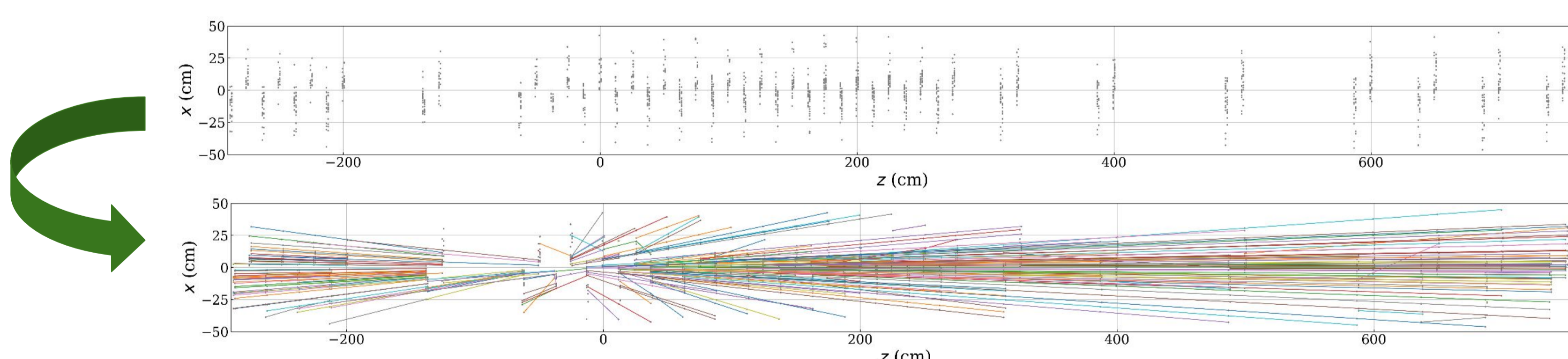
Steps of the pipeline:
- Embedding
- Graph Construction
- GNN
- Triplets (not shown in the diagram)
- Score cut

The efficiency reached at this level is higher than Allen. The ETX4VELO pipeline is based on edge and triplet GNN model that can reconstruct harsh cases of shared hits between tracks.
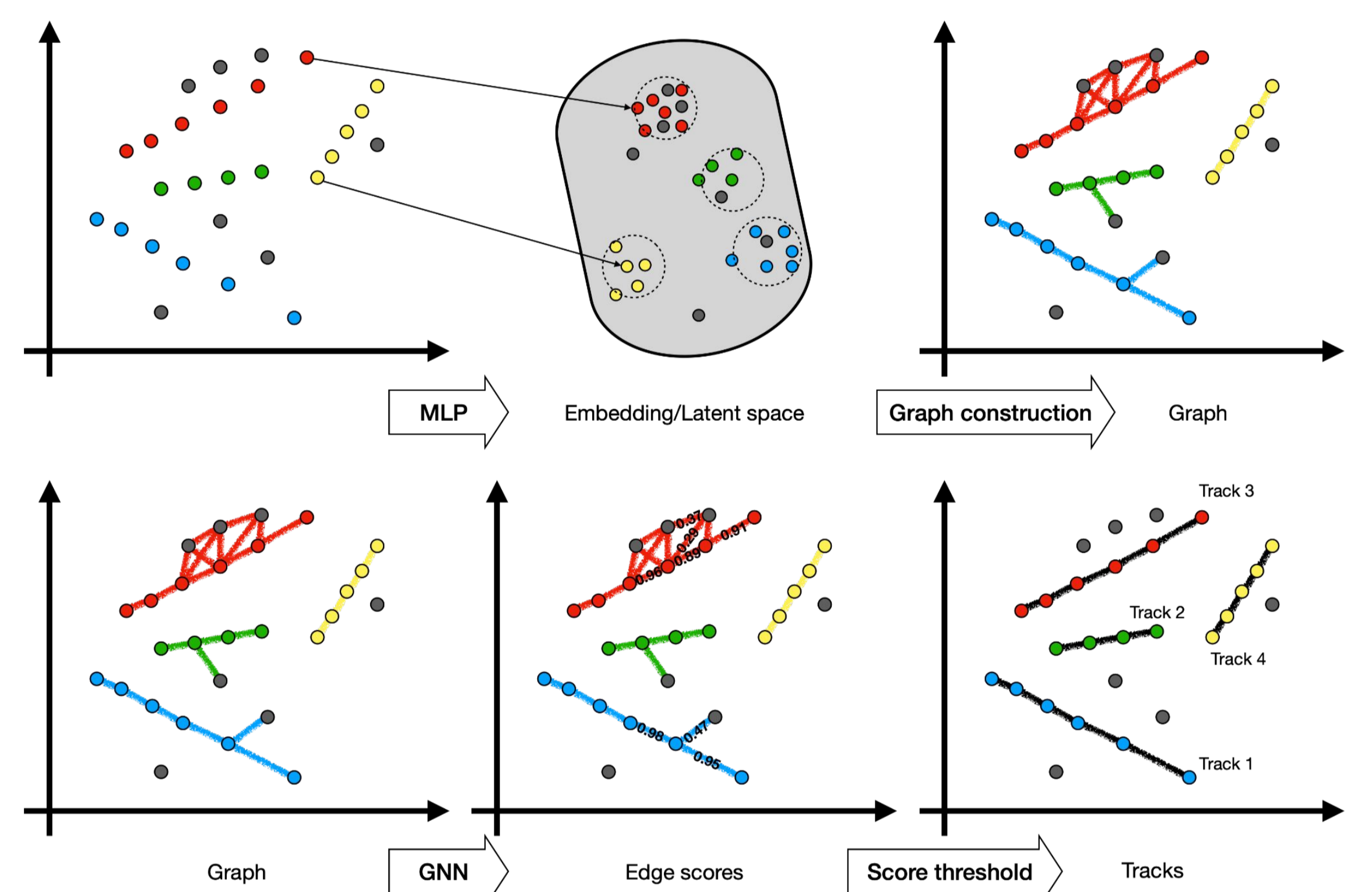

