Contribution ID: **2** Type: **Lightning Talk**

Using NVIDIA Triton Server for Inference-as-a-Service at Fermilab

Monday 25 September 2023 17:35 (5 minutes)

With machine learning gaining more and more popularity as a physics analysis tool, physics computing centers, such as the Fermilab LHC Physics Center (LPC), are seeing huge increases in their resources being used for such algorithms. These facilities, however, are not generally set up efficiently for machine learning inference as they rely on slower CPU evaluation, which has a noticeable impact on time-to-insight and is detrimental to computational throughput. In this work, we will discuss how we used the NVIDIA Triton Inference Server to re-optimize Fermilab's resource allocation and computing structure to achieve high throughput for scaling out to multiple users parallelizing their machine learning inference at the same time. We will also demonstrate how this service is used in current physics analyses and provide steps for how others can apply this tool to their analysis code.

Primary author: SAVARD, Claire (University of Colorado Boulder (US))

Presenter: SAVARD, Claire (University of Colorado Boulder (US))

Session Classification: Contributed Talks

Track Classification: Contributed Talks