**Ph.D. in Information Technology**
**Thesis Defense**

**April 11th, 2025**
**at 11:00**
**Aula Beta – edificio 24**

**Roberto ROCCO** – XXXVII Cycle

## An Advanced Framework for Fault Resiliency in HPC Applications Focusing on Novel MPI Features and Energy Implications

Supervisor: Prof. Gianluca Palermo

**Abstract:**

High-Performance Computing (HPC) has evolved from large mainframes to GPU-accelerated clusters, driven by the need to overcome physical limitations such as the end of Moore's law and Dennard scaling. While this shift has enabled unprecedented computational power, as demonstrated by the ExaFLOPS performance of the Frontier cluster, it has also led to significantly higher energy consumption. In such a context, maximizing efficiency is crucial, yet modern HPC systems face challenges beyond raw performance. One of these is fault management: as system complexity increases, so does the likelihood of faults, making them a growing concern for large-scale applications.

Despite the presence of various fault management techniques, the Message Passing Interface (MPI), the de-facto standard for inter-process communication in HPC, still lacks built-in fault management. Existing solutions such as Checkpoint and Restart (C/R) mitigate the issue but introduce performance overhead and scalability concerns. Recent efforts have explored alternatives like User-Level Fault Mitigation (ULFM) and Reinit, which allow MPI applications to continue execution after faults occur. However, due to their complexity, these solutions are rarely integrated into real-world HPC workloads.

This thesis takes a different approach by focusing on fault resilience: it extends our previous work on the Legio framework, which combines ULFM with graceful degradation, allowing applications to recover from faults more efficiently than traditional C/R methods. This approach trades some result accuracy for significantly lower recovery time and energy consumption, making it particularly suitable for embarrassingly parallel applications.

Beyond proposing a practical solution, this thesis addresses key gaps in the literature. First, while most research focuses on execution time overhead, we explicitly consider the energy impact of fault management to quantify the amount of energy wasted properly. Second, we extend Legio applicability past embarrassingly parallel applications, dealing with the concept of critical processes. Third, we extend fault resilience mechanisms to newer MPI features like group collective communicator creation and the Session model, ensuring compatibility with the evolving standard. Finally, we analyze the validity of approximate results produced under fault conditions, assessing when recomputation can be avoided to optimize energy efficiency further.

By tackling these challenges, this thesis analyses the energy efficiency of MPI-based HPC workloads even in the presence of faults, bridging the gap between theoretical fault management techniques and their practical adoption in large-scale computing environments.

## PhD Committee

Prof. Cristina Silvano**, Politecnico di Milano**

Prof. Stefano Markidis**, KTH**

Prof. Andrea Bartolini**, Università degli Studi di Bologna**