PULP PLATFORM

Open Source Hardware, the way it should be!

Open, Parallel Ultra-Low Power Platforms for Extreme Edge AI

Luca Benini <lbenini@iis.ee.ethz.ch>

European Research Council

European Commission

Horizon 2020
European Union funding
for Research & Innovation

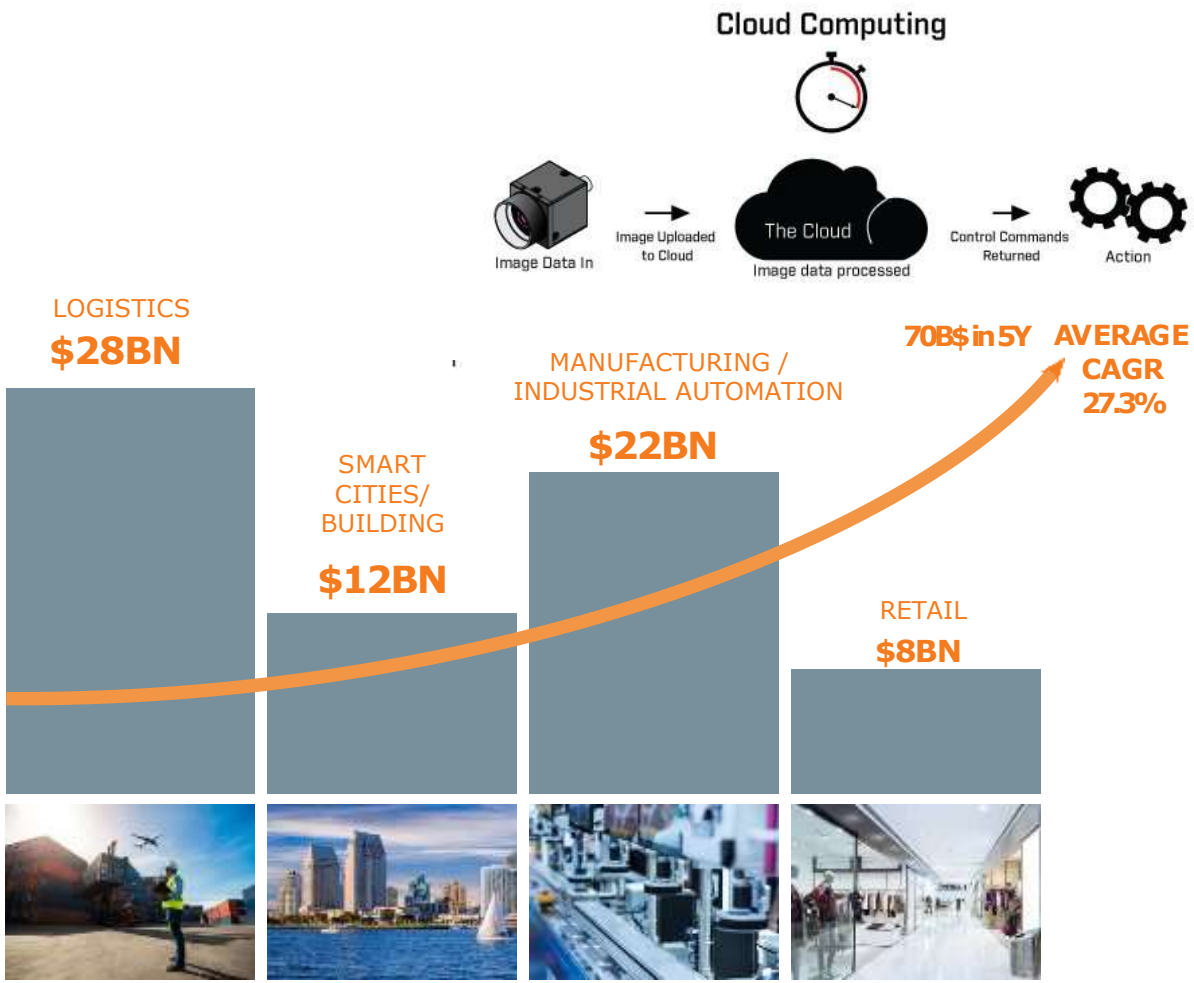
nano-tera.ch

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

http://pulp-platform.org @pulp_platform

ETH zürich

Cloud → Edge → Extreme Edge AI a.k.a. TinyML



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?*

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

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

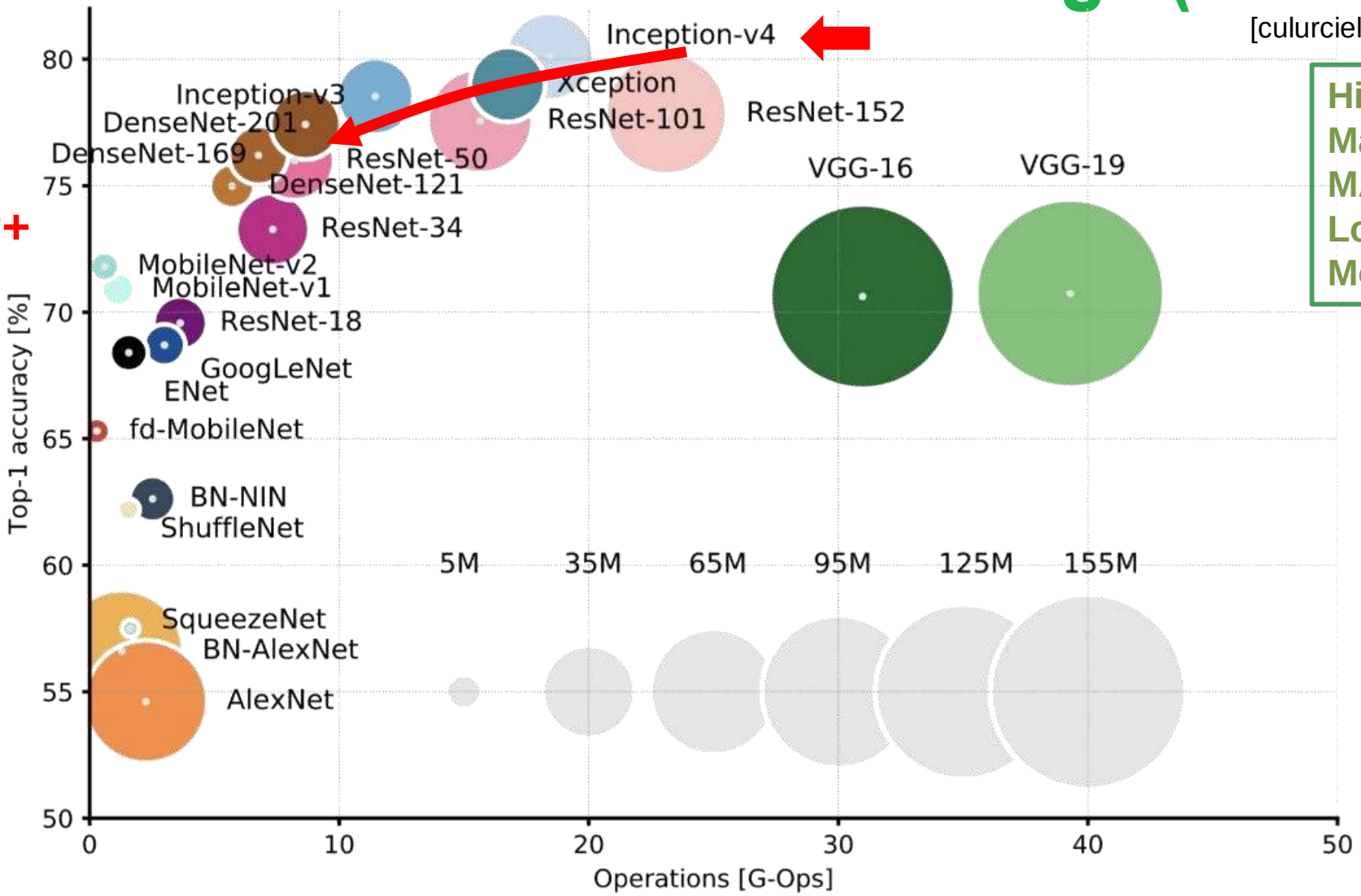
Extreme edge AI challenge

AI capabilities in the power envelope of an MCU:

100mW peak (1mW avg)

AI Workloads from Cloud to Edge (Extreme?)

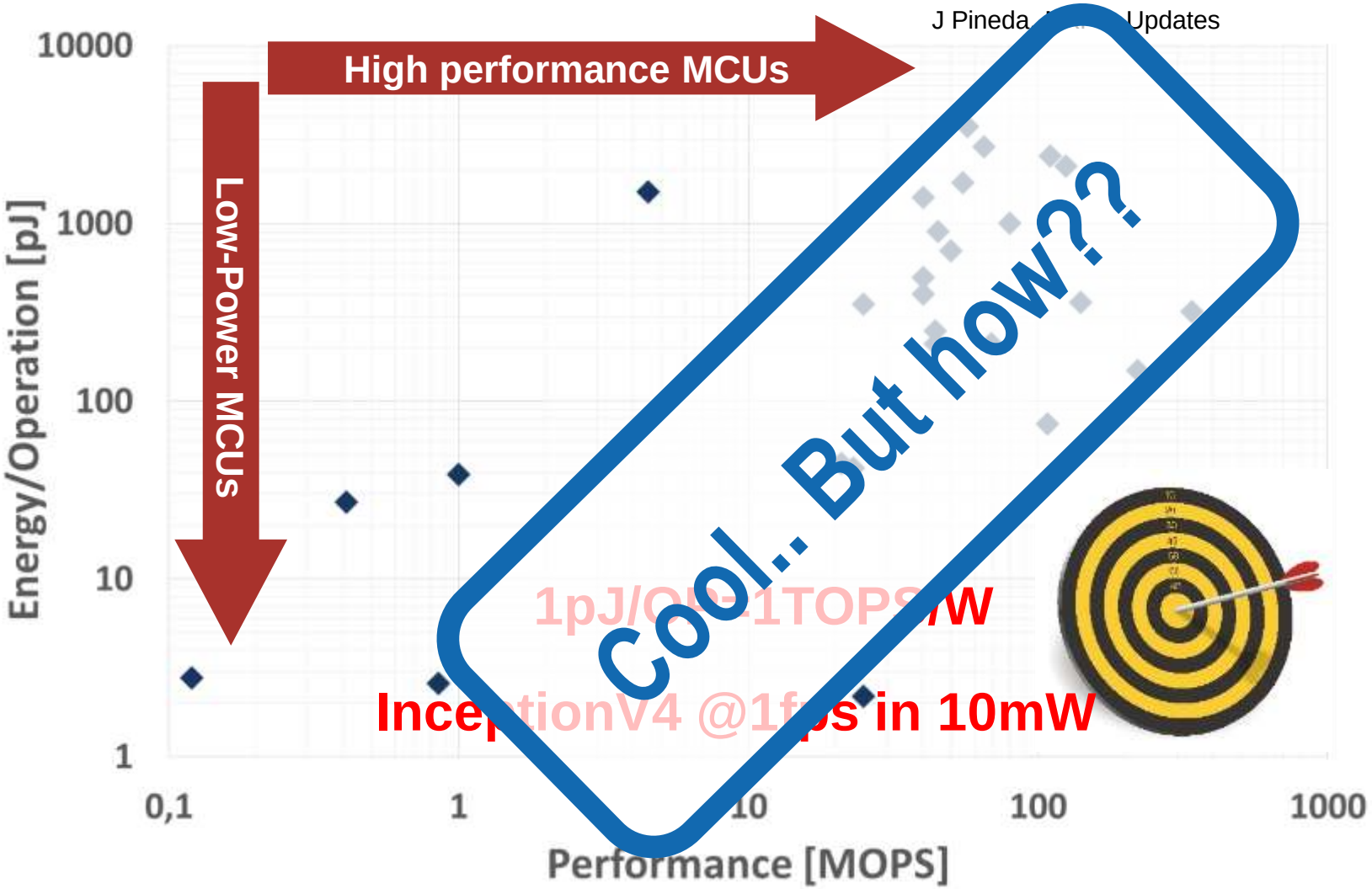
[culurciello18]



