

Diploma Thesis

Microprocessors and
Digital Systems
Laboratory



Power and Energy Efficiency of LLM Inference through Power Capping and DVFS

Large language models (LLMs) such as GPT-4, Llama 3, and DeepSeek-V3 have become a central component of modern artificial intelligence, enabling conversational assistants, code generation, and knowledge retrieval. With models now reaching hundreds of billions of parameters, serving them has become one of the dominant workloads in datacenters, and LLM inference is now a first-class topic at major systems venues [1, 2]. Inference consists of two distinct phases: a compute-bound prefill phase that processes the input prompt, and a memory-bound decode phase that generates output tokens one at a time. These phases are increasingly served on heterogeneous GPU fleets that mix hardware generations with different compute, memory, and power characteristics.

These demands translate into high power consumption and significant energy costs, raising concerns about infrastructure capacity, operating expenses, and sustainability. Recent studies show that inference servers often draw well below their provisioned power and that substantial energy savings are possible without violating latency targets [3, 4]. Exploiting this requires understanding the trade-offs among energy efficiency, latency, and throughput.

This thesis will explore power capping and dynamic voltage and frequency scaling (DVFS) to improve the energy efficiency of LLM inference. Because prefill and decode stress different resources, they respond differently to frequency changes. This motivates finer-grained management that tunes compute and memory frequencies independently, per phase and per GPU type. The objective is to identify strategies that reduce energy consumption while maintaining acceptable performance, contributing to more sustainable deployment of LLM-based systems.

RELATED MATERIAL:

- [1] P. Patel, E. Choukse, C. Zhang, A. Shah, Í. Goiri, S. Maleki, and R. Bianchini, "Splitwise: Efficient Generative LLM Inference Using Phase Splitting," in Proc. ISCA, 2024.
- [2] Y. Zhong, S. Liu, J. Chen, J. Hu, Y. Zhu, X. Liu, X. Jin, and H. Zhang, "DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving," in Proc. OSDI, 2024.
- [3] P. Patel, E. Choukse, C. Zhang, Í. Goiri, B. Warriar, N. Mahalingam, and R. Bianchini, "Characterizing Power Management Opportunities for LLMs in the Cloud," in Proc. ASPLOS, 2024.
- [4] A. K. Kakolyris, D. Masouros, P. Vavaroutsos, S. Xydis, and D. Soudris, "throttLL'eM: Predictive GPU Throttling for Energy Efficient LLM Inference Serving," in Proc. HPCA, 2025.

CONTACT INFORMATION:

Christos Kavallaris, Ph.D. Student, NTUA: (ckavallaris@microlab.ntua.gr)
Achilleas Tzenetopoulos, Post Doc, NTUA
Dimosthenis Masouros, Post Doc NTUA
Sotirios Xydis, Ass. Prof., Microlab NTUA
Dimitrios Soudris, Professor Microlab NTUA