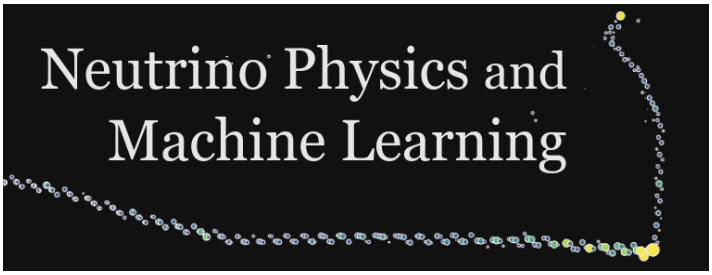

Contribution ID: **37** Type: **Individual talk**

GPU as a Service for Accelerating Machine Learning Applications in the Reconstruction Workflows of Neutrino Experiments

Wednesday, 22 July 2020 13:40 (25 minutes)

The employment of machine learning (ML) techniques has now become commonplace in the offline reconstruction workflows of modern neutrino experiments. Since such workflows are typically run on CPU-based high-througput computing (HTC) clusters with limited or no access to ML accelerators like GPU or FPGA coprocessors, the ML algorithms, for which CPUs are not the best suited platform, tend to dominate the total computational time of the workflows. In this talk we explore a computing model that provides GPUs as a Service (GPUaaS), where ML algorithms in offline neutrino reconstruction workflows running on typical HTC clusters can send inference requests to and receive the results from remote GPU-based inference servers running in the cloud, in a completely seamless fashion. We demonstrate a proof-of-principle using the full ProtoDUNE reconstruction chain, where we are able to acclerate the ML portion of the workflow by more than an order of magnitude, resulting in an overall 2-3x speed improvement. We also present scaling studies where we measure the performance as a function of the number of simultaneous clients.

Primary authors: HAWKS, Benjamin (Fermilab); HOLZMAN, Burt (Fermilab); PEDRO, Kevin (Fermilab); ACOSTA FLECHAS, Maria (Fermilab); WANG, Michael (Fermilab); TRAN, Nhan (Fermilab); YANG, Tingjun (Fermilab)

Presenter: YANG, Tingjun (Fermilab)

Session Classification: Day 4 Afternoon