Contribution ID: **6** Type: **not specified**

ADVANCING DATA CENTER SUSTAINABILITY: Empirical Measurements of AI Training Power Demand on a GPU-Accelerated Node

Data center sustainability, a phenomenon that has grown in focus due to the continuing evolution of Artificial intelligence (AI)/High Performance Computing (HPC) and High Throughput Computing (HTC) systems; furthermore, the rampant increase in carbon emissions resulted due to unprecedented rise in Thermal Design Power (TDP) of the computer chips. With the exponential increase of demand towards the usage of such systems, major challenges have surfaced in terms of productivity, Power Usage Effectiveness (PUE), and thermal/scheduling management. Deploying AI/HPC infrastructure in data centers will require substantial capital investment.

This study at the ATLAS Tier-1 site, Scientific Data and Computing Center (SDCC), Brookhaven National Laboratory (BNL) quantified the energy footprint of this infrastructure by developing models based on the power demands of AI hardware during training. We measured the instantaneous power draw of an 8-GPU NVIDIA H100 HGX node while training open-source models, including the ResNet image classifier and the Llama2-13b large language model. The peak power draw observed was about 18% below the manufacturer's rated TDP, even with GPUs near full utilization. For the image classifier, increasing the batch size from 512 to 4096 images reduced total training energy consumption by a factor of four when model architecture remained constant. These insights can aid the scientific data center facilities such as CERN identify the 'stranded power'within existing facilities and assist in capacity planning and provide researchers with energy use estimates. Future studies will explore the effects of liquid cooling technologies and carbon-aware scheduling on AI workload energy consumption. These results can help ATLAS in the development of the ATLAS software or operational model which may significantly reduce the carbon footprint at their data centers and identify opportunities for heat reuse.

Author: LATIF, Imran (Brookhaven National Laboratory)

Presenter: LATIF, Imran (Brookhaven National Laboratory)

Session Classification: Hardware and Fabrics

Track Classification: All contributions