Connecting
The Dots 2022
7th International CTD Workshop
Princeton University, Princeton, USA
May 31 - June 2, 2022

Contribution ID: **36**

Type: **Poster**

# ExaTrkX as a Service

*Tuesday, 31 May 2022 12:30 (3 minutes)*

Particle tracking plays a pivotal role in almost all physics analyses at the Large Hadron Collider. Yet, it is also one of the most time-consuming parts of the particle reconstruction chain. In recent years, the Exa.TrkX group has developed a promising machine learning-based pipeline that performs the most computationally expensive part of particle tracking, the track finding. As the pipeline obtains competitive physics performance on realistic data, accelerating the pipeline to meet the computational demands becomes an important research direction, that can be categorized as either software-based or hardware-based. Software-based inference acceleration includes model pruning, tensor operation fusion, reduced precision, quantization, etc. Hardware-based acceleration explores the usage of different coprocessors, such as GPUs, TPUs, and FPGAs.

In this talk, we describe the Exa.TrkX pipeline implementation as a Triton Inference Server for particle tracking. Clients will send track-finding requests to the server and the server will return track candidates to the client after processing. The pipeline contains three discrete deep learning models and two CUDA-based algorithms. Because of the heterogeneity and dependency chain of the pipeline, we will explore different server settings to maximize the throughput of the pipeline, and we will study the scalability of the inference server and time reduction of the client.

## Consider for young scientist forum (Student or postdoc speaker)

No

**Primary authors:** KHODA, Elham E (University of Washington (US)); ALINA, Lazar (Youngstown State University); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); FENG, Yongbin (Fermi National Accelerator Lab. (US))

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

**Session Classification:** Poster session