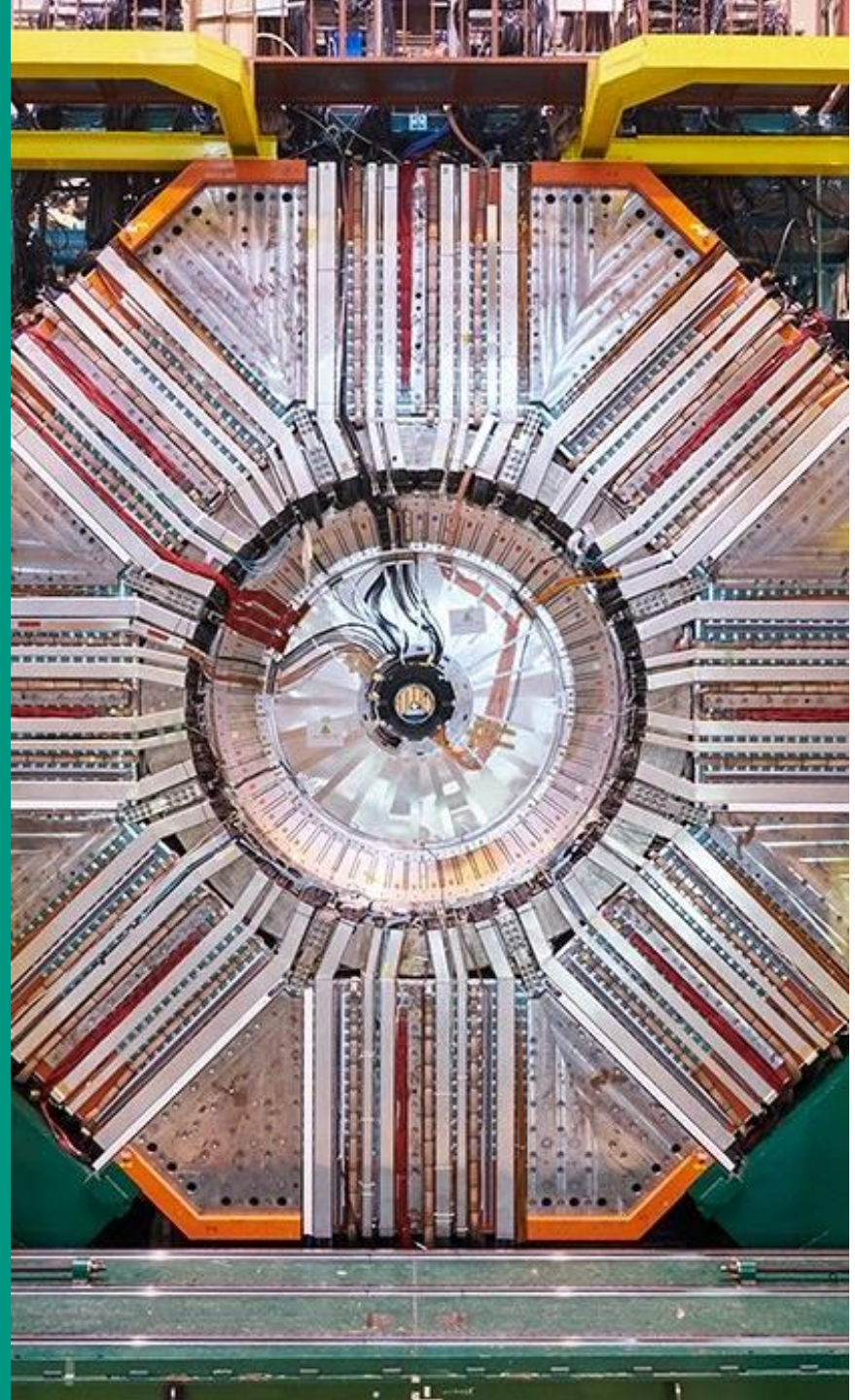
Commissioning and Low Latency Operation of the Graph Neural Network Electromagnetic Calorimeter Trigger at the Belle II Experiment



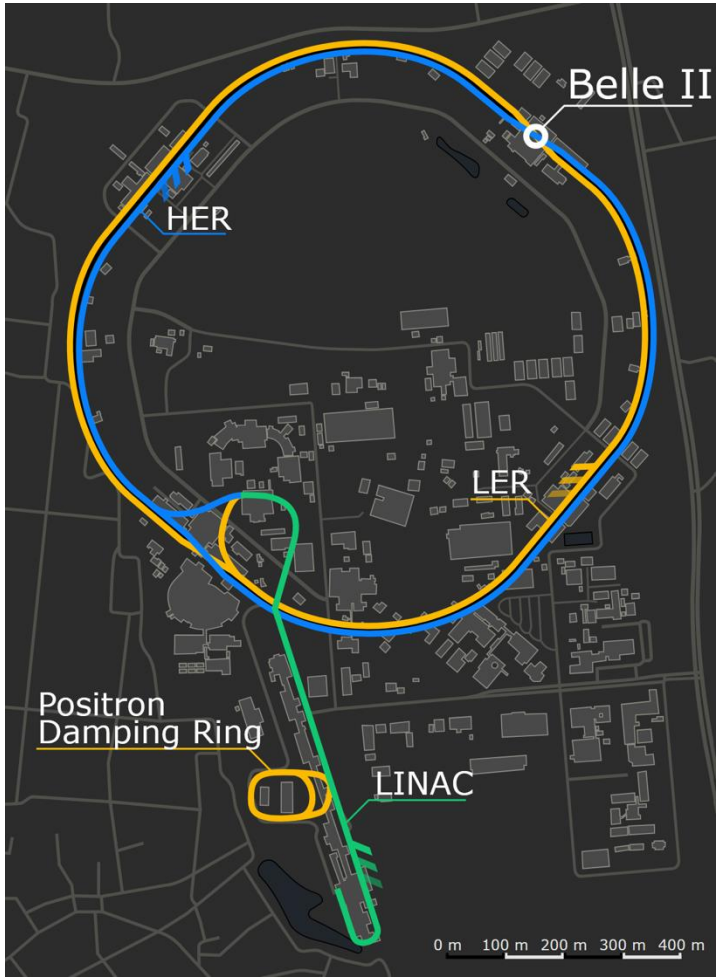
Marc Neu, Frank Baptist, Jürgen Becker, Torben Ferber, Isabel Haide, Yuuji Unno

marc.neu@kit.edu

25th IEEE Real Time Conference

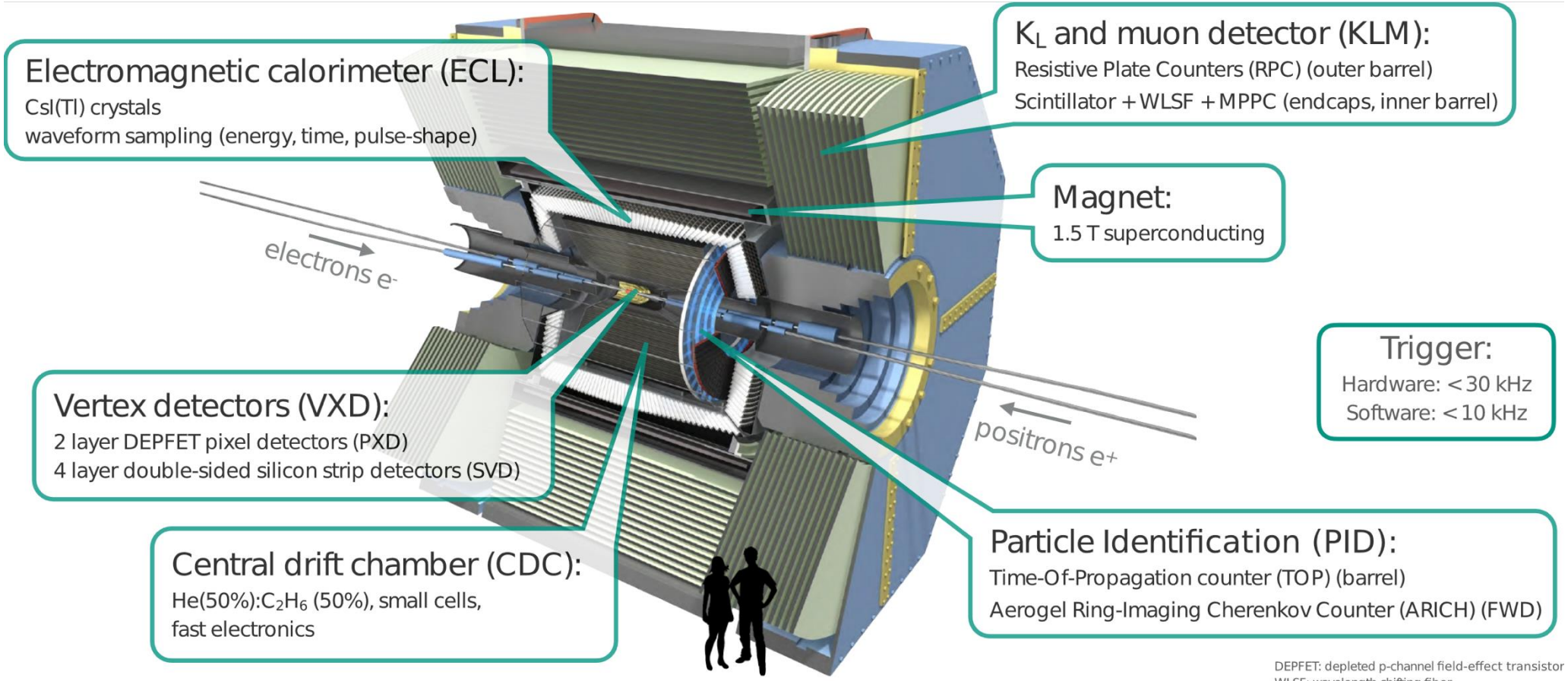
La Biodola, Elba, Italy, 28.05.2026

Belle II Experiment at SuperKEKB



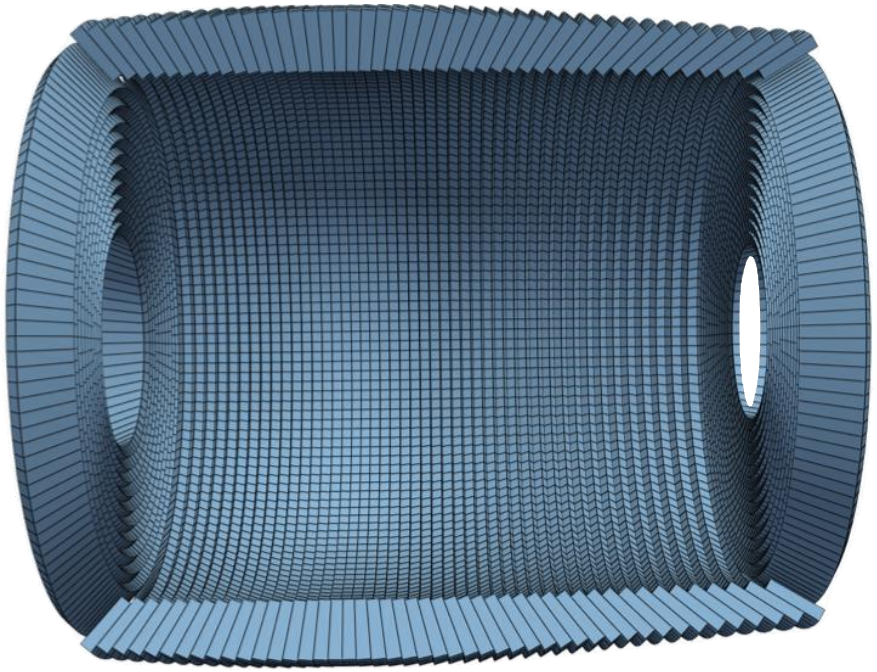
- SuperKEKB in Tsukuba, Japan operates at a bunch crossing rate of up to 250 MHz
- The electromagnetic calorimeter detector reads out 8 million detector snapshots per second
- However, the Data Aquisition System of the Belle II Detector only supports a continous readout rate of up to 30 kHz
- **Real-Time Hardware Trigger is required for Online Event Selection**

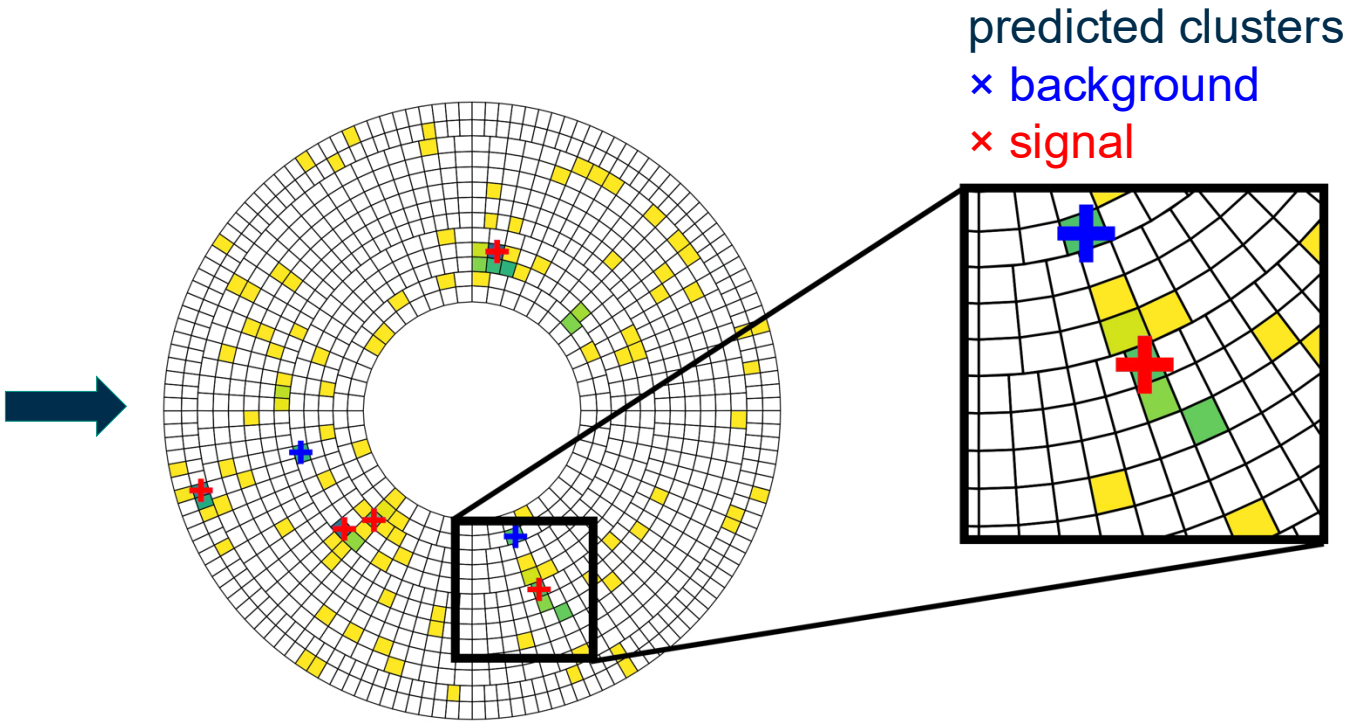
Belle II Detector – Electromagnetic Calorimeter (ECL)



DEPFET: depleted p-channel field-effect transistor
WLSF: wavelength-shifting fiber
MPPC: multi-pixel photon counter

