ETH zürich

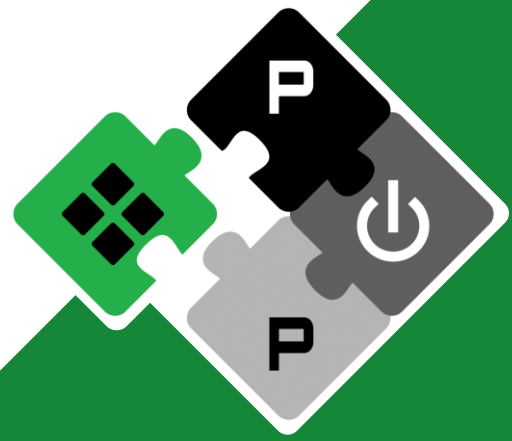
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

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

Integrated Systems Laboratory (ETH Zürich)

Luca Colagrande colluca@iis.ee.ethz.ch
Lorenzo Leone lleone@iis.ee.ethz.ch
Chen Wu chenwu@iis.ee.ethz.ch
Tim Fischer fischeti@iis.ee.ethz.ch
Raphael Roth raroth@student.ethz.ch
Prof. Dr. Luca Benini lbenini@iis.ee.ethz.ch

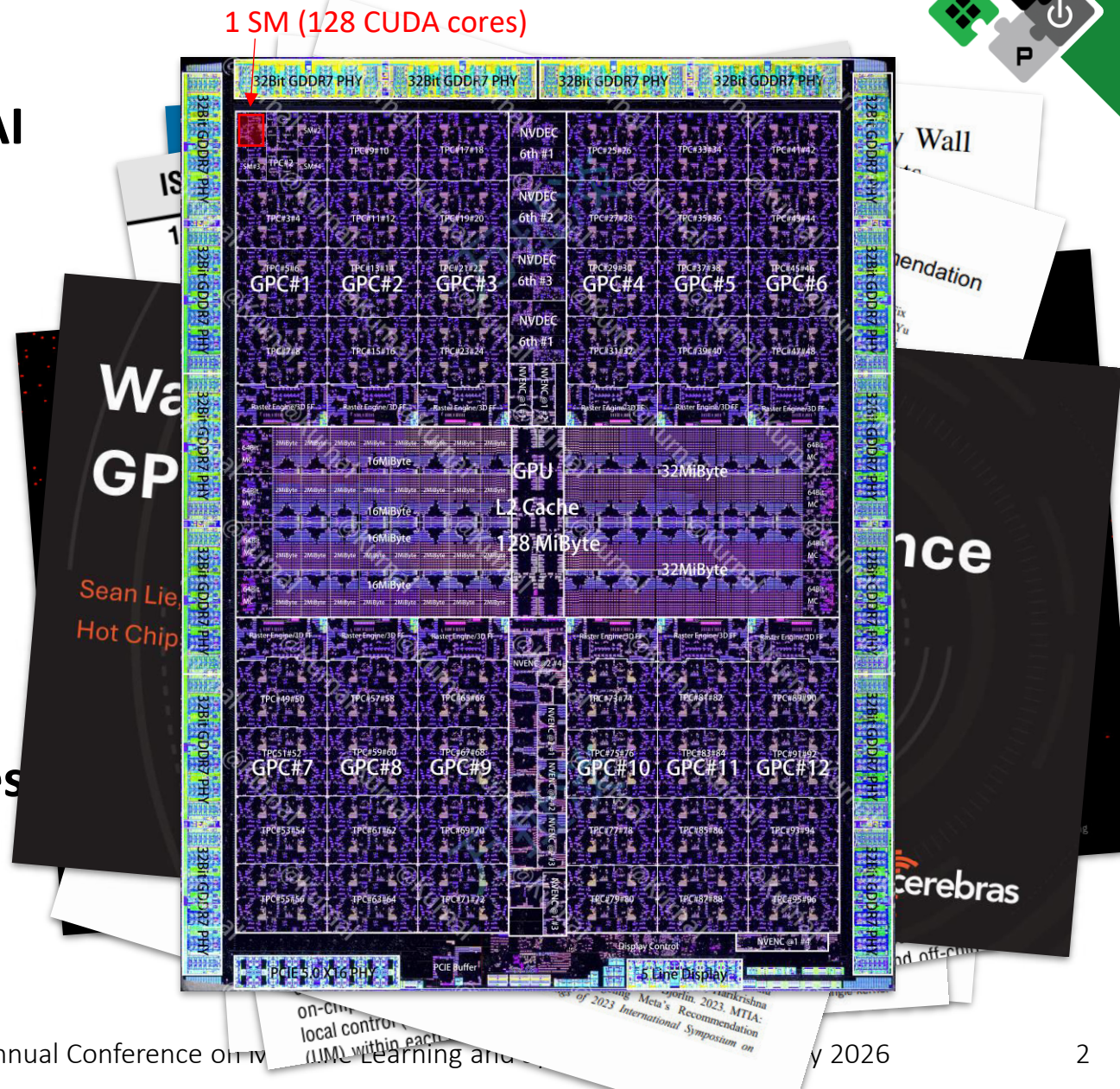
PULP Platform
Open Source Hardware, the way it should be!



