Contribution ID: **44** Type: **Talk**

Building the tools to run large scale machine learning with FPGAs with two new approaches: AIGEAN and FAAST

Tuesday 1 December 2020 15:28 (8 minutes)

FPGA programming is becoming easier as the vendors begin to provide environments, such as for machine learning (ML), that enable programming at higher levels of abstraction.The vendor platforms target FPGAs in a single host server.To scale to larger systems of FPGAs requires communication through the hosts, which has a significant impact on performance. We demonstrate the deployment of ML algorithms on single FPGAs through FAAST a newly developed FPGA based infrastructure framework. We also present a new Framework, AIGEAN, to run multiple FPGA and CPU heterogeneous system that can leverage direct FPGA-to-FPGA communication links. AIgean and FAAST, take as input an ML algorithm created with a standard ML framework and a specification of the available FPGA and CPU resources. The outputs are software and hardware cores that can compute one or more ML layers. These layers can be distributed across a heterogeneous cluster of CPUs and FPGAs for execution. As part of this work we present an optimized FPGA implementation of a CNNs. We show that in some cases FPGAs can exceed the performance of other accelerators, including GPUs.

Authors: RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Presenter: TARAFDAR, Naif (University of Toronto)