Contribution ID: **52** Type: **Oral Presentation**

Deep Learning fast inference on FPGA for CMS Muon Level-1 Trigger studies

Friday, 26 March 2021 14:00 (30 minutes)

Machine and Deep Learning techniques experienced an explosion in the adoption of a variety of HEP applications, ranging from event selection in trigger operations to end-user physics data analysis, as well as computing metadata based optimisations.
The range of applicability of such techniques in the High Energy Physics (HEP) context – with a particular accent on the 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 the ones usually exploited (CPUs/)GPUs - or even Google TPUs.
An area of application in particle physics, for instance, is the domain of FPGA-based Trigger/DAQ, characterized by even sub-microsecond latency requirements: stringent prerequisites 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 required to be prepared for this future are the availability of adequate hardware resources and expertise, as well as the capability to streamline the process of building and testing ML/DL models into FPGA firmware.

This paper presents and discusses the activity running at the University of Bologna and INFN-Bologna: pioneered by members of the CMS experiment together with a cross-experiment working group, including LHC physicists and electronics experts, the work is focused on the development of Neural Network models for HEP use-cases as well as prototyping the exploitation of various FPGA resources for the model inference. In particular, the target hardware used consisted in a Xilinx ZCU102 Evaluation Board featuring a Zynq Ultrascale+ MPSoC, mounted in a test bed controlled environment. To synthesize the high-level network architectures, written in Python, in a language oriented to a more circuital description (HDL), the High-Level Synthesis (HLS) toolkit was used - called HLS4ML - developed very closely to the HEP needs.
Hardware and software set-up, and performances on various baseline models used as benchmarks, will be presented. In particular, a specific real-life case study on the CMS Muon Level-1 Trigger will be presented, implementing a Deep Neural Network for the assignment of muon transverse momentum (pT) in an FPGA readable firmware. The possible advantages of performing such neural networks on dedicated programmable hardware, in terms of latency reduction as well as the tuning of inference performances with parallelisation and quantisation capabilities, for this class of problems will also be discussed.
As a future development for this study, a comparison in terms of performance and resource usage can be performed, using different hardware solutions; in particular, a newly acquired Vadatech ATC136 board hosting a Xilinx Virtex-7 FPGA, mounted on a ATCA crate and installed in the INFN-CNAF Tier-1 data center.

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

Co-authors: Dr BATTILANA, Carlo (University of Bologna and INFN); Prof. BONACORSI, Daniele (University of Bologna and INFN); Dr LORUSSO, Marco (University of Bologna); Dr TRAVAGLINI, Riccardo (INFN Bologna)

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

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

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