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

Open source HW solutions for EdgeAI

PULP platform in action

European Nanoelectronic Applications, Design & Technology Conference 2021

Frank K. Gürkaynak
<kgf@iis.ee.ethz.ch>

http://pulp-platform.org @pulp_platform

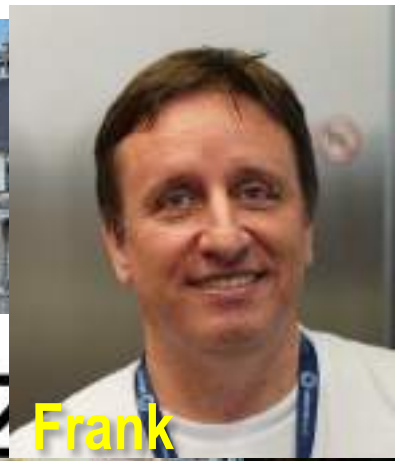
ETH zürich

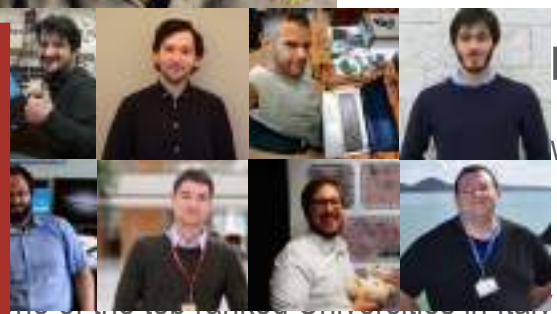
Parallel Ultra Low Power project in a nutshell

- Started in 2013, after Luca joined ETH Zürich
- We wanted to design energy efficient computing systems
  - Equally efficient for IoT and HPC over a wide range
- Key points
  - Parallel processing
  - Near threshold computing
  - Efficient switching between operating modes
  - Making best use of technology
  - Heterogeneous acceleration
  - And **open source** using a permissive license

Build programmable devices that can do more per Joule

Who is behind PULP?





In total about 60 people work on projects related to PULP in Zurich and Bologna

https://pulp-platform.org/team.html

STUDIORUM DI BOLOGNA

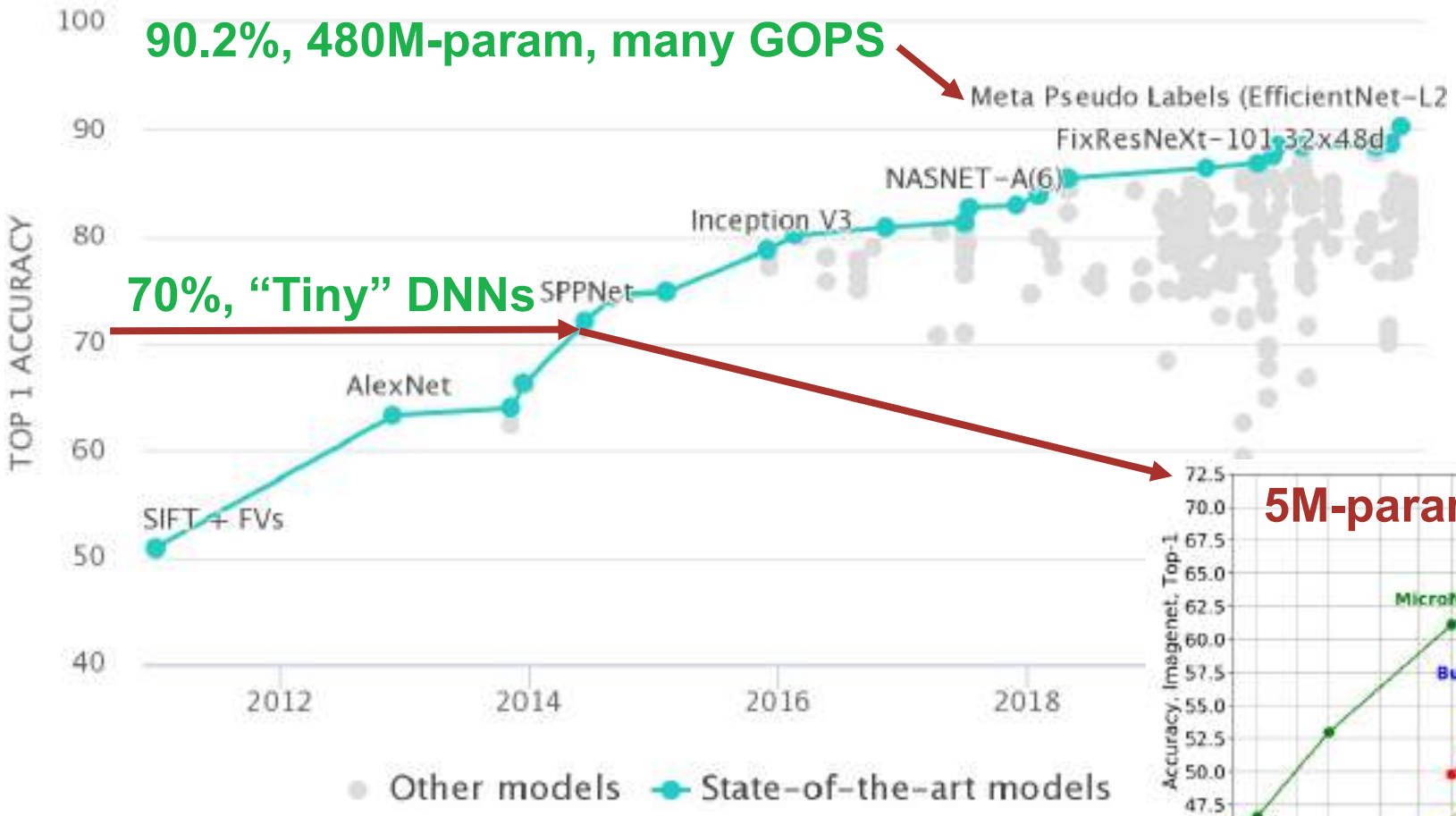
na

world

and

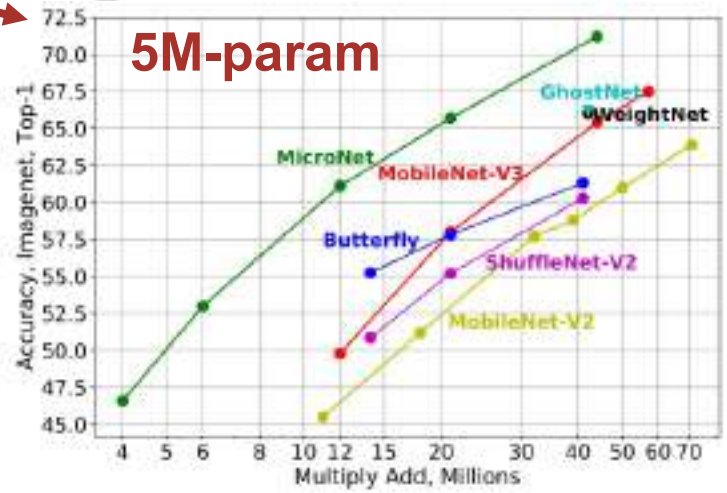
urope

AI: where GOPS are needed for better results



High OP/B ratio
Massive Parallelism
MAC-dominated
Low precision OK

Model redundancy



Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, Quoc V. Le, "Meta Pseudo Labels", arXiv:2003.10580 [cs.LG]

The more we can do on the EDGE the better

- **Less energy used for communication**
  - Major contribution to energy is the communication overhead.
- **Latency / independence**
  - Communication to cloud brings issues in getting responses in time, What happens if communication is disturbed?
- **More privacy & security**
  - Acquiring and consuming private data locally reduces attack surface

- **Question is how much can we afford to do on EDGE**

This is how we get more out of PULP

AI related optimizations

- Energy efficient RISC-V core (20 pJ/op - 8 bit)
- ISA extensions for DSP (1-2 pJ/op – 8 bit) XPULP
- Configurable datapath (50-100 fJ/op – 4 bit) RBE, NE
- Fully specialized datapath (5-10 fJ/op – ternary) XNE

And there is more

- Efficient parallelization, Smart wakeup, Better I/O

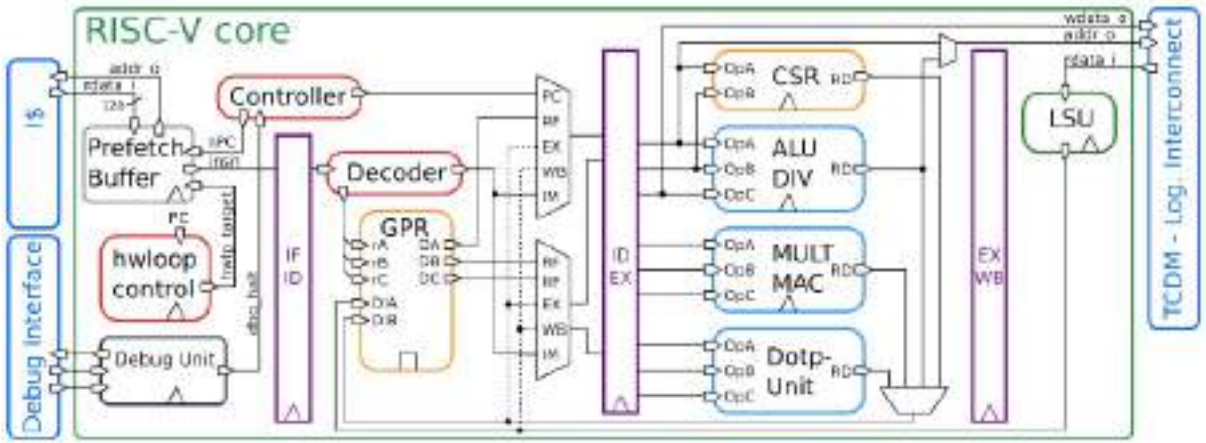
ETH zürich

Numbers using GF22FDX, Near Threshold @0.6V, High core utilization, minimal I/O & overhead

