Contribution ID: **92** Type: **Oral**

Empowering AI Implementation: The Versatile SLAC Neural Network Library (SNL) for FPGA, eFPGA , ASIC

Tuesday, 7 November 2023 16:20 (20 minutes)

This paper presents the SLAC Neural Network Library (SNL), a specialized set of extensible libraries designed in High-Level Synthesis (HLS) for deploying machine learning structures on Field Programmable Gate Arrays (FPGAs), eFPGAs and ASICs. Positioned at the edge of the data chain, SNL aims to create a high-performance, low-latency FPGA implementation for AI inference engines. Utilizing the Xilinx's High-Level Synthesis (HLS) framework, SNL offers an API modeled after the widely used Keras interface to TensorFlow. The primary objective of SNL is to deliver a high-performance, low-latency FPGA implementation of an AI inference engine capable of handling moderately sized networks. SNL allows for dynamic reloading of weights and biases without re-synthesis, enhancing adaptability, and facilitating experimentation. Moreover, SNL supports a modular approach, enabling the implementation of novel and custom ML layers for FPGAs and ASICs. The framework facilitates a standard interface for storing weights and biases, such as HDF5. SNL not only demonstrates its capability to attain higher data throughput but also contributes to meeting experiment-specific latency constraints.

Early Career

Primary authors: DAVE, Abhilasha (SLAC); Dr DRAGONE, Angelo (SLAC); DOERING, Dionisio (SLAC); RUSSELL, J.J. (SLAC); RUCKMAN, Larry (SLAC); COFFEE, Ryan (SLAC); HERBST, Ryan (SLAC)

Presenter: DAVE, Abhilasha (SLAC)

Session Classification: RDC5

Track Classification: RDC Parallel Sessions: RDC5: Trigger and DAQ