GOP+
MB+

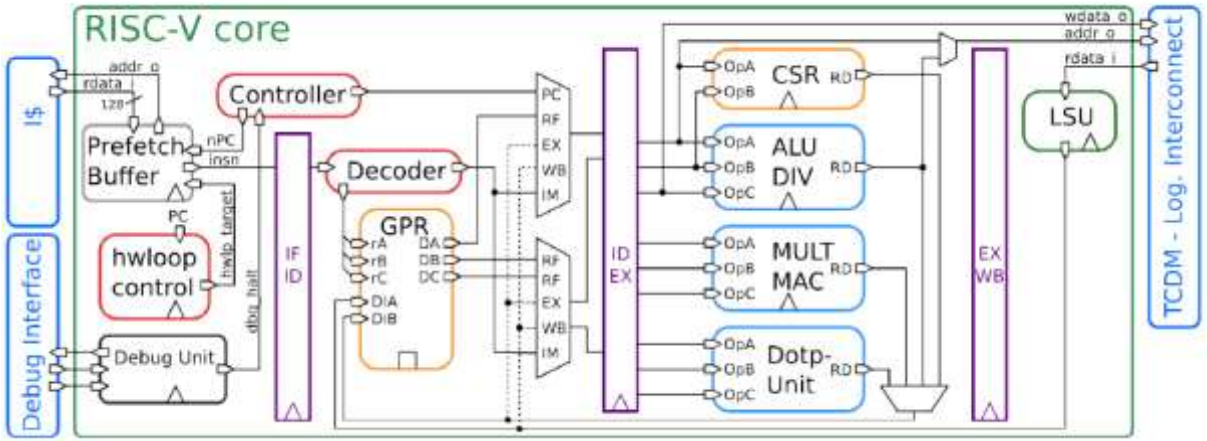
- High OP/B ratio
- Massive Parallelism
- MAC-dominated
- Low precision OK
- Model redundancy

Energy efficiency @ GOPS is THE Challenge

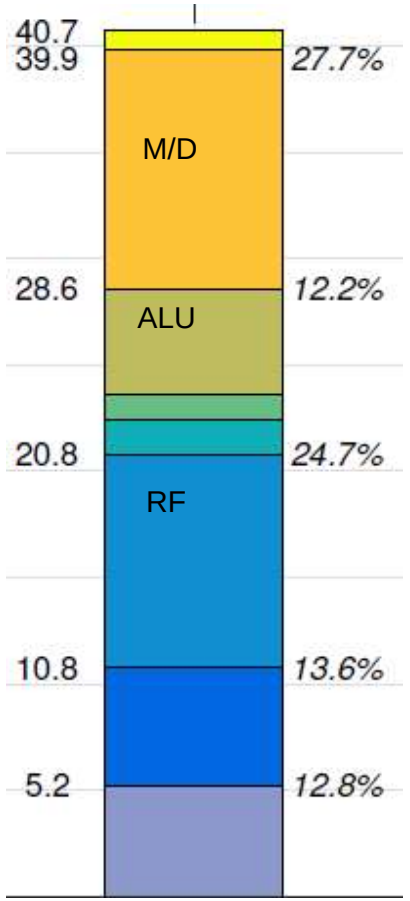


RI5CY – An Open MCU-class RISC-V Core for EE-AI

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



40 kGE
70% RF+DP

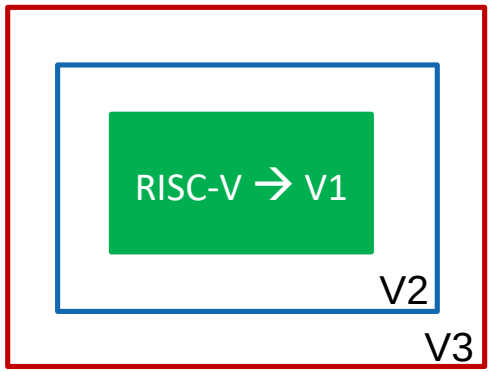


RISC-V ISA is extensible *by construction* (great!)

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)

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

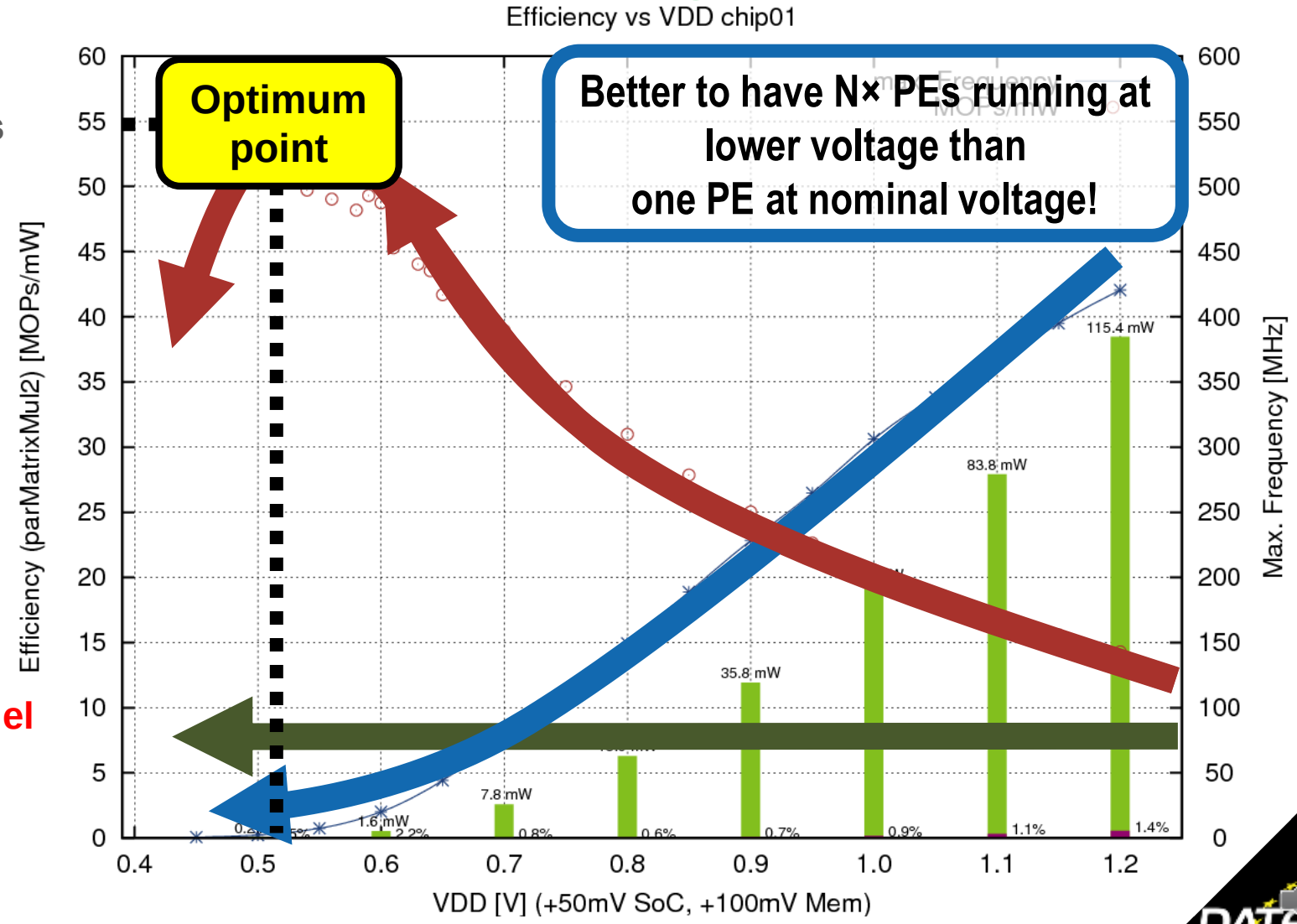
Nice – But what about the GOPS?
Faster+Superscalar is not efficient!

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

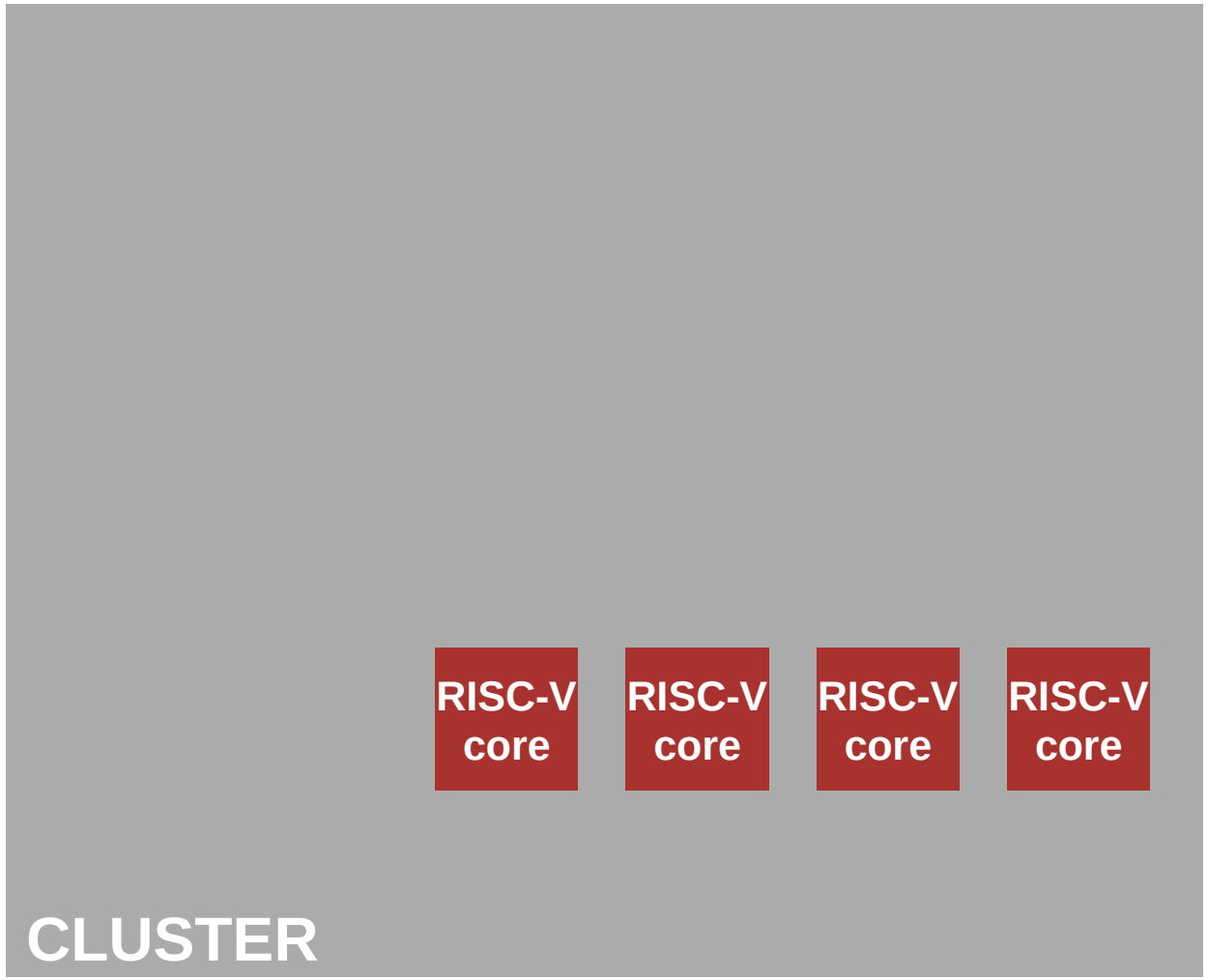
ML & Parallel, Near-threshold: a Marriage Made in Heaven