RI5CY* – An Open MCU-class RISC-V Core for EE-AI

** Now called CV32E40P by OpenHW group*

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

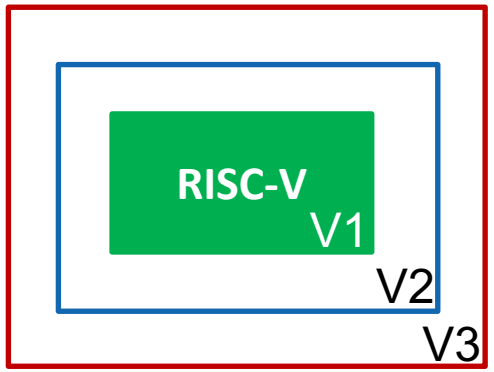


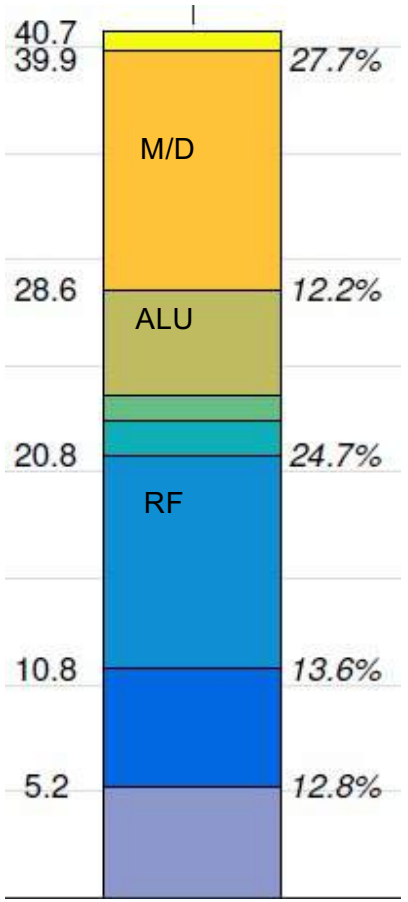
RISC-V ISA is extensible *by construction* (great!)

V1 Baseline RV32IMC **(not so good for ML)**
HW loops

V2 Post modified Load/Store
MAC operations
SIMD 2/4 + Dot Product + Shuffling

V3 Bit manipulation unit
Lightweight fixed-point support





40 kGE (+60%)

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, doi: 10.1109/TVLSI.2017.2654506.

PULP-NN: Xpulp ISA exploitation

RV32IMC

8-bit Convolution

N

```
addi  a0,a0,1
addi  t1,t1,1
addi  t3,t3,1
addi  t4,t4,1
lbu   a7,-1(a0)
lbu   a6,-1(t4)
lbu   a5,-1(t3)
lbu   t5,-1(t1)
mul   s1,a7,a6
mul   a7,a7,a5
add   s0,s0,s1
mul   a6,a6,t5
add   t0,t0,a7
mul   a5,a5,t5
add   t2,t2,a6
add   t6,t6,a5
bne   s5,a0,1c000bc
```

RV32IMCXpulp

N/4

```
lp.setup
p.lw  w1, 4(a0!)
p.lw  w2, 4(a1!)
p.lw   x1, 4(a2!)
p.lw   x2, 4(a3!)
pv.sdotsp.b    s1, w1, x1
pv.sdotsp.b    s2, w1, x2
pv.sdotsp.b    s3, w2, x1
pv.sdotsp.b    s4, w2, x2
end
```

HW Loop

LD/ST with post increment

8-bit SIMD sdotp

9x less instructions than RV32IMC

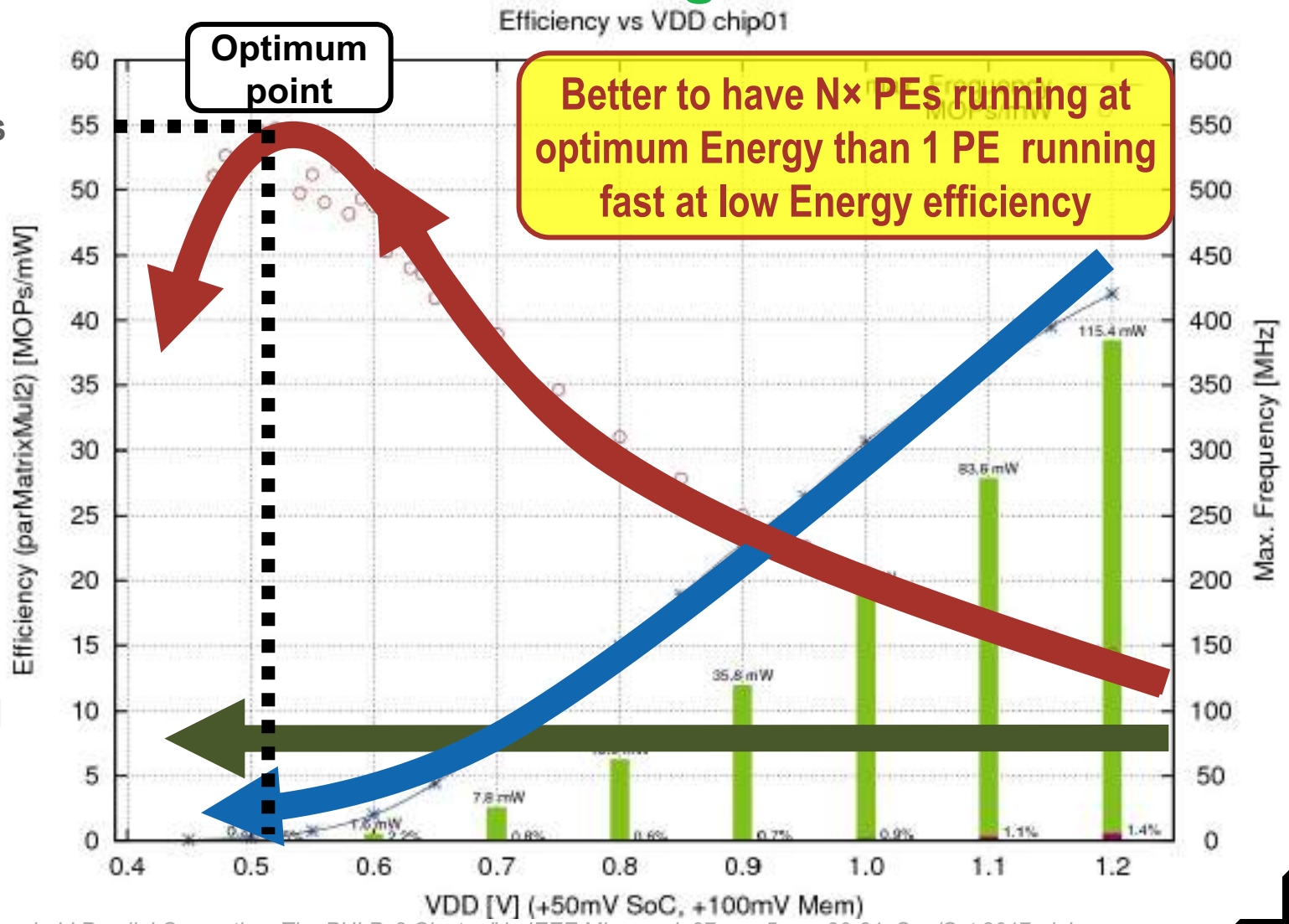
P↑ T↓↓↓ so, E=P×T↓↓ Nice!
But what about the GOPS?
Faster + Superscalar is not efficient!

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

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

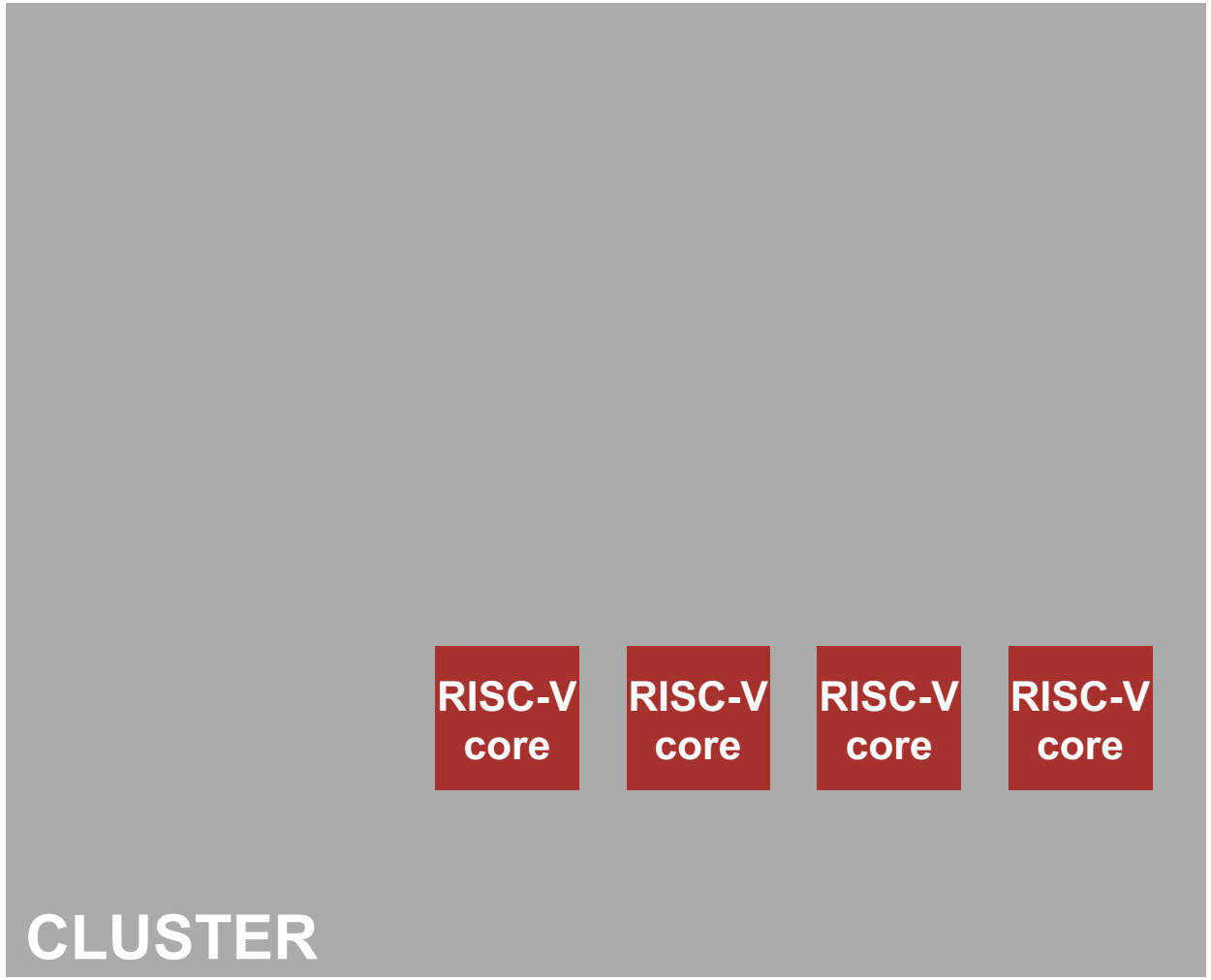
- As VDD decreases, operating speed decreases
- However efficiency increases→ more work done per Joule
- Until leakage effects start to dominate
- Put more units in parallel to get performance up and keep them busy with a parallel workload

