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

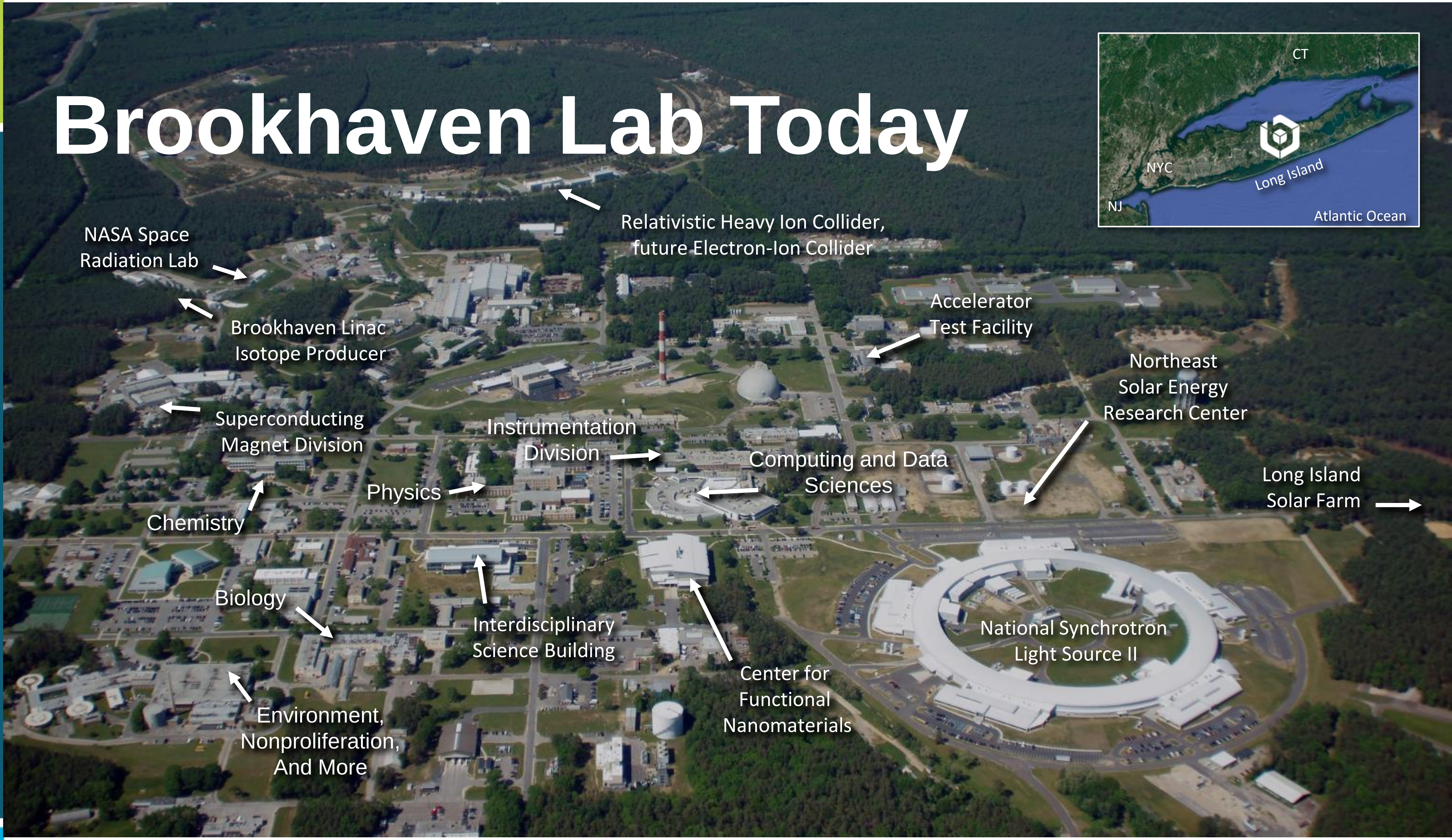
Imran Latif

Shigeki Misawa

11-07-24

@BrookhavenLab

Brookhaven Lab Today

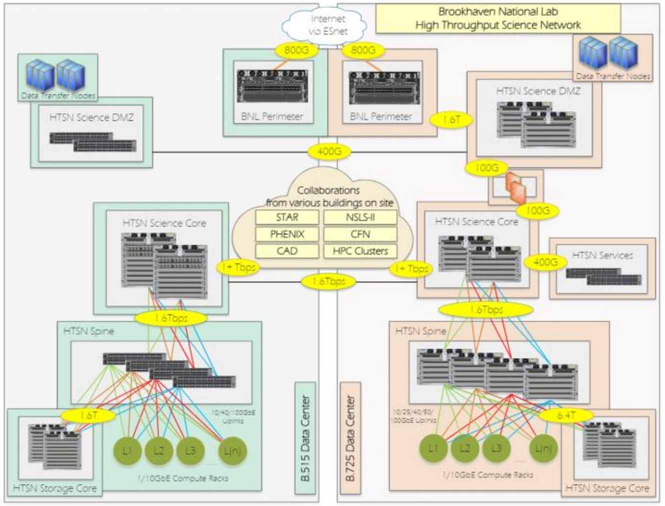


Computing and Data Sciences Directorate

***Provides Brookhaven Lab with leading-edge computing capabilities that enhance its ability to fulfill its mission today and into the future by*:**

- Conducting mission-informed foundational research in computer science, applied mathematics, and quantum information science
- Creating early hardware and software prototypes in close collaboration with the other Lab science directorates and the broader DOE community
- Providing computing and data services to enable research for the Lab's broad scientific user community through the Scientific Data and Computing Center (SDCC)
- Acting as a coordination point for cross-cutting computer science, applied mathematics, and computational science research at Brookhaven Lab
- Developing computing-oriented strategic initiatives for the Laboratory

SDCFD By the Numbers



SDFCD today is a leading computing center for High-Throughput Computing and Scientific Data, supports HEP and NP experiments, as well as other BNL, US and international projects : EIC, LQCD, NSLS II, CFN, BES, WLCG

One of the top-5 scientific data centers in the world

- The largest NP and HEP data archive in U.S.A.
- ~300 PB of data archived (exabytes by 2030)
- 160+ PB on disks
- *The mass storage HPSS system is used in Data Carousel mode, when data are actively migrated between disk and tape*

