Contribution ID: **174** Type: **Poster**

Portable acceleration of CMS computing workflow with coprocessors as a service

Thursday, 14 March 2024 16:10 (30 minutes)

Computing demands for large scientific experiments, such as the CMS experiment at the CERN LHC, will increase dramatically in the next decades. To complement the future performance increases of software running on central processing units (CPUs), explorations of coprocessor usage in data processing hold great potential and interest. Coprocessors are a class of computer processors that supplement CPUs, often improving the execution of certain functions due to architectural design choices. In this talk, I will introduce the approach of Services for Optimized Network Inference on Coprocessors (SONIC) and discuss the study of the deployment of this as-a-service approach in large-scale data processing.

In the studies, we take a data processing workflow of the CMS experiment and run the main workflow on CPUs, while offloading several machine learning (ML)inference tasks onto either remote or local coprocessors, specifically graphics processing units (GPUs). With experiments performed at Google Cloud, the Purdue Tier-2 computing center, and combinations of the two, we demonstrate the acceleration of these ML algorithms individually on coprocessors and the corresponding throughput improvement for the entire workflow. We will also show this approach can be easily generalized to different types of coprocessors and deployed on local CPUs without decreasing the throughput performance.

Significance

References

Experiment context, if any

The CMS Experiment

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

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

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research