Contribution ID: **1280** Type: **Parallel session talk**

Fast ML inference framework for real-time analysis at LHCb

Friday 19 July 2024 15:04 (17 minutes)

The LHCb detector generates vast amounts of data (5 TB/s), necessitating efficient algorithms to select data of interest and reduce the bandwidth to acceptable levels in real time. Deploying machine learning (ML) models for inference at all trigger stages is challenging, as the models need to fulfill strict throughput requirements.

To achieve the throughput aims, optimized batched evaluation for both GPU and CPU architectures is developed, used at the first and second trigger stages, respectively. Furthermore, the aim is to reduce the maintenance burden and turnaround time of retraining models and allow flexibility of training platforms by factorizing inference from training software.

This talk provides an overview of the real-time ML inference framework integrated into the software of the LHCb experiment, covering training and testing pipelines, alongside throughput evaluations of typical ML models at both stages of the trigger.

Alternate track

1. Operation, Performance and Upgrade (incl. HL-LHC) of Present Detectors

I read the instructions above

Yes

Authors: VAN VEGHEL, Maarten (Nikhef National institute for subatomic physics (NL)); AAIJ, Roel (Nikhef National institute for subatomic physics (NL))

Presenter: VAN VEGHEL, Maarten (Nikhef National institute for subatomic physics (NL))

Session Classification: Computing and Data handling

Track Classification: 14. Computing, AI and Data Handling