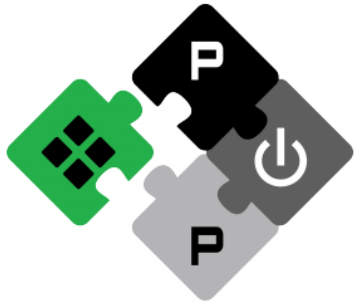


PULP PLATFORM

Open Source Hardware, the way it should be!

Open Source On-Chip Communication from Edge to Cloud: the PULP experience

Davide Rossi <davide.rossi@unibo.it>

Thomas Benz <tbenz@iis.ee.ethz.ch>

Luca Bertaccini <lbertaccini@iis.ee.ethz.ch>

Florian Zaruba <zarubaf@iis.ee.ethz.ch>

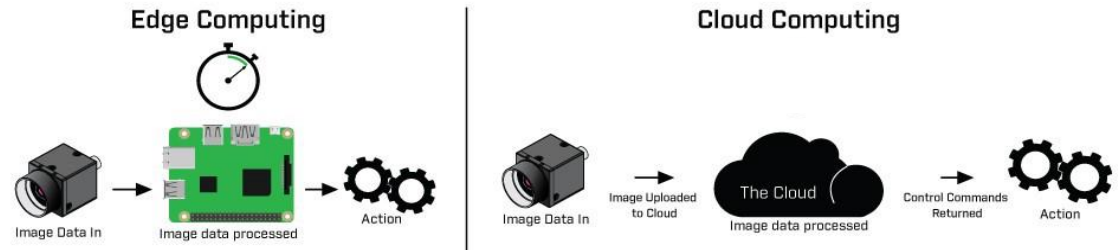
ETH zürich

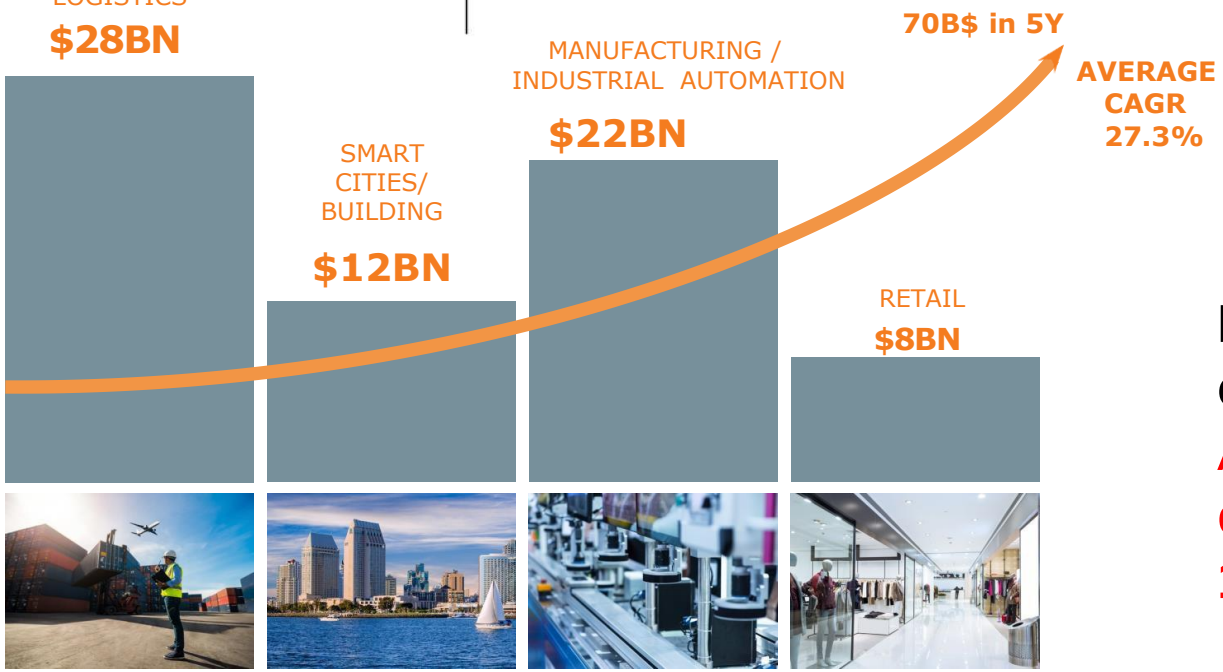
 http://pulp-platform.org  @pulp_platform  https://www.youtube.com/pulp_platform

Agenda

- *Davide Rossi (UNIBO): "PULP: An Open-Source RISC-V Based Multi-Core Platform for In-Sensor Analytics"*
- Thomas Benz (ETHZ): "An Open-Source Platform for High-Performance Non-Coherent On-Chip Communication"
- Luca Bertaccini (ETHZ): "HERO: A Heterogenous Research Platform to Explore HW/SW Codesign and RISC-V manycore accelerators"
- Florian Zaruba (ETHZ): "Manticore as an NoC Case Study: A 4096 Chiplet-based Architecture for Ultra-Efficient Floating-Point Computing"

IoT: Cloud → Edge → Near-Sensor Computing





THE SILENT INTELLIGENCE E. Gousev, Qcomm research

#1 Customer Question on Amazon.com (out of 1,000+):

1. *I don't want any of my (private, personal) videos on any servers not in my control. Is this possible?*

Source: www.amazon.com/ask/questions/asin/B01M3VHG87/

#2 Customer Question on Amazon.com (out of 1,000+):

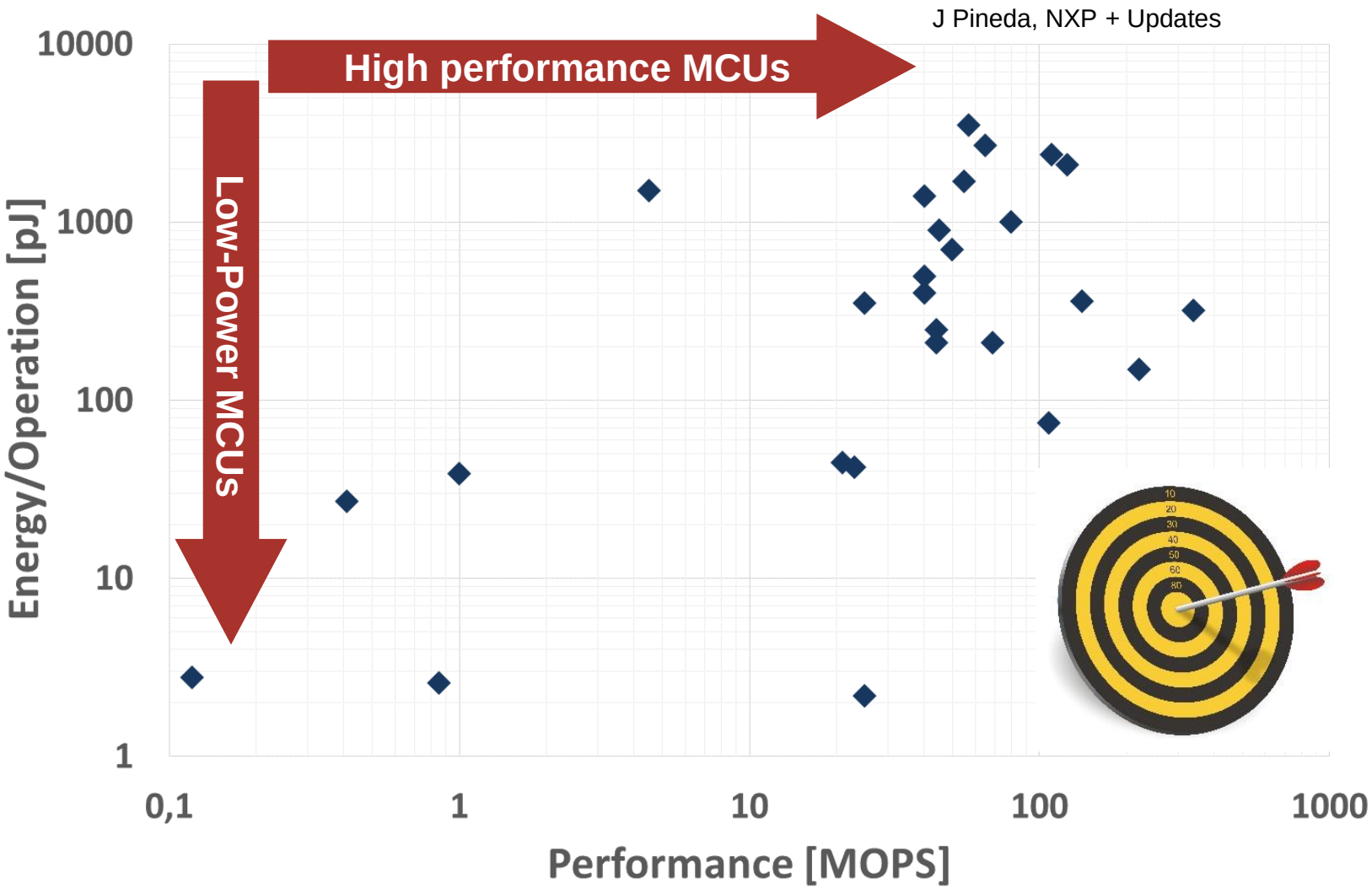
2. *How long does the battery charge last?*

Source: www.amazon.com/ask/questions/asin/B01M3VHG87/

Near-Sensor Computing challenge

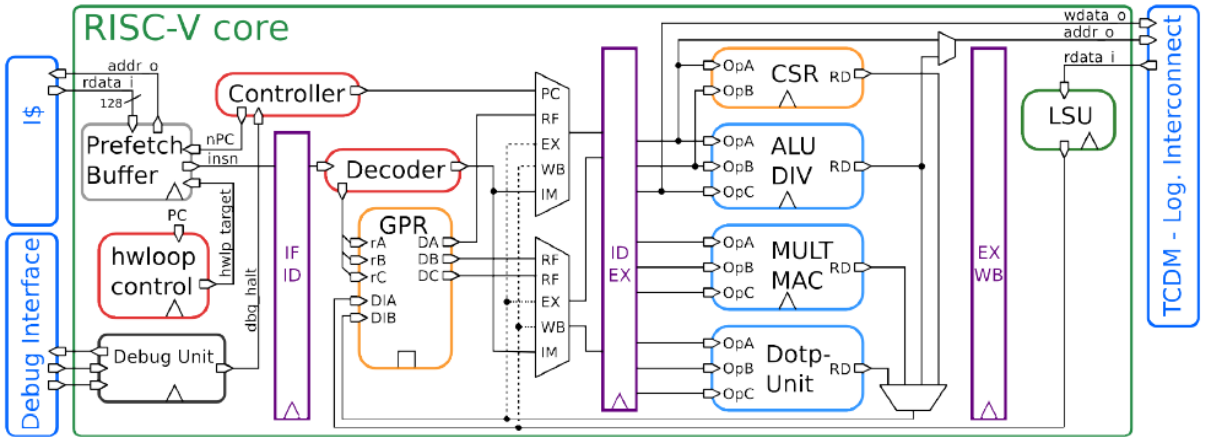
AI capabilities in the power envelope of an MCU:
100mW peak (10mW avg)

Energy efficiency @ GOPS is THE Challenge



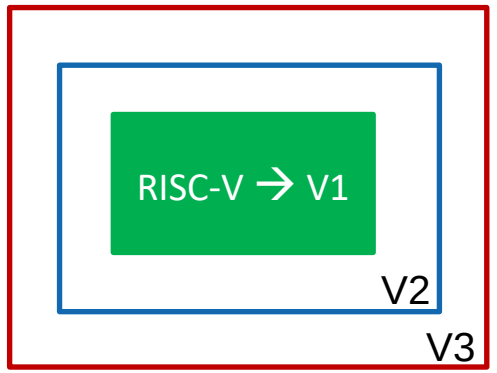
RI5CY Processor

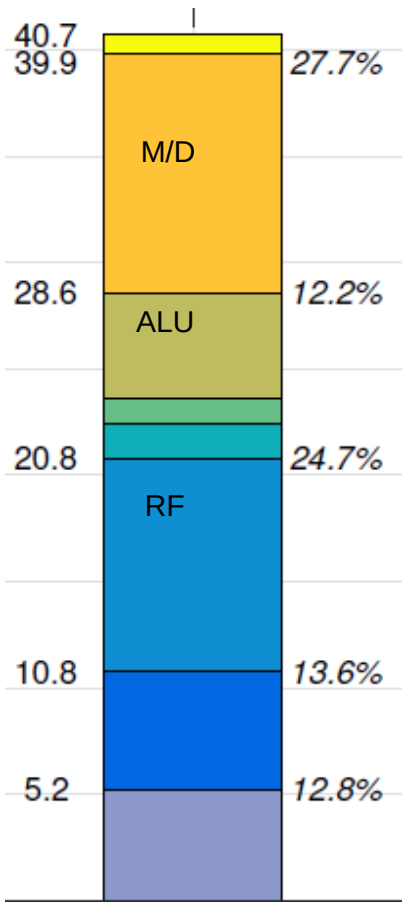
3-cycle ALU-OP, 4-cyle MEM-OP→IPC loss: LD-use, Branch



40 kGE
70% RF+DP

RISC-V core





V1 Baseline RISC-V RV32IMC (not good for ML)

V2 HW loops, Post modified Load/Store, Mac

V3 SIMD 2/4 + DotProduct + Shuffling
Bit manipulation, Lightweight fixed point

XPULP 25 kGE → 40 kGE (1.6x) but 9+ times DSP!

Nice – But what about the GOPS?
Faster+Superscalar is not efficient!

