Angelo Garofalo, P.h.D.
ARCES
Viale Carlo Pepoli 3/2
40122, Bologna (BO), Italy
e-mail: **angelo.garofalo@unibo.it**

Curriculum Vitae

Present Position

- *Junior Assistant Professor ("Ricercatore a Tempo Determinato di Tipo A (RTD-A)")* with the Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi: (DEI) of the University of Bologna (February 2023).
- *Post-DoC Researcher* at the Integrated System Laboratory (IIS Lab) of the Eidgenössische Technische Hochschule Zürich (ETHZ) (October 2022);

Experience

- *Visiting PhD Researcher* at the "In-Memory Computing" research group of IBM Research Zurich, led by Dr. Abu Sebastian (November 2020 – April 2021);
- *PhD Student* at the Energy Efficient Embedded Systems Laboratory (EEES Lab), Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI) of the University of Bologna (2018 – 2022).

Education

- *Ph.D.* in Electronics, Telecommunications, and Information Technologies Engineering, University of Bologna, 2022;
- *M.Sc.* in Electronics Engineering, University of Bologna, 2018 (grade: 110/110 with honors);
- *B.Sc.* in Electronics Engineering, University of Bologna, 2016 (grade: 110/110 with honors).

Scientific Profile

Research Activities

My research activity starts early in my career with a bachelor's thesis in the field of Structural Health Monitoring (SHM). The research topic was the design of signal processing algorithms in the time-frequency (or time-scale) domain to detect and localize impacts on aluminum and composite plates, targeting avionic applications. This work also continued during part of my master's studies, leading to the publication of two conference papers.

As a graduate researcher, I started my PhD at the Energy-Efficient Embedded System Laboratory (EEES Lab) of the University of Bologna, in 2018. My research focused on flexible computing systems for AI acceleration at the extreme edge of the IoT. On this topic, I made significant contributions in the field of

hardware-efficient machine learning, energy-efficient ultra-low-power embedded processors, and heterogeneous in-memory computing architectures.

I explored several multi-disciplinary approaches. On the one hand, I worked on hardware-software co-design of ultra-low-power multi-core processors based on the open-source RISC-V Instruction Set Architecture (ISA). In this context, I researched novel techniques to boost the efficiency of artificial intelligence (AI) kernels executed on this class of devices. Leveraging the concept of Quantized Neural Networks (QNNs), I designed optimized low-level-software routines and sets of domain-specific extensions to the RISC-V Instruction Set Architecture (ISA) for low-bitwidth integer computation.

On the other hand, I researched new architectural solutions to boost AI workloads by exploiting computing system heterogeneity. On this aspect, I focused on integrating diversified compute units within a tightly coupled system and on reconfigurable SIMD/MIMD programmable RISC-V architectures. Furthermore, I studied the emerging analog in-memory computing paradigm, which promises outstanding acceleration and energy-efficiency on Matrix-Vector-Multiplications (MVMs), the compute building block of AI workloads. To deepen my knowledge on this topic, I spent six months as visiting PhD researcher with the In-Memory Computing group at the IBM Research Lab, leading the research on Phase-Change-Memory (PCM) based in-memory computing acceleration. During this period, I contributed to bring-up the world's first in-memory computing chip based on phase-change memory (PCM) technology to map a DNN model on such an SoC.

My interest for this subject has constantly grown, and I focused on researching a new generation of highly-efficient heterogeneous processing systems that aim to tight integrate general-purpose and specialized digital processors with emerging analog in-memory computing accelerators. Currently, I am exploring the scalability of heterogeneous in-memory computing architectures to High-Performance Computing (HPC) systems, addressing challenges and opportunities from a system and communication perspective.

During the years of the Ph.D., I designed and taped-out two chips (as the principal architect and chip designer) and closely supervised the design and tape-out of one additional chip. Furthermore, I contributed to an SoC tape-out in collaboration with the company "Dolphin Integration", where I was the coordinator for the design of the SoC programmable multi-core accelerator for AI and DSP applications. This collaboration resulted in co-authoring a conference paper that will be presented at the following IEEE International Solid-State Circuits Conference (ISSCC) 2023.

At the end of my PhD, I decided to continue my academic research activity as *Post-Doc* to deepen my research on hardware-efficient machine learning, design of integrated digital systems, emerging heterogeneous in-memory computing architectures and high energy-efficient parallel computing systems, covering aspects from low-level-software to silicon implementation.