- 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

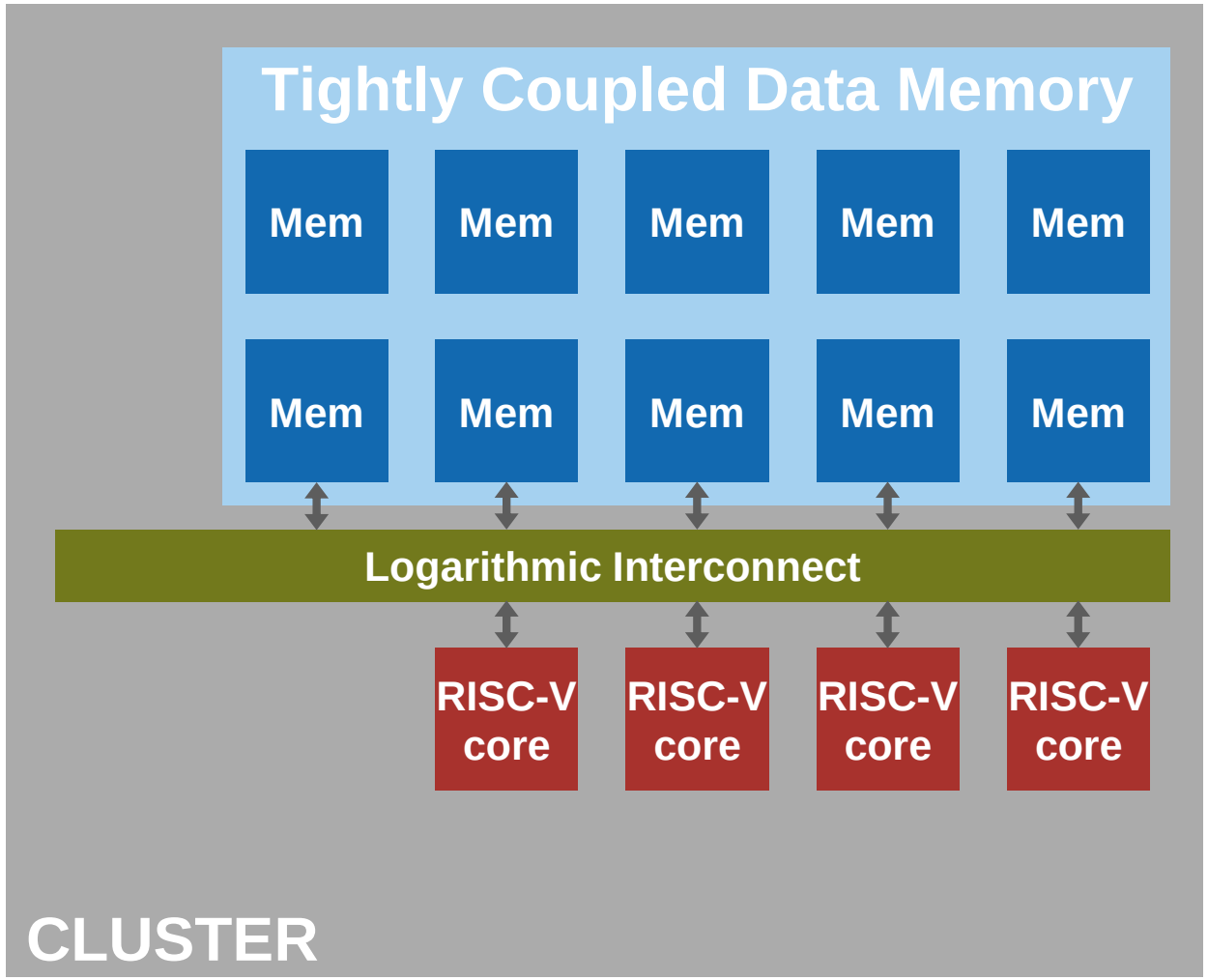
ML is massively parallel and scales well (P/S ↑ with NN size)



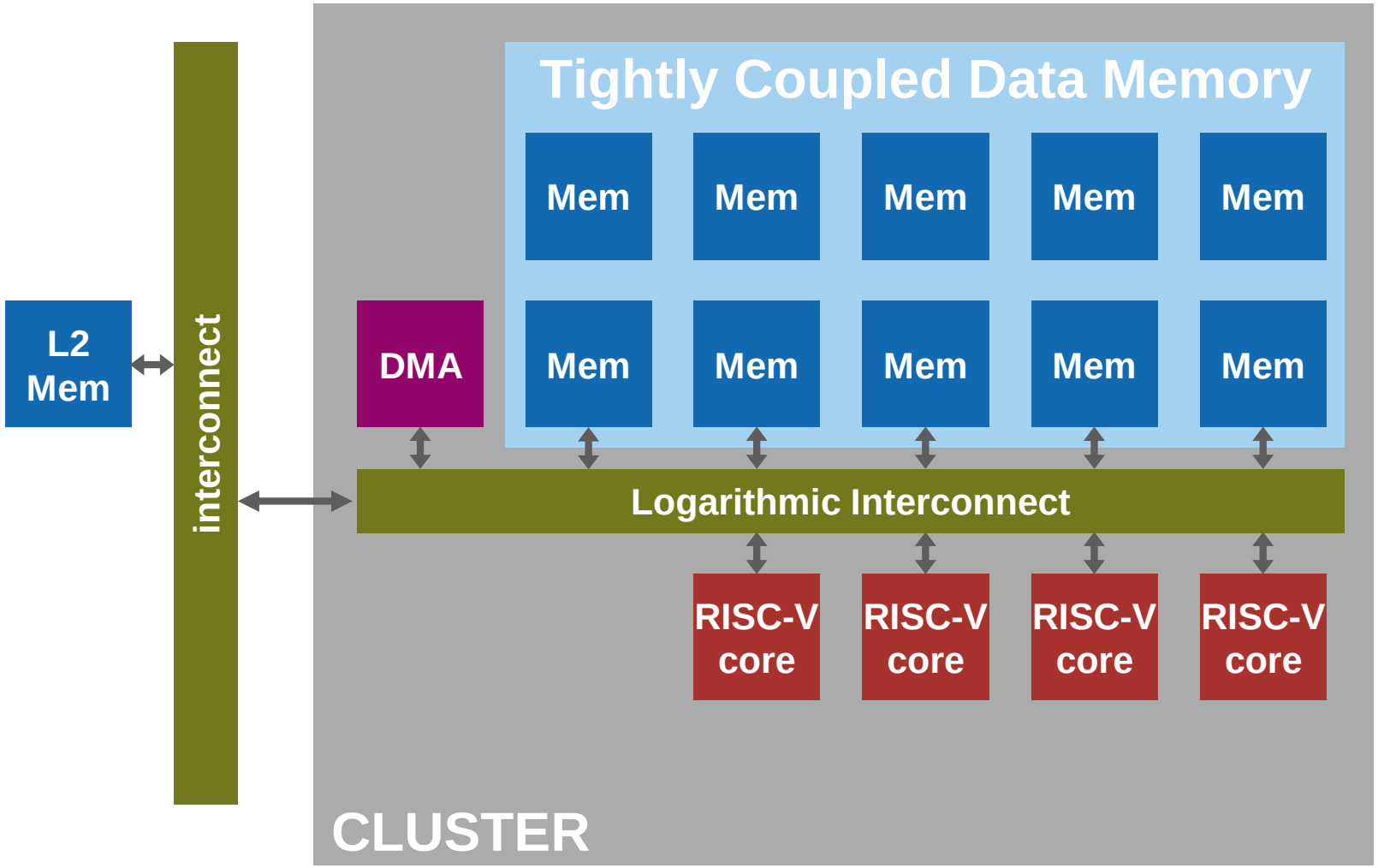
Multiple RI5CY Cores (1-16)



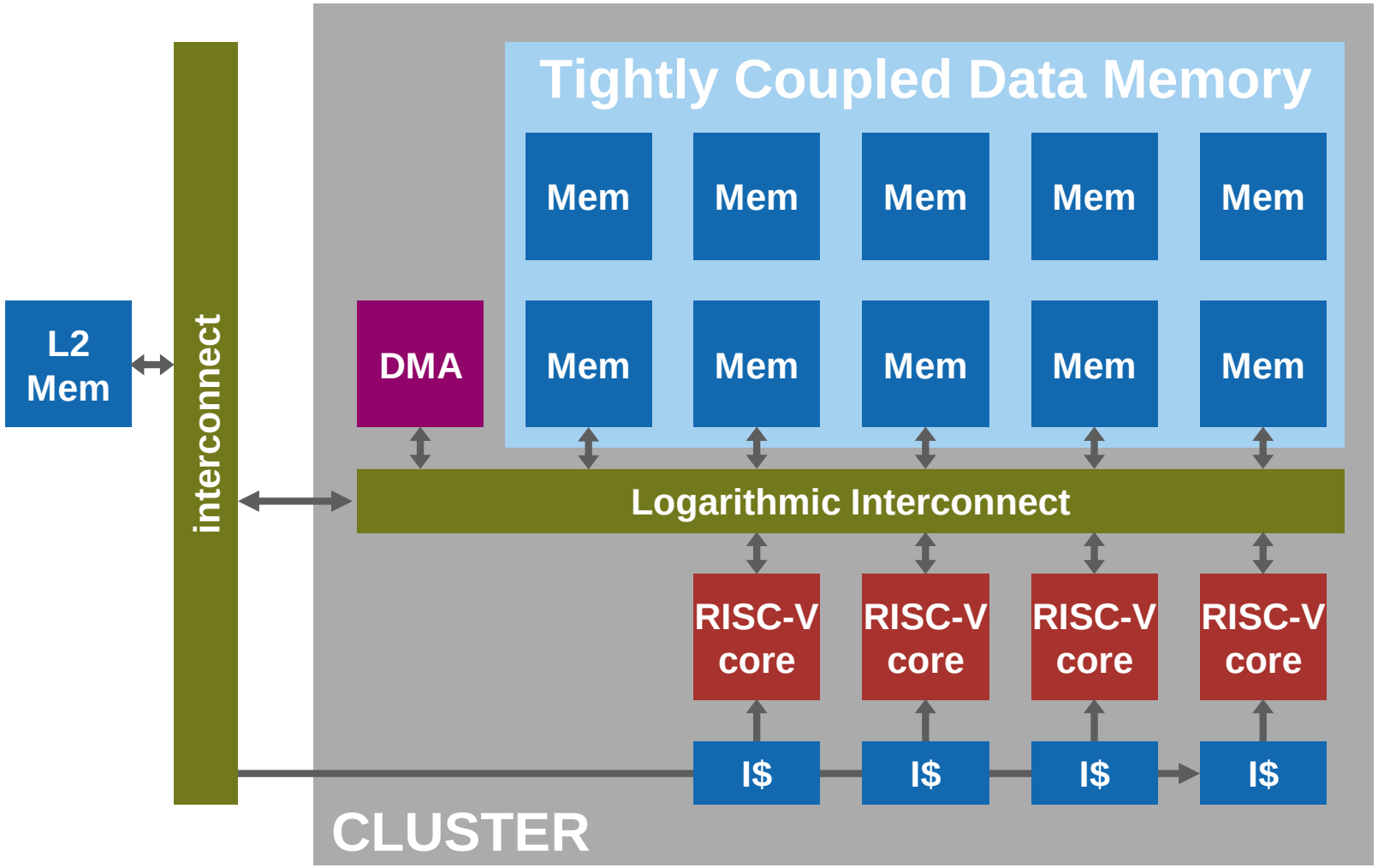
Low-Latency Shared TCDM



DMA for data transfers from/to L2

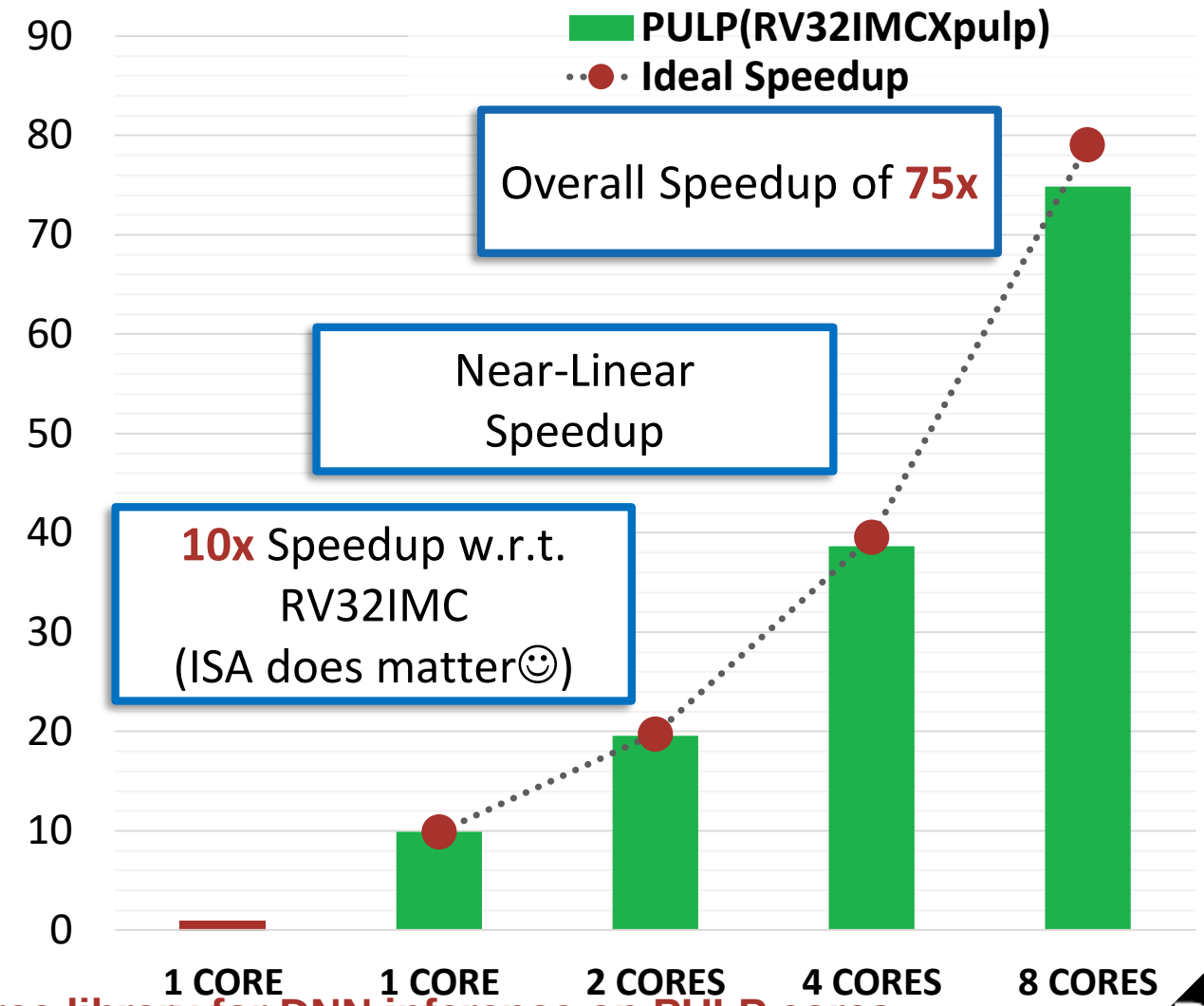


Shared instruction cache with private “loop buffer”

