POLITECNICO DI MILANO

PhD in INGEGNERIA DELL'INFORMAZIONE / INFORMATION TECHNOLOGY - 41st cycle

Research Area n. 1 - Computer Science and Engineering

THEMATIC Research Field: AI FOR HETEROGENEOUS HARDWARE ACCELERATOR DESIGN

Monthly net income of PhDscholarship (max 36 months)
1600.0 In case of a change of the welfare rates during the three-year period, the amount could be modified.

Context of the research activity	
Motivation and objectives of the research in this field	FPGAs are increasingly used as flexible accelerators for modern AI workloads that combine dense tensor operations, irregular memory access patterns, and graph-like computations under strict performance and energy constraints. High-Level Synthesis (HLS) has narrowed the productivity gap, yet current flows struggle to fully exploit the structure of models developed in frameworks such as PyTorch and TensorFlow, and to naturally target heterogeneous, multi-kernel FPGA architectures. The emergence of MLIR as a multi-level, extensible compiler infrastructure provides an ideal backbone for representing AI computations, ranging from high-level tensor graphs to hardware-oriented dialects and HLS-compatible code, enabling systematic optimizations at the appropriate abstraction level. mlir.llvm.org Recent work on MLIR-based HLS frameworks and hierarchical dataflow compilation, such as StreamHLS, ScaleHLS, and HIDA, has demonstrated the effectiveness of multi-level IRs and structured dataflow transformations for FPGA synthesis, while AutoScaleDSE and related approaches underline the importance of scalable design space exploration tightly integrated with the compilation flow. Within this landscape, Large Language Models (LLMs) are emerging both as relevant AI benchmarks (e.g., transformer-based inference) and as intelligent assistants capable of

	suggesting code transformations, pragma configurations, and design variants, but their integration into a principled FPGA compilation methodology is still largely unexplored. The project is motivated by the need to close the gap between rich AI software specifications and efficient heterogeneous FPGA implementations, reducing manual tuning while targeting architectures that go beyond single accelerators and embrace multi-threaded, streaming, and multi-kernel organizations. The core objective is to define a complete methodology, from PyTorch/TensorFlow models to FPGA-based heterogeneous accelerators, built on an MLIR-centric flow and advanced HLS backends. Concretely, the PhD will design techniques to extract parallelism and dataflow structure from AI models, map them to multi-level MLIR dialects, and generate efficient hardware through tools such as Vitis HLS and PandA-bambu, complemented when needed by new MLIR-based synthesis components. The research will pursue quantitative improvements in performance, resource efficiency, and design time, while demonstrating the approach on representative AI workloads, including LLM components and vision/ML kernels relevant for HPC and edge scenarios.
Methods and techniques that will be developed and used to carry out the research	The candidate will develop an end-to-end compilation and synthesis framework that couples MLIR-based intermediate representations with HLS tools and custom backends. This includes front-end pipelines from PyTorch/TensorFlow graphs to MLIR; analysis and transformation passes that expose coarse- and fine-grain parallelism, derive streaming and hierarchical dataflow architectures, and orchestrate heterogeneous multi-kernel designs; and backend flows that generate optimized HLS or RTL for FPGA targets using Vitis HLS, PandA-bambu (with dedicated extensions), and MLIR-driven code generation. Building on ideas from StreamHLS, ScaleHLS, HIDA and AutoScaleDSE, the methodology will embed design space exploration as a first-class component, combining analytical and empirical models to explore choices such as unrolling, tiling, pipelining, memory partitioning, and inter-kernel communication. LLMs will be integrated as optimization partners: they will

	be used to propose transformation sequences, pragma settings, code refactorings and architectural templates, and to refine the exploration strategy based on previous synthesis results, under the supervision of verifiable, rule-based compiler passes that guarantee correctness and reproducibility.
Educational objectives	The PhD will train the candidate at the intersection of reconfigurable computing, compilers, and AI. By the end of the program, the student is expected to master FPGA and heterogeneous architectures; to be fully proficient with HLS methodologies and tools (including industrial solutions and open-source frameworks like PandA-bambu); to gain hands-on expertise with MLIR and LLVM-based compiler development; and to build a solid understanding of contemporary AI models, including LLMs, from both algorithmic and implementation viewpoints. The project will also foster competencies in performance modeling, design space exploration, and the responsible use of LLM-based tools within rigorous engineering workflows, as well as experience in scientific publication, open-source development, and international collaborations.
Job opportunities	The profile developed in this PhD is highly aligned with current demands in both academia and industry. Graduates will be competitive for academic and research positions in computer architecture, electronic design automation, reconfigurable computing, and high-performance AI systems. In industry, the expertise in MLIR/LLVM-based compilation, HLS, FPGA accelerators, and AI-driven optimization is directly relevant for semiconductor companies, FPGA and EDA vendors, cloud and HPC providers, and R&D labs developing domain-specific accelerators and advanced heterogeneous platforms. The combination of methodological skills, practical tooling experience, and familiarity with AI workloads will also be valuable in applied sectors (automotive, telecommunications, finance, healthcare, robotics, and space) that increasingly rely on custom high-performance and energy-efficient computing solutions.

Composition of the research group	0 Full Professors 1 Associated Professors 1 Assistant Professors 2 PhD Students
Name of the research directors	Prof. Fabrizio Ferrandi

Contacts
fabrizio.ferrandi@polimi.it

Additional support - Financial aid per PhD student per year (gross amount)	
Housing - Foreign Students	--
Housing - Out-of-town residents	--

Scholarship Increase for a period abroad	
Amount monthly	800.0 €
By number of months	6

Additional information: educational activity, teaching assistantship, computer availability, desk availability, any other information
EDUCATIONAL ACTIVITIES (purchase of study books and material, including computers, funding for participation in courses, summer schools, workshops and conferences): financial aid per PhD student. TEACHING ASSISTANTSHIP: availability of funding in recognition of supporting teaching activities by the PhD student. There are various forms of financial aid for activities of support to the teaching practice. The PhD student is encouraged to take part in these activities, within the limits allowed by the regulations. COMPUTER AVAILABILITY: 1st year: Yes 2nd year: Yes 3rd year: Yes