PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Mirko Gagliardi

XXXI Cycle

Training and Research Activities Report – Third Year

Tutor: Alessandro Cilardo

Information

- Mirko Gagliardi, master's degree in Computer Engineering in 2015 from the University of Naples Federico II.
- XXXI Cycle – ITEE
- Doctoral Fellowship funded by CeRICT
- Tutor: Alessandro Cilardo
- Collaboration with CeRICT in the context of European project MANGO. "MANGO: exploring Manycore Architectures for Next-GeneratiOn HPC systems" is a large-scale European project exploring heterogeneous manycore architectures for HPC, and innovative architectural mechanisms for High Performance, Power-efficiency and time-Predictability tomorrow HPC systems.

Study and Training activities

- MS Modules:
  i. *Advanced Computer Architecture and GPU Programming*, Alessandro Cilardo, 07/01/2016, 6 CFU;
  ii. *Intelligenza Artificiale*, Flora Amato, 26/09/2016, 6 CFU;
- Ad Hoc Modules:
  i. *Communicating and Disseminating your Research Work*, Mo Mansouri, 16/03/2016, 3 CFU;
  ii. *Scientific Writing*, Paolo Russo, 24/05/2016, 5 CFU;
  iii. *Testing Automation,* Porfirio Tramontana, 12/01/2017, 3 CFU;
  iv. *Le imprese e la ricerca, Sforza Antonio,* 28/02/2016, 4 CFU;
  v. *Interoperability, Semantic Technologies and Applications,* Flora Amato, 24/03/2017, 2 CFU;
  vi. *Compiler and code optimization, Edoardo Fusella, 17/05/2018, 3 CFU;*
- Seminars
  i. *Networks-On-Chip: Introduction And Advanced Topics*, Josè Flitch, 16/11/2015 – 18/11/2015, 1.8 CFU (1 extra credit for supplementary activities);
  ii. *Adversarial Testing of Protocol Implementations,* Prof. Cristina Nita Rotaru, 23/09/16, 0.4 CFU;
  iii. *Programmable Network conjugation*, Roberto Bifulco, 26/02/2016, 0.4 CFU;
  iv. *Speech Technologies At Trinity College*, Loredana Cerrato, 02/03/2016, 0.2 CFU;
  v. *Challenging Real-Time Measurement Systems For Immersive Life-Size Augmented Environment*, Giovanni Caturano, 29/04/2016, 0.5 CFU;
  vi. *Methodologies for embedded software validation*, Diego Tornese, 25/05/2016, 0.2 CFU;
  vii. *Internet: la dimensione immateriale dell'esistenza*, Stefano Quintarelli, 19/05/2016, 0.4 CFU;

viii. *DDoS Detection in Cloud and Campus Networks*, Jill Jermin, 06/06/2016, 0.2 CFU;
ix. *An Overview on Image Forensics with emphasis on physics-based scene verification*, Christian Riess, 18/05/2016, 0.2 CFU;
x. *Half Day EMC Design and troubleshooting Course*, Arturo Mediano, 29/09/2016, 0.8 CFU;
xi. Modelling Critical Infrastructures (Cis), Peter Popov, 3-11-2016, 0.4 CFU;
xii. Security Operations in una Telco, esperienze e riflessioni dal campo, Pietro Fabio Zamparelli, 11-11-2016, 0.4 CFU;
xiii. Minix 3: A reliable and Secure Operating System, A. Tanembaum, 30-11-2016, 0.4 CFU;
xiv. L'AIS ed il SIstema Nazionale per il monitoraggio marittimo, Massimo Marazzo, 16-12-2016, 0.4 CFU;
xv. Video Coding / Transcoding in Heterogeneous HPC, Mario Kovac, 7-12-2016, 0.4 CFU;
xvi. IBM Cognitive Computing: Challenges and Opportunities in Building an Artificial Intelligence Platform for Business, Pietro Leo, 17-02-2017, 0.4 CFU;
xvii. Cognitive Computing and da Vinci robot: Research Proposal and Discussion, Paolo Maresca, 17-02-2017, 0.2 CFU;
xviii. Smart Nanodevices for Theranostics, Ilaria Rea, 24-02-2017, 0.3 CFU;
xix. Fuzzy Logic, Genetic Algorithms and their application to Next Generation Networks, Leonard Bapolli, 14-03-2017, 0.8 CFU;
xx. From Mathematical formalization to Artificial Visual-Attention: Toward a Human-Like Robot Vision, Kurosh Madani, 04-04-17, 0.4 CFU;
xxi. Living bots and alter ego, Tamburrini Guglielmo, 04-04-2017, 0.4 CFU;
xxii. Dataflow SuperComputing for BigData, Flora Amato, 12-04-2017, 0.6 CFU;

