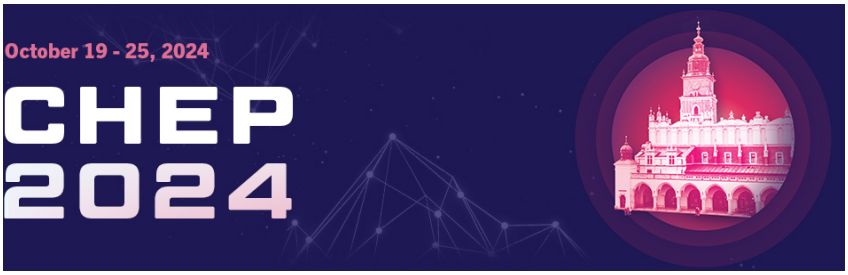

Contribution ID: **222** Type: **Talk**

Improving overall GPU sharing and usage efficiency with Kubernetes

Wednesday 23 October 2024 14:42 (18 minutes)

GPUs and accelerators are changing traditional High Energy Physics (HEP) deployments while also being the key to enable efficient machine learning. The challenge remains to improve overall efficiency and sharing opportunities of what are currently expensive and scarce resources.

In this paper we describe the common patterns of GPU usage in HEP, including spiky requirements with low overall usage for interactive access, as well as more predictable but potentially bursty workloads including distributed machine learning. We then explore the multiple mechanisms to share and partition GPUs, covering time slicing, virtualization, physical partitioning (MIG) and MPS for Nvidia devices.

We conclude with the results of an extensive set of benchmarks for multiple representative HEP use cases, including traditional GPU usage as well as machine learning. We highlight the limitations of each option and the use cases where they fit best. Finally, we cover the deployment aspects and the different options available targeting a centralized GPU pool that can significantly push the overall GPU usage efficiency.

Primary authors: GOLUBOVIC, Dejan; GAPONCIC, Diana (IT-PW-PI); Mr TOMAS GUERRA, Diogo Filipe (CERN); ROCHA, Ricardo (CERN)

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

Session Classification: Parallel (Track 7)

Track Classification: Track 7 - Computing Infrastructure