The Belle II Electromagnetic Calorimeter





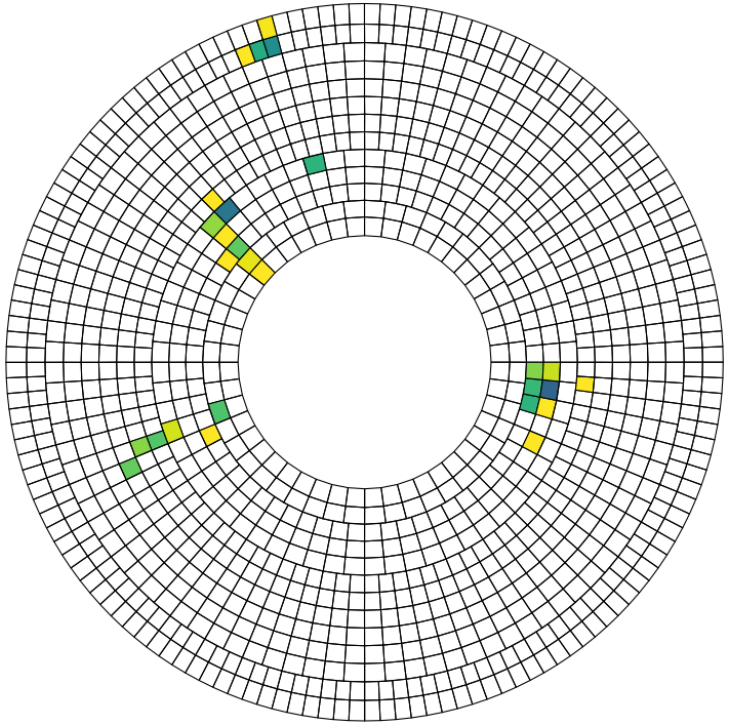
Belle II Electromagnetic Calorimeter

8736 CsI(Tl) Crystals

Figure by Thomas Lobmaier

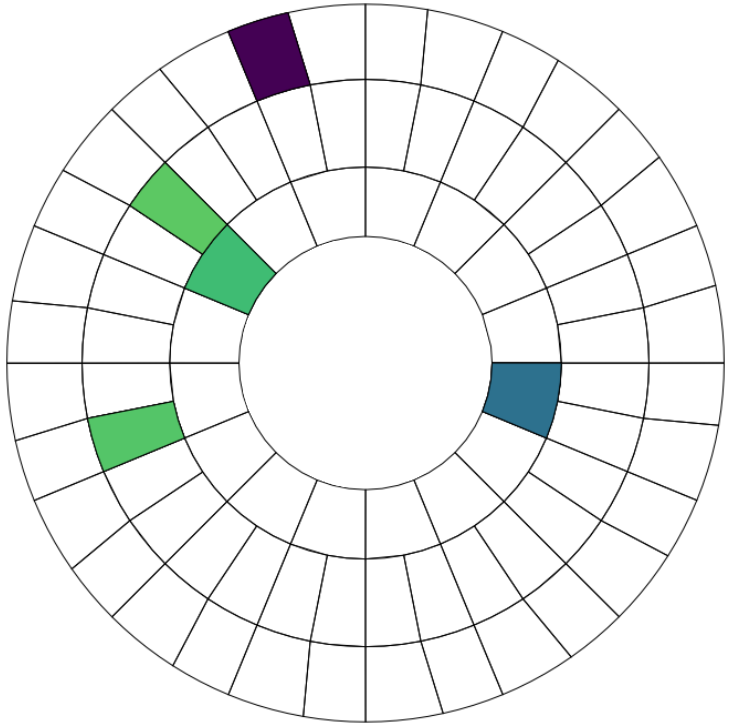
The Belle II Calorimeter – Crystal Readout on Trigger Level

Offline Resolution



- Full readout with **8736 crystals**
- Lower Energy Threshold

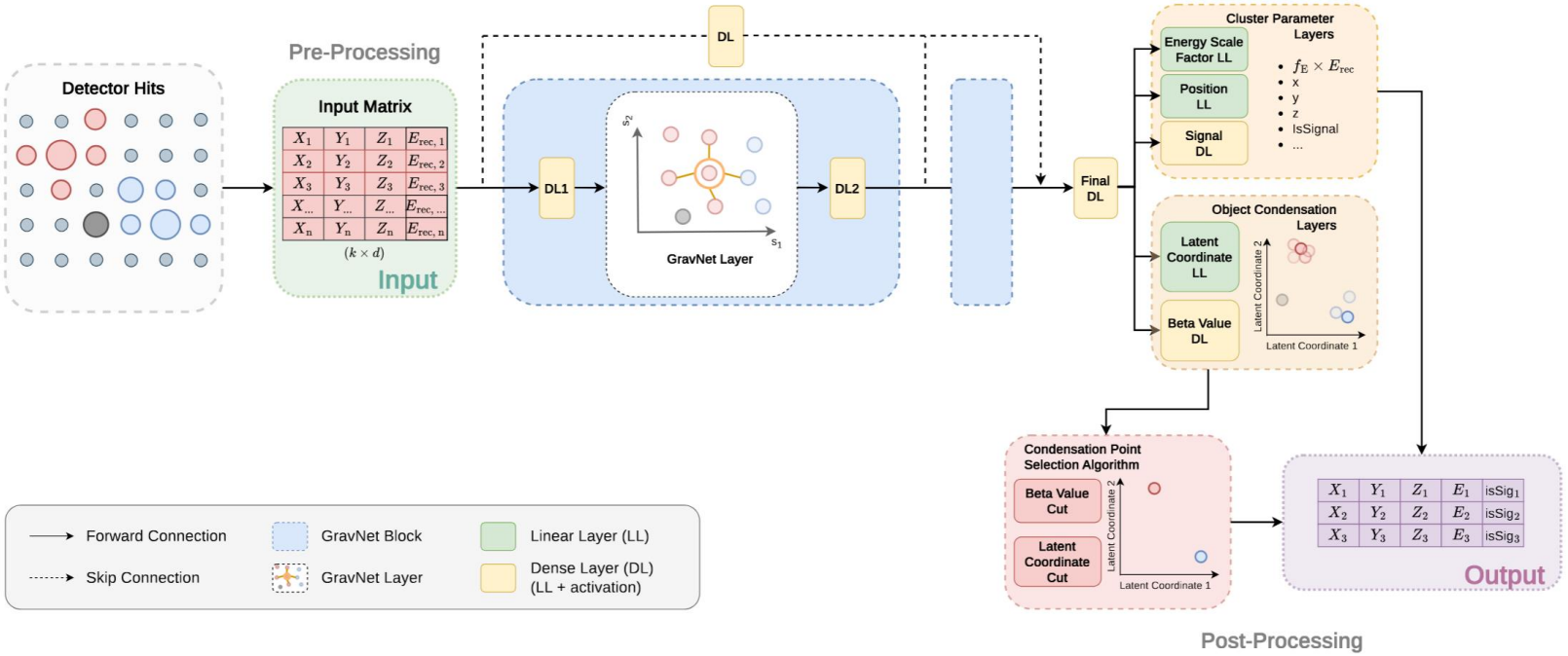
Trigger Resolution



- Reduced readout with **576 trigger cells**
- Data reduction through 4x4 analog sum
- Typically, less than **32 active trigger cells**

Dynamic Graph Neural Networks for Calorimeter Clustering: CaloClusterNet

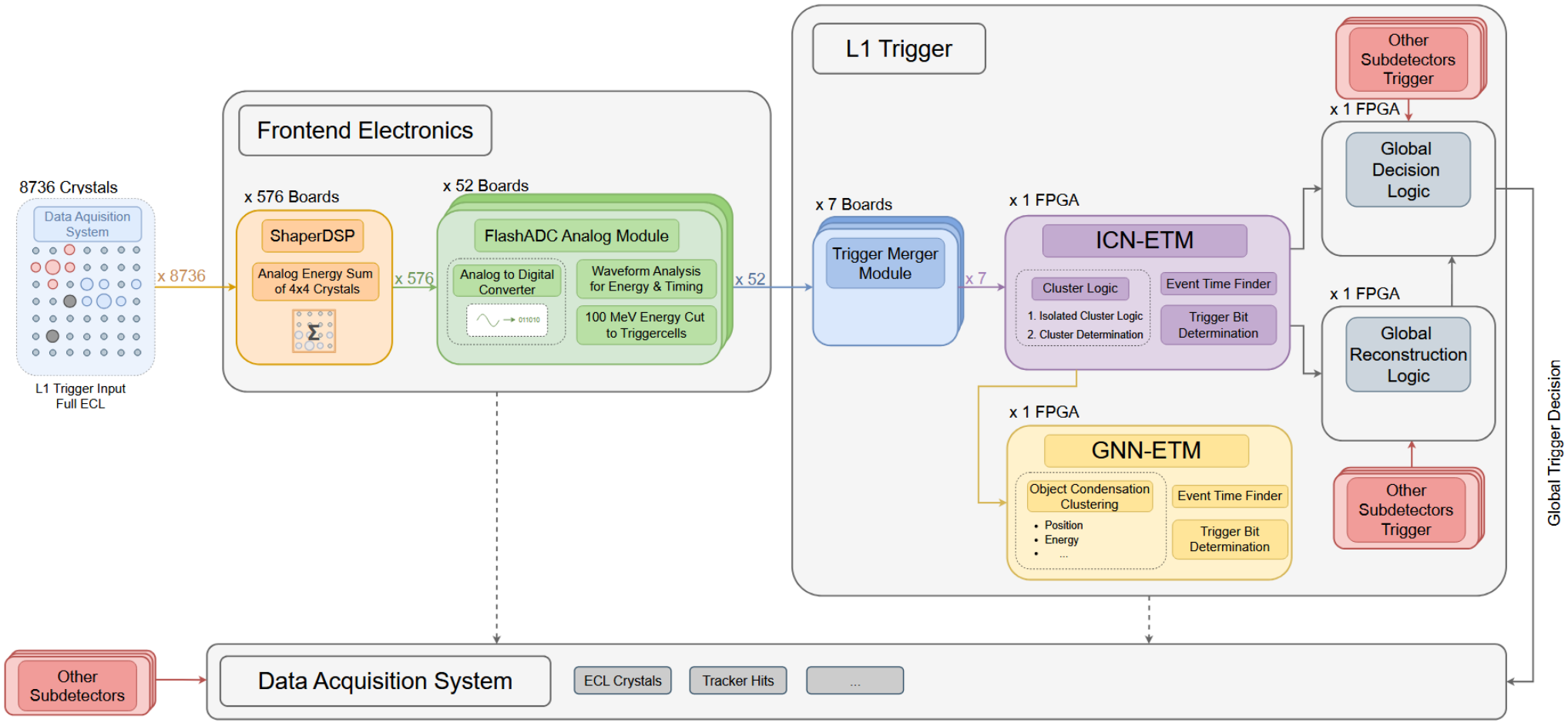
- For our work, we have developed the CaloClusterNet architecture [Hai26]
- GraVNet Layer for dynamic graph building via k-Nearest-Neighbour algorithm in a learned representation space [1]
- Selection of an unknown number of clusters via the Condensation Point Selection algroithm [2]
- Prediction of position, energy, and signal / background classification per cluster



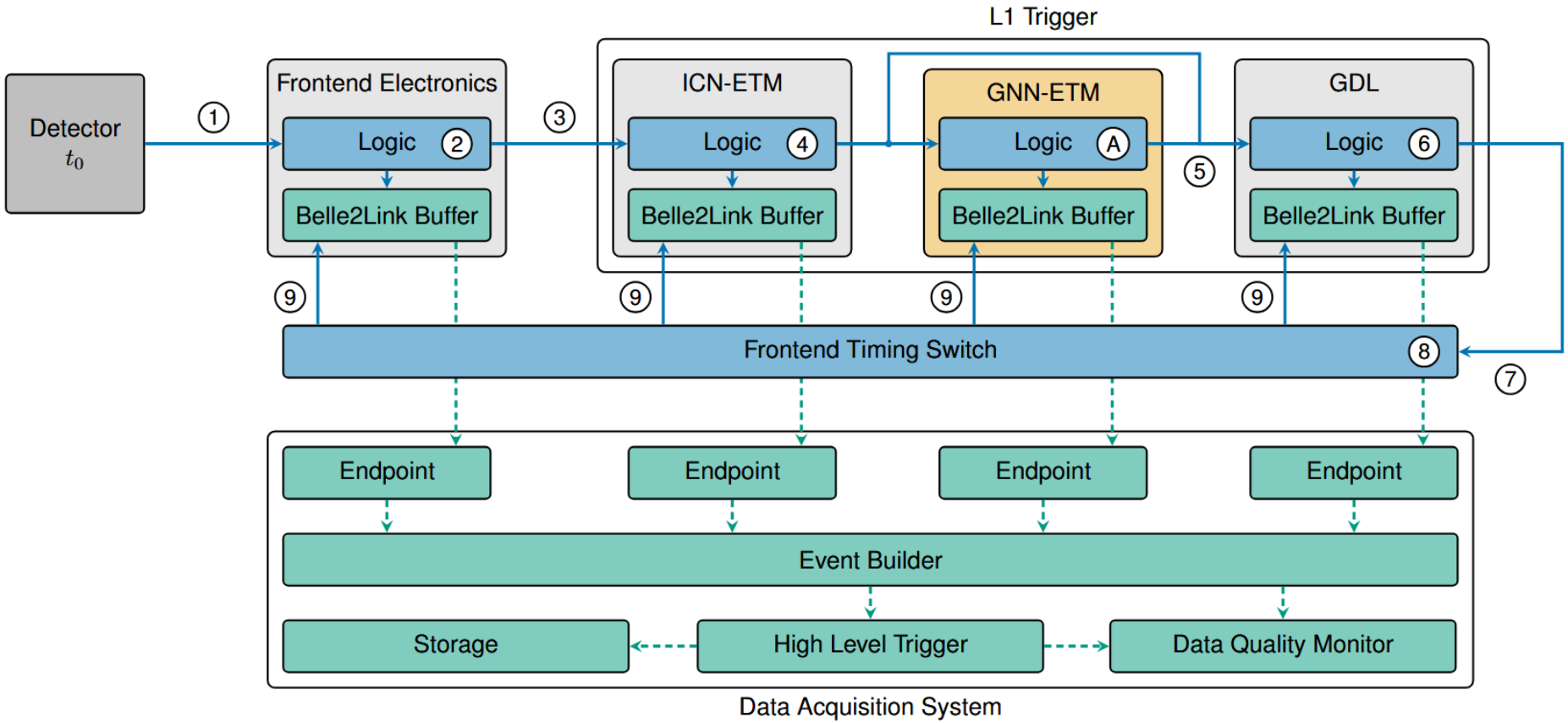
[1] S. R. Qasim et al., "Learning representations of irregular particle-detector geometry with distance-weighted graph networks," *Eur. Phys. J. C*, 2019.

[2] J. Kieseler, "Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph and image data," arXiv, 2020.