Overall, on the topics listed in this summary, I co-authored 6 journal papers and 11 conference proceedings in top-ranked journals and conferences of the sector. I strongly believe in collaborative and multi-disciplinary research. As a tasks coordinator or technical contributor, I am actively collaborating with several European partners from industry and academia in the context of European projects.

A) Teaching

Support Activities

During the last years I have thought in 1 course at the University of Bologna been didactic tutor for two bachelor courses, co-advisor of 6 master theses. More information about teaching support activities can be found below.

Tutor Activities

- 28651 - ELETTRONICA T-A (Tutor). University of Bologna, Academic Year 2021-2022. Aim of this course is to provide basic knowledge on the manufacturing processes and operation of electronic devices, as well as on the analysis of analog and digital circuits.
- 73731 - ARCHITETTURE E PROGRAMMAZIONE DEI SISTEMI ELETTRONICI T-A - Modulo 2 (29035 - LABORATORIO DI ARCHITETTURE E PROGRAMMAZIONE DEI SISTEMI ELETTRONICI INDUSTRIALI T-A- Modulo 2) (Tutor). University of Bologna. Academic year 2018-2019. Aim of this course is to teach the architecture of micro - controller based systems using ARM cortex M cores, and firmware programming for industrial applications. It covers both theoretical and practical aspects related to architecture and programming of ARM Cortex M microcontrollers.

Co-advisor of Master Thesis

- Gianmarco Ottavi, Thesis Title: “Sviluppo e Ottimizzazione di un Processore Configurabile con Unità di Calcolo a Precisione Variabile”, 19/12/2019;
- Nazareno Bruschi, Thesis Title: “Accelerating Mixed-Precision Quantized Neural Networks on Parallel Ultra Low Power - IoT End Nodes”, 11/03/2020;
- Ilario Coppola, Thesis Title: “A Reconfigurable MIMD/SIMD RISC-V Cluster for Energy-Efficient Parallel Computing”, 11/03/2020;
- Mattia Sinigaglia, Thesis Title: “Progettazione ed implementazione di un sistema on chip per applicazioni audio”, 21/07/2021;
- Alessandro Nadalini, Thesis Title: "Progettazione ed ottimizzazione di un processore dedicato per accelerazione di reti neurali quantizzate a precisione mista", 07/10/2021;
- Maicol Ciani, Thesis Title: “System-Level Integration of a Security Enclave on a RISC-V based Embedded System On Chip”, 05/10/2022.

B) Research Activity

Impact of Publications

In the last years I have published (international, peer-reviewed) **6 journal papers and 11 conference papers**. My **h-index is 9, i-10 index is 7** (from Google Scholar, November 2nd, 2022). The total number of citations is **278** and the number of citations per year is growing year-by-year: **7** in 2019, **41** in 2020, **115** in 2021, **109** in 2022 (from Google Scholar, November 2nd, 2022).

Scopus Page: **https://www.scopus.com/authid/detail.uri?authorId=57192061020#**
Google Scholar Page: **https://scholar.google.com/citations?user=K-dssAoAAAAJ&hl=it&oi=ao**

Participation to International Conferences and Workshops as Speaker

- A. Garofalo, M. Perotti, L. Valente, Y. Tortorella, A. Nadalini, L. Benini, D. Rossi, F. Conti, *DARKSIDE: 2.6GFLOPS, 8.7mW Heterogeneous RISC-V Cluster for Extreme-Edge On-Chip DNN Inference and Training* at the IEEE 48th European Solid-State Circuits Conference, September 2022;
- A. Garofalo, G. Ottavi, A. Di Mauro, F. Conti, G. Tagliavini, L. Benini, D. Rossi, *A 1.15 TOPS/W, 16-Cores Parallel Ultra-Low Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode* at the IEEE 47th European Solid State Circuits Conference (ESSCIRC), September 2021;
- A. Garofalo, G. Tagliavini, F. Conti, D. Rossi and L. Benini, *Xpulpnn: Enabling energy efficient and flexible inference of quantized neural networks on risc-v based iot end nodes* at the 28th IEEE International Symposium on Computer Arithmetic, June 2021;
- A. Garofalo, G. Tagliavini, F. Conti, D. Rossi and L. Benini, *XpulpNN: Accelerating quantized neural networks on RISC-V processors through ISA extensions,* at the Design, Automation & Test in Europe Conference & Exhibition (DATE), March 2020;
- A. Garofalo, M. Rusci, F. Conti, D. Rossi and L. Benini, *Pulp-nn: A computing library for quantized neural network inference at the edge on risc-v based parallel ultra-low power clusters*, at the 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), November 2019.
- A. Garofalo, M. Rusci, F. Conti, D. Rossi and L. Benini , *PULP-NN: Open-Source Library for QNNs Inference on RISC-V Based PULP Cluster,* at the International RISC-V Workshop, Zurich, June 2019.

