Contribution ID: **76** Type: **Oral**

ACTS as a Service

Thursday 14 March 2024 14:30 (20 minutes)

Recent advancements in track finding within the challenging environments expected in the High-Luminosity Large Hadron Collider (HL-LHC) have showcased the potential of Graph Neural Network (GNN)-based algorithms. These algorithms exhibit high track efficiency and reasonable resolutions, yet their computational burden on CPUs hinders real-time processing, necessitating the integration of accelerators like GPUs. However, the substantial size of the involved graphs, with approximately 300k nodes and 1M edges, demands significant GPU memory, posing a challenge for facilities lacking high-end GPUs such as NVIDIA A100s or V100s. These computing challenges must be addressed to deploy GNN-based track finding or any algorithm that requires coprocessors, into production.

To overcome these challenges, we propose the as-a-service approach to deploy the GNN-based track-finding algorithm in the cloud or high-performance computing centers such as the NERSC Perlmutter system with over 7000 A100 GPUs. In addressing this, we have developed a tracking-as-a-service prototype within A Common Tracking Software (ACTS), an experiment-independent toolkit for charged particle track reconstruction.

The GNN-based track finding is implemented as a service within ACTS, showcasing its versatility as a demonstrator. Moreover, this approach is algorithm-agnostic, allowing the incorporation of other algorithms as new backends through interactions with the client interface implemented in ACTS.

In this contribution, we present the implementation of the GNN-based track-finding workflow as a service using the Nvidia Triton Inference Server within ACTS. The GNN pipeline comprises three distinct deep-learning models and two CUDA-based algorithms, enabling full tracking reconstruction within ACTS. We explore different server configurations to assess track-finding throughput and GPU utilization, exploring the scalability of the inference server across the NERSC Perlmutter supercomputer and cloud resources such as AWS and Google Cloud.

Significance

This contribution describes the work that uses the as-a-service computing model to accelerate track finding in dense environments for HL-LHC. The as-a-service method provides more flexibility to scale and balance computing resources using coprocessors for state-of-art tracking reconstruction.

Furthermore, this approach is independent of the underlying tracking algorithm. The GNN-based tracking finding is implemented in ACTS to provide the first demonstrator for such an approach to showcase the potential and explore scalability using supercomputers

References

Experiment context, if any

Primary authors: ZHAO, Haoran (University of Washington (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); YAO, Yao (Purdue University (US)); FENG, Yongbin (Fermi National Accelerator Lab. (US)); CHOU, Yuan-Tang (University of Washington (US))

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

Presenter: CHOU, Yuan-Tang (University of Washington (US))

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research