Contribution ID: **207** Type: **Short Talk**

Accelerating GAN training using highly parallel hardware on public cloud

Thursday, 20 May 2021 11:29 (13 minutes)

With the increasing number of Machine and Deep Learning applications in High Energy Physics, easy access to dedicated infrastructure represents a requirement for fast and efficient R&D. This work explores different types of cloud services to train a Generative Adversarial Network (GAN) in a parallel environment, using Tensorflow data parallel strategy. More specifically, we parallelize the training process on multiple GPUs and Google Tensor Processing Units (TPU) and we compare two algorithms: the TensorFlow built-in logic and a custom loop, optimised to have higher control of the elements assigned to each GPU worker or TPU core. The quality of the generated data is compared to Monte Carlo simulation. Linear speed-up of the training process is obtained, while retaining most of the performance in terms of physics results. Additionally, we benchmark the aforementioned approaches, at scale, over multiple GPU nodes, deploying the training process on different public cloud providers, seeking for overall efficiency and cost-effectiveness. The combination of data science, cloud deployment options and associated economics allows to burst out heterogeneously, exploring the full potential of cloud-based services.

Primary authors: DA COSTA CARDOSO, Renato Paulo (Universidade de Lisboa (PT)); GOLUBOVIC, Dejan (CERN); PELUAGA LOZADA, Ignacio (CERN); BRITO DA ROCHA, Ricardo (CERN); Dr VALLECORSA, Sofia (CERN); FERNANDES, João (CERN)

Presenter: DA COSTA CARDOSO, Renato Paulo (Universidade de Lisboa (PT))

Session Classification: Artificial Intelligence

Track Classification: Distributed Computing, Data Management and Facilities