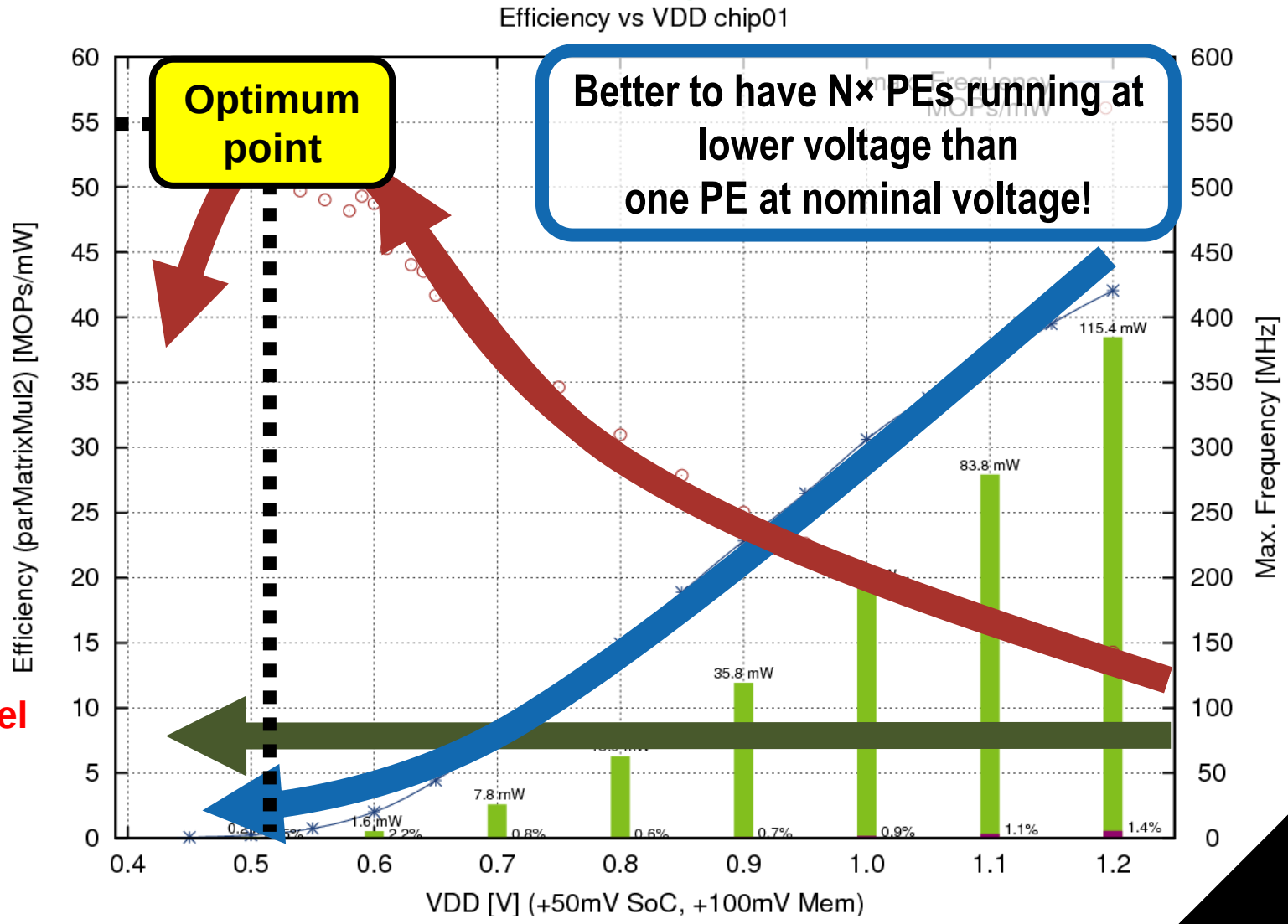
M7: 5.01 CoreMark/MHz-58.5 µW/MHz
M4: 3.42 CoreMark/MHz-12.26 µW/MHz

M. Gautschi *et al.*, "Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 10, pp. 2700-2713, Oct. 2017.

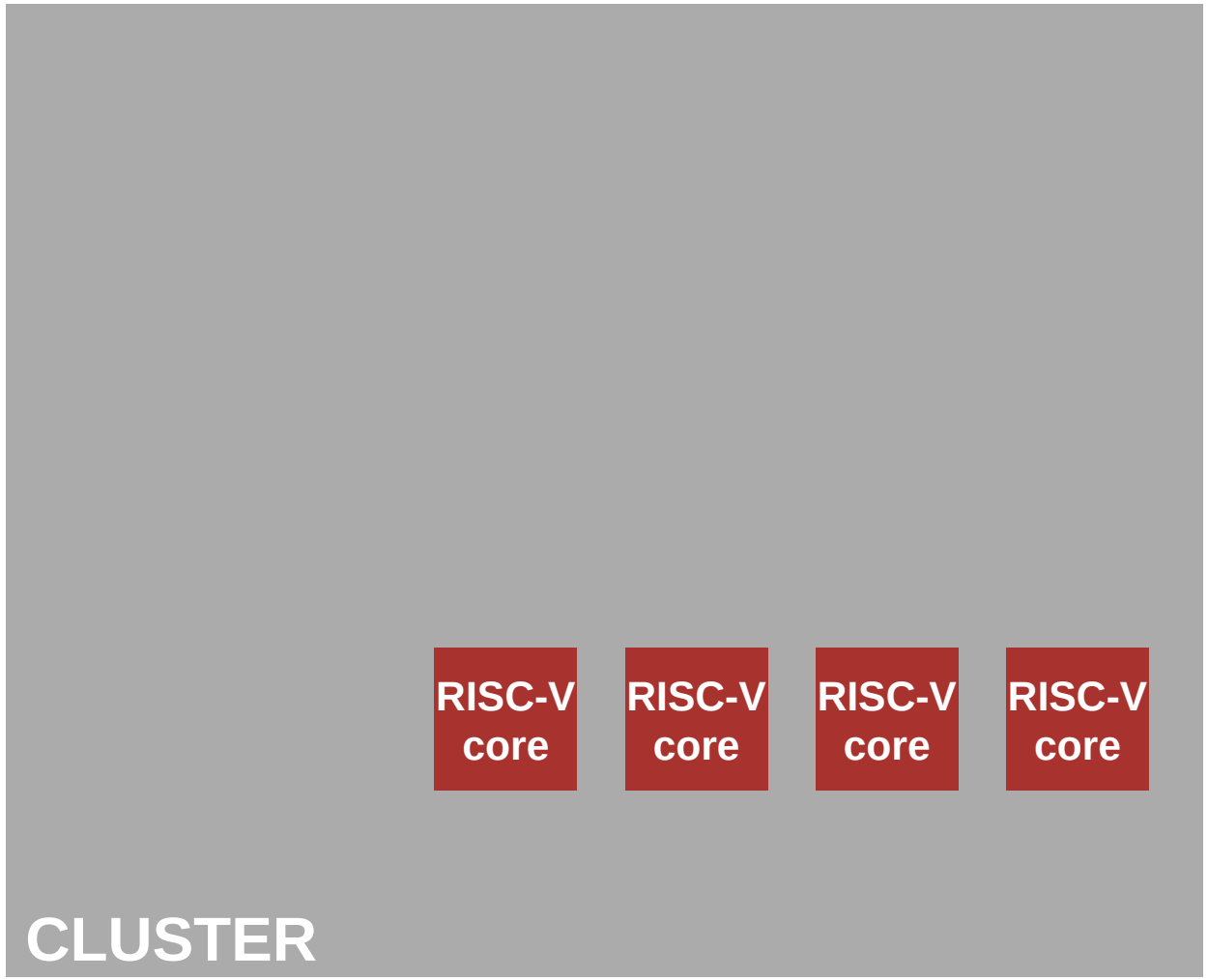
ML & Parallel, Near-threshold: a Marriage Made in Heaven

- As VDD decreases, operating speed decreases
- However efficiency increases→ more work done per Joule
- Until leakage effects start to dominate
- Put more units in parallel to get performance up and keep them busy with a parallel workload

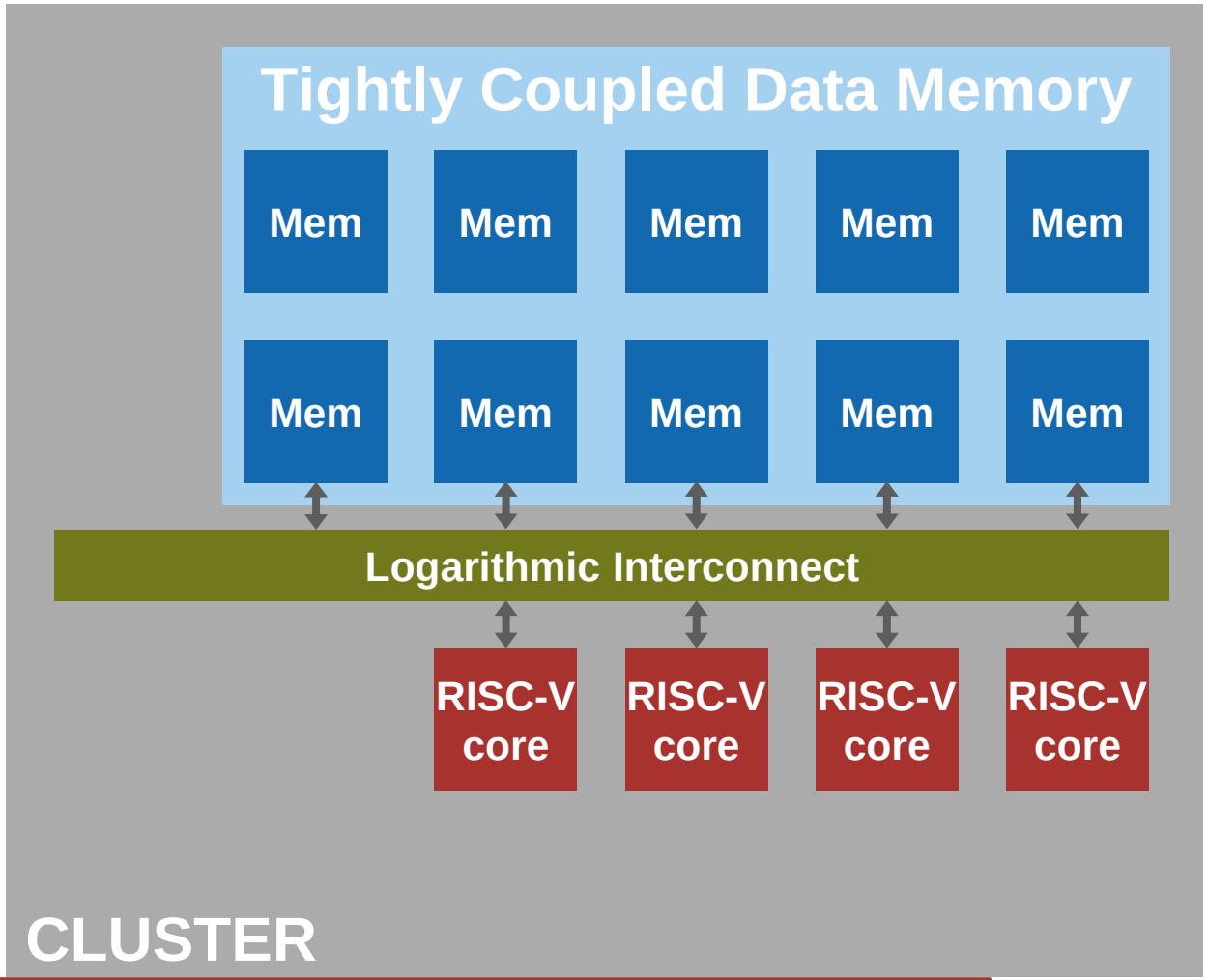
ML is massively parallel and scales well (P/S ↑ with NN size)



Multiple RI5CY Cores (1-16)

