HiGTR: A High-Performance FPGA Implementation for Complete GNN-Based Trajectory Reconstruction in High-Energy Physics

Yun-Chen Yang[*], Hao-Chun Liang[†], Bo-Cheng Lai[*]
Institute of Electronics[*], Institute of Pioneer Semiconductor Innovation[†]
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
science103555@gmail.com, bclai@nycu.edu.tw

SUMMARY

Charged-particle trajectory reconstruction is a cornerstone of modern collider physics. At the LHC, the CMS silicon tracker captures spatial hits across multiple layers of pixel and strip sensors, which must be associated in real time within the CMS Level-1 trigger (L1T) latency budget of 4 μs. This is achieved in a time-multiplexed architecture, sustaining a per-FPGA throughput of 2.22 MHz. The upcoming HL-LHC upgrade will increase the instantaneous luminosity to $7.5 \times 10^{34}\ \mathrm{cm^{-2}s^{-1}}$, resulting in an average pileup of ~200 interactions per bunch crossing—conditions that exceed the capabilities of conventional CPU/GPU-based solutions and necessitate deterministic-latency FPGA architectures for real-time tracking.

Graph Neural Networks (GNNs)—particularly the Interaction Network—excel at modeling spatial correlations among hits, delivering state-of-the-art reconstruction efficiency. A GNN-based tracker operates in three main stages: (1) graph construction, which maps hits to nodes and generates candidate edges between nearby detector layers; (2) edge classification, which estimates the probability that each edge belongs to a true track; and (3) track building, which assembles high-confidence edges into complete trajectories.

Existing FPGA-based prototypes are limited to edge classification, offloading graph construction and track assembly to external processors—introducing communication overhead and additional latency. We propose HiGTR (High-Performance GNN-based Trajectory Reconstructor), the first fully end-to-end, on-chip GNN tracker. Our key contributions are as follows:

1. **Geometry-Aware Graph Construction:** Introduce a graph constructor that prunes spurious edges early, leveraging geometric constraints to reduce candidate connections.

2. **Scheduling-Optimized GNN Engine:** Develop a fine-grained pipelined GNN architecture with resource sharing that cuts inference latency by 52.3 %.

3. **Lightweight Track Building:** Implement a linear-time track builder that removes clustering overhead, enhancing both resource efficiency and tracking accuracy.

4. **Scalable End-to-End Integration:** Provide three pipeline configurations to support spatial partitioning across multiple FPGAs, ensuring ample performance headroom for diverse HL-LHC scenarios.

Implemented on a Xilinx® Virtex UltraScale™ VU9P, our design delivers a 2.36 μs end-to-end latency and sustains 2.35 MHz throughput on the TrackML dataset—surpassing the HL-LHC Level-1 trigger requirements. Comprehensive benchmarks demonstrate its scalability, establishing this work as a template for next-generation AI acceleration in data-intensive scientific applications.

Keywords: FPGA, Hardware Accelerator, Graph Neural Network, High-Energy Physics, Particle Trajectory Reconstruction