Contribution ID: **7** Type: **Flash Talk**

# Ahead-of-time (AOT) compilation of Tensorflow models

*Friday 8 December 2023 12:12 (12 minutes)*

In a wide range of high-energy particle physics analyses, machine learning methods have proven as powerful tools to enhance analysis sensitivity.
In the past years, various machine learning applications were also integrated in central CMS workflows, leading to great improvements in reconstruction and object identification efficiencies.

However, the continuation of successful deployments might be limited in the future due to memory and processing time constraints of more advanced models evaluated on central infrastructure.

A novel inference approach for models trained with TensorFlow, based on Ahead-of-time (AOT) compilation is presented. This approach offers a substantial reduction in memory footprints while preserving or even improving computational performance.

This talk outlines strategies and limitations of this novel approach, and presents integration workflow for deploying AOT models in production.

**Primary authors:** WIEDERSPAN, Bogdan (UNI/EXP (Uni Hamburg, Institut fur Experimentalphysik)); RIEGER, Marcel (UNI/EXP (Uni Hamburg, Institut fur Experimentalphysik))

**Presenter:** WIEDERSPAN, Bogdan (UNI/EXP (Uni Hamburg, Institut fur Experimentalphysik))

**Session Classification:** Session II