Ph.D. in Information Technology: Thesis Defense

October 29th, 2020

online by Microsoft Teams – at 11.00

Rolando BRONDOLIN – XXXII Cycle

"ON THE MANAGEMENT OF POWER AND PERFORMANCE TRADE-OFFS IN DISTRIBUTED CLOUD-NATIVE INFRASTRUCTURES"

Advisor: Prof. **Marco Santambrogio**

Abstract:

CLOUD computing is now the de-facto standard for the deployment of complex and scalable applications and systems at scale. In the last few years, cloud computing applications shifted from the monolithic architecture to a more flexible microservice-based architecture with the so-called cloud-native ecosystem. This shift allowed to separate concerns among different development teams, increased the scalability of the cloud applications and allowed to develop, test, and deploy each functionality almost independently from the rest of the system. Cloud-native applications fostered even more the growth of cloud computing and, for this reason, cloud providers have to manage an unprecedented amount of applications for a huge amount of users. This trend poses new challenges in the management of data-centers. In particular, the expected energy usage of data-centers will reach 8% of the whole energy consumption of the world by 2030. Moreover, power consumption represents 20% of the Total Cost of Ownership (TCO) of a data-center. If we consider that the CPU is currently the most power-hungry component of a server, there is the need to optimize how cloud applications are executed within cloud infrastructures to keep the cloud-computing growth sustainable.

Within this context, the goal of this thesis work is the design and development of power management techniques able to sustain the performances requested by cloud-native applications and workloads while reducing as much as possible the power consumption such applications generate. Given the complexity that microservice-based applications bring, we decided to design a fully automated system to manage power consumption and performance leveraging the Observe Decide Act (ODA) autonomic control loop. This allowed us to focus on how to measure and monitor in a fine grain way performance and power consumption, on how to allocate power and

performance, and on how to actuate the control decisions defined by the ODA loop. Within this thesis, we designed a fully-black-box approach to attribute power consumption, measure resource usage, and monitor network performance. We leveraged these metrics to define reactive control policies able to maintain CPU usage and latency near a user-specified target. Then, we enforced in a timely and precise way the power budgets derived by the performance constraints.

Finally, we explored how we can improve the energy efficiency of microservice based applications by introducing heterogeneous architectures, merging together the elasticity of cloud-native applications and the performance and energy-efficiency of Field Programmable Gate Arrays (FPGAs). This work represents an interesting initial study and paves the way for more extensive research work on how to accelerate microservices and cloud-native workloads.

Lorenzo DI TUCCI – XXXII Cycle

"**HUGENOMIC: EXPLOITING FPGAS AS HARDWARE ACCELERATORS IN THE GENOMIC DOMAIN**"

Advisor: Prof. **Marco Santambrogio**

Abstract:

IN current years, we are assisting to an ever-increase request of performance in the genomic field, where human genome research will likely transform medical practices. The information that could be obtained from the genetic profile of different species and the knowledge of molecular basis of biological processes are leading to the development of more precise therapies and novel treatments. Nonetheless, the high complexity of algorithms used to process such data, paired with the necessity to find new optimal computing architectures due to the end of Moore's law and Dennard Scaling require a computational effort that may limit improvements.

Within this context, Domain Specific Architectures (DSAs) implemented on reconfigurable hardware architectures, such as Field Programmable Gate Arrays (FPGAs), have proved to be efficient in increasing algorithms performance, while keeping a relative low power profile. The main drawback with FPGA architectures relies in their programmability, that considerably limits their adoption in High Performance Computing (HPC). Many efforts have been made in easing the adoption of FPGAs, with both commercial and academic solutions that ease the programming of the device, but as of today they still require FPGA-programming skills to obtain highly peformant architectures.

The main objective of this thesis is the development of Hugenomic, a framework that exploits reconfigurable hardware architectures (FPGAs), to speedup genomic data analysis. Hugenomic is designed to be easy-to-use, providing genomic researchers with little knowledge of FPGA programming, a library of the most used hardware accelerated algorithms in sequence alignment and bioinformatics, and a platform to accelerate custom codes on FPGA with no knowledge of the underneath hardware architecture. The algorithms available in the Hugenomic framework have been developed thanks to a methodology to systematically accelerate algorithms on FPGA based on an adaptation of the Berkeley Roofline Model that accounts for the non-fixed architecture of FPGAs.

PhD Committee
Ing. **Francesco Trovo'**, DEIB
Prof. **Jose L. Ayala ,** Complutense University of Madrid
Prof. **Dionisios N. Pnevmatikatos**, National Technical University of Athens