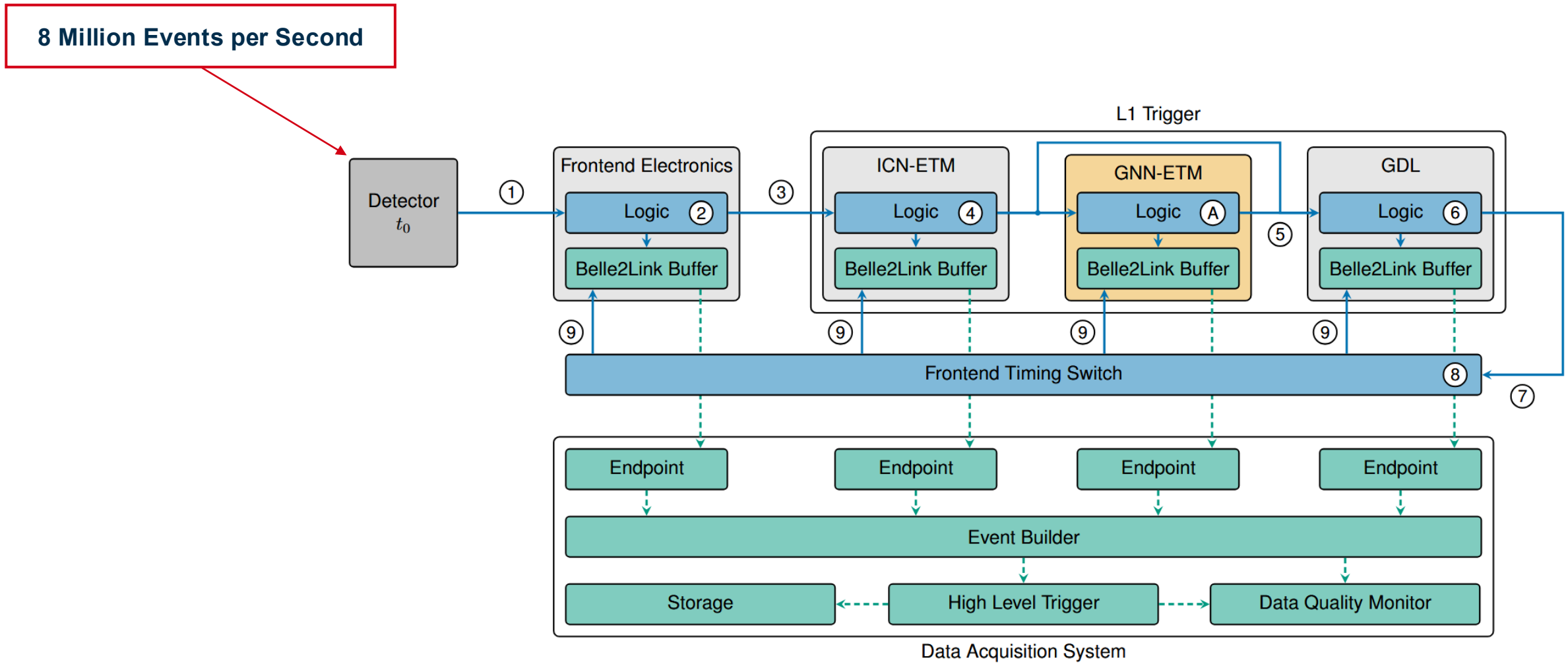
Integration into the first-level Trigger System at Belle II



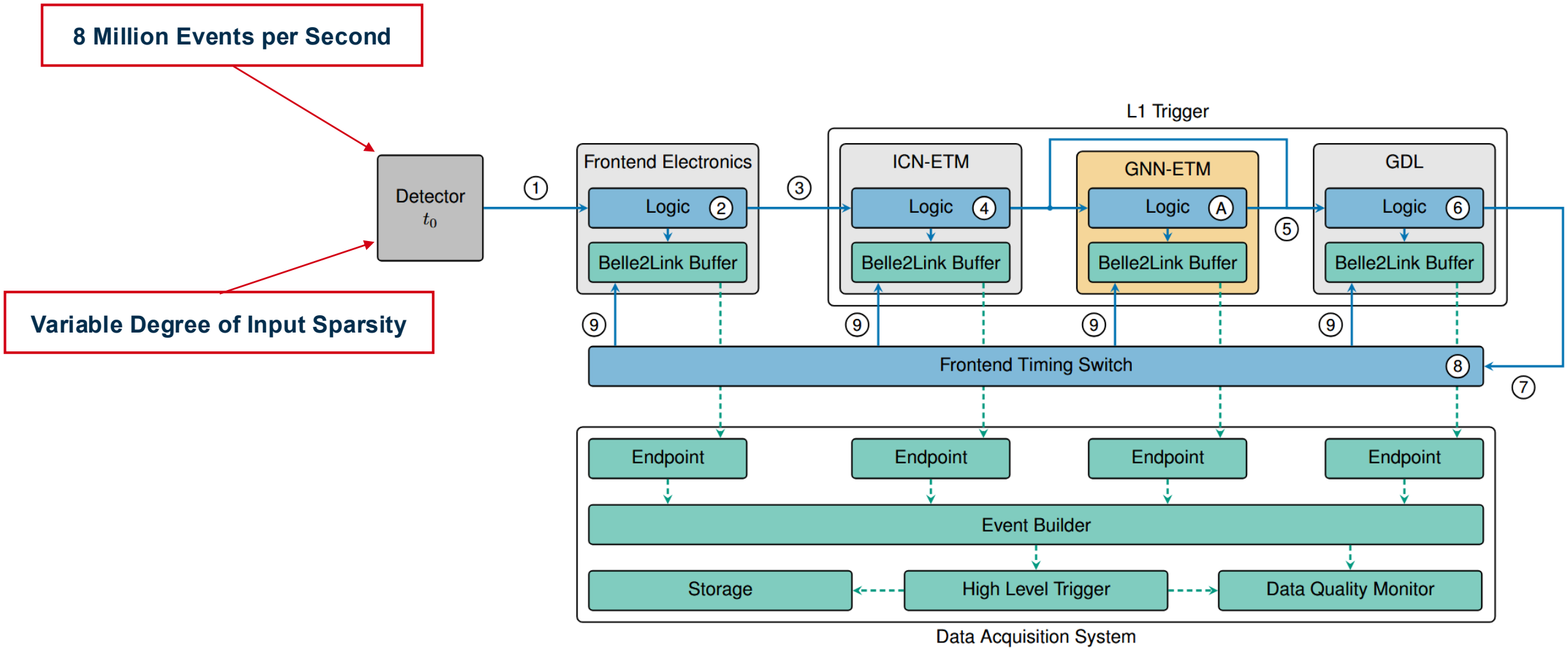
Integration into the first-level Trigger System at Belle II



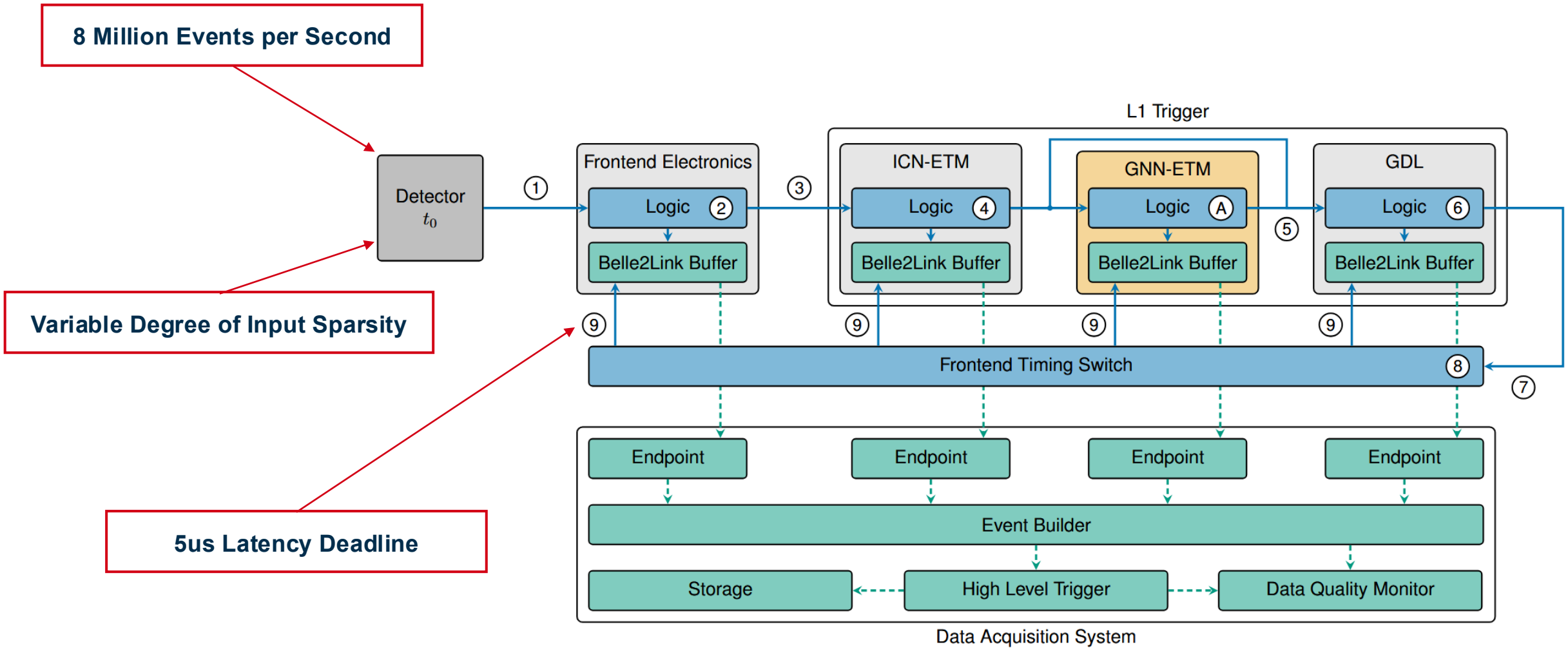
Integration into the first-level Trigger System at Belle II



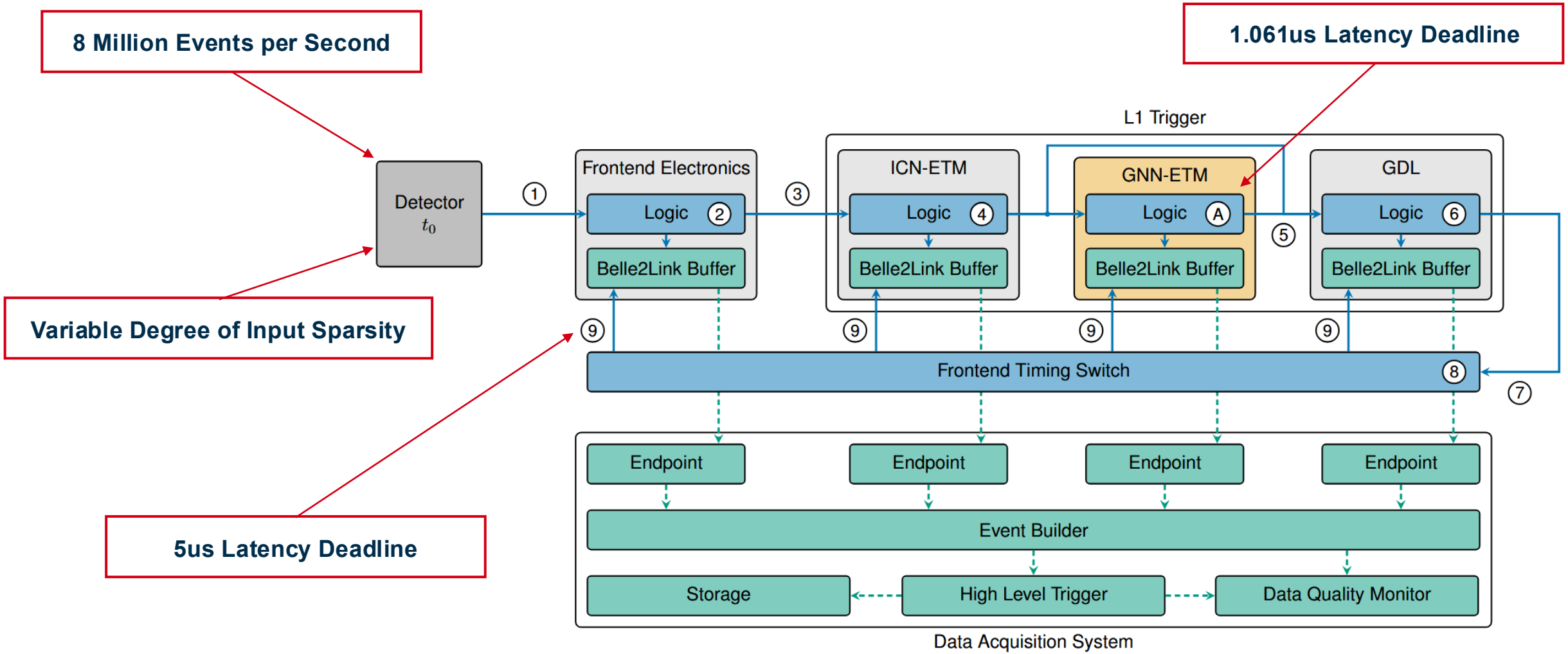
Integration into the first-level Trigger System at Belle II



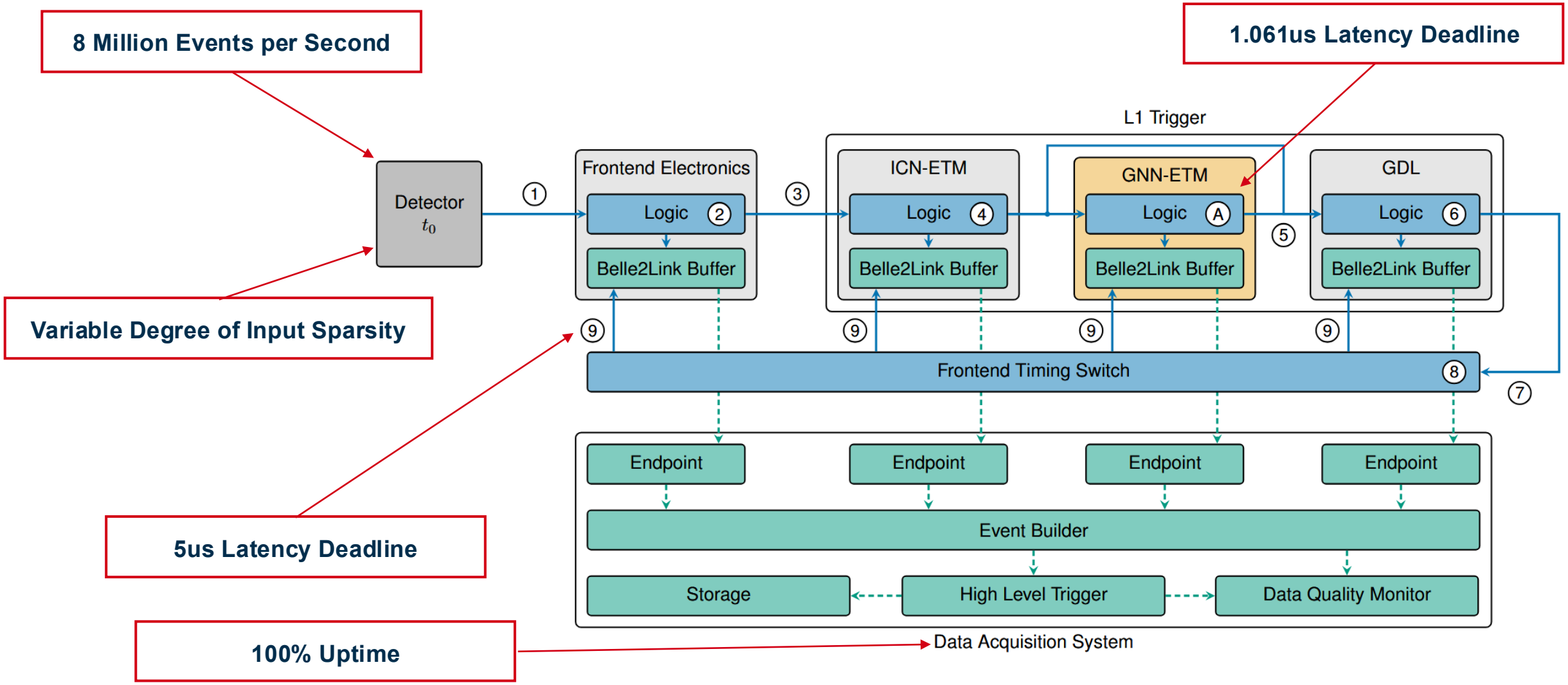
Integration into the first-level Trigger System at Belle II



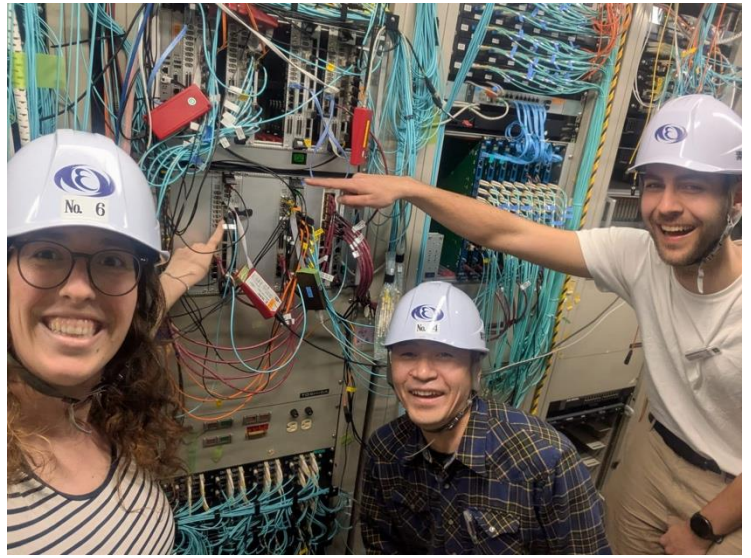
Integration into the first-level Trigger System at Belle II



