

OpenForBC, GPU partitioning made easy

Thursday, March 24, 2022 3:50 PM (20 minutes)

In recent years, compute performances of GPUs (Graphics Processing Units) dramatically increased, especially in comparison to those of CPUs (Central Processing Units). Large computing infrastructures such as HPC (High Performance Computing) centers or cloud providers offer high performance GPUs to their users. GPUs are nowadays the hardware of choice for several applications involving massive parallel operations, such as deep learning (DL) and Artificial Intelligence (AI) workflows. However the programming paradigms for GPUs significantly vary according to the GPU model and vendor, often posing a barrier to their use in scientific applications. In addition, powerful GPUs such as those available in HPCs or data centers are hardly saturated by typical computing applications. The OpenForBC (Open For Better Computing) project was born in this context, and aims to ease the use of GPUs in an efficient manner. OpenForBC is an open source software framework that allows for effortless partitioning of GPUs from different vendors in Linux virtualized environments. OpenForBC supports dynamic partitioning for various configurations of the GPU, which can be used to optimise the utilisation of GPU kernels from different users or different applications. For example training complex DL models may require a full GPU, but inference may need only a fraction of it. In this contribution we describe the most common GPU partitioning options available on the market, discuss the implementation of the OpenForBC interface, and show results of benchmark tests in typical scenarios.

Primary authors: Mr BORRIERO, Alessio (INFN); LEGGER, Federica (Istituto Nazionale di Fisica Nucleare); Mr MONTELEONE, Daniele (INFN); Dr FRONZÉ, Gabriele Gaetano (INFN); Dr VALLERO, Sara (INFN); Dr BAGNASCO, Stefano (INFN); Dr LUSSO, Stefano (INFN)

Presenter: Dr FRONZÉ, Gabriele Gaetano (INFN)

Session Classification: Converging High Performance infrastructures: Supercomputers, clouds, accelerators

Track Classification: Track 9: Converging High Performance Computing Infrastructures: Supercomputers, clouds, accelerators