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

In-Sensor Machine Learning

Heterogeneous computing in a mW

Luca Benini
<lbenini@iis.ee.ethz.ch,luca.Benini@unibo.it>

Prof. of Digital Circuit and Systems @ ETHZ and UNIBO. h-index=109, 53'000+ citations, 1'000+ publications, fellow IEEE, ACM, Chief Architect in STMicroelectronics (2009-2012) Group of 80+ people

European Research Council

European Commission | Horizon 2020 European Union funding for Research & Innovation

nano-tera.ch

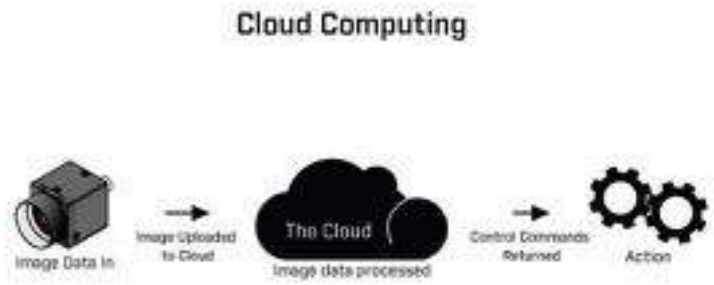
FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

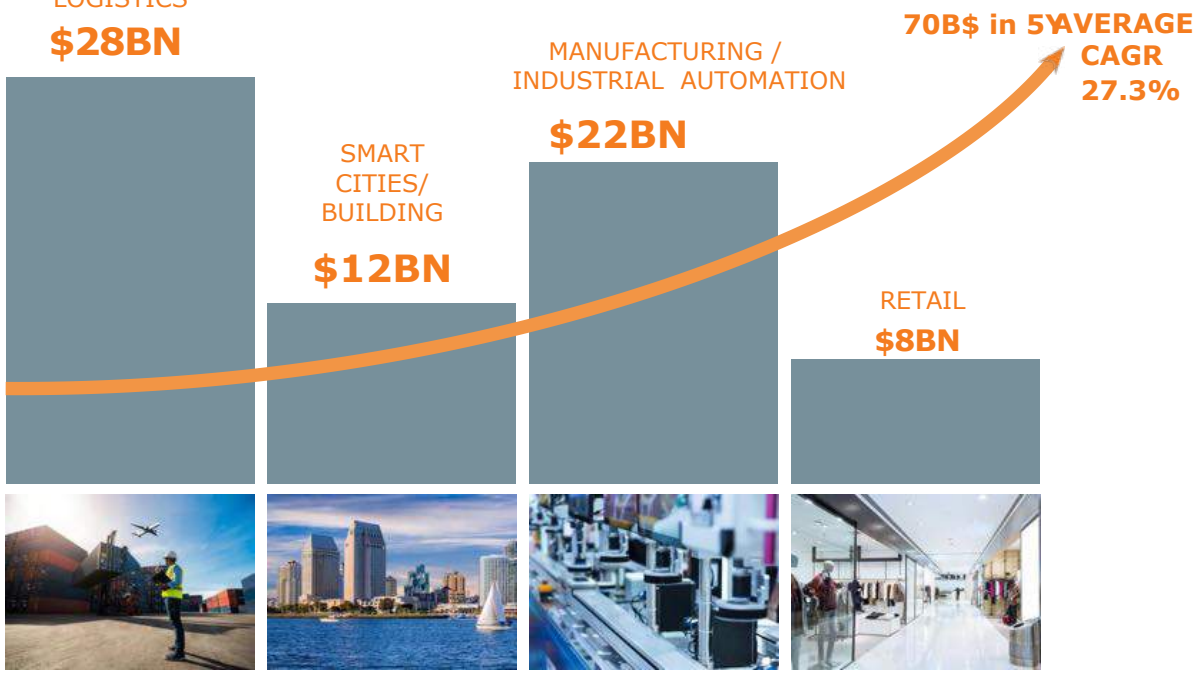
ETH zürich

Alma Mater Studiorum A.D. 1088

http://pulp-platform.org @pulp_platform

Cloud → Edge →Near-Sensor AI a.k.a. TinyML





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/

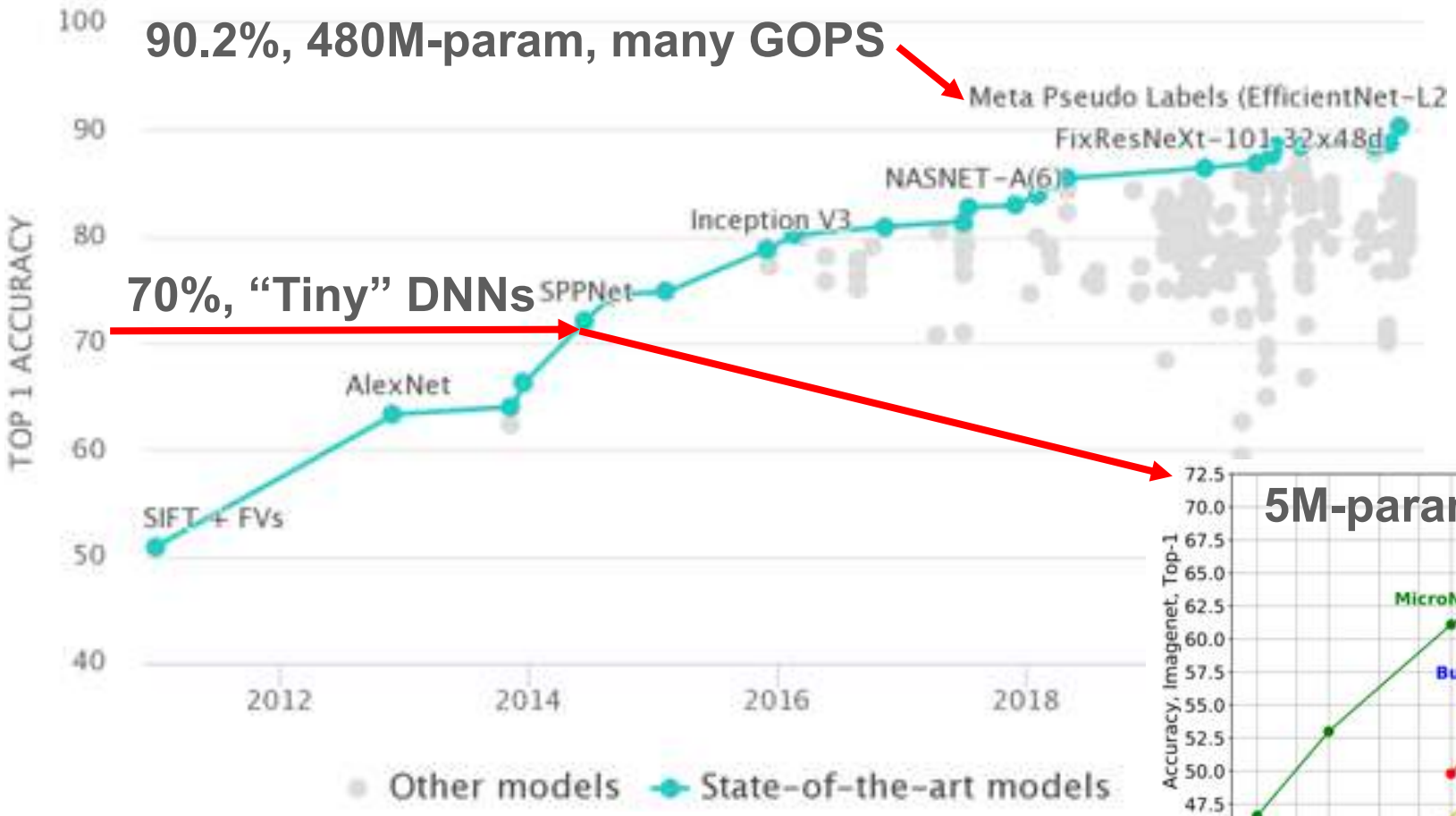
Near-Sensor AI challenge

AI capabilities in the power envelope of an MCU:

100mW peak (10mW avg)

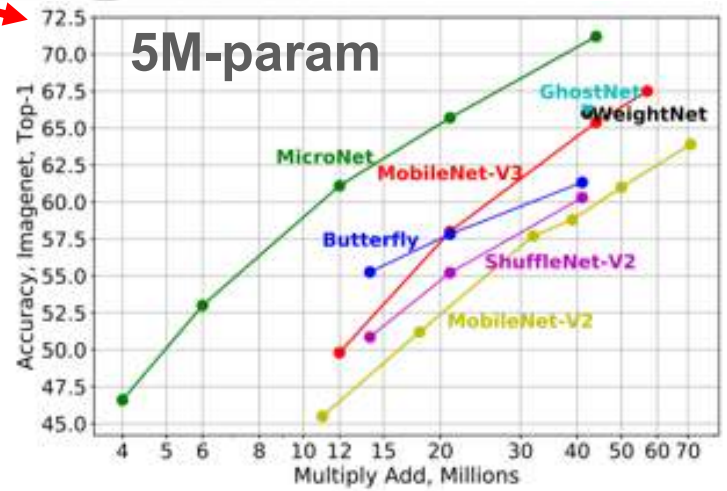
AI Workloads - DNNs

H Pham 2021(Google) arXiv:2003.10580v3

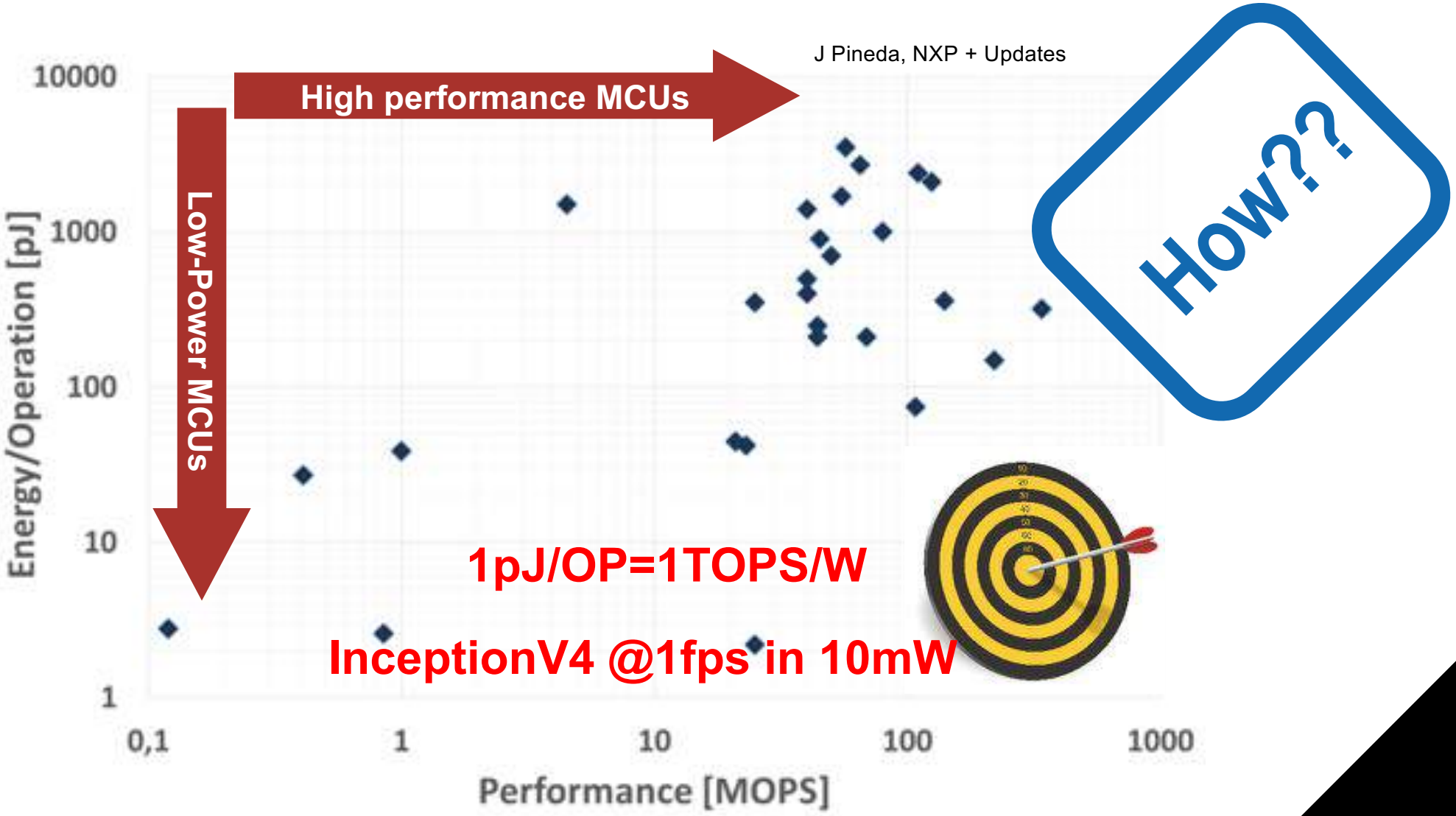


High OP/B ratio
Massive Parallelism
MAC-dominated
Low precision OK

Model redundancy

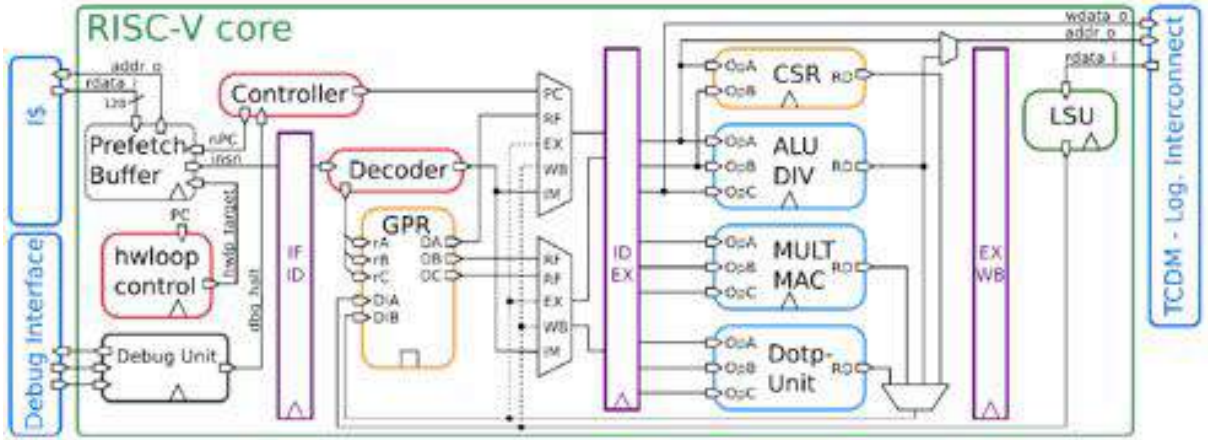


Energy efficiency @ GOPS is THE Challenge

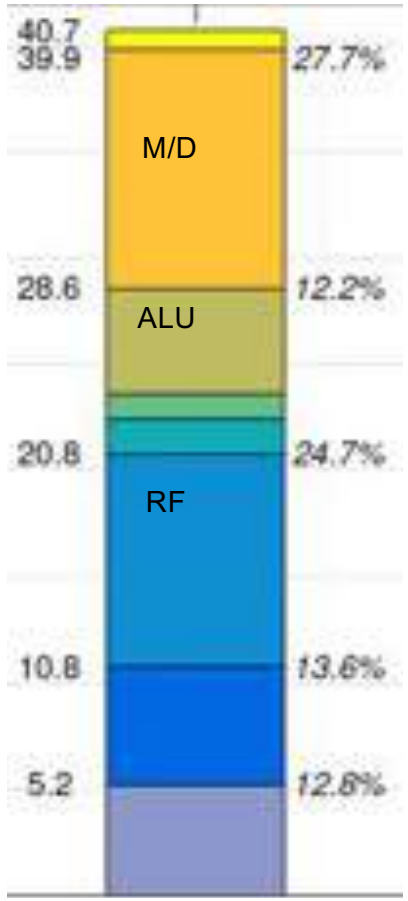


RI5CY – An Open MCU-class RISC-V Core for EE-AI

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



40 kGE
70% RF+DP

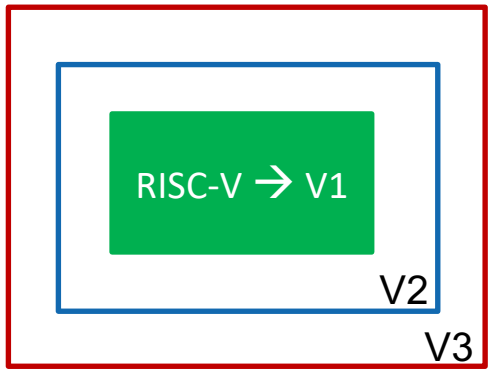


RISC-V ISA is extensible *by construction* (great!)