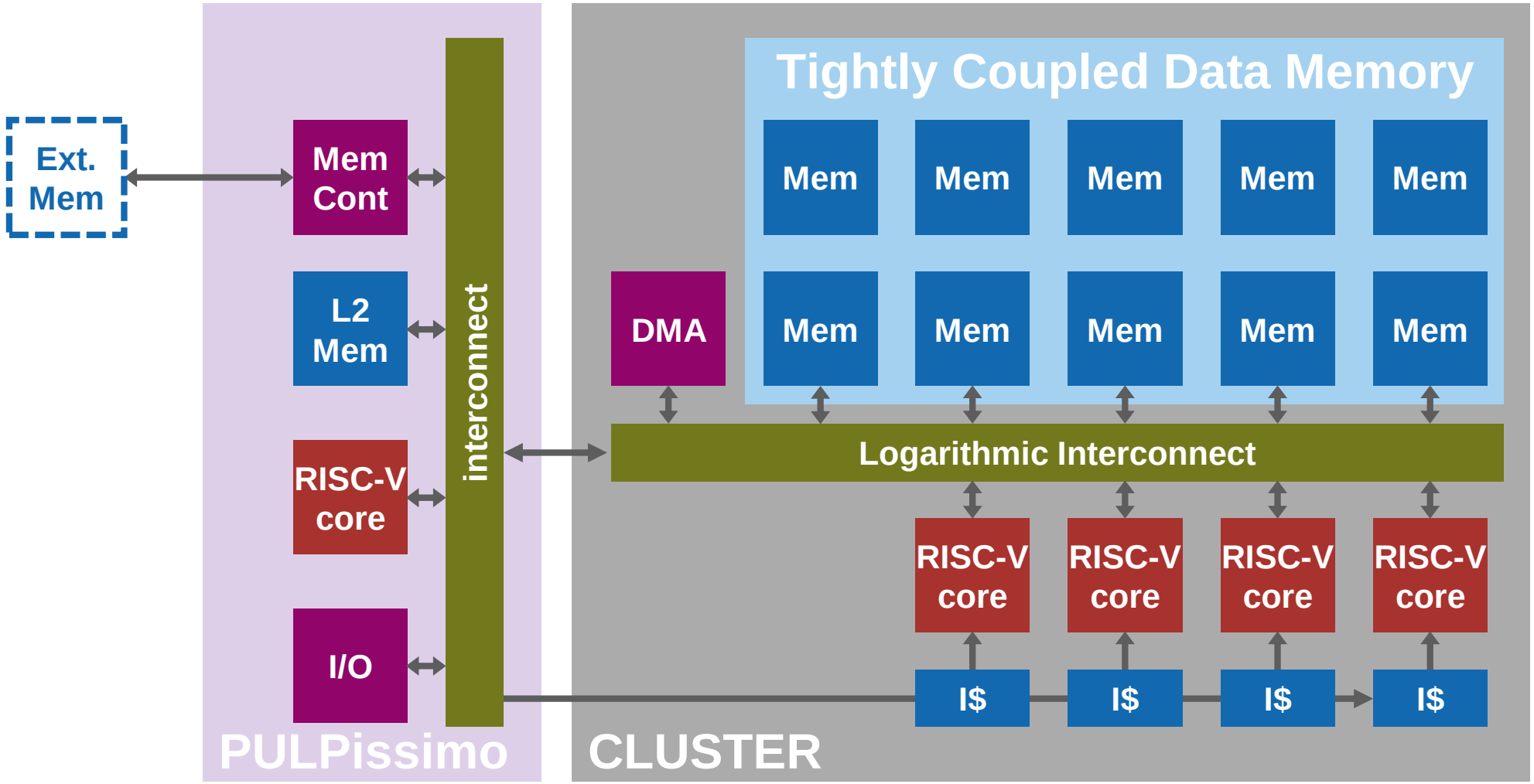


Results: RV32IMCXpulp vs RV32IMC

- 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



An additional I/O controller is used for IO

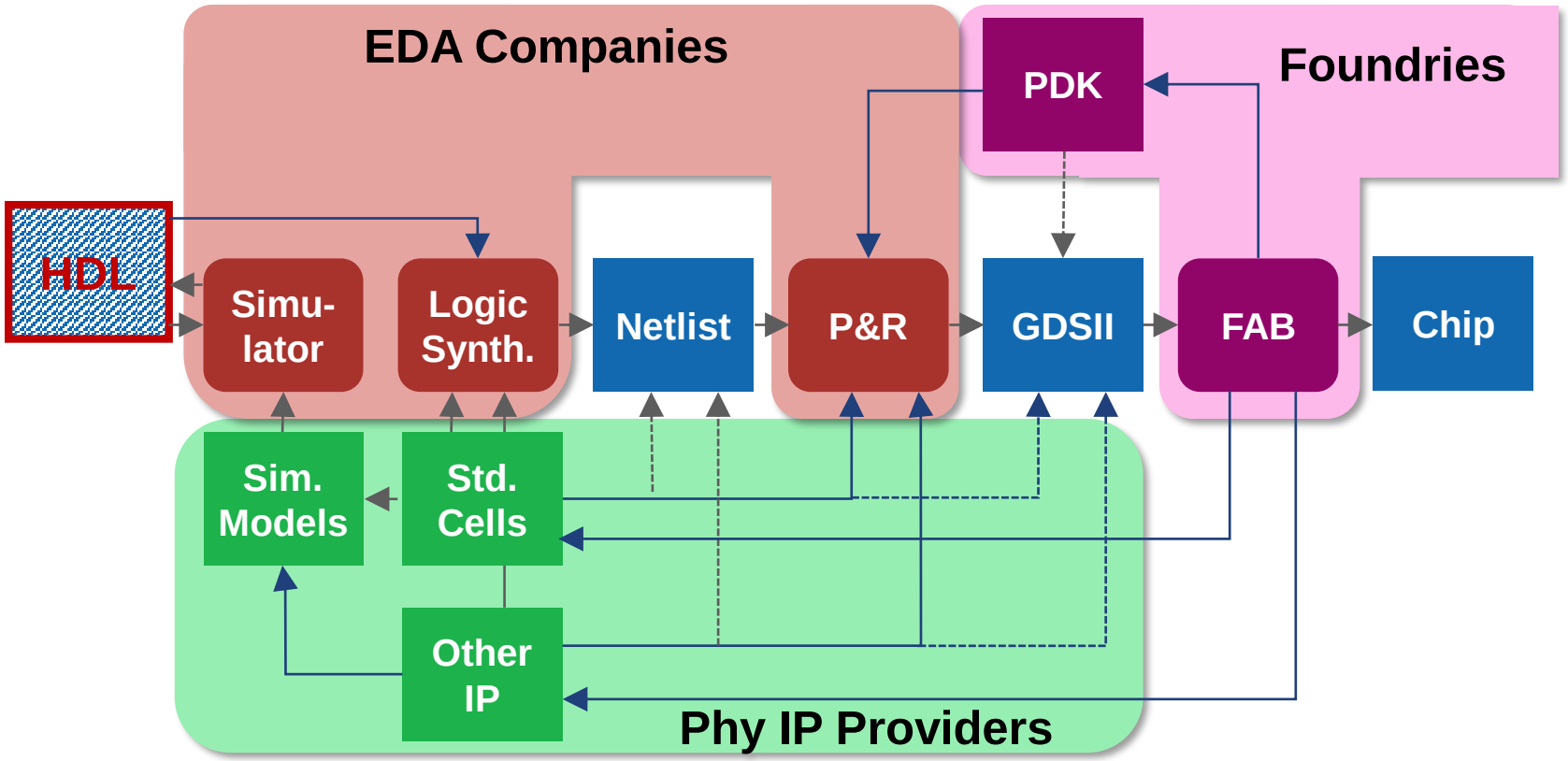


All this is Open Source HW – PULP open

Nice, but what exactly is "open" in Open Source HW?

Nice, but what exactly is “open” in Open Source HW?

- Only the first stage of the silicon production pipeline can be open HW
  → **RTL source code** (in an HDL such as SystemVerilog)
- Later stages contain closed IP of various actors + tool licensing issues



