End-to-End DNN Inference on a Massively Parallel Analog In-Memory Computing Architecture

Nazareno Bruschi, Giuseppe Tagliavini, Angelo Garofalo, Francesco Conti, Irem Boybat, Luca Benini, Davide Rossi
Introduction
Introduction
Contributions

• Present a general-purpose system based on RISC-V cores and nvAIMC tiles interconnected through a scalable hierarchical network-on-chip

• Present an efficient data-flow execution model

• Evaluate all the system inefficiencies performing ResNet-18 inference of a batch of 16 256x256 images

• Open-source the hardware (simulator) and the software (source code)
Analog In-Memory Computing
Analog In-Memory Accelerator

- 130 ns Computation Latency (Analog)
- 256x256 Crossbar Size
- 64 byte/cycle Streamers Bandwidth
- Digital (data streams) and analog (MVM) tasks can be overlapped [3]

[3] A. Garofalo et al., “A Heterogeneous In-Memory Computing Cluster for Flexible End-to-End Inference of Real-World Deep Neural Networks”
Cluster

- 1 IMA
- 16 RISC-V cores (with DSP features [4])
- 1 MB of L1 Memory (Tightly-Couple Data Memory)
- 1 DMA
- Cores, DMA and IMA tasks can be overlapped [5]

[4] M. Gautschi et al., “Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices”
[5] N. Bruschi et al., “Scale up your In-Memory Accelerator: Leveraging Wireless On-Chip Communication for AIMC-based CNN Inference”
• 512 clusters (CL) @ 1 GHz
• On-Chip Hierarchical Interconnect [6]
  ▪ 4 interconnect levels (L1, L2, L3, Wrapper)
  ▪ 512 Gbit/s bandwidth each
  ▪ 4 cycles latency each
• Off-Chip HBM link 512 Gbit/s, 100 cycles latency
• Heterogenous (Analog/Digital) General Purpose Accelerator Architecture

[6] A. Kurth et.al., “An Open-Source Platform for High-Performance Non-Coherent On-Chip Communication”
Motivation – ResNet-18
Computational model

Pipeline stages

\[
\begin{array}{c|c|c|c|c|c|c}
\hline
 & \text{batch ID} & \rightarrow & \text{one new batch after MAX(in, compute, out)} \\
\hline
0 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
\text{conv} & \text{compute} & \text{out} \\
\hline
1 & 0 & 1 & 2 & 3 & 4 & 3 & 4 \\
\hline
\text{pool} & \text{compute} & \text{out} \\
\hline
2 & 0 & 1 & 2 & 3 & 2 & 3 & 2 \\
\hline
\text{conv} & \text{compute} & \text{out} \\
\hline
\end{array}
\]

\[t\]
ResNet-18 Implementation

data-replication and parallelization

H_in_tile*C_in

H_out_tile*C_in

H_in_tile*C_in_tile

H_out_tile*C_out

Reduction

IMAs / CORES

IMAs

CORES

pipeline stages (reduction tree)
Results [7]

[7] N. Bruschi et al., “GVSoC: A Highly Configurable, Fast and Accurate Full-Platform Simulator for RISC-V based IoT Processors”
Results [7]

[7] N. Bruschi et al., “GVSoC: A Highly Configurable, Fast and Accurate Full-Platform Simulator for RISC-V based IoT Processors”
Results [7]

N. Bruschi et al., “GVSoC: A Highly Configurable, Fast and Accurate Full-Platform Simulator for RISC-V based IoT Processors”
[7] N. Bruschi et al., “GVSoC: A Highly Configurable, Fast and Accurate Full-Platform Simulator for RISC-V based IoT Processors”
Results [7]

[7] N. Bruschi et al., “GVSoC: A Highly Configurable, Fast and Accurate Full-Platform Simulator for RISC-V based IoT Processors”
Conclusions

- General-purpose heterogeneous (analog/digital) multi-core architecture
- Pipelining execution flow
- Techniques to increase the parallelism and split the workload among the nvAIMC cores.
- 20.2 TOPS and 6.5 TOPS/W for a 480 mm2 architecture on the ResNet-18 inference
- Exhaustive performance analysis
- Insights to drive the usage of nvAIMC as a general-purpose platform for DNN acceleration.
Paper ID: 885
nazareno.bruschi@unibo.it

Nazareno Bruschi, Giuseppe Tagliavini, Angelo Garofalo, Francesco Conti,
Irem Boybat, Luca Benini, Davide Rossi

ETH Zürich
IBM