pulp-platform.org
@pulp_platform
company/pulp-platform
youtube.com/pulp_platform

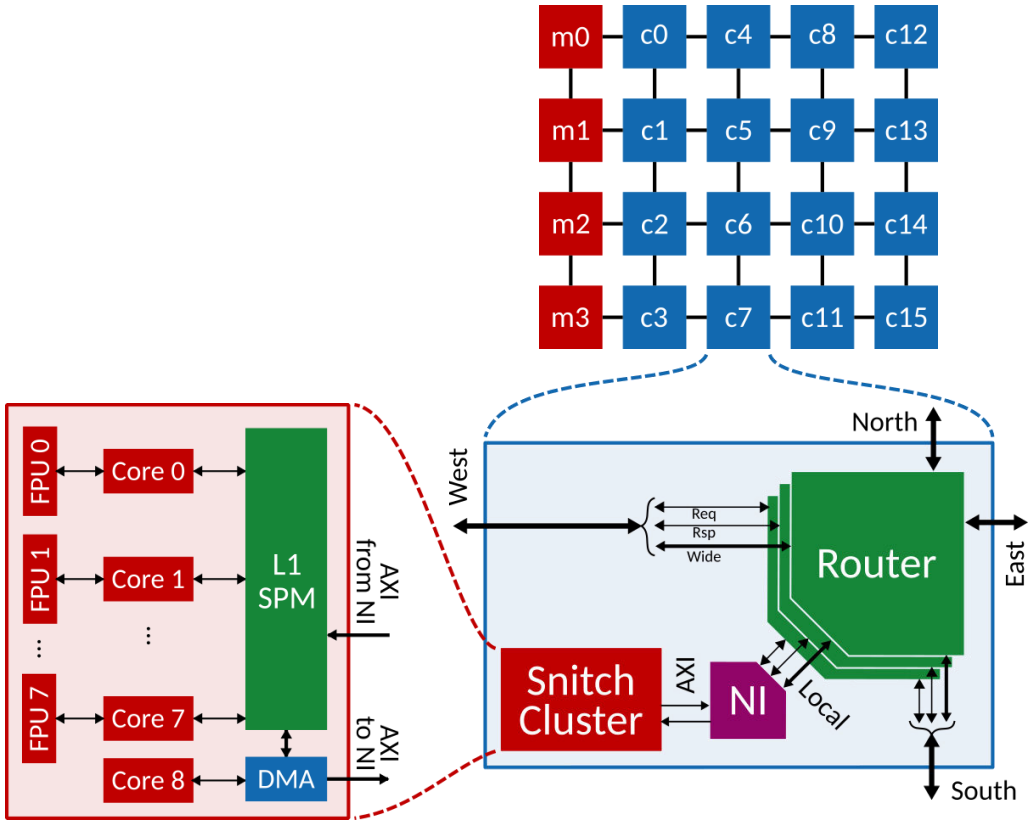
Motivation

- **Growing computational demands of AI**
- **Plethora of manycore accelerators**
  - Greater peak performance than CPUs
- **Greater on-chip parallelism**
  - Thousands of PEs on a single chip
- **Collective communications**
  - Also involve thousands of PEs on-chip
  - Increasingly expensive
- **NoC capable of accelerating collectives**
  - Can be scaled to thousands of cores



Background

- **Open source NoC: FlooNoC[1]**
- **Tile-based SoC similar to FlooOccamy[1]**
  - 4x4 compute tile array + 4 L2 memory tiles
  - Two networks: wide (512b) and narrow (64b)
  - Routers connect tiles with their neighbours
  - Network interfaces (NIs) connect the routers to their local endpoints (AXI interface)
- **Compute tile based on Snitch cluster[2]**
  - 8 64-bit compute cores + 1 DMA core

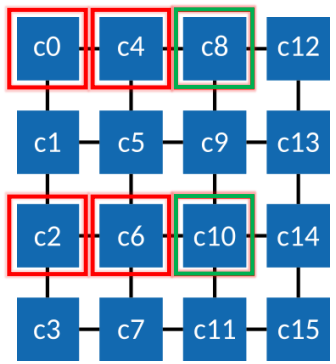


[1] T. Fischer, et al. "**FlooNoC: A 645-Gb/s/link 0.15-pJ/B/hop Open-Source NoC With Wide Physical Links and End-to-End AXI4 Parallel Multistream Support**".
IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 33, no. 4, pp. 1094-1107, April 2025.
[2] F. Zaruba, et al. "**Snitch: A Tiny Pseudo Dual-Issue Processor for Area and Energy Efficient Execution of Floating-Point Intensive Workloads**".
IEEE Transactions on Computers, vol. 70, no. 11, pp. 1845-1860, Nov. 2021.