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

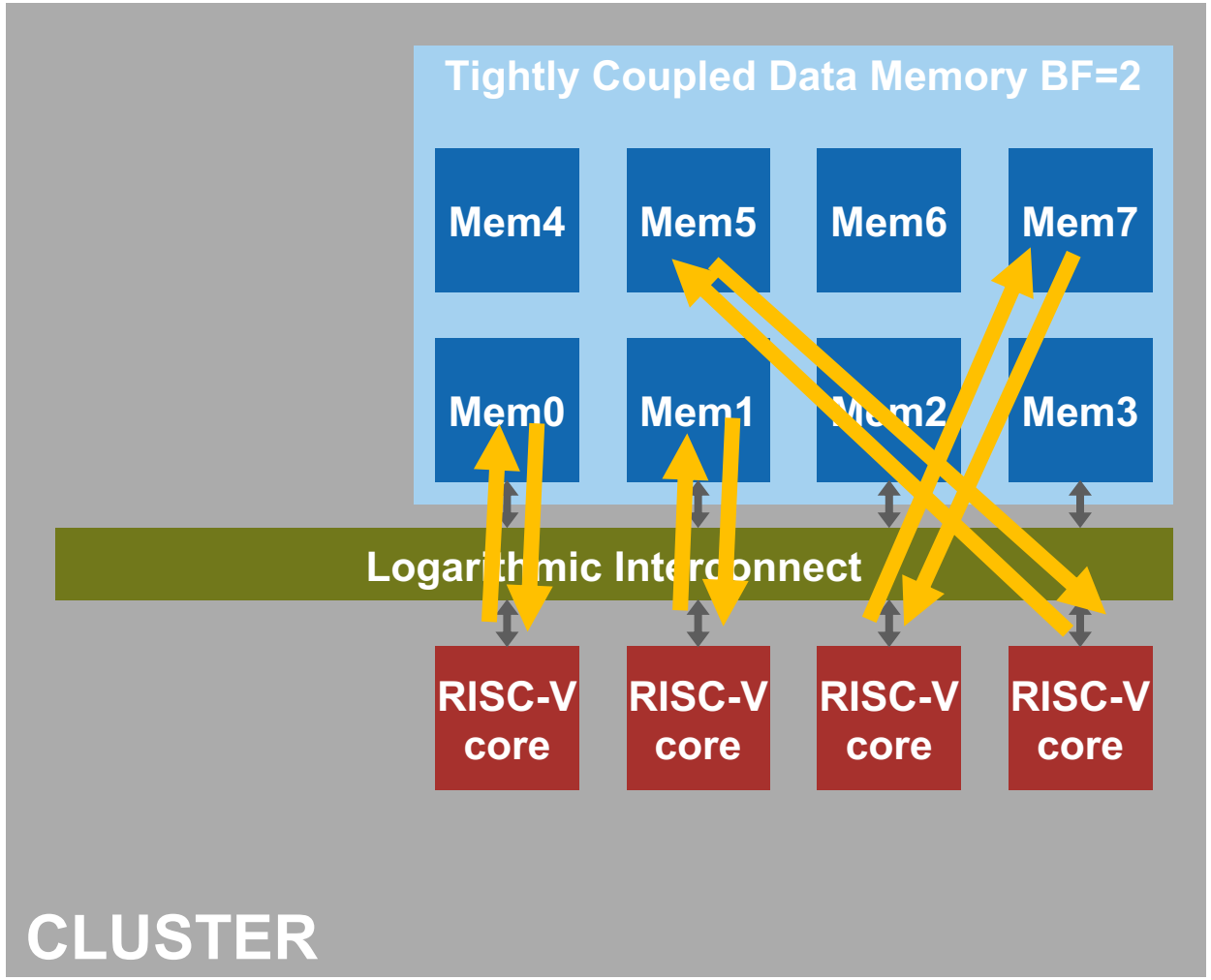


D. Rossi et al., "Energy-Efficient Near-Threshold Parallel Computing: The PULPv2 Cluster," in IEEE Micro, vol. 37, no. 5, pp. 20-31, Sep/Oct 2017, doi: 10.1109/MM.2017.3711645.

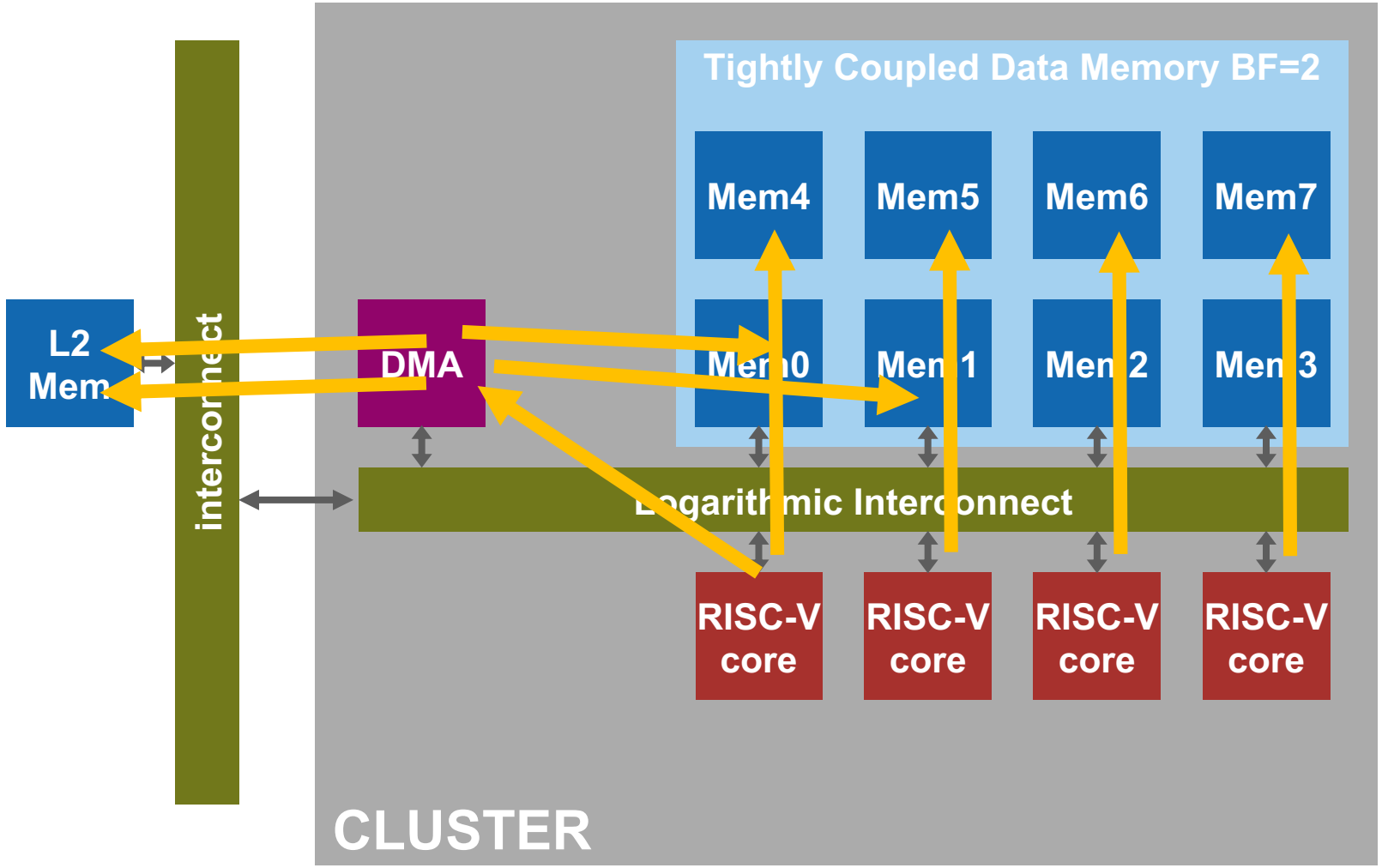
Multiple RI5CY Cores (1-16) in one cluster