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

V2 Post modified Load/Store
Mac

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



XPULP extensions: 25 kGE → 40 kGE (1.6x)

ETH zürich

PULP-NN: Xpulp ISA exploitation

8-bit Convolution

RV32IMC

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

Pooling & ReLu

HW loop
LD/ST with post-increment
8-bit SIMD max, avg INSNS

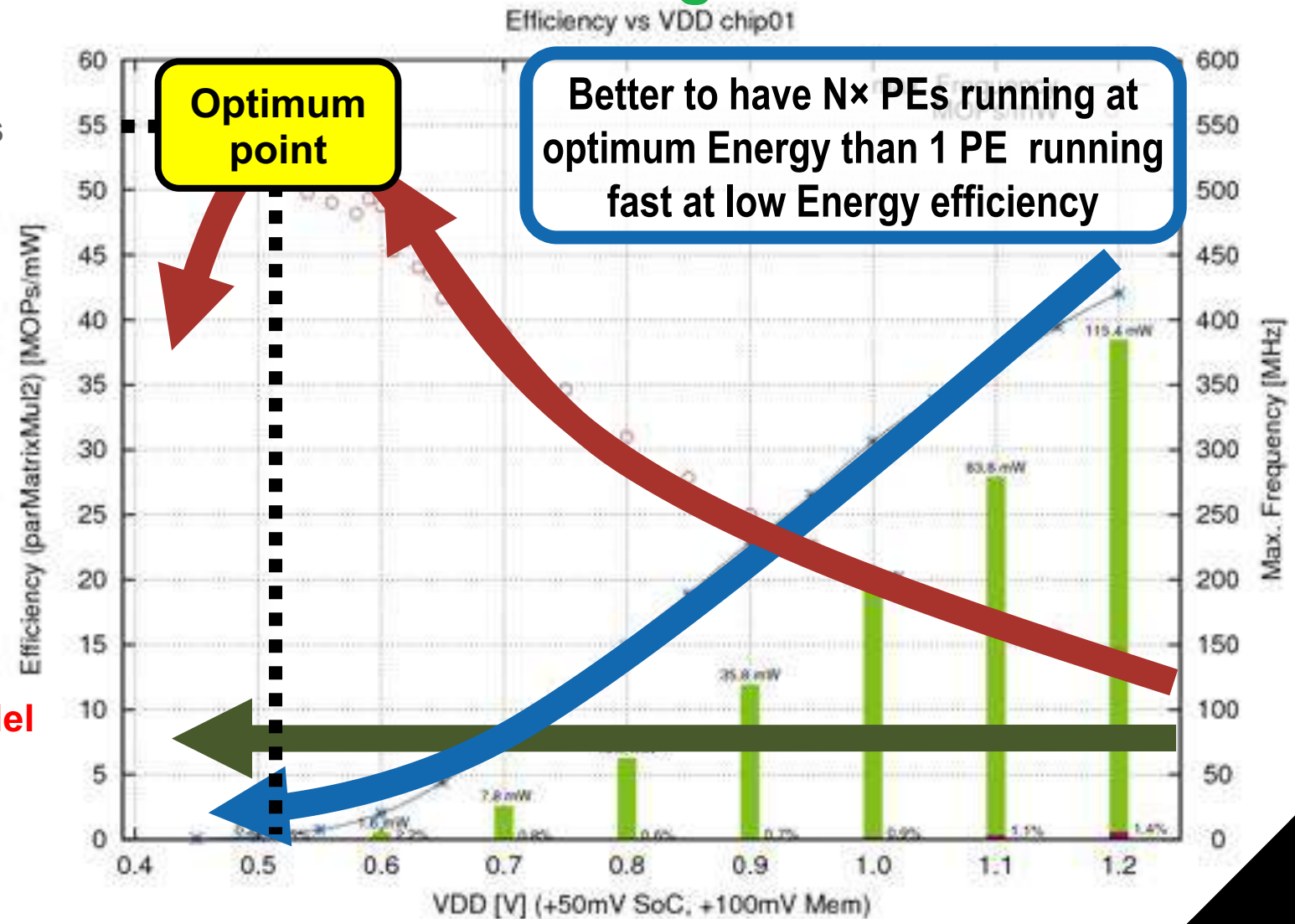
P↑ T↓↓↓ so, E=P*T↓↓ Nice!
But what about the GOPS?
Faster+Superscalar is not efficient!

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

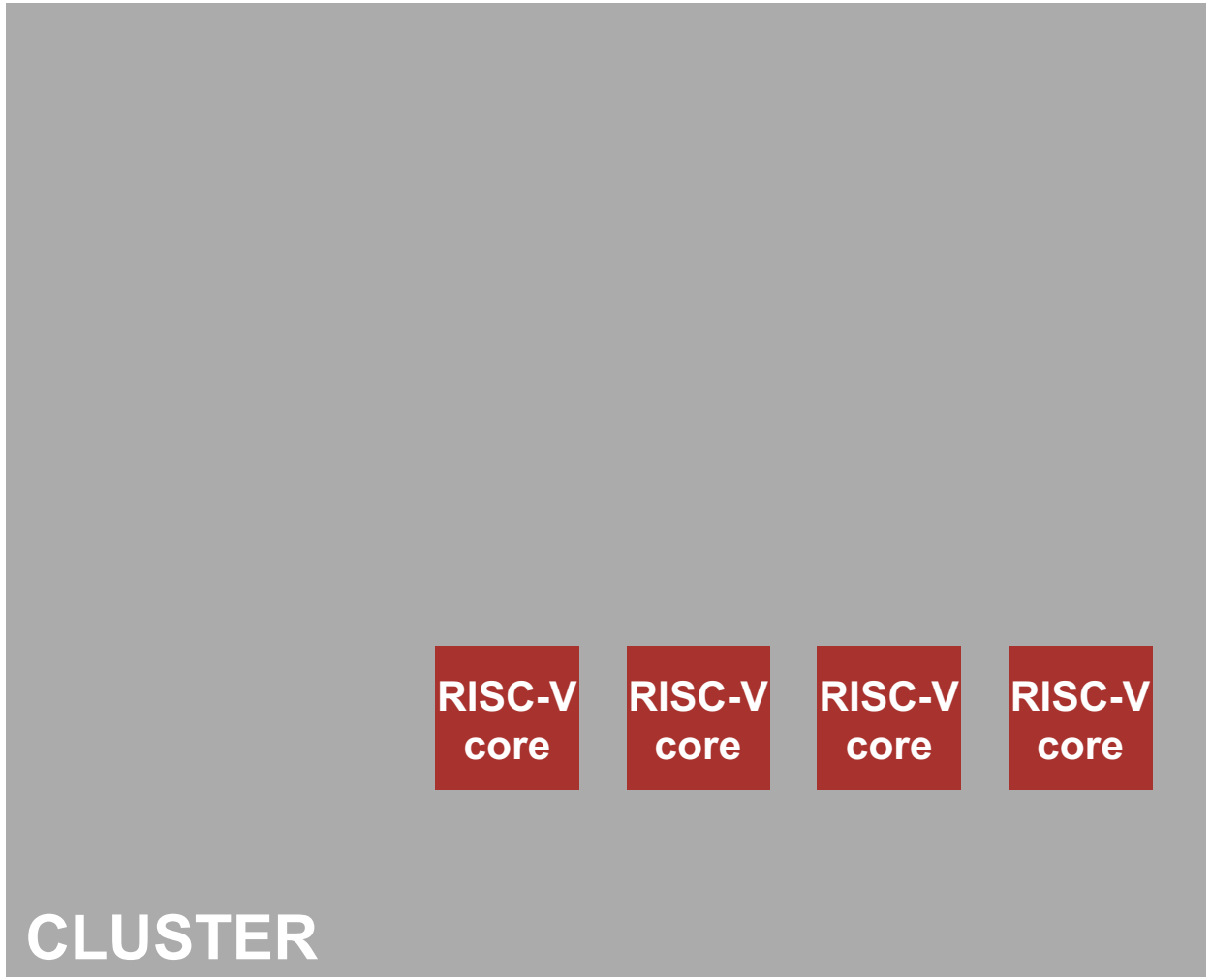
ML & Parallel, Near-threshold: a Marriage Made in Heaven

- 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

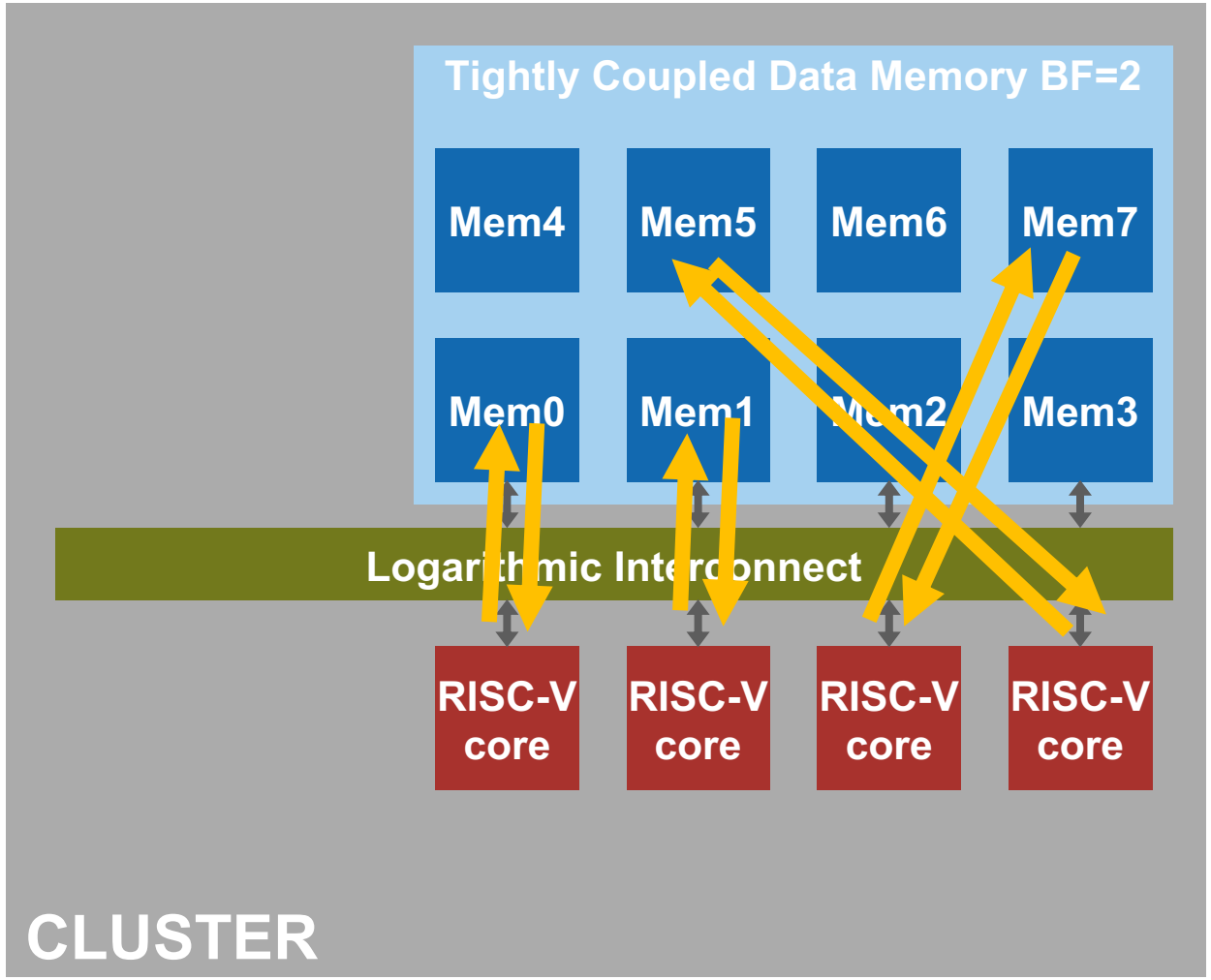
ML is massively parallel and scales well (P/S ↑ with NN size)



