Contribution ID: **765** Contribution code: **contribution ID 765** Type: **Oral**

Accelerating the Inference Time of Machine Learning-based Track Finding Pipeline

Thursday 2 December 2021 12:00 (20 minutes)

Recently, graph neural networks (GNNs) have been successfully used for a variety of reconstruction problems in HEP. In this work, we develop and evaluate an end-to-end C++ implementation for inferencing a charged particle tracking pipeline based on GNNs. The pipeline steps include data encoding, graph building, edge filtering, GNN and track labeling and it runs on both GPUs and CPUs. The ONNX Runtime C++ API is used to run PyTorch deep learning models converted to ONNX. The implementation features an improved GPU-based fixed radius nearest neighbor search for identifying edges and a weakly connected component algorithm for the labeling step. In addition, complete conversion to C++ allows integration with existing tracking software, including ACTS. We report the memory usage, average event latency, and the efficiency and purity tracking performance of our implementation applied to the TrackML benchmark dataset. The GPU-based implementation provides considerable speed-ups over the CPU-based execution and can be extended to run on multiple GPUs.

Significance

Deep learning inference runs predominantly on the GPU, therefore there is a lot to be gained by running it in parallel as a multi-threaded event-parallel task farm. Unfortunately, Python's threading model is limited by the Global Interpreter Lock, slowing down throughput and increasing latency. Converting the pipeline to C++, we overcome Python threading drawbacks, and provide an efficient mechanism to integrate the pipeline with C++-based event reconstruction workflows.

References

Ju, X., Murnane, D., Calafiura, P., Choma, N., Conlon, S., Farrell, S., Xu, Y., Spiropulu, M., Vlimant, J.R., Aurisano, A. and Hewes, J., 2021. Physics and Computing Performance of the Exa. TrkX TrackML Pipeline. arXiv preprint arXiv:2103.06995. Submitted to Europhysics Journal C.

Speaker time zone

Compatible with America

Primary authors: LAZAR, Alina (Youngstown State University); MURNANE, Daniel Thomas (Lawrence Berkeley National Lab. (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US))

Presenter: LAZAR, Alina (Youngstown State University)

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research