**PhD in Information Technology and Electrical Engineering**
**Università degli Studi di Napoli Federico II**

# PhD Student: Manuel Maddaluno

**Cycle: XXXIX**

## Training and Research Activities Report

## Academic year: 2024-25 - Year: Second

**Tutor: Prof. Alessandro Cilardo**

**Date: December 11, 2025**

ALESSANDRO
CILARDO
UNIVERSITÀ DEGLI
STUDI DI NAPOLI
FEDERICO II
11.12.2025 16:08:49
GMT+01:00

## 1. Information:

- **PhD student: Manuel Maddaluno**
- **DR number: DR997203**
- **Date of birth: 05/07/1999**
- **Master Science degree: Computer Engineering**
- **University: Università degli Studi di Napoli Federico II**
- **Doctoral Cycle: XXXIX**
- **Scholarship type: PNRR – Centro Nazionale CN1 - National Centre for High-Performance Computing, Big Data and Quantum Computing – Spoke: Future HPC**
- **Tutor: Prof. Alessandro Cilardo**

## 2. Study and training activities:

| Activity | Type[1] | Hours | Credits | Dates | Organizer | Certificate[2] |
|---|---|---|---|---|---|---|
| Solid State Transformers: Fundamentals, Insights and New Trends | Seminar | 2 | 0.4 | 20/12/2024 | Ing. Luigi Pio Di Noia | Y |
| Can we Rely on AI? Reliability Issues in Artificial Neural Networks and Potential Solutions for Autonomous Vehicles | Seminar | 1 | 0.2 | 16/01/2025 | Dr. Edoardo Giusto | Y |
| The Good, the Bad, and the Ugly in Quantum Computing: Computational Power, Intrinsic Noise, and Transient Faults | Seminar | 1 | 0.2 | 17/01/2025 | Dr. Edoardo Giusto | Y |
| Dynamic Risk Assessment: Industrial Applications Leveraging Bayesian Inference for Enhanced Decision Making | Seminar | 1 | 0.2 | 04/03/2025 | Prof. Simone Guarino | Y |
| 5G & Digital Transformation: A View From An Unconventional Perspective | Seminar | 4 | 0.8 | 14/03/2025 | Prof. Antonia Maria Tulino | Y |
| PhD Survival Strategies | Seminar | 1.5 | 0.3 | 30/05/2025 | Dr. Pietro Liguori | Y |
| Sovranità digitale cos'è e quali sono le principali | Seminar | 1.5 | 0.3 | 23/06/2025 | Prof. Marcello | Y |

| | | | | | | |
|---|---|---|---|---|---|---|
| minacce al cyberspazio nazionale | | | | | Cinque | |
| Superconducting Radio Frequency Cavities for Quantum Computing and Communication | Seminar | 1 | 0.2 | 24/06/2025 | Prof. Edoardo Giusto | Y |
| Trusted Execution Environments for QPUs | Seminar | 1 | 0.2 | 27/06/2025 | Prof. Edoardo Giusto | Y |
| Guardians Of Threats? AI at the Frontlines of Cybersecurity | Seminar | 4 | 0.8 | 17/10/2025 | Prof. Antonia Maria Tulino | Y |
| How to boost your PhD | Course | 20 | 5 | Jan. 08, 15, 22, 29 2025 | Prof. Antigone Marino | Y |
| Innovation and Entrepreneurship | Course | 12 | 3 | June 05, 12, 19, 26 2025 | Prof. Pierluigi Rippa | Y |
| High-Performance and Quantum Computing | Course | 48 | 6 | Sep. to Dec. 2025 | Prof. Alessandro Cilardo | Y |