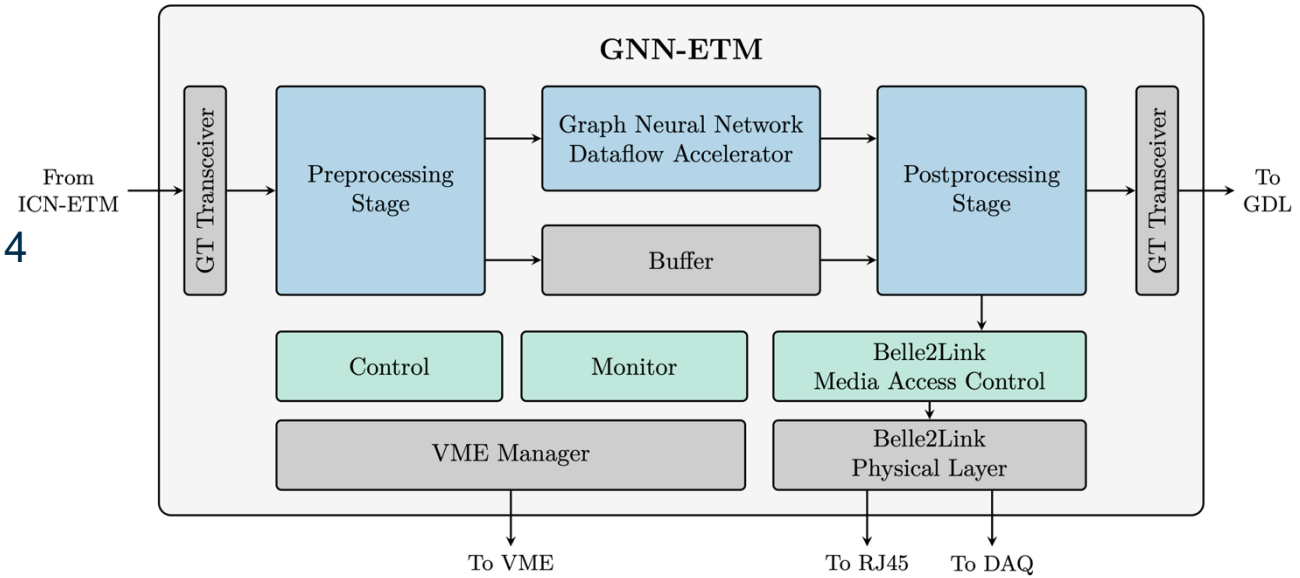
Integration into the first-level Trigger System at Belle II



The GNN-ETM



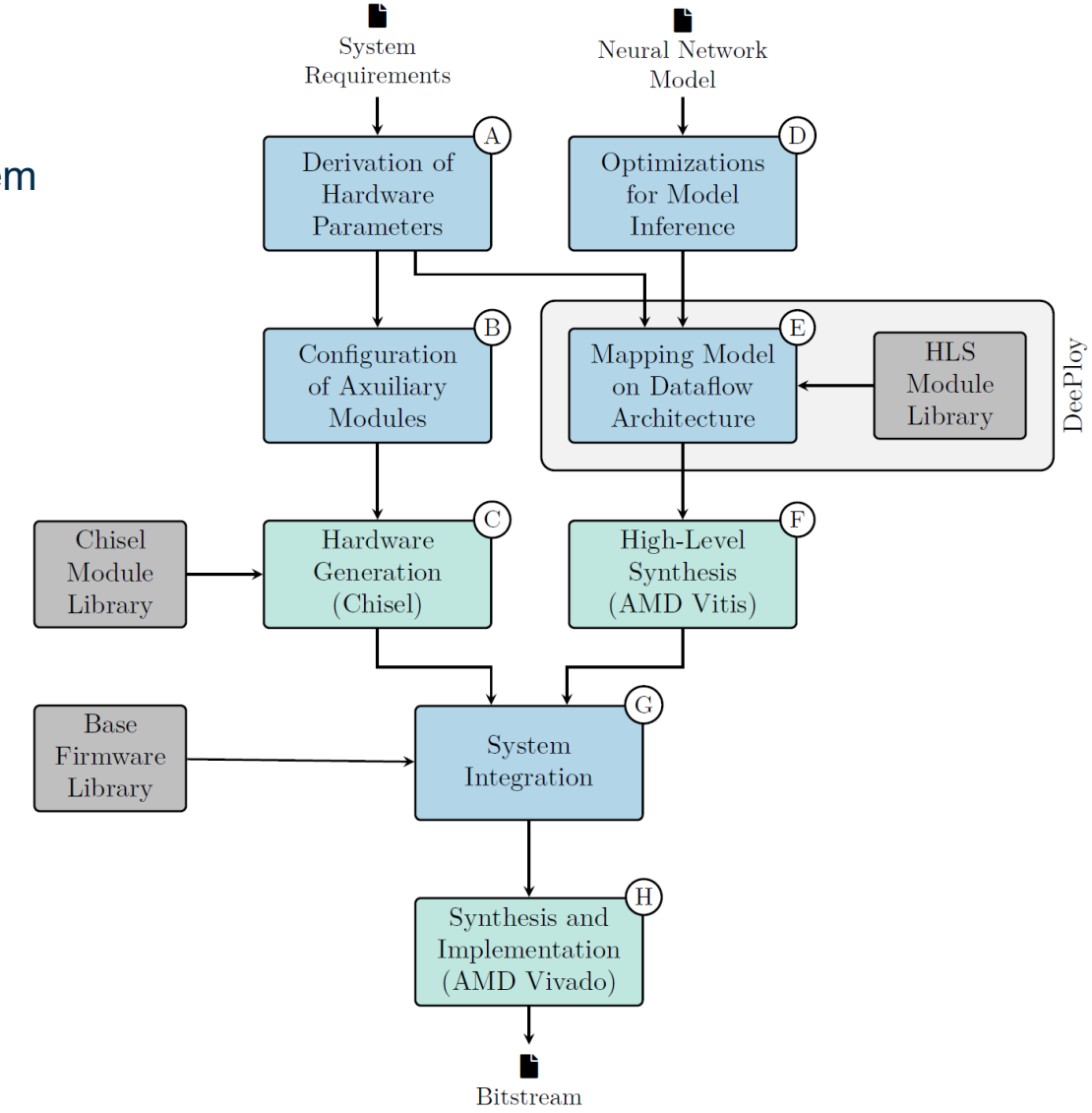
- Preprocessing stage is implemented in Chisel and facilitates the reduction of 576 trigger cells to less than 32 trigger cells through stream compaction [1]
- For the deployment of the CaloClusterNet we use our template-based approach and AMD Vitis HLS 2024.2 as backend
- The postprocessing stage is also implemented in Chisel
- GNN-ETM is deployed on the Universal Trigger Board 4 (AMD Ultrascale XCVU 190)
- Data Readout via monitoring protocol (Belle2Link) similar to Ethernet



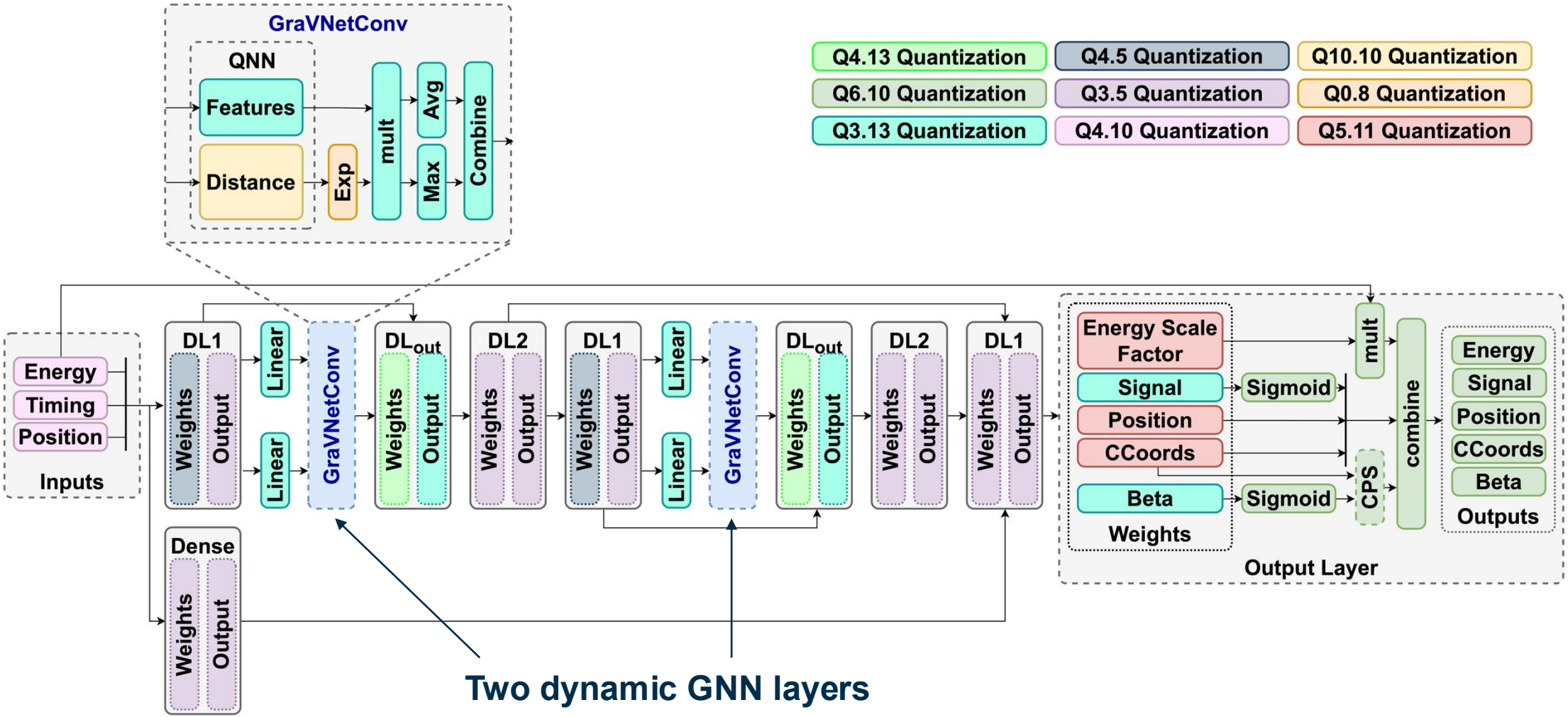
[1] M. Neu et al., „Real-Time Stream Compaction for Sparse Machine Learning on FPGAs", arXiv, 2026.

Deployment Methodology

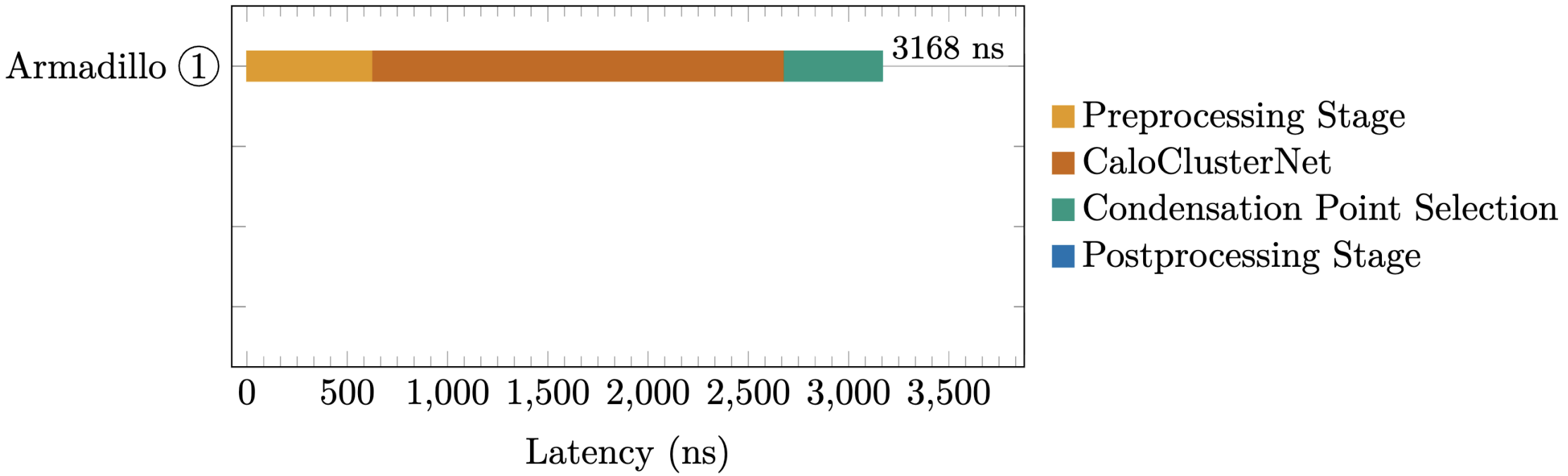
- We develop an end-to-end deployment methodology for our system
- User provides
  - System Requirements (Latency, Throughput)
  - Neural Network Model (QKeras)
- Semi automated flow in Python combining
  - Chisel3 for generation of RTL modules
  - AMD Vitis HLS for our GNN dataflow accelerator
  - FuseSoC for system integration
  - CoCoTB for RTL simulation
  - AMD Vivado for synthesis and implementation



Baseline CaloClusterNet (Armadillo)

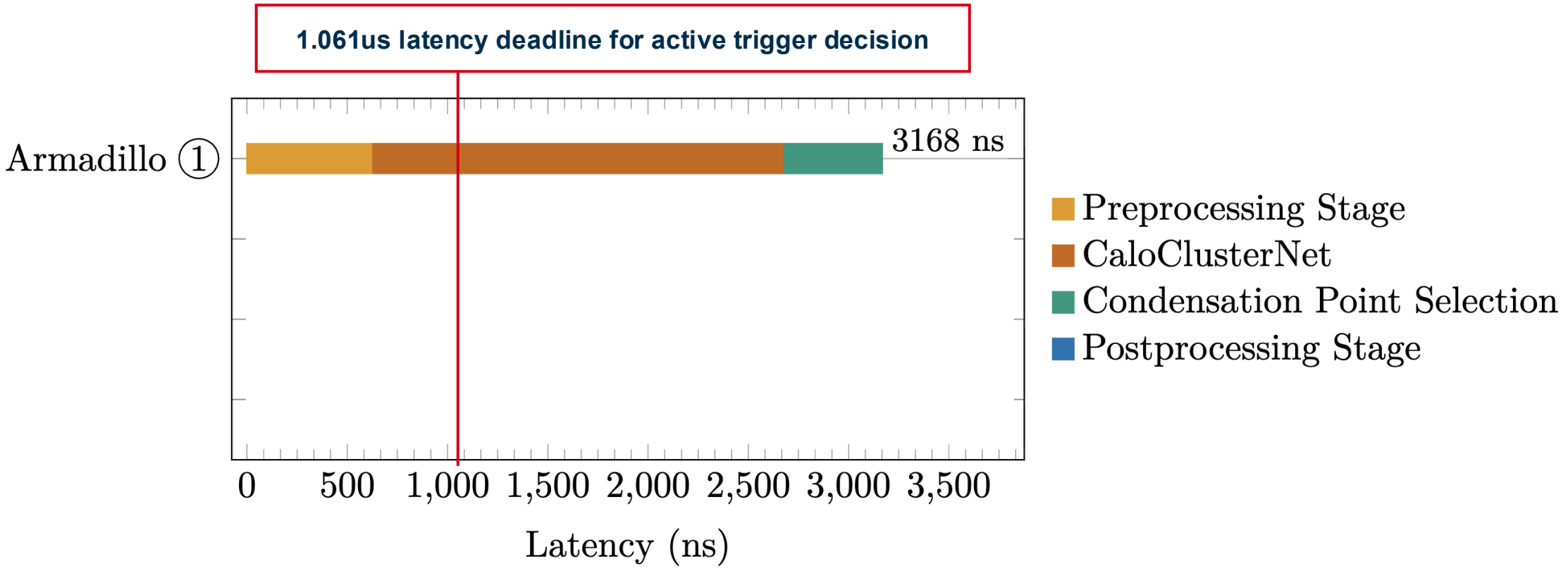


Baseline CaloClusterNet (Armadillo)



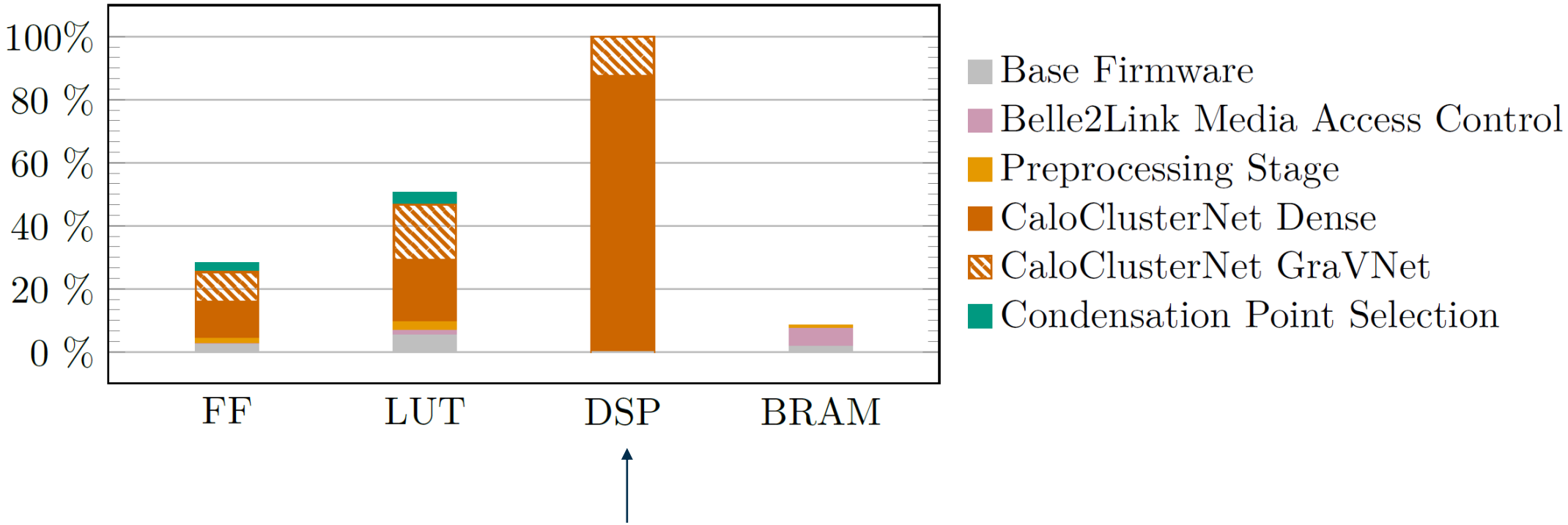
Can we optimize our deployment to reach the latency requirement without reducing the algorithmic performance of our model?