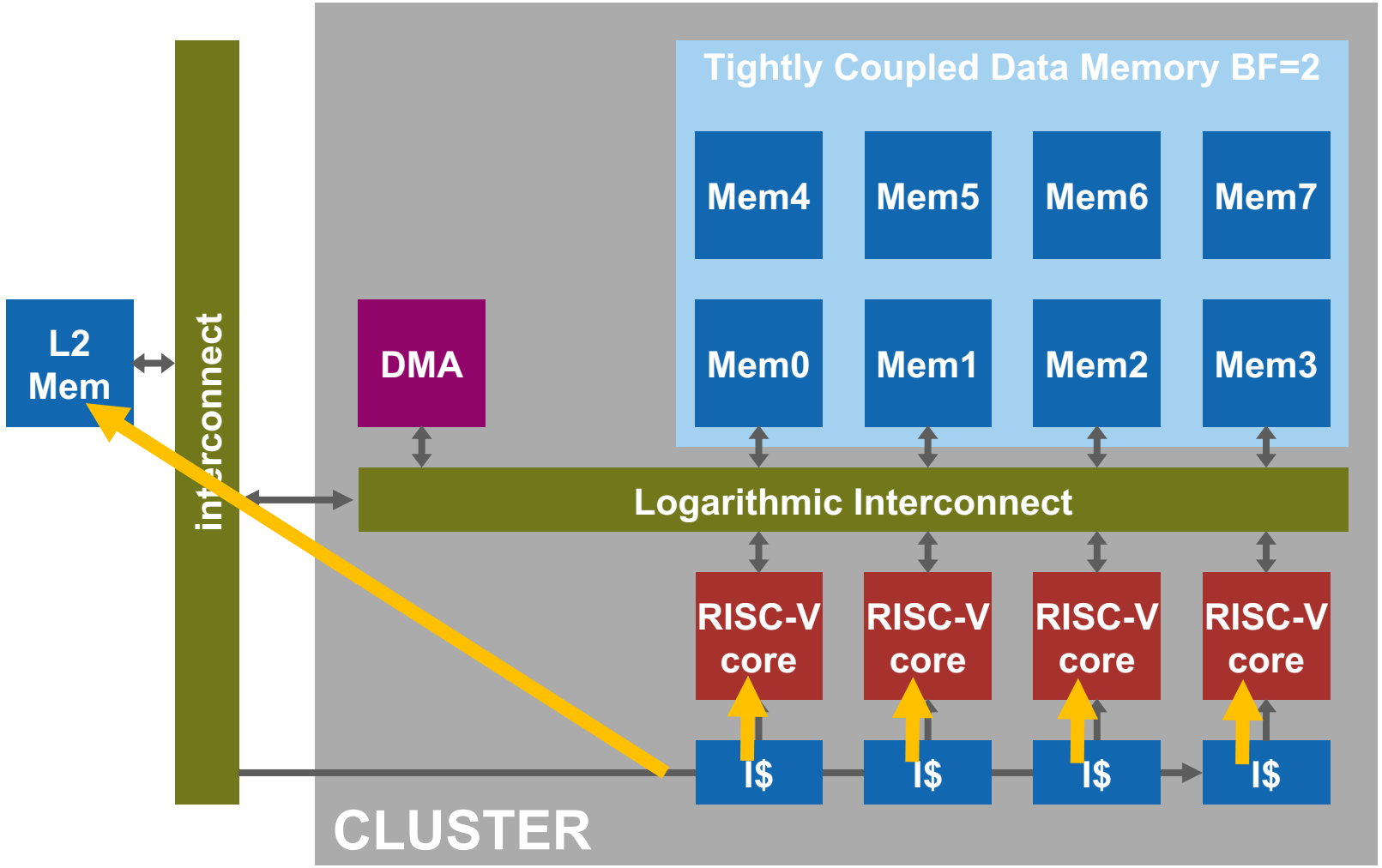
Low-Latency Shared TCDM



DMA for data transfers from/to L2

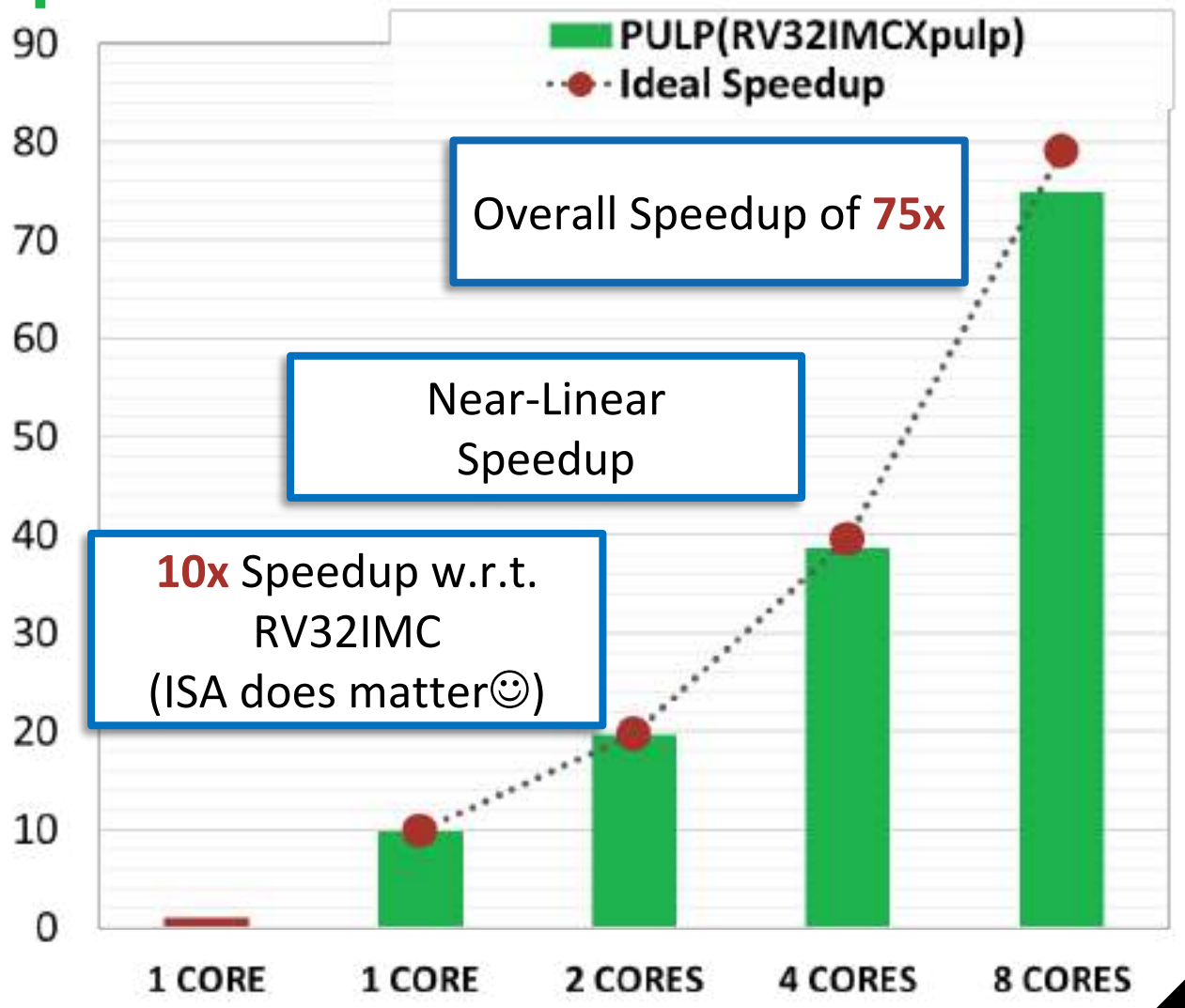


Shared instruction cache with private “loop buffer”



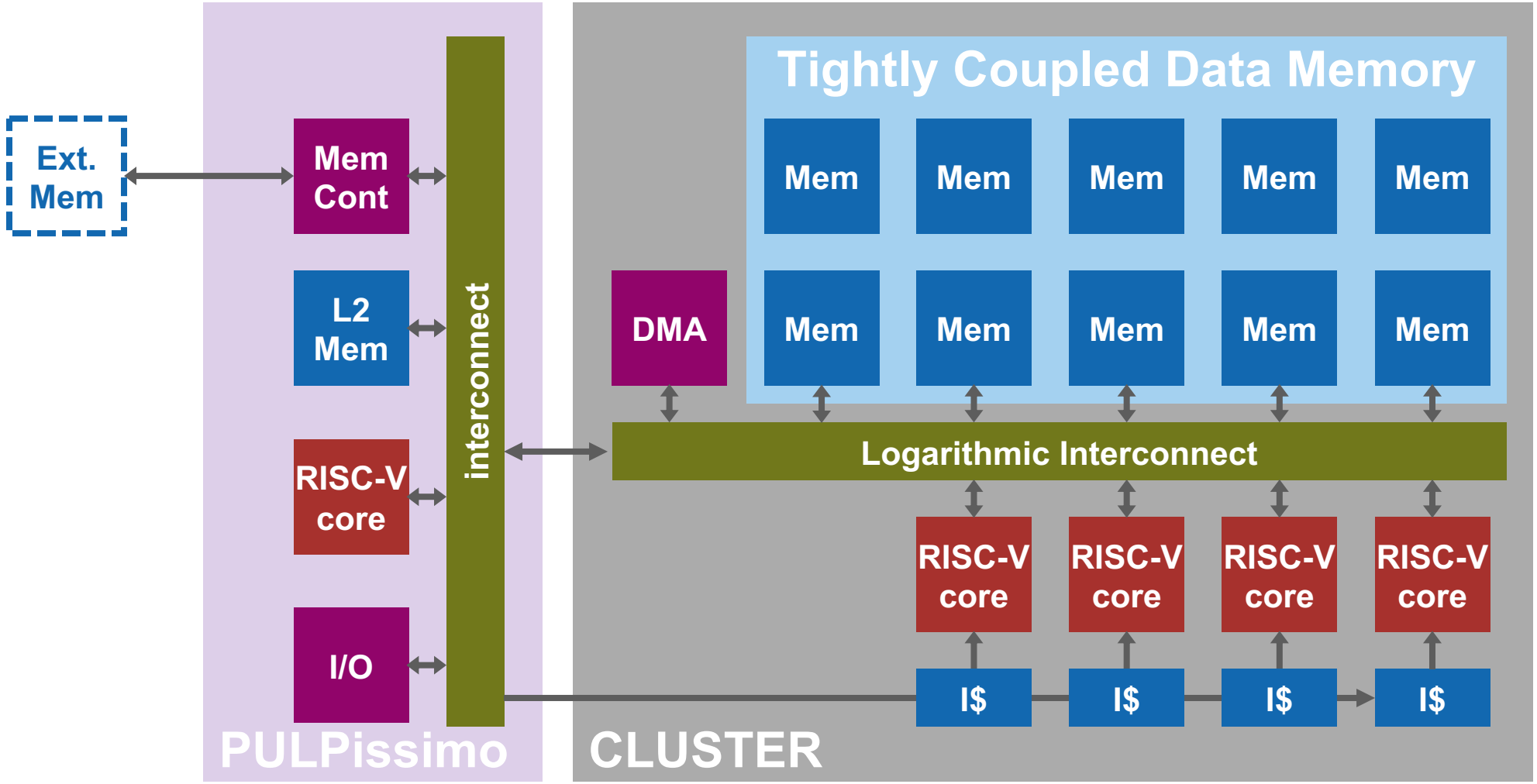
Results: RV32IMCXpulp vs RV32IMC

- **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**
  - Sub-byte: **x2 - x4** better
  - Mixed precision supported (more later)