	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	20	0	6	3	5	0	6	20	10	0	7	2	0	0	0	9	0	0	0	0	3	0	0	3	32	30-70
Seminars	5	1.8	0.8	0.8	1.2	0	0.8	5.4	5	2	0.9	2.2	0	0	0	5.1	0	0	0	0	0	0	0	0	11	10-30
Research	35	7	4	6	5	8	5	35	45	6	4	6	10	9	10	45	60	10	10	10	10	10	10	60	140	80-140
	60	8.8	10.8	9.8	11.2	8.0	11.8	60.4	60	8	12	10	10	9	10	59	60	10	10	10	13	10	10	63	183	180

- ## Research Activities
  - Title: *A Reconfigurable and Extensible Exploration Platform for Future Heterogeneous Systems*

- Study: Today many sectors, such as digital signal processing, scientific computing, computer graphics, and other application areas, have evolved to the point where their functionality requires performance levels that are not attainable on traditional CPU-based systems. Advancements of processors are largely driven by Moore's Law, which predicts that the number of transistors per silicon area doubles every 18 months. While Moore's Law is expected to continue for a few years, computer architects are moving along a fundamental shift in how the large amounts of available transistors are used to increase performance. Historically, performance improvements of microprocessors came from both increasing the frequency at which the processors run, and by increasing the amount of work performed in each cycle. This increasing need for resource- and power-efficient computing has stimulated the emergence of compute platforms with moderate or high levels of parallelism, like GPU, SIMD, and manycore processors in a variety of application domains. Furthermore, the ultimate technologies now allow designers to place CPU, GPU and DSP elements onto a single System on Chip (SoC). This allows smaller devices, reduces cost and saves power, moreover on-chip communication uses far less energy than off-chip connections.

- Research description: These factors have led to the use of heterogeneous computing, with specialized accelerators that complement general purpose CPU, and act as coprocessors for parallel workloads, to provide both power and performance benefits. Accelerator-based --or heterogeneous-- computing has become increasingly important in a variety of scenarios, ranging from High-Performance Computing (HPC) to embedded systems. In particular, modern many-core systems are based on a considerable number of lightweight processor cores typically connected through a Network-on-Chip (NoC), providing a scalable approach to the interconnection of parallel on-chip systems. To maximize resource and power efficiency, accelerator architectures tend to rely on parallelism to improve performance, with multi/many-core accelerators being today commonplace.

The main issue of heterogeneous computing model is that a programmers can only choose proprietary parallel programming languages. Existing programming toolkits have either been limited to a single product family, limiting the application (and the developers skills) on a specific vendor platform. Running the application on another system means to rewrite it. These limitations make it difficult for a developer to achieve the full compute power of heterogeneous computing model and a parallel code developer must well know the accelerator structure on which their application will be running. To gain the full benefits of this integration, separate components need shared access to the data they process. Getting these diverse accelerators to work as a team is no easy task. We need a unique model that presents these features in a manner comprehensible to mainstream software developers, supported by their already existing development environments and by future and current hardware accelerators. In most heterogenous computing platforms the memory on the accelerator is completely separate from host memory, hence parallel software developers are forced to manage main memory within their own application programs, and all data movement between host memory and device memory must be managed by the programmer through platform specific function and libraries that explicitly move data between the separate memories. Therefore, in an accelerator-targeted region, the programmers must orchestrate the execution by allocating memory on the accelerator device, initiating data transfer, sending the code to the accelerator, passing arguments to the compute region, queuing the device code, waiting for completion, transferring results back to the host, and deallocating memory. The concept of separate host and accelerator memories is very apparent in low-level accelerator programming languages such as CUDA or OpenCL, in which routine calls for data movement between the memories can dominate user code and a programmer cannot be unaware of memory structure. Knowing the specific accelerator structure has become an important requirement for the programmers. An ideal heterogenous computing model should allow programmers to create applications capable of using accelerators, with data movement between the host and accelerator implicit and managed by the compiler, without explicit accelerator startup and shutdown. An application designed for this model should be compatible for a wide assortment of data-parallel and task-parallel architectures and not restricted to a specific vendor platform (e.g. CUDA). At last, a heterogenous computing model should take into account the need to reduce system power consumption, even for devices that get their power from the wall. Such motivations gave birth to the **Horizon 2020 MANGO** project, which aims at exploring deeply heterogeneous accelerators for use in high-performance computing systems running multiple applications with

