Contribution ID: **56** Type: **Poster**

Graph Neural Network-based Tracking as a Service

Tuesday, 10 October 2023 19:18 (3 minutes)

Recent studies have shown promising results for track finding in dense environments using Graph Neural Network (GNN)-based algorithms. These algorithms not only provide high track efficiency but also offer reasonable track resolutions. However, GNN-based track finding is computationally slow on CPUs, necessitating the use of coprocessors like GPUs to accelerate the inference time. Additionally, due to the substantial graph size typically involved (consisting of approximately 300k nodes and 1M edges), significant GPU memory is required to ensure efficient computation. Not all computing facilities used for particle physics experiments are equipped with high-end GPUs such as NVIDIA A100s or V100s, which can meet the computational requirements. These computing challenges must be addressed in order to deploy GNN-based track finding into production. We propose addressing these challenges by establishing the GNN-based track finding algorithm as a service hosted either in the cloud or high-performance computing centers.

In this talk, we will describe the implementation of the GNN-based track finding workflow as a service using the Nvidia Triton Inference Server. The pipeline contains three discrete deep-learning models and two CUDA-based algorithms. Because of the heterogeneity in the workflow, we explore different server configurations? to maximize the throughput of track finding and the GPU utilization. We also study the scalability of the inference server using the Perlmutter supercomputer at NERSC and cloud resources like AWS and Google Cloud.

Primary authors: ZHAO, Haoran (University of Washington (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US))

Co-authors: NAYLOR, Andrew (Lawrence Berkeley National Lab); RANKIN, Dylan Sheldon (University of Pennsylvania (US)); KHODA, Elham E (University of Washington (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); FARRELL, Steven (Lawrence Berkeley National Laboratory); MCCORMACK, William Patrick (Massachusetts Inst. of Technology (US)); FENG, Yongbin (Fermi National Accelerator Lab. (US))

Presenter: JU, Xiangyang (Lawrence Berkeley National Lab. (US))

Session Classification: Poster