~2000 nodes (84k CPU cores and 350 GPU)

1.5 EB of data analyzed

BNL network externally connected to Global network at 1.6 Tbps

- Close to and/or over 1 PiB/day of data are transferred in and out of BNL
- Internally connected with various experiments and collaborators over Tbps

SDCFD supports experiments throughout the entire data collection and processing cycle, including data analysis. The data center serves ~2000 users from more than 20 projects and experiments.

State-of-the-Art SDCC Data Center

SDCFD Infrastructure: **59,000 sq-ft^2 Data Center** opened for operations in September 2021.

Running community services for more than 20 projects and experiments:

- RHIC (STAR and sPHENIX), ATLAS, Belle II, NSLS-II, DUNE, EIC, CFN, LQCD, IBM-Q Hub

Data Center Key Features

- Modular, Scalable and Robust design– 9.6MW of ultimate IT payload capacity
- Currently, 3.6MW (UPS/Generator) backed up power available
- Additional 1.2MW power block to be available in early FY-26
- Cooling with high efficiency chillers and Rear Door Heat Exchangers
- Liquid (Direct to Chip/Immersion) cooling ready for latest GPU based IT hardware deployment for AI, HPC and Digital Twin applications
- Energy efficient data center with 1.3 current PUE, aimed for 1.2 with full IT payload deployment
- Streamlined operations through Data Center Infrastructure Management (DCIM) including node level electric billing, asset management, environmental monitoring and capacity planning

