Diploma Thesis

Microproccessors and Digital Systems Laboratory

Improving Performance and Energy Efficiency of Large Language Models Inference Serving Systems

Large Language Models (LLMs) have surged in popularity over the last couple of years, driven by advancements in deep learning and highlighted by the remarkable success of ChatGPT from OpenAI [1]. However, despite their impressive capabilities, LLMs present significant challenges in terms of computational requirements, latency, and energy consumption, particularly when deployed in real-time inference serving systems. The complexity and size of these models necessitate vast computational resources, often resulting in high operational costs and limited scalability. As demand for AI-driven services increases, optimizing both the performance and energy efficiency of LLM inference systems becomes critical to sustaining their widespread adoption. Optimizing LLM inference systems has garnered significant attention, with numerous studies investigating ways to enhance performance while reducing energy consumption [2,3,4]. Yet, there remains a large unexplored space that remains untapped, particularly in terms of balancing trade-offs between model accuracy, latency, and energy efficiency.

In this direction, we propose a set of available theses that explore innovative strategies to reduce inference latency, improve resource utilization, and minimize energy consumption. These theses will focus on how to efficiently schedule incoming queries and requests across a pool of heterogeneous resources. Typically, LLM providers, such as OpenAI, manage vast arrays of specialized hardware accelerators, including various types of GPUs. Properly scheduling these incoming requests to optimize the utilization of this diverse hardware pool is crucial for enhancing overall performance and reducing operational costs. In the context of LLMs, this is particularly important because the type of input query can significantly impact the performance of the inference serving system. As a result, there is a need for intelligent scheduling algorithms that not only account for the available hardware but also consider the specific characteristics and demands of each query. By carefully balancing these factors, the system can simultaneously optimize latency, energy efficiency, and resource usage. This series of theses investigates various scheduling techniques, such as workload-aware scheduling, resource heterogeneity management, and dynamic adaptation to query types, with the goal of improving LLM inference performance. Through experimental evaluation and theoretical analysis, the work aims to provide insights into how scheduling strategies can be tailored to maximize efficiency and scalability in large-scale LLM deployments.

RELATED MATERIAL:

[1] https://chatgpt.com/

[2] Patel, P., Choukse, E., Zhang, C., Shah, A., Goiri, Í., Maleki, S., & Bianchini, R. (2024, June). Splitwise: Efficient generative llm inference using phase splitting. In 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA) (pp. 118-132). IEEE.

[3] Kakolyris, A. K., Masouros, D., Vavaroutsos, P., Xydis, S., & Soudris, D. (2024). SLO-aware GPU Frequency Scaling for Energy Efficient LLM Inference Serving. arXiv preprint arXiv:2408.05235.

[4] Patel, P., Choukse, E., Zhang, C., Goiri, Í., Warrier, B., Mahalingam, N., & Bianchini, R. (2023). Polca: Power oversubscription in llm cloud providers. arXiv preprint arXiv:2308.12908.

CONTACT INFORMATION:

Dimosthenis Masouros Ph.D.: (dmasouros@microlab.ntua.gr)

Prof. Dimitrios Soudris: (dsoudris@microlab.ntua.gr)