different Quality of Service (QoS) levels. The main goal of the project is to exploit customization to adapt computing resources to reach the desired QoS. For this purpose, it explores different but interrelated mechanisms across the architecture and system software. Along its path, the project involves different, and deeply interrelated, mechanisms at various architectural levels, from the heterogeneous computing cores, up to the memory architecture, the interconnect, the run-time resource management, power monitoring and cooling, also evaluating the implications on programming models and compilation techniques. Modern high-performance computing applications present a gap between the applications demand and the underlying architecture. Enabling a deeper customization of architectures to applications will eventually lead to computation efficiency, since it allows the computing platform to approximate the ideal system, featuring a fine-grained adaptation, or customization, used to tailor and/or reserve computing resources only driven by the application requirements. Finally, in such scenarios, coherent shared memory could be an important facility acting as a key enabler for programmer-friendly models exposed to the software as well as for the effective adaptation of existing parallel applications. However, unlike general-purpose architectures, hardware-managed coherence poses a major challenge for accelerators, due to the cost of the coherence infrastructure as well as the possible limitations in terms of scalability and performance. Full implementation of standard coherence protocols can induce significant overheads even when there is essentially no data sharing, e.g. when handling a nonshared block eviction. In fact, in many workloads a significant fraction of blocks are private to a single processing unit requiring in principle no coherence maintenance. While such problems have been widely investigated in the area of conventional homogeneous architectures, existing solutions do not always fit heterogeneity, moreover many-core accelerator-based systems pose special requirements and constraints, requiring further exploration of both hardware and software techniques.

My research activity investigates the architectural requirements of emerging HPC applications and the most suitable power efficient and high-performance solutions. The main goal is to define and pose a methodology for exploring novel solutions targeting heterogeneous systems. Future many-cores require exploration over new infrastructures typical of this novel paradigm (such as NoCs and sparse directories), which involves both hardware and software mechanisms in order to exploit scalability and higher efficiency, with customization has key-enablers to achieve

such desirable features. In this context, our research group founded the **Nu+ platform**.

Nu+ platform is an open-source NoC-based platform compliant with modern heterogeneous system trends. The platform aims to be highly modular, deeply customizable, meant to be easily extendible on both hardware and software levels, essential features for architectural exploration. This full-system enables us to better understand application-specific requirements through hardware customization, and also to evaluate the proposed solutions on a real system, running significant kernels extracted from typical workloads.

- Collaborations: My PhD work is part of a larger H2020 project named MANGO. The main MANGO objective is to define new-generation high-performance, power-efficient, deeply heterogeneous architectures.

Products

- Publications:
  i. Cilardo Alessandro, Gagliardi Mirko. "*Customizable Heterogeneous Acceleration for Tomorrow's High-Performance Computing.*" High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conference on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on. IEEE, 2015.
  ii. Cilardo Alessandro, Gagliardi Mirko, Donnarumma Ciro, "*A Configurable Shared Scratchpad Memory for GPU-like Processors*." International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Springer International Publishing, 2016.
  iii. Cilardo, Alessandro, Mirko Gagliardi, and Daniele Passaretti. "NoC-Based Thread Synchronization in a Custom Manycore System." International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Springer, Cham, 2017.
  iv. Gagliardi, Mirko, Edoardo Fusella, and Alessandro Cilardo. "Improving Deep Learning with a customizable GPU-like FPGA-based accelerator." 2018 14th Conference on Ph. D. Research in Microelectronics and Electronics (PRIME). IEEE, 2018.
  v. Zoni, Davide, Luca Cremona, Alessandro Cilardo, Mirko Gagliardi, and William Fornaciari. "PowerTap: All-digital power meter modeling for run-time power monitoring." Microprocessors and Microsystems 63 (2018): 128-139.

vi. Cilardo, Alessandro, Mirko Gagliardi, and Daniele Passaretti. “Hardware support for thread synchronization in an experimental manycore system”. International Journal of Grid and Utility Computing. Inderscience Publishers, 2018.
vii. Mirko Gagliardi, Alessandro Cilardo, Vincenzo Scotti. “Lightweight hardware support for selective coherence in heterogeneous manycore accelerators” DATE 2019 conference, ongoing peer review.

Conferences and Seminars

- Conferences:
  i. 14th Conference on PhD Research in Microelectronics and Electronics (PRIME) 2018, 2nd – 5th of July 2018, Prague, Czech Republic

Activity abroad

- Details: Internship at Arm Ltd. In the RD group, as Research Engineer Intern.
- Dates: from 05/06/2017 to 17/11/2017
- Activity description: Develop and maintain models and/or prototypes in support of Arm-based research enablement and in close collaboration with internal stakeholders. Main topic of interest: SoC design and FPGA Prototyping, Simulation and Modelling and IoT and Cloud Computing.
- Contribution: whitepaper “Cortex-M-based SoC Design and Prototyping using Arm DesignStart”.
- Contact: Ashkan Tousimojarad, Ashkan.Tousimojarad@arm.com