Garofalo et al. "PULP-NN: Accelerating Quantized Neural Networks on Parallel Ultra-Low-Power RISC-V Processors arxiv.org:1908.11263 [cs.NE]

An additional I/O controller is used for IO

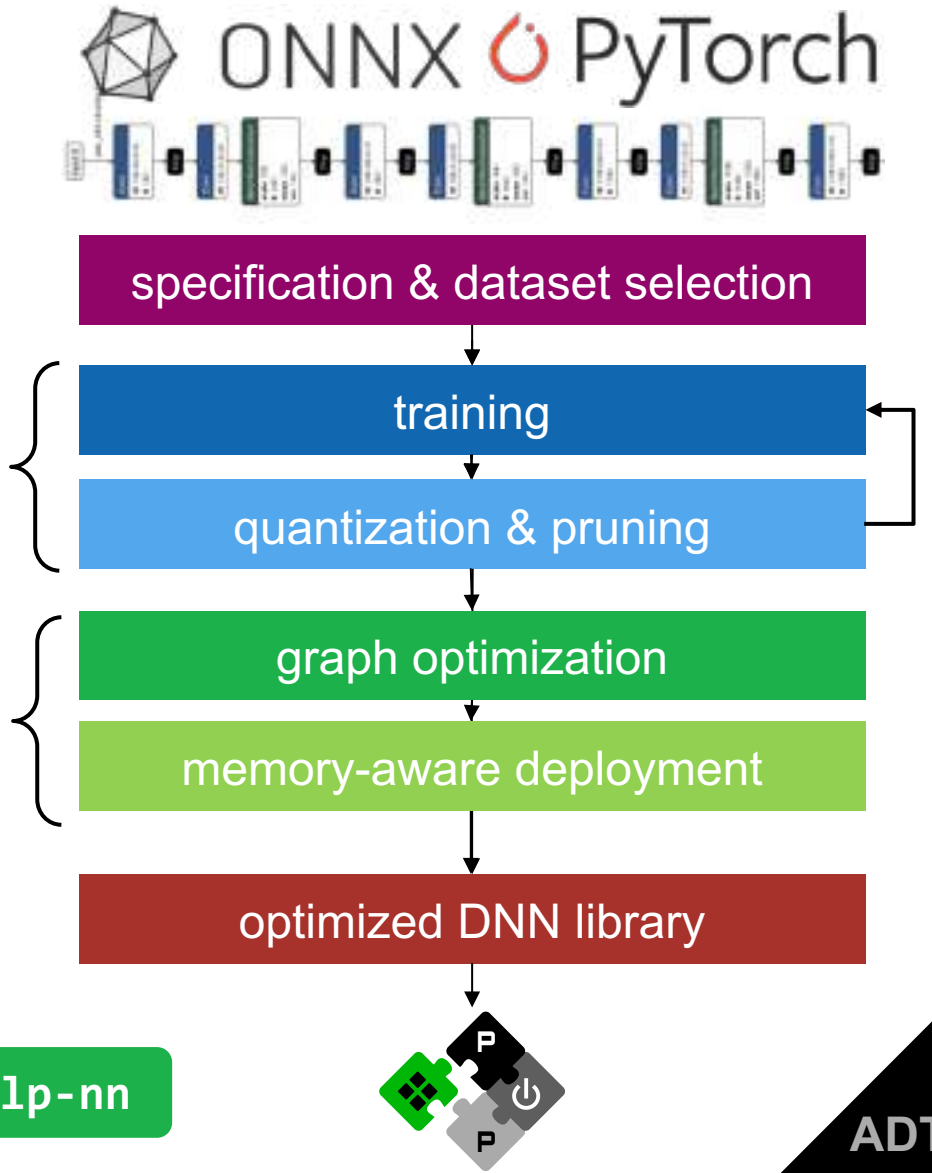


Open sourced since 2017: `github.com/pulp-platform/pulp`

Deploying DNNs on PULP

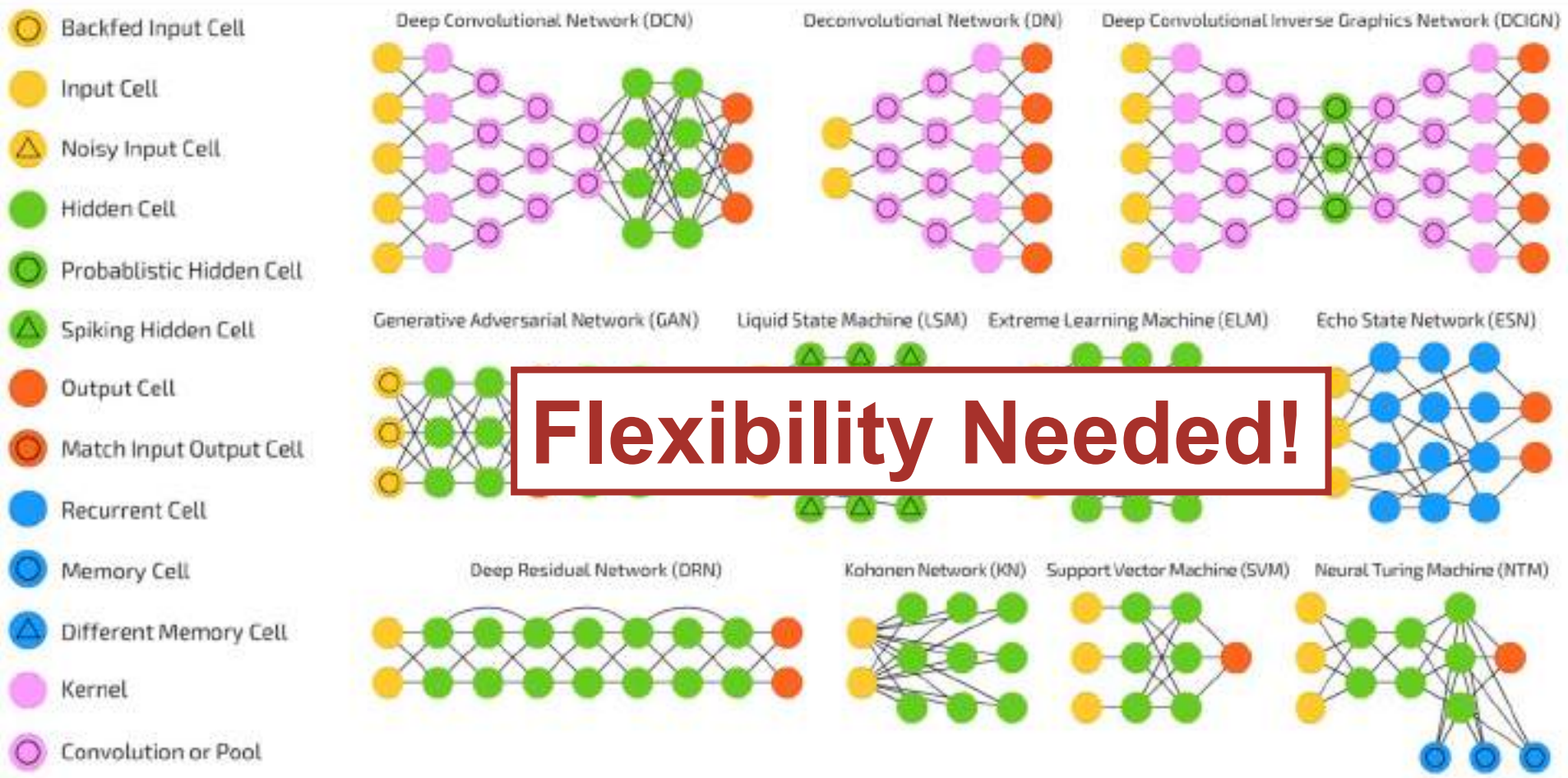
- **QuantLab**
  - Quantization Laboratory
- **NEMO**
  - NEural Minimization for pytOrch
- **DORY**
  - Deployment Oriented to memoRY
- **PULP-NN**
  - PULP Neural Network backend

`github.com/pulp-platform/nemo, /dory, /pulp-nn`



Burrello et al. "DORY: Automatic End-to-End Deployment of Real-World DNNs on Low-Cost IoT MCUs" arxiv.org:2008.07127 [cs.DC]