Permissive, Copyright (e.g APACHE) License is Key for industrial adoption

PULP includes Cores+Interco+IO → and Open Platform

RISC-V Cores

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

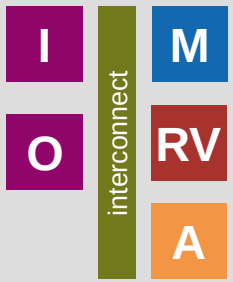
Peripherals

JTAG	SPI
UART	I2S
DMA	GPIO

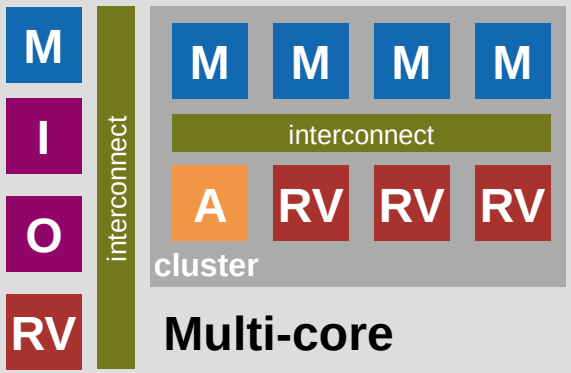
Interconnect

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

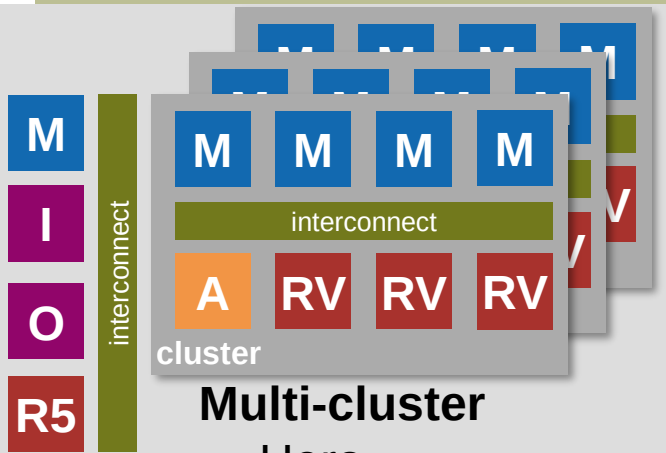
Platforms



Single Core
- PULPino
- PULPissimo



Multi-core
- Fulmine
- Mr. Wolf

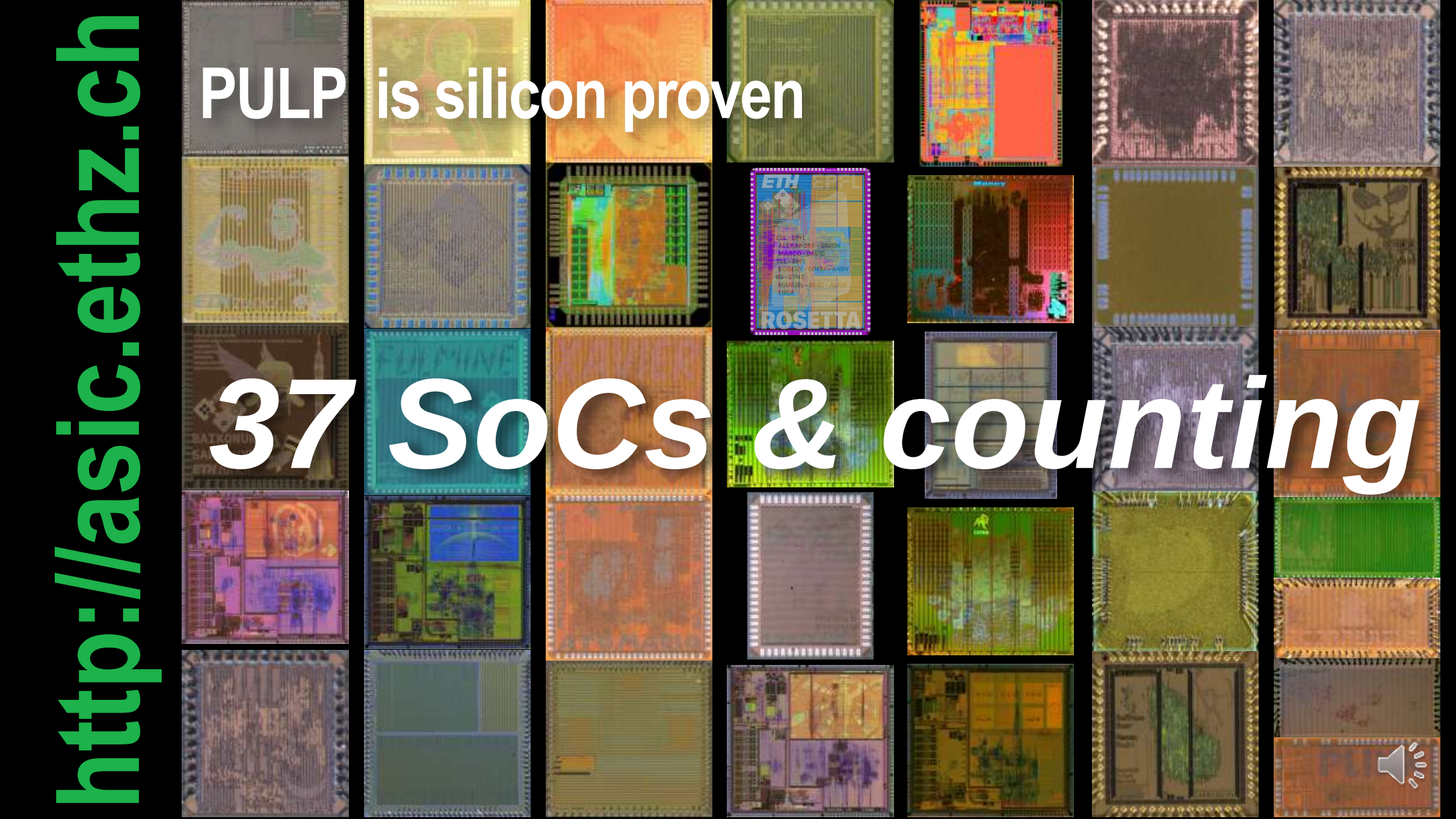


Multi-cluster
- Hero
- Open Piton

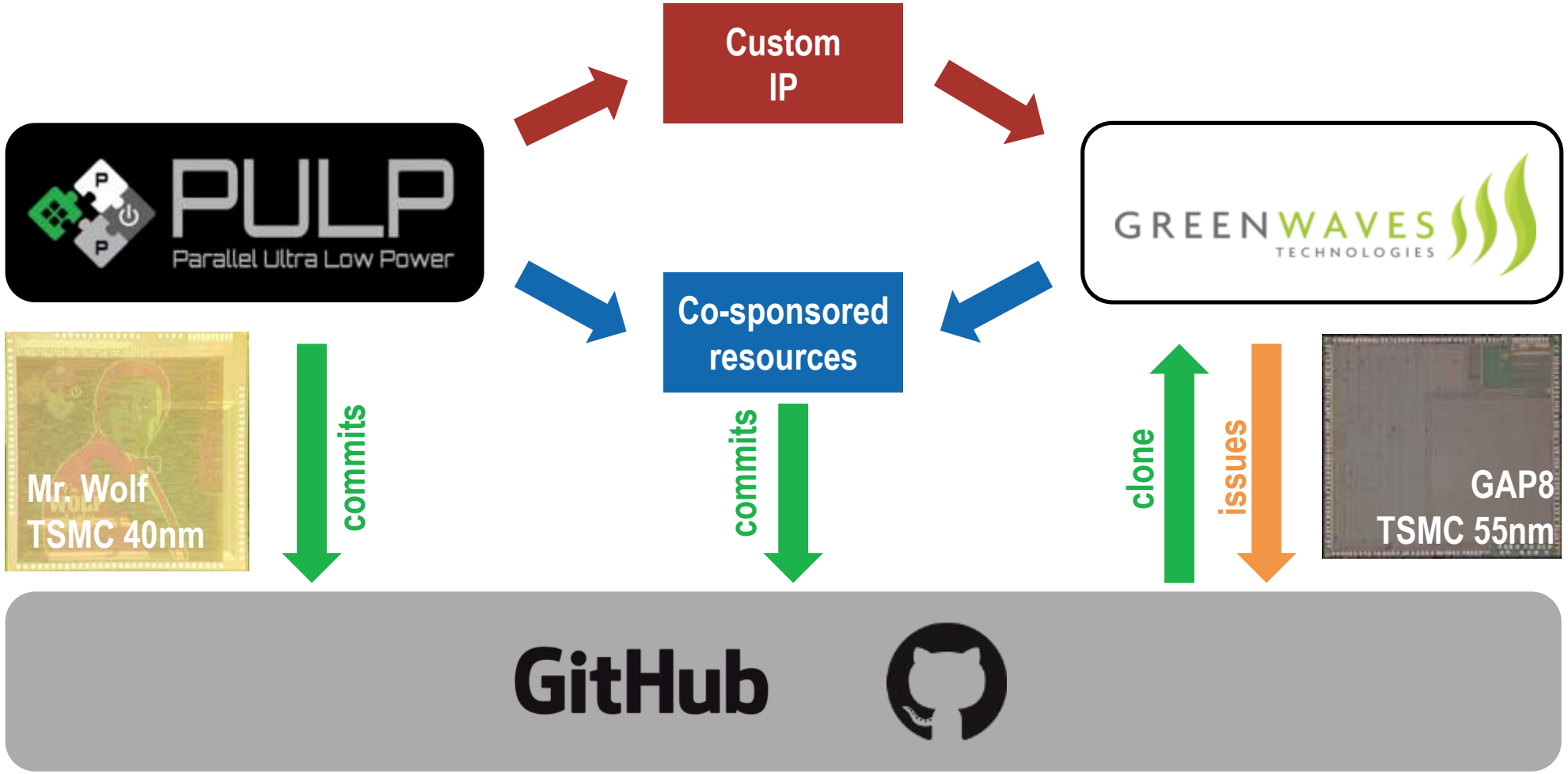
IOT → **HPC**

Accelerators

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

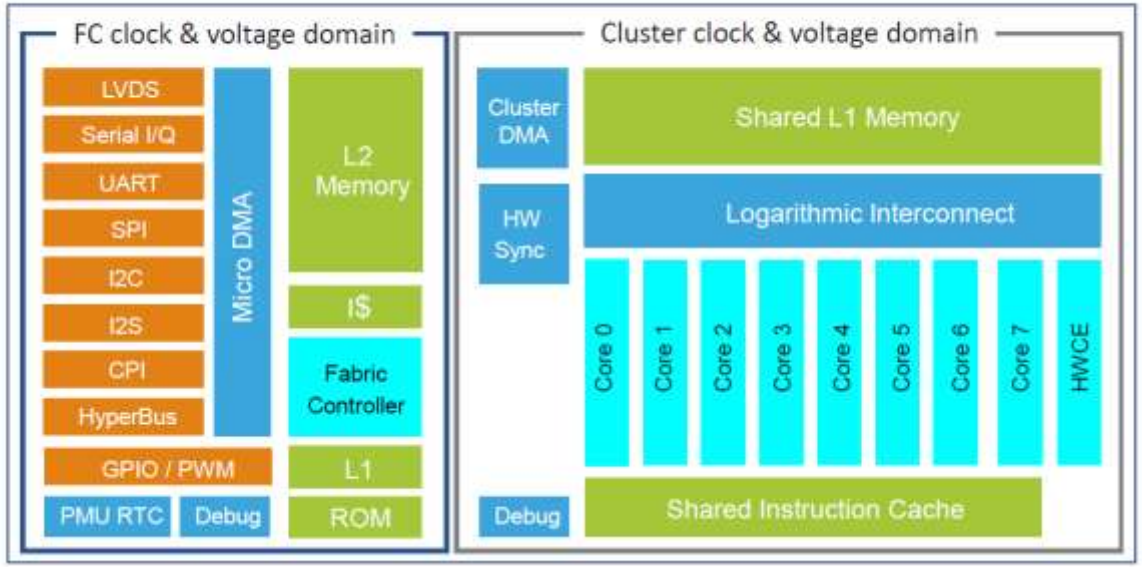

http://asic.ethz.ch
PULP is silicon proven
ROSETTA
37 SoCs & counting

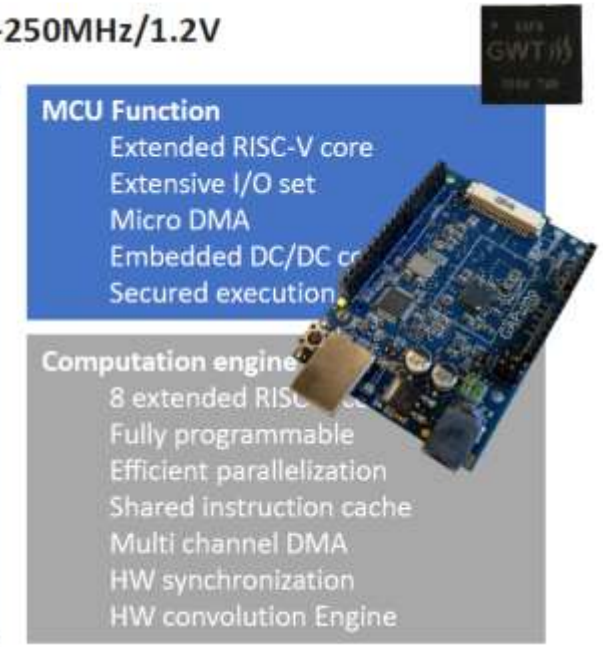
Open source collaboration with industry



Successful product development: GWT's GAP8

Two independent clock and voltage domains, from 0-133MHz/1V up to 0-250MHz/1.2V





What	Freq MHz	Exec Time ms	Cycles	Power mW
40nm Dual Issue MCU	216	99.1	21 400 000	60
GAP8 @1.0V	15.4	99.1	1 500 000	3.7
GAP8 @1.2V	17.5	8.7	1 500 000	70
GAP8 @1.0V w HWCE	4.7	99.1	460 000	0.8

