Contribution ID: **381** Type: **Oral**

Highly performant, Deep Neural Networks with sub-microsecond latency on FPGAs for trigger applications

Thursday, 14 March 2019 16:30 (20 minutes)

Artificial neural networks are becoming a standard tool for data analysis, but their potential remains yet to be widely used for hardware-level trigger applications. Nowadays, high-end FPGAs, as they are also often used in low-level hardware triggers, offer enough performance to allow for the inclusion of networks of considerable size into these system for the first time. Nevertheless, in the trigger context, it is necessary to highly optimize the implementation of neural networks to make full use of the FPGA capabilities.

We implemented the processing data and control flow of typical NN layers, taking into account incoming data rates of up to multiple tens of MHz and sub-microsecond latency limits, but also aiming at an efficient use of the resources of the FPGA. This resulted in a highly optimized neural network implementation framework, which typically reaches 90 to 100 % computational efficiency, requires few extra FPGA resources for data flow and controlling, and achieves latencies in the order of only tens to few hundreds of nanoseconds for entire (deep) networks. The implemented layers include 2D convolutions and pooling (both with multi-channel support), as well as dense layers, all of which play a role in many physics-/detector-related applications. Significant effort was put especially into the 2D convolutional layers, to achieve a fast implementation with minimal resource usage.

A toolkit is provided which automatically creates the optimized FPGA implementation of trained deep neural network models. Results are presented, both for individual layers as well as entire networks created by the toolkit.

Primary authors: NOTTBECK, Noel Aaron (Johannes Gutenberg Universitaet Mainz (DE)); BUESCHER, Volker (Johannes Gutenberg Universitaet Mainz (DE)); SCHMITT, Christian (Johannes Gutenberg Universitaet Mainz (DE))

Presenter: NOTTBECK, Noel Aaron (Johannes Gutenberg Universitaet Mainz (DE))

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research