Contribution ID: **32** Type: **not specified**

Services for Machine Learning applications (part 2 of 3)

Thursday 7 November 2024 13:30 (1 hour)

This session will focus on the infrastructure and low level tools required to efficiently deploy machine learning applications. In particular, it will cover:

- The different data types and how they can impact ML workloads, as well as support in different types of hardware and software libraries
- Key differences between CPUs and GPUs and how they impact ML workloads (training and serving)
- The available techniques in IT services for GPU sharing and partitioning. In particular, it will cover how applications can build on the existing Kubernetes service to simplify these operations
- Hands-on exercises on using GPUs for different types of workloads

Presenter: GAPONCIC, Diana (IT-PW-PI)