Virtual Benchmarking of LLMs using ExaDigiT and Calculon

Srishti Kalepu*
Georgia Institute of Technology
Atlanta, USA
skalepu3@gatech.edu

Wesley Brewer
Oak Ridge National Laboratory
Oak Ridge, USA
brewerwh@ornl.gov

Matthias Maiterth
Oak Ridge National Laboratory
Oak Ridge, USA
maiterthm@ornl.gov

Richard Vuduc
Georgia Institute of Technology
Atlanta, USA
richie@cc.gatech.edu

***Abstract*—As the use of LLMs, their size, and the required training compute increase, the infrastructure needed to support them is also expanding. The AI infrastructure being built by national labs and industry involves hundreds of thousands of GPUs, leading to massive energy and power consumption. Our approach uses digital twins and performance modeling tools to provide insight into the energy consumption of LLM training prior to AI infrastructure deployment, enabling analysis of the power load of the data centers and supercomputers.**

***Index Terms*—predictive modeling, energy efficiency, digital twins**

I. Introduction

LLMs are being used in nearly every aspect of our lives—from chatbots that write emails to recommendation models on Netflix that suggest our next watch [1]. This wide range of applicability has led companies to build their own AI infrastructures to keep up with the massive growth of LLMs and their training compute requirements. While companies continue to expand their infrastructures, there is growing concern about the power draw and energy consumption of data centers, along with their societal and environmental impacts.

One way to analyze the impact of data centers is by modeling their energy consumption. There is ongoing work on analytical energy models; however, these models do not account for power fluctuations across the various components of a data center. Another approach is to run workloads directly on the data center to obtain accurate energy consumption measurements. However, this approach is counterproductive, as it requires massive amounts of energy and cannot be used for broader energy consumption analysis. Our work focuses on estimating the energy consumption of LLM workloads without running them on a physical system. We leverage two tools: a digital twin of data centers and performance modeling tools of workloads to analyze the energy consumption of LLMs. The digital twin framework, ExaDigiT [2], simulates the data center, runs the workload, and provides estimates of power and energy utilization while accounting for the entire system. To provide the workload to the digital twin, we use performance modeling tools for LLMs that capture system utilization and runtime. In particular, we use Calculon [3], an analytical LLM performance modeling tool, which provides Model FLOPS Utilization (MFU), memory utilization, optimal hyperparameter configurations, and batch runtime of LLM training. We then use Calculon's output to generate synthetic LLM workloads that can be run on the digital twin framework.

By combining the digital twin of the data center with system utilization metrics from performance modeling tools, we have created a framework capable of predicting the energy consumption of training an LLM model on a data center. This framework is discussed further in Section IV.

II. ExaDigiT

ExaDigiT [2] is a digital twin framework of liquid-cooled supercomputers with three components: (1) a cooling model, (2) a resource allocator and power simulator (RAPS), and (3) a visual analytics model. The cooling model helps evaluate the effect of cooling on power use and overall energy consumption. The RAPS module schedules workloads, either from telemetry data or synthetic workloads, and can also replay telemetry data to show the energy and power use of past runs. The visual analytics model uses AR/VR to showcase the entire data center system. In this work, we focus on the RAPS module.

The RAPS module can use real-time telemetry data or synthetic workload traces to schedule workloads on the digital twin. While telemetry data shows the actual power and energy use, our project focuses on synthetic workloads to predict future energy consumption. To do this, RAPS uses a workload description that includes the number of nodes, job duration, and CPU/GPU utilization traces. These traces are created by sampling system utilization at fixed time intervals while running the workload. The utilization values are then mapped to node power using linear interpolation, and combined with

*This research was supported in part by an appointment to the Oak Ridge National Laboratory GRO Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education.