Participation to International Workshops and Forums as Invited Speaker

- A. Garofalo, *Enabling End-to-End QNN Execution on PULP,* at the OpenHW TV (season 3 episode 2), February 2022. (Invited Talk);
- A. Garofalo, Is an AI Accelerator All You Need? Overcoming Amdahl's Law With Tightly-Coupled Specialized Accelerators, in Forum 6 of the 2023 *IEEE International Solid-State Circuits Conference 2023* (ISSCC 2023).

Activities in National and European Research Projects

- *Task Leader (September 2022 -).* Task 3.2 - Design of AI specialized RISC-V based local digital processing unit for the IMNPU – of NeuroSoC (A multiprocessor system on chip with in-memory neural processing unit, HORIZON-RIA, Overall founding: € 7.952.677). The project aim is to develop an advanced Multi-Processor System on Chip prototype in FD-SOI 28nm CMOS technology that tightly integrates an AIMC IMNPU unit, a local digital processing subsystem, and functional safe

multiprocessor host subsystems based on an enhanced version of existing RISC-V microprocessor implementation, while covering IMNPU security aspects holistically to tackle the requirements of a wide set of edge-AI applications.

- *Task Leader* (*January 2023 –*). Task 2.3 – Embedded Mid-end processors – of TRISTAN (Together for RISC-V Technology and Applications, European KDT-JU programme). The project aim is to expand, mature and industrialize the European RISC-V ecosystem so that it is able to compete with existing commercial/proprietary alternatives.
- *Work Package Leader (January 2022 -).* WP4 – Architecture Design – of WiPLASH: "Architecting More Than Moore – Wireless Plasticity for Heterogeneous Massive Computer Architectures" (GA 863337, overall project founding: € 3 M). The project aims to pioneer an on-chip wireless communication plane able to provide architectural plasticity, reconfigurability and adaptation to the application requirements with near-ASIC efficiency but without loss of generality. In this project I am responsible for the heterogeneous architecture design in WP4.

Other Collaborations with International Research Centers and Companies

- *Dolphin Integration* (2019 – 2022): AI capable edge processor. This project, in collaboration with ETH Zurich and *Dolphin Integration* company, aimed at developing a low-power digital signal processor for embedded video and audio applications featuring capabilities of running embedded machine learning and artificial intelligence workloads. I was the main responsible for the architecture design of the programmable multi-core accelerator of the system on chip (SoC). The SoC has been taped-out in GF22FDX technology node in July 2021 and has been tested throughout year 2022. This collaboration resulted into the co-authorship of a conference paper that will be presented at the next *IEEE International Solid-State Circuits Conference 2023 (ISSCC 2023)*.

Professional Services

- Member of the technical program committee for the international conference DATE 2023;
- Member of the technical program committee for the international conference EWC 2023;
- Regular reviewer for several international journals, including IEEE TCAS-I, IEEE TCAD, IEEE Access.

List of Publications

Journal Papers

1. Garofalo, A., Tortorella, Y., Perotti, M., Valente, L., Nadalini, A., Benini, L., ... & Conti, F. (2022). Darkside: A Heterogeneous RISC-V Compute Cluster for Extreme-Edge On-Chip DNN Inference and Training. *IEEE Open Journal of the Solid-State Circuits Society*;
2. Garofalo, A., Ottavi, G., Conti, F., Karunaratne, G., Boybat, I., Benini, L., & Rossi, D. (2022). A Heterogeneous In-Memory Computing Cluster For Flexible End-to-End Inference of Real-World Deep Neural Networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*;

3. Montagna, F., Mach, S., Benatti, S., Garofalo, A., Ottavi, G., Benini, L., ... & Tagliavini, G. (2021). A Low-Power Transprecision Floating-Point Cluster for Efficient Near-Sensor Data Analytics. *IEEE Transactions on Parallel and Distributed Systems*, *33*(5), 1038-1053;
4. Burrello, A., Garofalo, A., Bruschi, N., Tagliavini, G., Rossi, D., & Conti, F. (2021). Dory: Automatic end-to-end deployment of real-world dnns on low-cost iot mcus. *IEEE Transactions on Computers*, *70*(8), 1253-1268;
5. Garofalo, A., Tagliavini, G., Conti, F., Benini, L., & Rossi, D. (2021). Xpulpnn: Enabling energy efficient and flexible inference of quantized neural networks on risc-v based iot end nodes. *IEEE Transactions on Emerging Topics in Computing*, *9*(3), 1489-1505.;
6. Garofalo, A., Rusci, M., Conti, F., Rossi, D., & Benini, L. (2020). PULP-NN: accelerating quantized neural networks on parallel ultra-low-power RISC-V processors. *Philosophical Transactions of the Royal Society A*, *378*(2164), 20190155.