Encoding

- **Multiaddress encoding for multicast[3]**
- **Extended for 2D meshes and reductions**
- **Properties**
  - **Scalable**
    - Logarithmic w/ size of NoC
    - Independent of number of destinations
  - **"Reasonably" flexible**
    - Cannot represent arbitrary address sets

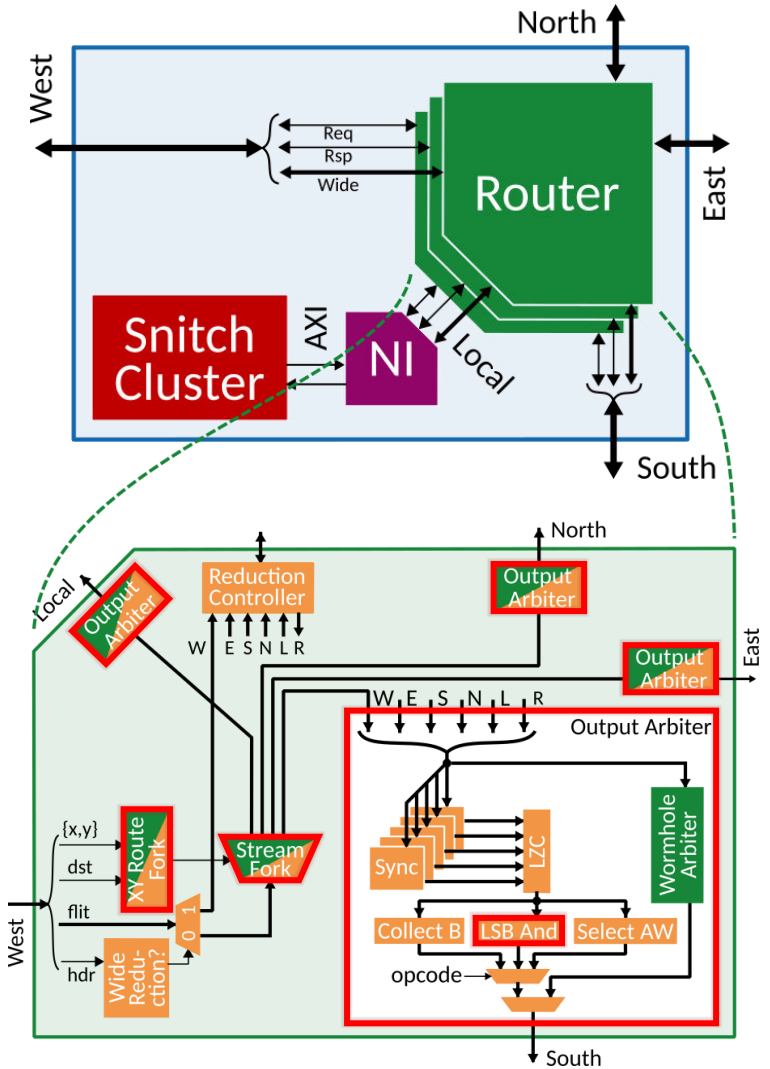

```
addr set = {
   0x10001b00, 0x10101b00,
   0x10801b00, 0x10901b00
}
```



[3] L. Colagrande and L. Benini, "**A Multicast-Capable AXI Crossbar for Many-core Machine Learning Accelerators**". *IEEE 7th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2025.

