Exploration of Beyond von Neumann Computing to solve the Memory-Wall issue

Andrea Coluccio

Supervisor
Prof. Mariagrazia Graziano

Committee:

Prof. Alberto Bosio, Full Professor, INL – Ecole Centrale de Lyon, Lyon, France
Prof. Giovanni Finocchio, Associate Professor, Università degli studi di Messina, Dipartimento di Scienze matematiche e informatiche, scienze fisiche e scienze della terra, Italy

Summary

The impressive growth in complexity of transistor technology has been the driving force behind modern electronics. Many applications (e.g., Neural Networks), which have become increasingly popular over the years, require processing enormous datasets quickly, placing stringent requirements on the hardware. Many computer architectures employed today are mainly based on a Central Processing Unit (CPU) and memories: the CPU executes the instructions composing programs, takes data from memory, and, once the processing is over, stores back the outcomes in the memory. As a result, these CPU-Memory structures, known as von Neumann architectures, require frequent data exchanges, wasting time and power. In addition, CPU and memory have not followed the same trend, resulting in an increasingly wider performance gap that requires the CPU to wait for memory constantly. This problem, known as Memory-Wall, is the most significant bottleneck preventing modern systems from keeping up with the performance demands of the latest applications. Therefore, a complete redefinition of the computing paradigms is required to overcome the limitations of von Neumann structures. A possibility lies in Beyond von Neumann Computing (BvNC), where part of the computational elements is moved close or even inside the memory, aiming to reduce the data traffic and execute tasks in parallel, achieving energy and time savings. However, developing new computing methods often requires a comprehensive rethinking of the design paradigm: for this reason, researchers have developed specialized software and CADs to assist designers with new computing paradigms or technologies. These tools generally specialize in one or more types of BvNCs and state-of-the-art architectural templates and focus either on simulations, performance estimations, or both. This thesis work presents a tool known as Design Explorer for In-Memory Architectures (DExIMA). Differently from existing tools, the idea is to define the architecture with high flexibility, going through the whole design flow up to automatic simulation, performance estimation, and comparison with von Neumann architectures. DExIMA maintains architectural-level descriptions, so estimations can be done on any technology if implemented inside the tool. The framework allows the designer to develop BvNC architectures in a simple and guided manner by providing a schematic editor, supported implementation of algorithms and control units, Register Transfer Level (RTL) simulation, circuit performance estimation with DExIMA-Backend (an ad-hoc tool implemented in C++), and comparisons with von Neumann architectures using the Gem5 and Cacti by HP tools. At the end of the design flow, DExIMA will give an indication of the performance obtained in the BvNC case to understand how effective this type of solution is compared to a classical implementation. DExIMA also provides the architecture's RTL code that can be synthesized with classical EDA tools. Each step is guided by DExIMA, equipped with a PyQT5-based Graphical User Interface that implements all the previous routines. On the user side, the effort is significantly reduced and consists of defining the algorithm and the architecture. Different benchmarks are proposed that confirm the effectiveness of BvNC and show how, with DExIMA, the user has the possibility to control every step of the design with simplicity, considerably speeding up the whole procedure.

The contributions of this work include a study of the state-of-the-art on BvNC proposals, an overview of existing EDA tools applied to BvNC, proposals of standalone BvNC architectures with demonstrated effectiveness, implementation of the DExIMA tool capable of modeling BvNC structures and estimating performance, validation of DExIMA compared to commercial EDA tools, and evaluation of the tool's versatility through the implementation of various benchmarks and analysis of their impact on von Neumann architectures.