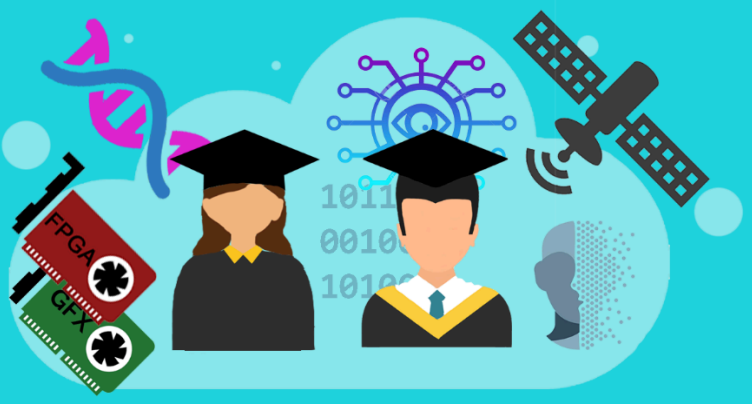
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Multi-CGRA SoC Architecture for Efficient LLM Inference

Motivation

The rapid adoption of Large Language Models (LLMs) has created an increasing demand for energy-efficient hardware capable of supporting their substantial computational and memory requirements. While GPUs provide high computational throughput, the energy and performance costs associated with data movement and general-purpose execution motivate the exploration of more specialized architectures. **Coarse-Grained Reconfigurable Architectures (CGRAs)** represent a promising alternative, combining the efficiency of specialized accelerators with sufficient programmability to support a diverse range of computational kernels [1].

However, most existing CGRAs target considerably smaller workloads and do not inherently provide the **computational throughput, memory bandwidth, and degree of parallelism required for efficient LLM inference**. A straightforward solution would be to increase the size of a single CGRA fabric. Scaling toward increasingly large monolithic fabrics, however, introduces challenges related to **kernel mapping, resource utilization, communication distance, and compilation complexity [2]**.

An alternative is a **multi-instance CGRA paradigm**, in which several smaller reconfigurable fabrics cooperate within the same SoC. This can provide greater aggregate computational throughput while avoiding some of the scalability limitations of a single large fabric [3]. It also introduces a new architectural challenge: **how should LLM kernels and their data be partitioned, mapped, and scheduled across multiple CGRA instances?**

Scope

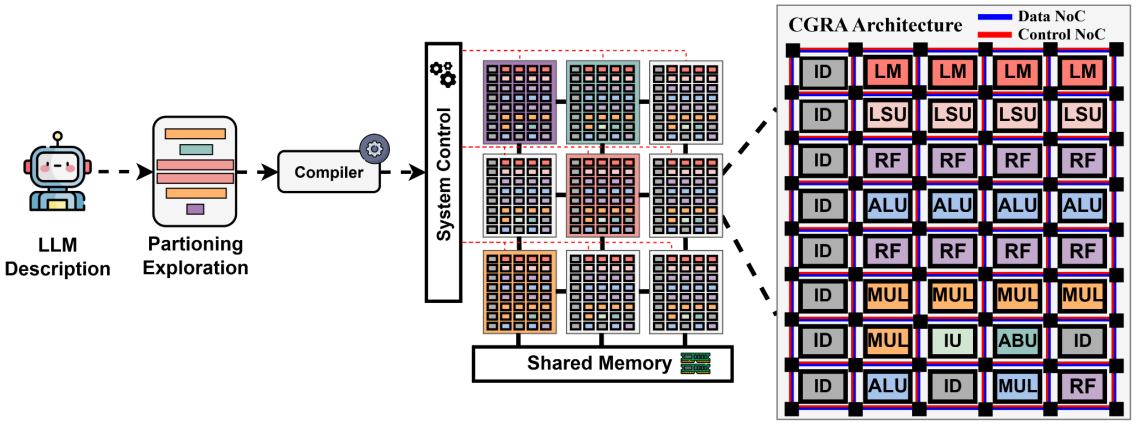


Figure 1: Abstraction of the Thesis Flow

The objective of this diploma thesis, briefly described in Figure 1, is to design and evaluate a multi-CGRA System-on-Chip (SoC) architecture specialized for efficient LLM inference. The proposed system will integrate multiple interconnected CGRA accelerators and investigate how LLM workloads can be partitioned, mapped, and scheduled across these computational resources. The thesis will initially characterize representative LLM inference workloads, identifying their dominant computational kernels, data dependencies, memory requirements, and available parallelism. In parallel, the student will investigate the scalability **limits of conventional single-fabric CGRAs,** studying how increasing fabric size affects mapping efficiency, utilization, communication requirements, and achievable performance.

Based on these observations, different multi-CGRA organizations will be explored, considering the **number and size of CGRA instances, memory hierarchy, inter-CGRA communication, workload partitioning, and scheduling**. An important question will be whether several appropriately sized CGRAs can provide better utilization and scalability than an equivalently sized monolithic fabric. A central research direction will also be the investigation of **heterogeneous** CGRA specialization. Rather than requiring every instance to provide identical capabilities, individual CGRAs may be optimized for different classes of LLM operations. The thesis will investigate whether such specialization can improve utilization and energy efficiency while retaining the programmability advantages of reconfigurable architectures.

RELATED MATERIAL:

[1] Barry de Bruin, Kanishkan Vadivel, Mark Wijtvliet, Pekka Jääskeläinen, and Henk Corporaal. 2024. R-Blocks: an Energy-Efficient, Flexible, and Programmable CGRA. ACM Trans. Reconfigurable Technol. Syst

[2] G. Alexandris *et al.*, "Voltage Aware Approximate CGRA Synthesis for Energy Efficient DNN Inference," *2026 Design, Automation & Test in Europe Conference (DATE)*

[3] Cheng Tan, Miaomiao Jiang, Yuqi Sun, Ruihong Yin, Yanghui Ou, Qing Zhong, Lei Ju, and Jeff Zhang. 2026. Neura: A Unified Framework for Hierarchical and Adaptive CGRAs. 2026, ASPLOS.

Contact Information:

- Alexis Maras, Ph.D Student, NTUA: (amaras@microlab.ntua.gr)
- Georgios Alexandris, Ph.D Student, NTUA: (galexandris@microlab.ntua.gr)
- Panagiotis Chaidos, Ph.D Student, NTUA: (pchaidos@microlab.ntua.gr)
- Georgios Mytilinaios, Ph.D Student, NTUA: (gmytilinaios@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)