ELECTRICAL, ELECTRONICS AND COMMUNICATIONS ENGINEERING

DET - Reconfigurable and Energy-Efficient Hardware Architectures for Neural Networks and Large Language Models Using Low-Precision and Ternary Arithmetic

Funded By	Dipartimento DET
Supervisor	CASU MARIO ROBERTO - mario.casu@polito.it
Contact	CASU MARIO ROBERTO - mario.casu@polito.it LAZARESCU MIHAI TEODOR - mihai.lazarescu@polito.it LAVAGNO LUCIANO - luciano.lavagno@polito.it
Context of the research activity	Design of reconfigurable systolic and CGRA-based hardware for neural networks and LLMs using unconventional arithmetic schemes: ternary weights (-1, 0, +1) and small floating-point formats (FP8–FP4). Hardware/software co-design and NN2FPGA-based automation for FPGA and ASIC prototypes with high efficiency and scalability.
	The proposed PhD project investigates new paradigms for the hardware implementation of deep neural networks and large language models (LLMs) based on low-precision arithmetic and reconfigurable architectures. The continuous scaling of AI models has dramatically increased computational and energy requirements. While GPUs dominate current deep-learning workloads, their general-purpose nature limits efficiency for emerging numeric formats such as FP8, FP6, FP4, or ternary arithmetic. Conversely, reconfigurable hardware platforms like FPGAs and coarse-grained reconfigurable arrays (CGRAs) can tailor datapaths and arithmetic precision to specific workloads, offering a path toward sustainable and adaptive AI hardware. The research will focus on the design of systolic and CGRA architectures capable of executing neural computations with reduced numerical precision. The goal is to determine how far arithmetic precision can be reduced without compromising model accuracy and to identify efficient hardware structures to implement such arithmetic. Special attention will be devoted to ternary-weight networks, in which weights assume values -1, 0, or +1. This representation allows multiplication to be replaced with simple additions and subtractions, drastically simplifying datapaths while reducing memory footprint and power consumption. Parallel to this, the project will explore small-size floating-point formats, evaluating their numerical properties and efficiency for various neural workloads, including transformer-based models used in LLMs. These

Objectives	formats promise significant memory savings while maintaining numerical stability better than integer quantization. The methodology will rely on a hardware/software co-design approach, combining algorithmic modeling with low-level implementation in RTL and HLS. A central element will be the extension of the NN2FPGA framework, an automatic flow that translates trained neural networks into FPGA implementations. The framework will be enhanced to support small floating-point and ternary arithmetic, enabling the automatic generation of optimized compute arrays and memory hierarchies for different precisions. This approach will facilitate rapid evaluation of hardware trade-offs and accelerate design-space exploration. The PhD activity will proceed through three main phases. The first phase will focus on the modeling and simulation of low-precision arithmetic to assess trade-offs between accuracy, robustness, and hardware cost. The second phase will address the design and implementation of parameterizable systolic and CGRA-based accelerators supporting mixed-precision computation, validated on FPGA platforms. In the final phase, the candidate will integrate the proposed units into complete accelerator prototypes, evaluate them on representative DNN and LLM workloads, and perform ASIC synthesis to estimate power, area, and scalability in advanced technology nodes. Expected outcomes include new energy-efficient architectures for AI acceleration using unconventional numeric representations, an extended NN2FPGA design flow supporting ternary and small floating-point models, and FPGA/ASIC prototypes demonstrating substantial gains in performance per watt compared with FP16-based systems. The project aims to contribute to the advancement of low-power, reconfigurable, and sustainable hardware solutions for next-generation artificial intelligence.

Skills and competencies for the development of the activity	Solid background in RTL design (VHDL, Verilog, or SystemVerilog), experience with High-Level Synthesis (HLS) tools (Xilinx Vitis HLS, Intel HLS), and familiarity with hardware/software co-design. Basic understanding of ASIC design flow (logic synthesis, timing, place and route) and of machine learning frameworks (PyTorch, TensorFlow). Strong interest in low-power and reconfigurable architectures for AI acceleration is desirable.