Contribution ID: **36**   Type: **not specified**

Case Study: AI Training Power Demand on a GPU-Accelerated Node

Thursday 7 November 2024 09:45 (20 minutes)

Data center sustainability, a phenomenon that has grown in focus due to the continuing evolution of Artificial intelligence (AI)/High Performance Computing (HPC) systems; furthermore, the rampant increase in carbon emissions resulted in an unprecedented rise in Thermal Design Power (TDP) of the computer chips at the Scientific Data and Computing Center (SDCC) at Brookhaven National Laboratory (BNL). 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 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 image classifier and the large language model. The peak power draw observed nearly 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 data center operators in capacity planning and provide researchers with energy use estimates. Future studies will explore the effects of cooling technologies and carbon-aware scheduling on AI workload energy consumption.

Desired slot length

12

Speaker release

Yes

Author: LATIF, Imran (Brookhaven National Laboratory)

Co-author: MISAWA, Shigeki (Brookhaven National Laboratory (US))

Presenters: LATIF, Imran (Brookhaven National Laboratory); MISAWA, Shigeki (Brookhaven National Laboratory (US))

Session Classification: Topical Session: Carbon & Sustainability in Data Centers

Track Classification: IT Facilities, Business Continuity and Green IT