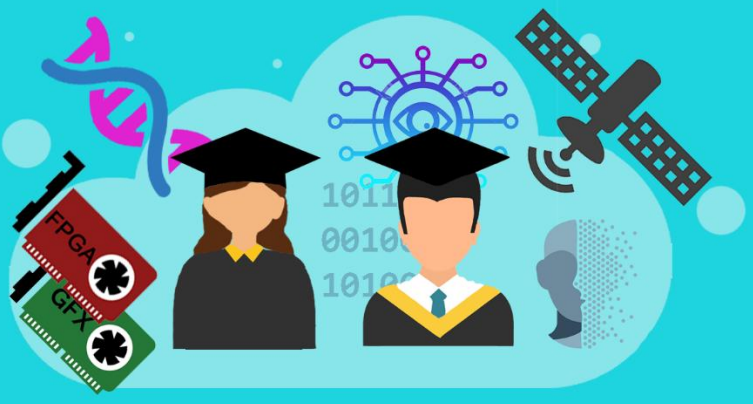
Diploma Thesis

Microproccessors and Digital Systems Laboratory

&



MAX PLANCK INSTITUTE FOR INFORMATICS

Heterogeneous Accelerator Design for Disaggregated RAG-based Large Language Models

Retrieval-Augmented Generation (RAG) based Large Language Models (LLMs) represent a paradigm shift in how AI systems access and utilize information. These models rely on large vector datasets that encode knowledge as high-dimensional embeddings, enabling them to retrieve relevant information during inference through similarity search in vector space. RAG-based LLMs work by first retrieving contextually relevant documents or texts from vector databases using query embeddings, and then feeding this retrieved information as additional context to the base LLM model for the final output generation. This approach allows the model to access up-to-date, domain-specific information that was not present during training, while maintaining the model's reasoning and generation capabilities.

RAG-based LLMs demonstrate superior efficiency compared to traditional LLMs that rely on massive parameter counts to encode knowledge directly in model weights. Instead of burdening the language model with enormous parameters to store factual information, RAG systems offload information storage to external vector databases. This separation reduces the computational overheads associated with massive parameter matrices. However, RAG systems introduce their own challenges, as they require substantial memory resources (memory capacity) to store and efficiently search through large vector datasets and embedding tables. This may create memory bottlenecks on modern GPUs that can limit system scalability and performance.

This research project will investigate the design of disaggregating the two distinct computational phases of RAG-based LLMs—retrieval and attention-based inference—by executing them on heterogeneous accelerators that match their distinct compute-memory characteristics. The retrieval phase, which involves memory-intensive vector similarity searches, will be mapped to FPGAs that excel at memory bandwidth and custom compute patterns, while the compute-intensive attention kernels will be executed on GPUs that have massive parallel processing capabilities. We will investigate and implement efficient vector search algorithms, including both exact k-nearest neighbors and approximate nearest

neighbors methods, optimized for FPGA architectures. Additionally, we will propose system-level optimizations for load balancing across heterogeneous resources, intelligent search space exploration, and efficient communication strategies that minimize data transfer overheads between FPGA and GPU devices. The final goal is to propose an efficient RAG-based LLM serving system for FPGA-GPU platforms.

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] Wenqi Jiang, Marco Zeller, Roger Waleffe, Torsten Hoefler, and Gustavo Alonso, “Chameleon: a Heterogeneous and Disaggregated Accelerator System for Retrieval-Augmented Language Models”, VLDB 2025

[2] Chaoqiang Liu, Haifeng Liu, Dan Chen, Yu Huang, Yi Zhang, Wenjing Xiao, Xiaofei Liao, Hai Jin, “HeterRAG: Heterogeneous Processing-in-Memory Acceleration for Retrieval-augmented Generation”, ISCA 2025

[3] Wenqi Jiang, Hang Hu, Torsten Hoefler, and Gustavo Alonso, "Fast Graph Vector Search via Hardware Acceleration and Delayed-Synchronization Traversal", VLDB 2025

[4] Michael Shen, Muhammad Umar, Kiwan Maeng, G. Edward Suh, Udit Gupta, “Hermes: Algorithm-System Co-design for Efficient Retrieval-Augmented Generation At-Scale”, ISCA 2025

Contact Information:

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

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

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