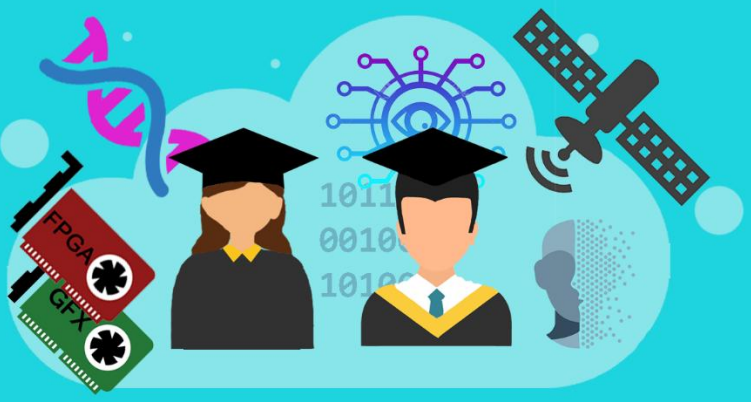
# Diploma Thesis

Microproccessors and Digital Systems Laboratory

&




MAX PLANCK INSTITUTE FOR INFORMATICS

## Energy-Efficient Inference Serving for Mixture of Experts (MoE) Large Language Models

Large Language Models (LLMs) have become ubiquitous, powering everything from conversational AI and content generation to complex reasoning tasks in enterprise applications. Among various model architectures, Mixture of Experts (MoE) models have emerged as a superior paradigm compared to traditional dense transformers, offering dramatically improved quality by selectively activating subsets of model parameters during inference, i.e., some of the experts are activated during inference. This sparse activation pattern on model's experts allows MoE models to achieve the capabilities of much larger dense models, while requiring significantly fewer computational resources, making them increasingly attractive for large-scale deployment.

However, MoE architectures introduce unique operational challenges that complicate their efficient deployment. The sparse and dynamic nature of expert activation creates highly irregular computational workloads, where some experts are activated far more frequently than others, leading to load imbalance across computational resources. This heterogeneous activation pattern results in unpredictable memory access patterns, suboptimal hardware utilization, and complex routing decisions that must be tackled in real-time. In addition, the deployment of these large-scale models is performed on power-hungry GPUs that consume substantial energy both during manufacturing and operation. This has contributed to a dramatic increase in energy consumption and carbon footprint of data centers, raising critical sustainability concerns as AI workloads continue to scale exponentially.

This thesis will explore comprehensive strategies to reduce energy consumption and carbon footprint in MoE inference serving systems. We will first assess, measure, and characterize how application MoE execution impacts energy consumption and carbon emissions in modern GPU systems, in order to identify key sources of inefficiency. Then, we will investigate system-level optimizations including intelligent workload placement algorithms that consider both computational requirements and energy efficiency, dynamic scheduling mechanisms that adapt to real-time expert activation patterns, as well as

microarchitectural features such as Dynamic Voltage and Frequency Scaling (DVFS) to optimize power consumption based on workload characteristics. By developing novel techniques that bridge the gap between model architecture awareness and energy-efficient system design, this project aims to enable sustainable deployment of MoE models, while maintaining their superior performance characteristics.

*This diploma thesis has the possibility of being conducted as part of an internship at the Max Planck Institute for Software Systems (MPI-SWS) in Germany.*

References:

[1] Andreas Kosmas Kakolyris, Dimosthenis Masouros, Petros Vavaroutsos, Sotirios Xydis, Dimitrios Soudris, “throttLL’eM: Predictive GPU Throttling for Energy Efficient LLM Inference Serving”, HPCA 2025

[2] J. Stojkovic, C. Zhang, Í. Goiri, J. Torrellas, E. Choukse, "DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency", HPCA 2025

[3] Yichao Yuan, Lin Ma, Nishil Talati, “MoE-Lens: Towards the Hardware Limit of High-Throughput MoE LLM Serving Under Resource Constraints”, arXiv 2025

[4] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, Kim Hazelwood, “Sustainable AI: Environmental Implications, Challenges and Opportunities”, MLSys 2022

Contact Information:

Prof. Christina Giannoula: (cgiannoula@mpi-sws.org)

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

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