

Do We Really Need All of that Scale?

Presenter: Grant Wilkins
CRSID: gfw27



340TWh

The estimated energy consumption of datacenters globally, 1.5% of global energy
(International Energy Agency)




Problem Space



- Data centers have grown in node count to parallelize large applications that process and evolve large volumes of data.
- With more components, nodes, and devices the energy and water consumption of systems increases, leading to greater environmental disruption to a local area.
- Largely distributed systems also face the trouble of concurrency problems, I/O wait times, and communication across large areas.
- Advances in single-chip systems and more powerful node architectures, beg the question: how much do we really need this scale?

Our goal is to quantify the energy usage of different data center architecture for distributed tasks to inform future system designs.





Methodology

- Three systems:
 - Macbook Pro (M1 Pro Chip with 32GB RAM and 14-core GPU)
 - Argonne Swing System (6 nodes, 8xA100 GPUs & 128-core Intel CPU per node)
 - Argonne ThetaGPU System (24 nodes, A100 GPU & 64-core AMD CPU per node)
- Run Llama 2 (7B), Mistral (7B), Orca-2 (7B) inference on each system with energy and performance counters
 - PAPI performance counters for CPU & DRAM energy consumption
 - nvidia-smi performance counters for GPU
 - Powermetrics for Macbook Pro
- Compare system performance on the basis of energy consumption, memory bandwidth, job completion, tokens per second, and I/O wait time



Hypotheses

1. Nodes that are “larger” will be more energy efficient than jobs that require coordination across multiple nodes.
2. We expect to see an increase in latency when scaling training across multiple nodes, similar to the work presented in the “PyTorch Distributed” paper.
3. The MacBook testing will have comparable performance and likely be the most energy-efficient, but not more performant.
4. The data center systems may have a lower mean time between failure compared to the MacBook Pro.



Expected Outcomes

- Be able to recommend system design considerations for energy efficiency.
- Begin to approach the problem of how to include system information in scheduling large-scale jobs across geographically distributed data centers to minimize environmental impact.
- Quantify scalability of PyTorch for LLM inference on different kinds of architectures.
- Start towards recommendations for research in selecting hardware configurations that optimize for both performance and energy efficiency, contributing to the development of green AI.



References

International Energy Agency. Data Centres and Data Transmission Networks. Retrieved November 27, 2023, from

<https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks>

Jiang, A. Q., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., & Chintala, S. (2020). PyTorch Distributed: Experiences on Accelerating Data Parallel Training. arXiv preprint arXiv:2006.15704.

Touvron, H., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

Mitra, A., et al. (2023). Orca 2: Teaching Small Language Models How to Reason. arXiv preprint arXiv:2311.11045.