Contribution ID: **177** Type: **Oral**

FPGA-accelerated machine learning inference as a service for particle physics computing

Tuesday, 5 November 2019 14:00 (15 minutes)

Large-scale particle physics experiments face challenging demands for high-throughput computing resources both now and in the future. New heterogeneous computing paradigms on dedicated hardware with increased parallelization, such as Field Programmable Gate Arrays (FPGAs), offer exciting solutions with large potential gains. The growing applications of machine learning algorithms in particle physics for simulation, reconstruction, and analysis are naturally deployed on such platforms. We demonstrate that the acceleration of machine learning inference as a web service represents a heterogeneous computing solution for particle physics experiments that requires minimal modification to the current computing model. As examples, we retrain the ResNet50 convolutional neural network to demonstrate state-of-the-art performance for top quark jet tagging at the LHC and apply a ResNet50 model with transfer learning for neutrino event classification. Using Microsoft Azure Machine Learning deploying Intel FPGAs to accelerate the ResNet50 image classification model, we achieve average inference times of 60 (10) milliseconds with our experimental physics software framework deployed as a cloud (edge or on-premises) service, representing an improvement by a factor of approximately 30 (175) in model inference latency over traditional CPU inference in current experimental hardware. A single FPGA service accessed by many CPUs achieves a throughput of 600-700 inferences per second using an image batch of one, comparable to large batch-size GPU throughput and significantly better than small batch-size GPU throughput. Deployed as an edge or cloud service for the particle physics computing model, coprocessor accelerators can have a higher duty cycle and are potentially much more cost-effective.

Consider for promotion

Yes

Primary authors: TRAN, Nhan Viet (Fermi National Accelerator Lab. (US)); PEDRO, Kevin (Fermi National Accelerator Lab. (US))

Presenter: PEDRO, Kevin (Fermi National Accelerator Lab. (US))

Session Classification: Track X – Crossover sessions

Track Classification: Track X – Crossover sessions from online, offline and exascale