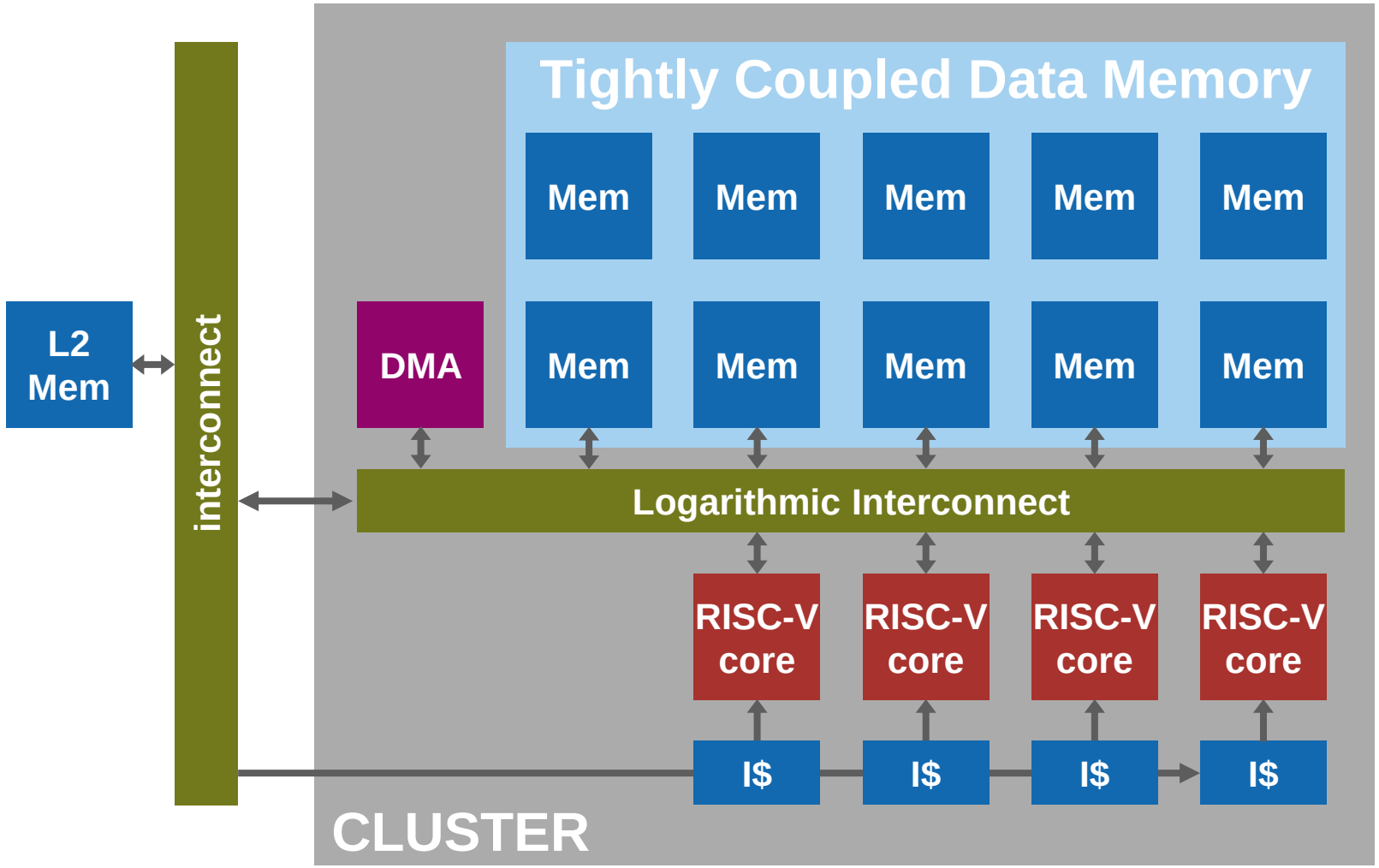


Low-Latency Shared TCDM



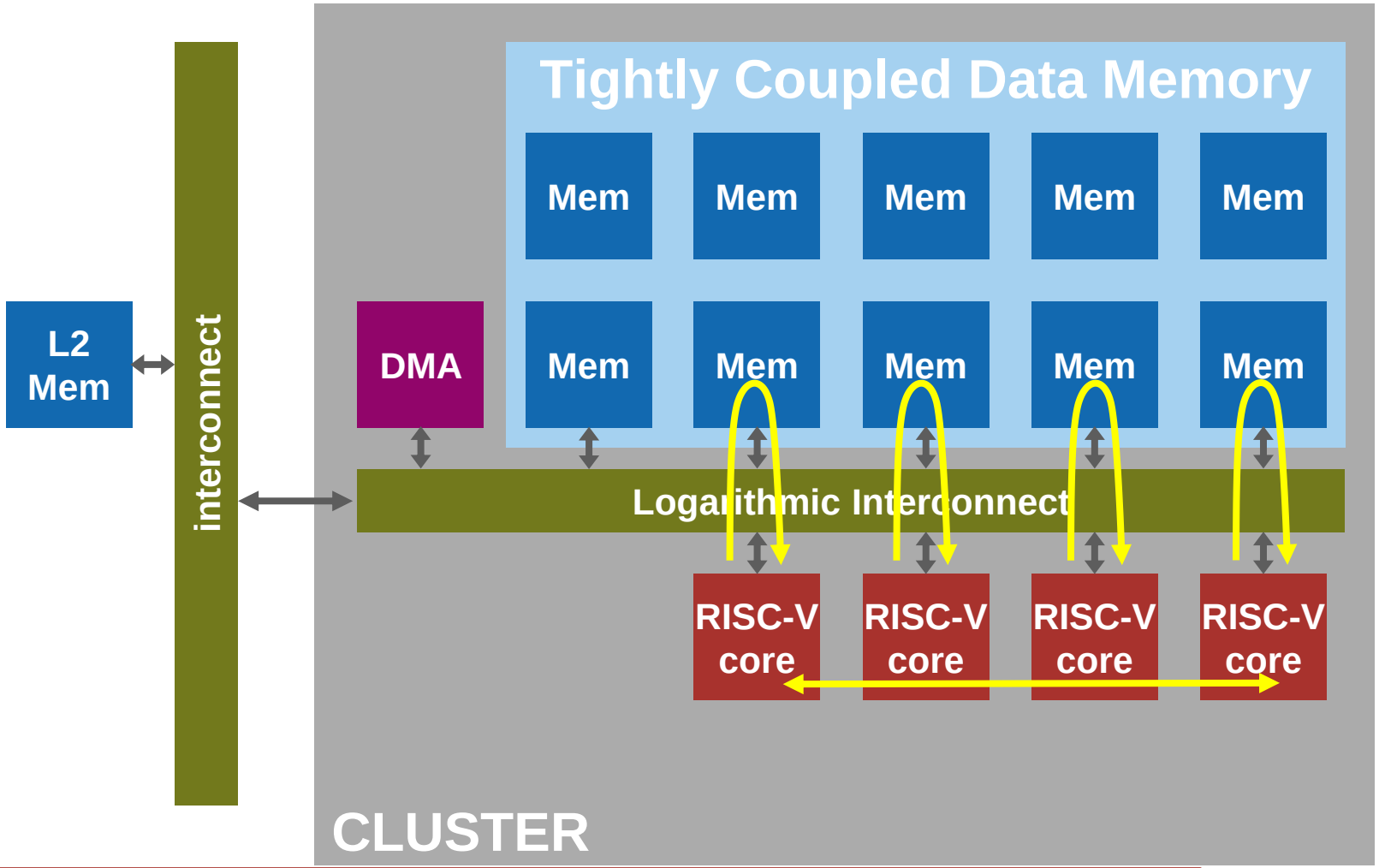
A. Rahimi, I. Loi, M. R. Kakoee and L. Benini, "A fully-synthesizable single-cycle interconnection network for Shared-L1 processor clusters," *2011 Design, Automation & Test in Europe*, 2011, pp. 1-6.

Shared instruction cache with private “loop buffer”



I. Loi, A. Capotondi, D. Rossi, A. Marongiu and L. Benini, "The Quest for Energy-Efficient I$ Design in Ultra-Low-Power Clustered Many-Cores," in *IEEE Transactions on Multi-Scale Computing Systems*, vol. 4, no. 2, pp. 99-112, 1 April-June 2018.

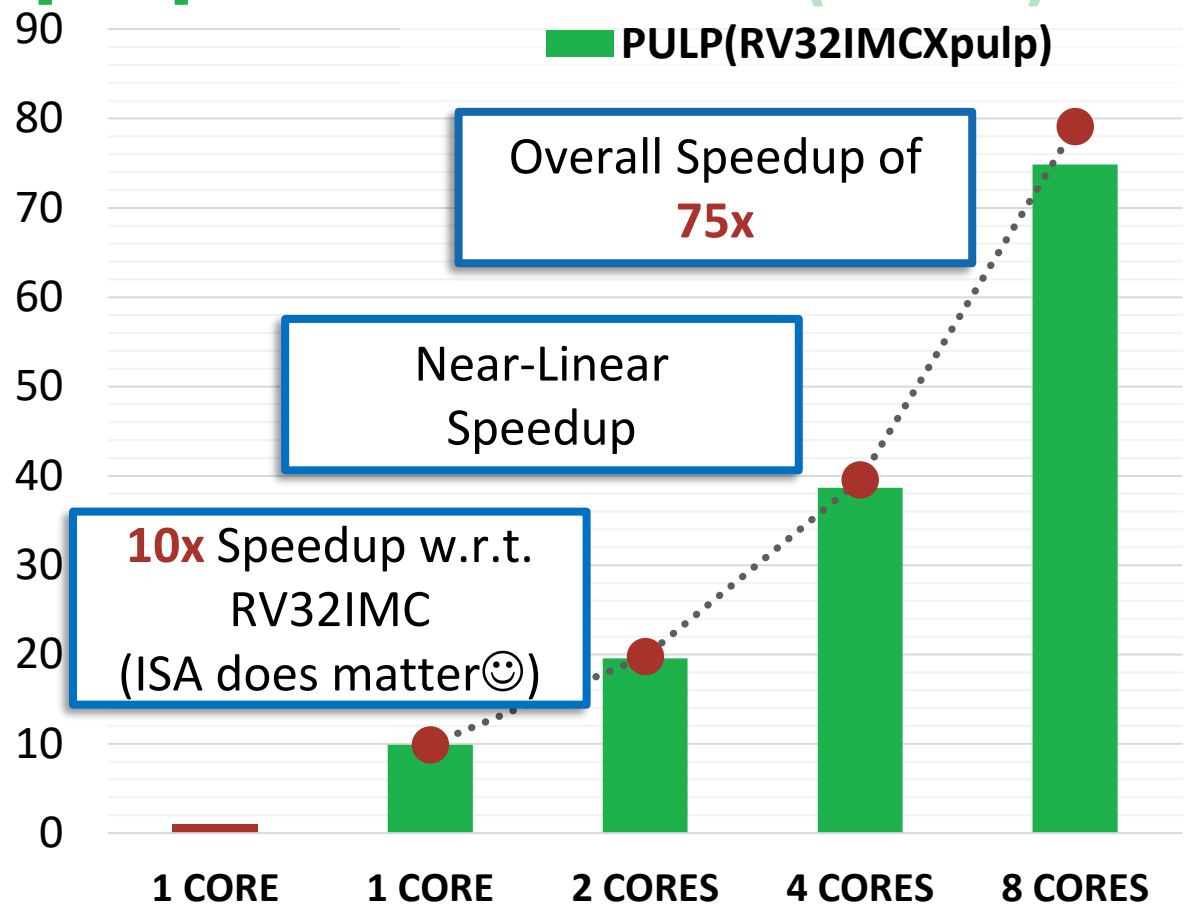
Fast synchronization and Atomics



F. Glaser, G. Tagliavini, D. Rossi, G. Haugou, Q. Huang and L. Benini, "Energy-Efficient Hardware-Accelerated Synchronization for Shared-L1-Memory Multiprocessor Clusters," in *IEEE* TPDS, vol. 32, no. 3, pp. 633-648, 1 March 2021.

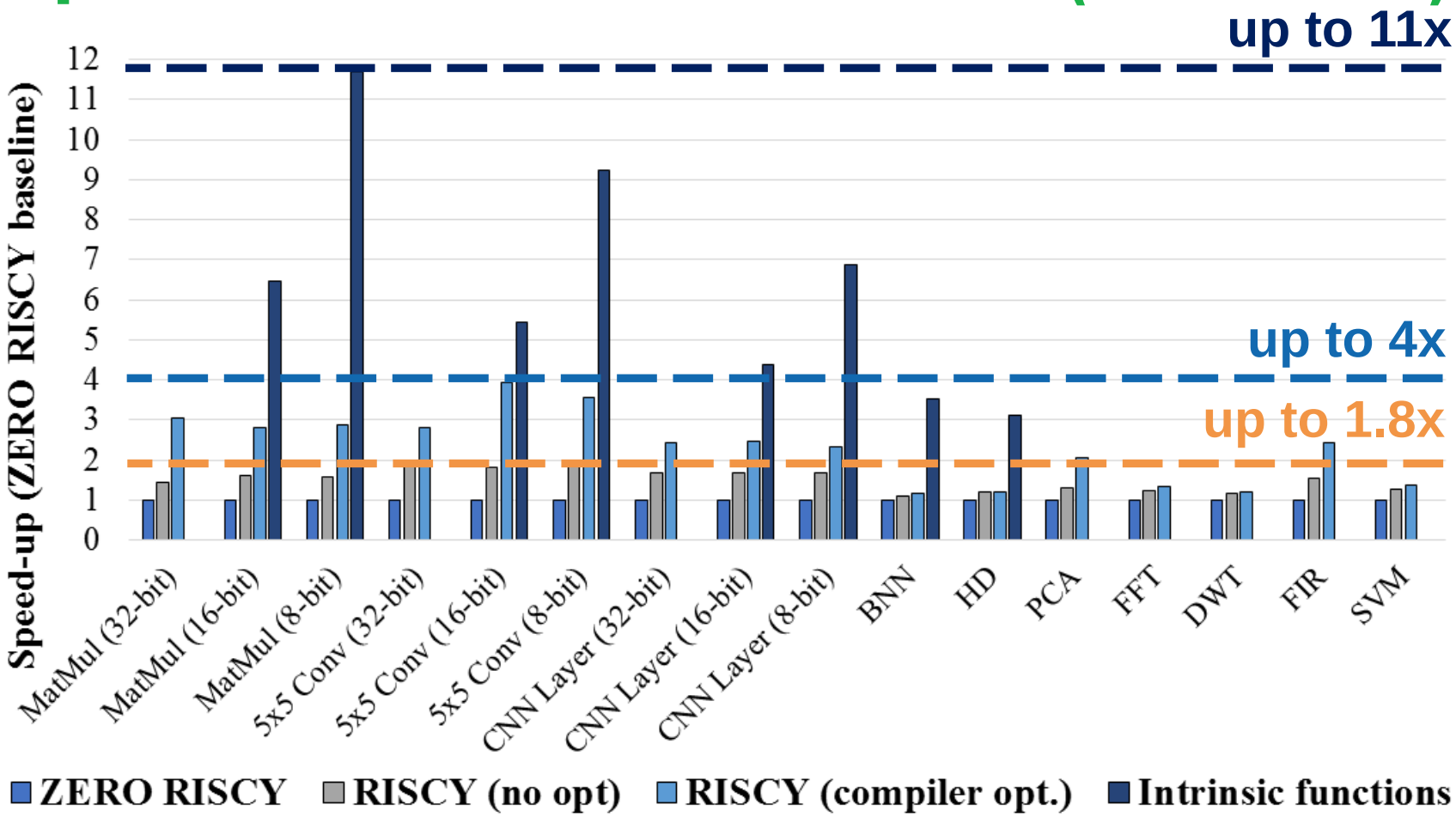
Results: RV32IMCXpulp vs RV32IMC (DNN)

- 8-bit convolution
  - Open source DNN library
- 10x through xPULP
  - Extensions bring real speedup
- Near-linear speedup
  - Scales well for regular workloads.
- 75x overall gain

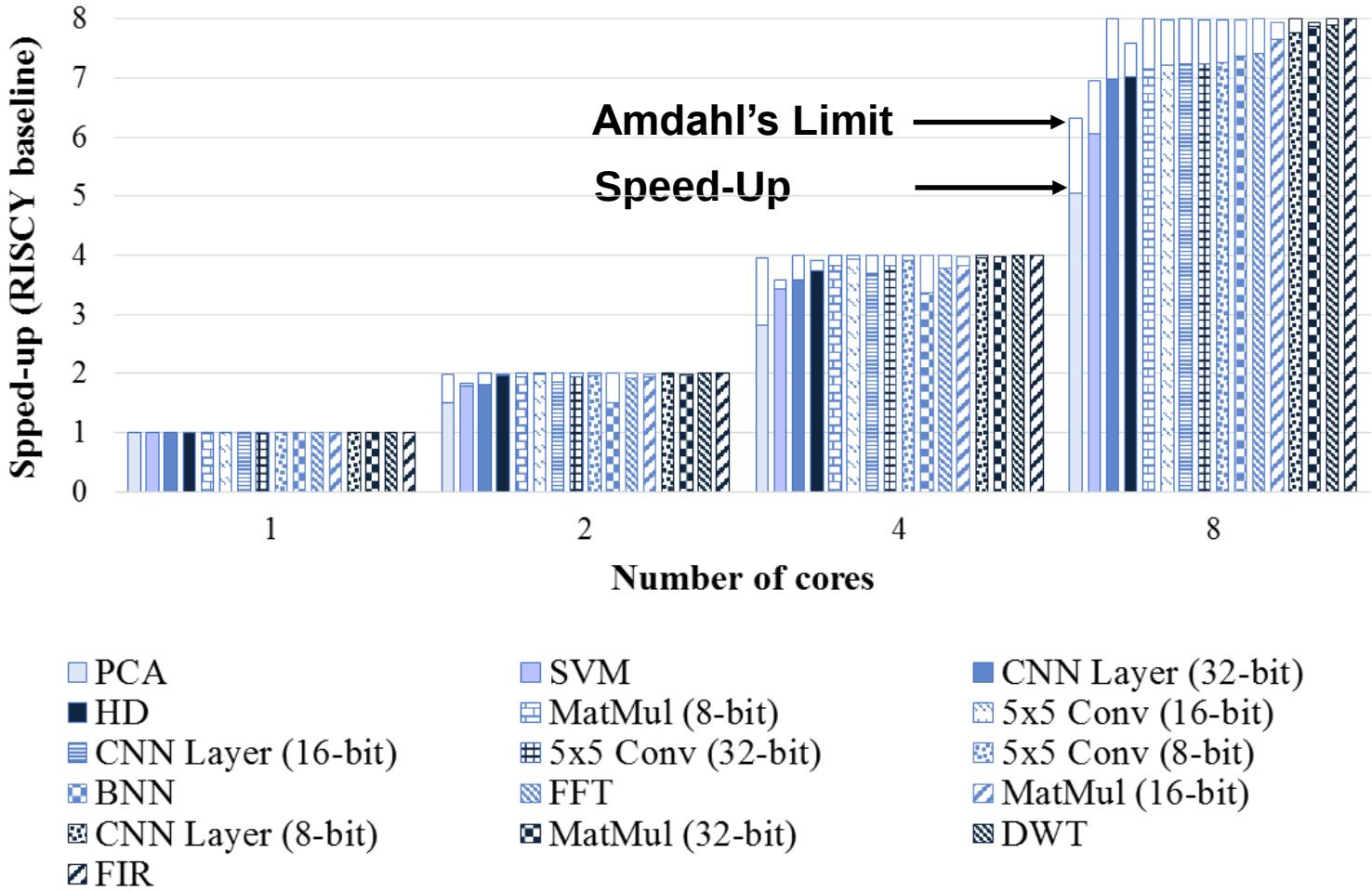


Garofalo, Angelo et al. "PULP-NN: Accelerating Quantized Neural Networks on Parallel Ultra-Low-Power RISC-V Processors." Philosophical Transactions of the Royal Society A

Xpulp Extensions Performance (non-DNN)