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

### 2.1. Study and training activities - credits earned

| | Courses | Seminars | Research | Tutorship | Total |
|---|---|---|---|---|---|
| Bimonth 1 | 5 | 0.8 | 10 | 0 | 15.8 |
| Bimonth 2 | 0 | 1 | 9 | 0 | 10 |
| Bimonth 3 | 0 | 1 | 9 | 0 | 10 |
| Bimonth 4 | 3 | 0 | 7 | 0 | 10 |
| Bimonth 5 | 0 | 0.8 | 9.2 | 0 | 10 |
| Bimonth 6 | 6 | 0 | 4 | 0 | 10 |
| **Total** | 14 (32 Total) | 3.6 (10.2 Total) | 48.2 (86.2 Total) | 0 | 65.8 (128.4 Total) |
| **Expected** | **10 - 20 (30–70 Total)** | **5 - 10 (10-30 Total)** | **30 - 45 (80-140 Total)** | **0 - 1.6 (0 – 4.8 Total)** | |

## 3. Research activity:

**Research topic:**

The second year of my PhD further deepened my investigation into High-Performance Computing (HPC) architectures, with a continued focus on latency reduction and latency control. Building on the foundations of the first year - where I developed predictable HBM controllers and low-latency SmartNIC architectures - my research evolved toward a broader and more integrated theme: workload-driven architectural optimization combined with reconfigurable hardware.

This evolution stemmed from both the practical experience accumulated and the increasing need for architectures that can adapt to the diverse requirements of modern HPC applications, including High-Frequency Trading (HFT), Distributed Learning, Distributed Algorithms, and Quantum Error Correction (QEC).

A central piece of the second-year work has been the characterization of workloads across several domains such as financial computing (HFT), large-scale distributed machine learning, consensus and coordination protocols in distributed systems, and timing-critical quantum control loops.

Across such domains, modern workloads exhibit highly diverse behaviors in terms of memory access patterns, communication intensity, and latency constraints. Yet the design of today's accelerators is typically static, assuming fixed usage conditions.

The central idea driving my research is the opposite:

Hardware should not be static - hardware should adapt to the workload.

This requires understanding the workload first, and designing flexibility directly into the architecture.

This philosophy led me to formalize a four-stage methodology that I applied consistently across all the architectural blocks I worked on.

**Methodology**:

1. Understanding the workload - Workload Characterization

The process begins with characterizing the workload independently of any specific architecture.

In particular, the workload can be:

- Compute intensive
- Memory intensive
- Communication intensive

Or in a combination of the three.

These analyses were performed using traces, mathematical modeling, and full-system simulations.

The important aspect is that the workload, not the hardware, is the starting point.

This approach applies equally well to:

- model updates in distributed learning,
- transaction bursts in HFT,
- synchronization primitives in distributed algorithms,
- syndrome extraction loops in quantum error correction.

All these domains share the same objective: understanding how the workload behaves under pressure.

2. Architecture Modeling and Flexibility Points

Once the workload is characterized, the second step is to define an architectural model that exposes flexibility points, i.e., parameters or components that can be configured or swapped depending on the workload class.

Examples include:

- configurable scheduling policies in the HBM controller,
- tunable arbitration mechanisms,
- parameterized pipeline stages in the SmartNIC architecture,
- modular accelerators in a System-on-Chip (SoC).

The novelty in the second year is that flexibility is not treated as an afterthought.

It is built into the model from the beginning, guided by what was learned during workload characterization.

Server-grade FPGA accelerators such as the AMD Alveo cards, enable this paradigm by giving the possibility of reconfiguring the hardware architecture.

3. Co-Simulation for Workload-Driven Design (CrossSim)

A major step forward this year has been the extensive use and enhancement of CrossSim, the co-simulation environment combining Gem5 and QuestaSim.

CrossSim serves as the core tool in my methodology because it enables:

- executing full software workloads in Gem5,
- capturing realistic traces and timing,
- injecting these into cycle-accurate RTL simulations in QuestaSim,
- closing the loop between workload behavior and hardware implementation.

Rather than testing hardware in isolation, CrossSim allows me to validate how a flexible HBM controller, a SmartNIC pipeline, or a reconfigurable SoC behaves in the context of a full system running a real workload.

4. Implementation, Deployment, and Iterative Adaptation

The final phase involves:

- implementing the architecture in RTL,
- synthesizing on FPGA,
- measuring actual latency and throughput,
- and iteratively refining the design.

What distinguishes the second-year approach is that adaptation is iterative and guided by workload feedback. If measurements under a specific workload reveal a bottleneck, the architectural flexibility points allow the design to be re-tuned without re-architecting the entire system.

