Ph.D. in Information Technology
Thesis Defense

October 15th, 2024
at 14:00
Seminar Room Alessandra Alario – building 21

Stephanie SOLDAVINI – XXXVI Cycle

Design Methods for Simplifying the Creation of High-Performance FPGA Memory Architectures

Supervisor: Prof. Christian Pilato

Abstract:

Specialized hardware accelerators are becoming important for increasingly many applications. Thanks to specialization, these accelerators can achieve high performance and energy efficiency but their design is complex and time consuming. This problem is exacerbated when large amounts of data must be processed, like in modern big data and machine learning applications. The designer has not only to optimize the accelerator logic but also produce efficient memory architectures. This requires hardware platform design expertise and months of effort and to obtain truly efficient designs specific application design knowledge is needed as well, and rarely does the kernel developer have both hardware and application expertise. To simplify this process, in this thesis a multi-level compilation flow is proposed that specializes a customizable memory template to match data, application, and technology requirements. The Olympus system generation tool was developed to integrate several memory and data transfer based optimizations automatically, to relieve an application designer from requiring deep hardware design or platform knowledge. Olympus is built using the multi-level intermediate representation (MLIR) compiler framework to promote extensibility of the tool, allowing compatibility with many compiler-front end tools. Designs generated by the Olympus tool integrating these optimizations were demonstrated to achieve 2.7× speedup and 7.0× energy efficiency over a highly optimized vectorized CPU implementation of a computational fluid dynamics (CFD) application and 6.39× speedup and 5.07× energy efficiency over the CPU implementation of a real industrial traffic modeling use case. These designs are generated from a simple description of the accelerator kernels, allowing an application designer to easily create efficient hardware accelerators. This simplicity also allows the designer to perform design space exploration (DSE) as well, notably in the area of data representations. The CFD application achieved 6.7× speedup and 24.5× energy efficiency over the CPU implementation with data representation tuning while incurring only a minimal loss in result accuracy.

PhD Committee

Prof.**ssa Cristina Silvano, Politecnico di Milano**

Prof. Luciano Lavagno, Politecnico di Torino

Prof.ssa Lana Josipovic, ETH Zurich