Contribution ID: **61**

Type: **Poster**

Accelerating CMS Workflow with ML Inference as-a-service on Perlmutter

As scientific experiments are generating increasingly larger and more complex datasets, the need to accelerate scientific workflows becomes ever more pressing. Recent advancements in machine learning (ML) algorithms, combined with the power of cutting-edge GPUs, have led to significant performance gains. However, optimizing computational efficiency remains crucial to minimize processing latency and resource consumption. To address these challenges, the CMS experiment is exploring ML inference-as-a-service (IaaS) to better utilize hardware and meet rising computational demands within budget constraints. CMS has integrated the approach of Services for Optimized Network Inference on Coprocessors (SONIC) into its core software stack, leveraging the open-source Nvidia Triton inference server software.

The Perlmutter supercomputer, featuring over 7000 Nvidia A100 GPUs, is designed to advance and accelerate scientific discoveries through its state-of-the-art performance and advanced capabilities. This study investigates a CMS data processing workflow that offloads machine learning tasks to GPUs running as a service on Perlmutter via the Nvidia Trion inference server. This talk will present the current status of this study, exploring the performance, challenges faced, and future directions. By leveraging Perlmutter's capabilities, we aim to significantly enhance the efficiency and scalability of CMS workflows.

Focus areas

HEP

Author: NAYLOR, Andrew (Lawrence Berkeley National Lab)

Presenter: NAYLOR, Andrew (Lawrence Berkeley National Lab)