Available MSc theses

Last update: March 10, 2025

Compilers and computer architectures for deep learning

1.1 - Design coherent RISC-V multicores

Cache coherence is the uniformity of shared resource data stored in multiple local caches. When multiple processors in a multiprocessor system maintain caches of a shared memory resource, problems may arise with incoherent data.

The thesis foresees integrating a set of already-developed cores within a multiprocessor system on an FPGA target and developing a hardware mechanism for maintaining cache coherence within the multiprocessor, leveraging existing CPU and SoC developed in-house.

1.2 - End-to-end compiler from PyTorch to RISC-V

While the inference of large ML and DL models is achieved by means of massive GPUs, the quest to maximize computational efficiency and the need to make such models available at the edge fueled the investigation on novel programmable platforms. In this scenario, the royalty-free RISC-V ISA offers an interesting opportunity to explore novel and efficient computing platforms while the compiler support is paramount to support such research wave.

The thesis aims to investigate and optimize our in-house end-to-end compiler that allows us to generate RISC-V binaries from a PyTorch description of the model.

The thesis allows the student to understand the internals of modern compilers as well as the optimization opportunities when the compiler can take advantage of the available architectural and microarchitectural information.

1.3 – Programmable RISC-V-based accelerator for LLMs

Large language models (LLMs) and foundation models require massive computation due to their deep transformer-based architectures, high memory bandwidth demands, and irregular data dependencies. These characteristics make them challenging yet rewarding targets for acceleration, as conventional hardware struggles with both efficiency and flexibility. Given the rapid evolution of AI models, accelerators must also be programmable to support diverse workloads and future algorithmic advancements.

This thesis explores the design of a programmable RISC-V-based accelerator optimized for LLM inference, balancing performance, energy efficiency, and adaptability. The research will focus on optimizing tensor processing, memory hierarchies, and instruction extensions while ensuring programmability for different AI frameworks. The accelerator's efficiency will be evaluated through real-world AI benchmarks and hardware prototyping.

Students will gain expertise in AI hardware acceleration, RISC-V microarchitecture, and workload-driven optimization, and they will develop hands-on skills in accelerator design and evaluation.

1.4 – FPGA-based accelerator of neural networks for event cameras

Event cameras capture sparse, asynchronous brightness changes, enabling low-latency, power-efficient vision processing. FARSE-CNN, an open-source hybrid RNN-CNN model, efficiently processes these event streams but remains computationally demanding for real-time inference.

This thesis focuses on designing an FPGA accelerator for FARSE-CNN, implementing hybrid recurrent-convolutional units and optimizing dataflow to leverage input sparsity. The design will include efficient memory architectures and be evaluated against CPU and GPU baselines for latency, throughput, and energy efficiency.

Students will gain expertise in FPGA-based deep learning acceleration, RTL design, and sparse data processing, and they will develop hands-on skills in SystemVerilog, FPGA toolchains, and performance benchmarking for hardware acceleration and embedded AI systems.

See 3.1 for a complementary MSc thesis on the learning-side aspects of FARSE-CNN

1.5 – Design of RISC-V vector co-processors for deep learning

Vector processing is key to efficient acceleration of deep-learning models, which heavily rely on matrix-matrix and matrix-vector multiplications, and RISC-V's open ISA offers unique opportunities for optimizing vector co-processors.

This thesis explores the design, implementation, and evaluation of RISC-V-based vector processing units for deep-learning workloads. The research will analyze existing architectures, identify bottlenecks, and propose optimizations for performance and efficiency. It will focus on workload-specific tuning, including memory access, parallelism, and instruction-level improvements. The final design will be validated through simulation and FPGA prototyping.

Students will gain expertise in RISC-V vector extensions, hardware acceleration, and performance benchmarking, and they will develop hands-on skills in simulation, hardware description languages, and FPGA-based design.

1.6 – Programmable RISC-V-based interconnect for real-time scenarios

Efficient data movement is critical in real-time computing, where deterministic performance and low-latency communication are essential. A programmable interconnect for multi-core RISC-V-based systems can enable dynamic adaptation to application workloads to ensure real-time execution of critical tasks.

This thesis will focus on the design of a scalable interconnect with configurable communication policies, allowing workload-aware routing, quality-of-service (QoS) enforcement, and predictable latency. The resulting RTL design will be validated through simulation and FPGA prototyping.

Students will gain expertise in on-chip communication, real-time constraints, and hardware-software co-design, and they will develop skills in interconnect modeling, FPGA implementation, and performance optimization for real-time multi-core systems.

Machine learning and analog circuit design

2.1 - Machine learning to optimize the layout of analog circuits

Traditionally, the layout of analog circuits is a manual and error-prone process preventing the scalability of the entire analog design flow. At present, a few seminal frameworks leveraging ILP and coarse-grained heuristics constitute the state of the art.

The thesis aims to investigate the use of machine learning techniques, e.g., deep learning and online learning, to automate and optimize the layout of complex analog circuits. The experimental evaluation will be carried out by integrating the ML algorithms into available analog design frameworks.

The thesis allows the student to experiment with industrial grade technology libraries as well as with commercial and open-source CAD tools.

2.2 - Device generator for industrial technology libraries targeting analog circuits [industrial internship]

The technology library defines the silicon components that can be implemented by the foundry to deliver the final tape out of any analog design. The use of novel technology libraries is paramount to maximize the efficiency of the final product. To foster the adoption of novel technology libraries, the creator of the library must provide a design toolkit to allow the circuit designer to use the technology library. The device generator is a key component of the design toolkit since it allows the EDA of choice to use the technology library.