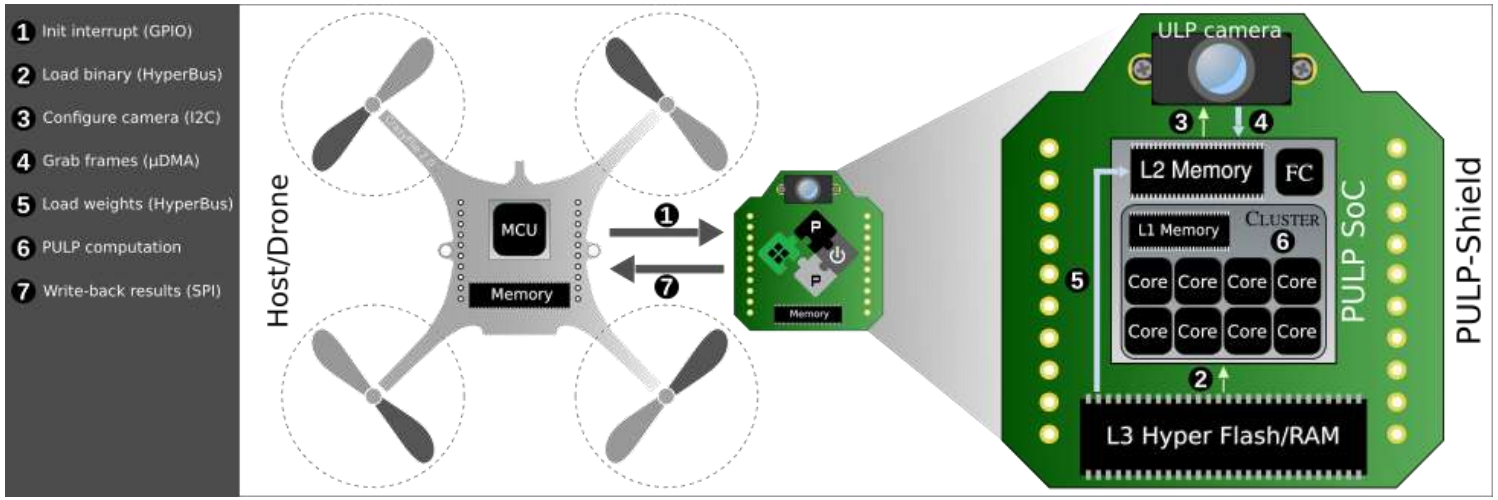
11 X

16 X

17
5

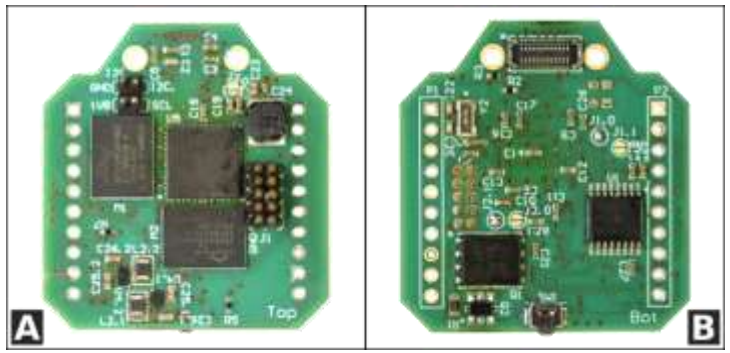
GREENWAVES TECHNOLOGIES

New Application Frontiers: DroNET on NanoDrone



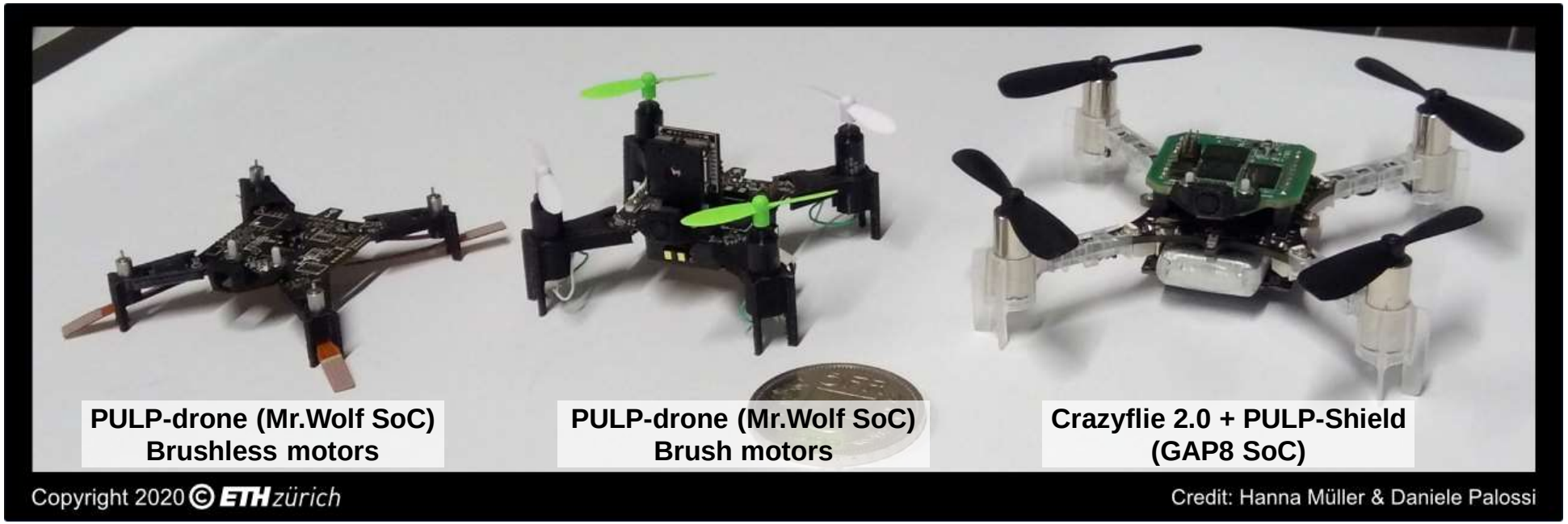
Pluggable PCB: PULP-Shield

- ~5g, 30×28mm
- GAP8 SoC
- 8 MB HDRAM
- 16 MB HFlash
- QVGA ULP HiMax camera
- Crazyflie 2.0 nano-drone (27g)



Only onboard computation for autonomous flight + obstacle avoidance
no human operator, no ad-hoc external signals, and no remote base-station!

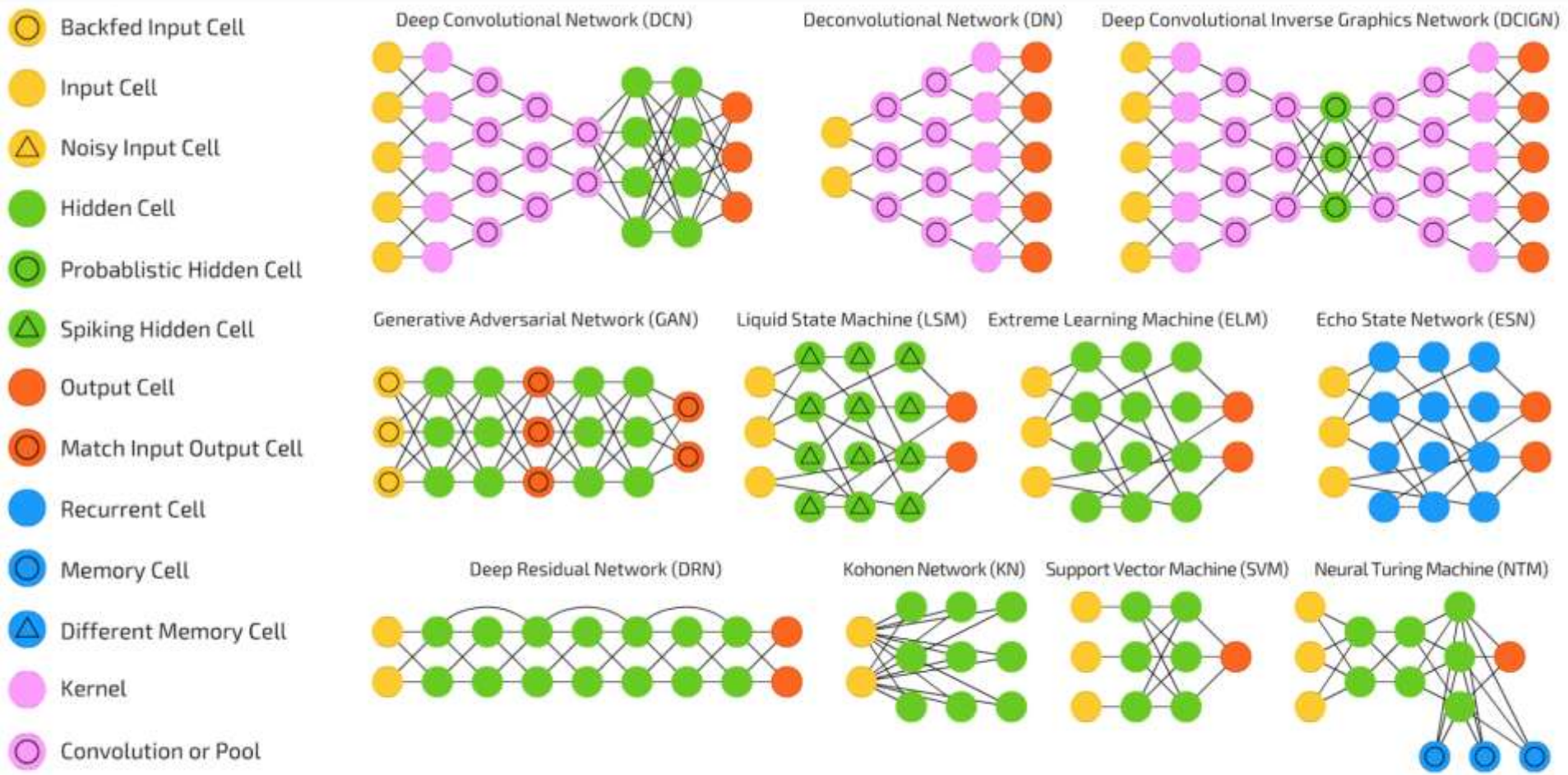
Outlook: PULP-based nano UAV family



- FrontNET [1]: hovering in front of a freely moving user
- End-to-end, RESnet-like CNN
- Runs with DroneNET
- AI+control: 5-10% of tot. Power
- Collaboration with IDSIA (USI-SUPSI) Lugano, Switzerland

[1] Mantegazza, Dario, et al. "Vision-based Control of a Quadrotor in User Proximity: Mediated vs End-to-End Learning Approaches." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

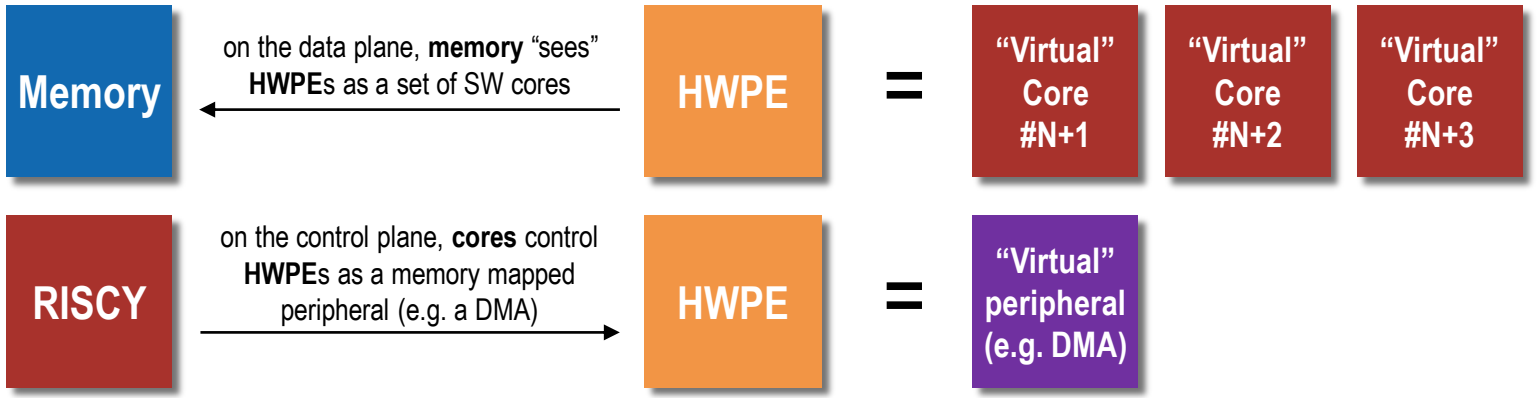
What's next? Sub-pJ/OP Accelerators, but flexibility needed!

