Contribution ID: **109** Type: **Poster**

Real-Time Stream Compaction for Sparse Machine Learning on FPGAs

Machine learning algorithms are being used more frequently in the {first-level triggers in collider experiments}, with Graph Neural Networks (GNNs) pushing the hardware requirements of FPGA-based triggers beyond the current state of the art. {As a first online event processing stage, first-level trigger systems process $O(10\ \mathrm{M})$ events per second with a hard real-time latency requirement of $O(1\ \mathrm{us})$.}
To meet the stringent demands of high-throughput and low-latency environments, we propose a concept for latency-optimized pre-processing of sparse sensor data, enabling efficient GNN hardware acceleration by removing dynamic input sparsity.

Our approach rearranges data coming from a large number of First-In-First-Out (FIFO) interfaces, typically sensor frontends, to a smaller number of FIFO interfaces connected to a GNN hardware accelerator.
In order to achieve high throughput while minimizing the hardware utilization, we developed a hierarchical stream compaction pipeline optimized for FPGAs.

We implemented our concept in the Chisel design language and integrate our open-source package as a parameterizable IP core with FuseSoC. For demonstration, we implemented one configuration of our IP core as pre-processing stage in a GNN-based first-level trigger for the {Electromagnetic Calorimeter (ECL) inside the Belle II detector}. Additionally we evaluate latency, throughput, resource utilization, and scalability for a wide range of parameters, to enable broader use for {other large scale scientific experiments}.

Significance

Our work presents a reusable and scalable data orchestration solution for real-time GNN applications in FPGA-based triggers for particle physics. It enables the hardware acceleration of GNNs on FPGAs in low-latency applications, by eliminating structured sparsity in large heterogeneous sensor arrays. Eliminating sparsity is an important step for real-time algorithms, as it greatly improves compute efficiency and allows for straightforward calculation of worst-case execution times.

References

Experiment context, if any

Belle II Experiment

Author: NEU, Marc

Co-authors: HAIDE, Isabel (Karlsruhe Institute for Technology); BECKER, Juergen; FERBER, Torben (KIT - Karlsruhe Institute of Technology (DE))

Presenter: NEU, Marc

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research