Baseline CaloClusterNet (Armadillo)



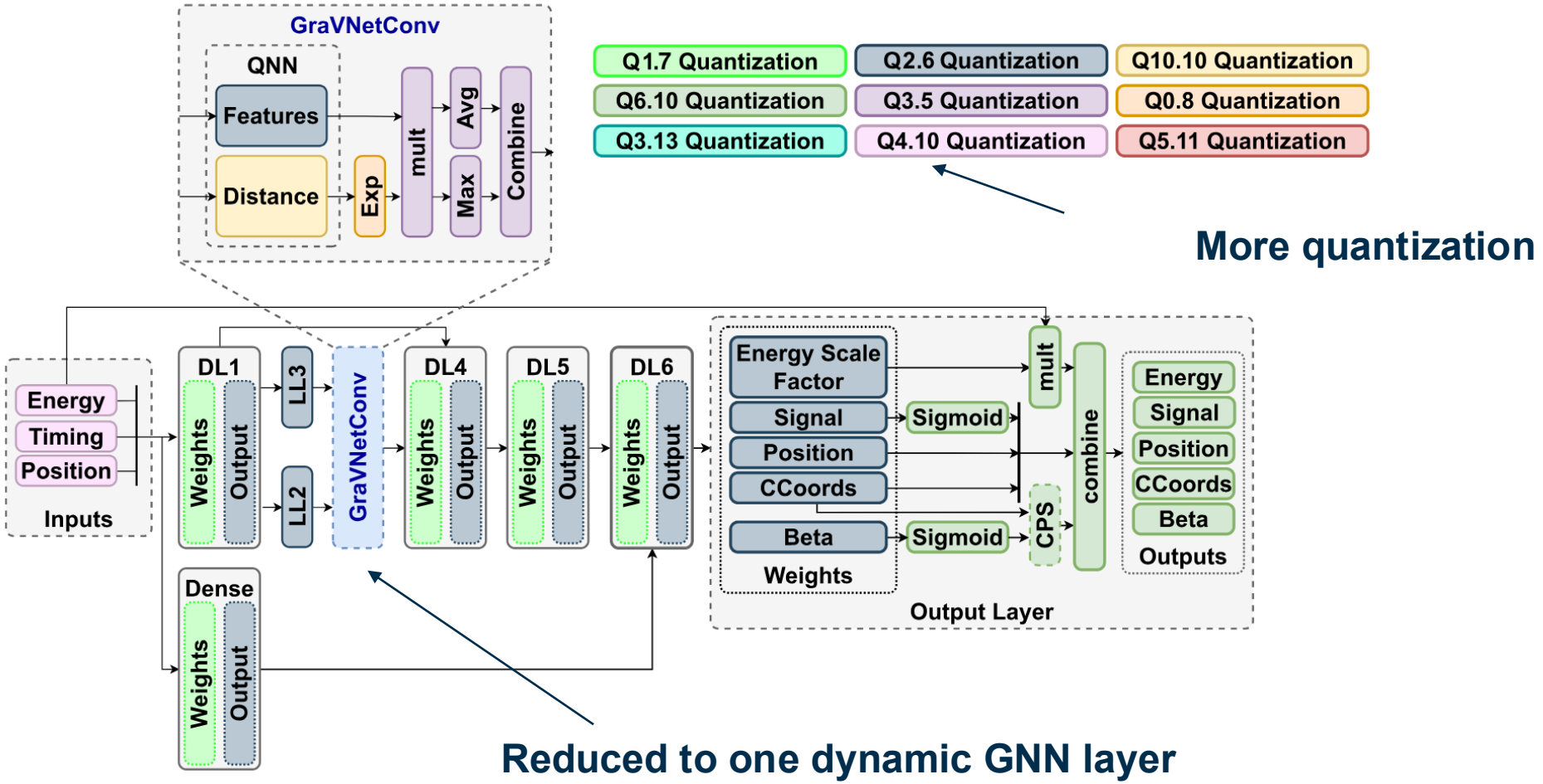
Can we optimize our deployment to reach the latency requirement without reducing the algorithmic performance of our model?

Baseline CaloClusterNet (Armadillo)

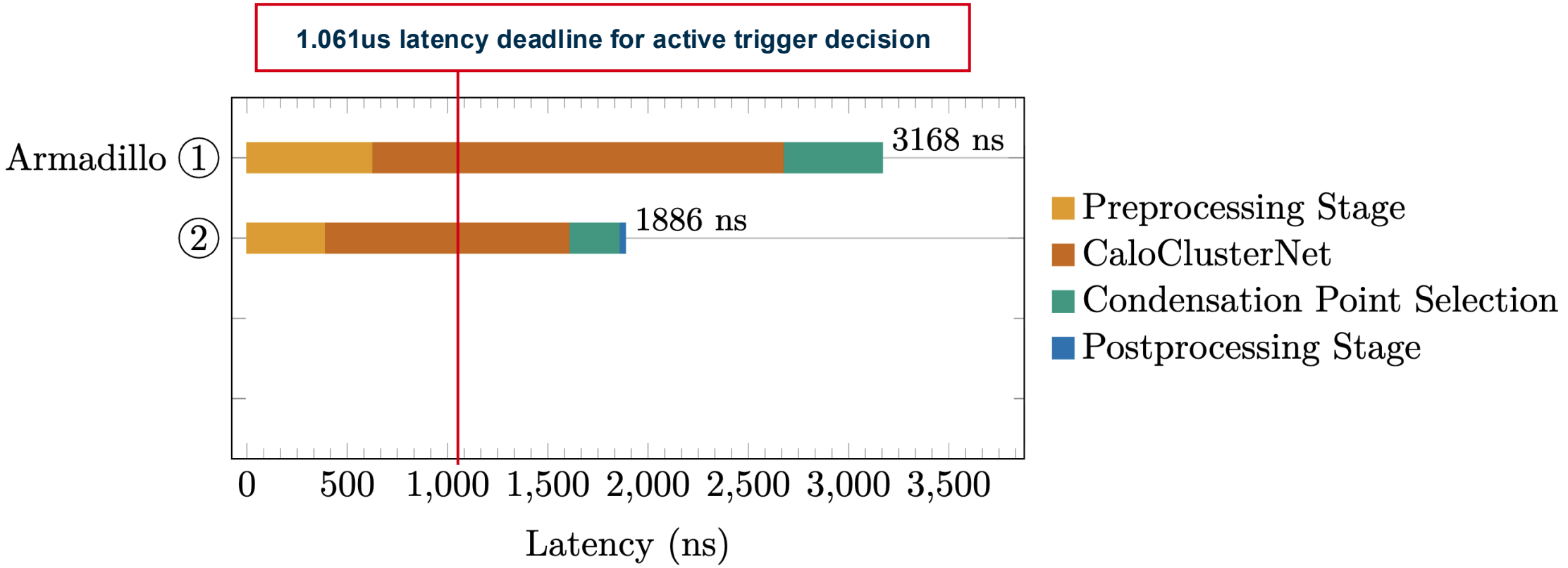


No additional parallelization with the current FPGA

Compressed CaloClusterNet (Sunset)



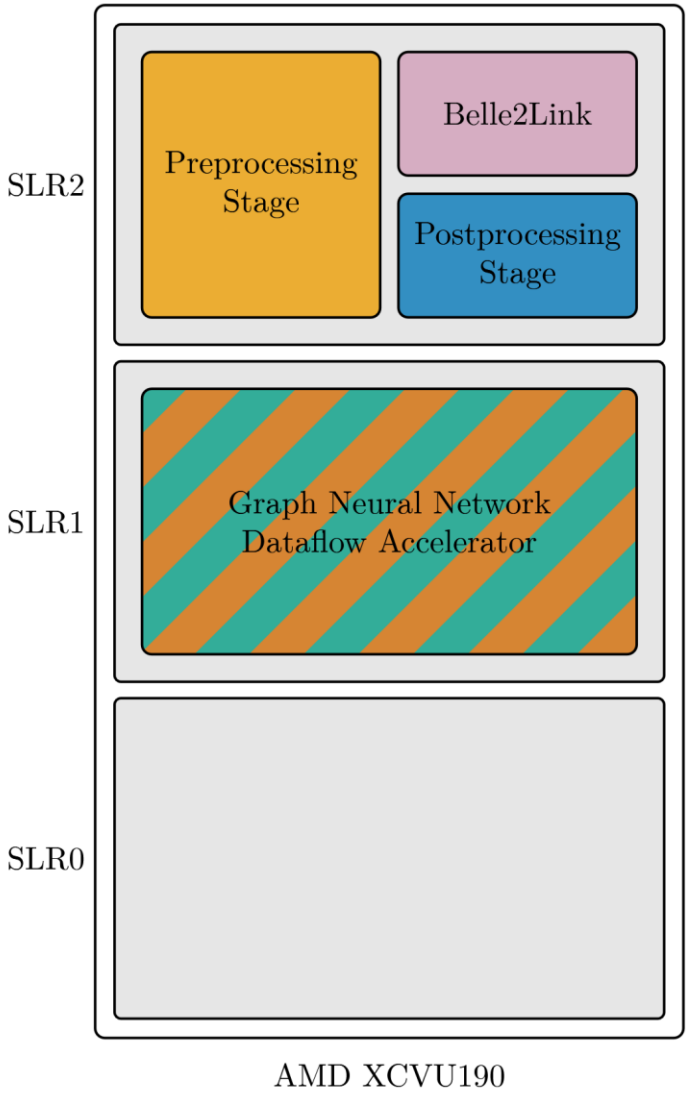
Compressed CaloClusterNet (Sunset)



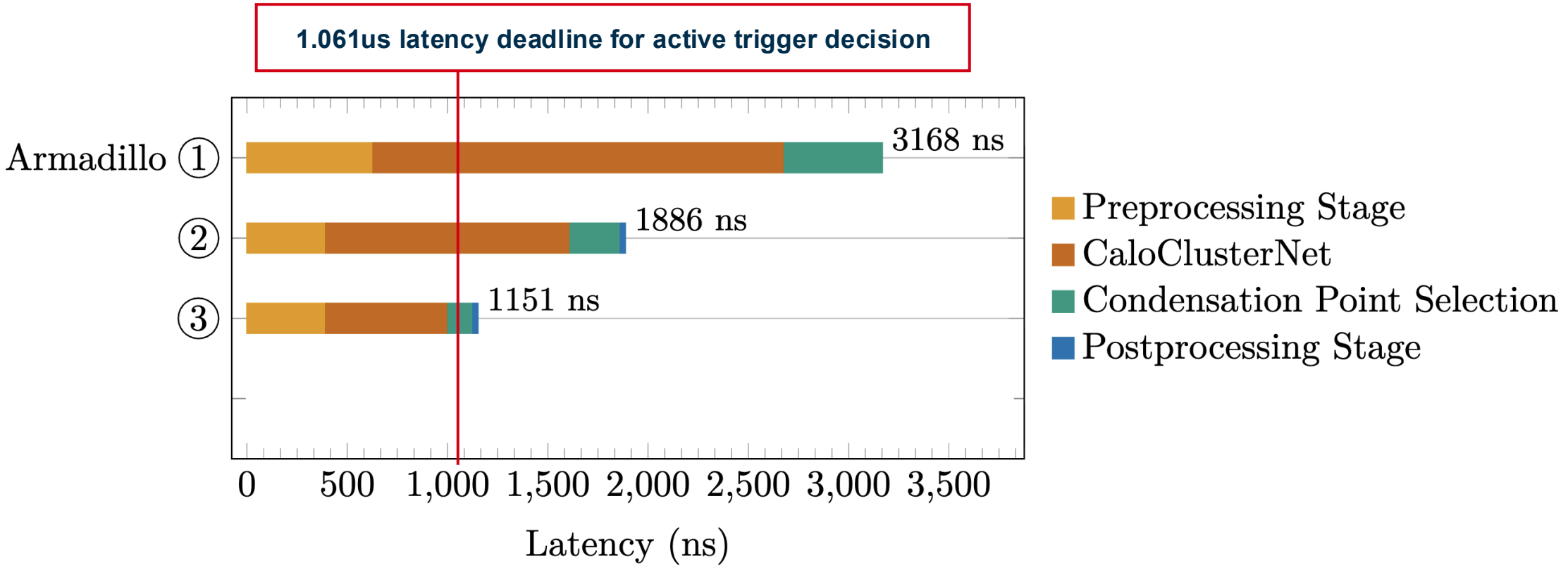
1282ns latency reduction through model compression

Physical Design Optimizations

- In initial versions, we only achieved up to 127 MHz on the AMD Ultrascale XCVU190
- Multiple Super Logic Regions on the FPGA introduced bottlenecks for timing and routing congestion
- Thus, we introduce a manual floorplanning approach
- As a result, we double the design frequency to 254 MHz without timing issues