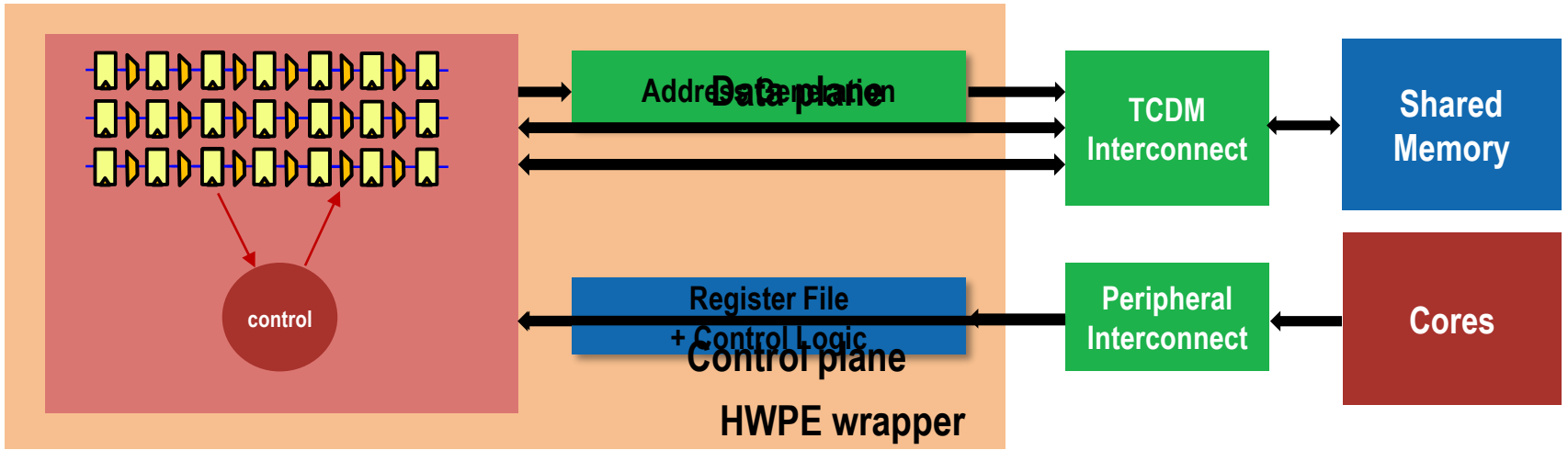


asimovinstitute.org/neural-network-zoo

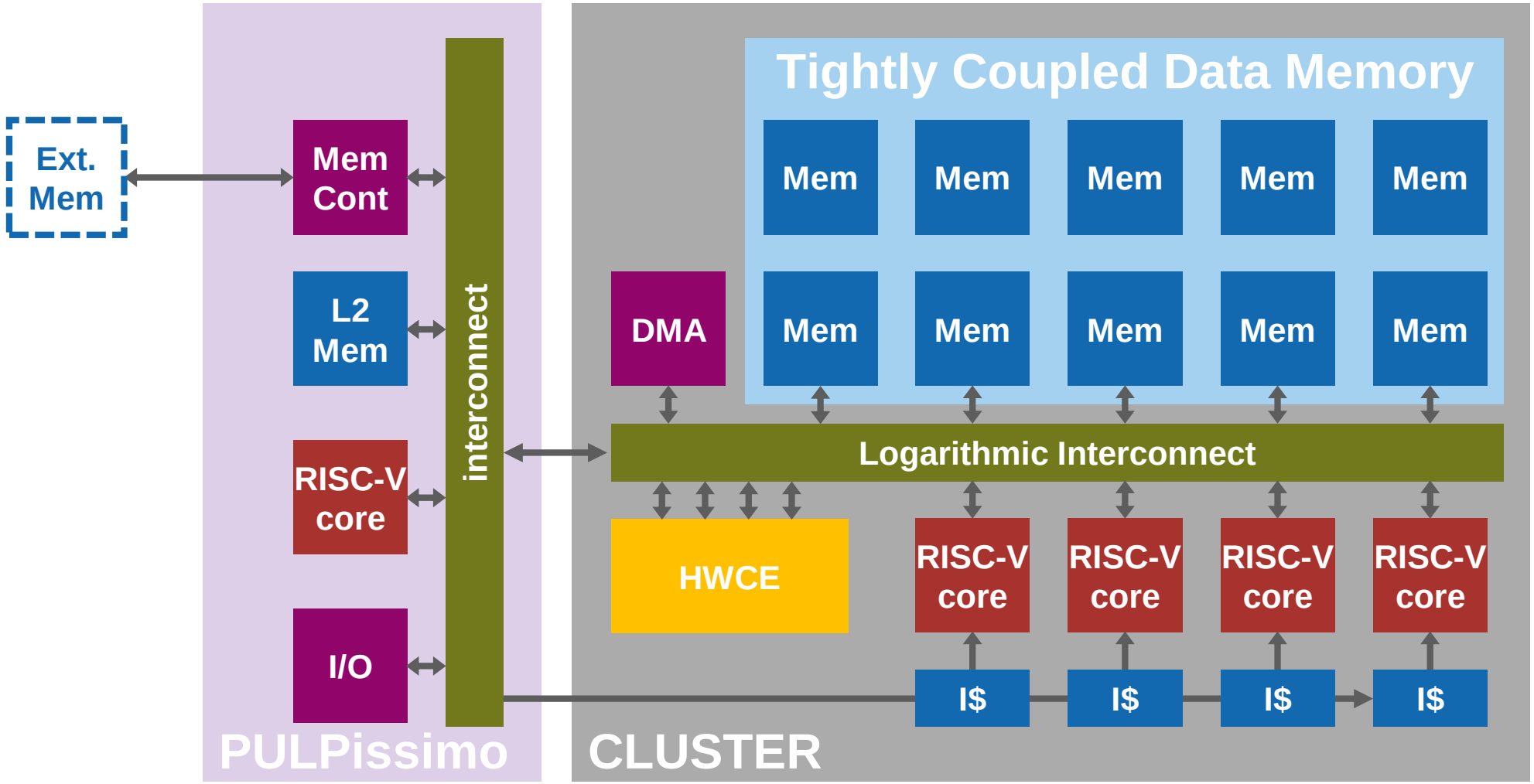
+ FFT, PCA, Mat-inv,…

Hardware Processing Engines (HWPEs)





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



Acceleration with flexibility: zero-copy HW-SW cooperation

More Efficiency (2): Extreme Quantization

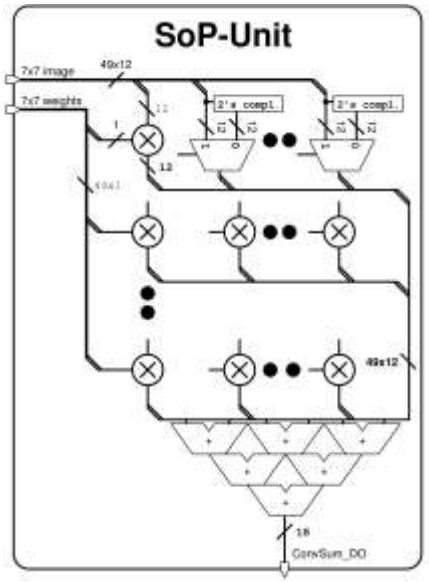
Low(er) precision: 8→4→2

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%

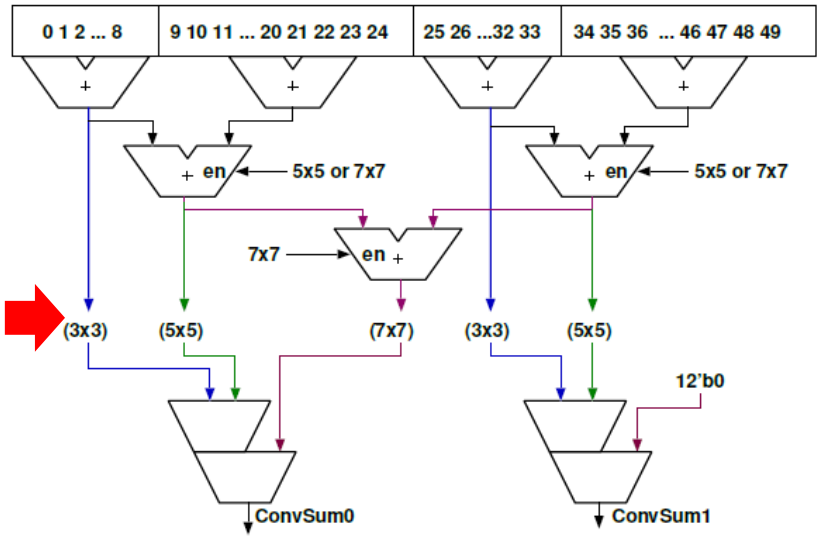
SOA INQ retraining

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

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

XNOR

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

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

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

Thresholding

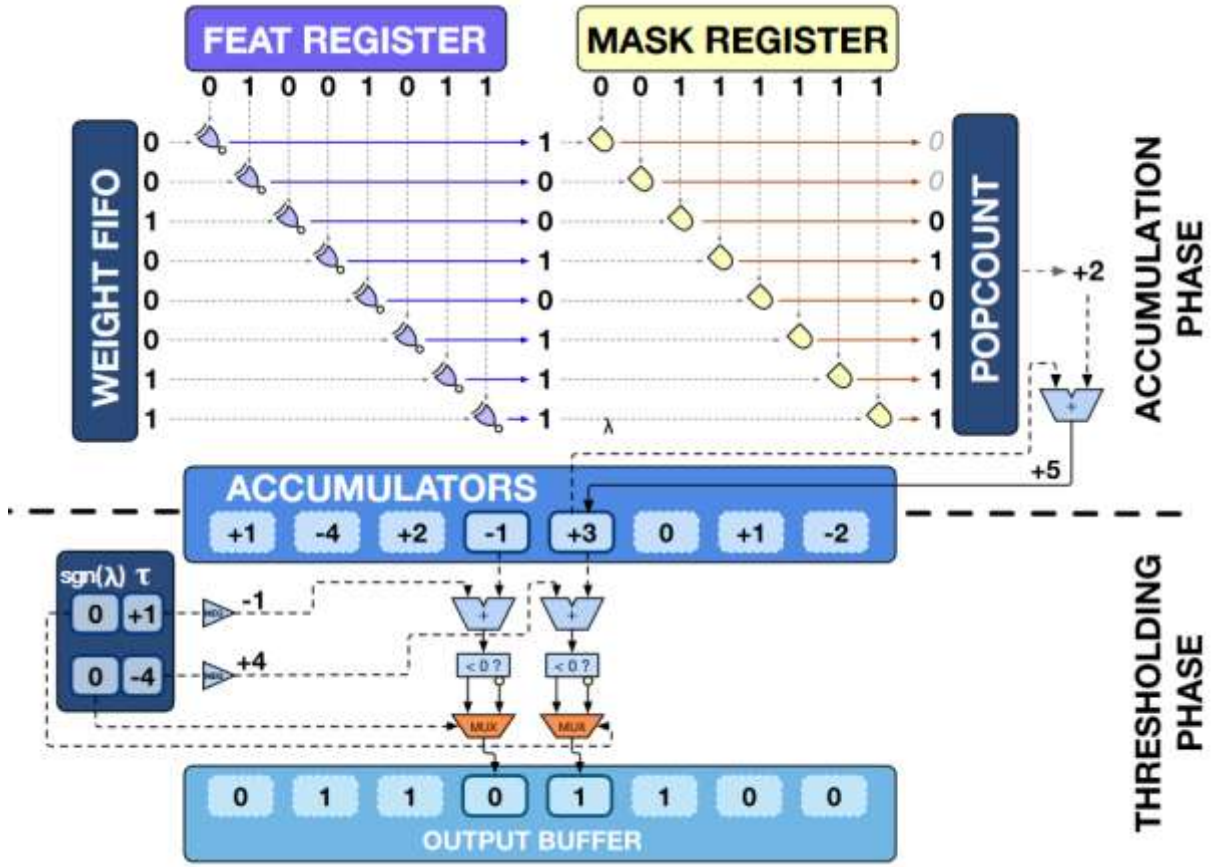
Multi-bit accumulation

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



BINCONV: Binary dot-product and thresholding logic array

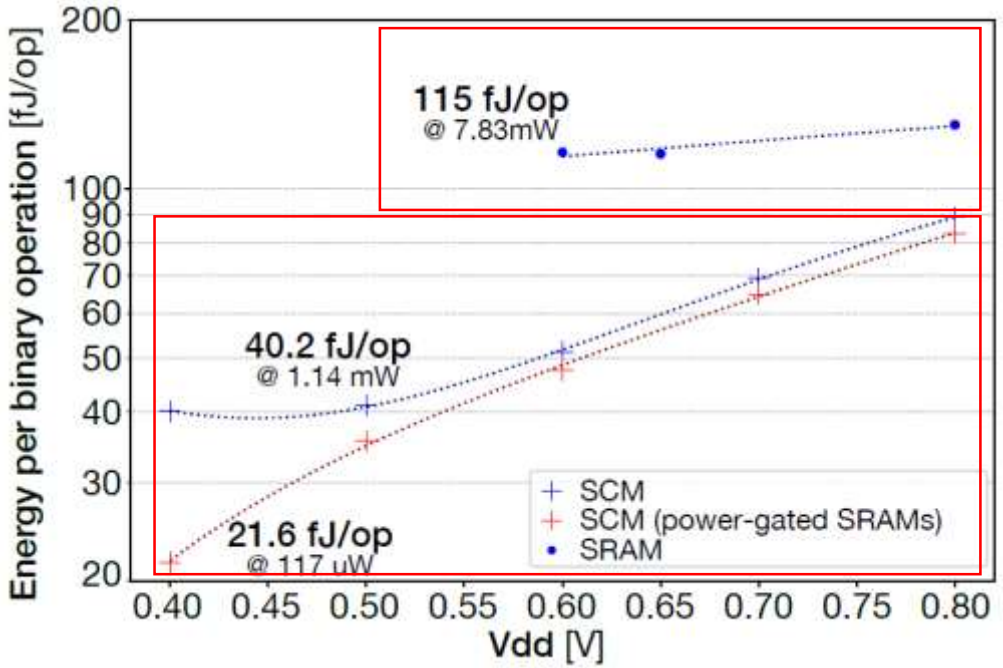
XNE Energy Efficiency

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**

100+ TOPS/W (10fJ/OP)
Is reachable!



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

Flexibility needed: Binary-Based Quantization (BBQ)

QNN layer :

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

INT32 accumulator

Q-bit output fmaps

M-bit weights

N-bit input fmaps

Many $\boldsymbol{M} \times \boldsymbol{N}$ bits products...

