# Towards the next generation Heterogeneous Multi-core Multi-accelerator Architectures for Machine Learning

Vikram Jain[1], Giuseppe Sarda[1,2], Pouya Houshmand[1], Marian Verhelst[1,2]

[1]MICAS, KU Leuven, Belgium, [2]Imec, Belgium; email: vikram.jain@kuleuven.be

The idea of "one size fits all" can no longer meet the requirements in computing as specialized workloads, like deep learning, continue to dominate in edge platforms. These specialized workloads are highly deterministic and can gain orders of magnitude performance, as well as energy efficiency, when mapped onto specialized hardware accelerators. Therefore, more and more heterogeneous and asymmetric systems that combine general purpose CPUs (like RISC-V), GPUs and specialized hardware accelerators are becoming common in current embedded platforms. Designing such heterogeneous multi-core processors, however, comes with its own set of challenges along the whole HW-SW stack: from memory hierarchy design and interface/network-on-chip design to resource partitioning and workload scheduling. This work presents two heterogeneous systems extending a RISC-V based system-on-chip (SoC) with dedicated and flexible hardware accelerators designed for machine learning workloads.

The first SoC, Versa, is designed for sub-mW operation and targets low power extreme edge/tinyML applications in sensor nodes like machine monitoring and keyword spotting. Versa is a highly adaptive SoC consisting of: 1) a RISC-V processor, 2) a flexible ML accelerator which supports efficient mapping of several ML workloads like convolution, fully connected, temporal convolution, auto-encoders with deconvolution and support vector machines 3) an embedded MRAM which acts as a non-volatile storage space for boot code and ML parameters, and 4) wake-up and sleep modes that enable always-on, on-demand or duty-cycled inferences essential in many tinyML applications. Diana, our second heterogeneous system, is designed for embedded applications like robotics, machine vision, etc., that require high performance at high energy efficiency. Diana is a single CPU core, multi-accelerator SoC. It consists of a RISC-V core, a flexible ML accelerator in digital domain (similar to Versa but with more processing elements) and an analog in-memory computing-based hardware accelerator. The key idea of the SoC is to exploit both digital and IMC ML acceleration concepts and to select the best accelerator based on the workload characteristics. The multi-accelerator system allows simultaneous execution of subsequent layers across the digital and analog cores, assigning high-precision layers and layers with limited IMC utilization (e.g. FC layers and layers with low channel count) to the digital core, and all other compute-intensive layers to the IMC core. This brings advantages in both throughput and energy efficiency, with the chip achieving 14.4 TOPS/W when running ResNet-20 with CIFAR-10.

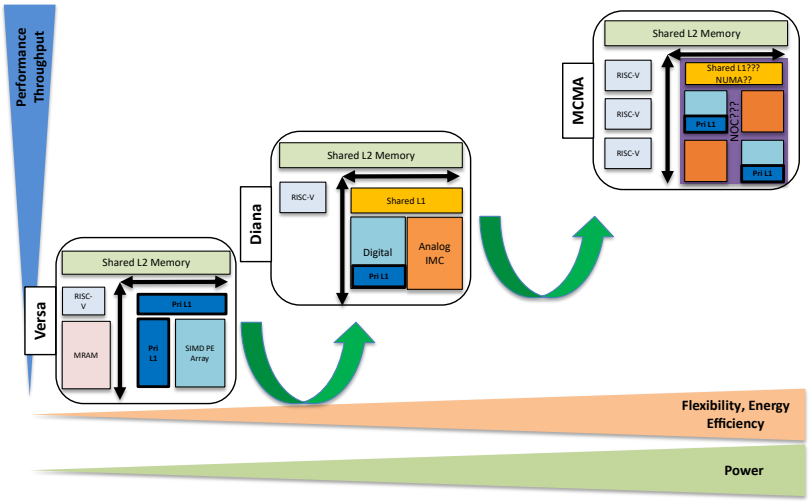


Fig. 1: Moving towards multi-core multi-accelerator (MCMA) architecture can bring improvement in overall performance.

State-of-the-art neural network models evolve very quickly, and future SoCs will have to keep up with the demand for higher flexibility, without compromising performance and efficiency. Our systems integrate only a single base CPU implementation of RISC-V, as we use it for the central control and for lightweight pre- and post- processing. But in tasks which would require complex and irregular operations, or tasks that are not supported by standard Machine Learning accelerators, a single core is not enough. The natural evolution, and our idea, would be to integrate multiple heterogeneous accelerators with multiple RISC-V CPUs to truly enable an omnipotent system, which can back up the always-growing demand from SW.