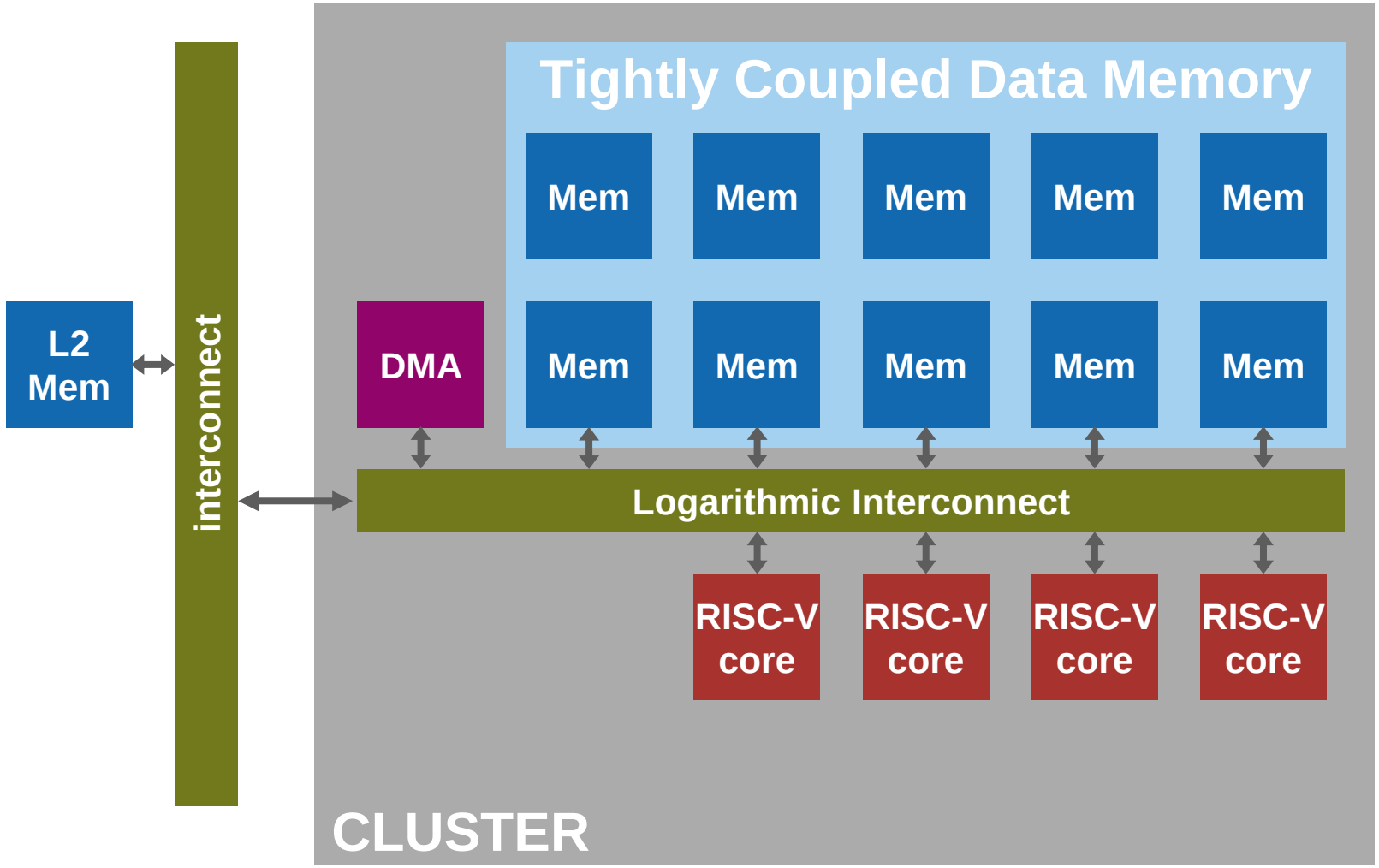


Parallel Speed-Up (non-DNN)



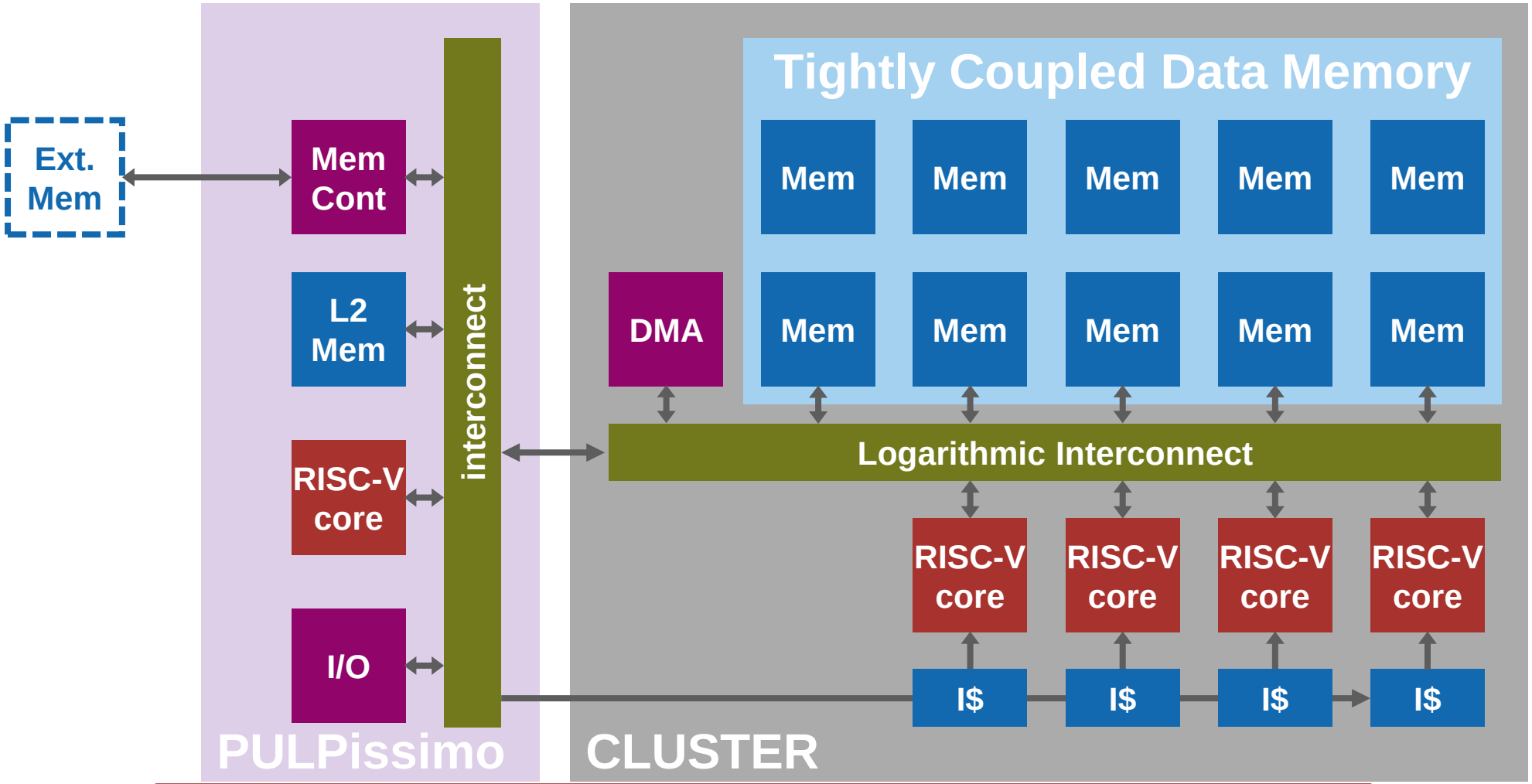
A. Pullini, D. Rossi, I. Loi, G. Tagliavini and L. Benini, "Mr.Wolf: An Energy-Precision Scalable Parallel Ultra Low Power SoC for IoT Edge Processing," in IEEE Journal of Solid-State Circuits, vol. 54, no. 7, pp. 1970-1981, July 2019.

DMA for data transfers from/to L2



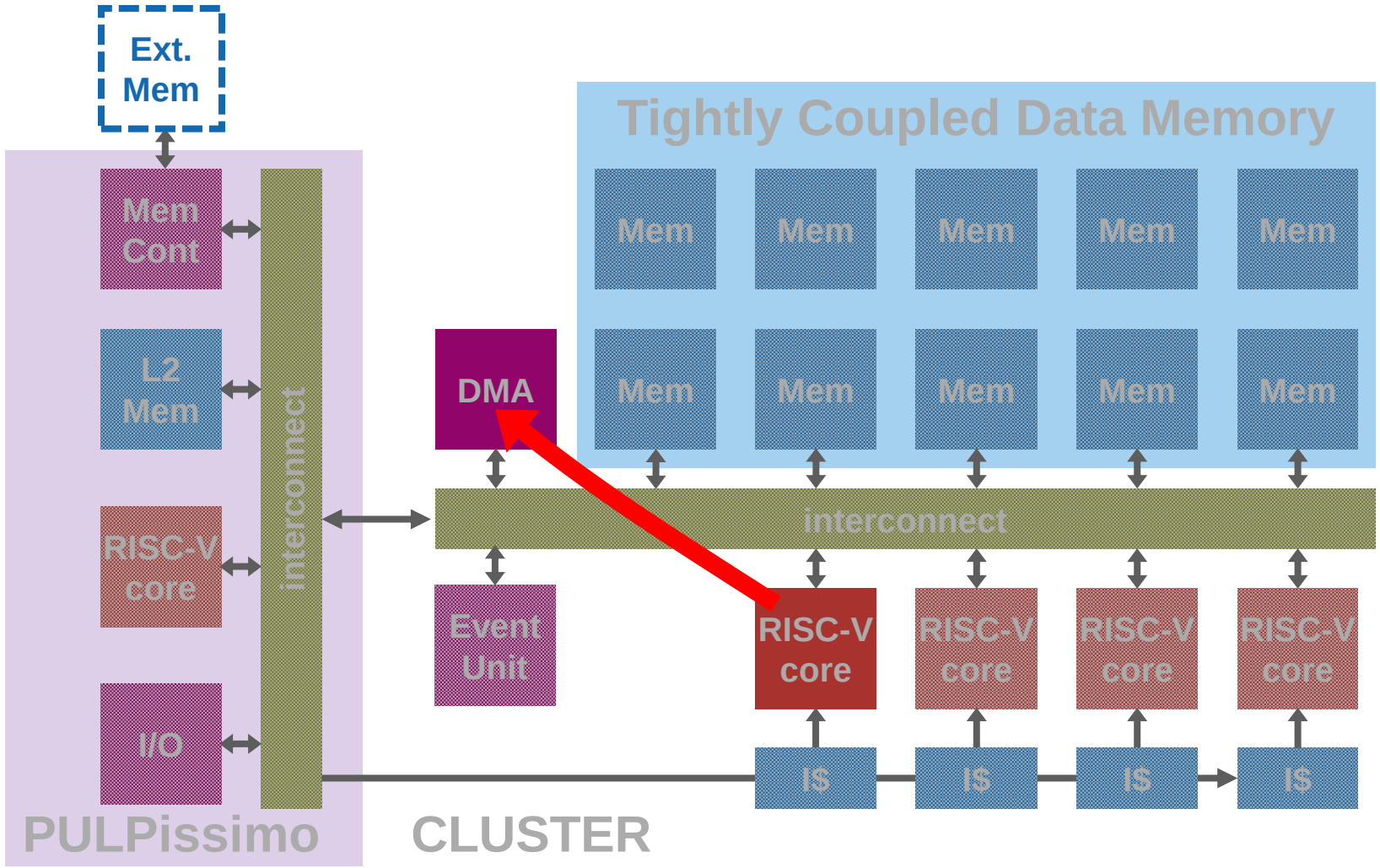
D. Rossi, I. Loi, G. Haugou, and L. Benini. 2014. Ultra-low-latency lightweight DMA for tightly coupled multi-core clusters. In Proceedings of the 11th ACM Conference on Computing Frontiers (CF '14).

An additional I/O controller is used for IO

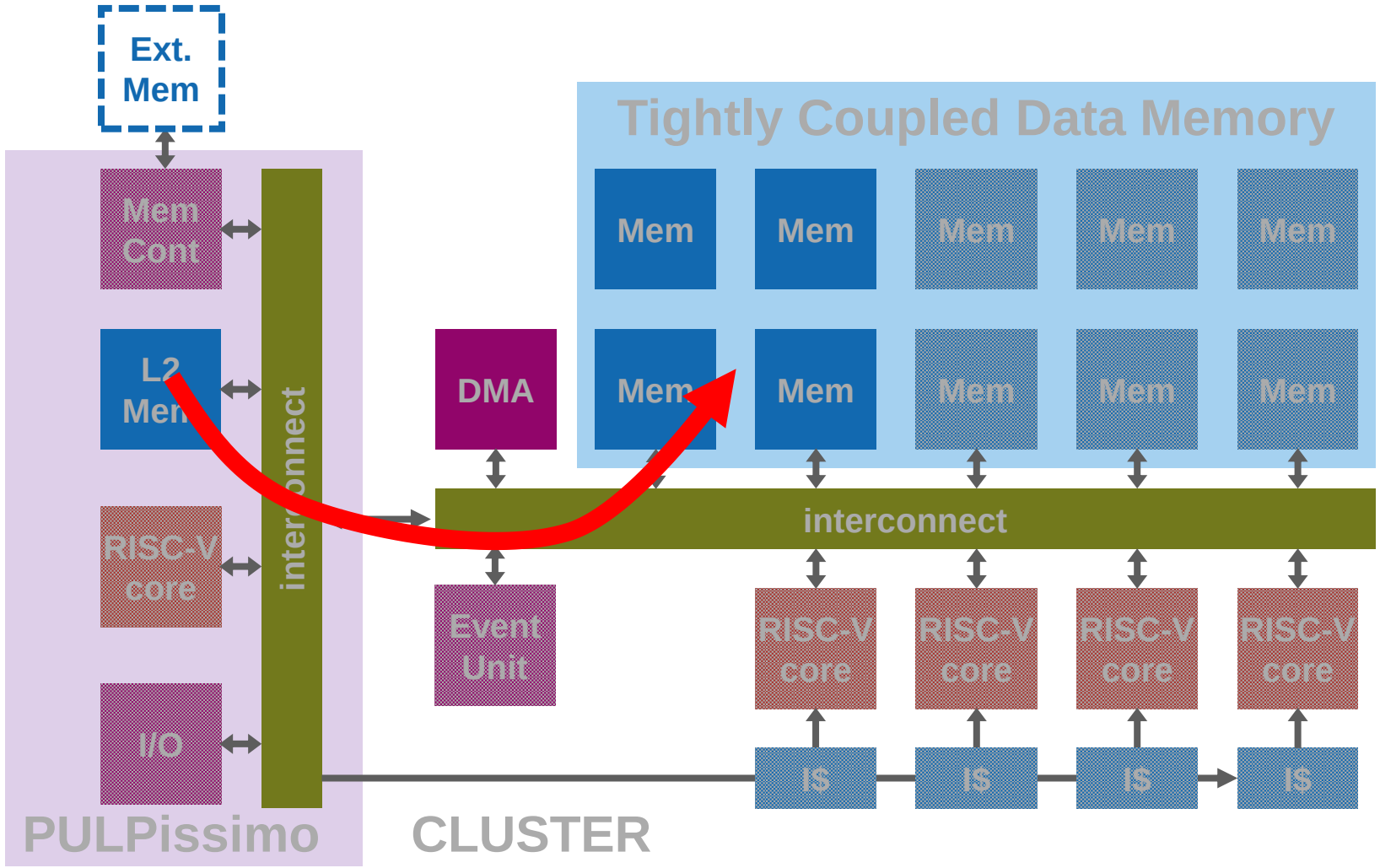


A. Pullini, D. Rossi, G. Haugou and L. Benini, "µDMA: An autonomous I/O subsystem for IoT end-nodes," *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, 2017.