The thesis aims to develop a novel device generator for an industry-grade technology library and its integration into available open-source analog CAD tools.

The thesis allows the student to understand the intimate details within modern technology libraries as well as the optimization of commercial and open-source CAD tools targeting analog circuit design.

2.3 - ML-driven automation for post-layout optimization of analog circuits

Fabrication is the ultimate purpose of analog circuits design, it highly depends on correct post-layout verification. This process consists of three primary verification steps:

- Design Rule Checking (DRC): Validates that the layout adheres to foundry-specific manufacturing rules, ensuring manufacturability and yield.

- Layout vs. Schematic (LVS): Confirms that the physical layout accurately represents the intended circuit design by comparing it to the schematic netlist.
- Testbench Simulation: Assesses circuit functionality and performance under different operating conditions to verify compliance with design specifications.

Despite the maturity of these verification techniques, their automation remains limited, often requiring manual intervention and interpretation. A significant challenge is the lack of an efficient framework to correlate design flaws with their root causes, making debugging and optimization time-consuming.

This thesis aims to develop an automated post-layout verification flow that integrates intelligent analysis techniques to enhance verification efficiency, by leveraging machine learning and algorithmic optimization.

The outcome is expected to reduce verification effort, improve design turnaround time, and enhance the reliability of complex analog circuits.

2.4 - Transfer function pruning in analog circuits with machine learning techniques

In analog design, the transfer function serves as a mathematical representation of a circuit's intended behavior. However, real-world layout generation introduces *stray capacitance,* which alters the transfer function, potentially leading to unexpected outputs and performance deviations.

A major challenge is obtaining a compact and interpretable comparison between the post-layout transfer function and its ideal counterpart. Existing approaches often fail to effectively highlight deviations and their root causes.

This thesis aims to develop a machine learning-based framework to analyze and prune the circuit's transfer function, correlating the new members to the circuit components that introduces them and making it more interpretable. By systematically comparing the extracted post-layout transfer function with the original design, the framework will identify key differences and potential sources of error. The pruning process will be circuit-aware, allowing for localized transfer function computation at different points within the circuit.

2.5 - Layout generation optimization by pre- and post-layout simulation comparison with ML

The primary objective of layout generation, whether performed manually or through automated techniques, is to achieve a physical design whose post-layout behavior closely matches its pre-

layout counterpart. However, this equivalence is often difficult to attain due to the impact of real-world effects introduced during the layout process. Consequently, designers must perform multiple iterative refinements to identify and mitigate the specific devices or interconnections that degrade circuit performance.

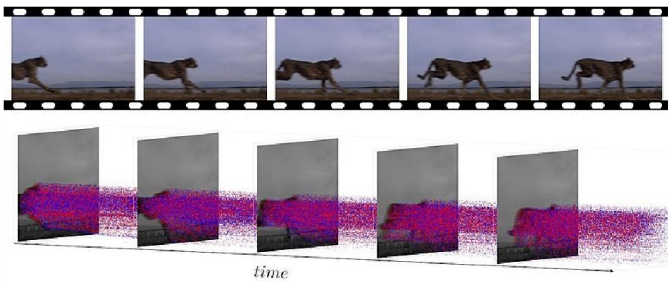
This thesis aims to develop novel methodologies for systematically identifying the circuit elements that most significantly influence post-layout deviations. By leveraging machine learning techniques, the proposed approach will analyze the discrepancies between pre- and post-layout simulations, providing insights into the most critical performance degradations.

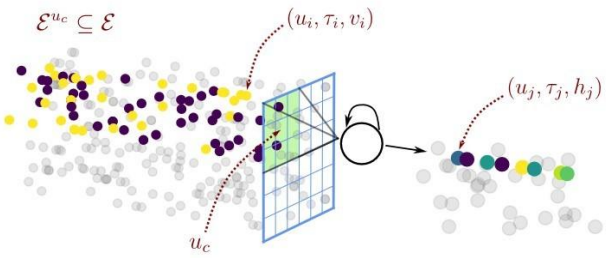
The expected outcome is a framework capable of guiding layout optimization by predicting potential issues early in the design process, thereby enhancing design convergence and overall performance reliability.

Collaboration with Prof. Matteucci

Contacts: matteo.matteucci@polimi.it, riccardo.santambrogio@polimi.it

3.1 - Asynchronous models for event-based vision





Event-based cameras are novel sensors that measure light variations with their independent pixels, generating a stream of data that is inherently sparse and asynchronous. Despite this, the standard approach for processing this data is to convert it into dense 2D or 3D representations and apply well-known models of frame-based computer vision. This approach achieves state-of-the-art results on many tasks and benchmarks, but it fails to exploit some of the key advantages of event cameras, such as low latency and efficiency in sparsity. Only recently, researchers started developing deep learning models tailored for event data, that exploit sparsity and process the streaming asynchronous events with minimal recomputation. These methods have shown very promising results, but their efficacy for various applications is still unclear, especially when compared to their dense counterparts.

The goal of this thesis is to explore the application of asynchronous models for event-based vision in different scenarios and tasks. These can include object detection, visual odometry, and more. In particular, the work will be focused on FARSE-CNN, an asynchronous model that we developed by combining the mechanisms of recurrent and convolutional neural networks, which has already proved effective on various benchmarks in object recognition, gesture recognition, and object detection.

Note that, while the goal of this work is not to develop a new asynchronous model, the application of existing architectures to different tasks might require considerable adaptations.

Suggested Skills: Python, PyTorch

See 1.4 for a complementary MSc thesis on the hardware acceleration of FARSE-CNN