Ph.D. in Information Technology
Thesis Defense

March 1st, 2024
at 15:00
Sala Seminari Nicola Schiavoni, Building 20

Nicola LEPRI – XXXVI Cycle

IN-MEMORY ANALOG ACCELERATION OF DEEP LEARNING INFERENCE

Supervisor: Prof. Daniele Ielmini

Abstract:

The impact that Deep Learning, as the most prominent subfield of AI, is having on our lives is already profound, and it is increasing week after week. However, with the end of Moore's law for transistor scaling and the rise of other physical bottlenecks, traditional computers based on von Neumann architecture seem unable to withstand the extraordinary pace of AI growth in the long term, risking a sudden brake for the exploding energy consumption.

In this framework, in-memory analog computing based on emerging memory technologies represents a potentially disruptive paradigm shift, that promises to satisfy the requirements of Deep Learning workloads, allowing to overcome the massive inefficiency of data transfer while further boosting the integration density and parallelism. Crosspoint arrays, which are a computing architecture based on a properly routed matrix of analog programmable memories, are key players in the expansion of the in-memory computing paradigm, thanks to their demonstrated integration density, low power consumption, and the ability to perform matrix operation with low latency and high throughput.

Clearly, several challenges lie in the path toward full technology maturity and mass diffusion. When considering workloads and computing architectures of relevant sizes, nonidealities such as parasitic resistances along wires (IR drop), device and process variations, peripheral circuitry overhead, and workload partitioning issues become significant and potential showstoppers.

In this scenario, this doctoral thesis mainly focuses on three directions of investigation, namely analysis and compensation of parasitic effects, exploration of memory devices, and circuital implementation. Concerning parasitic effects, a thorough study was carried out to understand, model, and assess various parasitic effects, enabling the development of compensation techniques involving architectural or algorithmic solutions. At device level, a subthreshold operated one-selector/one-resistor memory element was characterized, modeled, and validated for the application, showing outstanding resilience against parasitic IR drop. At circuital level, a PCM-based ASIC accelerator of Deep Learning inference was conceived, designed, and delivered for the tapeout, relying on an ad hoc computing architecture and embedding features that will help to further investigate circuital and device solutions. The research work tackled various challenges of the in-memory computing paradigm, exploring and addressing them from different standpoints, pivoting and exploiting the intrinsic multidisciplinarity of the field.

PhD Committee

Giacomo Langfelder, **Politecnico di Milano**
Eleonora Franchi Scarselli, **Università di Bologna**
Nini Pryds, **Technical University of Denmark**