What's next? Sub-pJ/OP Accelerators

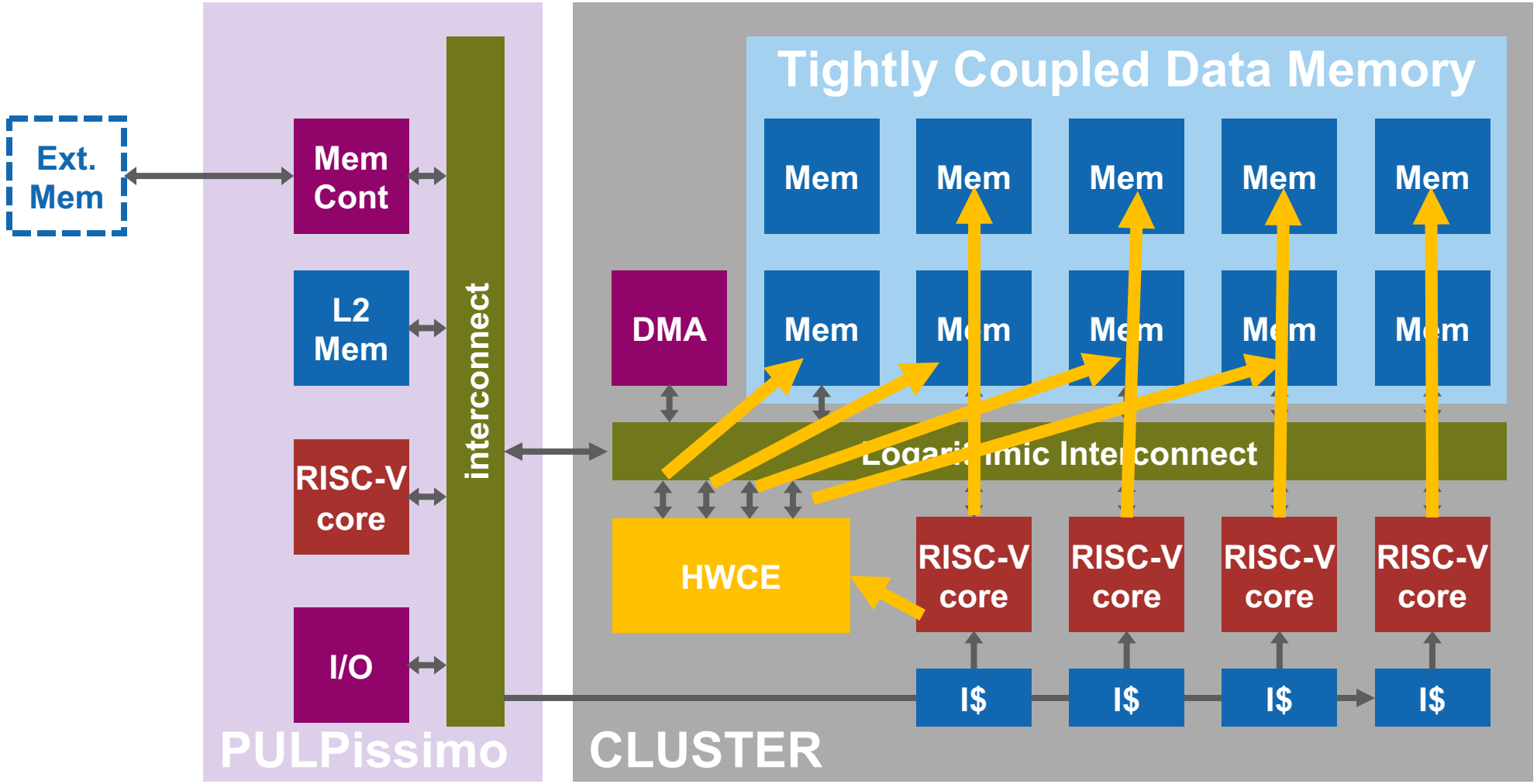


Flexibility Needed!

+ FFT, PCA, Mat-inv,…

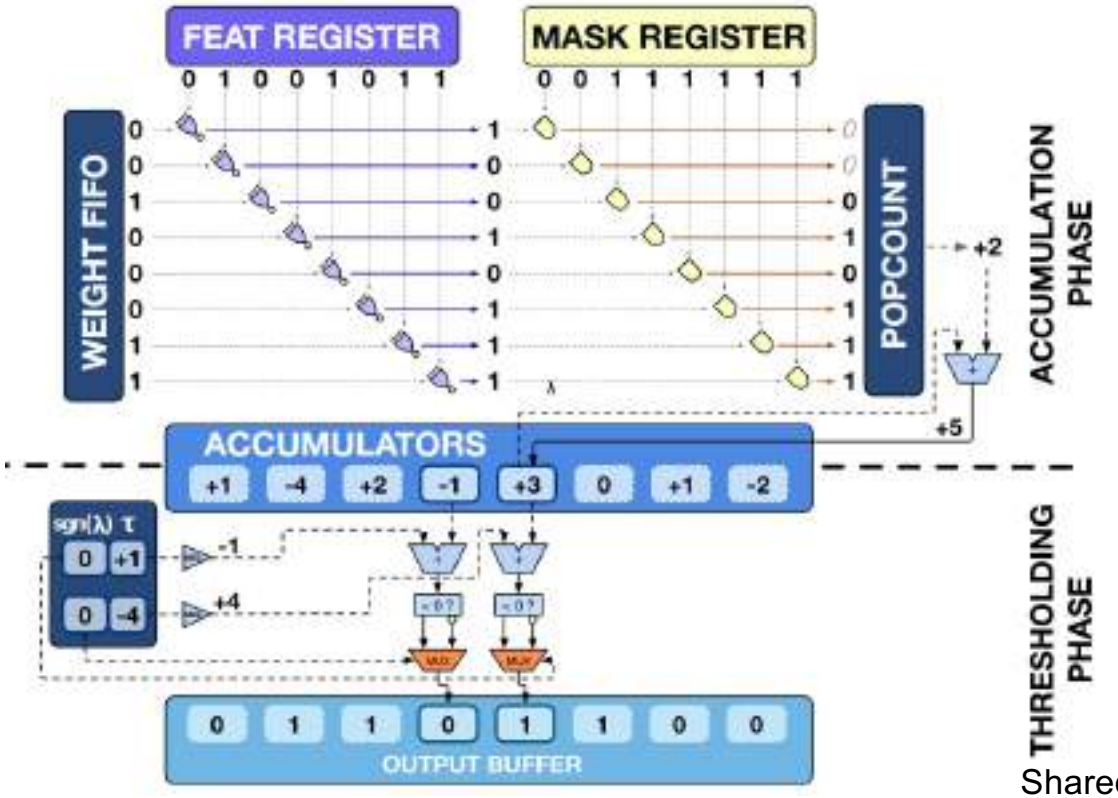
asimovinstitute.org/neural-network-zoo

Tightly-coupled HW Compute Engine

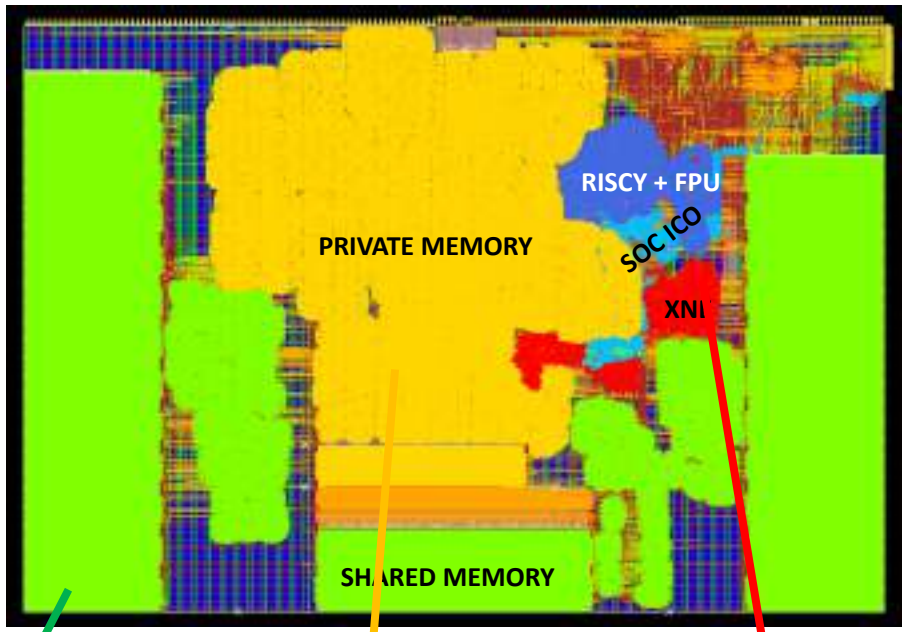


Acceleration with flexibility: zero-copy HW-SW cooperation

XNE: XNOR Neural Engine



Quentin in GlobalFoundries 22FDX



Shared memory is **56 KB** SRAM + **8 KiB** SCM

Private memory is **448 KiB** SRAM + 3r2w **8 KiB** SCM

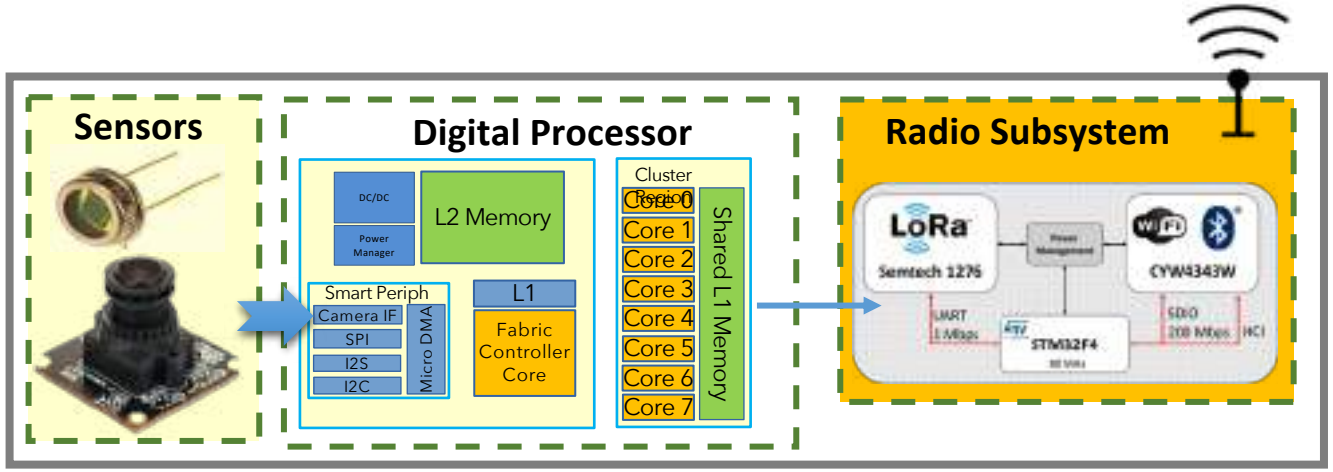
XNE area **~14000 µm2** (71 KGE, 72% RI5CY+FPU)