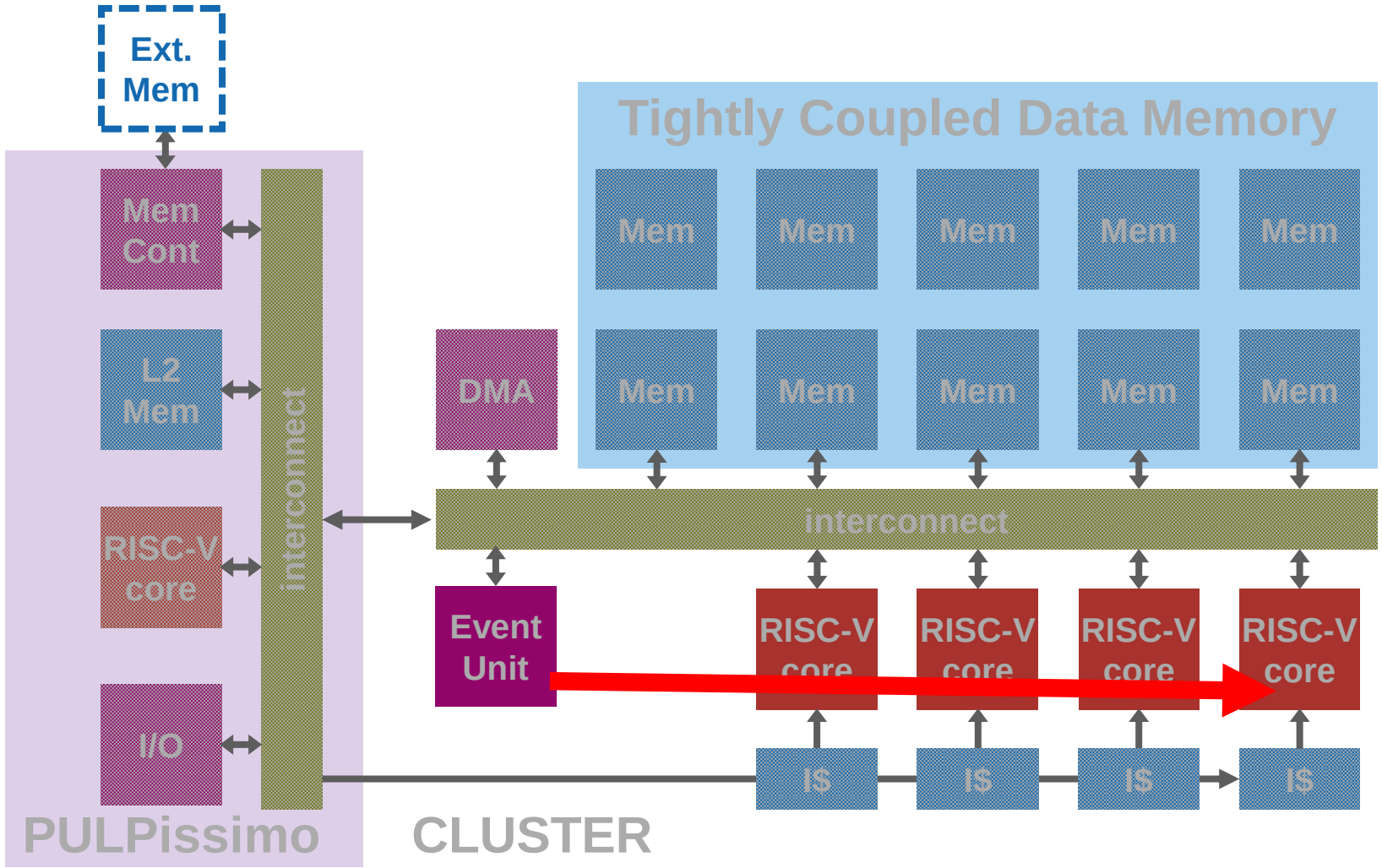
How do we work: Initiate a DMA transfer



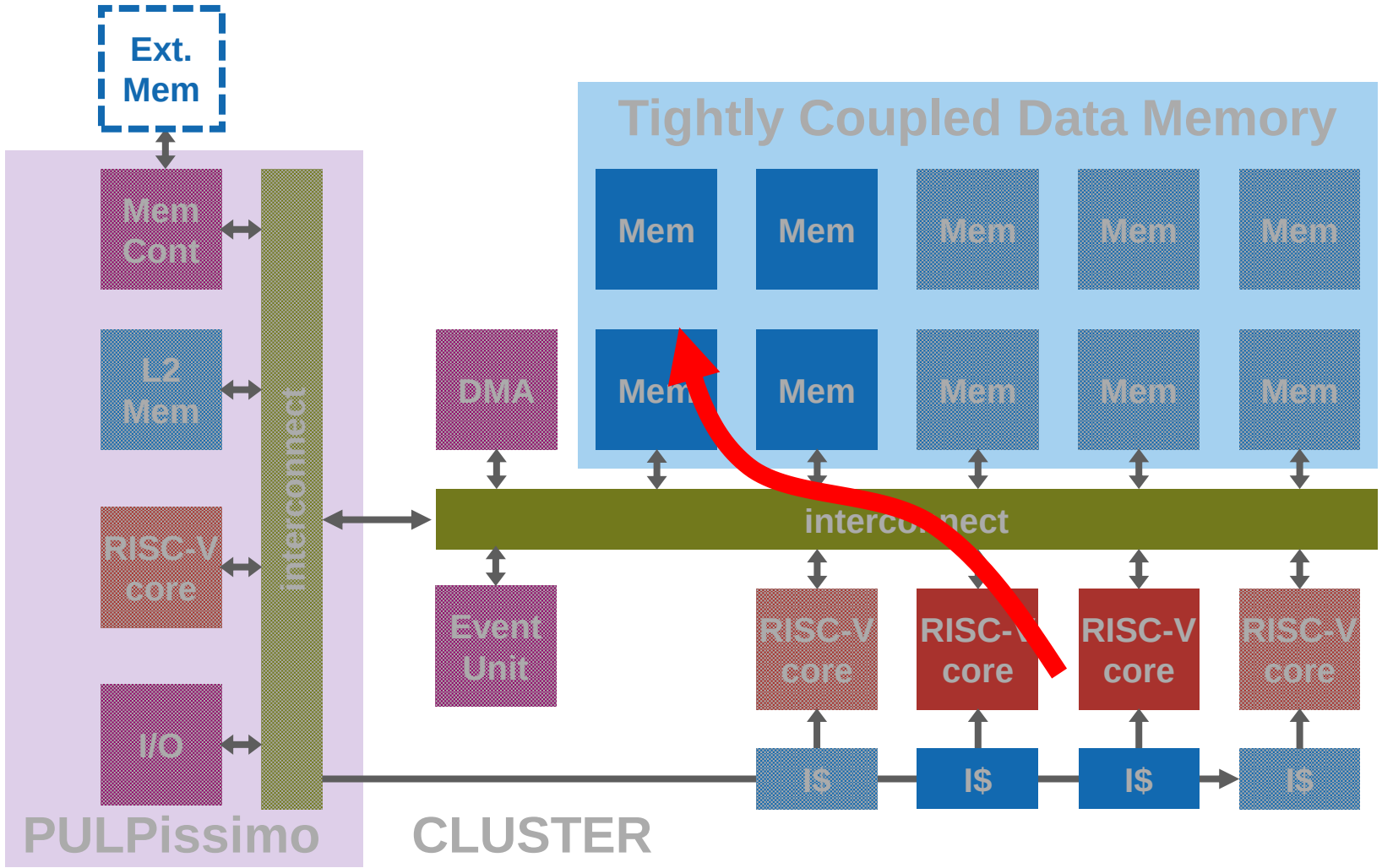
Data copied from L2 into TCDM



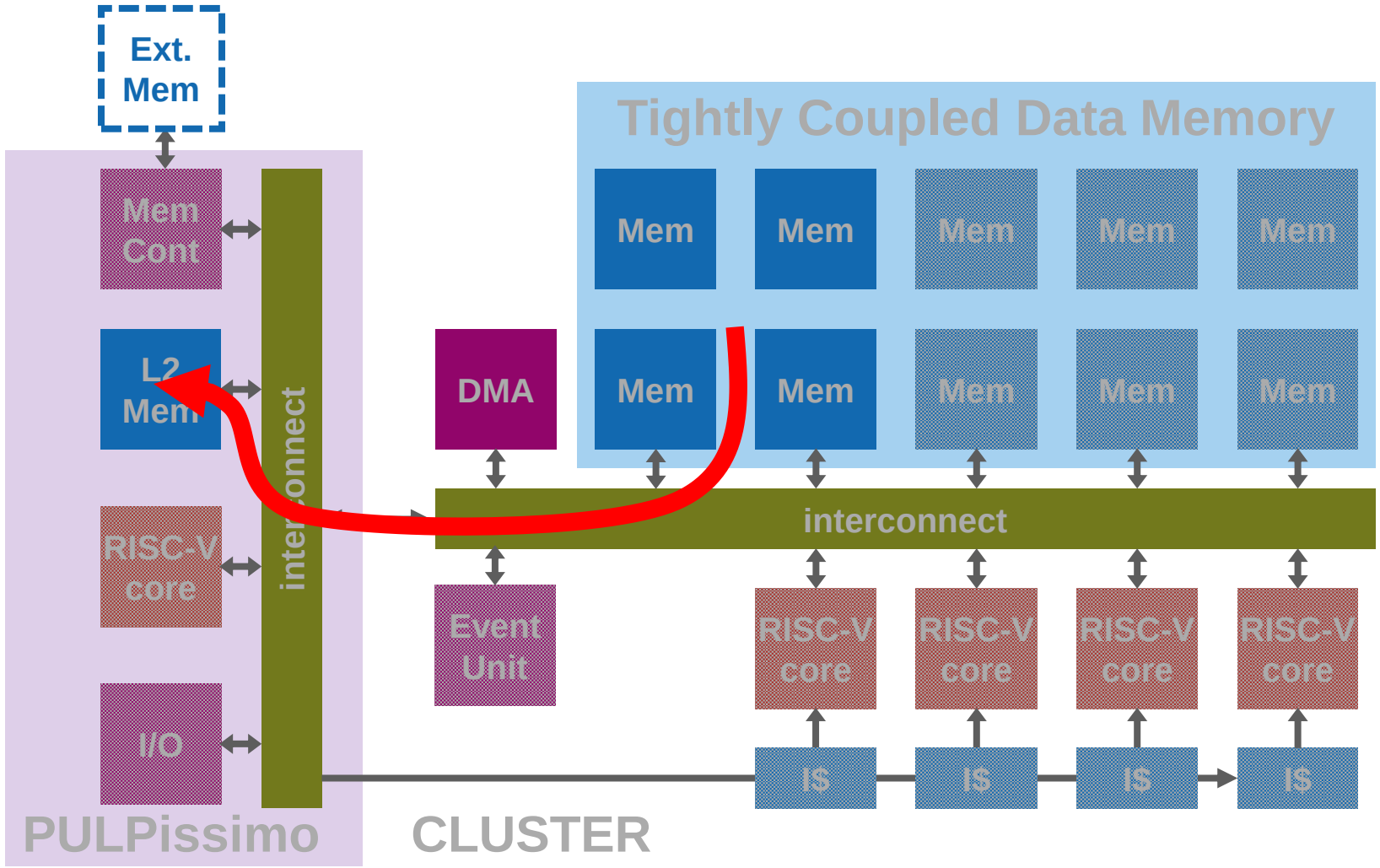
Once data is transferred, event unit notifies cores



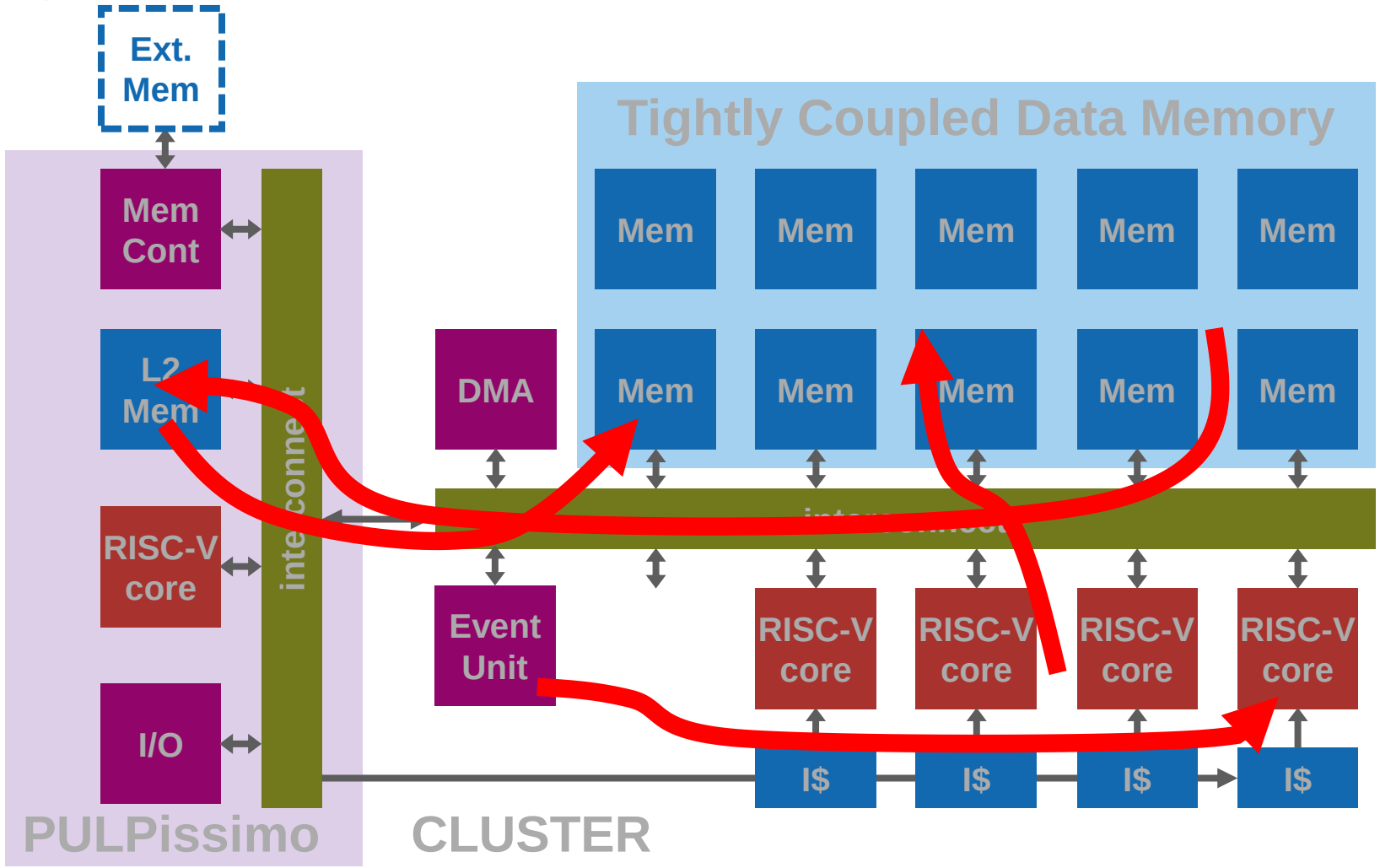
Cores can work on the data transferred



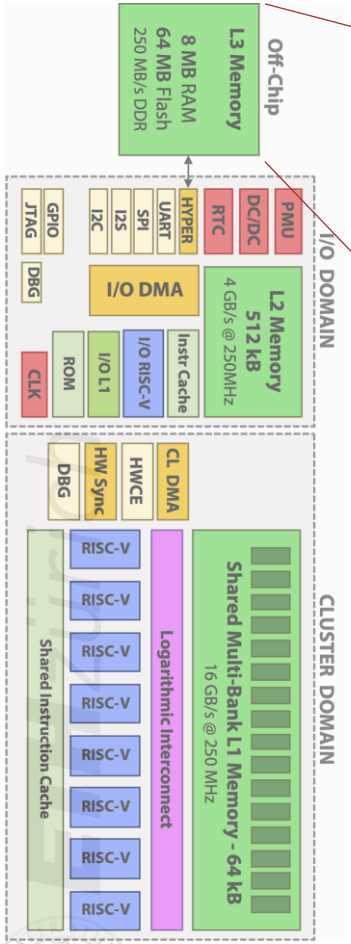
Once our work is done, DMA copies data back



