Contribution ID: **114** Type: **Poster**

Inference as a Service for HEP ML Models on AMD GPUs Using the SONIC Framework

One potential way to meet the quickly growing computing demands in High Energy Physics (HEP) experiments is by leveraging specialized processors such as GPUs. The "as a service"(AAS) approach helps improve utilization of GPU resources by allowing one GPU to serve a wide range of tasks, significantly reducing idle time. The SONIC project implements the AAS approach for a variety of widely used HEP algorithms and Machine Learning (ML) models by serving them using the NVIDIA Triton Inference Server framework. Focus has been primarily on serving models on NVIDIA GPUs, but the PyTriton package is flexible enough to allow Triton servers to be launched using AMD GPUs as well. This has been implemented, and the inference performance for two HEP ML models is compared across several AMD and NVIDIA GPUs.

Focus areas

HEP

Authors: COLBERT, Ethan (Purdue University (US)); FENG, Yongbin (Texas Tech University (US))

Co-author: LIU, Miaoyuan (Purdue University (US))

Presenter: COLBERT, Ethan (Purdue University (US))