BINCONV: Binary dot-product and thresholding logic array

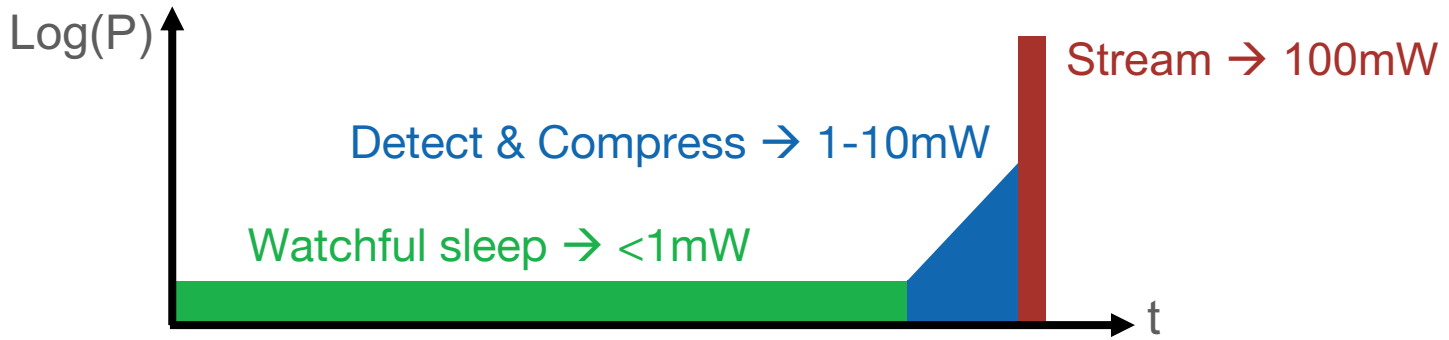
A. D. Mauro, F. Conti, P. D. Schiavone, D. Rossi and L. Benini, "Always-On 674µ W@4GOP/s Error Resilient Binary Neural Networks With Aggressive SRAM Voltage Scaling on a 22-nm IoT End-Node," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 11, pp. 3905-3918, Nov. 2020, doi: 10.1109/TCSI.2020.3012576.

But how to achieve sub-mW average power?

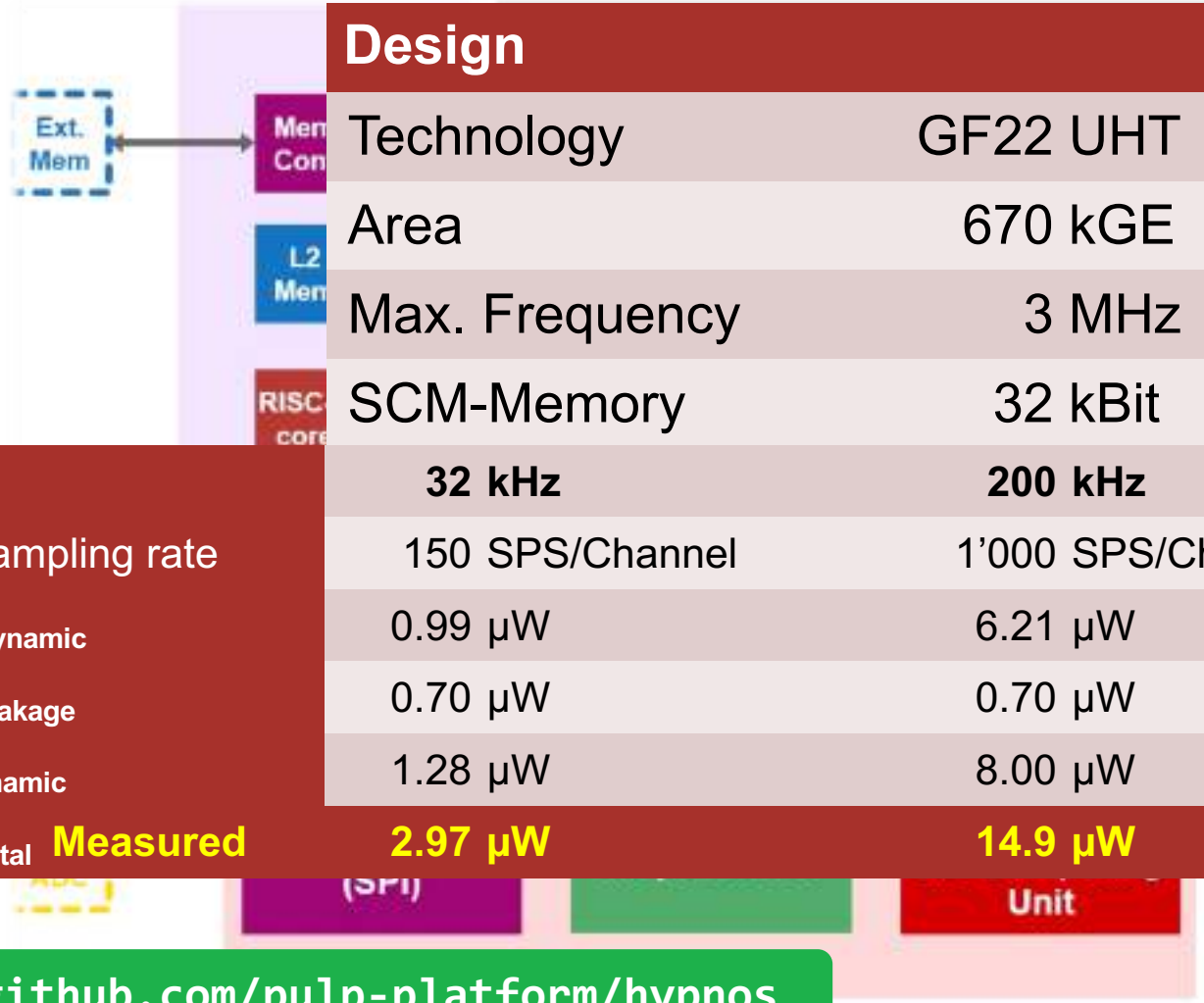
1mW average power with 10mW active power (10GOPS @ 1pJ/OP) → **sub mW sleep**



Duty cycling not acceptable when input events are asynchronous → **watchful Sleep**



HD-Based smart Wake-Up Module - Hypnos

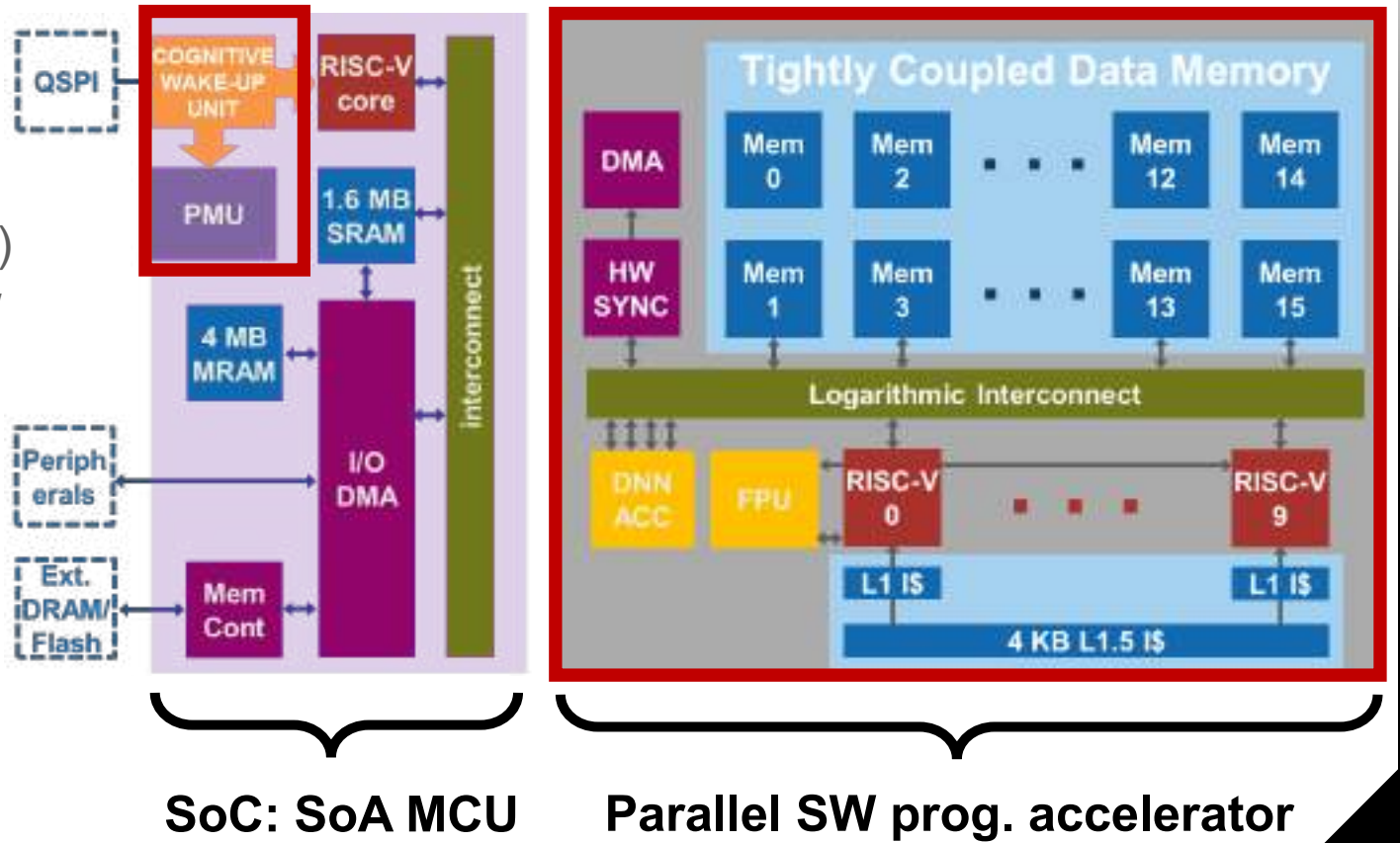


Design		
Technology	GF22 UHT	
Area	670 kGE	
Max. Frequency	3 MHz	
SCM-Memory	32 kBit	
f_{clk}	**32 kHz**	**200 kHz**
max. sampling rate	150 SPS/Channel	1'000 SPS/Channel
$P_{SWU, dynamic}$	0.99 µW	6.21 µW
$P_{SWU, leakage}$	0.70 µW	0.70 µW
$P_{SPI, dynamic}$	1.28 µW	8.00 µW
$P_{SWU, total}$ **Measured**	**2.97 µW**	**14.9 µW**