During normal operation all of these occur concurrently



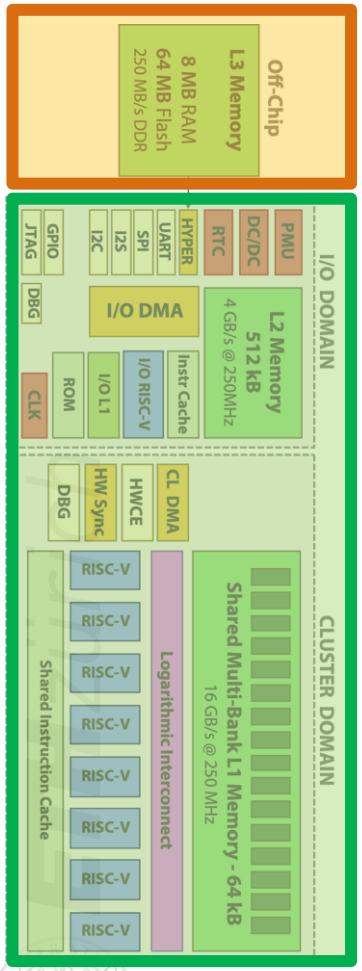
Explicit Memory Management: MobileNet Example



1.0-MobileNet-128
(59% top-1 accuracy on ImageNet)

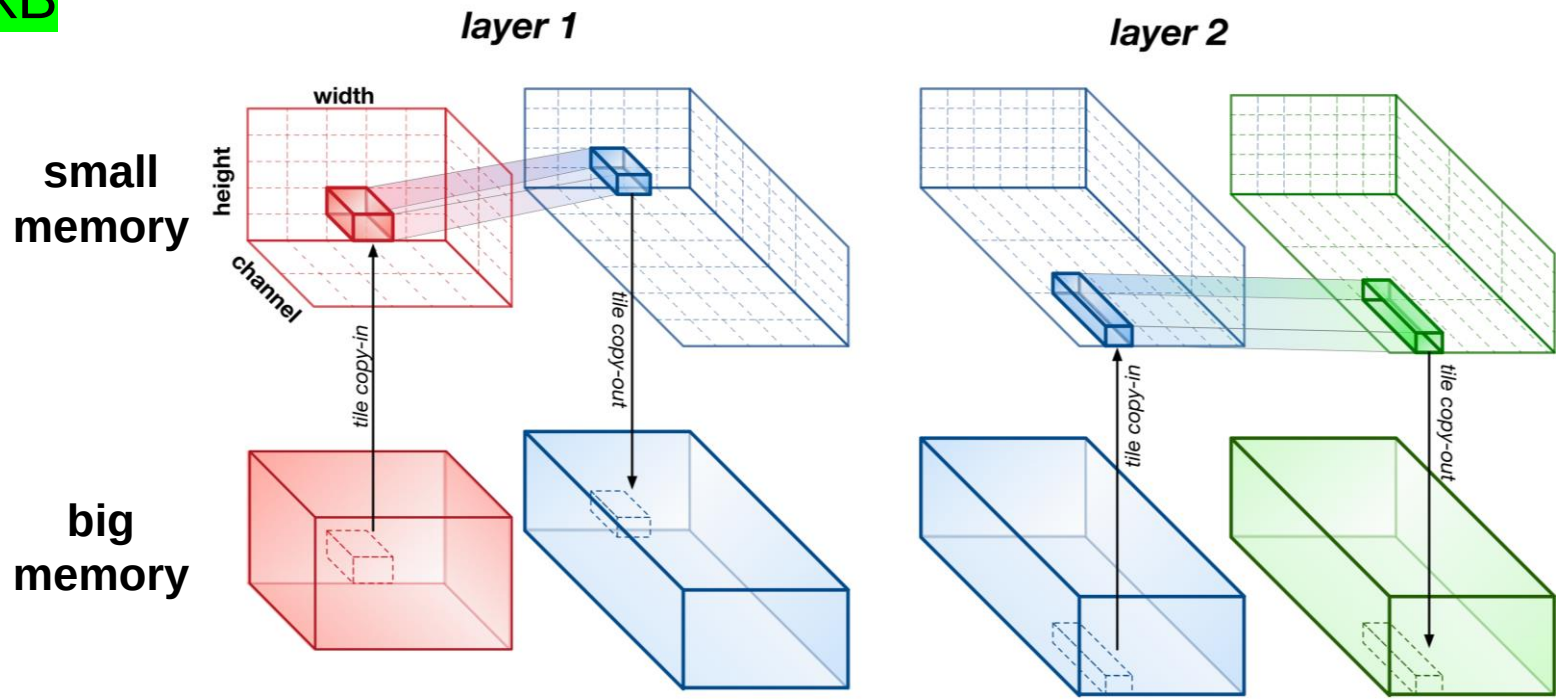
- ~4 Mparameters →need to store weights in off-chip memory (L3)
- L1 Bandwidth: 256 Gbit/s @ 250 MHz
- L2 Bandwidth: 32 Gbit/s @ 250 MHz
- L3 Bandwidth: 1.6 Gbit/s @ 100 MHz

