Low-power Acceleration of CNNs using Near Memory Computing on a RISC-V SoC

BY KRISTOFFER WESTRING, LINUS SVENSSON, PER ANDERSSON, MASOUD NOURIPAYAM, ARTURO PRIETO, SERGIO CASTILLO MOHEDANO, AND JOACHIM RODRIGUES

Department of Electrical and Information Technology, Integrated Electronic Systems, Lund University, Lund Sweden

{kristoffer.westring, linus.svensson, joachim.rodrigues}@eit.lth.se

1. Introduction

Edge devices are increasingly expected to perform data-intensive applications due to latency and privacy concern.

These devices are constrained by limited energy and performance, intensified by the von Nuemann bottleneck. Innovative architectures are required to overcome these challenges.

Near Memory Computing (NMC) addresses these limitations by bringing the computational units closer to the data storage, thereby reducing the system bus traffic and data movement energy, resulting in improved energy efficiency and performance.

2. Architecture

The system architecture is structured around the functionality of an existing accelerator, which handles all CNN operations except for the dense layer, which is computed in the RISC-V core.

RISC-V Core

RISC-V with custom instructions for Multiply-And-Accumulate (MAC), and specific instructions for Direct Memory Access.

Dual Data Busses

Cross connect and dual busses to RAM for parallell data transactions drastically improves performance.

Memory Integration

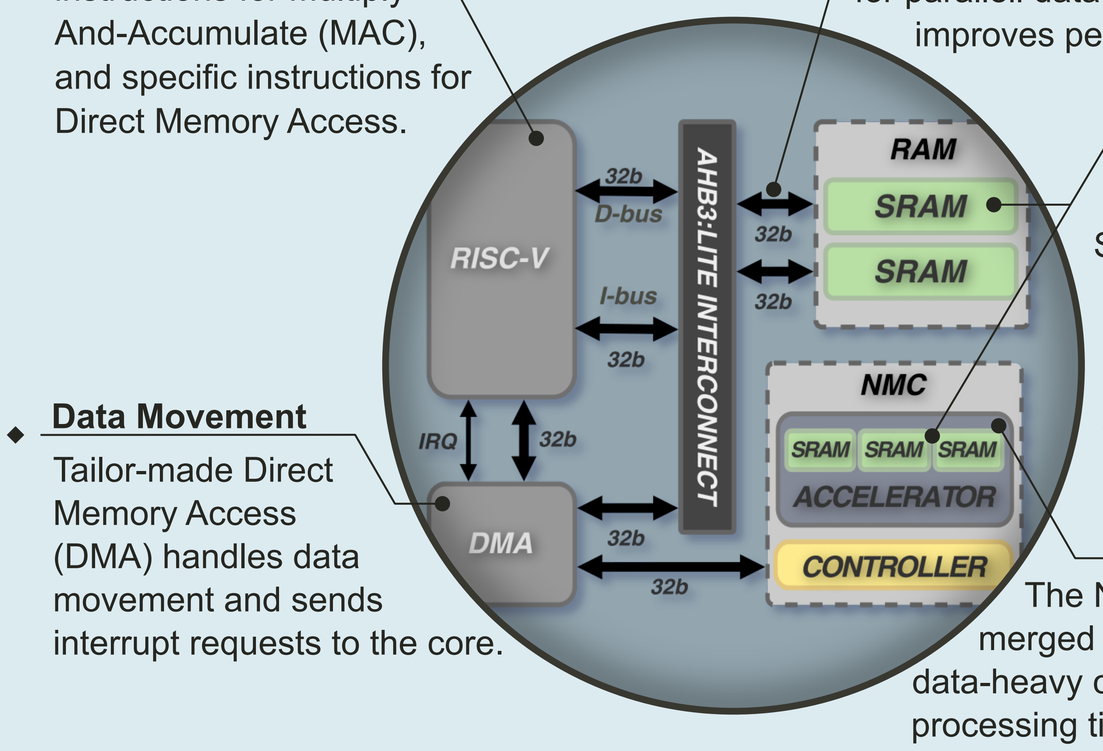
The design leverages the high density of standard SRAM, for seamless integration in both ASIC and FPGA, at the same time providing scalability and modularity.

Data Movement

Tailor-made Direct Memory Access (DMA) handles data movement and sends interrupt requests to the core.

