Contribution ID: **1297** Type: **Parallel Talk**

Accelerating Machine Learning inference using FPGAs: the PYNQ framework tested on an AWS EC2 F1 Instance

Friday, 8 July 2022 14:30 (15 minutes)

In the past few years, using Machine and Deep Learning techniques has become more and more viable, thanks to the availability of tools which allow people without specific knowledge in the realm of data science and complex networks to build AIs for a variety of research fields. This process has encouraged the adoption of such techniques: in the context of High Energy Physics, new algorithms based on ML are being tested for event selection in trigger operations, end-user physics analysis, computing metadata based optimizations, and more. Time critical applications can benefit from implementing algorithms on low-latency hardware like specifically designed ASICs and programmable micro-electronics devices known as FPGAs. The latter offers a unique blend of the benefits of both hardware and software. Indeed, they implement circuits just like hardware, providing power, area and performance benefits over software, yet they can be reprogrammed cheaply and easily to implement a wide range of tasks, at the expense of performance with respect to ASICs. In order to facilitate the translation of ML models to fit in the usual workflow for programming FPGAs, a variety of tools have been developed. One example is the HLS4ML toolkit, developed by the HEP community, which allows the translation of Neural Networks built using tools like TensorFlow to a High-Level Synthesis description (e.g. C++) in order to implement this kind of ML algorithms on FPGAs.
This paper presents and discusses the activity started at the Physics and Astronomy department of University of Bologna and INFN-Bologna devoted to preliminary studies for the trigger systems of the Compact Muon Solenoid (CMS) experiment at the CERN LHC accelerator. A broader-purpose open-source project from Xilinx (a major FPGA producer) called PYNQ is being tested combined with the HLS4ML toolkit. The PYNQ purpose is to grant designers the possibility to exploit the benefits of programmable logic and microprocessors using the Python language. This software environment can be deployed on a variety of Xilinx platforms, from IOT devices like the ZYNQ-Z1 board, to the high performance ones, like Alveo accelerator cards and on the cloud AWS EC2 F1 instances. The use of cloud computing in this work allows us to test the capabilities of this workflow, from the creation and training of a Neural Network and the creation of a HLS project using HLS4ML, to managing NN inference with custom Python drivers.
The main application explored in this work lives in the context of the trigger system of the CMS, where new reconstruction algorithms are being developed due to the advent of the High-Luminosity phase of the LHC (HL-LHC) and the related increase in instantaneous luminosity. Indeed, Machine Learning techniques have already shown promising results in the measurement of transverse momentum of muons at the Level-1 trigger. By implementing such a model on an FPGA, we can take advantage of the smaller latencies with respect to traditional inference algorithms running on CPU.

This work is also part of a project aimed at creating and offering an FPGA-as-a-Service as part of the INFN Cloud service catalogue. In this context, new Alveo boards have been purchased by INFN in order to test the feasibility of the proposal and to compare the performance obtained using FPGAs made available on INFN Cloud, with those obtained by using third party cloud computing, in particular Amazon EC2 F1 instances.

Hardware and software set-up, together with performance tests on various baseline models used as benchmarks, will be presented.The presence or not of some overhead causing an increase in latency will be investigated. Eventually, the consistency in the predictions of the NN, with respect to a more traditional way of interacting with the FPGA using C++ code, will be verified.

In-person participation

Yes

Primary authors: BONACORSI, Daniele (Istituto Nazionale di Fisica Nucleare); LORUSSO, Marco (Istituto Nazionale di Fisica Nucleare); SALOMONI, Davide (Istituto Nazionale di Fisica Nucleare); MICHELOTTO, Diego (CNAF); DUMA, Doina Cristina (Istituto Nazionale di Fisica Nucleare); VERONESI, Paolo (INFN-BOLOGNA); TRAVAGLINI, Riccardo (Istituto Nazionale di Fisica Nucleare)

Presenter: LORUSSO, Marco (Istituto Nazionale di Fisica Nucleare)

Session Classification: Computing and Data handling

Track Classification: Computing and Data handling