A Lightweight Collective-Capable NoC for Large-Scale ML Accelerators

Luca Colagrande*[1], Lorenzo Leone*[1], Chen Wu[1], Tim Fischer[1], Raphael Roth[1], Luca Benini[1,2]

*Equal contribution; [1]Integrated Systems Laboratory (IIS) ETH Zürich; [2]DEI, University of Bologna

1 Abstract

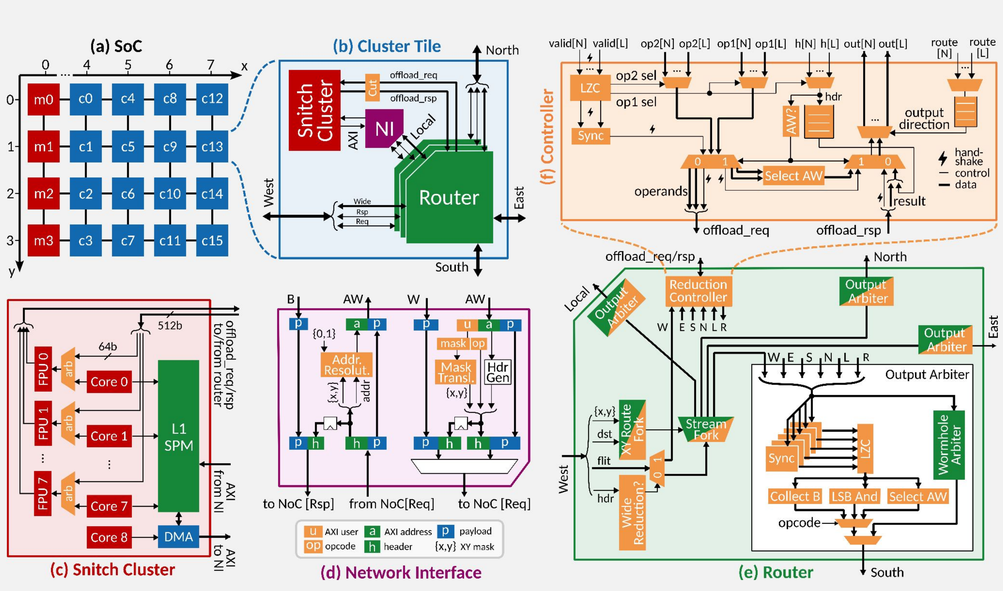
As modern ML accelerators integrate **thousands of PEs** on-chip, the boundary between distributed and on-chip systems has blurred, making efficient **on-chip collective communication** increasingly important. We present a lightweight, **scalable NoC** supporting efficient **barrier synchronization**, **multicast** and **high-throughput reduction** operations, optimized for the next generation of ML accelerators.

2 Implementation

Open source **4x4 compute mesh** built on FlooNoC[1], with **wide (512-bit) network** for high-bandwidth bullk data transfers and **narrow (64-bit) network** for latency-sensitive traffic.

We add **hardware support** for:

1. narrow and wide **multicast**
2. narrow **5-input parallel reductions** for barriers
3. high-throughput **2-input 512-bit FP reductions**



We introduce the **Direct Compute Access (DCA)** paradigm, which grants the NoC direct access to the clusters' FPUs.

Through DCA, we offload the **2-input 512-bit SIMD FP reductions** at each router to the resp. cluster's **8x 64-bit FPUs**.

3 Area Results

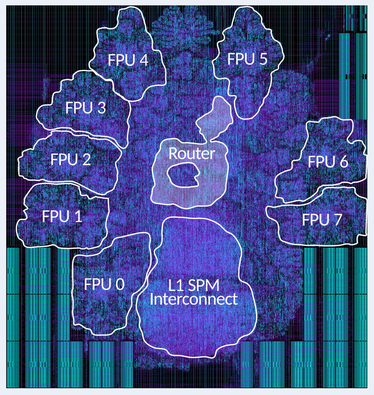
We implement the cluster tile in **GF12LP+** technology, w/ **1GHz** frequency target.

Our extensions introduce only:

+16.9% router **area**

+3.5% network interface **area**

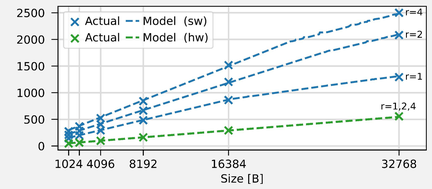
Resulting in **<1%** cluster tile **area overhead**



The 8x 64-bit FPUs occupy large part of the area. By reusing them, DCA enables **significant area savings**.

4 Microbenchmark Results

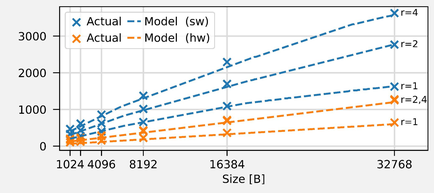
Runtime comparison (in cycles) of software and hardware **multicast** transfers for varying number of rows $r \in \{1,2,4\}$ and transfer sizes.



5.3x geomean speedup on full mesh

Constant runtime w/ **number of rows**

Same comparison for **wide reductions**:

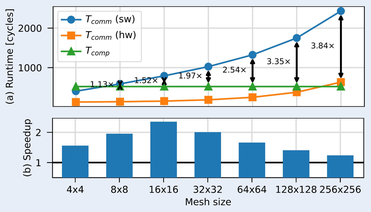


2.8x geomean speedup on full mesh

Constant runtime w/ **number of rows >1**

5 GEMM Results

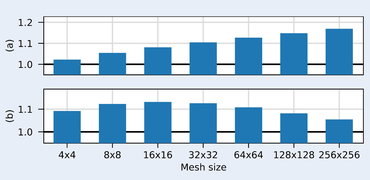
SUMMA[2] (a) and FCL[3] (b) **GEMM runtime comparison** w/ and w/o hardware collectives.



3.8x peak speedup on SUMMA

2.4x peak speedup on FCL

Same comparison for **energy saving**:



1.17x peak energy saving on SUMMA

1.13x peak energy saving on FCL

4 Summary

Scalable, **optimized multicast** implementation **for** regular **ML-oriented traffic**

First work demonstrating lightweight support for **high-throughput in-network reductions**, enabled by the **Direct Compute Access (DCA) paradigm**

5.9x and **2.8x speedup** on **broadcast** and **reduction**

3.8x speedups and **1.17x energy savings** on **GEMM**

<1% compute tile **area overhead**

References

[1] T. Fischer, et al. **"FlooNoC: A 645-Gb/s/link 0.15-pJ/B/hop Open-Source NoC With Wide Physical Links and End-to-End AXI4 Parallel Multistream Support"**. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 33, no. 4, April 2025.
[2] R. A. van de Geijn and J. Watts, **"SUMMA: Scalable Universal Matrix Multiplication Algorithm"**. *Technical Report*, 1995.
[3] V. Potocnik et al., **"Optimizing Foundation Model Inference on a Many-Tiny-Core Open-Source RISC-V Platform"**. *IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCASAI)*, vol. 1, no. 1, Sept. 2024.