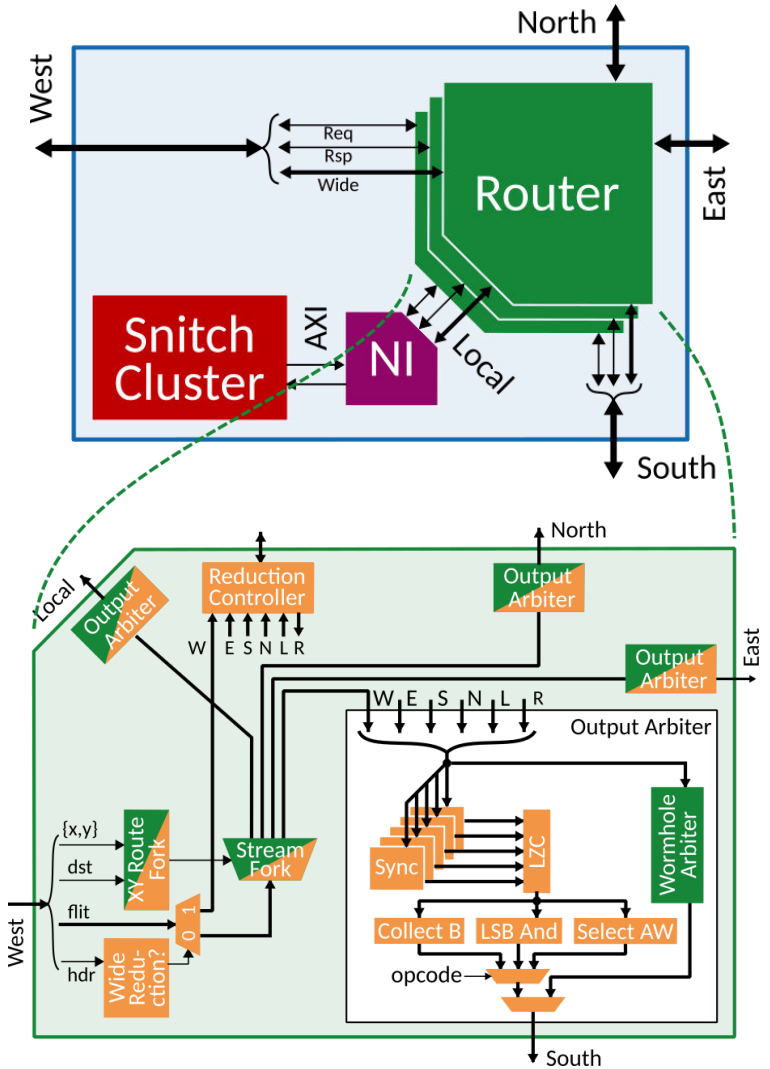
Router

- **Function**
  - Route packets from one port to another (one-to-one)
  - From one-to-many for multicast, from many-to-one for reductions
- **Multicast extension**
  - XY route and stream fork for one-to-many routing
- **Narrow (64-bit) reduction extension**
  - Reduce all inputs in parallel (via reduction tree)
  - Replicated reduction hardware at each output port



Router

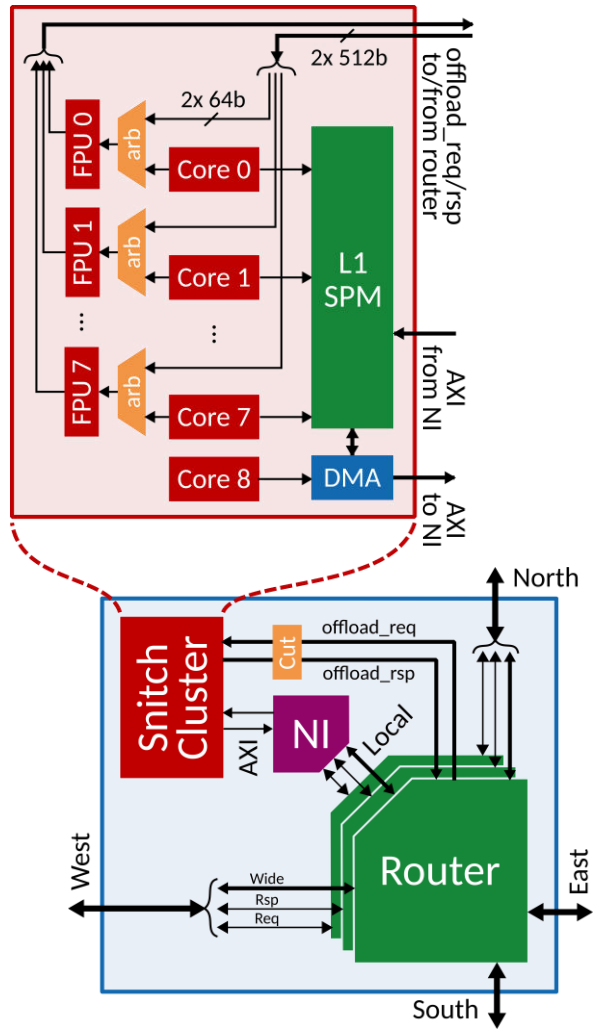
- **Wide (512-bit) reduction extension**



Router

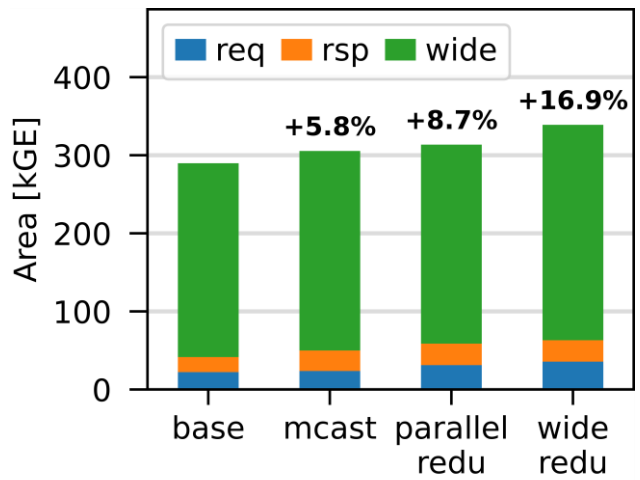
- **Wide (512-bit) reduction extension**
  - ~~5-input reduction tree~~ → 2-input reduction only
  - ~~Replicated per output port~~ → Single shared instance
  - ~~2-input 512-bit ALU~~ → Reuse cluster ALUs

- **Direct Compute Access (DCA)**
  - Grant the NoC direct access to the cluster's ALUs
  - Time share compute datapath between NoC and cores