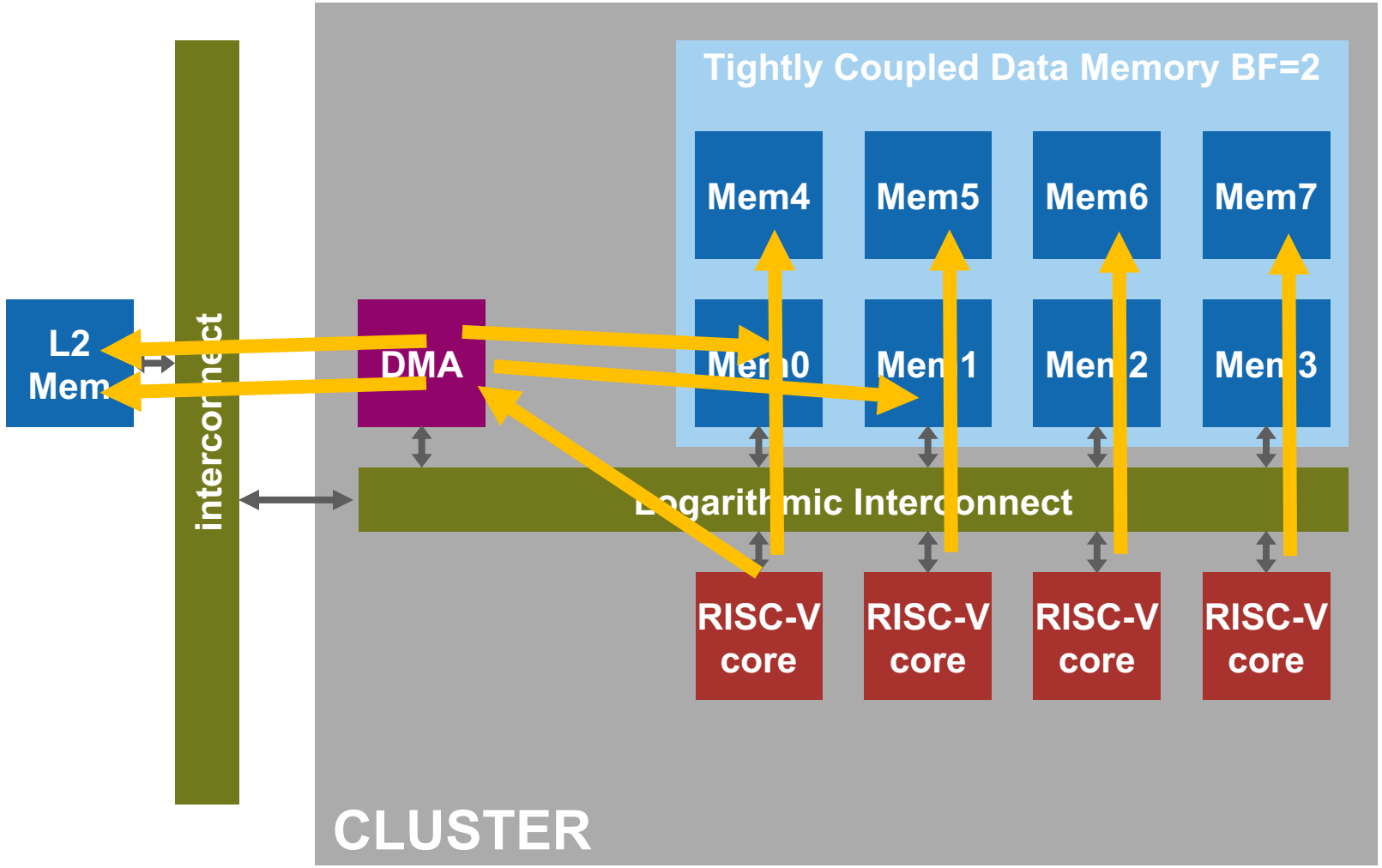
Multiple RI5CY Cores (1-16)



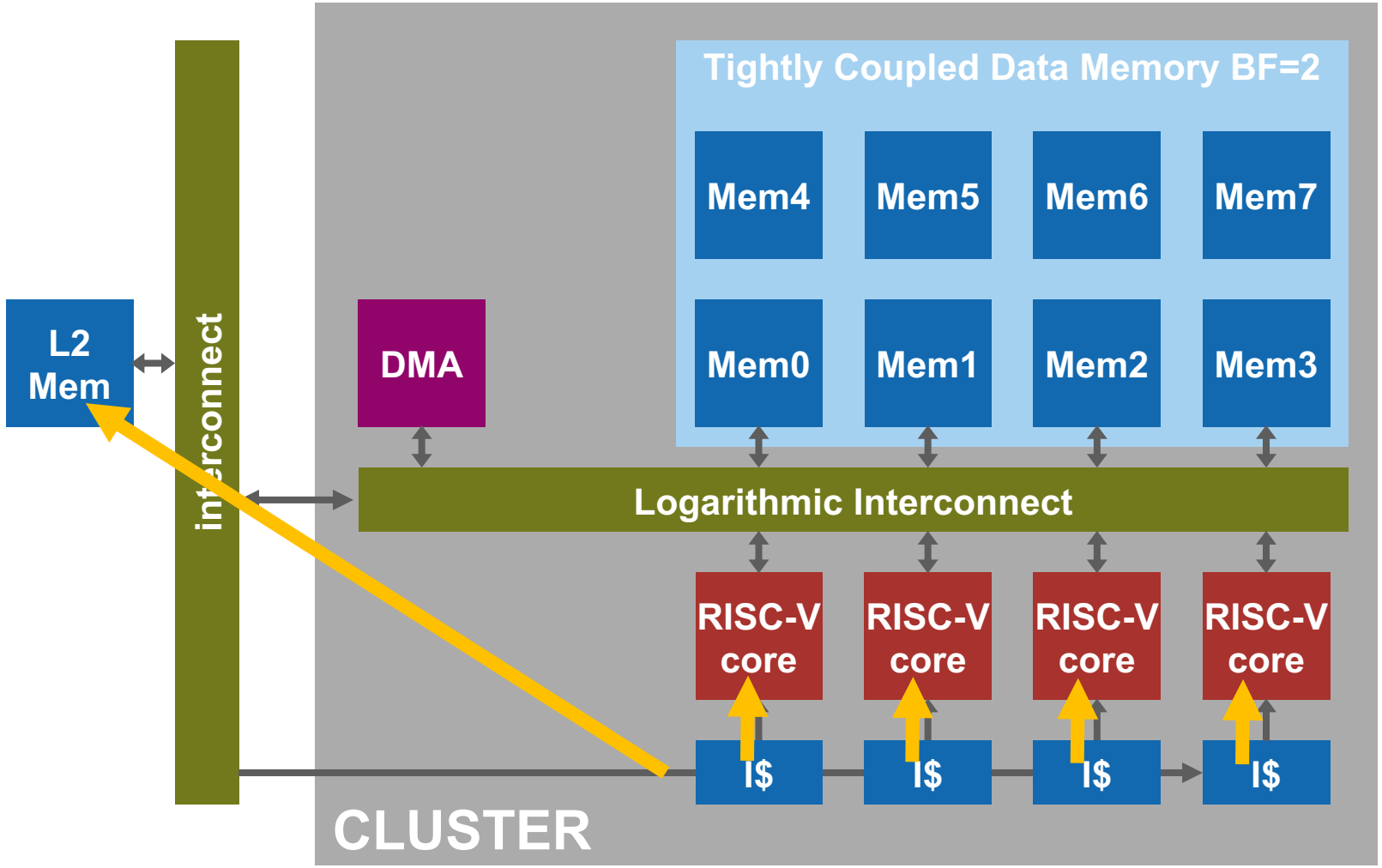
Low-Latency Shared TCDM



DMA for data transfers from/to L2

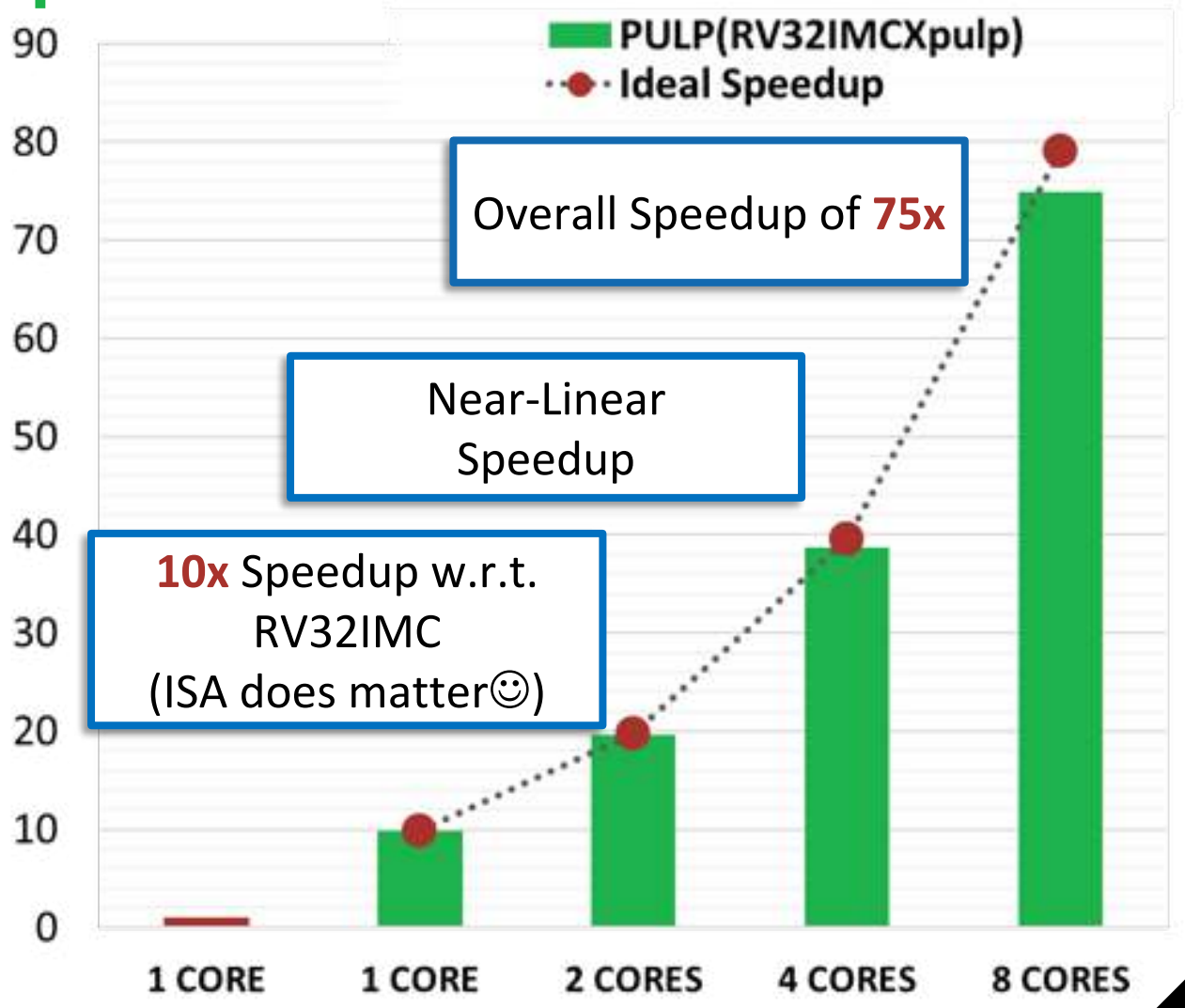


Shared instruction cache with private “loop buffer”

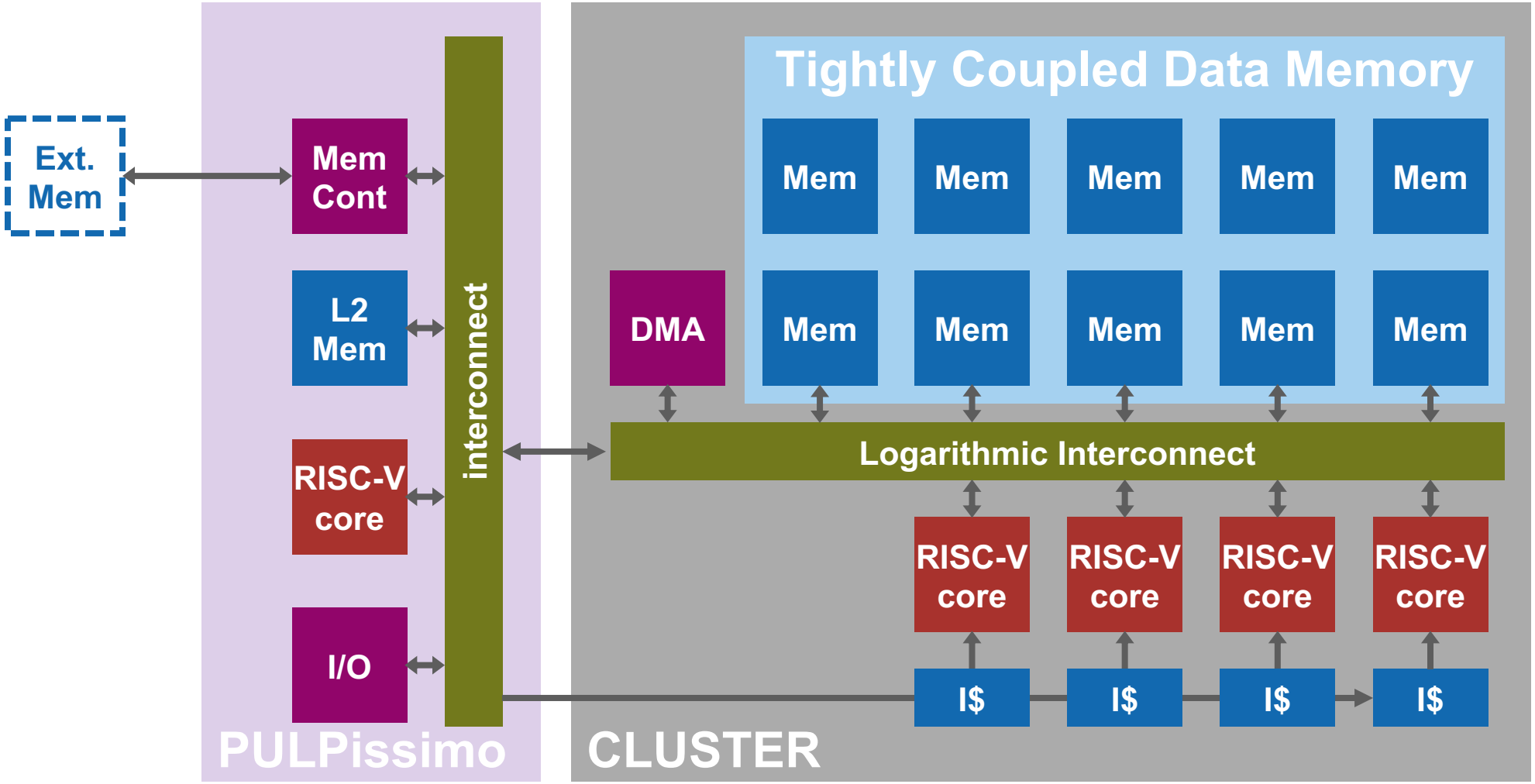


Results: RV32IMCXpulp vs RV32IMC

- 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: x2-4x better
  - Mixed precision supported