Tensor tiling

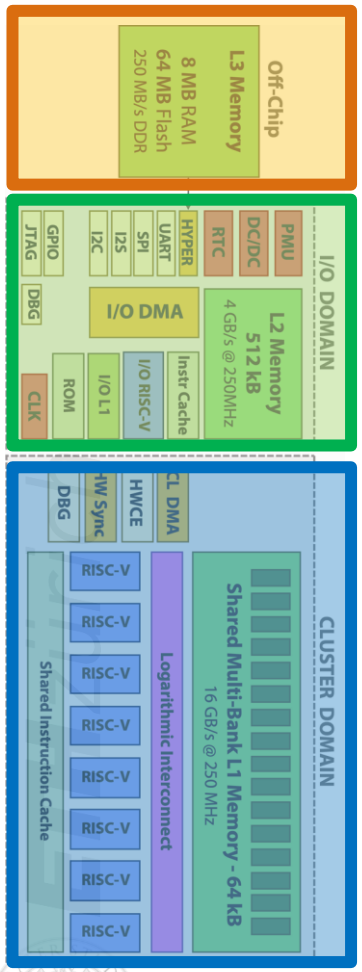


L3 / **L2** tiling

64 MB / 512 kB

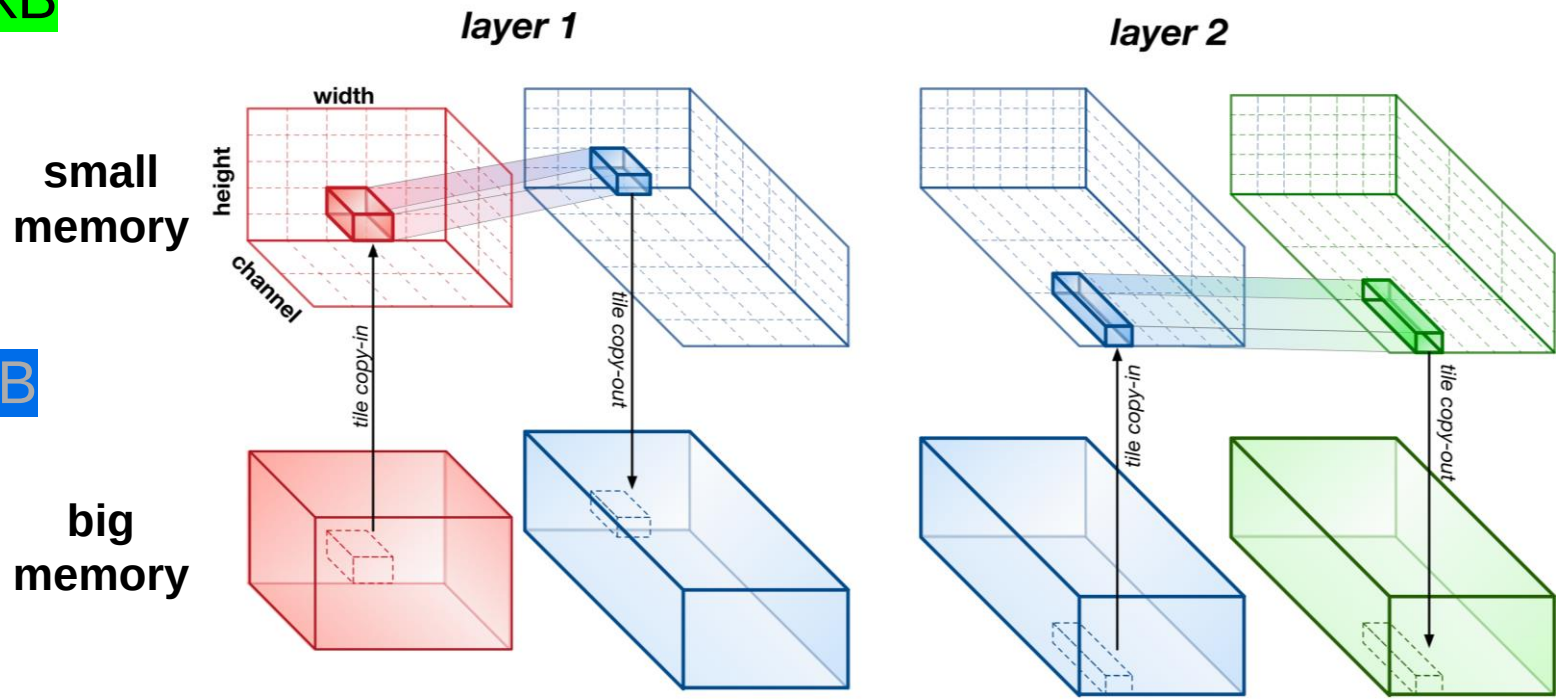


Tensor tiling

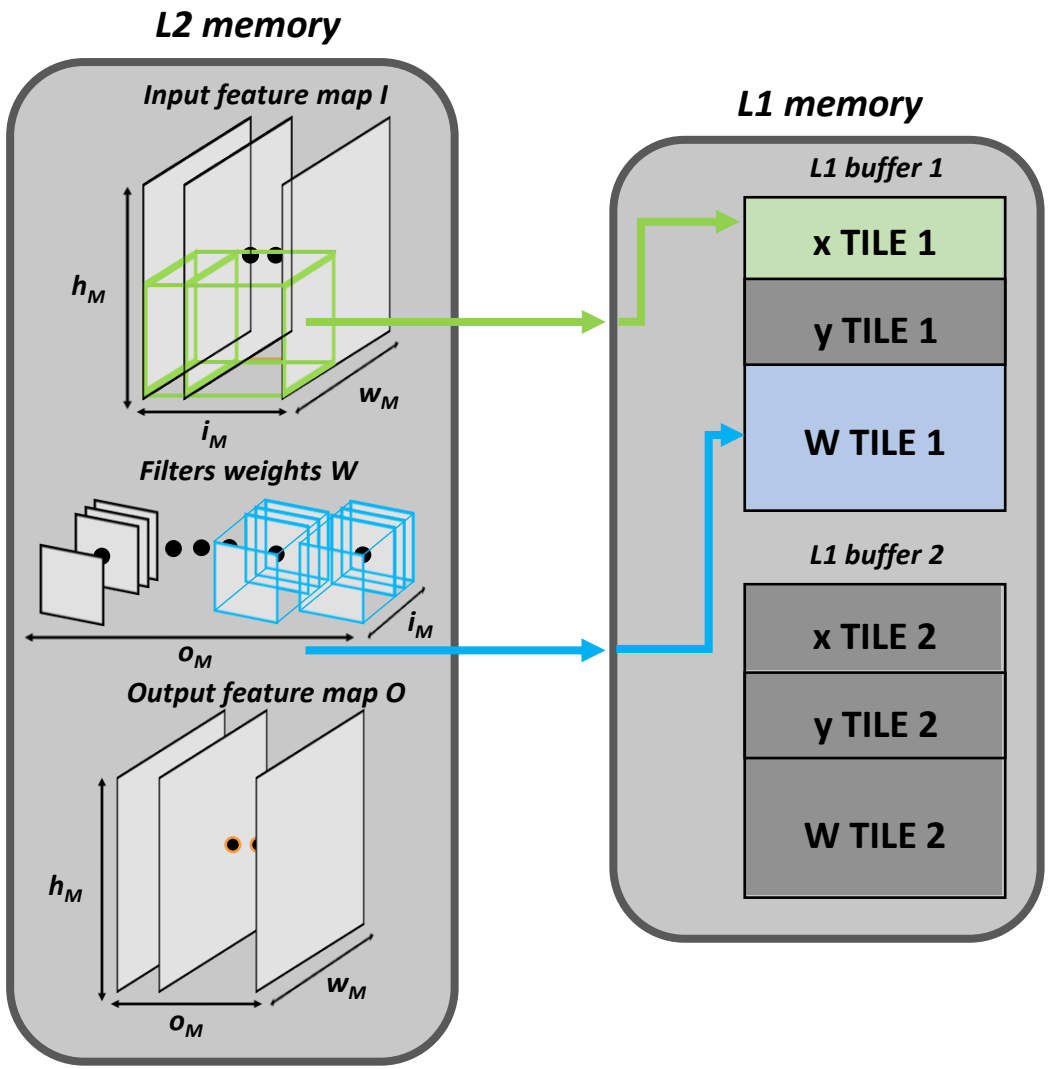


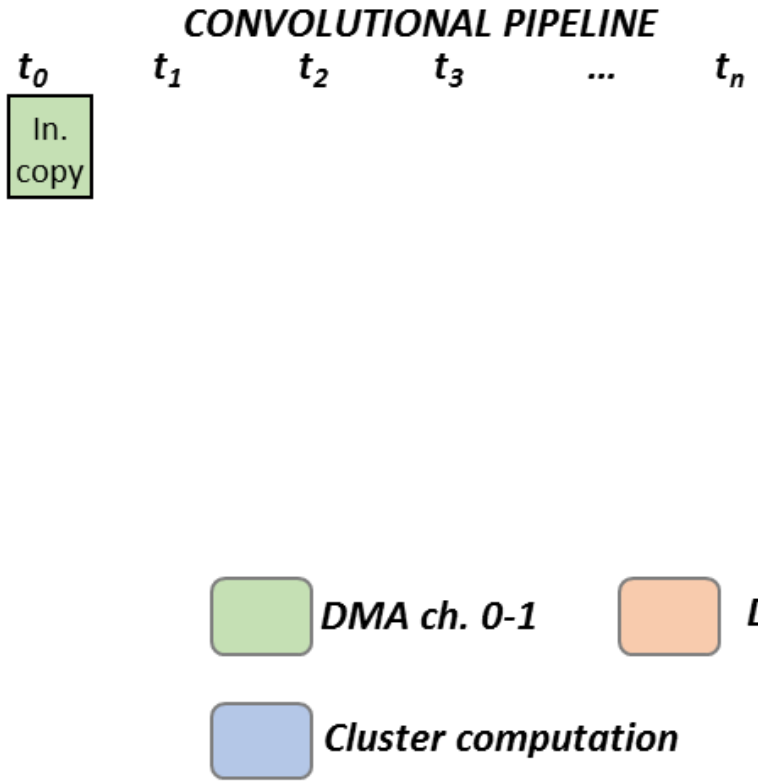
L3 / L2 tiling
64 MB / 512 kB

L2 / L1 tiling
512 kB / 64 kB

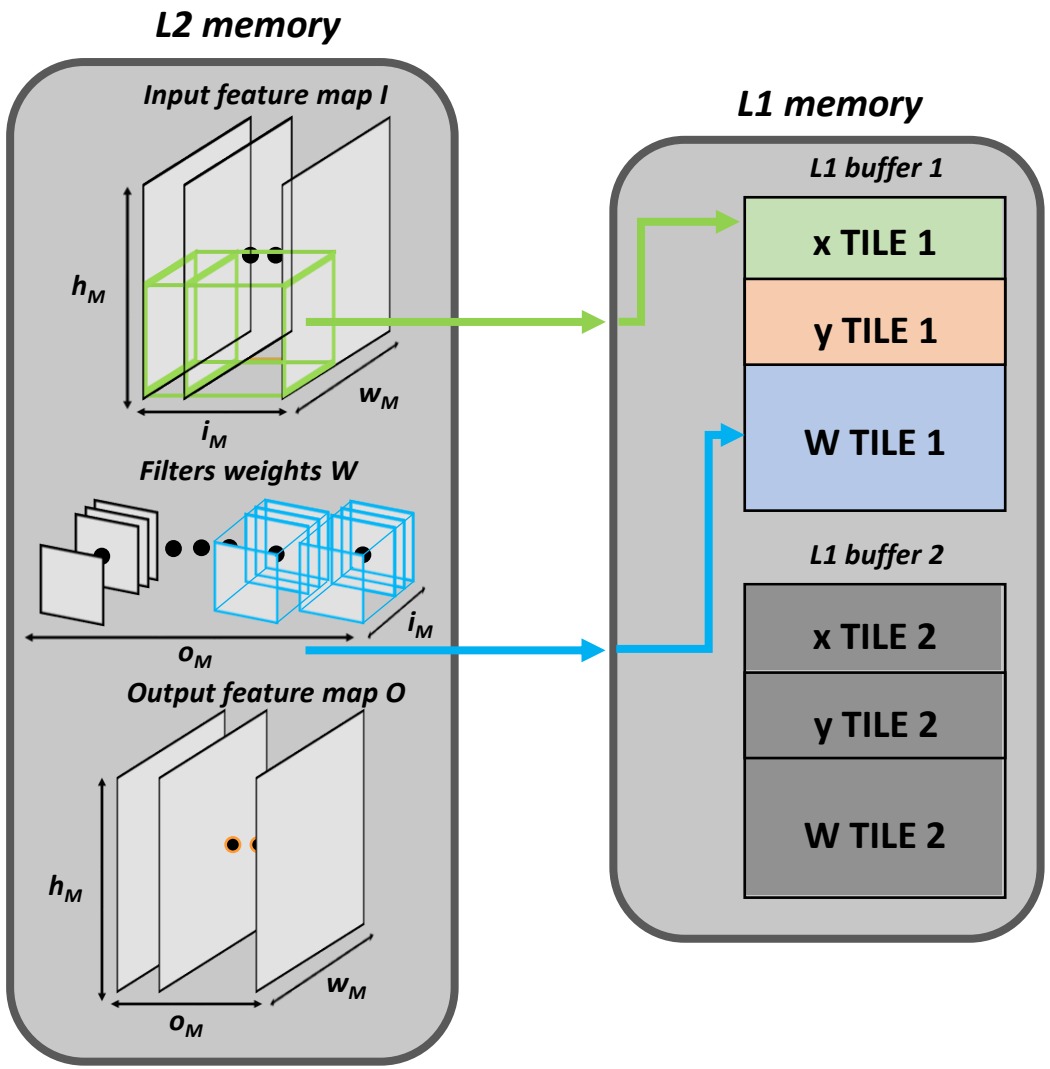


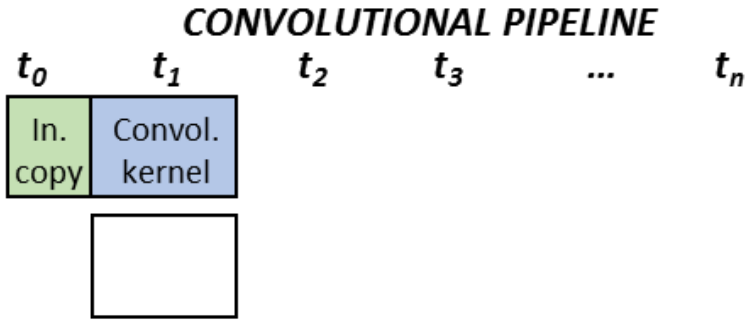
Tile Data Movement





Tile Data Movement



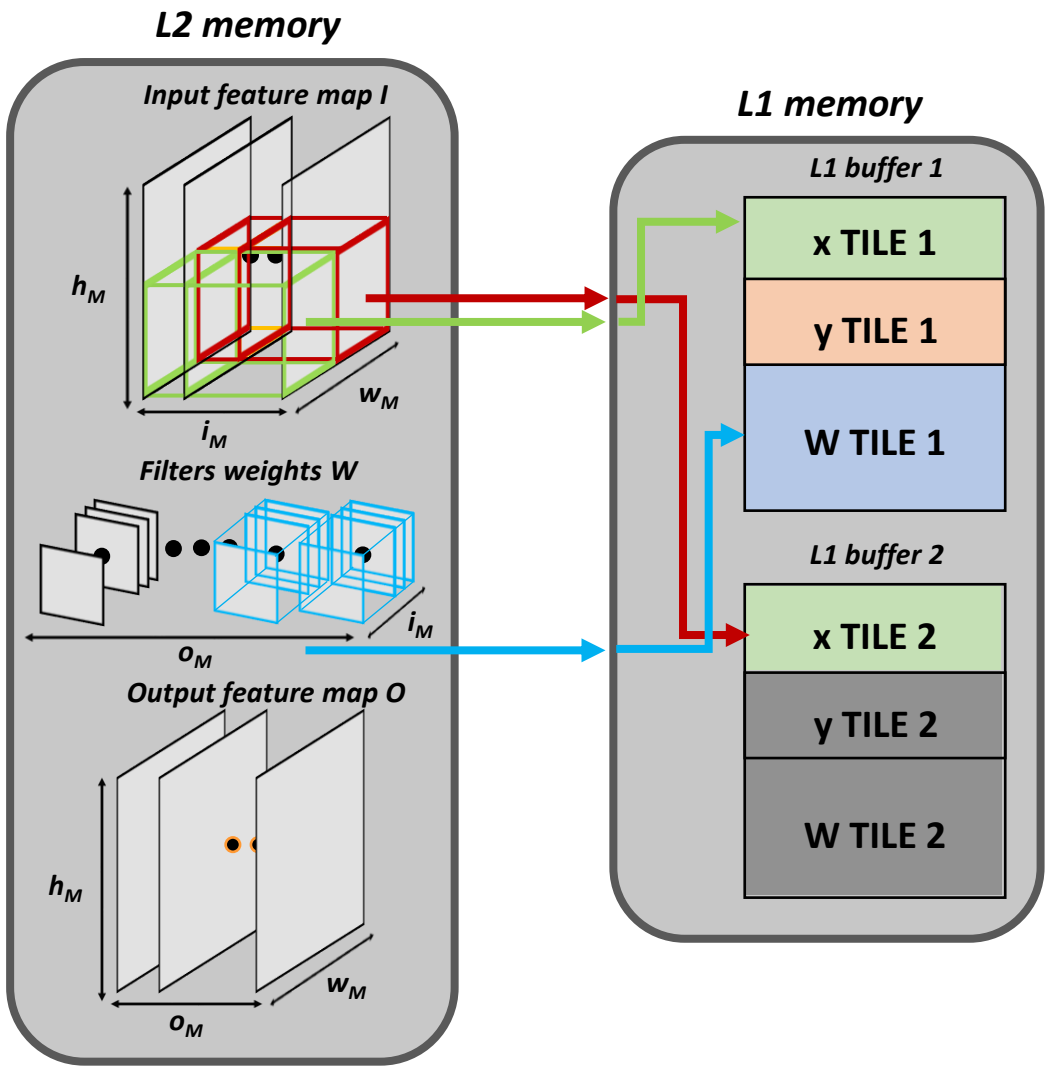


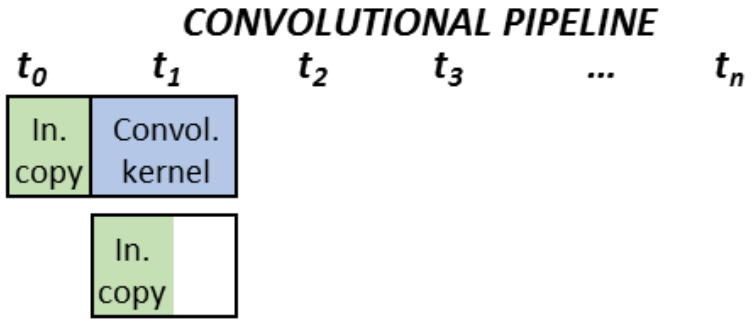
DMA ch. 0-1

DMA ch. 2

Cluster computation

Tile Data Movement



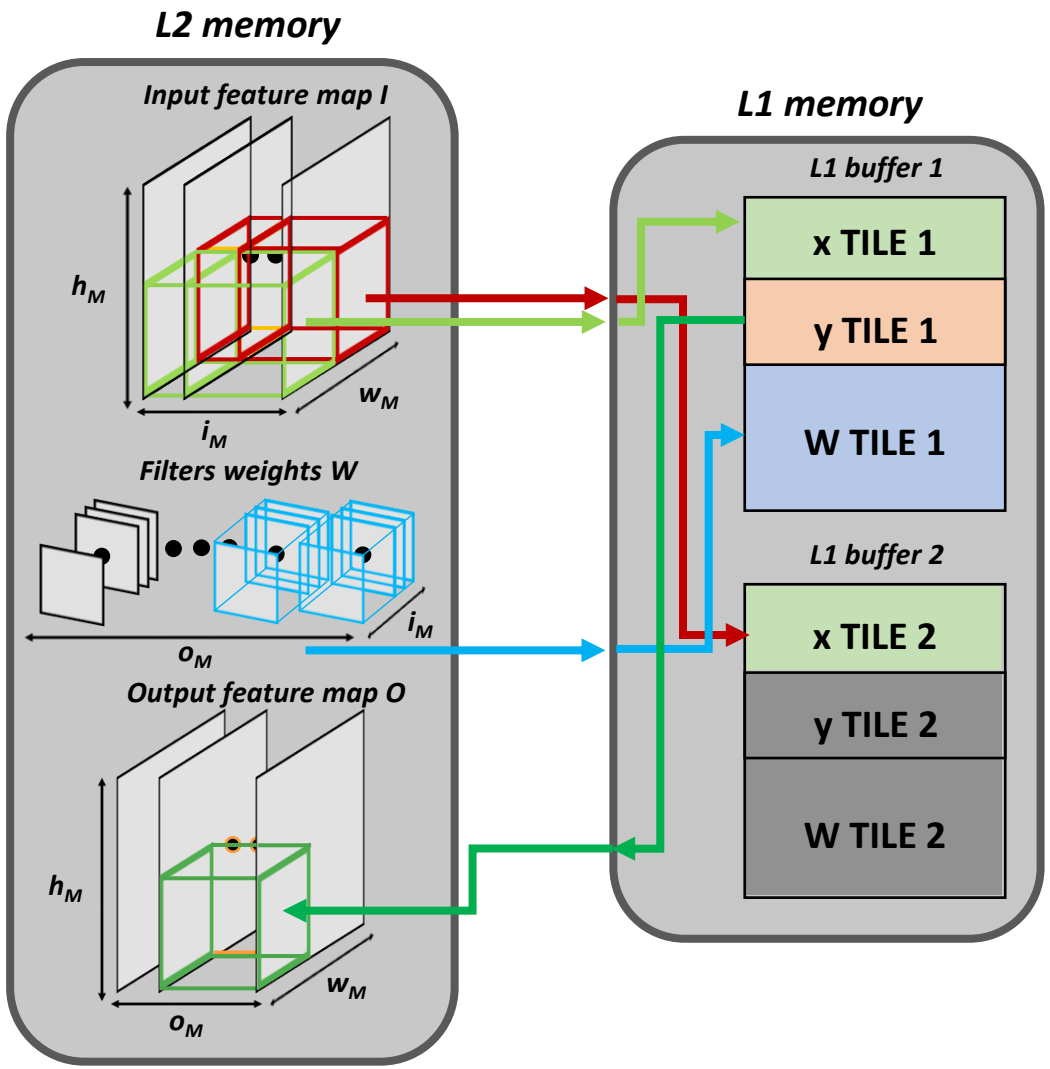


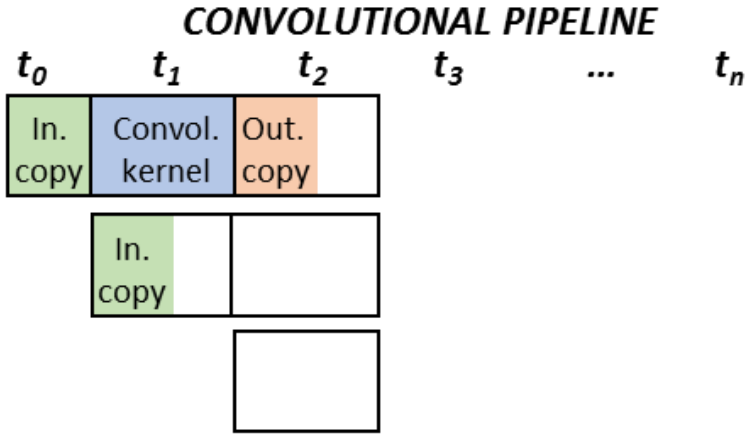
DMA ch. 0-1

DMA ch. 2

Cluster computation

Tile Data Movement



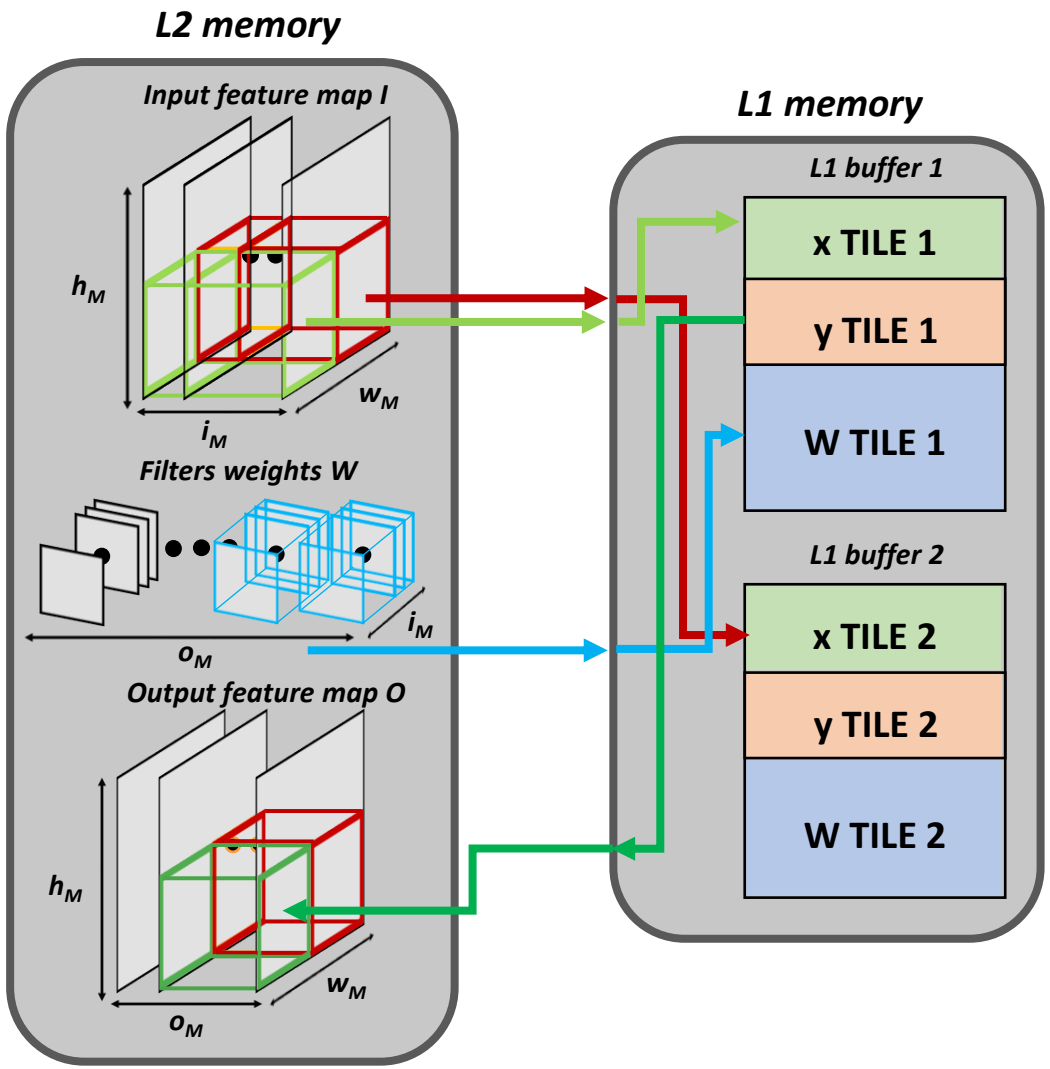


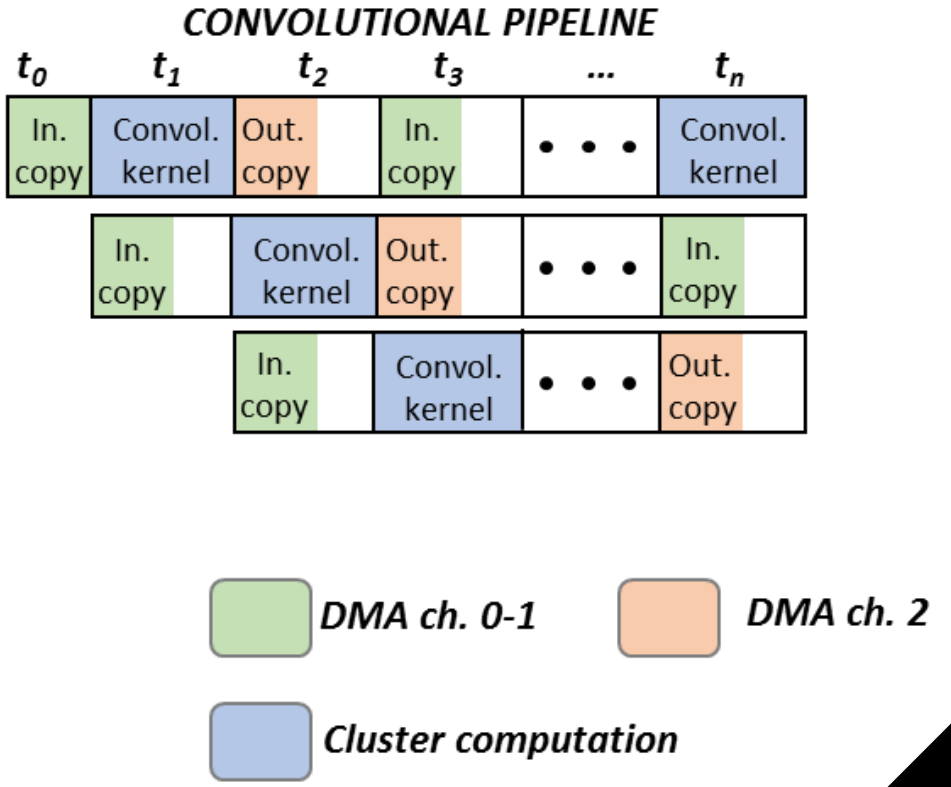
DMA ch. 0-1

DMA ch. 2

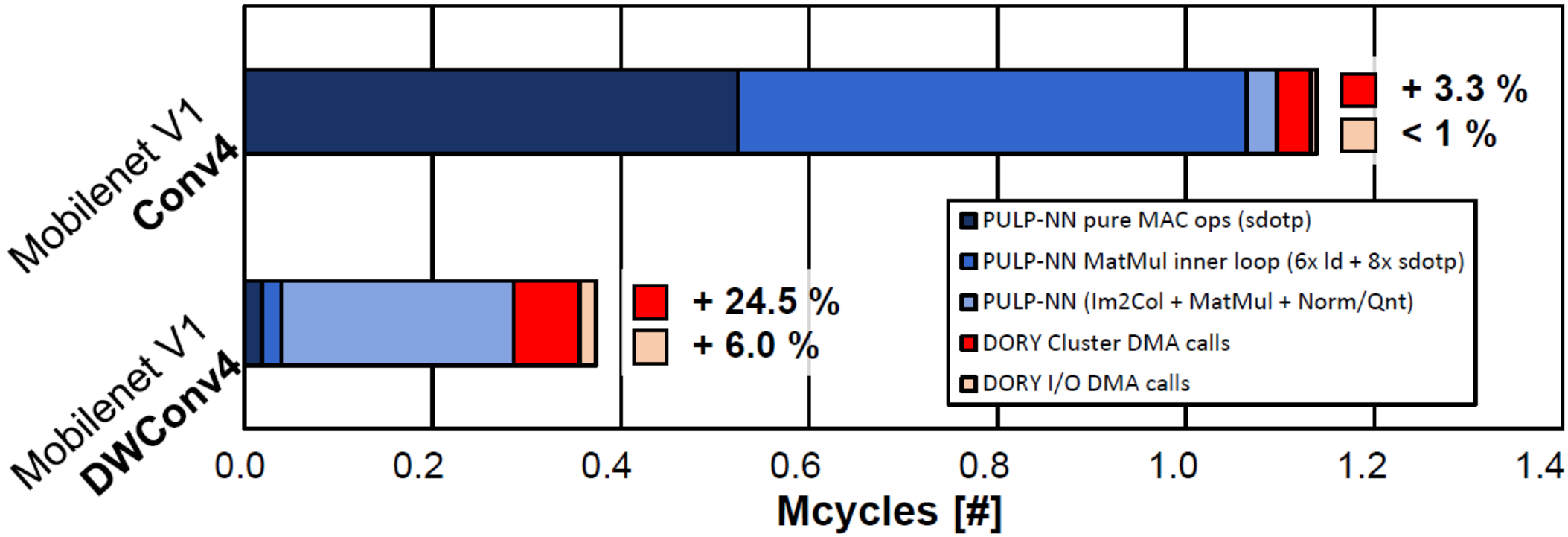
Cluster computation

Tile Data Movement





Tiling Overhead on MobilenenetV1