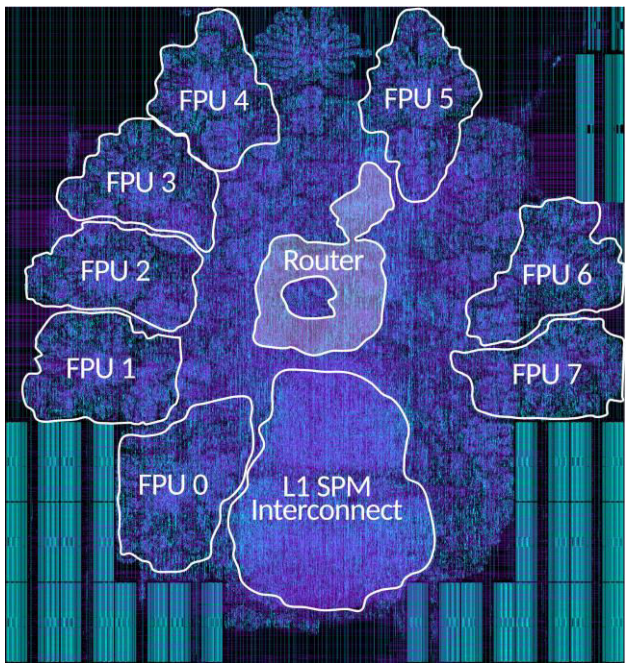


Area Results

- **Router:**
  - **+5.8%** for narrow and wide multicast
  - **+2.9%** for narrow reduction (LsbAnd)
  - **+8.2%** for wide reduction

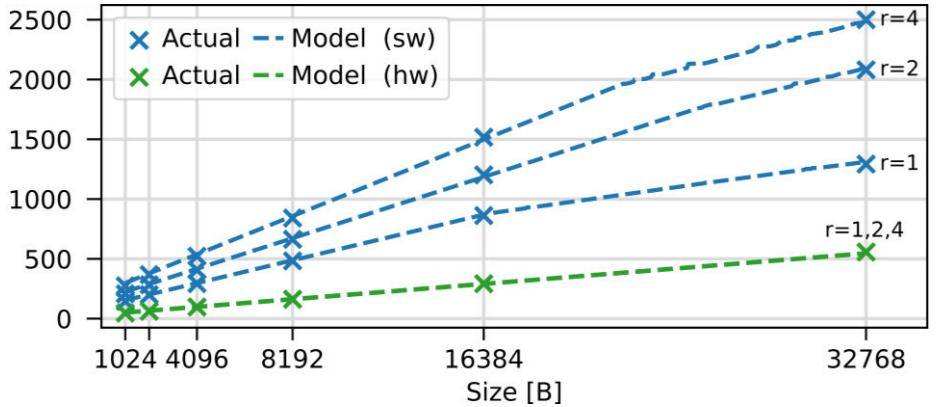


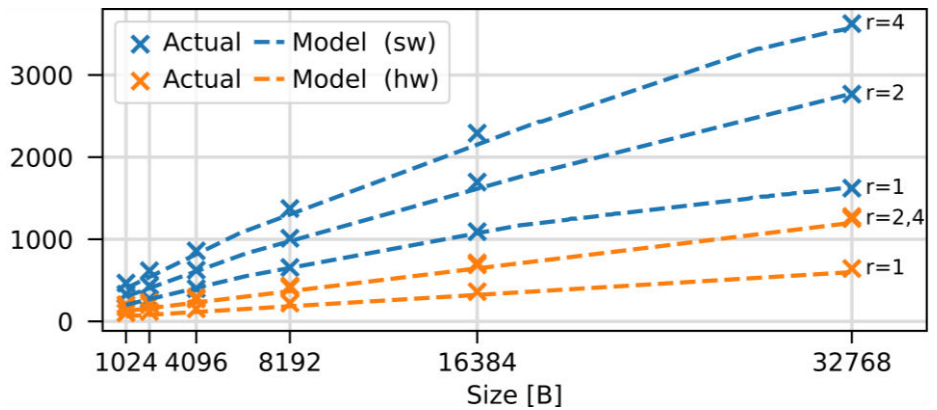
- **Network Interface: +3.5%**
- **Compute tile: <1%**



Performance Results

- **Multicast primitives**
  - Up to **5.3x** speedup on full 4x4 mesh
- **Reduction primitives**
  - Up to **2.8x** speedup on full 4x4 mesh
- **GEMM (two different dataflows)**
  - [4] Hardware multicast delivers up to **3.8x** speedup and **1.17x** energy saving
  - [5] Hardware reduction delivers up to **2.4x** speedup and **1.13x** energy saving





[4] R. A. van de Geijn and J. Watts, "**SUMMA: Scalable Universal Matrix Multiplication Algorithm**". *Technical Report*, 1995.
[5] 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.

