Contribution ID: **46** Type: **Oral**

Embedded Neural Networks on FPGAs for Real-Time Computation of the Energy Deposited in the ATLAS Liquid Argon Calorimeter

Wednesday 21 September 2022 12:00 (20 minutes)

At the HL-LHC, the number of proton-proton collisions in one bunch-crossing (called pileup) increases significantly, putting more stringent requirements on the LHC detectors electronics and real-time data-processing capabilities. The ATLAS LAr calorimeter measures with an excellent resolution the energy of particles produced in LHC collisions. The energy is computed in real-time using optimal filtering (OF) algorithms running on dedicated data-acquisition electronic boards based on FPGAs. However, with the increased pileup, the performance of these algorithms decreases significantly. Dedicated Neural networks (NNs) are found to outperform the OF algorithms. The architecture, performance, and firmware implementation for these NNs will be presented.

Summary (500 words)

The Phase-II upgrade of the LHC will increase its instantaneous luminosity by a factor of 7 leading to the HL-LHC. The ATLAS Liquid Argon (LAr) calorimeter measures the energy of particles produced in LHC collisions. In order to enhance the ATLAS physics discovery potential in the blurred environment created by the pileup, it is crucial to have an excellent energy resolution and an accurate detection of the energy-deposit time.

The energy computation is currently done using optimal filtering algorithms that assume a nominal pulse shape of the electronic signal. Up to 200 simultaneous proton-proton collisions are expected at the HL-LHC which leads to a high rate of overlapping signals in a given calorimeter channel. This results in a significant energy degradation especially for low time-gap between two consecutive pulses (figure *1). We developed several neural network (NN) architectures showing significant performance improvements with respect to the filtering algorithms. These NNs are capable to recover the degraded performance in the low-time gap region by using the information from past events as shown in figure* 1.

The energy computation is performed in real-time using dedicated electronic boards based on FPGAs. FPGAs are chosen for their capacity to treat large amount of data (O(1Tb/s) per FPGA) with low latency (O(1000ns)). The back-end electronic boards for the Phase-II upgrade of the LAr calorimeter will use the next high-end generation of INTEL FPGAs with increased processing power. This is a unique opportunity to develop more complex algorithms on these boards. Several hundreds of channels should be treated by each FPGA and thus several hundreds of NNs should run on one FPGA. The energy computation should be done at a fixed latency of the order of 100 ns. The main challenge is to meet these stringent requirements in the firmware implementation.

Special effort was dedicated to minimize the needed computational operations while optimizing the NNs architectures. Each internal operation of the NNs is optimized during the firmware implementation. This includes complex mathematical functions implementation in LookUp Tables (LUTs), quantization of arithmetic operations using fixed-point representations and rounding, and optimisation of the usage of FPGA logic elements (figure*2). The firmware implementation results are compared to software and the resolution due to firmware approximations was found to be around 1% (figure* 3).

Five NN algorithms based on CNN, RNN, and LSTM architectures will be presented. The improvement of the energy resolution compared to the legacy filter algorithms will be discussed. The results of firmware

implementation in VHDL and Quartus HLS will be presented. The implementation results on Stratix 10 INTEL FPGAs, including the resource usage, latency, and operation frequency will be reported. Additionally a test on a Stratix 10 INTEL development kit of the RNN implementation will presented. Up to 36 RNN instances are fitted on the FPGA with a time multiplexing of 10 and no timing violation in the firmware. This allows to treat 370 channels on one FPGA assuming no other tasks are required on the FPGA.

*The figures are attached in the additional material.

Author: AAD, Georges (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR))

Presenter: AAD, Georges (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR))

Session Classification: Programmable Logic, Design Tools and Methods

Track Classification: Programmable Logic, Design Tools and Methods