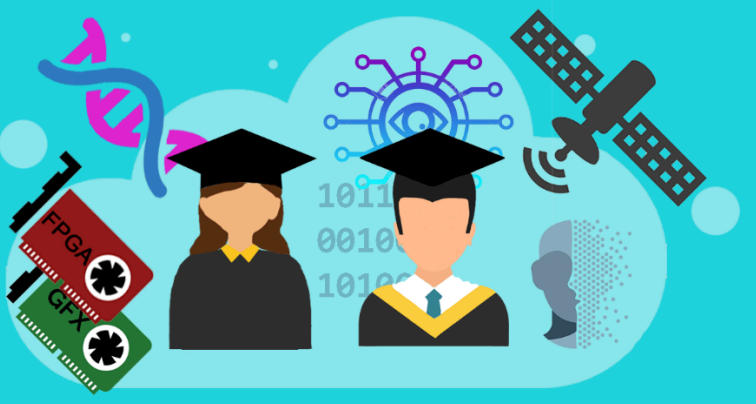
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Hardware-Aware LLM Model Compression

Large Language Models (LLMs) have rapidly become a cornerstone technology across industry and research, powering applications such as conversational agents, code generation tools, and search interfaces. Their adoption continues to grow as model capabilities improve, but this growth is coupled with increasing computational and memory requirements. As a result, deploying LLMs on edge devices, consumer hardware, or even cost-constrained data center environments remains challenging, motivating strong research interest in techniques that reduce model size and inference cost.

Current model compression techniques such as pruning, and quantization, achieve significant reductions in model parameters and memory footprint. However, state-of-the-art approaches typically evaluate compression quality using high-level metrics (parameter count, memory footprint) without considering the true performance impact on target hardware platforms. In practice, compressed models may still exhibit sub-optimal latency, memory bandwidth usage, or compute utilization, depending on the microarchitectural characteristics of the system. This disconnect limits the potential benefits of compression when applied to diverse hardware such as CPUs and GPUs.

This thesis will explore hardware-aware LLM compression, developing a methodology that tailors pruning and compression decisions to real, measurable performance characteristics of the underlying hardware. By integrating profiling data, microarchitectural bottleneck analysis, and feedback-driven optimization into the compression process, the resulting models will be smaller and more efficient in practice. The goal is to demonstrate improved latency, throughput, and energy efficiency across different hardware targets, bridging the gap between algorithmic compression techniques and hardware-level performance realities.

PREREQUISITES:

Knowledge of: ML and ML model compression, understanding of GPU architectures.

Desirable: Linux, Python scripting.

SKILLS YOU WILL LEARN:

The selected candidate will have the chance to familiarize with GPU architectures, profiling tools and methodologies, ML model compression techniques and LLM architectures.

RELATED MATERIAL:

CONTACT INFORMATION:

P. Eleftherakis (pelef@microlab.ntua.gr), A. Kapetanakis (akapetanakis@microlab.ntua.gr), Dr. K. Iliakis (kiliakis@microlab.ntua.gr), Asst. Prof. S. Xydis (sxydis@microlab.ntua.gr)