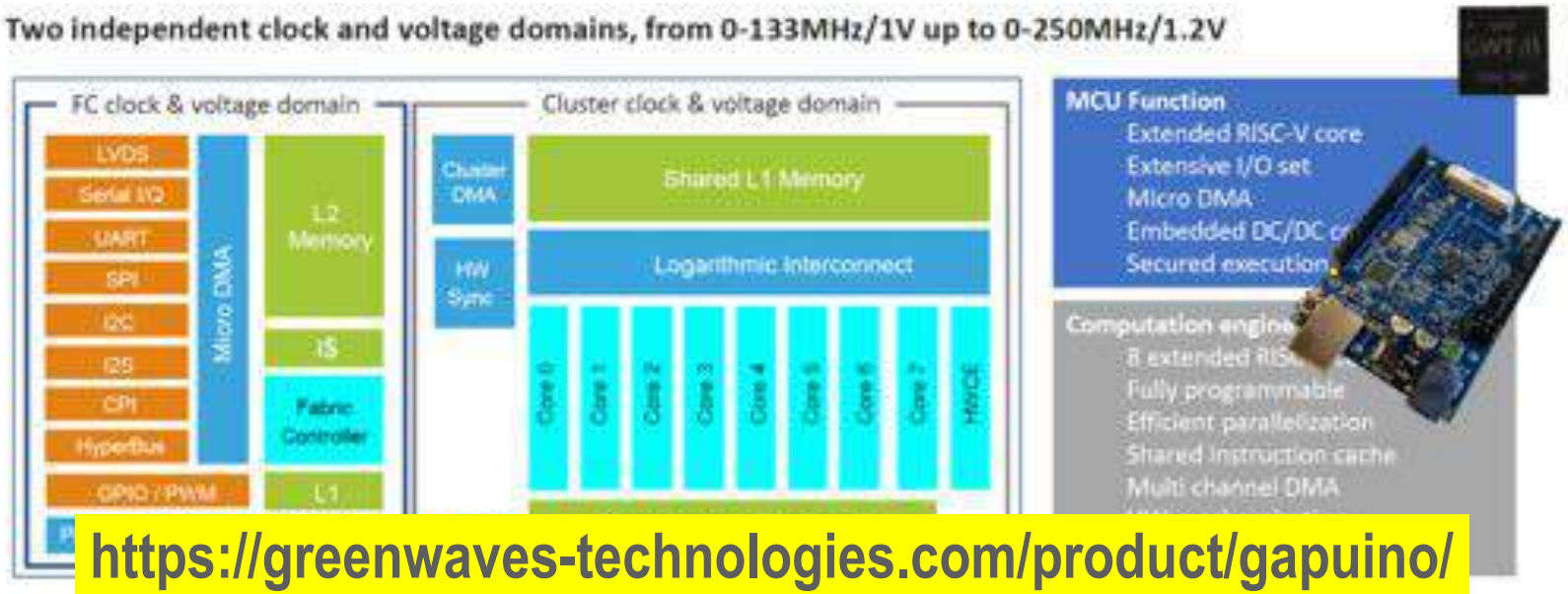


An additional I/O controller is used for IO



All this is Open Source HW – PULP open

Successful product development: GWT's GAP8



https://greenwaves-technologies.com/product/gapuino/

GREENWAVES TECHNOLOGIES

The evolution of the PULP species

2017: **GAP-8 55nm (TSMC):** 50 MOPS/mW (20pJ/OP @32bit 3.5GOPS)

2018: Wolf(8) 40nm (TSMC): 120 MOPS/mW (8pJ/OP @32bit +FP 7GOPS)

2019: **Vega(8) 22FDX**: 500 MOPS/mW (**2pJ/OP** @32bit, +FP, 10GOPS)

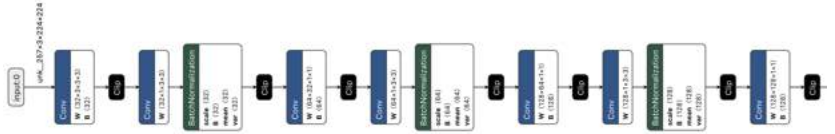
2020: **Marsellus(16) 22FDX**: 500+ MOPS/mW (pre-tapeout, **30GOPS)**

2x GOPS/W Y/Y

ETH zürich

Deploying DNNs on PULP

ONNX PyTorch



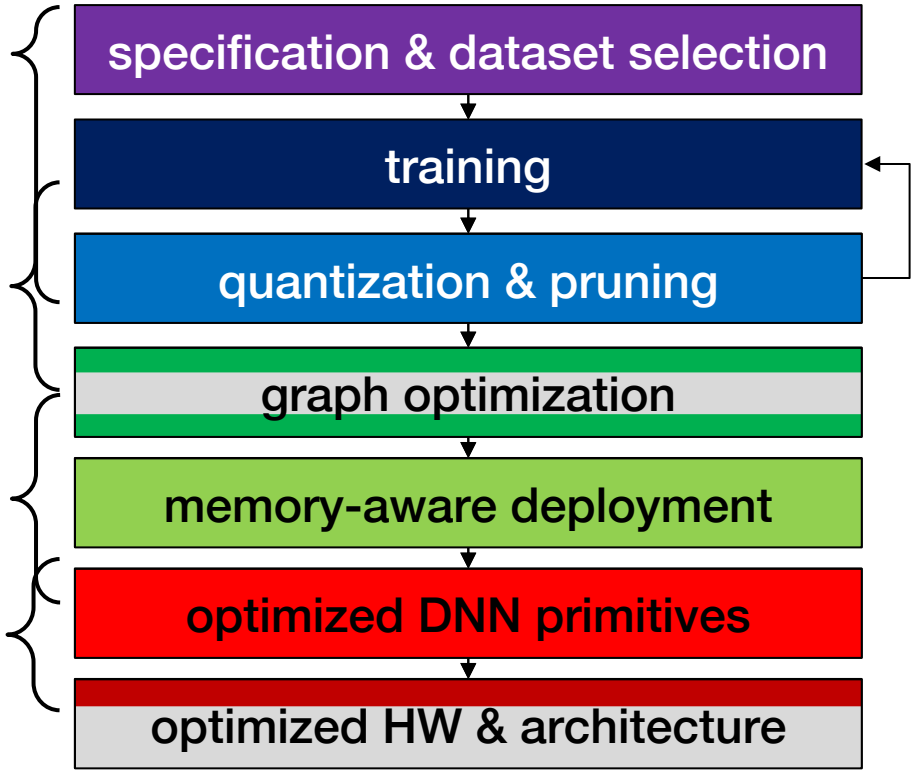
QuantLab
Quantization Laboratory

Automatic Mixed Prec

NEMO
***NE**ural **M**inimization for pyt**O**rch*

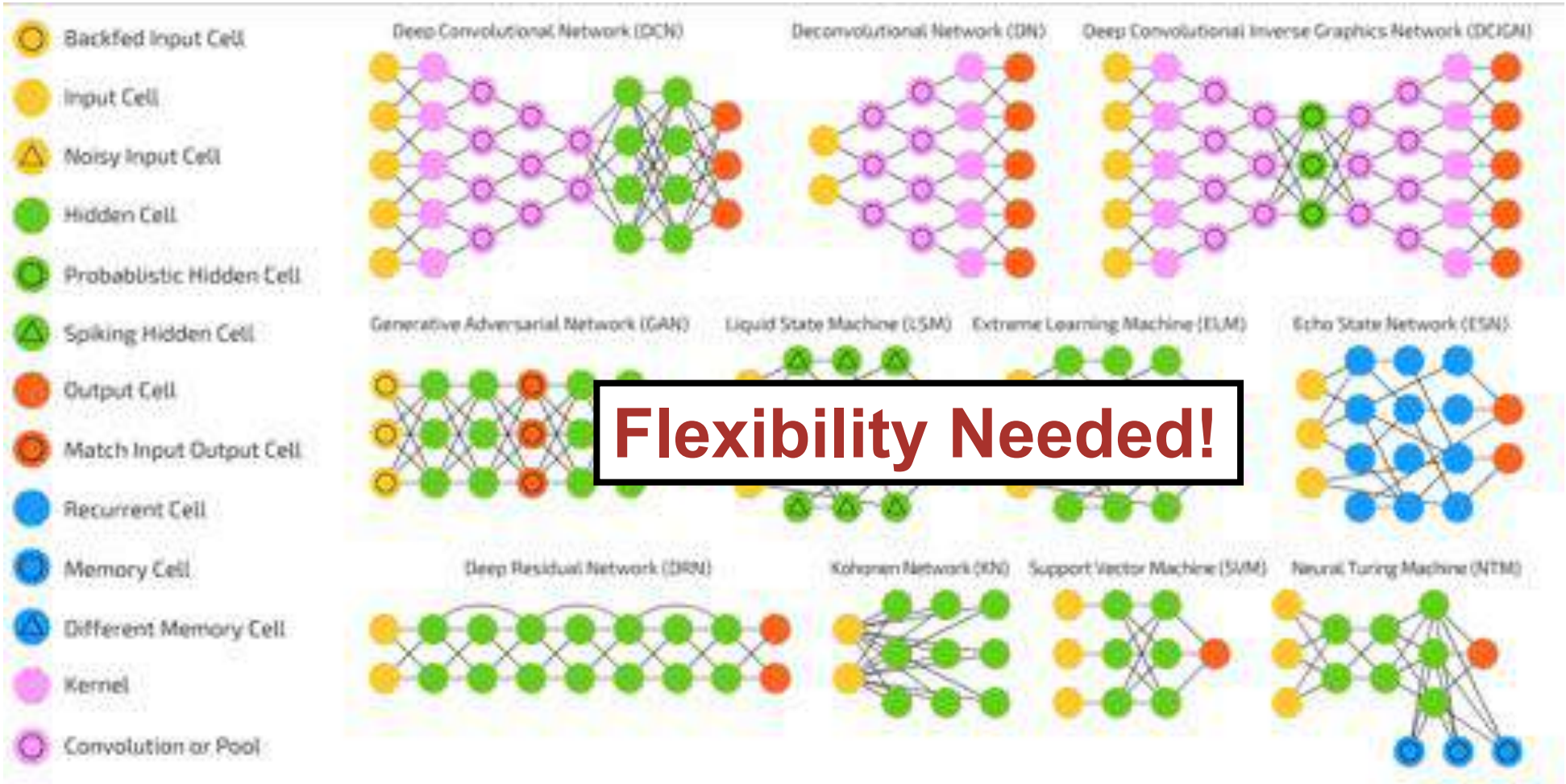
DORY
D**eployment **O**riented to memo**RY

PULP-NN
***PULP N**eural **N**etwork backend*



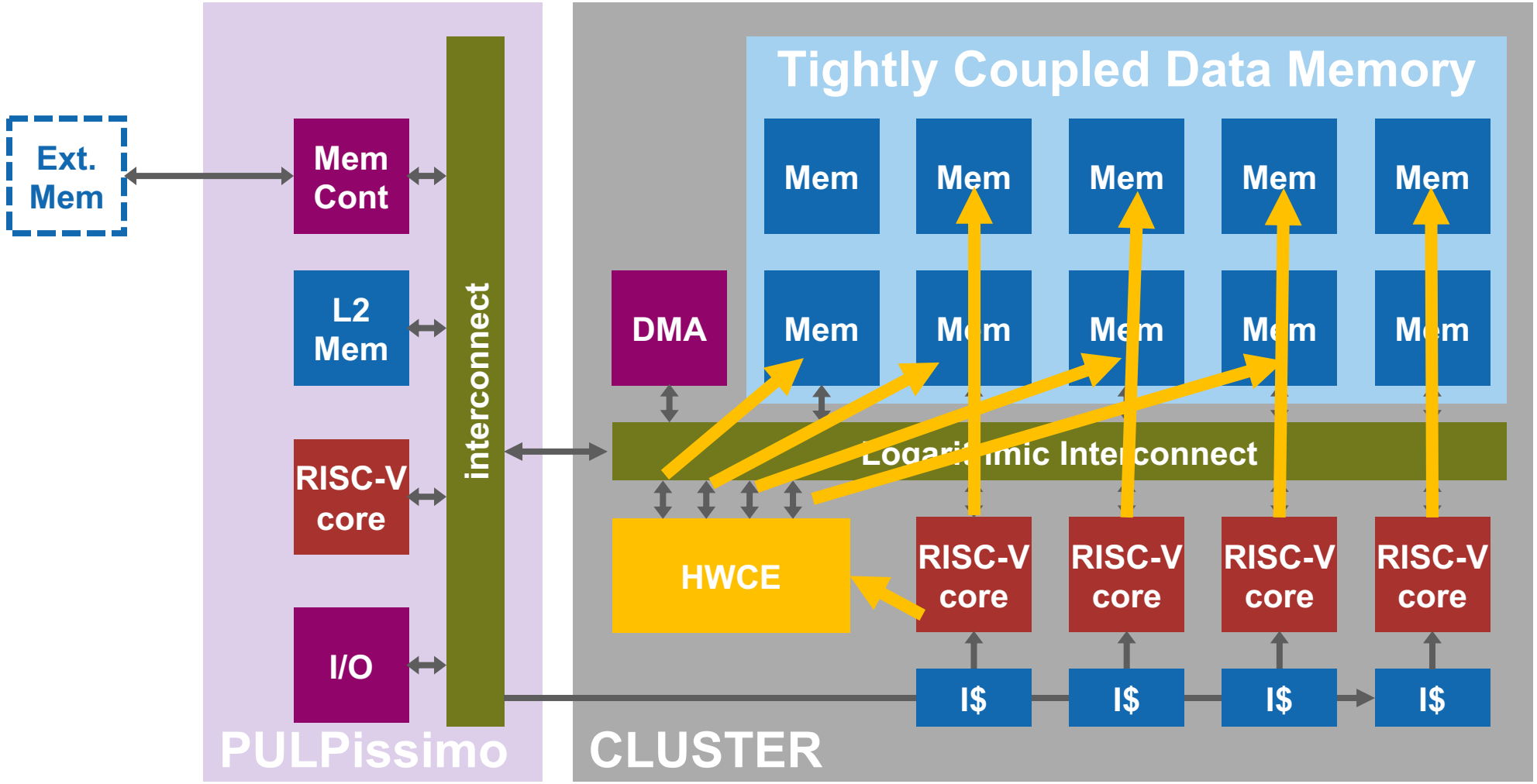
https://github.com/pulp-platform/nemo
https://github.com/pulp-platform/dory
https://github.com/pulp-platform/pulp-nn

What's next? Architecture: Sub-pJ/OP Accelerators



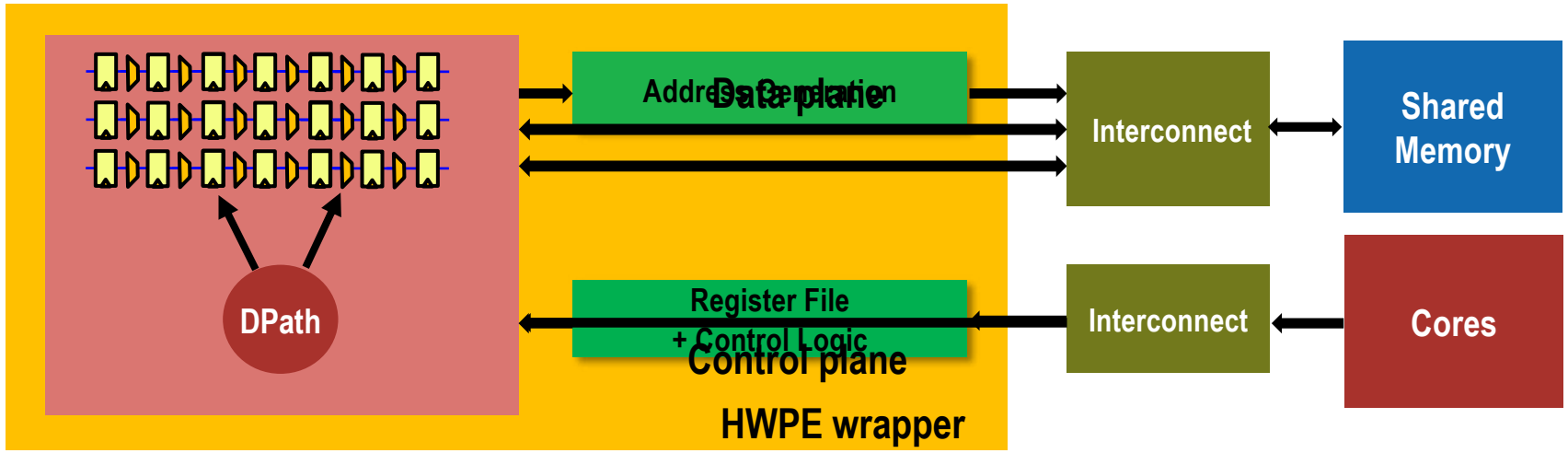
+ FFT, PCA, Mat-inv,…

Sub-pJ/W Accelerator; Tightly-coupled HW Compute Engine



Acceleration with flexibility: zero-copy HW-SW cooperation

Hardware Processing Engines (HWPEs)



