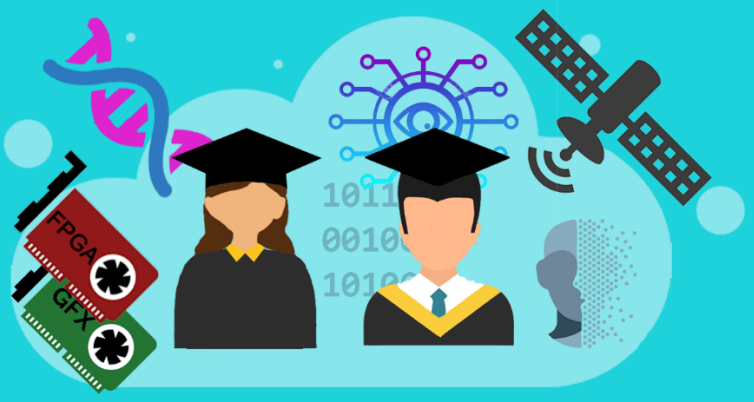
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Vectorized RISC-V for Efficient Deployment of DNN Models

The increasing demand for deploying Deep Neural Networks (DNNs) on edge devices has driven the need for hardware platforms that can deliver high computational efficiency under strict power and area constraints. RISC-V, an open and extensible instruction set architecture (ISA), has emerged as a promising candidate for such deployments due to its modularity and support for custom extensions. In particular, vector extensions to RISC-V offer a compelling approach to accelerate the parallel computations inherent in DNN workloads [1]. By leveraging data-level parallelism, vectorized RISC-V architectures can significantly improve inference tasks' throughput and energy efficiency.

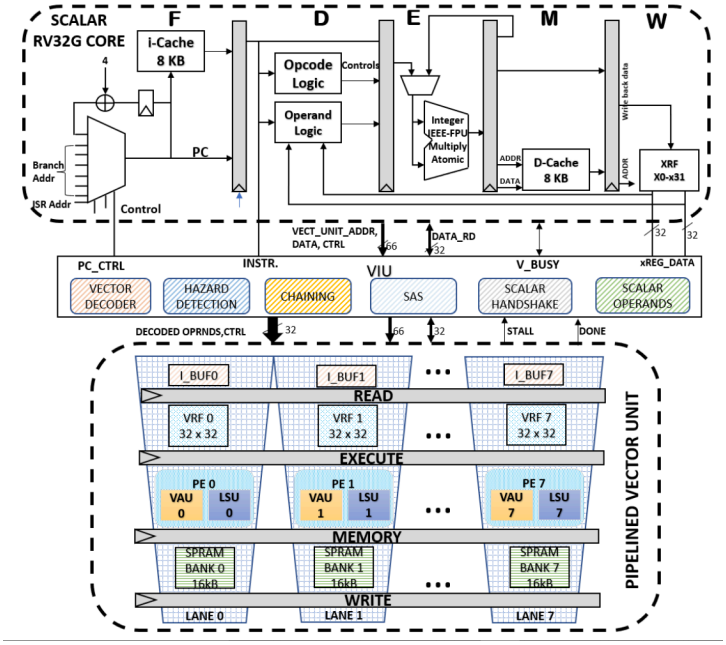


Figure 1: Overview of a Vectorized RISC-V processor

In recent years, transformer-based models have gained significant traction beyond their original applications in natural language processing. Vision Transformers (ViTs) have demonstrated competitive or superior performance compared to traditional

convolutional architectures in various computer vision tasks, while Large Language Models (LLMs) continue to push the boundaries of generative AI and natural language understanding. These models are characterized by their heavy reliance on matrix multiplications, making them well-suited for acceleration through vectorized computation. However, their computational and memory demands pose challenges for deployment on resource-constrained edge devices. This makes the exploration of vectorized RISC-V architectures particularly relevant, as they offer a customizable and energy-efficient platform for tailoring hardware to the specific needs of transformer-based inference [3]. By aligning vector extensions' capabilities with transformers' computational patterns, this thesis aims to unlock new opportunities for deploying state-of-the-art models at the edge.

The primary objective of this thesis is to investigate how vectorized RISC-V architectures can be leveraged to efficiently deploy deep neural network models, with a particular focus on transformer-based architectures such as ViTs and LLMs. To achieve this, the thesis (not limited to one person) will pursue the following goals:

- **Workload characterization** of such models on vectorized RISC-V processors
    - Identification of **bottlenecks**
- **Software-Hardware Co-design**: ISA extensions and design of specialized hardware computational units to optimize inference
- Apply compression techniques such as pruning and quantization to reduce model complexity and enable efficient deployment on resource-constrained RISC-V systems
    - **Trade-off** assessment between **accuracy** and **performance**
- **Explore architectural configurations**, including the:
    - number of RISC-V cores and
    - memory hierarchy organization,

  to identify optimal trade-offs between performance, area, and energy efficiency.

REFERENCES:

[1] GitHub - pulp-platform/ara

[2] V. Titopoulos, K. Alexandridis, C. Peltekis, C. Nicopoulos and G. Dimitrakopoulos, "Optimizing Structured-Sparse Matrix Multiplication in RISC-V Vector Processors," in *IEEE Transactions on Computers*, vol. 74, no. 4, pp. 1446-1460, April 2025, doi: 10.1109/TC.2025.3533083

[3] C. Wang, C. Fang, X. Wu, Z. Wang and J. Lin, "SPEED: A Scalable RISC-V Vector Processor Enabling Efficient Multiprecision DNN Inference," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 33, no. 1, pp. 207-220, Jan. 2025, doi: 10.1109/TVLSI.2024.3466224

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)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)