Related Work

- **Literature survey on hardware support for one-to-many collectives**
  - Most focus on irre…ations
  - Mandates flexible…
  - Many focus on so…
  - In turn, this requires sophisticated deadlock avoidance or recovery mechanisms

Scalable multicast implementation optimized for regular ML-oriented traffic

- **Literature survey on hardware support many-to-one collectives**
  - Very few works, lim…ages
  - Conventional wisd…

First work demonstrating lightweight support for high-throughput in-network reductions

Conclusions

A **scalable**, **high-throughput** multicast and reduction-capable NoC **optimized** for regular, **ML traffic**

5.3x and **2.8x** speedups on multicast and reduction

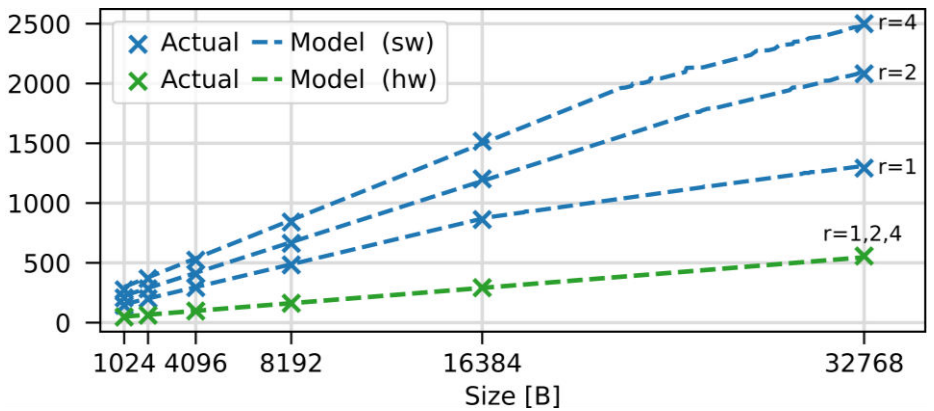
3.8x speedups and **1.17x** energy savings on GEMM

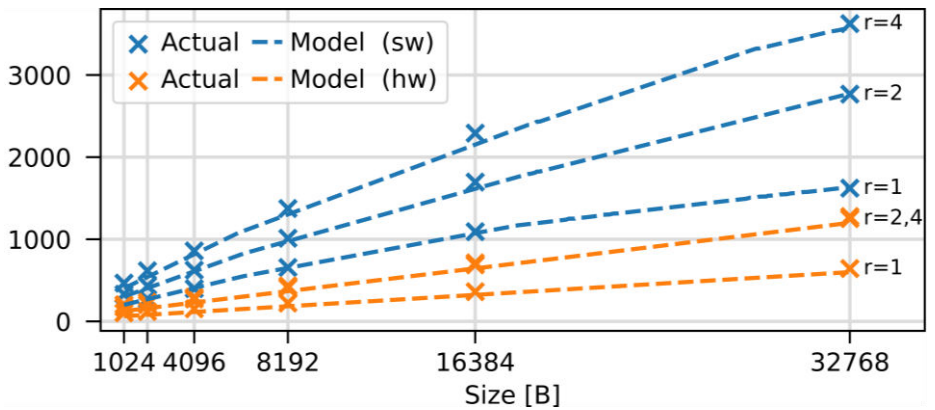
<1% compute tile area overhead

`github.com/pulp-platform/{FlooNoC,picobello}`

Performance Results - Microbenchmarks

- **Multicast:**
  - **2.9x** speedup on single row/column
  - **5.3x** speedup on full mesh
- **Reduction:**
  - **2.5x** speedup on single row/column
  - **2.8x** speedup on full mesh





