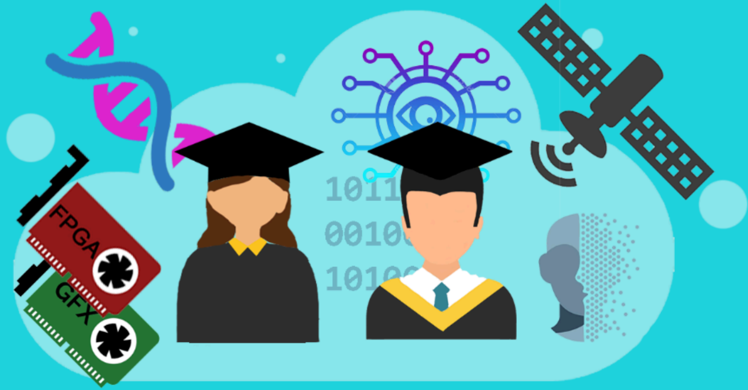
Diploma Thesis

Microproccessors and Digital Systems Laboratory



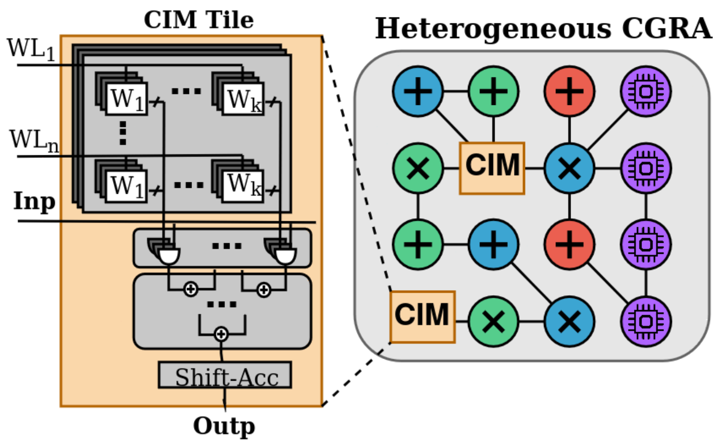
Compute-In-Memory Enabled CGRA Architectures for Next-Gen DNN Applications

Deep Neural Networks (DNNs) now power applications from image, video, and audio processing to text and music generation. With the rise of Edge AI, running inference close to the data source for IoT and near-sensor platforms, there is growing pressure to meet real-time constraints under tight energy budgets. Yet integrating DNN workloads at the edge remains difficult because DNNs are compute- and data-movement-intensive, demanding vast numbers of MAC operations and frequent memory accesses that quickly exhaust ultra-low-power envelopes.

- **SRAM Compute-In-Memory (CIM)** addresses this by executing arithmetic within or adjacent to memory arrays, dramatically reducing data transfers. Digital CIM designs, especially, provide great stability, predictable energy profiles and high flexibility when it comes to mapping complex AI applications[1].

- **Coarse-Grained Reconfigurable Architectures (CGRAs)** provide a compelling alternative to fixed-function tensor engines by combining ASIC-like efficiency with post-deployment flexibility across diverse neural workloads[2]. Heterogeneous PEs for DNN acceleration have attracted attention and attempts to produce efficient mappings for such architectures have been proposed[3], [4].

This thesis aims to fill this gap by embedding CIM tiles within a heterogeneous CGRA. CIM tiles act as high-throughput, low-data-movement unit for dense linear algebra (GEMM/conv/attention), while neighboring programmable tiles handle control, non-linear ops, sparsity, and precision adaptation. This reduces energy per inference by attacking the dominant cost, data movement, without sacrificing programmability or portability.



The diploma thesis focuses on the following areas (but is not limited to):

- Studying the execution patterns of State-of-the-Art LLMs and dynamic DNNs
- Modeling HW-Aware faults for modern DNN accelerators
- Conduct experiments in FPGA/ASIC flow

RELATED MATERIAL:

[1] J. Sun, P. Houshmand and M. Verhelst, "Analog or Digital In-Memory Computing? Benchmarking Through Quantitative Modeling," *2023 ICCAD*

[2] 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. https://doi.org/10.1145/3656642

[3] Tan, Cheng & Jiang, Miaomiao & Patil, Deepak & Ou, Yanghui & Li, Zhaoying & Ju, Lei & Mitra, Tulika & Park, Hyunchul & Tumeo, Antonino & Zhang, Jeff. (2024). ICED: An Integrated CGRA Framework Enabling DVFS-Aware Acceleration. 1338-1352. 10.1109/MICRO61859.2024.00099.

[4] Georgios Alexandris, Panagiotis Chaidos, Alexis Maras, Barry de Bruin, Manil Dev Gomony, Henk Corporaal, Dimitrios Soudris, Sotirios Xydis, "A Unified Framework for Mapping and Synthesis of Approximate R-Blocks CGRAs", https://doi.org/10.48550/arXiv.2505.23553

PREREQUISITES:

- Familiarity with Machine Learning/Neural Network Models Architectures
- Python, Pytorch/ONNX, Bash, Digital Design

CONTACT INFORMATION:

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