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

- Quantifying the energy footprint of computational infrastructure requires models parameterized by the power demand of AI hardware during training
- Empirical measurement of the instantaneous power draw of an 8-GPU NVIDIA H100 HGX node obtained during the training of open-source **image classifier (ResNet)**, **large-language models (Llama2-13b)** and **GPU Burn stress test**
- All training runs were conducted on a single node system
- The active components within AI training servers include GPU’s, supervisory CPU’s, memory, storage drives, interconnect, and fans
- While holding the active components of a given server constant, there may be interaction effects between the IT and cooling systems

Hardware Information

- Supermicro HGX 8U air cooled node with (8) H100 GPU’s
- CPU : AMD EPYC 9354 32-Core Processor with 1.5 TB memory
- GPU: NVIDIA H100 with 80 GB memory
- Rated (node) TDP 10.2kW

HGX Node

Power and Cooling

- Standard 19" wide, 42U rack deployment
- Cooling provided by Rear-Door Heat Exchangers (RDHx)
- The RDHx is supplied with 15.5 °C facility chilled water
- 23.8 °C ambient cooling temperature maintained
- Power to the HGX node is provided via rack mounted power
- distribution units (PDU’s) using six standard C13 outlets
- Power monitoring available via DCIM system by NLYTE

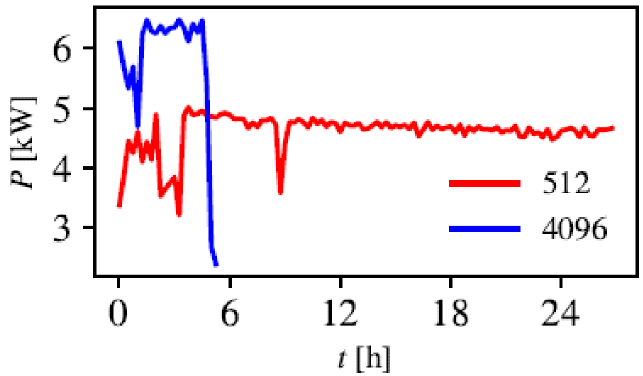


FIGURE 1. Instantaneous pre-training power demand measured in kW. Data corresponds to training ResNet with a batch size of 512 and 4096 images. Total energy usage for these two cases were computed by integrating the curves and are 123 and 30 kWh, respectively.

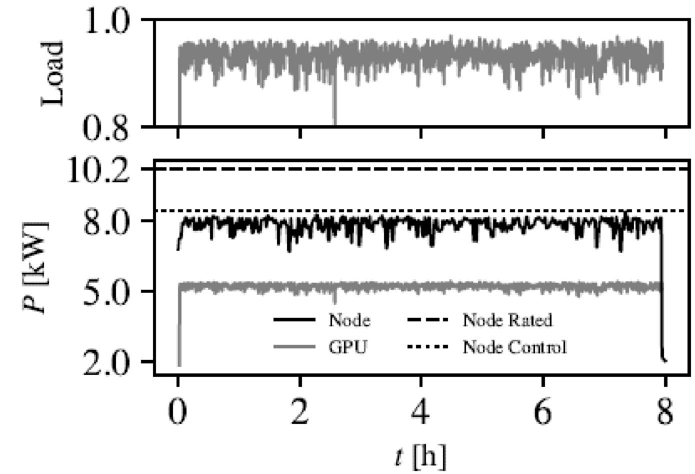


FIGURE 2. Instantaneous pre-training power demand measured in kW, and average total GPU load, for the Llama2-13b parameter model. The rated power of the system (10.2 kW) is shown as a horizontal dashed line. The median node power draw from the GPU+CPU burn control experiment is shown as a horizontal dotted line (at 8.43 kW).

TABLE 1. Power was sampled at 1-minute intervals over the duration of training. Batch size is measured in images and tokens for the ResNet and Llama models, respectively.