A. Burrello, A. Garofalo, N. Bruschi, G. Tagliavini, D. Rossi and F. Conti, "DORY: Automatic End-to-End Deployment of Real-World DNNs on Low-Cost IoT MCUs," in *IEEE Transactions on Computers*, vol. 70, no. 8, pp. 1253-1268, 1 Aug. 2021.

PULP includes Cores+Interco+IO+HWCE → Open Platform

RISC-V Cores

- RI5CY 32b
- Ibex 32b
- Snitch 32b
- Ariane + Ara 64b

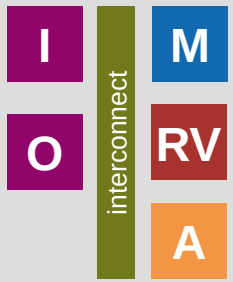
Peripherals

- JTAG
- SPI
- UART
- I2S
- DMA
- GPIO

Interconnect

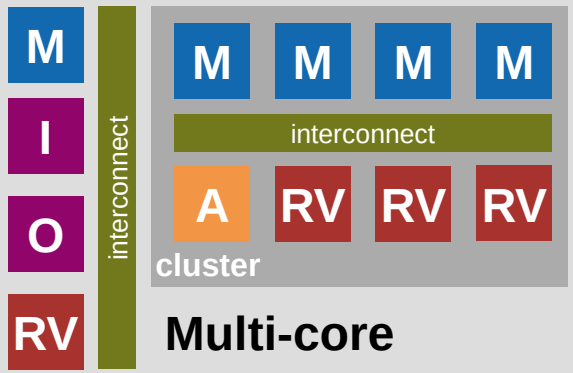
- Logarithmic interconnect
- APB – Peripheral Bus
- AXI4 – Interconnect

Platforms



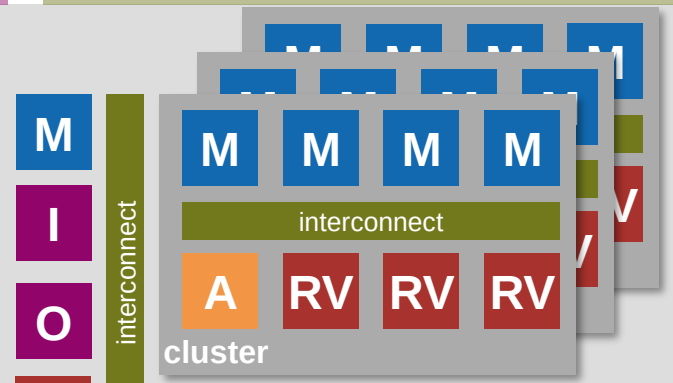
Single Core

- PULPino
- PULPissimo



Multi-core

- Fulmine
- Mr. Wolf



Multi-cluster

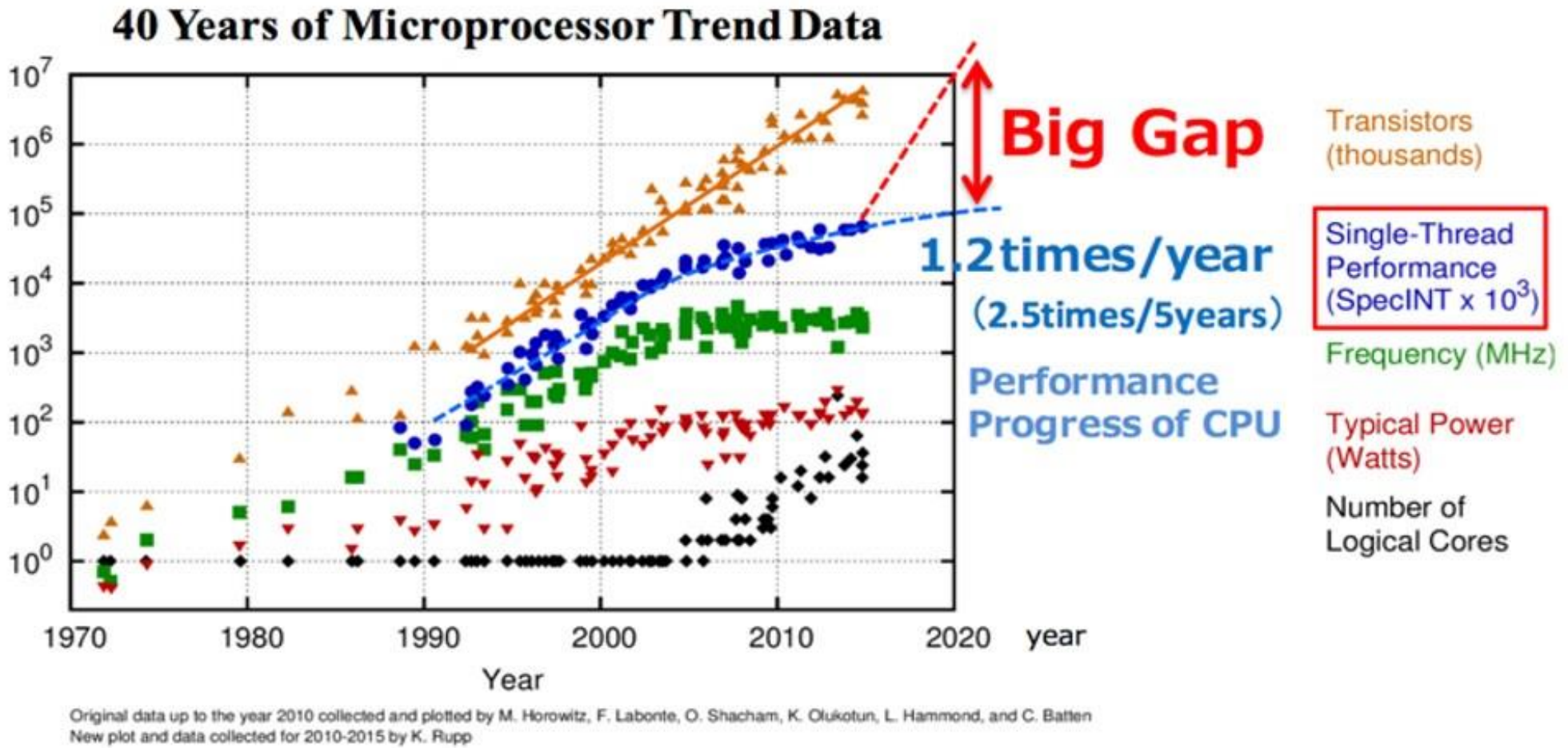
- Hero
- Manticore

IOT → HPC

Accelerators

- HWCE (convolution)
- Neurostream (ML)
- HWCrypt (crypto)
- PULPO (1st ord. opt)

From Edge to Cloud: The Communication Wall



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Ref.: Ahmet Ceyhan, INTERCONNECTS FOR FUTURE TECHNOLOGY GENERATIONS—CONVENTIONAL CMOS WITH COPPER/LOW-κ AND BEYOND, Fig 2, 9