HWPE efficiency

1. Specialized datapath (e.g. systolic MAC) & internal storage (e.g. linebuffer, accum-regs)
2. Dedicated control (no I-fetch) with shadow registers (overlapped config-exec)
3. Specialized high-BW interco into L1 (on data-plane)

More HWPE Efficiency: Extreme Quantization

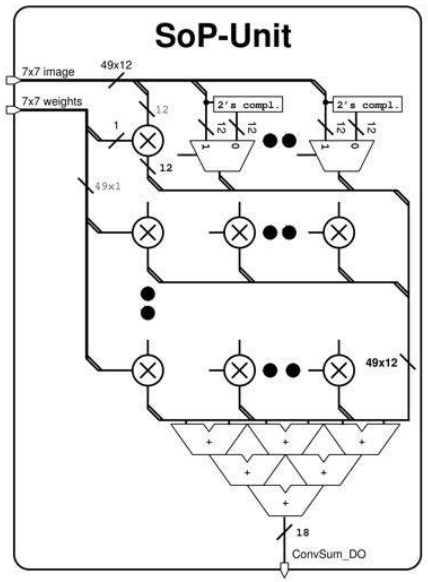
Low(er) precision: 8→4→2

Model	Bit-width	Top-1 error
ResNet-18 ref	32	31.73%
INQ	5	31.02%
INQ	4	31.11%
INQ	3	31.92%
INQ	2 (ternary)	33.98%

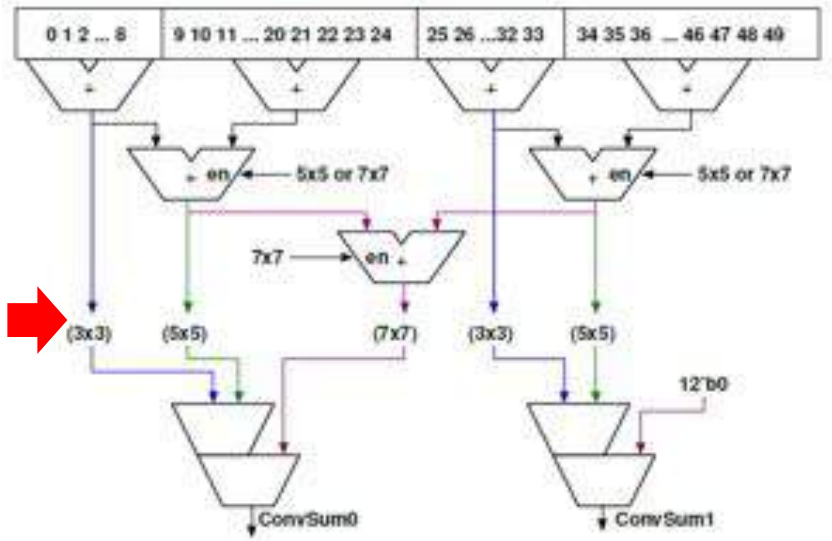
SOA INQ retraining

2.2% loss → 0% with 20% larger net

MULT → MUX



Equivalent for 7x7 SoP



ImageBank

Image Mapping (3x3, 5x5, 7x7)

1 MAC Op = 2 Op (1 Op for the “sign-reverse”, 1 Op for the add).

From +/-1 Binarization to XNORs

$$\mathrm{y}(k_{out}) = \mathrm{binarize}_{\pm 1}\left(\mathbf{b}_{k_{out}} + \sum_{k_{in}}\left(\mathbf{W}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$

XNOR

$$\mathrm{binarize}_{\pm 1}(t) = \mathrm{sign}\left(\gamma \frac{t-\mu}{\sigma} + \beta\right)$$

$$\mathrm{binarize}_{0,1}(t) = \begin{cases} 1 \text{ if } t \geq -\kappa/\lambda \doteq \tau, \text{ else } 0 & (\text{when } \lambda > 0) \\ 1 \text{ if } t \leq -\kappa/\lambda \doteq \tau, \text{ else } 0 & (\text{when } \lambda < 0) \end{cases}$$

$$\mathrm{y}(k_{out}) = \mathrm{binarize}_{0,1}\left(\sum_{k_{in}}\left(\mathbf{W}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$

Thresholding

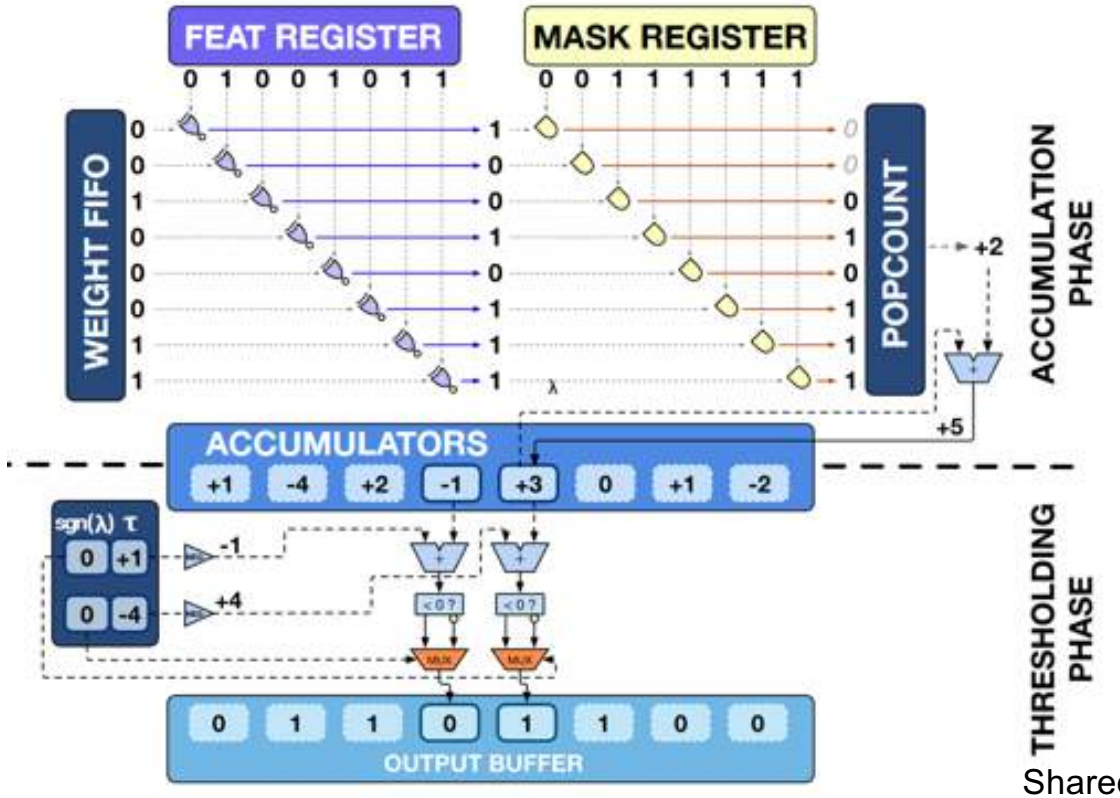
Multi-bit accumulation

Binary product → XOR

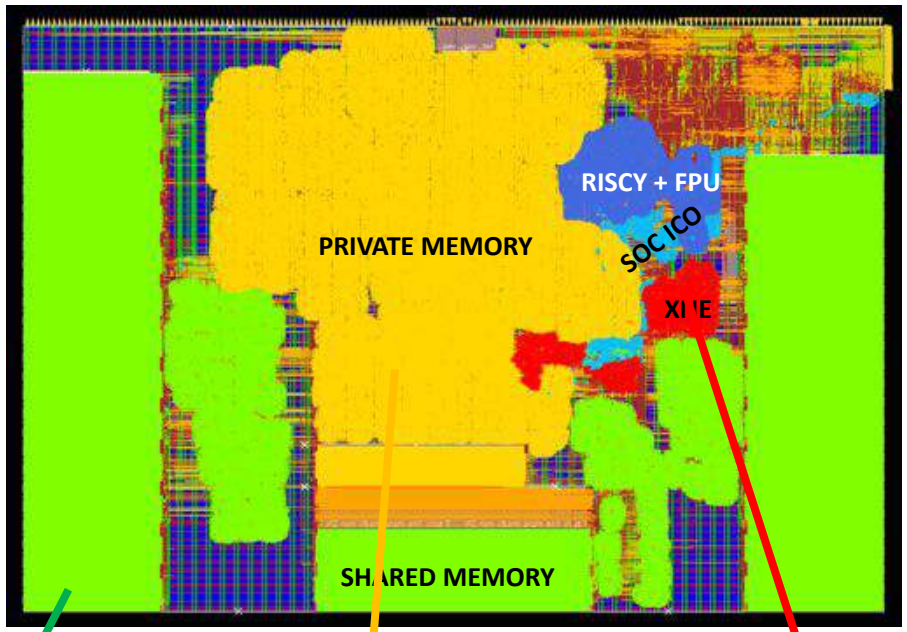
A	B	out
-1	-1	+1
-1	+1	-1
+1	-1	-1
+1	+1	+1

A	B	out
0	0	1
0	1	0
1	0	0
1	1	1

XNE: XNOR Neural Engine



Quentin in GlobalFoundries 22FDX



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

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

XNE area is **~14000 um2** (71 KGE, 72% Riscy+FPU)

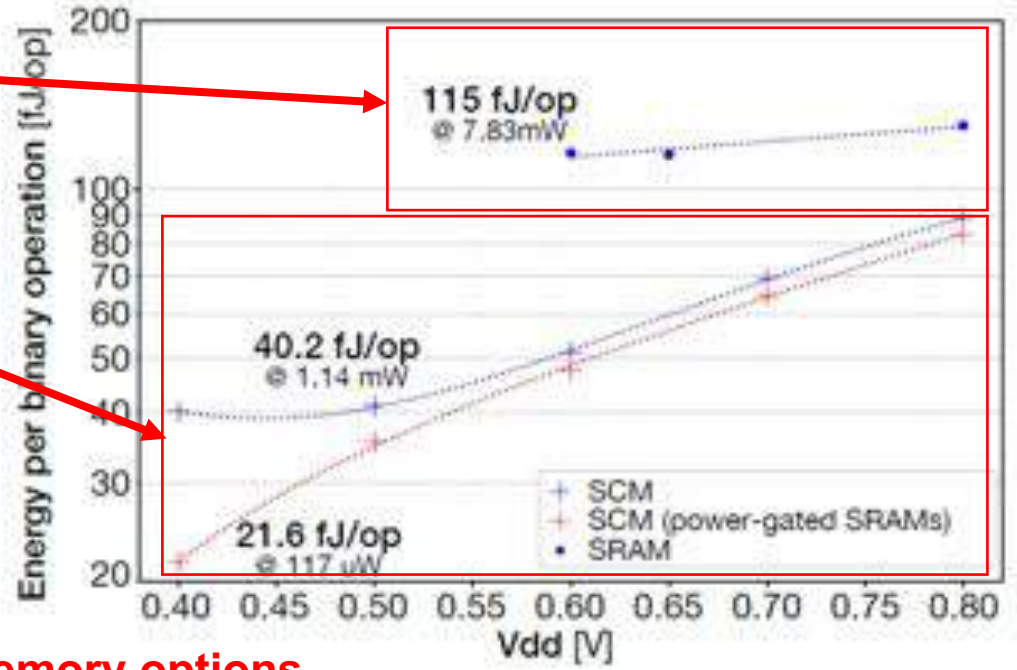
BINCONV: Binary dot-product and thresholding logic array

XNE Energy Efficiency

22 FDX measured silicon

With SRAMs, max eff
@ 0.65V **8.7 Top/s/W**

With SCMs, max eff
@ 0.5V **46.3 Top/s/W**



Note: All Memory on chip (max:MBs)

The importance of on-chip memory options
L1 SCM, L2 high-density, low leakgage SRAM (activations), MRAM (weights)

But… Accuracy Loss is high even with retraining **(10%+)**
Need **flexible precision tuning!**

Flexibility needed: Binary-Based Quantization (BBQ)

QNN layer :

$$\mathbf{y}(k_{out}) = \textit{quant}\left(\sum_{k_{in}} \left(\mathbf{W}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$

INT32 accumulator

Q-bit output fmaps

M-bit weights

N-bit input fmaps

Many $M \times N$ bits products...

