Contribution ID: **186** Type: **Short Talk**

Training and Serving ML workloads with Kubeflow at CERN

Thursday, 20 May 2021 11:16 (13 minutes)

Machine Learning (ML) has been growing in popularity in multiple areas and groups at CERN, covering fast simulation, tracking, anomaly detection, among many others. We describe a new service available at CERN, based on Kubeflow and managing the full ML lifecycle: data preparation and interactive analysis, large scale distributed model training and model serving. We cover specific features available for hyper-parameter tuning and model metadata management, as well as infrastructure details to integrate accelerators and external resources. We also present results and a cost evaluation from scaling out a popular ML use case using public cloud resources, achieving close to linear scaling when using a large number of GPUs.

Primary authors: GOLUBOVIC, Dejan (CERN); ROCHA, Ricardo (CERN)

Presenter: GOLUBOVIC, Dejan (CERN)

Session Classification: Artificial Intelligence

Track Classification: Distributed Computing, Data Management and Facilities