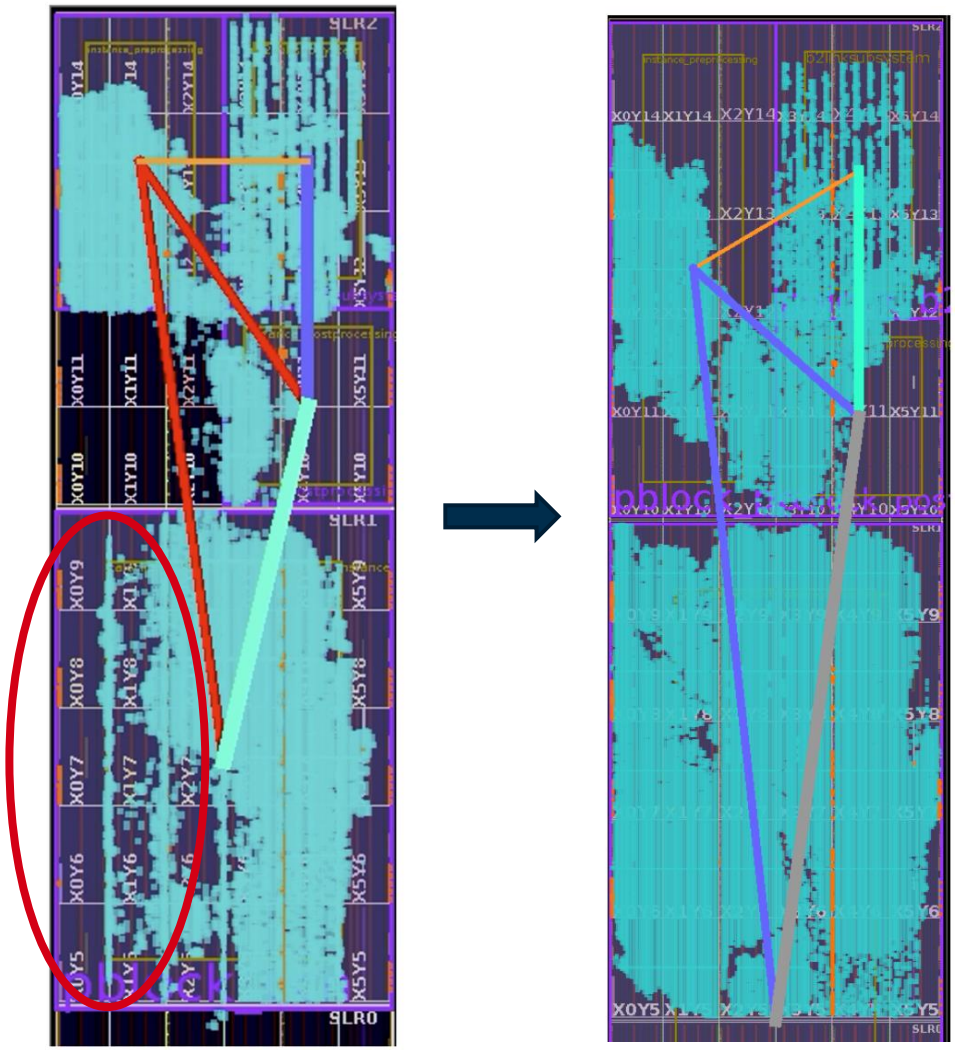
Physical Design Optimizations



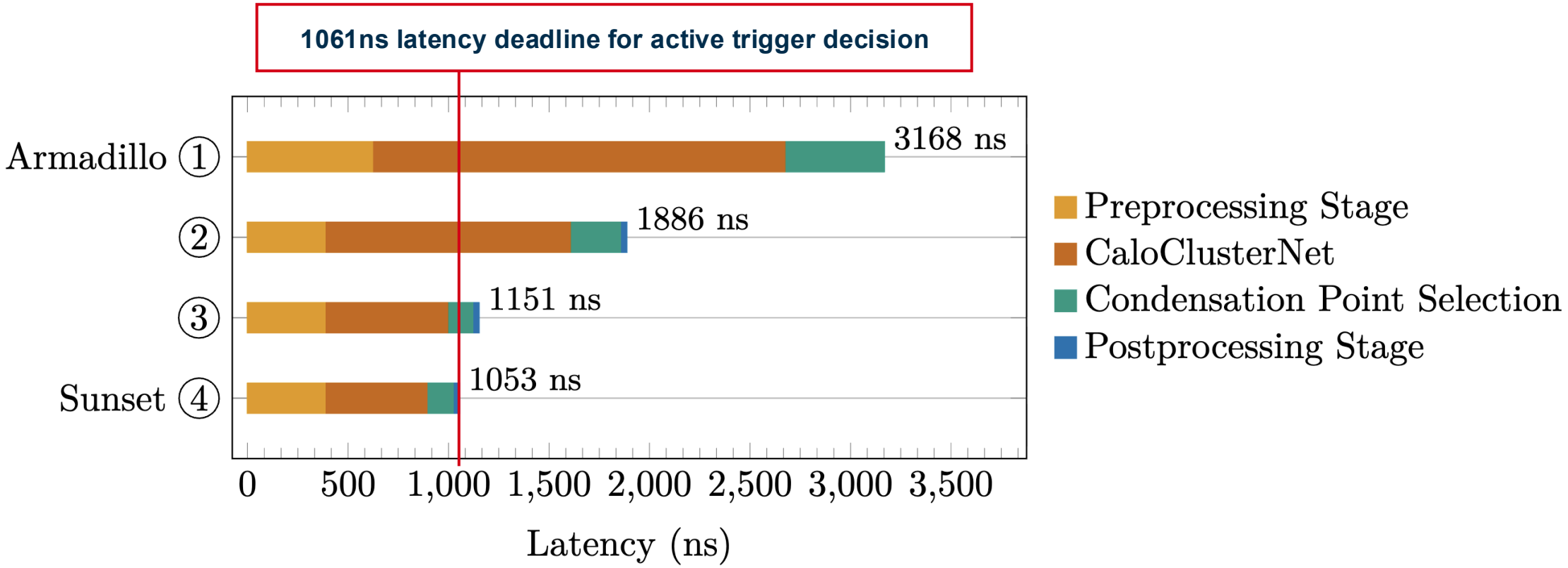
735ns latency reduction through physical design optimizations

Improving HLS Directives

- Initially, we aimed to maximize the DSPs utilization
- However, DSP Macros require extensive pipeline register for higher frequencies
- For example on AMD Ultrascale, up to 3 clock cycles are required
- Turns out, removing them improves latency significantly ☺

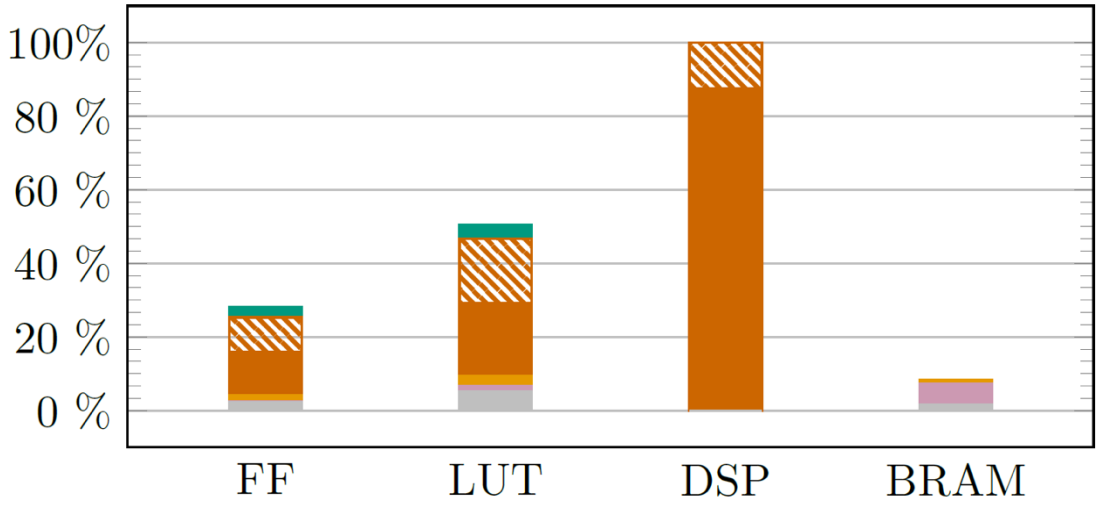


GNN-ETM Final Design

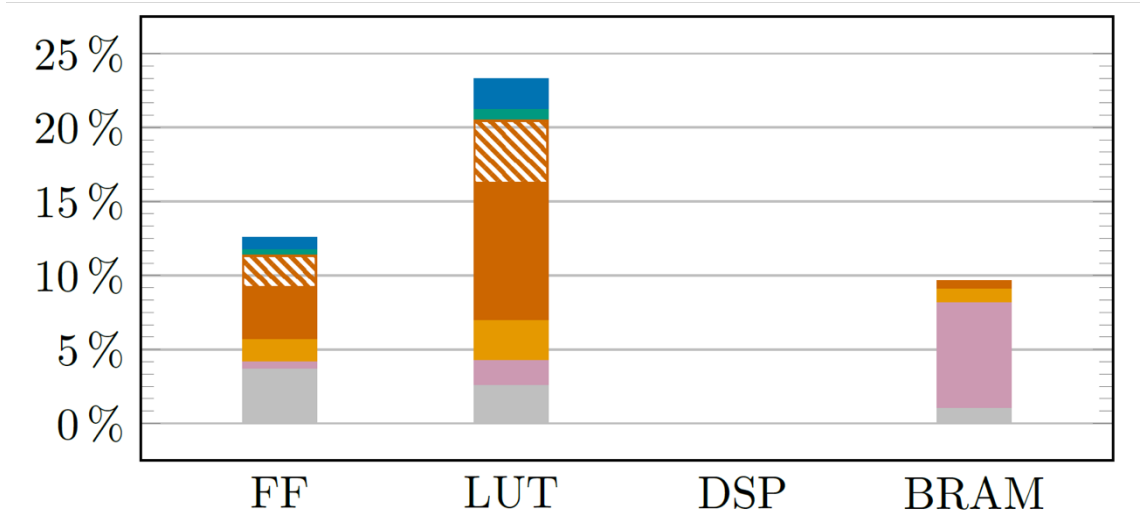


Latency deadline reached!

GNN-ETM Final Design

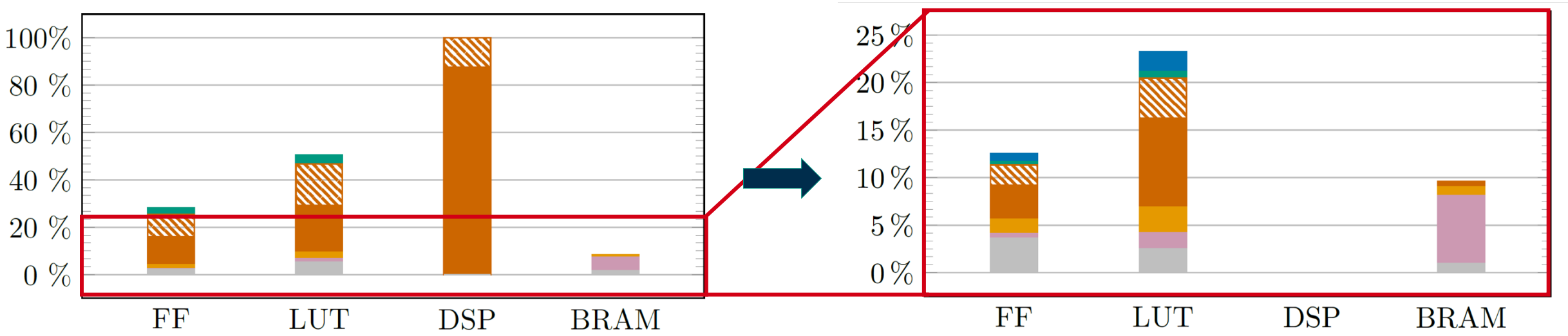


Initial Armadillo Model



Optimized Sunset Model

GNN-ETM Final Design

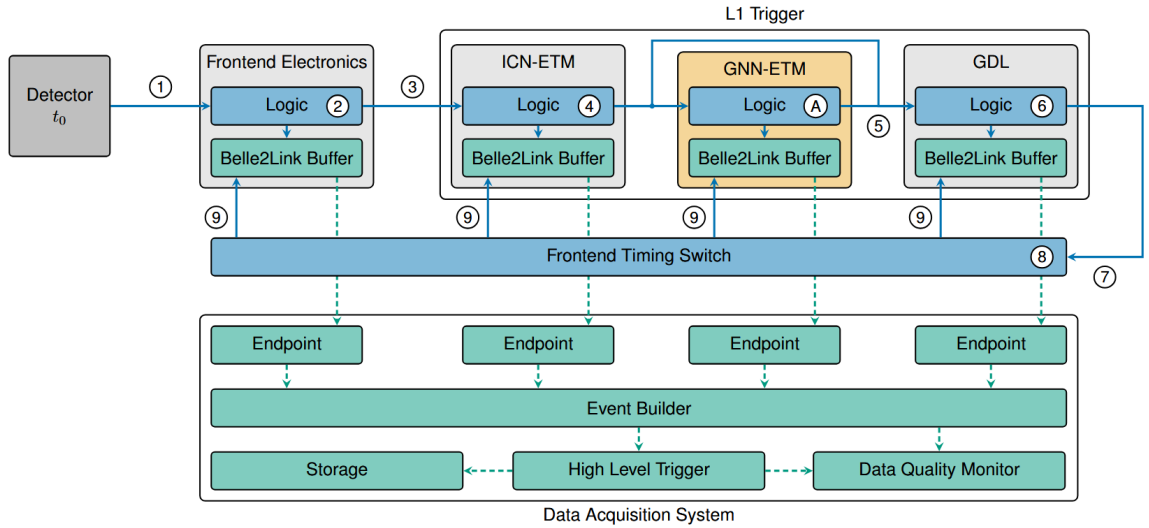


Initial Armadillo Model

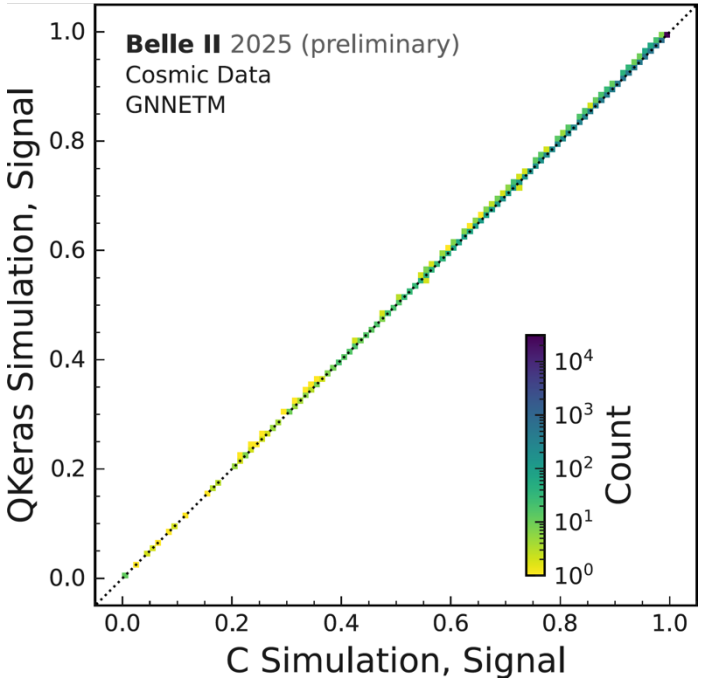
Optimized Sunset Model

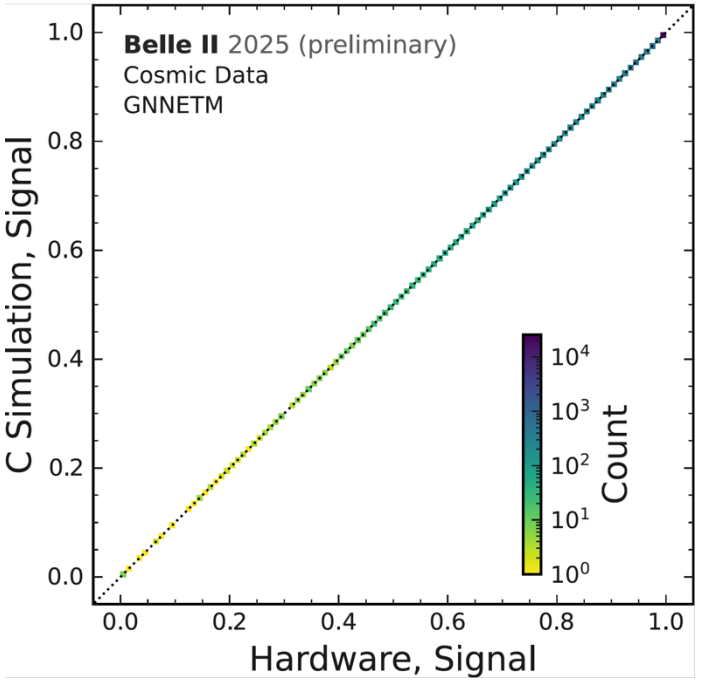
GNN-ETM Comissioning and Operation

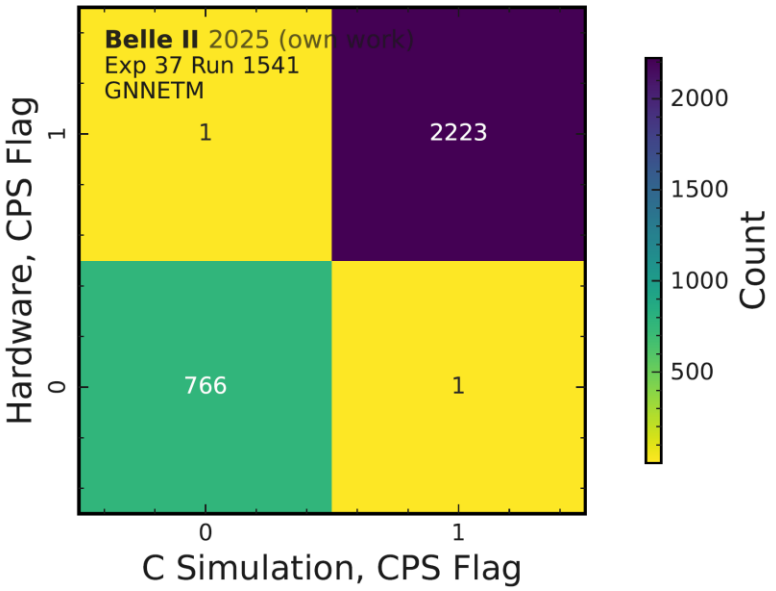
- For comissioning, we validate the following
  - **Data-Level Validation** Compare offline simulation results with recorded debug data from the system installed in the first-level trigger
  - **Reliability Analysis** Stress test of the integrated design (Poisson Trigger Rate)
  - **Trigger Rate Monitoring** Monitor and log the trigger rate during operation. Compare the trigger rate the existing ICN-ETM.