... but one $M \times N$ product is the superposition of $M \times N$ **1-bit** products!

$$\mathbf{y}(k_{out}) = \textit{quant}\left(\sum_{i=0..M} \sum_{j=0..N} \sum_{k_{in}} 2^i 2^j \left(\mathbf{W}_{\mathbf{bin}}(k_{out}, k_{in}) \otimes \mathbf{x}_{\mathbf{bin}}(k_{in})\right)\right)$$

power-of-2 scaling factors

Q-bit output fmaps

1-bit weights

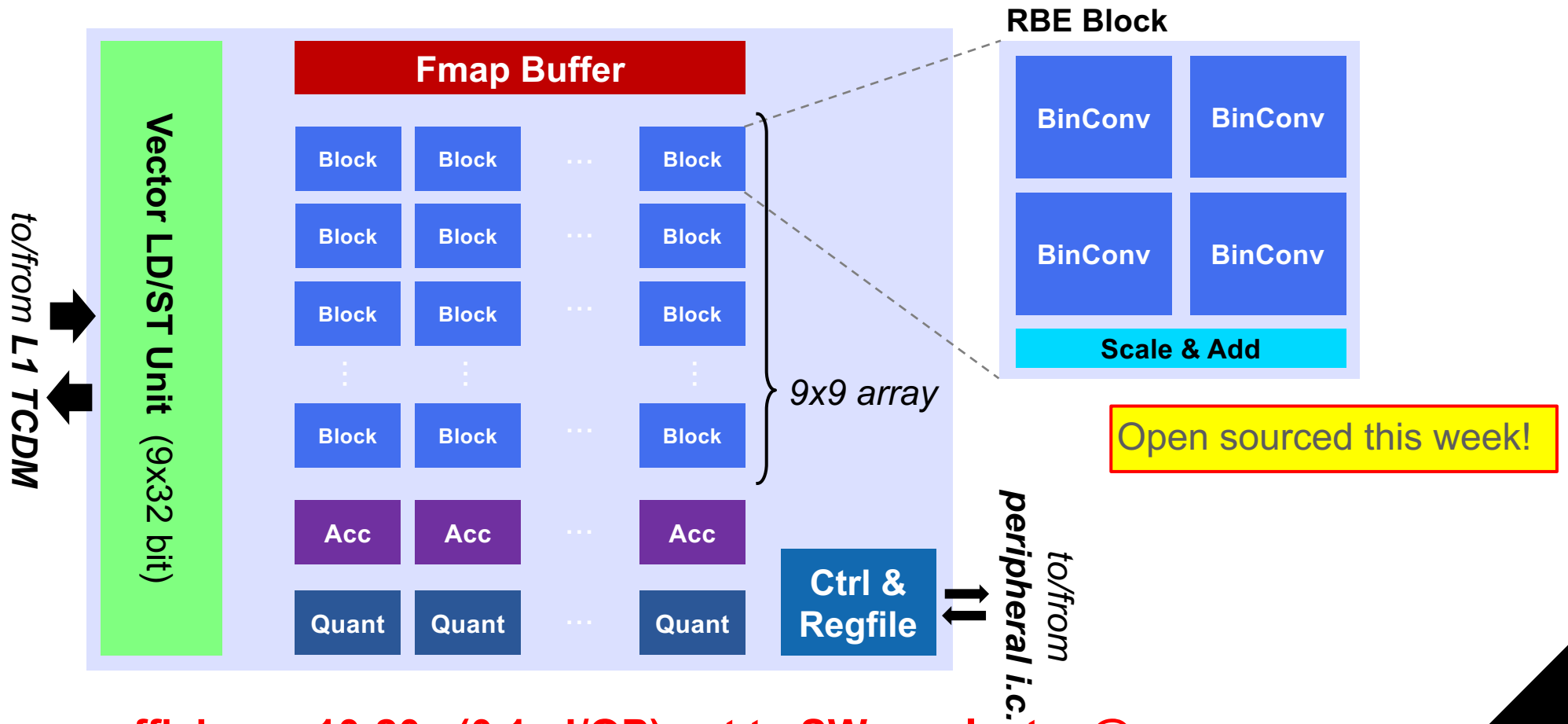
1-bit input fmaps

One quantized NN can be emulated by superposition of power-of-2 weighted $M \times N$ binary NN

Reconfigurable Binary Engine

$$y(k_{out}) = quant\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(W_{bin}(k_{out}, k_{in}) \otimes x_{bin}(k_{in})\right)\right)$$

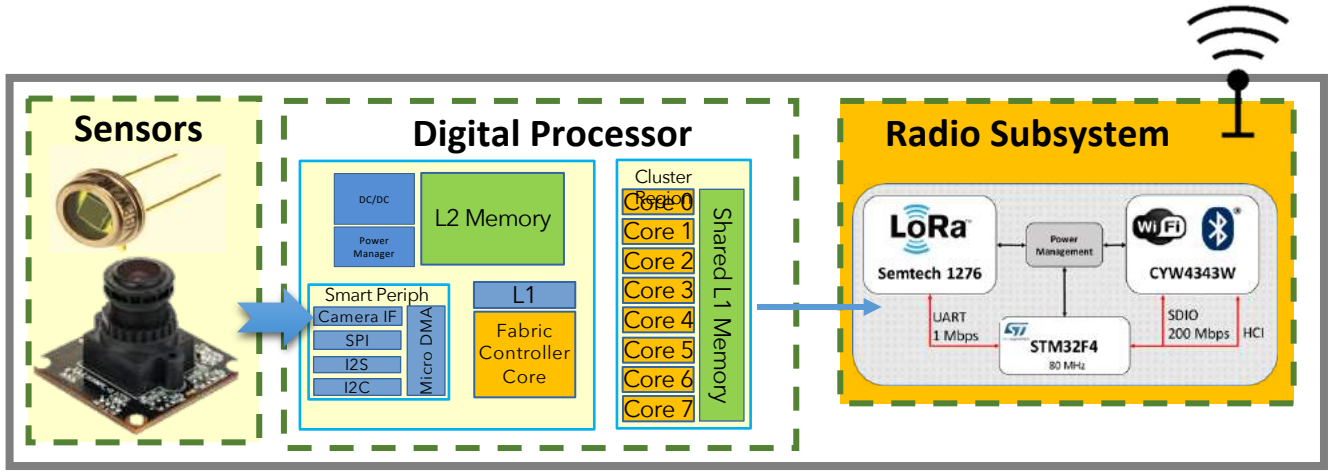
e.g. a 3x3 conv
with **N**=4 bits



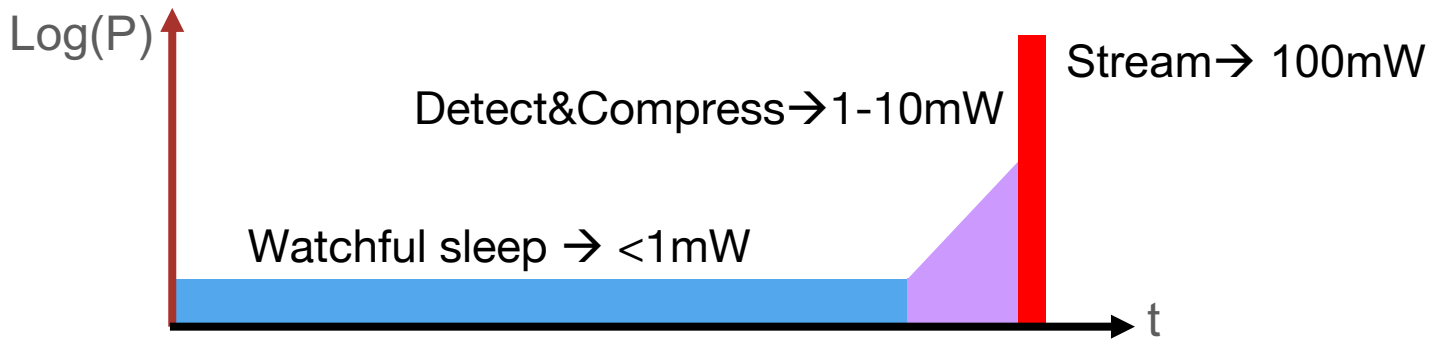
Energy efficiency 10-20x (0.1pJ/OP) wrt to SW on cluster @same accuracy

Towards In-Sensor: Achieving sub-mW average power?

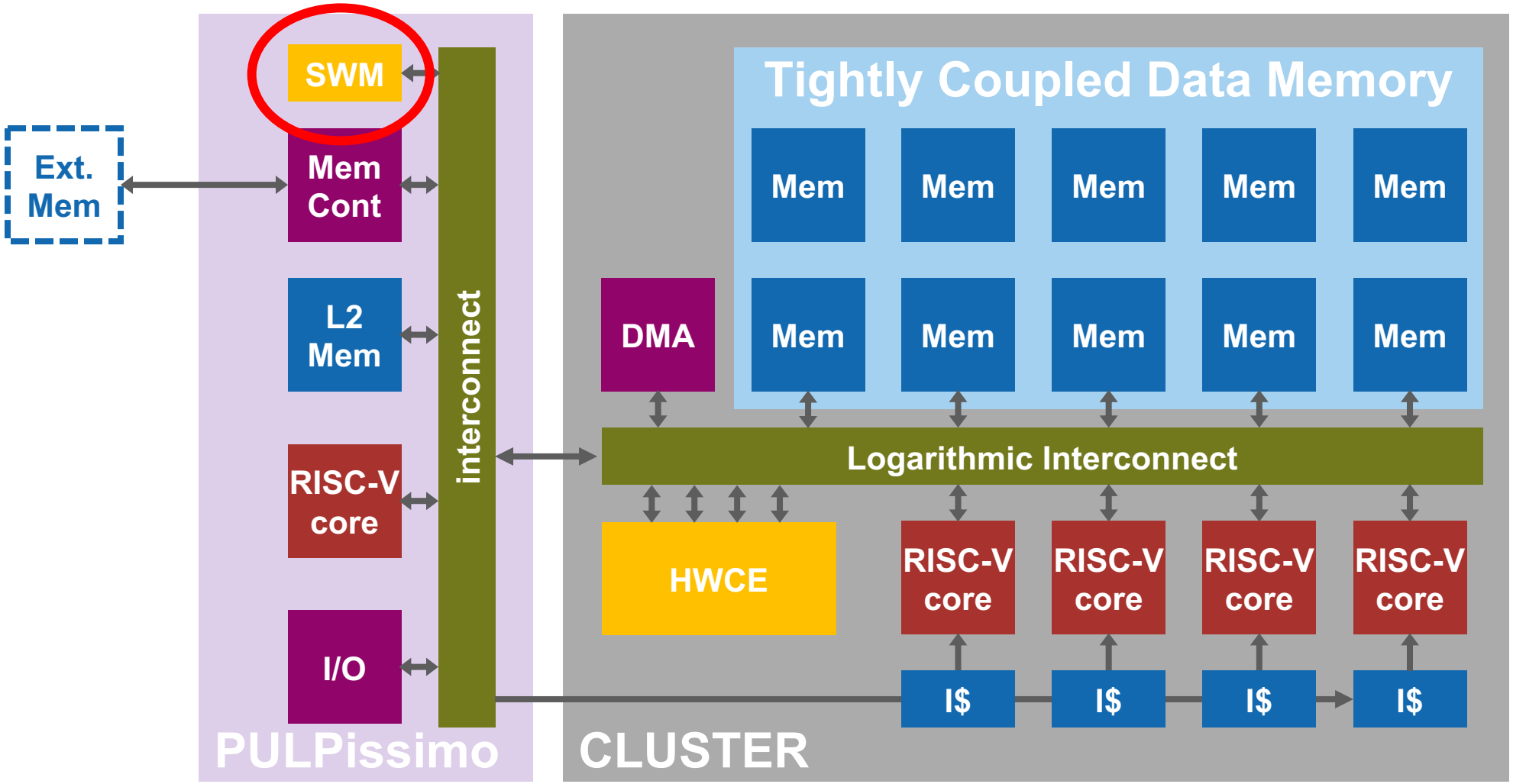
1mW average power with 10mW active power (10GOPS @ 1pJ/OP) → **sub mW sleep**