The cycle "characterise → model → simulate → adapt" is the backbone of my current research.

**Results:**

Leveraging this methodology, together with the developed reconfigurable SoC platform Simply-V, I was able to rapidly design and deploy an ad-hoc, ultra-low-latency architecture tailored for the HFT domain. The structured workflow allowed the customization process to remain straightforward and systematic, despite the stringent latency and determinism requirements typical of HFT workloads. Experimental evaluations using realistic order-flow patterns show that the resulting architecture achieves sub-microsecond end-to-end latency, confirming the benefits of domain specialization applied through the proposed methodology. The flexibility of the platform further enabled controlled experimentation with memory and communication subsystems, which was essential for evaluating different design trade-offs.

In addition, the same methodology in combination with CrossSim provided the ideal environment to develop, integrate, and test the developed HBM controller. The platform allowed me to explore and tune the controller's internal mechanisms, most importantly the address-mapping policy, to match specific workload characteristics. By analyzing the access patterns generated by the target application, I was able to refine the controller configuration to maximize sustained throughput while preserving predictability, which is a critical requirement for latency-sensitive and real-time systems.

This iterative optimization process validated both the flexibility of the platform and the relevance of the proposed methodology for systematically guiding performance-driven architectural specialization.

## 4. Research products:

- *Alessandro Cilardo, **Manuel Maddaluno** - Exploring time predictability for High-Bandwidth Memory technology - IEEE Transactions on Emerging Topics in Computing (ACCEPTED)*

- *Vincenzo Maisto, Stefano Mercogliano, **Manuel Maddaluno**, Alessandro Cilardo (2025). Simply-V: A RISC-V Reconfigurable Playground Soft-SoC for Open Hardware Research and Fast Prototyping. WiPiEC Journal - Works in Progress in Embedded Computing Journal, 11(1), https://doi.org/10.64552/wipiec.v11i1.86 (PUBLISHED)*

- *Vincenzo Maisto, Stefano Mercogliano, **Manuel Maddaluno**, Alessandro Cilardo - The Simply-V Framework: an Extensible RISC-V Reconfigurable Soft-SoC forOpen Research and Fast Prototyping - ACM Transactions on Design Automation of Electronic Systems (SUBMITTED-GOT Major Revision)*

- *Alessandro Cilardo, **Manuel Maddaluno** - A Customizable Open-Source HBM Controller - ACM Transactions on Design Automation of Electronic Systems (SUBMITTED)*

## 5. Conferences and seminars attended

-

## 6. Activity abroad:

I will be joining Advanced Micro Devices (AMD) in Dublin as part of the 2026 Internship Program, from January 2026 to August 2026.

## 7. Activity in partner companies:

-

## 8. Tutorship:

-

## 9. Plan for year three:

Future developments will focus on extending the proposed methodology targeting new domains where specialization, predictability, and scalability are equally essential. In particular, two promising directions are:

- Distributed Machine Learning, where architectures must manage high-bandwidth, low-latency gradient exchanges across large GPU/accelerator clusters.

- Quantum Error Correction, where deterministic, ultra-fast classical control must interact continuously with quantum hardware under strict timing guarantees.

Both domains present unique computational patterns and communication constraints that align well with the capabilities of reconfigurable architectures and the design flow explored during this research.

A practical challenge encountered during the second year concerns access to realistic workloads and traces. While the ideal evaluation method would involve executing real applications at scale, this is often infeasible due to limited computational resources or the unavailability of representative datasets. To overcome this limitation, an important future research direction is the development of a workload-driven traffic generator.

This tool would operate by capturing the essential characteristics of selected applications, once they can be run at least in a reduced or emulated form, and then generating synthetic but statistically accurate workload traces that can be used to feed and evaluate new architectures.

Such a generator would enable systematic, reproducible, and domain-specific benchmarking under realistic scenarios, greatly enhancing the applicability and robustness of the overall methodology. Ultimately, this extension would allow the approach to scale to multiple research domains, enabling faster architectural exploration grounded in realistic operational constraints.