Performance and Energy Results - GEMM

- **Speedup:**

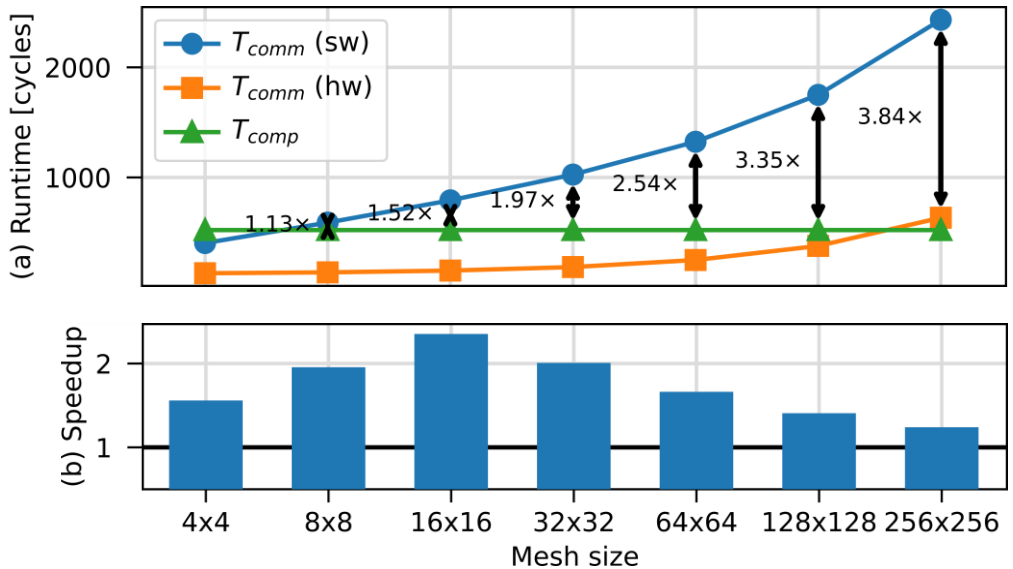
  Up to **3.8x** on SUMMA[4] GEMM

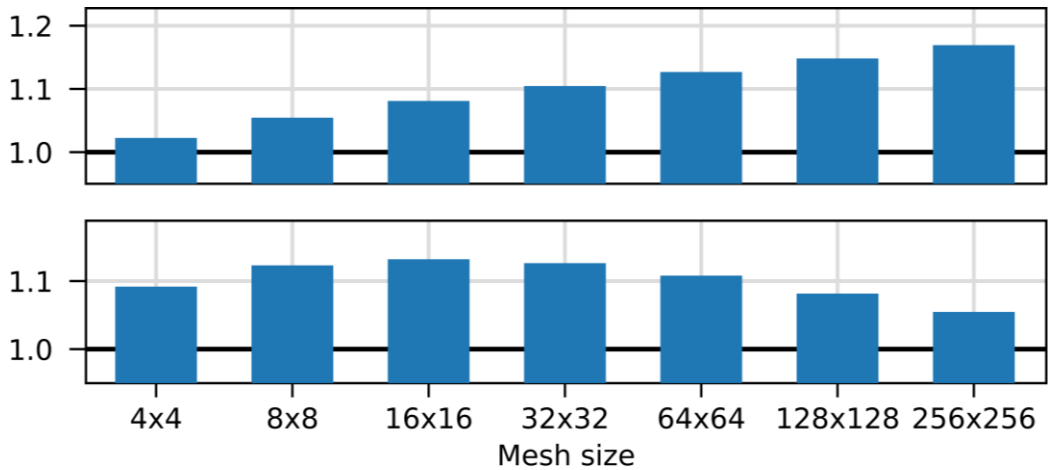
  Up to **2.4x** on FusedConcatLinear[5] GEMM

- **Energy saving:**

  Up to **1.17x** on SUMMA GEMM

  Up to **1.13x** on FusedConcatLinear GEMM

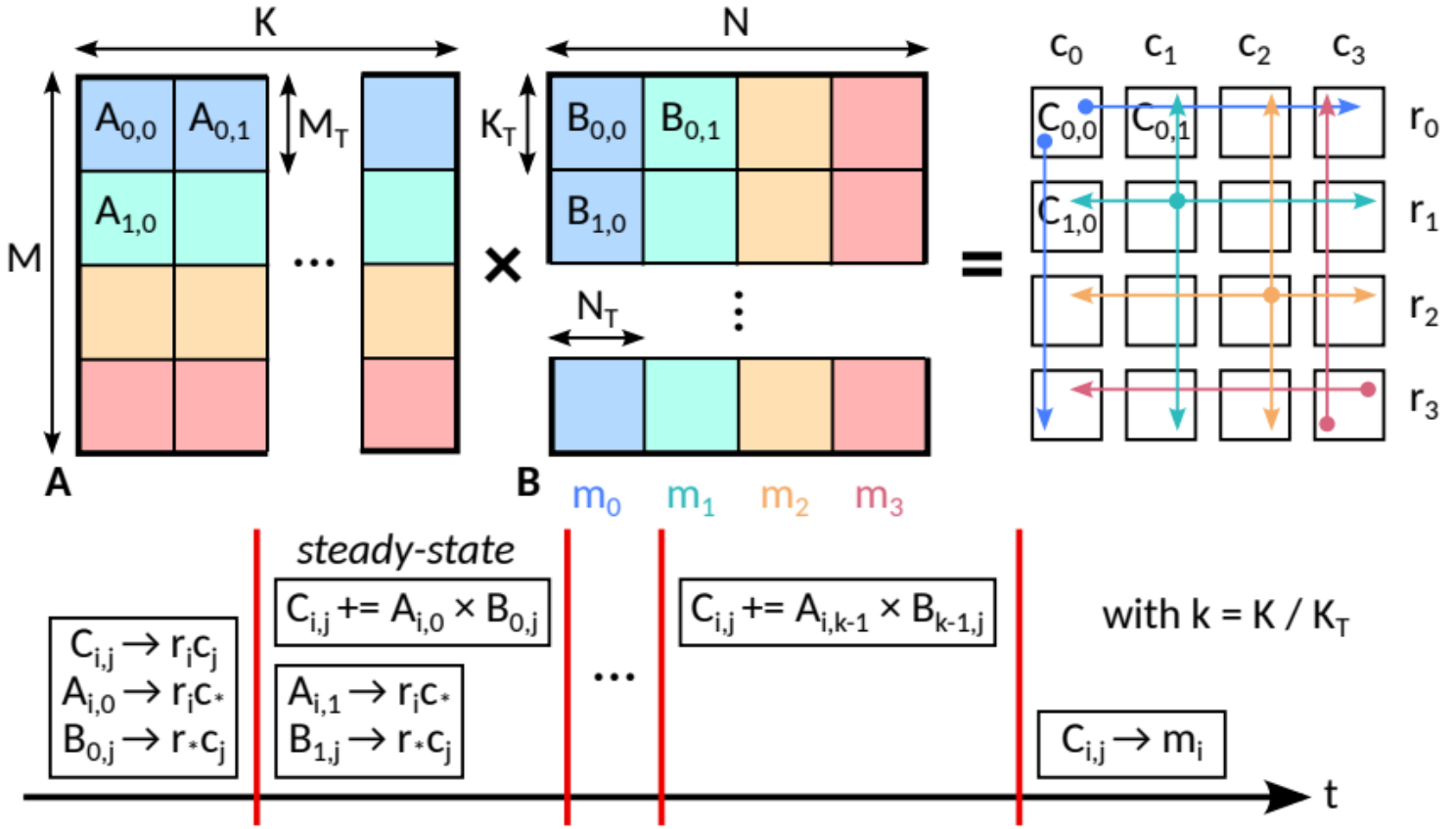




[4] R. A. van de Geijn and J. Watts, "**SUMMA: Scalable Universal Matrix Multiplication Algorithm**". *Technical Report*, 1995.