Category	Metric	Allen	ETX4VELO *smaller graph*	ETX4VELO *Larger graph*
Long, no electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Reconstructible in the SciFi ✓ Not an electron	Efficiency	99.26%	99.28%	99.51%
	Clone rate	2.54%	0.96%	0.89%
	Hit efficiency	96.46%	98.73%	98.90%
	Hit Purity	99.78%	99.94%	99.94%
Long electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Reconstructible in the SciFi ✓ Electron	Efficiency	97.11%	98.80%	99.22%
	Clone rate	4,25%	7.42%	7.31%
	Hit efficiency	95.24%	96.54%	96.79%
	Hit purity	97.11%	98.46%	98.46%
Long, from strange ✓ In acceptance ✓ Reconstructible in the velo ✓ Decays from a strange *Good proxy for displaced tracks*	Efficiency	97.69%	97.50%	98.06%
	Clone rate	2.50%	0.92%	0.81%
	Hit efficiency	97.69%	98.22%	98.77%
	Hit purity	99.34%	99.68%	99.68%
X	Ghost rate	2.18%	0.76%	0.81%

Inference of the ETX4VELO Models on GPUs and FPGAs

Then we focused on running these algorithms on architectures with more potential for performance.

- Integration inside the LHCb's first-level GPU trigger (HLT1)
- Inference of the MLP and GNN on GPU
- Inferece of the MLP on FPGA
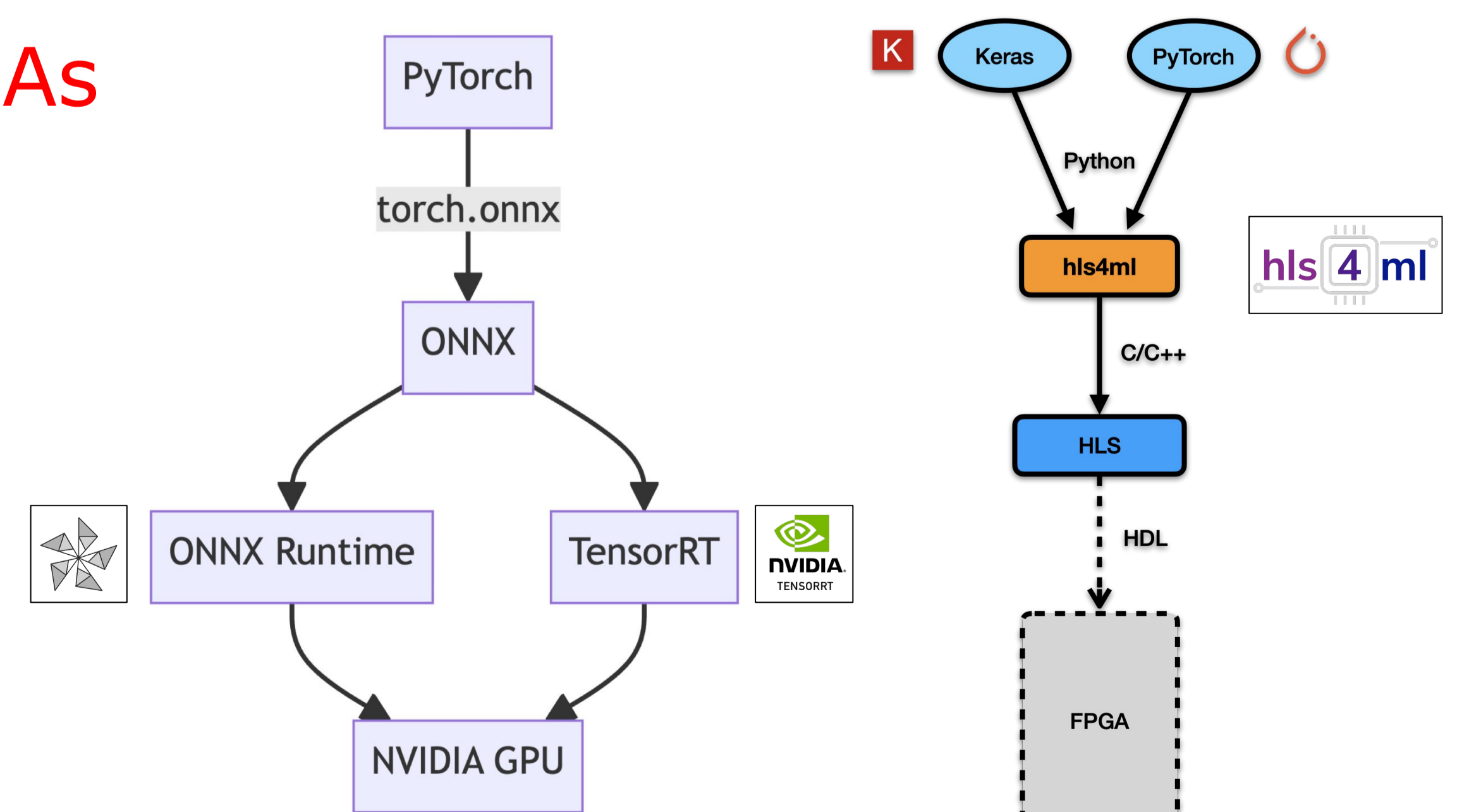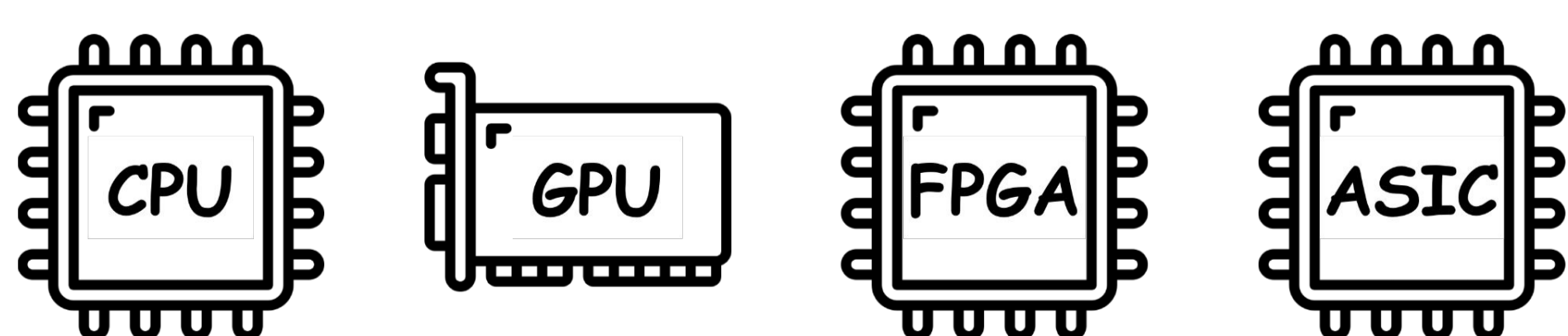
Pipeline	Step	Throughput		
		ONNX Runtime (FP32)	TensorRT (FP32)	TensorRT (INT8)
ETX4VELO	MLP	50K events/s	260K events/s	**520K events/s**
	GNN	0.31K events/s	0.86K events/s	-
	Velo tracks	-	**0.83K events/s**	-
ALLEN	Velo tracks	**510K events/s**		