

3rd International Workshop on Machine Learning and Quantum Computing Applications in Medicine and Physics



Contribution ID : 107

Type : not specified

Optimizing PyTorch AI Models with GPU Profiling and Triton Kernels - Part 1

Friday, 11 September 2026 13:40 (120)

This training provides a practical introduction to optimizing PyTorch-based AI models on GPUs, with a strong focus on performance profiling and custom Triton kernels. The aim is to equip attendees with the skills needed to identify performance bottlenecks and accelerate GPU-based computations. By the end of the training, attendees will be able to:

- manage CPU–GPU memory transfers and reason about performance,
- profile GPU code and interpret traces to spot bottlenecks,
- understand the motivation and principles behind writing custom GPU kernels,
- write simple custom kernels in Triton and integrate them into PyTorch workflows,
- compare custom kernel performance against built-in PyTorch operations.

Target audience: Users who already work with PyTorch and want to accelerate and optimize their GPU-based numerical or AI computations.

Requirements: Working knowledge of PyTorch tensors and basic GPU concepts; Python proficiency; familiarity with undergraduate-level linear algebra.

Required tools: Bring your own laptop — a remote Jupyter Notebook environment with everything needed will be provided.

Primary author(s) : Dr KLIMASZEWSKI, Konrad (NCBJ); OBARA, Michał (NCBJ)

Presenter(s) : Dr KLIMASZEWSKI, Konrad (NCBJ); OBARA, Michał (NCBJ)

Session Classification : Machine Learning Tutorial