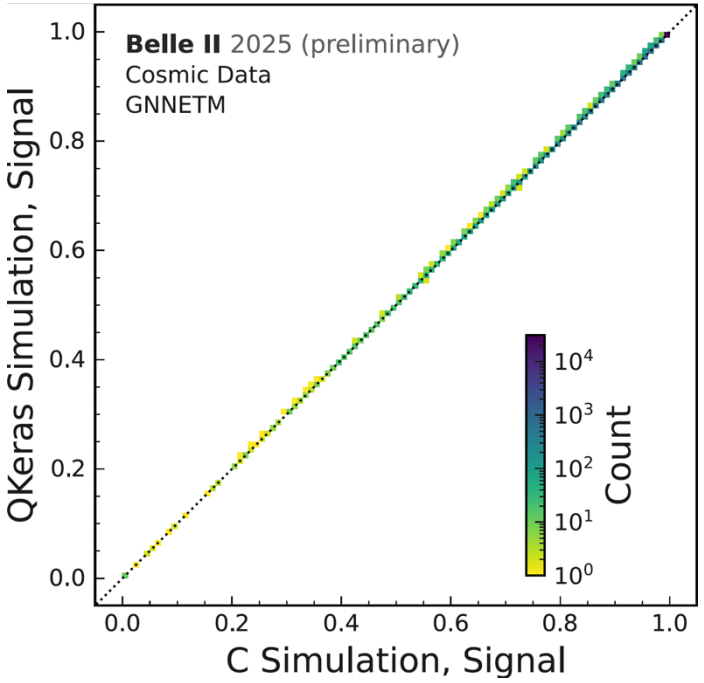
Data Validation

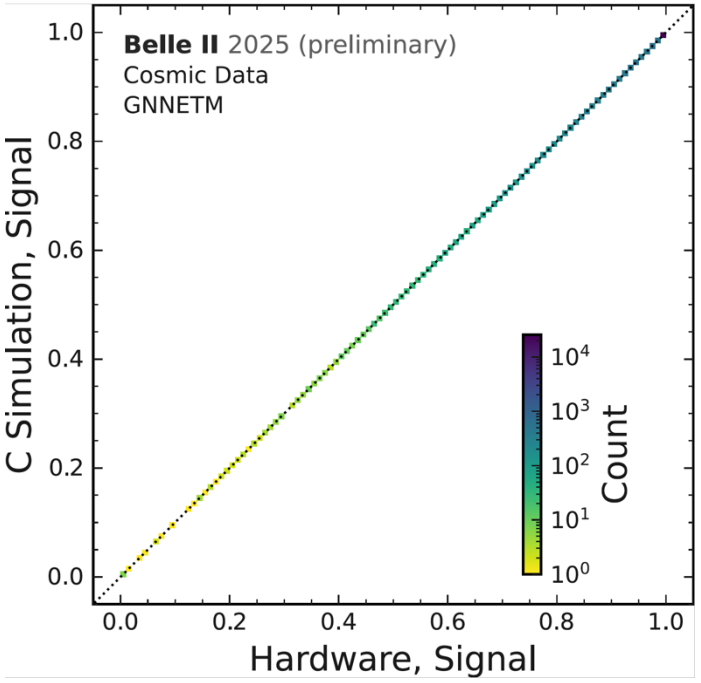


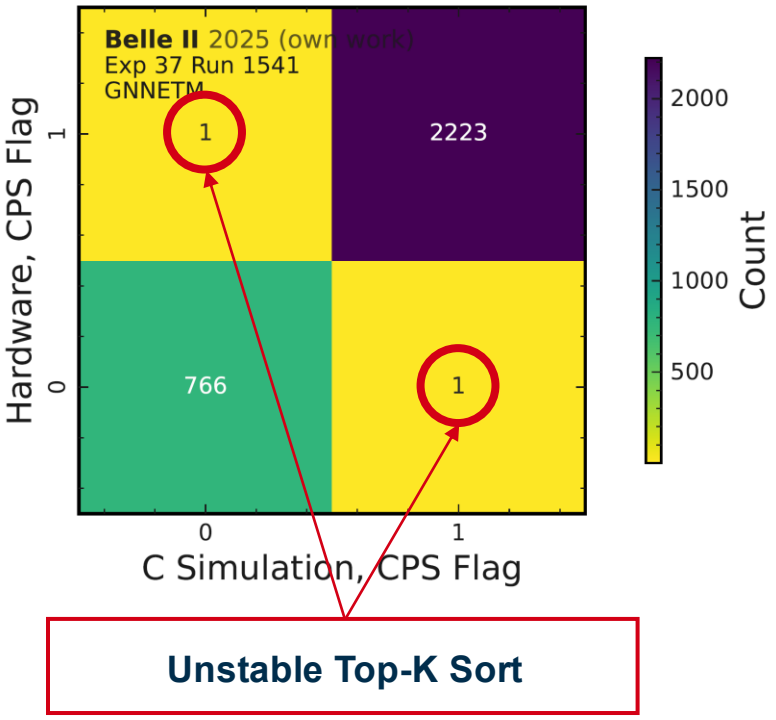




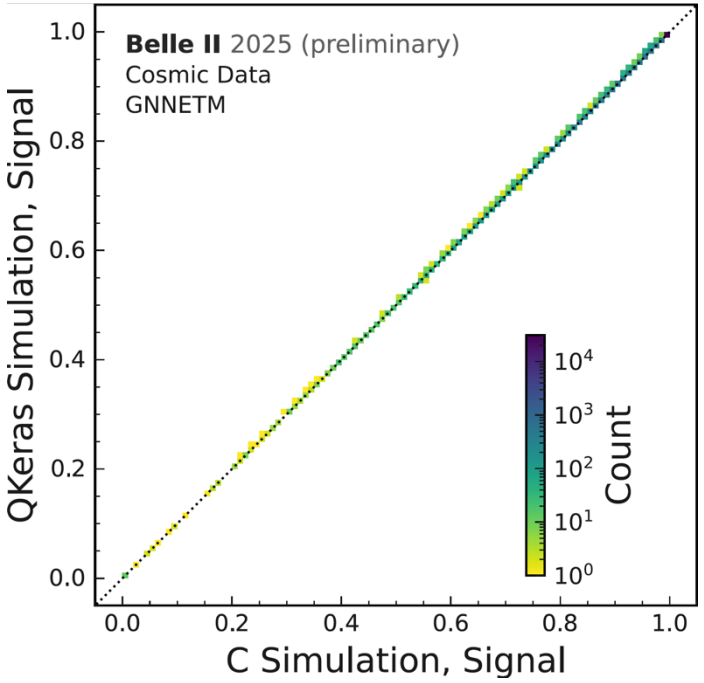
Data Validation

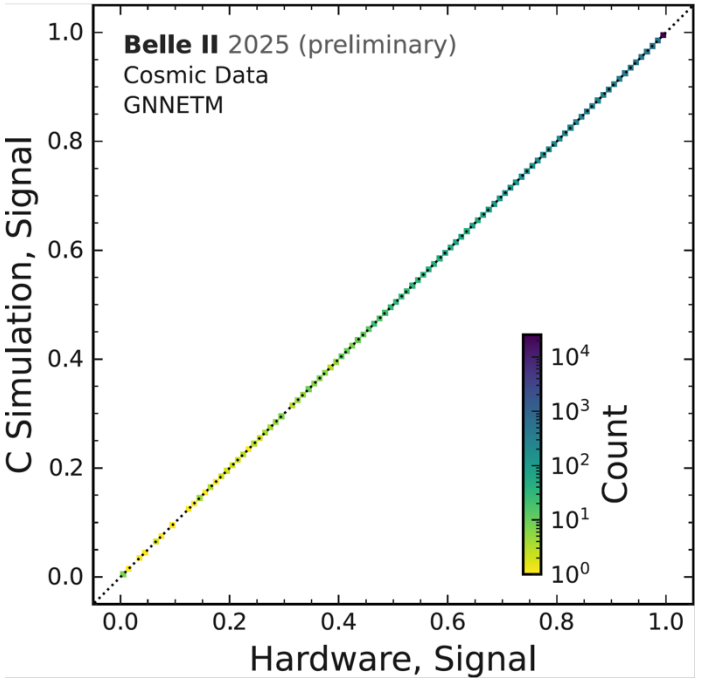


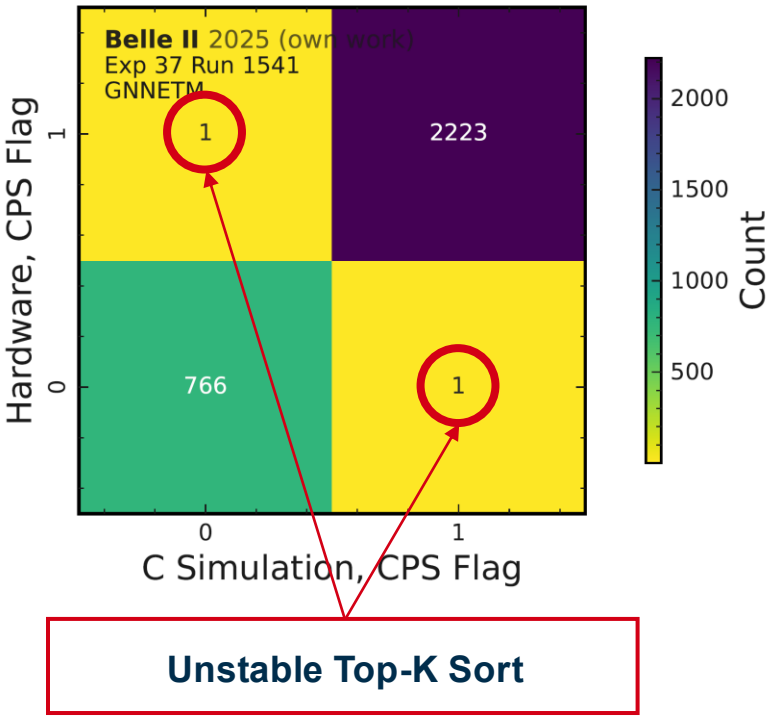




Data Validation



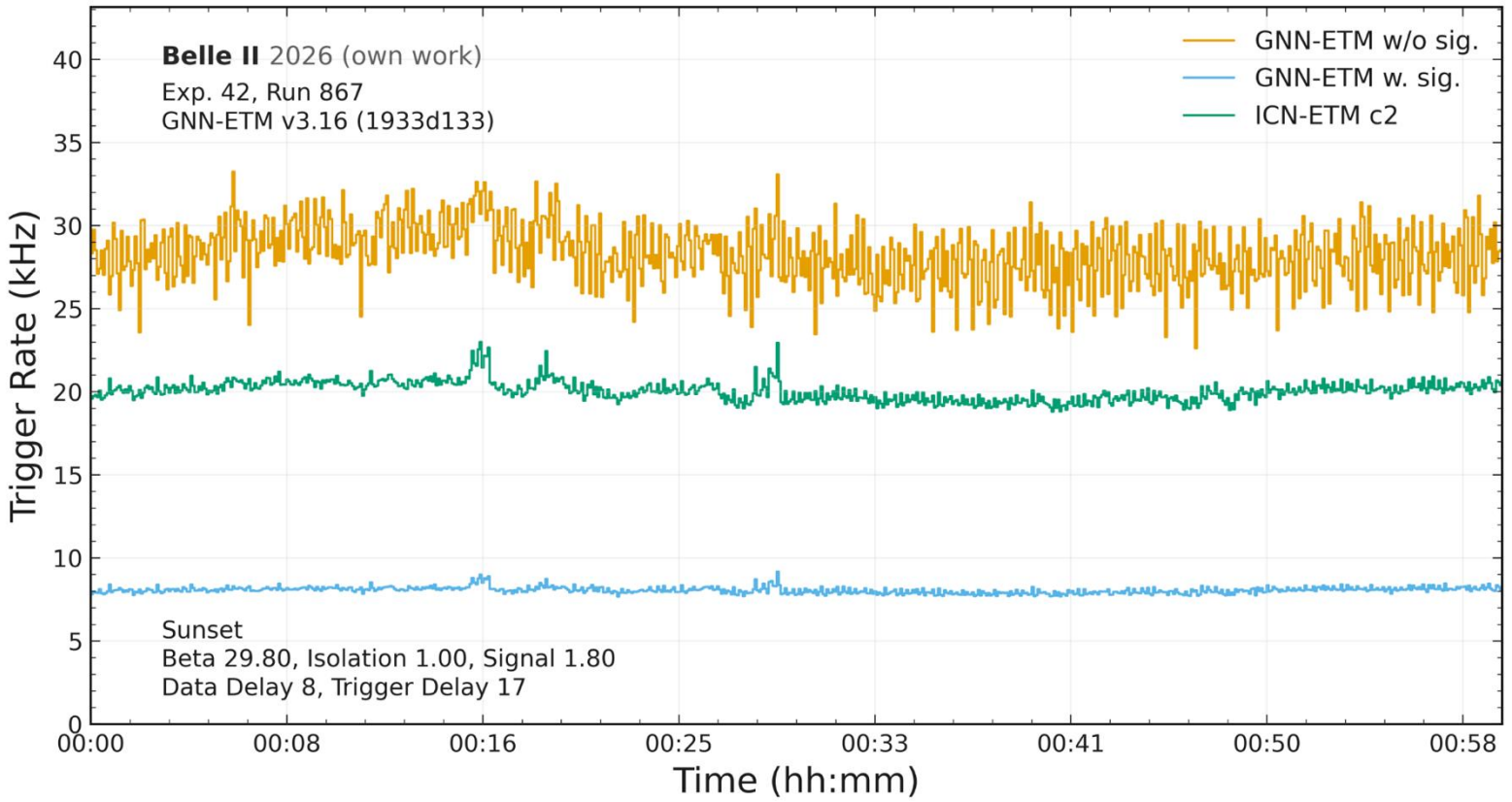




Bit accurate agreement between C simulation and hardware deployment

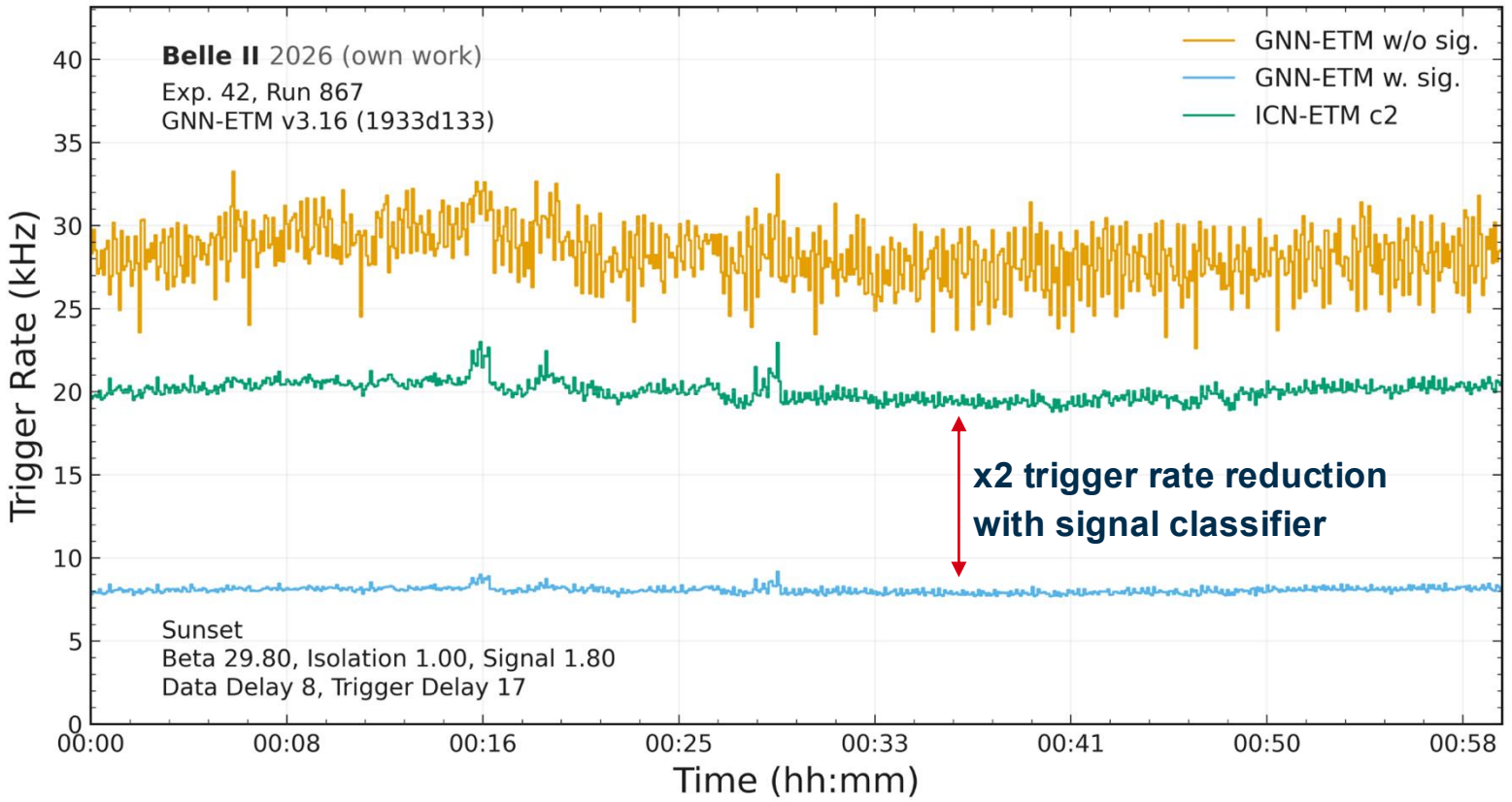
Trigger Rate Monitoring

- For trigger rate monitoring, we choose a simple trigger bit
- „At least two clusters in the inner region of the Belle II ECL detector have an energy above 100MeV“
- Trigger rates are monitored via slow control and stored in the Belle II EPICS database
- We compare trigger rates of the ICN-ETM and the GNN-ETM after applying injection and bhabha vetos



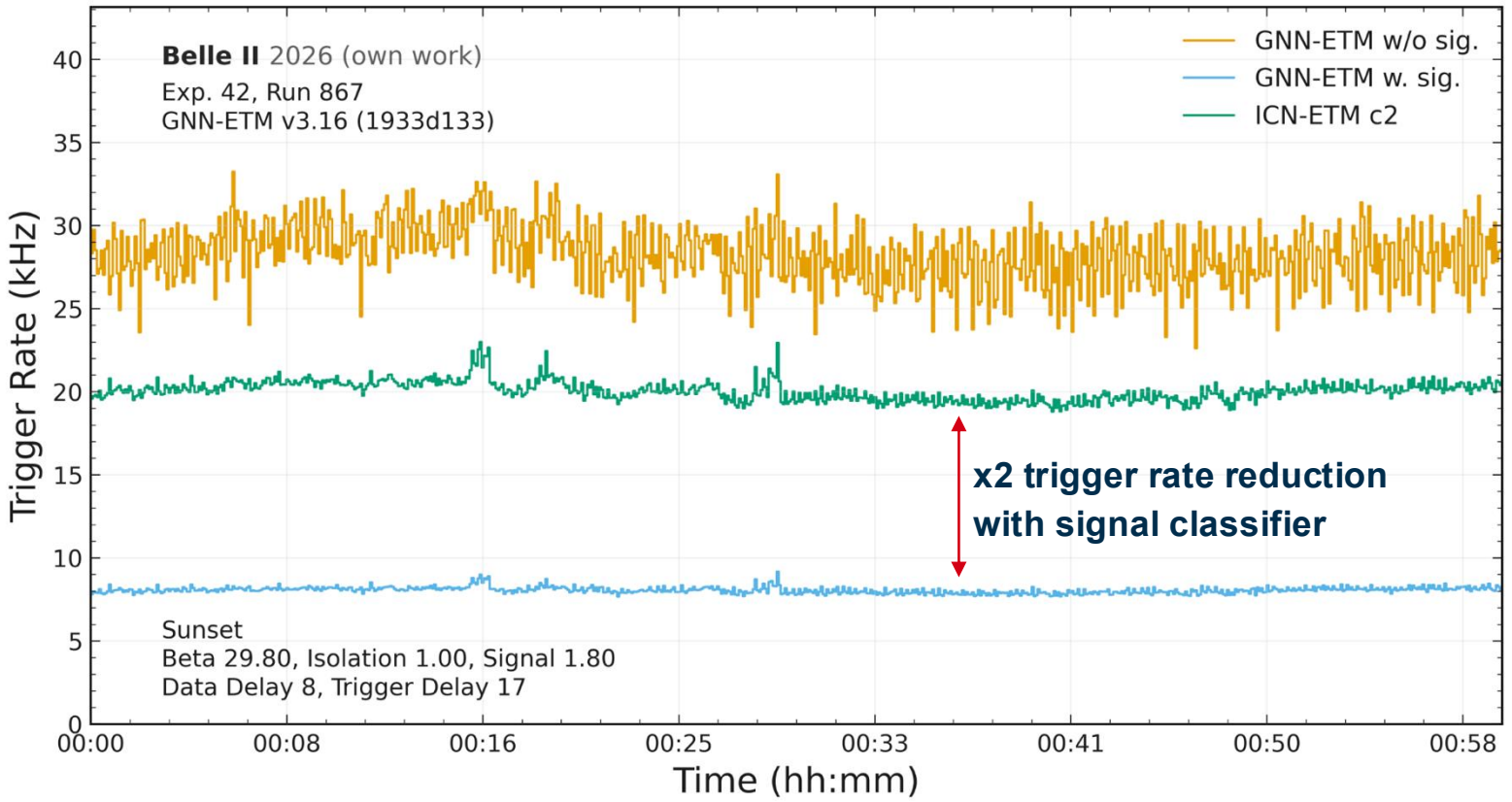
Trigger Rate Monitoring

- For trigger rate monitoring, we choose a simple trigger bit
- „At least two clusters in the inner region of the Belle II ECL detector have an energy above 100MeV“
- Trigger rates are monitored via slow control and stored in the Belle II EPICS database
- We compare trigger rates of the ICN-ETM and the GNN-ETM after applying injection and bhabha vetos



Trigger Rate Monitoring

- For trigger rate monitoring, we choose a simple trigger bit
- „At least two clusters in the inner region of the Belle II ECL detector have an energy above 100MeV“
- Trigger rates are monitored via slow control and stored in the Belle II EPICS database
- We compare trigger rates of the ICN-ETM and the GNN-ETM afte applying injection and bhabha vetos



Monitoring demonstrates a significant reduction of the C2 trigger rate – Validation ongoing

Conclusion

