Contribution ID: **164** Type: **Oral presentation**

hls4ml: low latency neural network inference on FPGAs

Tuesday 23 April 2024 14:50 (20 minutes)

Machine learning is becoming increasingly prevalent in High Energy Physics (HEP), offering significant potential for enhancing trigger and Data Acquisition (DAQ) performance, as well as other real-time control applications. However, the exploration of these techniques in low latency/power Field-Programmable Gate Arrays (FPGAs) is still in its early stages. We introduce hls4ml, a user-friendly software based on High-Level Synthesis (HLS), designed specifically for deploying network architectures on FPGAs. To demonstrate the features of hls4ml we will show several case studies at the Large Hadron Collider (LHC) and analyze resource usage and latency in relation to different network architectures. Additionally, we report on the progress of new developments in hls4ml, particularly focusing on newer neural network architectures graph neural networks, transformers, symbolic regression, support for QONNX and discuss their potential for use in future HEP applications and beyond.

Minioral

No

IEEE Member

No

Are you a student?

No

Author: LONCAR, Vladimir (Massachusetts Inst. of Technology (US))

Presenter: LONCAR, Vladimir (Massachusetts Inst. of Technology (US))

Session Classification: Oral presentations

Track Classification: AI, Machine Learning, Real Time Simulation, Intelligent Signal Processing