Training Workload	Model Task	Median Power (kW)	Batch Size	Duration (min)	Total Energy (kWh)	Average GPU Utilization (%)
Idle Node	N/A	1.86	N/A	N/A	N/A	N/A
Resnet-512	Image	4.68	512	1,605	123	36
Resnet-4096	Image	6.26	4,096	315	30	77
Llama2-13b	VQA	7.92	128	480	62	93
GPU Burn (no CPU)	Stress test	7.97	N/A	60	8	100
GPU Burn (+CPU)	Stress test	8.43	N/A	77	10.5	100

ResNET Image Classifier

- The training of the ResNet image classifier with a batch size of 512 images took over 26 hours
- The median node power demand was 4.68 kW, with a maximum power demand measured during this training of 5.02 W
- This was followed by the training of ResNet with a batch size of 4096 images, which completed in 5 hours
- The node demanded a median power draw of 6.26 kW during this workload, with a maximum measured demand of 6.48 kW
- The instantaneous power demand for these experiments is showcased in Fig. 2

Llama2-13b

- The instantaneous power demand results for training Llama2-13b are shown in Fig. 2
- Llama2-13b was trained with a batch size of 128 image question-answer pairs, completing in 8 hours
- The median node power consumption was 7.92 kW, with a maximum measured power demand of 8.42 kW
- GPU-level utilization during training (shown as the “GPU" label in Fig. 2) was close to maximum for all 8 chips

GPU Burn

- The dotted line in Fig. 2 (“Node Control") shows the total node power consumption during the maximally intensive GPU/CPU Burn stress
- tests, which maximized available GPU memory and tensor core operations GPU and CPU stress tests ran concurrently, the median
- power demand was 8.43 kW only ~400 W greater than the median demand of the Llama workload

Results and Conclusion

- The maximum observed power draw during a GPU Burn stress was approximately 8.4 kW, 18% lower than the manufacturer-rated 10.2 kW at 100% GPU utilization
- Holding model architecture constant, increasing batch size from 512 to 4096 images for ResNet used 1kW higher power on average but reduced total training energy consumption by a factor of 4
- Llama training GPU load was 93% on average, with a median power draw of 7.9 kW, indicating, also well below the rated maximum
- The minimal difference in power consumption between the GPU stress test and the real-world Llama training workload demonstrates that our empirical measurements offer a reliable characterization of the power requirements for computationally demanding workloads on this hardware
- The rack-level power distribution at the SDCC supplies 70 kW (14kwx5) of power to the racks which contain (5) Nvidia HGX nodes
- These racks were designed according to the manufacturer rated maximum, with a margin of redundant power of approximately 27%
- **These results suggest that existing infrastructure could support an additional 8-GPU node while maintaining design-specified redundancy**

Future research

- These findings can inform capacity planning for data center operators and energy use estimates by researchers
- Future work will investigate the impact of cooling technology and carbon-aware scheduling on AI workload energy consumption
- Future work should evaluate impact of cooling technology on node-level power demand, as well as other configuration, hyper-parametric, and infrastructure determinants of workload energy use.
- More effective cooling technologies, such as direct-to-chip liquid cooling, could allow the same computational utilization with lower node level energy usage – much needed !!!
- Full visibility into the component-level share of node power draw is another opportunity for future research efforts
- While technically challenging, detailed sub-metering of all active components in the node would identify the sensitivities of instantaneous power draw within different workloads, enabling hardware optimization
- There is an opportunity for future work to extend these analyses to multi-nodal training

References

https://github.com/wilicc/gpu-burn

A. S. Luccioni, S. Viguier, and A.-L. Ligozat, “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model,” Journal of Machine Learning Research, vol. 24, no. 253, pp. 1–15, 2023. [Online]. http://jmlr.org/papers/v24/23-0069.html

A. de Vries, “The growing energy footprint of artificial intelligence,” Joule, Oct. 2023. [Online]. https://www.sciencedirect.com/science/article/pii/S2542435123003653