ADVANCING DATA CENTER SUSTAINABILITY: CARBON-AWARE COMPUTING UTILIZING AI AUTOMATION

Imran Latif†
Computational Science Initiative
US DOE, Office of Science,
Brookhaven National Laboratory
Upton, New York, USA
ilatif@bnl.gov

Marwan Ruby
Mechanical Engineering
SUNY Farmingdale State College
Farmingdale, New York, USA
rubyma@farmingdale.edu

Mohtadi Mahim
Engineering and Applied Sciences
SUNY Stony Brook University
Stony Brook, New York, USA
mohtadi.mahim@stonybrook.ed

ABSTRACT

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, a rampant increase in carbon emissions resulted from 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.

This abstract aims to benchmark the carbon footprint of AI/HPC workloads using a single Supermicro HGX node with - Nvidia H-100 GPUs and analyzing potential improvements from diverse types of cooling systems. Our current studies show performance optimization issues involving the inability to reach the maximum rated TDP. Based on the analysis, a conversion from the existing air-cooled system to liquid-cooled one is proposed to reduce the carbon footprint. In addition to environmental benefits, the proposed changes should bring scalability, availability, and reliability improvements to BNL's SDCC datacenter.

KEYWORDS

Carbon Awareness, Data Center Sustainability, High-Performance Computing, Liquid Cooling, Workload Scheduling

Power intensive workloads have reinforced the need for a transformation from the current air-cooling system, utilizing the Rear Door Heat exchangers (RDHx), to a Direct-To-Chip (DTC) liquid cooling system at SDCC. According to Goldman Sachs, the power consumption for AI computing alone will see an increase of 160% by 2030, inevitably increasing carbon footprints [1]. The research proposes repeating benchmarks on a liquid cooled node, of similar hardware specifications to the identified H-100 powered by 8 Nvidia GPUs (700W each), for analysis on the reduction of power consumption and carbon emissions at no compromise towards performance [2]. Neighboring innovative data centers have shown DTC cooling systems, installed to the most heat-producing components (GPUs/CPUs), are more effective than air-cooled systems by 20-25% [3]. Furthermore, the increase in thermal capacity provides greater hardware density per rack, refining space efficiency and enhancing scalability [4]. Other dimensional enhancements involve automating task scheduling by shifting heavier workloads to off-peak hours for a reduction in energy usage and carbon pricing.

To qualify the proposed approach, various test runs, and carbon monitoring are performed to collect data before and after applying a DTC cooling system. The outline of the research is to collect data through experimental stress runs, while simultaneously tracking carbon emissions, CDU and GPU power consumptions, and PUE through the utilization of CodeCarbon, Pytorch, Grafana, ResNET 152/CIFAR10, and LlaMA2 13B on a single node [5]. Finally, creating a scheduling AI model to schedule runs based on various parameters [6].

Currently, with over 55,000 PUE inputs over this fiscal year, displayed in Figure 1, SDCC sits at an average PUE of about 1.3, indicating a significant level of energy efficiency already. Our data from AI training stress tests have revealed that at 100% GPU/CPU utilization, the nodes operate at only 80% of their intended TDP.

The proposed research will allow the owners/operators in BNL's hyper scale data center to conserve power, resulting in increased power and cooling capacity, and ultimately decrease carbon emissions. A potential power savings of 20-25% can be used for scalability by introducing high density HPC racks. The integration of DTC cooling and workload management systems will result in the optimization of power consumption efficiency, carbon emissions, and operation expenditure without having to sacrifice performance required for ever-evolving computational workloads.

∗DCS: A multi-dimensional approach towards carbon awareness

†Imran Latif of Brookhaven National Laboratory's Office of Science

REFERENCES

[1] Goldman Sachs. 2023. AI Poised to Drive 160% Increase in Power Demand. In Goldman Sachs Insights. Goldman Sachs, New York, NY, USA. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

[2] NVIDIA. 2023. NVIDIA H100 Tensor Core GPU: Unprecedented Performance, Scalability, and Security for AI Workloads. In NVIDIA Data Center Solutions. NVIDIA Corporation, Santa Clara, CA, USA. https://www.nvidia.com/en-us/data-center/h100/

[3] TECHEASE. 2022. How Data Centers Work. In YouTube Video. Google LLC, USA. https://www.youtube.com/watch?v=KQGonFB74ow

[4] ASUS. 2023. Server Liquid Cooling Solution. In ASUS Event. ASUSTeK Computer Inc., Taipei, Taiwan. https://www.asus.com/event/server-liquid-cooling-solution/

[5] Xingyu Lu. 2023. CIFAR10-ResNet152-Batchsize 1024-8 Epoch 200. In Weights & Biases Reports. Weights & Biases, San Francisco, CA, USA. https://wandb.ai/xyu1/cifar10-ResNet152-batch_1024/reports/CIFAR10-ResNet152-batchsize-1024-8-epoch200--Vmlldzo5Mzk1OTA5

[6] Ali Borji. 2022. A Categorical Archive of ChatGPT Failures. In arXiv preprint arXiv:2211.02001. 10 pages. https://arxiv.org/pdf/2211.02001

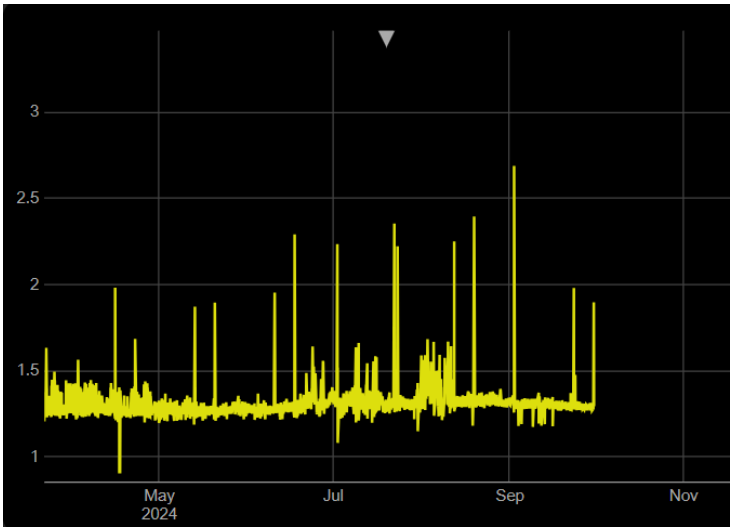

Figure 1: Power Usage Effectiveness (PUE) tracked over the past fiscal year.