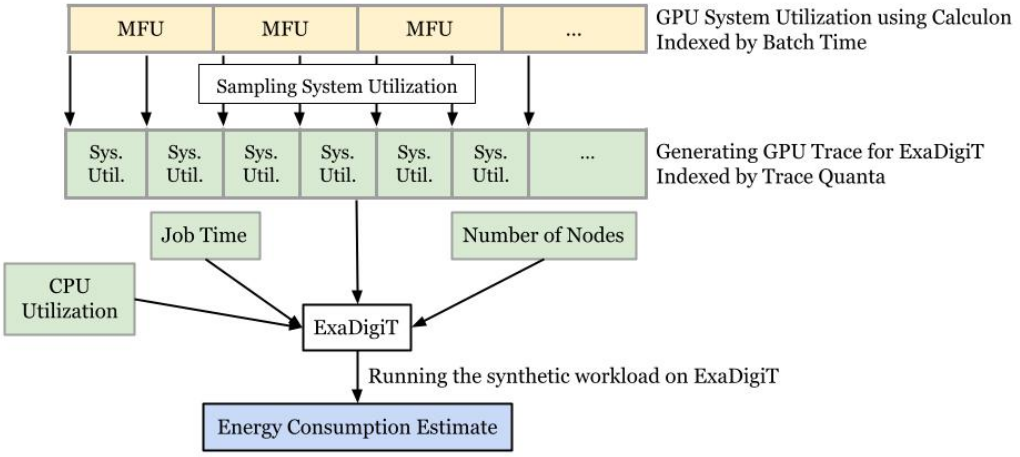

Fig. 1. Energy Consumption Estimation Framework

energy conversion losses in RAPS to estimate overall energy consumption.

To generate synthetic workloads, we need to provide system utilization at fixed intervals, job time, and number of nodes. Calculon [3] provides the system utilization and batch time of LLM training runs in a straightforward manner, which can be integrated directly into the RAPS module to generate synthetic workloads.

III. Calculon

Calculon is an analytical performance modeling tool that takes the system configuration of the GPU, the number of GPUs, the LLM model, and the hyperparameter configuration (which can also be given by Calculon if the user does not provide it) as input to provide the batch time, the Model FLOPS Utilization (MFU) of the batch, and the memory consumption of the LLM. MFU is a metric used to understand GPU system utilization while running LLM training, which is represented as:

$$\text{MFU} = \frac{\text{utilized FLOPS}}{\text{theoretical peak FLOPS}}$$

. Using the iterative nature of batched LLM training, we can make an array of MFU values indexed by batch times. The array of MFU values can be used as GPU utilization traces for input as part of the synthetic LLM workload into ExaDigiT, which can be used to predict the energy consumption. The integration of the two tools will be discussed in the next section.

IV. Virtual Benchmarking Framework

ExaDigiT estimates the energy and power utilization of a workload based on its CPU/GPU utilization traces. Using the batch time and MFU of LLM training provided by Calculon, we generate GPU utilization traces of running an LLM model, as shown in Figure 1. However, we still need CPU utilization for the synthetic workload in ExaDigiT. Since this is not directly provided, we are exploring optimal solutions. Currently, ExaDigiT generates CPU utilization traces by fluctuating between 0 (idle) and 1 (peak utilization).

ExaDigiT then uses the GPU utilization traces from Calculon, along with the generated CPU utilization traces, job time, and number of nodes, as input into the RAPS module. It runs the synthetic workload by processing the traces and calculating power values at one-second intervals, using linear interpolation between utilization and power. At the end of the run, ExaDigiT provides the total energy consumption by all the jobs as well as per-job information. This allows us to predict the energy consumption of synthetic LLM workloads without running them on physical supercomputers.

For validation, we plan to run the same LLM training configuration workloads with Calculon and ExaDigiT, and on physical supercomputers. The results of both will be overlapped to validate the framework against real-time data.

Our framework can be used to study the trade-offs between scaling the LLMs on different architectures, adding energy consumption as a metric to study the optimal hyper-parameter configuration for LLMs, or to understand the environmental effect of LLMs by analyzing their energy consumption.

V. Conclusion

Using the digital twin framework ExaDigiT and the LLM performance modeling tool Calculon, we have built a framework to estimate the energy consumption of LLM training without requiring physical deployment. Calculon provides system utilization information, which we use to generate a synthetic LLM workload. This synthetic workload is then executed on ExaDigiT, which simulates the data center environment and produces energy consumption estimates while accounting for the entire system. In this way, our framework enables detailed analysis of the energy consumption of LLMs without incurring any energy penalty from actual runs.

REFERENCES

[1] Hsiao, K.-J., Feng , Y., and Lamkhede, S. (2025). Foundation model for personalized recommendation. Retrieved from https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39

[2] W. Brewer, M. Maiterth, V. Kumar, R. Wojda, S. Bouknight, J. Hines, W. Shin, S. Greenwood, D. Grant, W. Williams, and F. Wang, "A digital twin framework for liquid-cooled supercomputers as demonstrated at exascale," in Proc. SC24: Int. Conf. High Perform. Comput., Netw., Storage and Anal., Nov. 2024, pp. 1–18. [Online]. Available: http://dx.doi.org/10.1109/SC41406.2024.00029

[3] M. Isaev, N. Mcdonald, L. Dennison, and R. Vuduc, "Calculon: A methodology and tool for high-level co-design of systems and large language models," in Proc. Int. Conf. High Perform. Comput., Netw., Storage and Anal. (SC '23), Denver, CO, USA, 2023, Art. no. 71, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3581784.3607102