Contribution ID: **230** Type: **Parallel Talk**

GPU computation energy-efficiency: from lattice QCD to large language model training

Thursday, 3 August 2023 16:20 (20 minutes)

In the current climate and energy crisis context, it is crucial to study and optimise the energy efficiency of scientific software used at large scale computing facilities. This supports moving toward net-zero computing targets, and reduce the negative impact of growing operational costs on the production of scientific data.
The energy efficiency of a computation is generally quantified as an amount of work performed per unit of energy spent. The study presented here was comissioned by the national UK STFC DiRAC facility, and performed on the Edinburgh "Tursa" supercomputer based on 724 NVIDIA A100 GPUs. We study how the energy efficiency of various workflows varies as we down-clock the frequency of the GPUs. From lattice QCD benchmarks (Grid & QUDA) to large language model training (GPT), we observe that lower frequencies than the default one lead to an increase of the GPU energy efficiency by 20-30%, with a reasonable impact on performances. This study led to a modification of the default GPU frequencies on Tursa in December 2022, resulting in an estimated saving to date of 60 MWh.

Topical area

Software Development and Machines

Primary author: PORTELLI, Antonin (University of Edinburgh)

Presenter: PORTELLI, Antonin (University of Edinburgh)

Session Classification: Software Development and Machines