

Vrije Universiteit Amsterdam
<https://www.overleaf.com/read/mbzddrhpdvvr#e96bec>



Honours Programme, Problem Statement

Precision Capacity Planning: Simulating the Cost-Energy Trade-offs of
LLM Training on Homogeneous Clusters

Author: Thijs van den Heuvel (2849429)

1st supervisor: Prof. Dr. Ir. Cav. Alexandru Iosup
daily supervisor: MSc. Dante Niewenhuis

*A report submitted in fulfillment of the requirements for the Honours Programme,
which is an excellence annotation to the VU Bachelor of Science degree in
Computer Science/Artificial Intelligence/Information Sciences
version 1.0*

December 9, 2025

Abstract

The training of Large Language Models (LLMs) has evolved into a massive industrial operation, where single training runs are costing tens of millions of dollars and consuming gigawatt-hours of energy. As models scale from billions to trillions of parameters, the financial and environmental sustainability of AI is becoming a critical and global challenge. Datacenters training these kinds of massive LLMs rely currently on a "good enough" human judgment or massive over-provisioning to manage these workloads, leading to significant inefficiencies in capital and energy use. Simulation offers a way to optimize these systems without the cost of physical experimentation. However, state-of-the-art simulators like OpenDC currently lack the specialized workload models required to accurately capture the non-functional properties of modern LLM training loops, such as gradient synchronization latency and checkpointing overheads.

This report proposes to bridge this gap by developing a parametric LLM training workload model for OpenDC. We aim to simulate the cost and energy trade-offs of training on homogeneous clusters, enabling datacenter operators to identify the best possible configuration where performance is maximized before diminishing returns and excessive carbon emissions set in.

Keywords

Large Language Models, LLMs, Simulation, Datacenter, Cost-Energy

1 Introduction

Artificial Intelligence (AI) has shifted from a niche research field to a primary driver of global datacenter demand. This AI era has introduced a new class of workloads, like the training of Large Language Models (LLMs). These type of workloads are not short-lived inference tasks or traditional web services, but they are synchronous, monolithic, and resource-intensive training jobs that occupies thousands of GPUs for months at a time [3, 9]. Estimating the carbon footprint of models like BLOOM reveals that training alone can emit over 50 tonnes of CO₂[7]. Efficiency in these systems is not just a performance metric, but a necessity for both financial and environmental sustainability.

Training is estimated to account for 10-40% of the total energy consumed by an LLM in its lifetime[2]. This is substantial and optimizing this to reduce even a few percent would cut costs and emissions. Discrete-event simulators could get every bit of performance out of a system resulting into millions of dollars in operational savings and significant reduction in carbon emissions. These simulators Act as a critical lever in this optimization, enabling companies to identify these efficiency gains without the capital risk of physical experimentation.

Datacenter configurations where all compute nodes (GPUs) are identical, an example being 1000 NVIDIA H100s, is the current industrial standard for training Frontier Models to minimize the "straggler effect" and synchronization latency found in heterogeneous environments. Even in these heterogeneous clusters, there is still a lot of time lost on

synchronization. A simple background process can stall thousands of others during gradient synchronization[6]. Optimizing LLM training is notoriously difficult because physical experimentation is expensive. Operators often over-provision hardware to ensure stability, avoiding the risk of potential downtime but leading to massive resource waste[10]. Current simulators also fail to model the "system-level" components of training, specifically Checkpointing mechanisms (saving model states to disk in case of failure) and gradient synchronization (network communication between clusters). Recent surveys insidate that 30-40% of training time can be lost due to these overheads if not optimized, but standard simulators still treat them as negligible[11, 12].

Key prior work relies on the "Scaling Laws" of Neural Networks[4], which which predict the raw compute requirements for training models based on their size. To meet these demands, industry has adopted distributed training systems like Megatron-LM[9], which split the workload across thousands of GPUs. In the field of simulation, tools like OpenDC[8] and ASTRA-sim have successfully modeled general cloud and network behaviors. Still, a gap remains: recent surveys[1] and production traces[5] show that current simulators fail to account for the "system-level" friction of training, specifically the massive overheads of checkpointing and synchronization. This is a critical gap, as existing tools cannot accurately predict the true energy and financial impact of large-scale training workloads.

So this leads us to the main research question: **How can we utilize discrete-event simulation to determine the cost-optimal and energy-optimal homogeneous cluster configuration for training Large Language Models?** This will become the first open-source LLM Training Workload Model for OpenDC that specifically accounts for reliability overheads like checkpointing and gradient synchronization. Allowing researchers to study "Green AI" infrastructure without needing access to supercomputers to do these experiments. This project will contribute to the scientific community by providing a new tool for simulating LLM training, enabling more efficient and sustainable AI development. It will also help me develop my skills in simulation, distributed systems, and AI, preparing me for a career in this rapidly evolving field.

2 Background

Explain the key concepts needed to understand this work. See also Section II of [?].

References

- [1] E. Dryden et al. A survey of end-to-end modeling for distributed dnn training: Workloads, simulators, and tco. *arXiv preprint arXiv:2506.09275*, 2025. Reference [6] in text.
- [2] Z. Ji and M. Jiang. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions. *Renewable and Sustainable Energy Reviews*, 225:116159, Jan. 2026.
- [3] Z. Jiang, H. Lin, Y. Zhong, Q. Huang, Y. Chen, Z. Zhang, Y. Peng, X. Li, C. Xie, S. Nong, Y. Jia, S. He, H. Chen, Z. Bai, Q. Hou, S. Yan, D. Zhou, Y. Sheng, Z. Jiang, H. Xu, H. Wei, Z. Zhang, P. Nie, L. Zou, S. Zhao, L. Xiang, Z. Liu, Z. Li, X. Jia, J. Ye, X. Jin, and X. Liu. Megascale: Scaling large language model training to more than 10, 000 gpus, 2024.
- [4] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.
- [5] J. Lin, Z. Jiang, Z. Song, S. Zhao, M. Yu, Z. Wang, C. Wang, Z. Shi, X. Shi, W. Jia, et al. Understanding stragglers in large model training using what-if analysis. *arXiv preprint arXiv:2505.05713*, 2025. Reference [13] in text (ByteDance Paper).
- [6] J. Lin, Z. Jiang, Z. Song, S. Zhao, M. Yu, Z. Wang, C. Wang, Z. Shi, X. Shi, W. Jia, Z. Liu, S. Wang, H. Lin, X. Liu, A. Panda, and J. Li. Understanding stragglers in large model training using what-if analysis, 2025.
- [7] A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model, 2022.
- [8] F. Mastenbroek, G. Andreadis, S. Jounaid, W. Lai, J. Burley, J. Bosch, E. van Eyk, L. Versluis, V. van Beek, and A. Iosup. Openc 2.0: Convenient modeling and simulation of emerging technologies in cloud datacenters. In *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 455–464, 2021.
- [9] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. A. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, and M. Zaharia. Efficient large-scale language model training on gpu clusters using megatron-lm, 2021.
- [10] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon emissions and large neural network training, 2021.
- [11] W. Xu, X. Huang, S. Meng, W. Zhang, L. Guo, and K. Sato. An efficient checkpointing system for large machine learning model training. In *Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis (SC-W)*, pages 896–900, 11 2024.
- [12] K. Zhang, Y. Chen, Z. Hu, W. Lin, J. Xu, and W. Chen. Gockpt: Gradient-assisted multi-step overlapped checkpointing for efficient llm training, 2025.