Contribution ID: **51** Type: **Oral Presentation**

Exploring Deep Learning fast inference on an ATCA processor board with Xilinx Virtex-7 FPGAs

Thursday, 27 August 2020 17:00 (30 minutes)

Machine and Deep Learning techniques experienced an explosion in the adoption in a variety of HEP applications, ranging from event selection to end-user physics data analysis, as well as computing metadata based optimizations.
The range of applicability of such techniques in the High Energy Physics (HEP) context – with a particular accent on experiments at the Large Hadron Collider (LHC) at CERN in Geneva – will extend to an even larger variety of applications, if low-latency hardware solutions are added to usually exploited (CPUs/)GPUs - or even Google TPUs.
One example area of application in particle physics is the domain of FPGA-based Trigger/DAQ, characterized by even sub-microsecond latency requirements: stringent requirements towards this solution come from the upcoming Run-3 needs and even more from the evolution towards the operational conditions foreseen for the High-Luminosity LHC (HL-LHC) phase.
Crucial ingredients to prepare for this future are the availability of adequate hardware resources and expertize, as well as the capability to streamline the process of build and testing ML/DL models into FPGA firmware.
This talk will present the work done and planned ahead in the University of Bologna and the INFN-Bologna cross-experiment working groups. The team is working on a customized ATC136 board hosting a Xilinx Virtex-7 FPGA with full-mesh backplane fabric connectivity, mounted on a ATCA crate installed in the INFN-CNAF Tier-1 data center. The High-Level Synthesis (HLS) toolkit was used - called hls4ml – developed very close to the HEP needs.
Hardware and software set-up, and performances on various baseline models used as benchmarks, will be presented. Real-life case studies for specific deep neural networks developed in the context of future evolutions of LHC Trigger systems will be also presented, and the possible advantages of performing neural network inference with FPGAs for this class of problems will be discussed.

Primary author: Mr DIOTALEVI, Tommaso (INFN and University of Bologna)

Co-author: Prof. BONACORSI, Daniele (University of Bologna)

Presenter: Mr DIOTALEVI, Tommaso (INFN and University of Bologna)

Session Classification: Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

Track Classification: Converging High Performance infrastructures: Supercomputers, clouds, accelerators