Contribution ID: **16** Type: **not specified**

Accelerating JEDI-net for jet tagging on FPGAs

Monday, 3 October 2022 14:30 (15 minutes)

This work proposes a novel reconfigurable architecture for reducing the latency of JEDI-net, a Graph Neural Network (GNN) based algorithm for jet tagging in particle physics, which achieves state-of-the-art accuracy. Accelerating JEDI-net is challenging since low latency is requited to potentially deploy the network on the online event selection systems at the CERN Large Hadron Collider. This presentation proposes a custom code transformation with strength reduction for matrix multiplication operations which avoids the costly multiplication of the adjacency matrix with the input feature matrix. It exploits sparsity patterns as well as binary adjacency matrices, and avoids irregular memory access, leading to a reduction in latency and improvement of hardware efficiency. We also introduce an outer-product based matrix multiplication approach which is enhanced by the strength reduction for low-latency design. Furthermore, a customizable template for this architecture has been designed and open-sourced, which enables the generation of low-latency FPGA designs with efficient resource utilization using high-level synthesis tools. Evaluation results show that our FPGA implementation is up to 9.5 times faster and consumes up to 6.5 times less power than a GPU implementation. Moreover, the throughput and latency of our FPGA design is sufficiently high to enable deployment of JEDI-net in a sub-microsecond, real-time collider trigger system, enabling it to benefit from improved accuracy.

Primary authors: QUE, Zhiqiang (Imperial College London); Prof. TAPPER, Alexander D (Imperial College London); Prof. LUK, Wayne (Imperial College London)

Presenter: QUE, Zhiqiang (Imperial College London)

Session Classification: Contributed Talks