Contribution ID: **69** Type: **Lightning talk**

GPU and FPGA as a Service for Machine Learning Inference Accelerations

Friday 23 October 2020 15:15 (5 minutes)

The data rate may surge after some planned upgrades for the high-luminosity Large Hadron Collider (LHC) and accelerator-based neutrino experiments. Since there is no enough storage to save all of the data, there is a challenging demand to process and filter billions of events in real-time. Machine learning algorithms are becoming increasingly prevalent in the particle reconstruction pipeline. Specially designed hardware can significantly accelerate the machine learning inference time compared to CPUs. Thus, we propose a heterogeneous computing framework called the Services for Optimized Network Inference on Coprocessors (SONIC) to accelerate machine learning inferences with various coprocessors. With a unified interface, the framework conveniently provides GPU as a service, using either the Nvidia Triton framework or the Microsoft Brainwave service as the backend. It also features the first open-source FPGA-as-a-service toolkit, using either our hls4ml framework or the Xilinx ML Suite as the backend. We demonstrated that our method could speed up one classification and two regression problems in the LHC experiments and ProtoDUNE-SP. By providing coprocessors as a service, our work may assist various other computing workflows across science.

Authors: LOU, Yu (University of Washington (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); KRUPA, Jeffrey; LIN, Kelvin (University of Washington (US)); PEDRO, Kevin (Fermi National Accelerator Lab. (US)); Dr KNOEPFEL, Kyle (Fermi National Accelerator Laboratory); ACOSTA FLECHAS, Maria (Fermi National Accelerator Lab. (US)); TRAHMS, Matthew (UW ACME Lab); LIU, Mia; WANG, Michael (Fermi National Accelerator Lab. (US)); SUAYSOM, Natchanon (University of Washington (US)); TRAN, Nhan Viet (Fermi National Accelerator Lab. (US)); HARRIS, Philip (Unknown); HAUCK, Scott (University of Washington); HSU, Shih-Chieh (University of Washington Seattle (US)); HO, Ta-Wei (National Tsing Hua University (TW)); KLIJNSMA, Thomas (Fermi National Accelerator Lab. (US)); YANG, Tingjun (Fermi National Accelerator Lab. (US)); HAWKS, Benjamin (Fermi National Accelerator Laboratory); Dr HOLZMAN, Burt (Fermi National Accelerator Lab. (US)); RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US)); DINSMORE, Jack

Presenter: LOU, Yu (University of Washington (US))

Session Classification: Workshop

Track Classification: 6 ML infrastructure : Hardware and software for Machine Learning