Contribution ID: **87** Type: **Oral**

Hardware-Accelerated GNN Hit Filtering for the Belle II Level-1 Trigger

Friday, 10 October 2025 10:30 (16 minutes)

We present a real-time hit filtering system based on Graph Neural Networks (GNNs), implemented on FPGAs for the Level-1 trigger of Belle II. The system processes raw data from 14,336 sense wires with a sustained throughput of 32MHz and sub-microsecond latency. It combines GNN inference with static graph-building logic in a latency- and resource-optimized FPGA pipeline. This work demonstrates a scalable, low-latency architecture for detector-level background suppression, enabling efficient real-time data reduction in high-rate collider environments.

Summary (500 words)

The Belle II experiment faces growing challenges in trigger rate control as a result of the increasing beam-induced background associated with high luminosity. To address this, we developed a GNN-based real-time hit filtering system, implemented on FPGAs for the Level-1 trigger. Designing such a system for a high-luminosity experiment like Belle II presents several challenges. Key constraints include a fixed maximum latency below 1us, continuous processing at 32 million events per second, and operation on 14,336 raw sense wire channels from the Central Drift Chamber (CDC), without intermediate reconstruction of higher-level objects. These constraints demand hardware-algorithm co-design across the model architecture, dataflow, and firmware.

To meet these goals, we implemented a GNN based on the Interaction Network model and optimized for FPGA deployment through model compression, fixed-point quantization, and static, deterministic graph construction on FPGA. Using operator fusion, we combine edge generation with message-passing logic, eliminating pre-processing overhead and enabling high-throughput, low-latency inference within tight real-time constraints.

Scalability is a central design consideration. The system is parallelized across spatial CDC sectors, each processed independently on a dedicated FPGA. Each sector handles up to 978 sense wires and processes 32 million events per second, aligned with the CDC trigger system clock. This sector-based solution enables full-detector coverage using 20 boards in parallel, handling the large number of sense wires and meeting throughput requirements.

Quantization to fixed-point arithmetic was optimized to maximize inference accuracy within the given hardware constraints. Trade-offs in network depth, width, and network pruning were explored to optimize both classification performance and implementation efficiency.

Our prototype has been implemented on the AMD XCVU160 FPGA used in the 4th generation of the Universal Trigger Board at Belle II, demonstrating full inference within the given latency and throughput constraints. The design is fully compatible with Belle II's Level-1 trigger infrastructure and intended for stable operation under high-rate running conditions.

Beyond performance, this work introduces a reusable framework for deploying real-time GNN-based filtering systems in collider experiments. The architecture is modular, allowing the integration of alternative GNN models or adaptation to other detectors, such as calorimeters or silicon trackers, with similar real-time constraints.

This project demonstrates the practical feasibility of deploying complex machine learning models like GNNs for real-time data reduction in high-energy physics experiments. By addressing both high-background chal-

lenges from the accelerator side and high-throughput and low-latency demands from the detector side, it bridges the gap between high-level machine learning algorithm design and low-level FPGA implementation.

Authors: Mr MAYER, Fabio (KIT); HEINE, Greta Sophie; Prof. BECKER, Juergen (KIT); NEU, Marc; FERBER, Torben (KIT - Karlsruhe Institute of Technology (DE))

Presenter: HEINE, Greta Sophie

Session Classification: Logic

Track Classification: Programmable Logic, Design and Verification Tools and Methods