Analog In-Memory Computing (AIMC) is a promising paradigm to overcome the “Memory Wall”.

Physical fabrication of the crossbar devices limits the full exploitation of this technology, which constrain the memory capacity of a single array.

Multi-AIMC architectures have been demonstrated only for tiny and custom CNNs or performing some layers off-chip.

**MOTIVATION**

- General-purpose system based on RISC-V cores for digital computations and nvAIMC cores for analog operations, such as 2D convolutions.
- A scalable hierarchical network-on-chip interconnects the system to maximize on-chip bandwidth and reduce communication latency.

**RESULTS**

- Experimental assessment on an extended version of an opensource system-level simulator.
- Up to 20.2 TOPS and 6.5 TOPS/W for the whole ResNet-18 inference of a batch of 16 256x256 images in 4.8 ms.