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

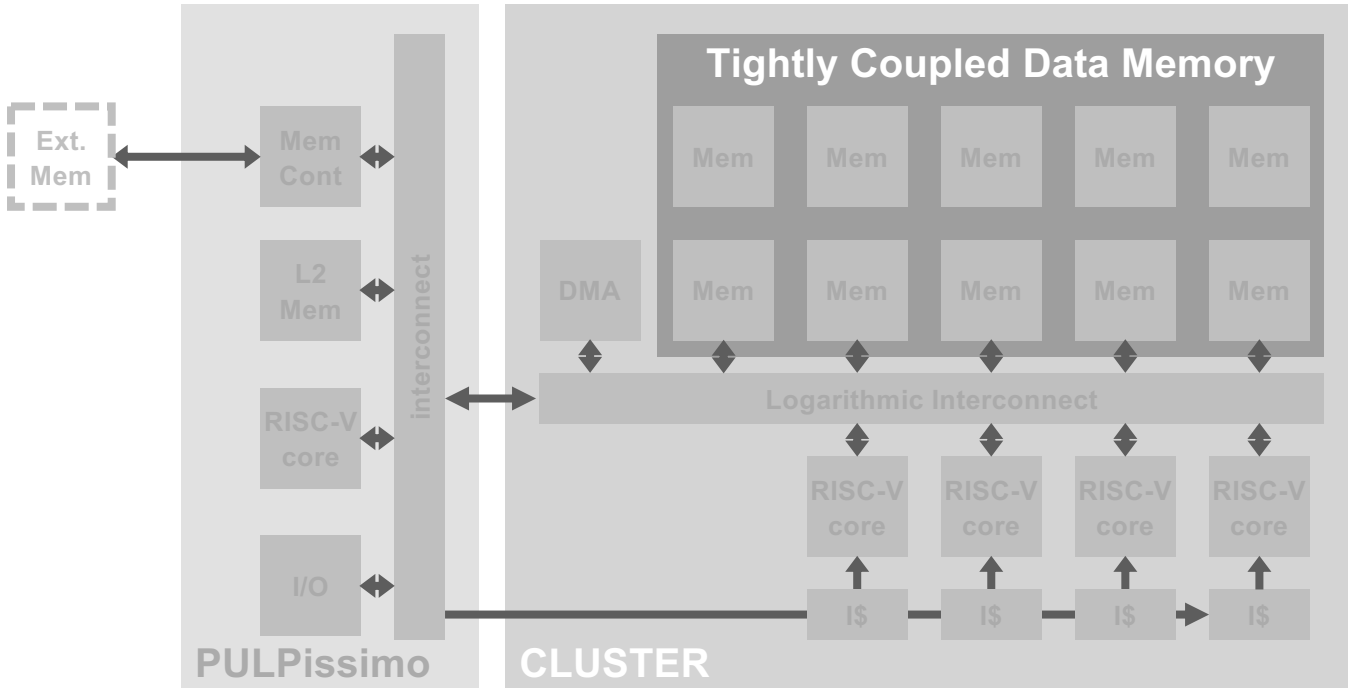


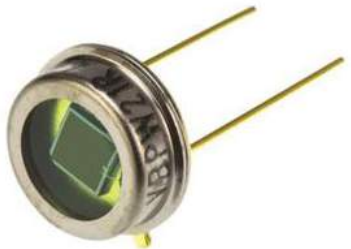
Need µW-range always-on Intelligence

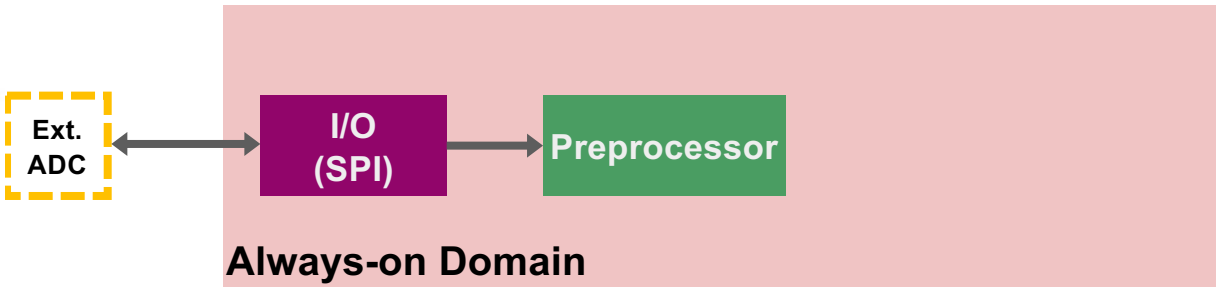


Smart Wakeup Module

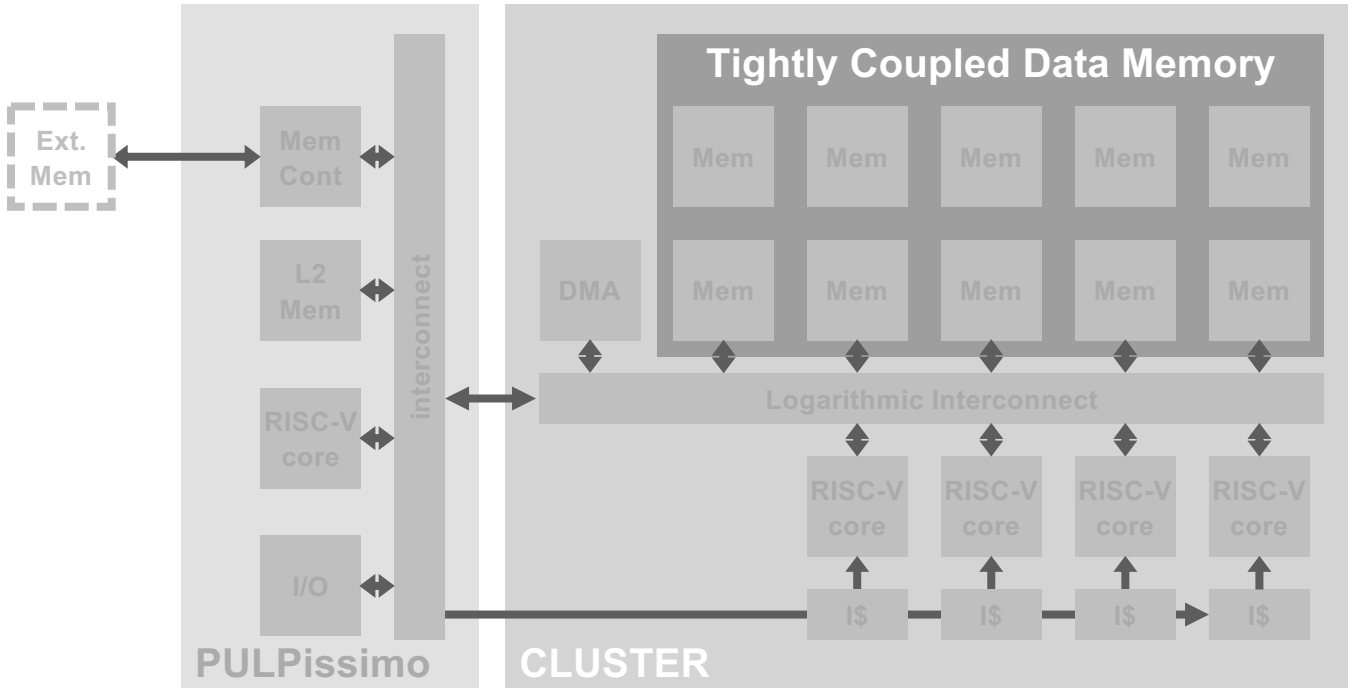
HD-Based smart Wake-Up Module

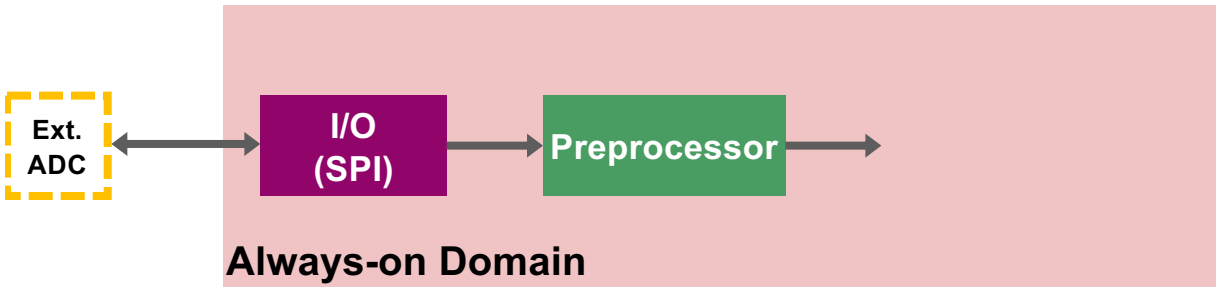




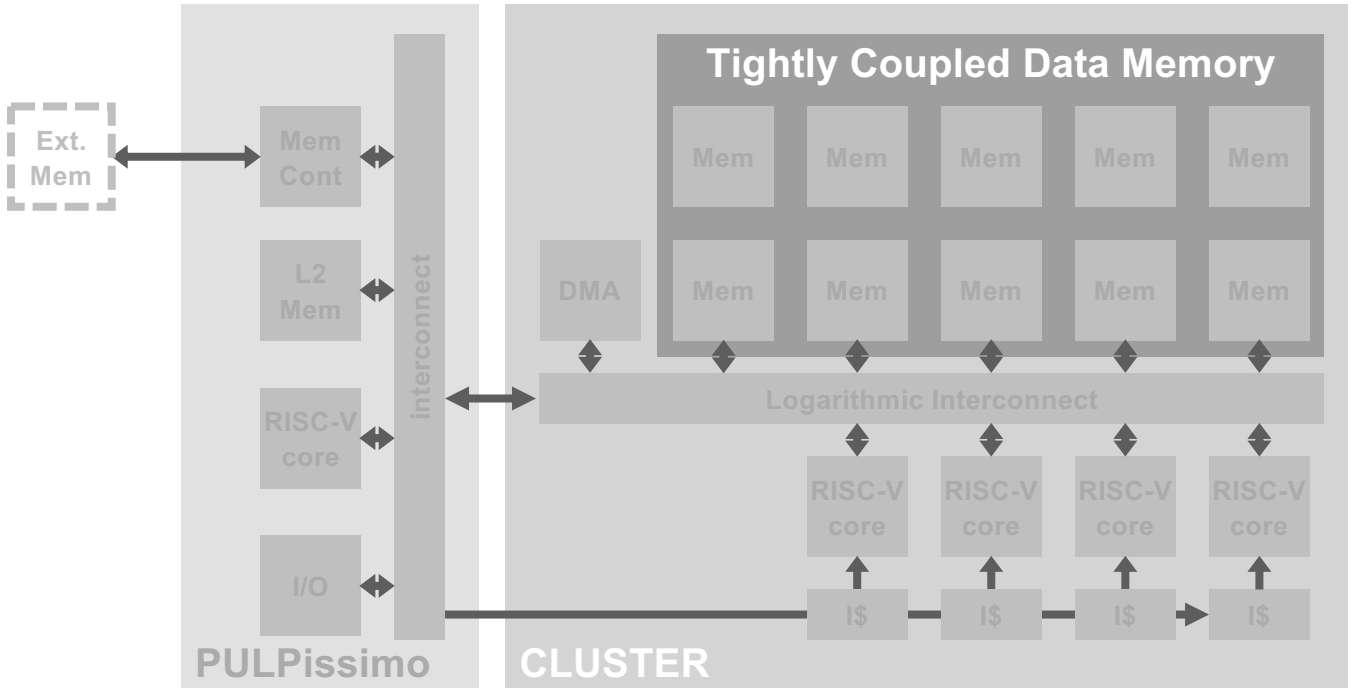


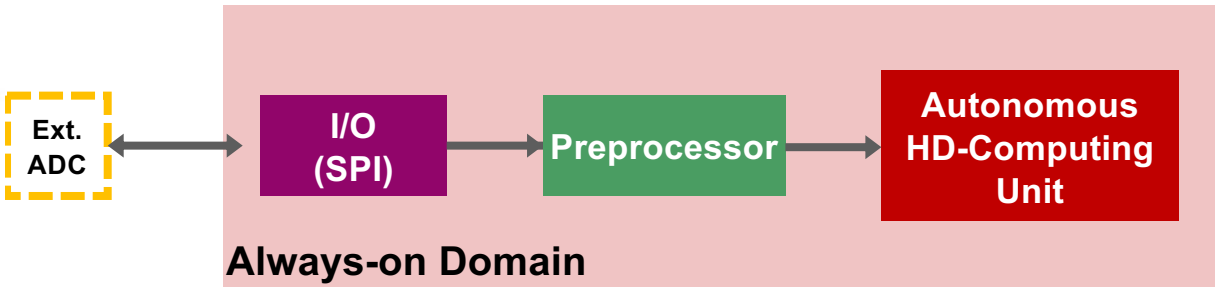
HD-Based smart Wake-Up Module



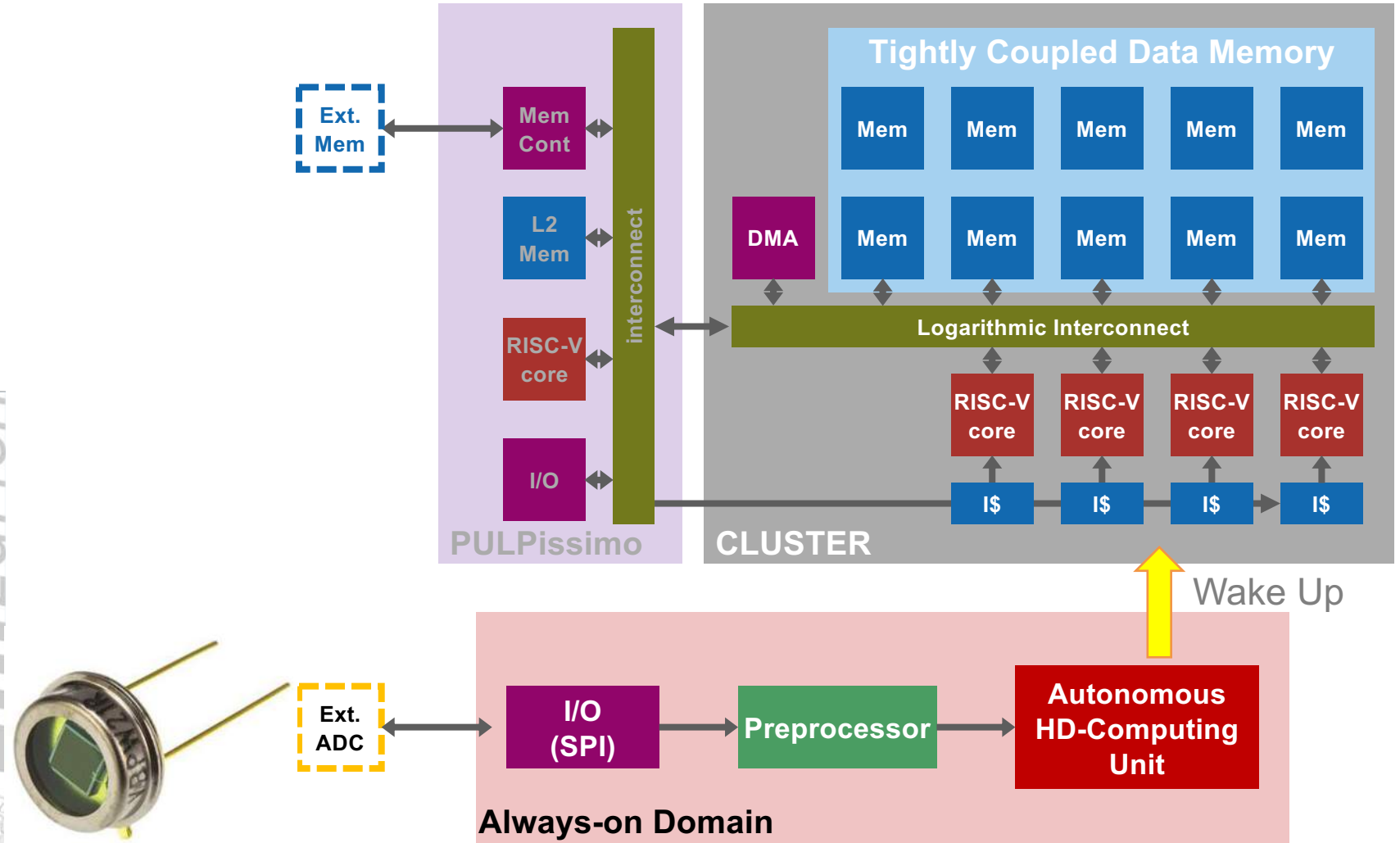


HD-Based smart Wake-Up Module

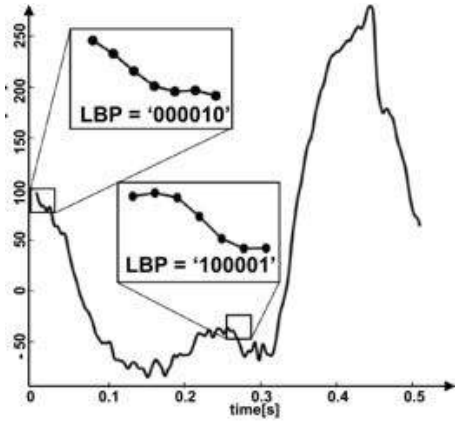




HD-Based smart Wake-Up Module



Not Only CNNs: Hyper-Dimensional Computing



Mapping →

1st 2nd 3rd 4th ... 1000th

[0 1 0 1 ..... 1]

Low Dimensional Input Data (e.g. 7-bit LBP)

[0 1 0 1 ..... 1]
[1 1 1 0 ..... 1]
[1 1 0 0 ..... 0]
[0 1 1 1 ..... 1]

HD-Encoding

- Component-wise Majority
- XOR
- Permutation

Search Vector

[1 1 0 1 ..... 1]

Similarity Search
(e.g. Hamming Distance)

Prototype Vectors

Associative Memory

[0 1 0 1 ..... 1]
[1 1 1 0 ..... 1]
[1 1 0 0 ..... 0]
[0 1 1 1 ..... 1]
[1 1 1 1 ..... 1]
[0 1 0 1 ..... 1]
[0 1 0 1 ..... 1]

