Doctoral Dissertation
Doctoral Program in Electrical, Electronics and Communication Engineering
(31.st cycle)

Exploring New Computing Paradigms for Data-Intensive Applications

Giulia Santoro
* * * * * *

Supervisor
Prof. Mariagrazia Graziano

Since the conception of the first stored-program computer by John von Neumann in 1945, processing units have undergone extraordinary transformations. Over the years, computing systems have pervaded numerous aspects of the human life and are now ubiquitous. The continuous evolution of processing systems has been mainly driven by three factors: *technological progress*, *architectural innovation* and an always growing demand for *computational power*. The combined action of these factors has led to the creation of extremely powerful computing systems and to the emergence of some fascinating applications that could not have existed without the support of such computational power. However, such advances do not come for free. *Power consumption* is, nowadays, one of the major concerns. Complexity at the architectural level of processing units, technological scaling and resource demanding characteristics of modern applications all have a large impact on power consumption. Moreover, applications are not only *resource demanding*, but also *data demanding*, and this has a strong effect on the role that the memory plays on power consumption. *Memory accesses* are extremely costly in terms of energy and are a performance bottleneck. In fact, while CMOS technology keeps scaling, memories have not made progress at the same pace. This has created a performance gap between processing units and memories that is best known as *von Neumann bottleneck* or *memory wall*. Memories are not able to provide data at the same rate as processing units are able to compute them. In addition to all these problems, *technological scaling* is also approaching a limit where it would not be possible to further progress because of fundamental physical, technological and economical limitations. The introduction of novel *beyond-CMOS technologies* based on new information-processing paradigms is a potential solution to the limitations of technological scaling. This thesis addresses different aspects of these problems.

In the first part of this research work the need for computational power and energy efficiency is targeted through a specific and widespread application: Convolutional Neural Networks (CNNs). Being resource and data demanding, CNNs require energy efficient hardware acceleration. For this aim, a custom-designed hardware accelerator is proposed. The *Deep Learning Processor* combines the quality of design achieved with the ASIC implementation flow with the reconfigurability of FPGAs. The accelerator is an array of Processing Elements interconnected through a Network-on-Chip. The Processing Element is the basic block of the whole accelerator and it has been designed to be *flexible* but at the same time *optimized for CNN-like workload*, *performance oriented* and *low power*. The Deep Learning Processor has been used as an architectural template for conducting a design space exploration that takes into account all the key features of the accelerator and defines the best configurations in terms of energy efficiency and throughput.

In the second part of this thesis, the limitations related to the technological scaling and the von Neumann bottleneck are targeted through the exploration of a *non-von-Neumann computing paradigm* that is *Logic-in-Memory*. This novel approach goes beyond the separation between computation and memory, typical of von Neumann processing systems, trying to fully integrate them in a single unit. Data are computed directly inside the memory without the need to move them. This approach has a twofold advantage: tearing down memory accesses (and the related power consumption) and demolishing the memory wall. This research work investigates the concept of Logic-in-Memory by presenting a novel *Configurable Logic-in-Memory Architecture* (CLiMA) that exploits the in-memory computing paradigm while also targeting flexibility and high performance. A version of CLiMA based on an emerging non-CMOS technology, namely Nano Magnetic Logic, is also presented. The effectiveness of the CLiMA approach is validated through comparisons with the non-LiM architecture presented in the first part of this thesis. Moreover, a *taxonomy* that classifies the main works found in literature regarding the in-memory processing topic is presented.