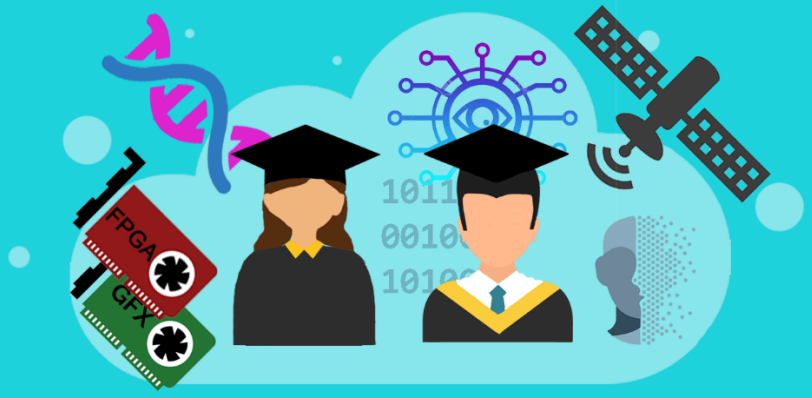


Diploma Thesis

Microprocessors and
Digital Systems
Laboratory



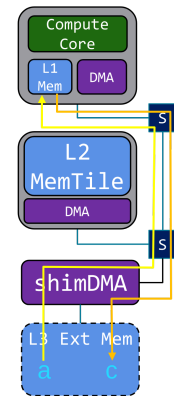
Data Reuse and Buffer Allocation in Software-Managed DNN Accelerators

The efficiency of modern DNN and LLM accelerators is dominated by data movement, where fetching an operand from off-chip DRAM costs orders of magnitude more energy than the MAC operation that consumes it [1]. Consequently, performance and energy depend on how well a mapping exploits data reuse across the memory hierarchy, and on how on-chip buffers are sized, partitioned and allocated to tensors. Systematic loop and data-layout transformations drastically reduce memory traffic and footprint [2]. While state-of-the-art DNN mapping frameworks [3], [4] usually assume idealized, homogeneous buffers, **this thesis will address platform-specific constraints of varying EDDO (Explicitly Decoupled Data Orchestration) architectures with explicitly managed memory types** (Scratchpads, FIFOs, discrete BRAM/URAM blocks and NVIDIA Buffets).

The **main goal** of the thesis is to **identify and exploit data reuse and buffer allocation opportunities on accelerators with software-managed memories**, combining loop-nest analysis with the specific characteristics and constraints of each platform.

Target Platforms:

- **AMD NPU w/ AI Engines:** AIEs constitute arrays of VLIW/SIMD tiles with explicitly managed local memories and memory tiles, where all data movement is programmed through DMAs and streams, using local and shared scratchpads. GEMM tiling on AIE-ML [5] and compiler flows that map loop nests directly onto AMD NPUs [6] show that reuse-aware buffer placement is critical for performance.
- **FPGAs:** On-chip memory consists of discrete BRAM/URAM blocks with fixed widths and depths, so buffer partitioning, banking and double-buffering decisions directly translate to resource utilization and parallelism. FPGAs also have the capability of instantiating fine grained FIFOs for memory management.
- **NVIDIA Buffets:** An efficient and composable storage idiom for EDDO [7], [8], offering fine-grained synchronization between filling and draining a buffer. Buffets enable reuse patterns that conventional double-buffering does not capture, by bringing together the random access nature of scratchpads with the low-cost control overhead of FIFOs.



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

- Analyzing temporal and spatial reuse in the loop nests of modern DNN workloads
- Automated exploration of tiling, loop ordering and buffer sizing/allocation
- Validation on real hardware (AMD Ryzen AI NPU, FPGA boards)

RELATED MATERIAL:

- [1] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," 2014 IEEE International Solid-State Circuits Conference (ISSCC), pp. 10–14, 2014.
- [2] Uday Bondhugula, et al, A practical automatic polyhedral parallelizer and locality optimizer. PLDI '08
- [3] A. Parashar et al., "Timeloop: A Systematic Approach to DNN Accelerator Evaluation," 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2019.
- [4] L. Mei, P. Houshmand, V. Jain, S. Giraldo and M. Verhelst, "ZigZag: Enlarging Joint Architecture-Mapping Design Space Exploration for DNN Accelerators," IEEE TC, 2021
- [5] K. M. Mhatre, E. Taka, and A. Arora, "GAMA: High-Performance GEMM Acceleration on AMD Versal ML-Optimized AI Engines," FPL, 2025.
- [6] Erwei Wang, et al., "From Loop Nests to Silicon: Mapping AI Workloads onto AMD NPUs with MLIR-AIR". ACM TRETs 2026, <https://doi.org/10.1145/3785670>
- [7] M. Pellauer et al., "Buffets: An Efficient and Composable Storage Idiom for Explicit Decoupled Data Orchestration," ASPLOS, 2019.
- [8] Angshuman Parashar, Prasanth Chatarasi, and Po-An Tsai, "Hardware Abstractions for targeting EDDO Architectures with the Polyhedral Model", IMPACT 2021

PREREQUISITES:

- Computer Architecture and memory hierarchies, familiarity with DNNs.
- C/C++, Python, Bash, HLS/FPGA design, MLIR/compiler

CONTACT INFORMATION:

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