Ph.D. in Information Technology: Theses Defenses

February 6th, 2019

Room Conferenze "Emilio Gatti" – 10.45 am

Emanuele DEL SOZZO – XXXI Cycle

"On How To Effectively Target FPGAs From Domain Specific Tools"

Advisor: Prof. **Marco Santambrogio**

Abstract:

Domain specific tools represent an efficient solution to both easily target heterogeneous architectures and increase productivity. Domain Specific Languages (DSLs) and Machine Learning (ML) frameworks are two significant examples. Both permit users to quickly and easily develop portable and efficient designs for multiple architectures. However, although DSLs and ML frameworks are highly effective in assisting users towards the generation of efficient designs for CPUs and GPUs, they still lack a concrete support for FPGAs. Indeed, the whole FPGA design process remains complex and the integration with high-productivity tools and languages is still limited. For these reasons, this research thesis focuses on the development of tools able to efficiently and easily target FPGAs from domain-specific scenarios. In particular, it consists in both a framework for the fast-prototyping and deployment of CNN accelerators on FPGA, and FROST, a unified backend to efficiently hardware-accelerate DSLs on FPGAs. On one hand, the goal of the CNN framework is to bridge the gap between high-productivity ML frameworks, like TensorFlow and Caffe, and FPGA design process. On the other, starting from an algorithm described in one of the supported DSLs, FROST translates it into its Intermediate Representation, performs a series of FPGA-oriented optimizations steps by means of a high-level scheduling co-language, and, finally, generates an optimized design suitable of FPGA tools.

Davide GADIOLI – XXX Cycle

"Dynamic Application Autotuning for Self-Aware Approximate Computing"

Advisor: Prof. **Gianluca Palermo**

Abstract:

In the autonomic computing context, the system is perceived as an ensemble of autonomous elements capable of self-managing, where end-users define high-level goals and the system shall adapt to achieve the desired behaviour. This runtime adaptation creates several optimization opportunities, especially if we consider approximate computing applications, where it is possible to trade off the result accuracy and the performance. Given that modern systems are limited by the power dissipated, autonomic computing is an appealing approach to increase the computation efficiency.

This PhD thesis focuses on a dynamic autotuning framework, named mARGOt, which aims at enhancing the target application with an adaptation layer to provide self-optimization capabilities at the production phase. In this context, the end-user might specify complex high-level requirements and the proposed approach automatically tunes the application accordingly, relying on reactive and proactive adaptation mechanisms.
The mARGOt framework has been evaluated by leveraging its features in two different scenarios. On one hand, we evaluated the orthogonality between resource managers and application autotuning. On the other hand, we proposed an approach to enhance the application with a kernel-level compiler autotuning and adaptation layer in a seamless way for application developers. Moreover, the autotuning framework has been deployed in two real-world application case studies, showing how it is possible to significantly improve computation efficiency, by applying approximate computing techniques and by using mARGOt to manage them.

Giuseppe NATALE – XXXI Cycle

"On How to Design Optimized Spatial Architectures: from Iterative Stencils to Convolutional Neural Networks"

Advisor: Prof. **Marco Santambrogio**

Abstract:

This thesis is focused on a set of algorithms sharing a similar computational pattern, namely iterative stencils and Convolutional Neural Networks (CNNs).
The key computation for both algorithms consists in sliding a filter on the input data, computing new elements using a submatrix of the input. Iterative stencil algorithms are heavily employed in physics, numerical solvers and even finance, while CNNs are one of the recently developed deep learning algorithms, currently used mainly to perform image classification and video analysis.
While High-Level Synthesis (HLS) capabilities have improved dramatically over the recent years, the synthesis tools have yet to reach the level of sophistication required to properly optimize these algorithms, and extract sufficient parallelism or generate highly scalable solution. This thesis objective is to improve the state of the art with respect to FPGA synthesis of these two algorithms.
For both iterative stencils and the features extraction stage of CNNs, we designed optimized spatial architectures that are able to exploit different sources of parallelism of such algorithms and reduce the cost of data movements alleviating the burden on external memory. Moreover, we address the scalability of the proposed solutions implementing the accelerators on very big Stacked Silicon Interconnect (SSI) FPGAs or even on custom-designed multi-FPGA systems.
The validation performed shows that for iterative stencils, the proposed solution is comparable with the state of the art for the single FPGA implementation, but we are able to perform substantially better on multi-FPGA,

thanks to an approximately linear scaling in performance. Moreover, in the case of CNNs our implementations for the well-known AlexNet and VGG-16 achieve a throughput of respectively 1.61 and 2.99 TOPS. For VGG-16 we are the second-best implementation -- by a narrow margin -- available in the State of the Art, while for AlexNet we substantially outperform the previous work.

Marco RABOZZI – XXXI Cycle

"CAOS: CAD as an Adaptive Open-platform Service for High Performance Reconfigurable Systems"

Advisor: Prof. **Marco Santambrogio**

Abstract:

In current years we are assisting at a new era of computer architectures, in which the need for energy-efficiency is pushing towards hardware specialization and the adoption of FPGAs.

Despite the potential benefits given by embracing reconfigurable hardware both in the HPC and cloud contexts, we notice that one of the main limiting factor to the widespread adoption of FPGAs is complexity in programmability, as well as the effort required to port a pure software solution to an efficient hardware-software implementation targeting reconfigurable heterogeneous systems.

The main objective of this dissertation is the development of CAD as an Adaptive Open-platform Service (CAOS), a platform able to guide the application developer in the implementation of efficient hardware-software solutions for high performance reconfigurable systems. The platform aims at assisting the designer starting from the high level analysis of the code, towards the definition of the functionalities to be accelerated on the reconfigurable nodes. Furthermore, the platform guides the user in the selection of a suitable architectural template for the FPGA devices and the final implementation of the system together with its runtime support. Finally, CAOS has been designed to facilitate the integration of external contributions and to foster research on the development of Computer Aided Design (CAD) tools for accelerating software applications on FPGA-based systems.

Alberto SCOLARI – XXXI Cycle

"Optimizing Data-Intensive Applications for Modern Hardware Platforms"

Advisor: Prof. **Marco Santambrogio**

Abstract:

Data-intensive workloads like Data Analytics (DA) and Machine Learning (ML) are increasingly important and have ever-growing demands. These applications often run in the cloud with multi-core, server-class CPUs, which have complex features to leverage for better performance but also have shared resouces that applications may contend. To abstract the hardware and the underlying software stack, cloud providers are

increasingly offering DA and ML services as "Frameworks" that automatically provision resources and even offload the most compute-intensive tasks to specialized hardware. This thesis investigates these aspects and proposes solutions to enhance the performance and predictability of these workloads. In particular, it investigates and evaluates the design principles for a framework runtime system that retains the programmability of existing approaches while thoroughly optimizing each application for the available hardware.

PhD Committee:
Prof. **Gianluca Palermo**, DEIB
Prof. **Josè L. Ayala**, Complutense University of Madrid
Prof. **Roberto Giorgi**, Universita' di Siena