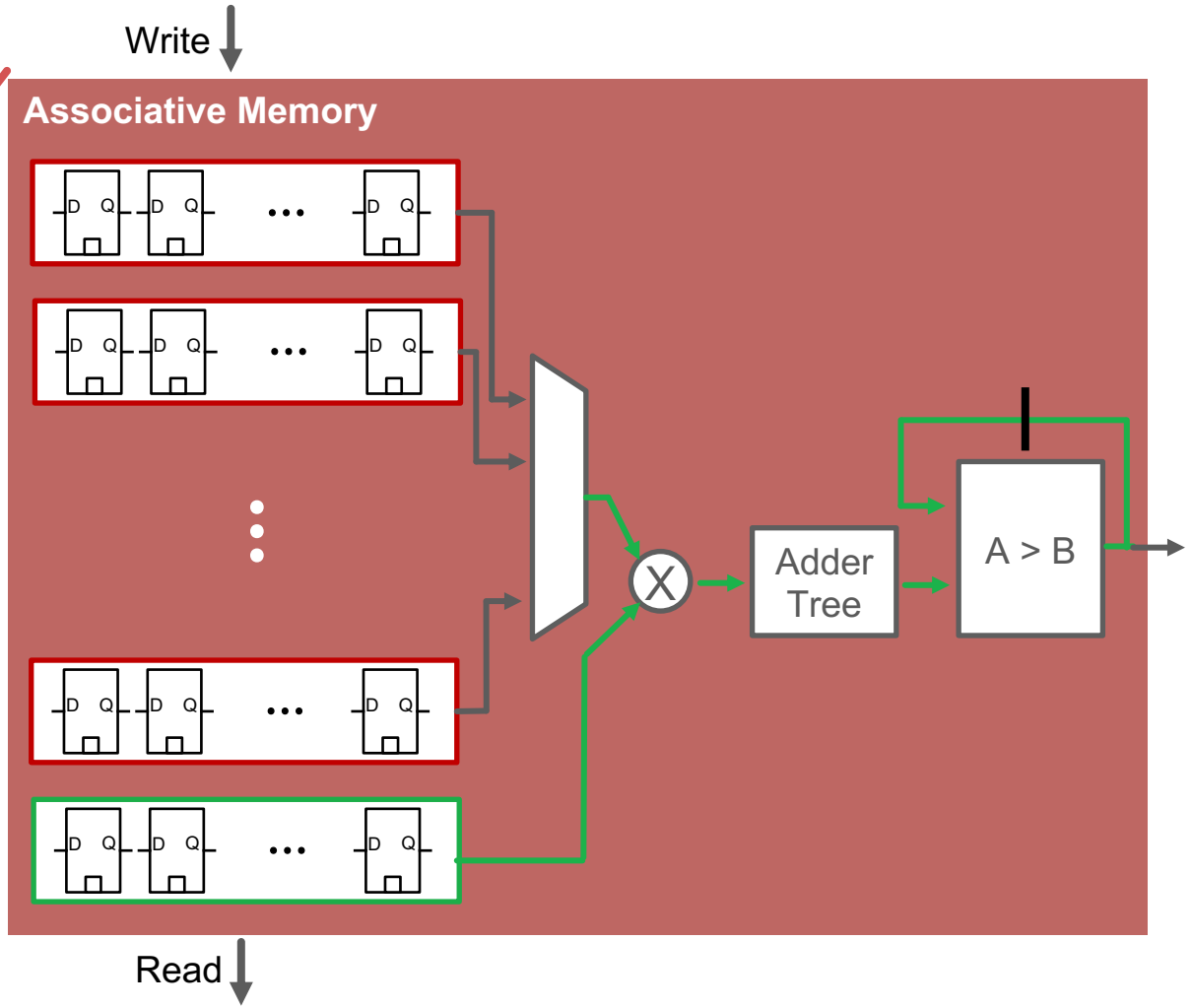
Highly parallel, fault-tolerand binary operators, assoc-min-distance search → Merge storage & computation i.e. **In-memory computing**

In-memory Hyperdimensional Computing

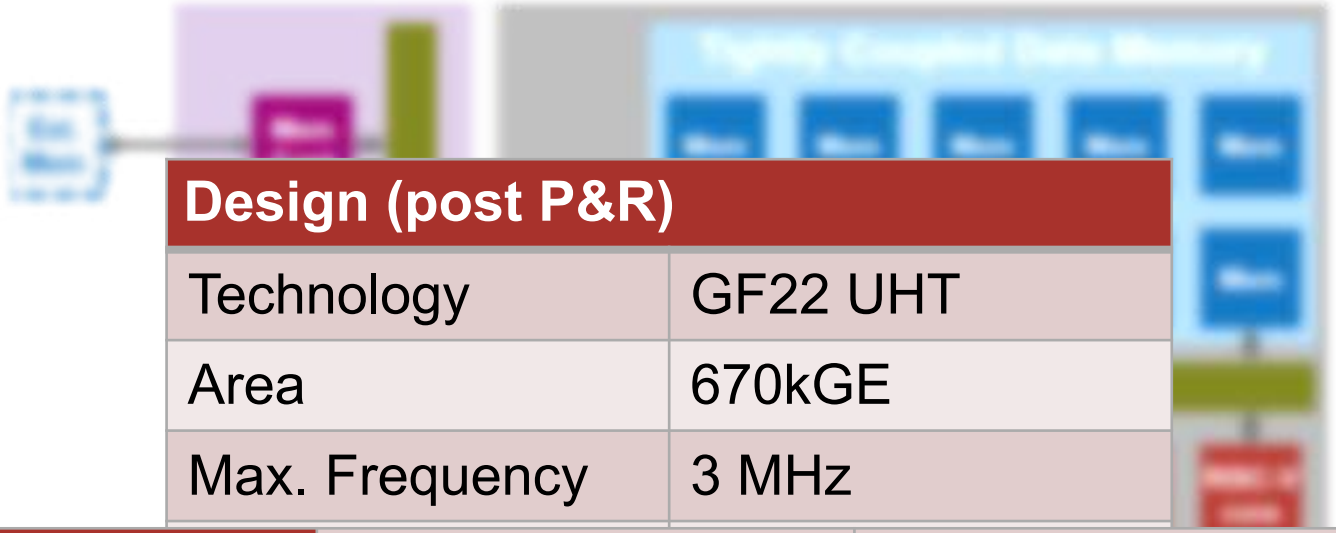
Associative Memory
(latch based SCM)

[0100010.....1]
[1000101.....1]
[0100101.....0]
⋮
[0100101.....0]

N_{CLASS} cycles



HD-Based smart Wake-Up Module



Design (post P&R)	
Technology	GF22 UHT
Area	670kGE
Max. Frequency	3 MHz

Implemented with lowest leakage cell library (UHVT)

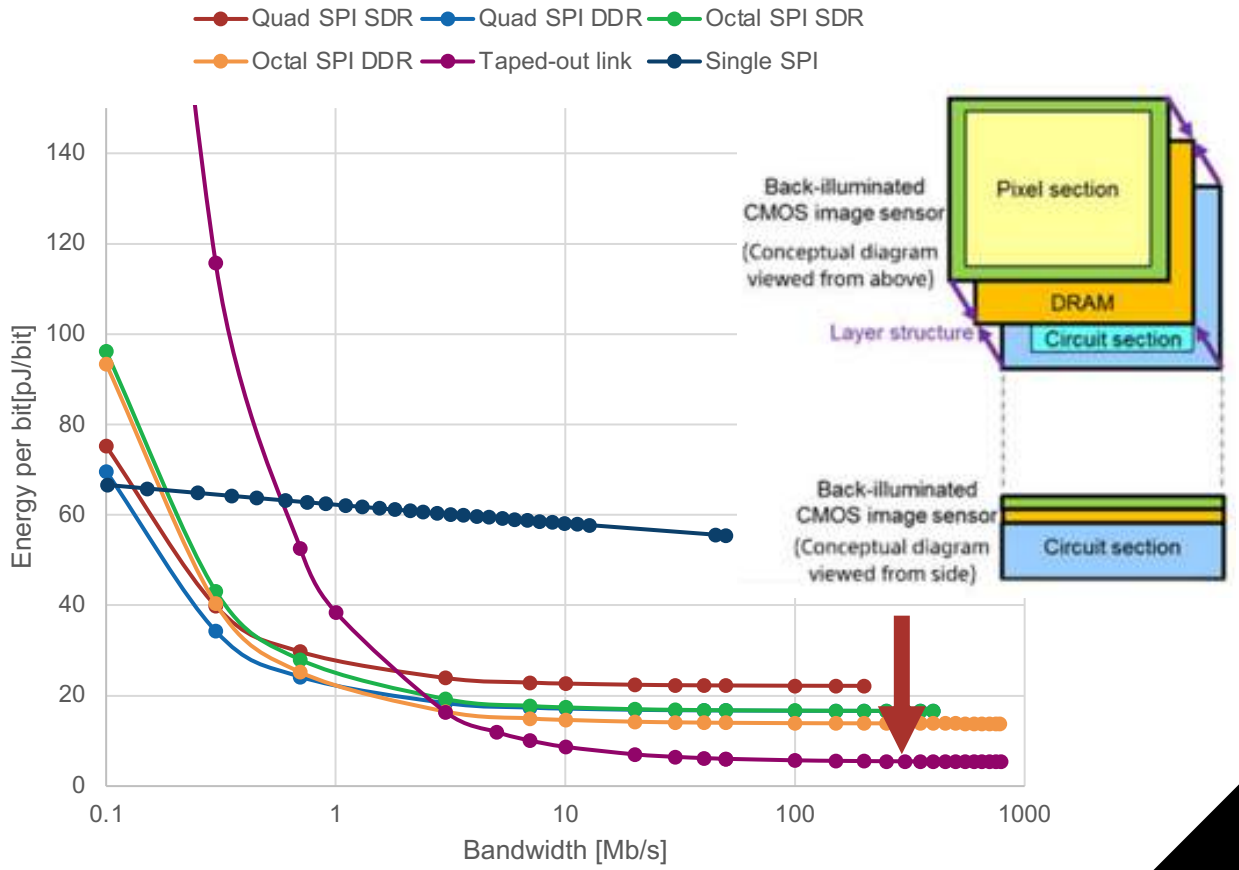
f_{clk}	32kHz	200kHz
max. sampling rate	150 SPS/Channel	1kSPS/Channel
$P_{SWU,\ dynamic}$	0.99uW	6.21uW
$P_{SWU,\ leakage}$	0.7uW	0.7uW
$P_{SPI,\ dynamic}$	1.28uW	8.00uW
$P_{SWU,\ total}$ **Measured**	**2.97uW**	**14.9uW**

To be open sourced in a few days!!

When you count mWatts, everything matters!

What about IO power? (Mem, Sensor)

- **SPIs**
  - I/O VDD=1.8V
  - fspi-max=50MHz,
  - Assuming duty-cycled operation @ various bandwidths
- **ULP serial link** (duty-cycled)
  - 10.2x less energy and 15.7x higher maximum BW compared to single SPI
  - 2.56x higher efficiency than the DDR Octal SPI @787Mbps
  - **5 → 3pJ/bit**
  - **However it's still 2mW@ 500Mbps**
- **3D integration: 0.15pJ/bit and below**



From near-sensor to in-sensor (3D IC)

Closing thoughts – Open Platform for near-sensor AI

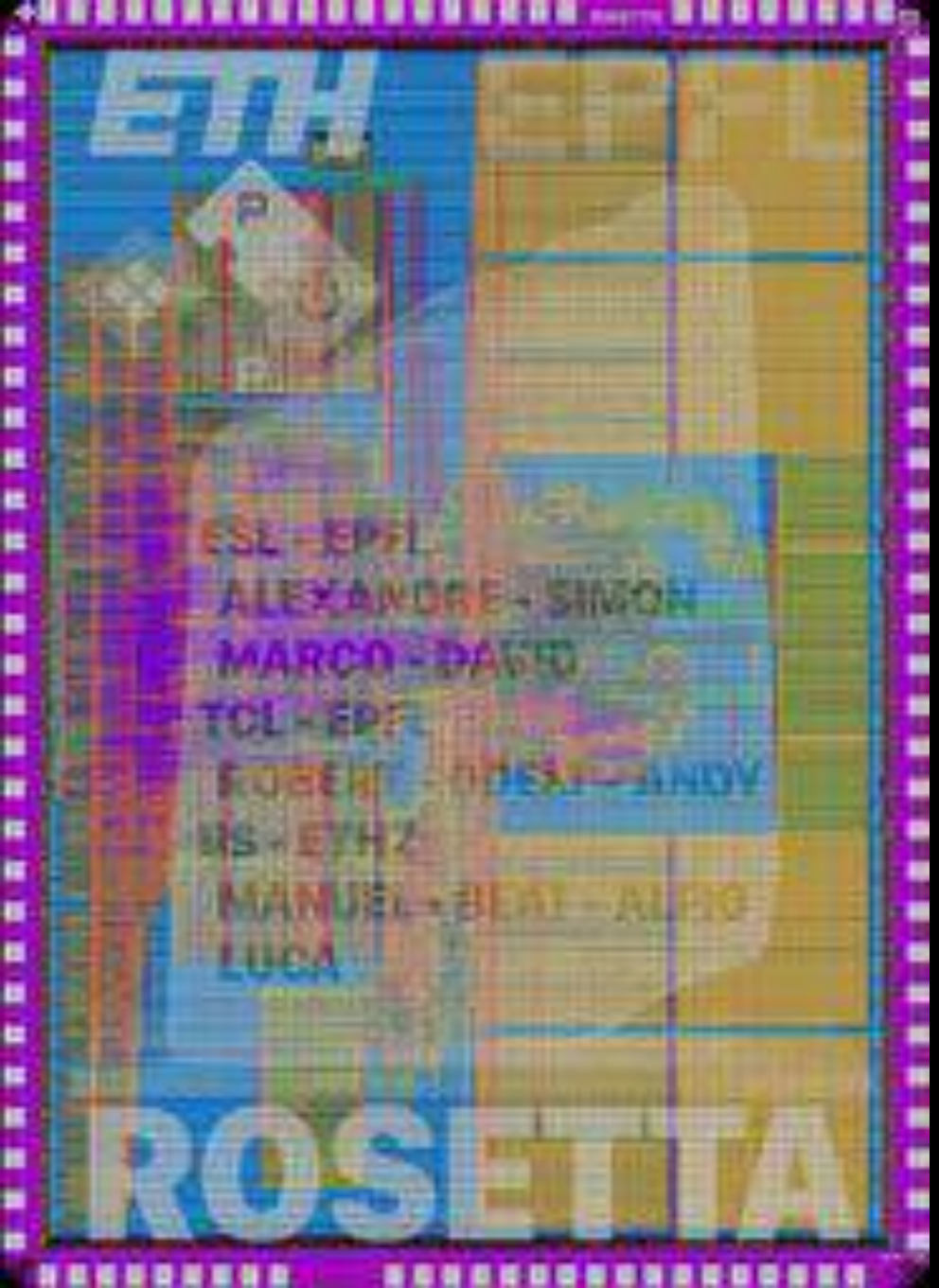
Open Platform

- 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)
  - Helps exchange of information across NDA walls
  - Great for Marketing & Training
  - More Secure, safe, auditable HW
  - Exemplary collaboration with GF (Quentin, Arnold, Vega…)

Posh Open Source Hardware (*POSH*):
An open source System on Chip (SoC) design and verification eco-system that enables cost effective design of ultra-complex SoCs

Heterogeneous & Flexible

- 1-2 orders of magnitude improvement by acceleration
  - Various flavors: number-crunching, always-on, reconfigurable
- 2 orders of magnitude improvement on IO energy (memory, sensor) needed to achieve pJ/OP @ full platform
  - 3D-IC technology is a key enabler



PULP

Parallel Ultra Low Power

Luca Benini, Davide Rossi, Andrea Borghesi, Michele Magno, Simone Benatti, Francesco Conti, Francesco Beneventi, Daniele Palossi, Giuseppe Tagliavini, Antonio Pullini, Germain Haugou, Lukas Cavigelli, Manuele Rusci, Florian Glaser, Renzo Andri, Fabio Montagna, Bjoern Forsberg, Pasquale Davide Schiavone, Alfio Di Mauro, Victor Javier Kartsch Morinigo, Tommaso Polonelli, Fabian Schuiki, Stefan Mach, Andreas Kurth, Florian Zaruba, Manuel Eggimann, Philipp Mayer, Marco Guermandi, Xiaying Wang, Michael Hersche, Robert Balas, Antonio Mastrandrea, Matheus Cavalcante, Angelo Garofalo, Alessio Burrello, Gianna Paulin, Georg Rutishauser, Andrea Cossettini, Luca Bertaccini, Maxim Mattheeuws, Samuel Riedel, Sergei Vostrikov, Vlad Niculescu, Frank K. Gurkaynak,
and many more that we forgot to mention

http://pulp-platform.org **@pulp_platform**