Conference Proceedings

1. Garofalo, A., Perotti, M., Valente, L., Tortorella, Y., Nadalini, A., Benini, L., ... & Conti, F. (2022, September). Darkside: 2.6 GFLOPS, 8.7 mW Heterogeneous RISC-V Cluster for Extreme-Edge On-Chip DNN Inference and Training. In *ESSCIRC 2022-IEEE 48th European Solid State Circuits Conference (ESSCIRC)* (pp. 273-276). IEEE;
2. Garofalo, A., Ottavi, G., Di Mauro, A., Conti, F., Tagliavini, G., Benini, L., & Rossi, D. (2021, September). A 1.15 TOPS/W, 16-Cores Parallel Ultra-Low Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode. In *ESSCIRC 2021-IEEE 47th European Solid State Circuits Conference (ESSCIRC)* (pp. 267-270). IEEE;
3. Montagna, F., Tagliavini, G., Rossi, D., Garofalo, A., & Benini, L. (2021, June). Streamlining the OpenMP Programming Model on Ultra-Low-Power Multi-core MCUs. In *International Conference on Architecture of Computing Systems* (pp. 167-182). Springer, Cham;
4. Ottavi, G., Garofalo, A., Tagliavini, G., Conti, F., Benini, L., & Rossi, D. (2020, July). A mixed-precision RISC-V processor for extreme-edge DNN inference. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (pp. 512-517). IEEE.;
5. Bruschi, N., Garofalo, A., Conti, F., Tagliavini, G., & Rossi, D. (2020, May). Enabling mixed-precision quantized neural networks in extreme-edge devices. In *Proceedings of the 17th ACM International Conference on Computing Frontiers* (pp. 217-220);
6. Garofalo, A., Tagliavini, G., Conti, F., Rossi, D., & Benini, L. (2020, March). XpulpNN: Accelerating quantized neural networks on RISC-V processors through ISA extensions. In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 186-191). IEEE;
7. Garofalo, A., Rusci, M., Conti, F., Rossi, D., & Benini, L. (2019, November). Pulp-nn: A computing library for quantized neural network inference at the edge on risc-v based parallel ultra low power clusters. In *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)* (pp. 33-36). IEEE.;

8. Burrello, A., Conti, F., Garofalo, A., Rossi, D., & Benini, L. (2019, October). Work-in-progress: DORY: Lightweight memory hierarchy management for deep NN inference on IoT endnodes. In *2019 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ ISSS)* (pp. 1-2). IEEE;
9. Ruospo, A., Cantoro, R., Sanchez, E., Schiavone, P. D., Garofalo, A., & Benini, L. (2019, October). On-line testing for autonomous systems driven by RISC-V processor design verification. In *2019 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)* (pp. 1-6). IEEE;
10. Garofalo, A., Testoni, N., Marzani, A., & De Marchi, L. (2017, July). Multiresolution wavelet analysis to estimate Lamb waves direction of arrival in passive monitoring techniques. In *2017 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS)* (pp. 1-6). IEEE;
11. Garofalo, A., Testoni, N., Marzani, A., & De Marchi, L. (2016, September). Wavelet-based Lamb waves direction of arrival estimation in passive monitoring techniques. In *2016 IEEE International Ultrasonics Symposium (IUS)* (pp. 1-4). IEEE.

Pre-Print Papers

1. Ottavi, G., Garofalo, A., Tagliavini, G., Conti, F., Di Mauro, A., Benini, L., & Rossi, D. (2022). Dustin: A 16-Cores Parallel Ultra-Low-Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode. *arXiv preprint arXiv:2201.08656*;
2. Montagna, F., Mach, S., Benatti, S., Garofalo, A., Ottavi, G., Benini, L., ... & Tagliavini, G. (2020). A transprecision floating-point cluster for efficient near-sensor data analytics. *arXiv preprint arXiv:2008.12243.*