ICHEP 2022 BOLOGNA

ICHEP 2022 6
XLI 13 07 2022
International Conference
on High Energy Physics
Bologna (Italy)

Contribution ID: **614** Type: **Parallel Talk**

OpenForBC, the GPU partitioning framework

Saturday, 9 July 2022 11:45 (15 minutes)

In recent years, compute performances of GPUs (Graphics Processing Units) dramatically increased, especially in comparison to those of CPUs (Central Processing Units). GPUs are nowadays the hardware of choice for scientific applications involving massive parallel operations, such as deep learning (DL) and Artificial Intelligence (AI) workflows. Large-scale computing infrastructures such as on-premises data centers, HPC (High Performance Computing) centers, and public or private clouds offer high performance GPUs to researchers. 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 are hardly saturated by typical computing applications. GPU partitioning may be the key to exploit GPU computing power in an efficient and affordable manner. Multiple vendors proposed custom solutions to allow for GPU partitioning, often with poor portability across different platforms and OSs (Operating Systems).

OpenForBC (Open For Better Computing) is an open source software framework that allows for effortless and unified partitioning of GPUs from different vendors in Linux KVM virtualized environments. OpenForBC supports dynamic partitioning for various configurations of the GPU, which can be used to optimize the utilization of GPU kernels from different users or different applications. For example training complex DL models may require a full GPU, but inference may only need a fraction of it, leaving free resources for multiple cloned instances or other tasks. In this contribution we describe the most common GPU partitioning options available on the market, discuss the implementation of the OpenForBC interface, and show the results of benchmark tests in typical use case scenarios.

In-person participation

Yes

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

Presenter: LEGGER, Federica (Istituto Nazionale di Fisica Nucleare)

Session Classification: Computing and Data handling

Track Classification: Computing and Data handling