github.com/pulp-platform/hypnos

VEGA: Extreme Edge IoT Processor

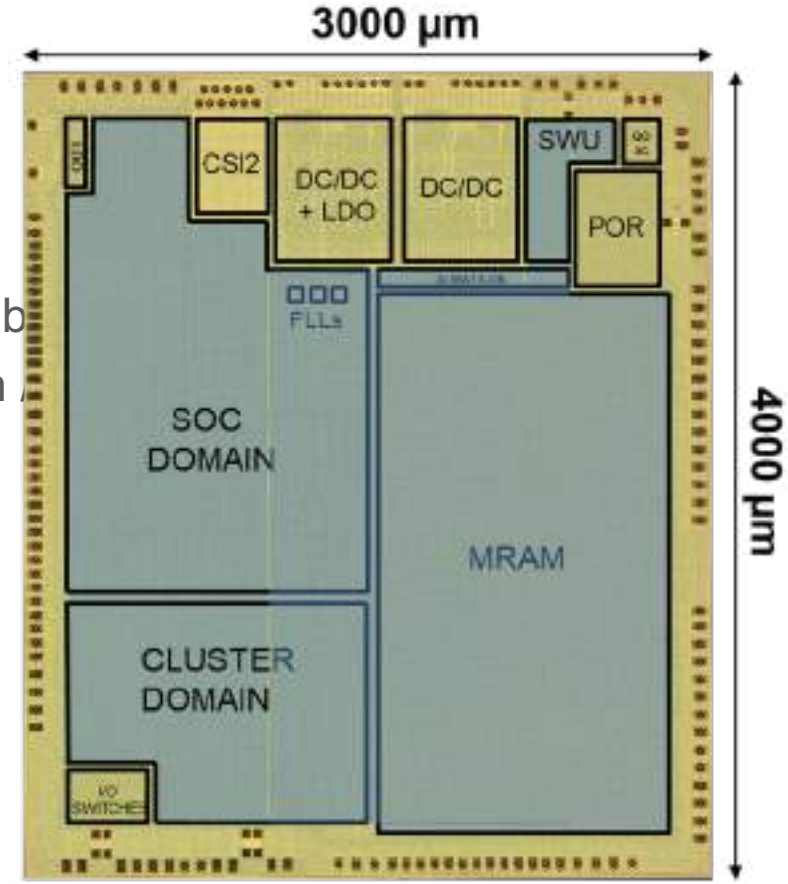
- RISC-V cluster (8cores +1) 614GOPS/W @ 7.6GOPS (8bit DNNs), 79GFLOPS/W @ 1GFLOP (32bit FP appl)
- Multi-precision HWCE(4b/8b/16b) 3×3×3 MACs with normalization / activation: 32.2GOPS and 1.3TOPS/W (8bit)
- 1.7 µW cognitive unit for autonomous wake-up from retentive sleep mode



D. Rossi et al., "4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7µW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode," 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021, pp. 60-62, doi: 10.1109/ISSCC42613.2021.9365939.

VEGA: Extreme Edge IoT Processor

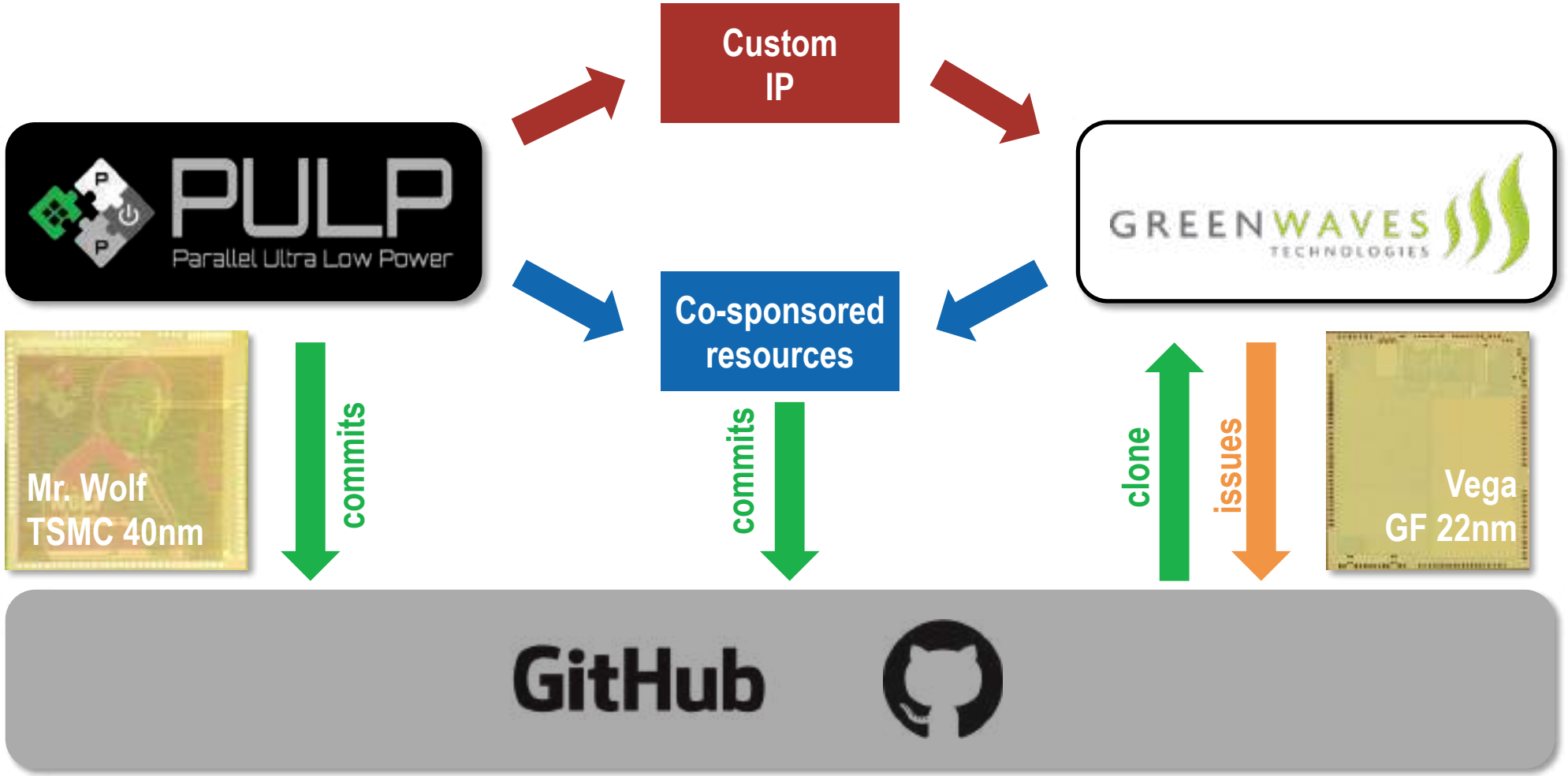
- RISC-V cluster (8cores +1) 614GOPS/W @ 7.6GOPS (8bit DNNs), 79GFLOPS/W @ 1GFLOP (32bit FP appl)
- Multi-precision HWCE(4b/8b/16b 3×3×3 MACs with normalization activation: 32.2GOPS and 1.3TOPS/W (8bit)
- 1.7 µW cognitive unit for autonomous wake-up from retentive sleep mode
- **Fully-on chip DNN inference with 4MB MRAM**



Technology	22nm FDSOI
Chip Area	12mm^2
SRAM	1.7 MB
MRAM	4 MB
VDD range	0.5V - 0.8V
VBB range	0V - 1.1V
Fr. Range	32 kHz - 450 MHz
Pow. Range	1.7 µW - 49.4 mW

D. Rossi et al., "4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7µW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode," 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021, pp. 60-62, doi: 10.1109/ISSCC42613.2021.9365939.

Open source collaboration scheme explained



PULP is an Open Platform

- **For science … fundamental "research infrastructure"**
  - Reduce "getting up to speed" overhead for partners
  - Enables fair and well controlled benchmarking
- **For Business … it is truly disruptive**
  - Reduces the NRE , faster innovation path for startups
  - New business models (for profit and non-for profit)
- **Heterogeneous & Flexible**
  - 1-3 orders of magnitude improvement (wrt to efficient RV) by acceleration
    - Achieved through efficient implementation, ISA extensions, heterogeneous accelerator combinations
  - To achieve true system-level sub pj/Op operation, everything little thing counts
    - Efficient I/O, 3D integration, sleep modes, power conversion

PULP

Parallel Ultra Low Power

Luca Benini, Alessandro Capotondi, Alessandro Ottaviano, Alessio Burrello, Alfio Di Mauro, Andrea Borghesi, Andrea Cossettini, Andreas Kurth, Angelo Garofalo, Antonio Pullini, Arpan Prasad, Bjoern Forsberg, Corrado Bonfanti, Cristian Cioflan, Daniele Palossi, Davide Rossi, Fabio Montagna, Florian Glaser, Florian Zaruba, Francesco Conti, Georg Rutishauser, Germain Haugou, Gianna Paulin, Giuseppe Tagliavini, Hanna Müller, Luca Bertaccini, Luca Valente, Manuel Eggimann, Manuele Rusci, Marco Guermandi, Matheus Cavalcante, Matteo Perotti, Matteo Spallanzani, Michael Rogenmoser, Moritz Scherer, Moritz Schneider, Nazareno Bruschi, Nils Wistoff, Pasquale Davide Schiavone, Paul Scheffler, Philipp Mayer, Robert Balas, Samuel Riedel, Segio Mazzola, Sergei Vostrikov, Simone Benatti, Stefan Mach, Thomas Benz, Thorir Ingolfsson, Tim Fischer, Victor Javier Kartsch Morinigo, Vlad Niculescu, Xiaying Wang, Yichao Zhang, Frank K. Gürkaynak, all our past collaborators and many more that we forgot to mention

http://pulp-platform.org @pulp_platform