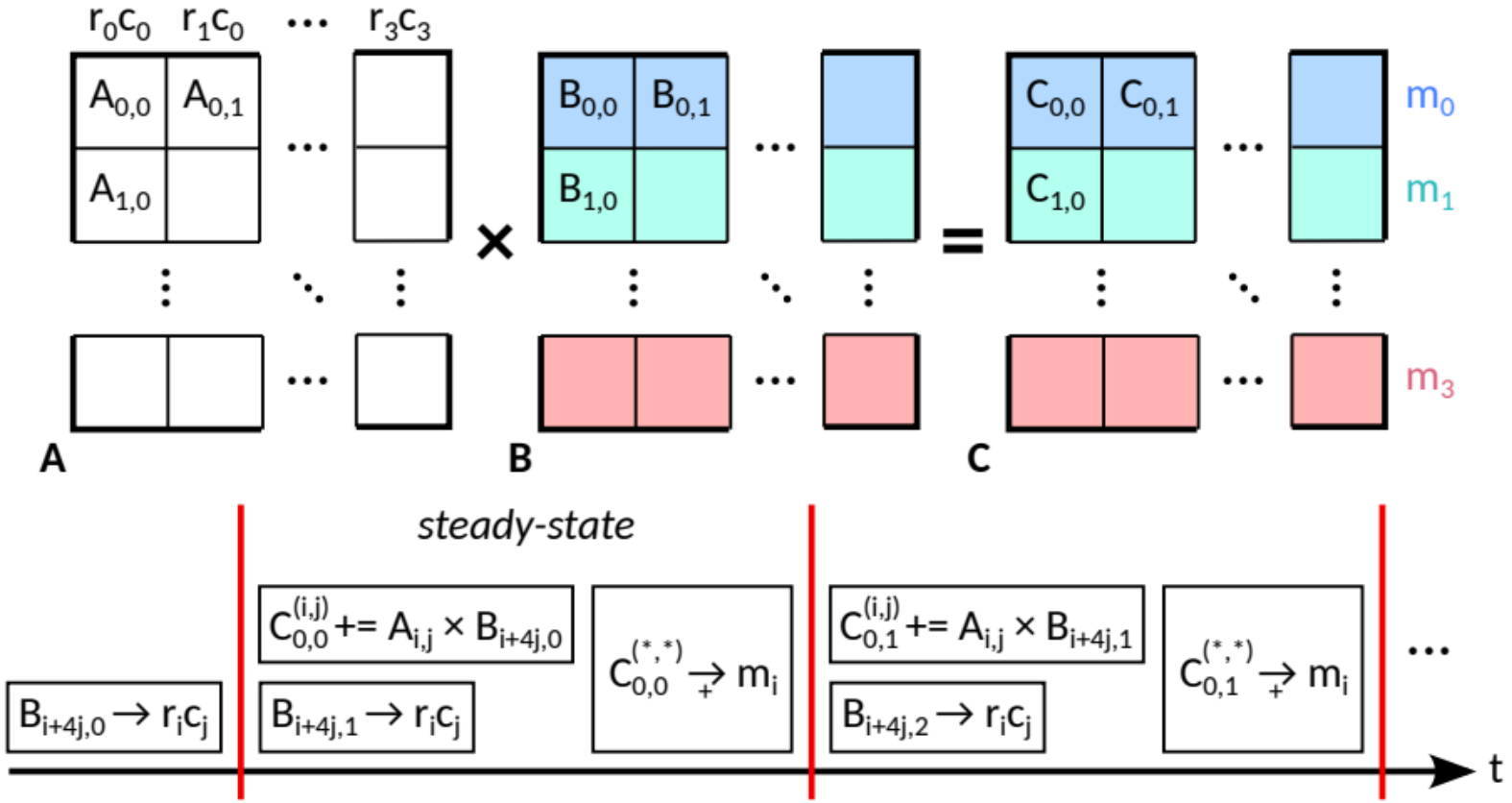
[5] V. Potocnik et al., "**Optimizing Foundation Model Inference on a Many-Tiny-Core Open-Source RISC-V Platform**". *IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCASAI)*, vol. 1, no. 1, Sept. 2024.

SUMMA GEMM[4]



[4] R. A. van de Geijn and J. Watts, "**SUMMA: Scalable Universal Matrix Multiplication Algorithm**". *Technical Report*, 1995.

FusedConcatLinear GEMM[5]



[5] 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.