Contribution ID: **32** Type: **not specified**

Exa.TrkX inference as-a-service

Monday, 3 October 2022 14:10 (5 minutes)

Particle tracking plays a crucial role in many particle physics experiments, e.g, the Large Hadron Collider. Yet, it is also one of the most time-consuming components in the whole particle reconstruction chain. The Exa.TrkX group has developed in recent years a promising and well-performed machine-learning-based pipeline that carries out the track finding, which is the most computationally expensive part of particle tracking. An important research direction is to accelerate the pipeline, via software-based approaches such as model pruning, tensor operation fusion, reduced precision, quantization, and hardware-based approaches such as usages of different coprocessors, such as GPUs, TPUs, and FPGAs.
In this talk, we will introduce our implementation of Exa.TrkX inference as-a-service through NVIDIA Triton Inference servers. Clients read data and send track-finding inference requests to (remote) servers; servers run the inference pipeline on different types of coprocessors and return outputs to clients. The pipeline running on the server side includes three discrete deep learning models and two CUDA-based domain algorithms. Compared with normal local inferences, this approach allows us more freedom to easily utilize different types of coprocessors more efficiently, while maintaining similar throughputs and latency. We will discuss in detail different server configurations explored in order to achieve this.

Primary authors: LAZAR, Alina (Youngstown State University); HSU, Shih-Chieh (University of Washington Seattle (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); FENG, Yongbin (Fermi National Accelerator Lab. (US))

Presenter: FENG, Yongbin (Fermi National Accelerator Lab. (US))

Session Classification: Contributed Talks