- ✓ We have developed a template-based deployment approach for dynamic GNNs on FPGAs. Our HLS templates are available in our open source kernel library [Neu25].
- ✓ We have designed and implemented a GNN-based clustering algorithm for the Belle II ECL trigger. This is the **first GNN-based reconstruction algorithm** running in a real-time environment at a particle physics experiment [Hai26].
- ✓ We have optimized the GNN-ETM to achieve **1.053us end-to-end latency** and comissioned the system inside the first-level trigger at Belle II, validating its operation.

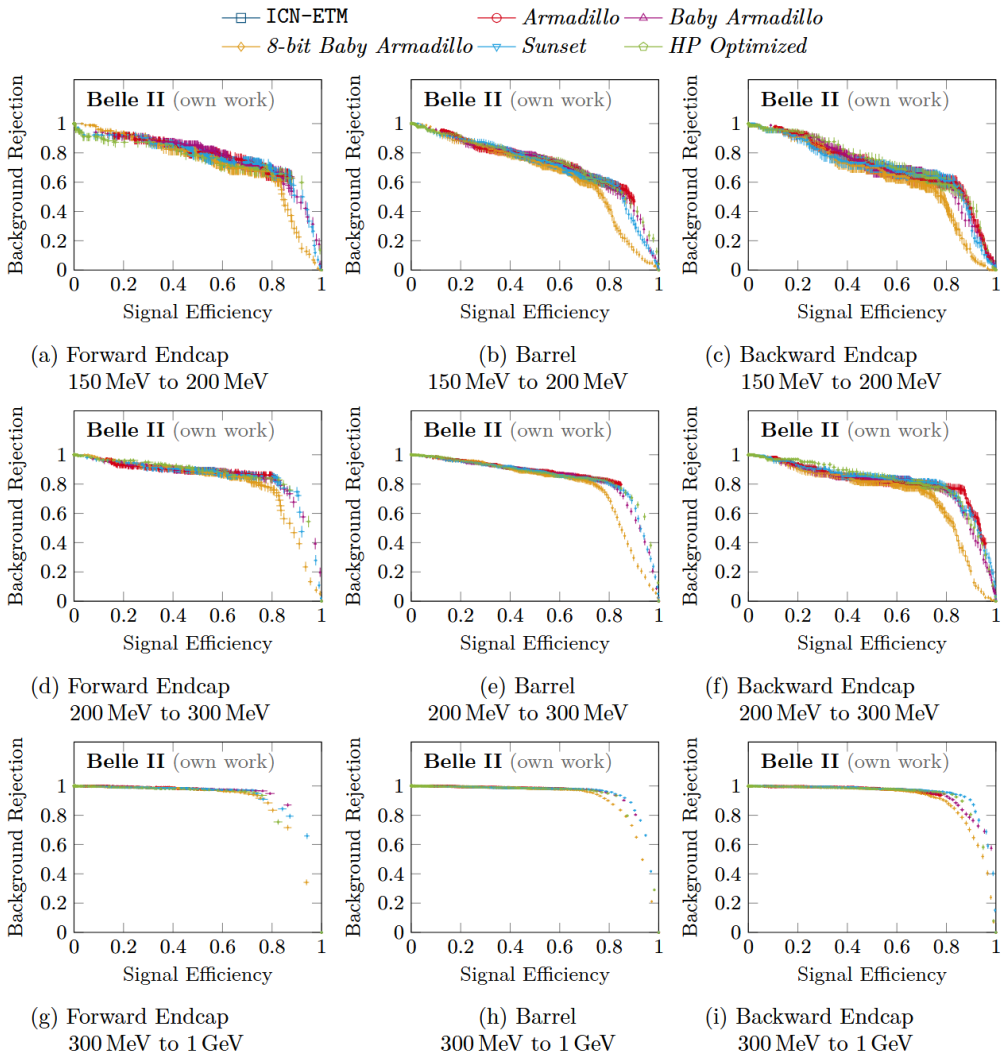
[Neu25] M. Neu et al., "Real-time graph-based point cloud networks on FPGAs via stall-free deep pipelining", IEEE SBCCI, 2025.

[Hai26] I. Haide et al., "Real-time graph neural networks on FPGAs for the Belle II electromagnetic calorimeter" accepted to JINST, 2026.

[Neu26] M. Neu et al., "Reconfigurable Computing Challenge: Real-Time Graph Neural Networks for Online Event Selection in Big Science", IEEE FCCM RCC, 2026.

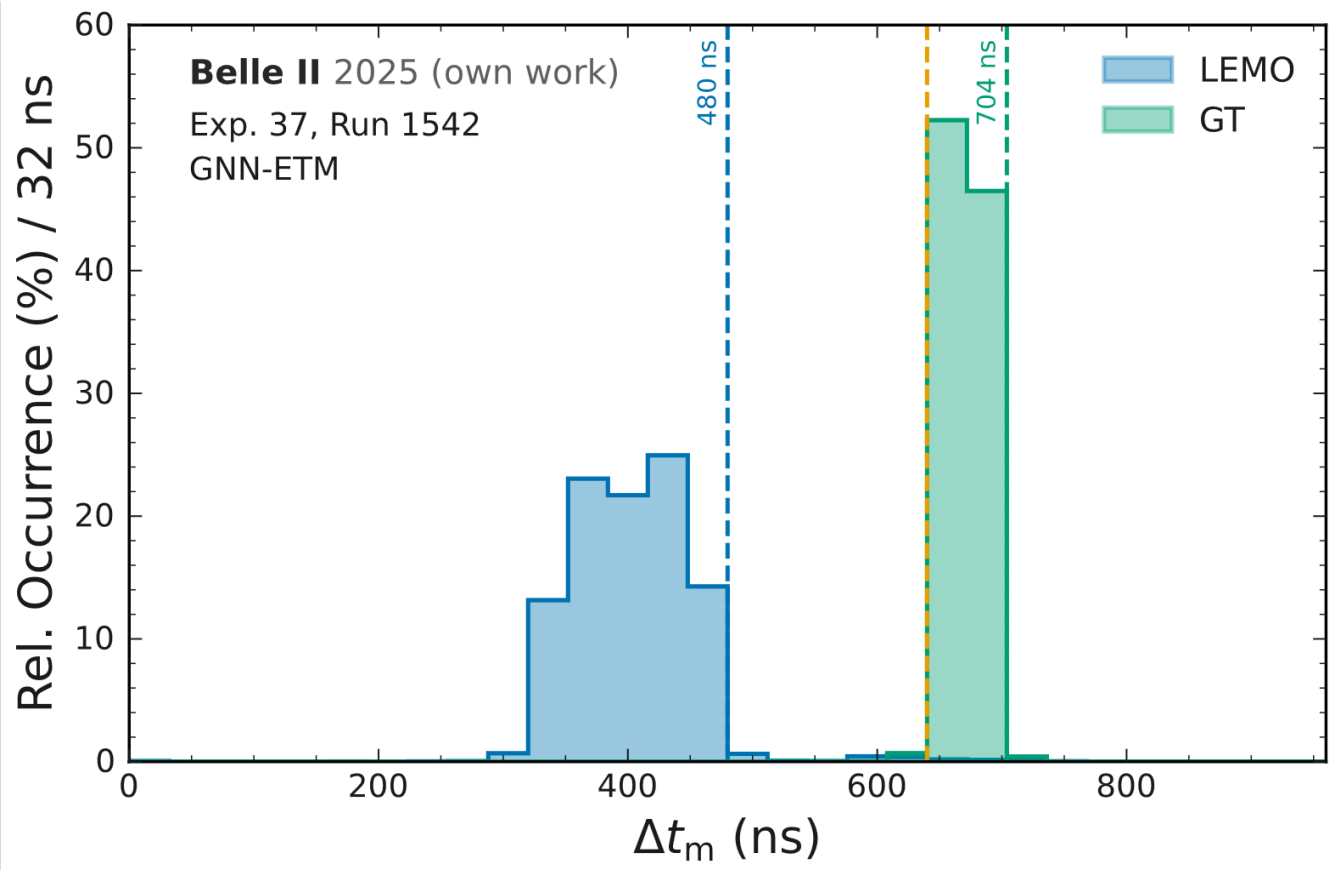
Algorithmic Performance

- Algorithmic performance of CaloClusterNet is continuously monitored to detect regressions
- The ROC curves shown evaluate a single-photon sample with exactly two offline clusters per event
- Energy and position resolution are tracked as additional regression indicators
- The compressed model achieves performance comparable to the baseline
- Multi-target optimization can shift the trade-off between energy and position resolution



F. Baptist, "Training and Optimization of a Graph Neural Network for Deployment on FPGA Hardware in the Belle II Level 1 Trigger," Master's thesis, Institute of Experimental Particle Physics (ETP), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, 2026.

Latency Validation



The Belle II Calorimeter – Crystal Readout