NMC Architecture

The NMC, an accelerator merged with three SRAMs, handles data-heavy convolutions, freeing up core processing time.



3. Challenges

- **Limited energy and area budget**
  Edge devices often run on limited power sources, requiring the architecture to be designed with energy efficiency in mind, especially for data-heavy computations.
- **CPU-Accelerator Cooperation**
  The RISC-V core and the CNN accelerator must work in tandem to ensure high throughput and efficient processing.
- **Modularity and Scalability**
  Flexability for multiple platforms, with a scalable system to fit the requirements for the specific application.

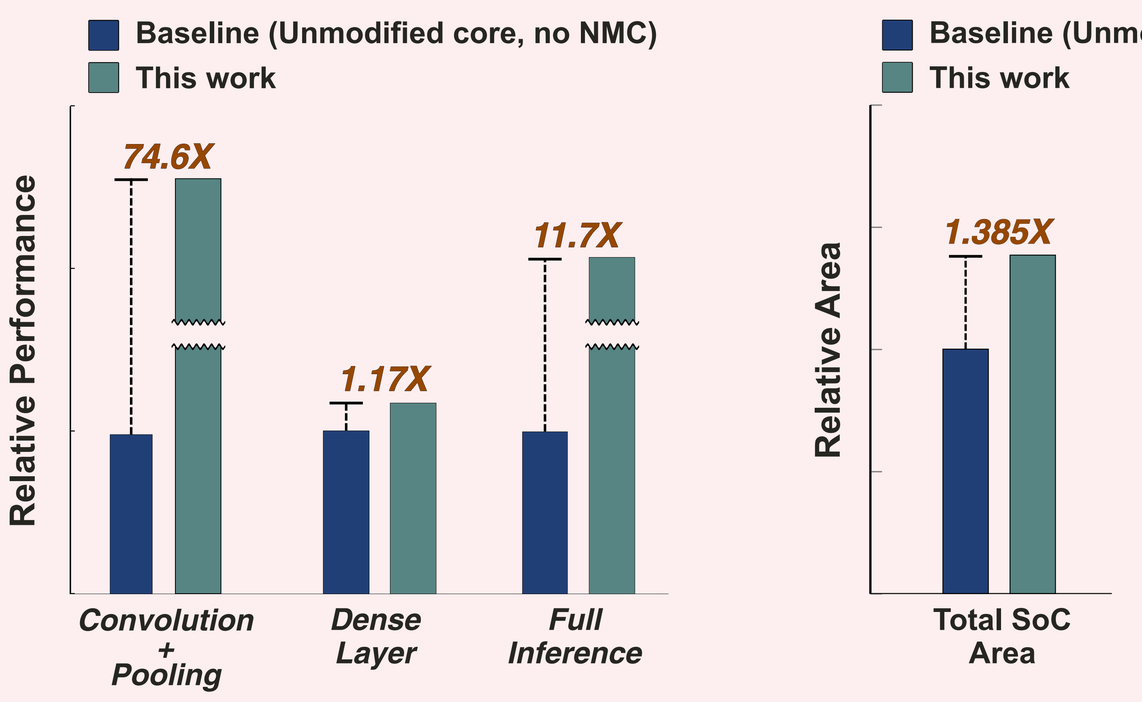
4. Results

Benchmarked using reduced MNIST dataset. The measurements are relative to the baseline case, i.e., SoC with unmodified RISC-V, not containing any NMC.

- Convolution and pooling: **74.7×** speedup compared to baseline.
- Dense Layer: **1.17×** speedup with custom MAC instruction.
- Overall Speedup: **11.7×** for complete CNN operations.
- Area overhead of **38.5%** compared to baseline.
- Performance density improvement: **8.4×**

5. Conclusion

Our results contribute to the growing field of edge computing, demonstrating that integrated circuits for edge devices can be tailor-made to fit specific applications. The proposed design flow provides a streamlined process for merging an existing accelerator with high-density SRAMs, and attaching the resulting NMC module to a RISC-V SoC. This approach includes verification on FPGA, ensuring a successful tapeout in 22FDX technology.



Contact

kristoffer.westring@eit.lth.se
Lund, Sweden

This research was enabled with the support of

Codasip

Learn more