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

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

power-of-2 scaling factors

Q-bit output fmaps

1-bit weights

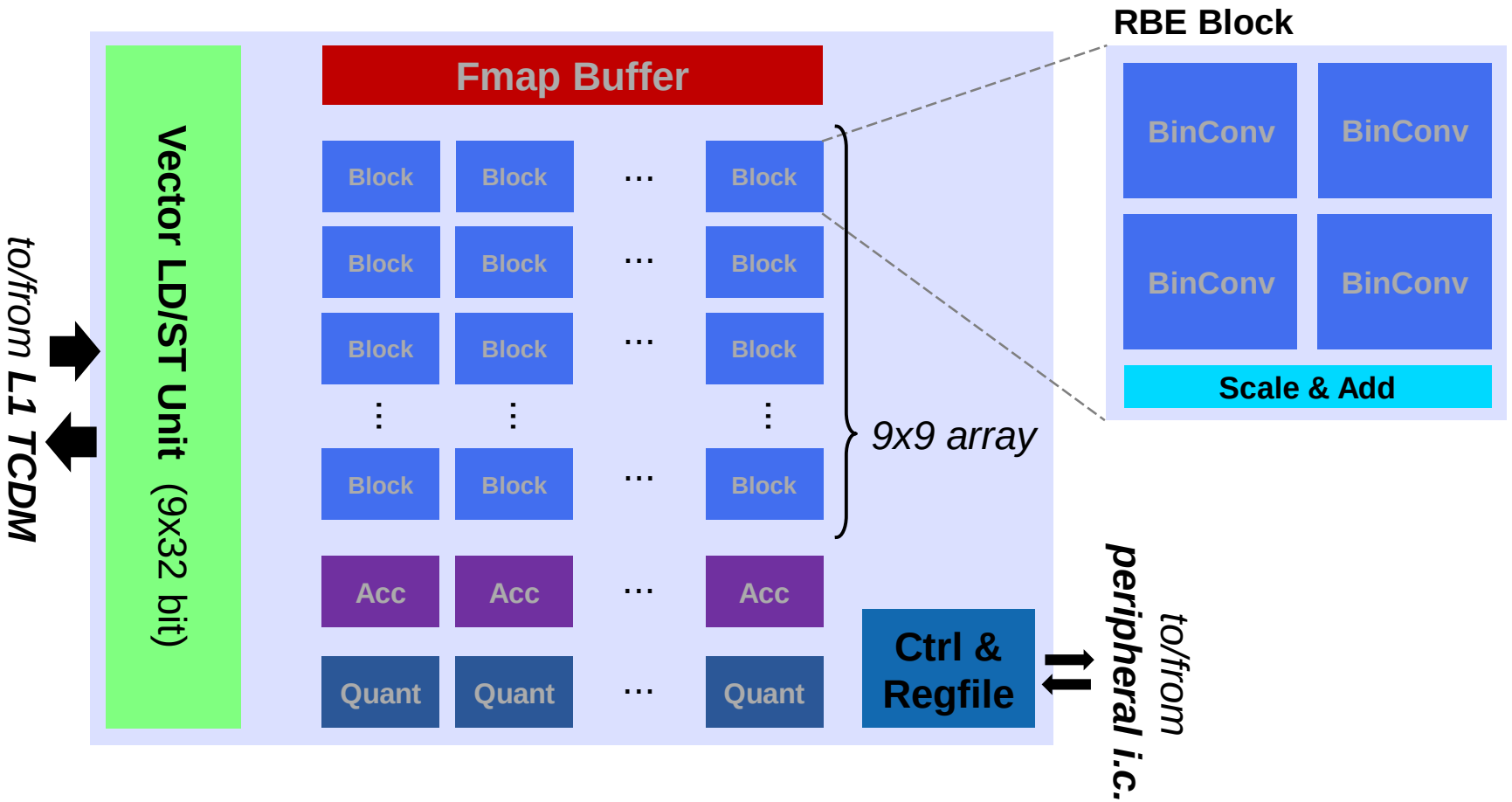
1-bit input fmaps

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

Reconfigurable Binary Engine

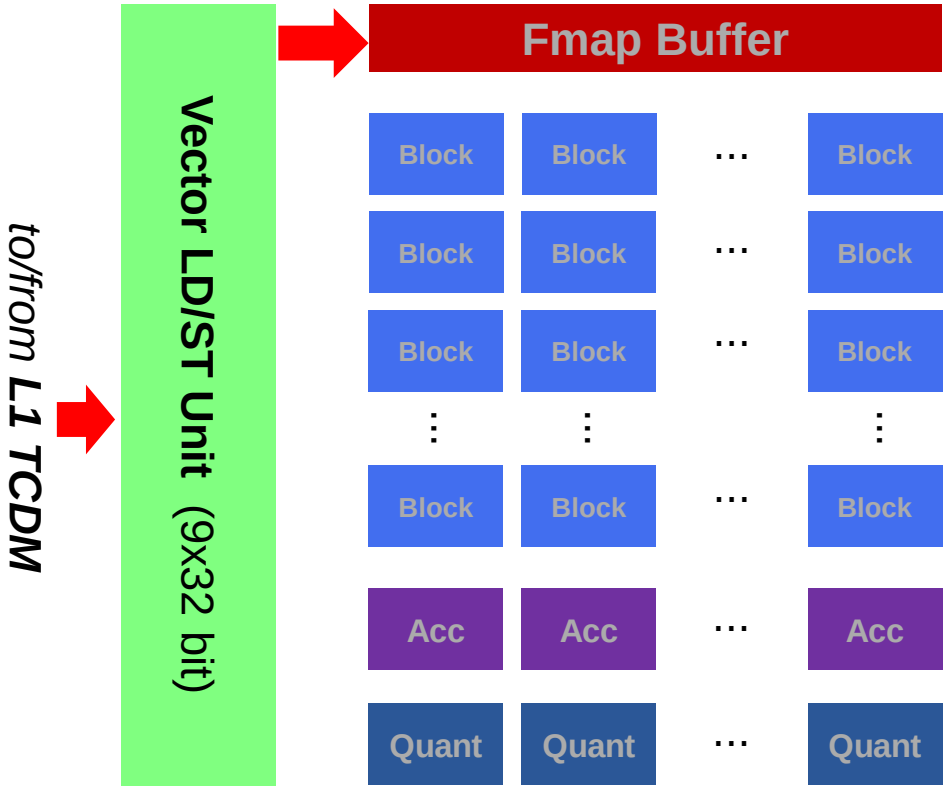
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$

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



Reconfigurable Binary Engine

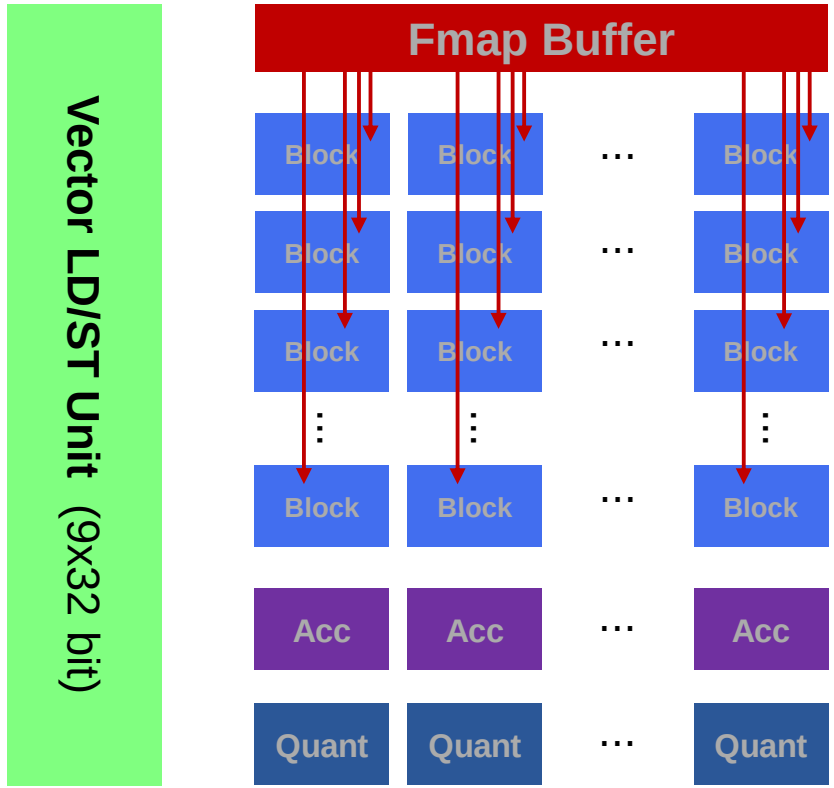
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$



1. *preload **fmap buffer** with input activations (up to 5x5 pixels, 32 channels, **N**=4 bits/pixel)*

Reconfigurable Binary Engine

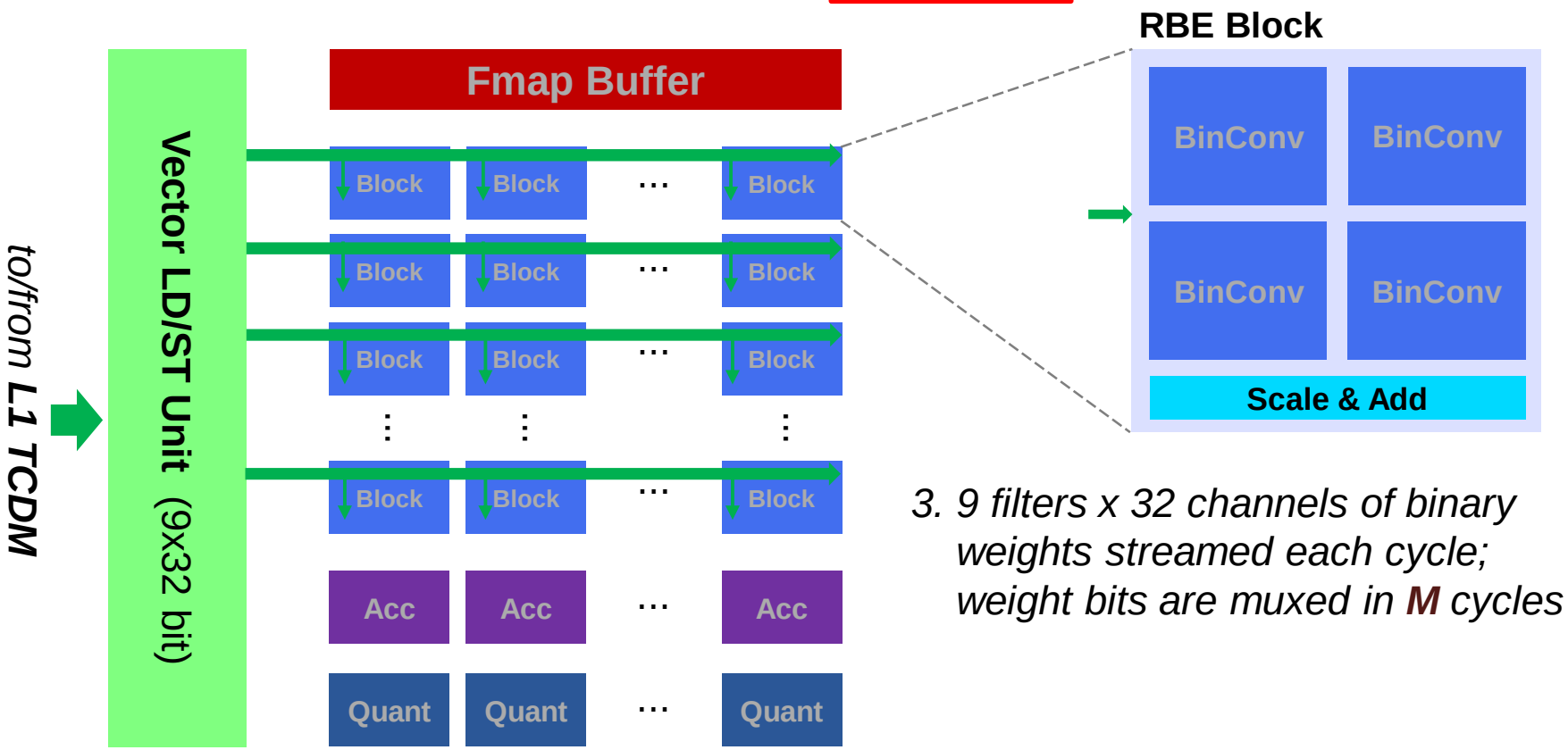
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$



1. preload **fmap buffer** with input activations (up to 5x5 pixels, 32 channels, 4 bits/pixel)
2. *each of the 81 blocks has as input one of the 25 px (32x4 bits)*

Reconfigurable Binary Engine

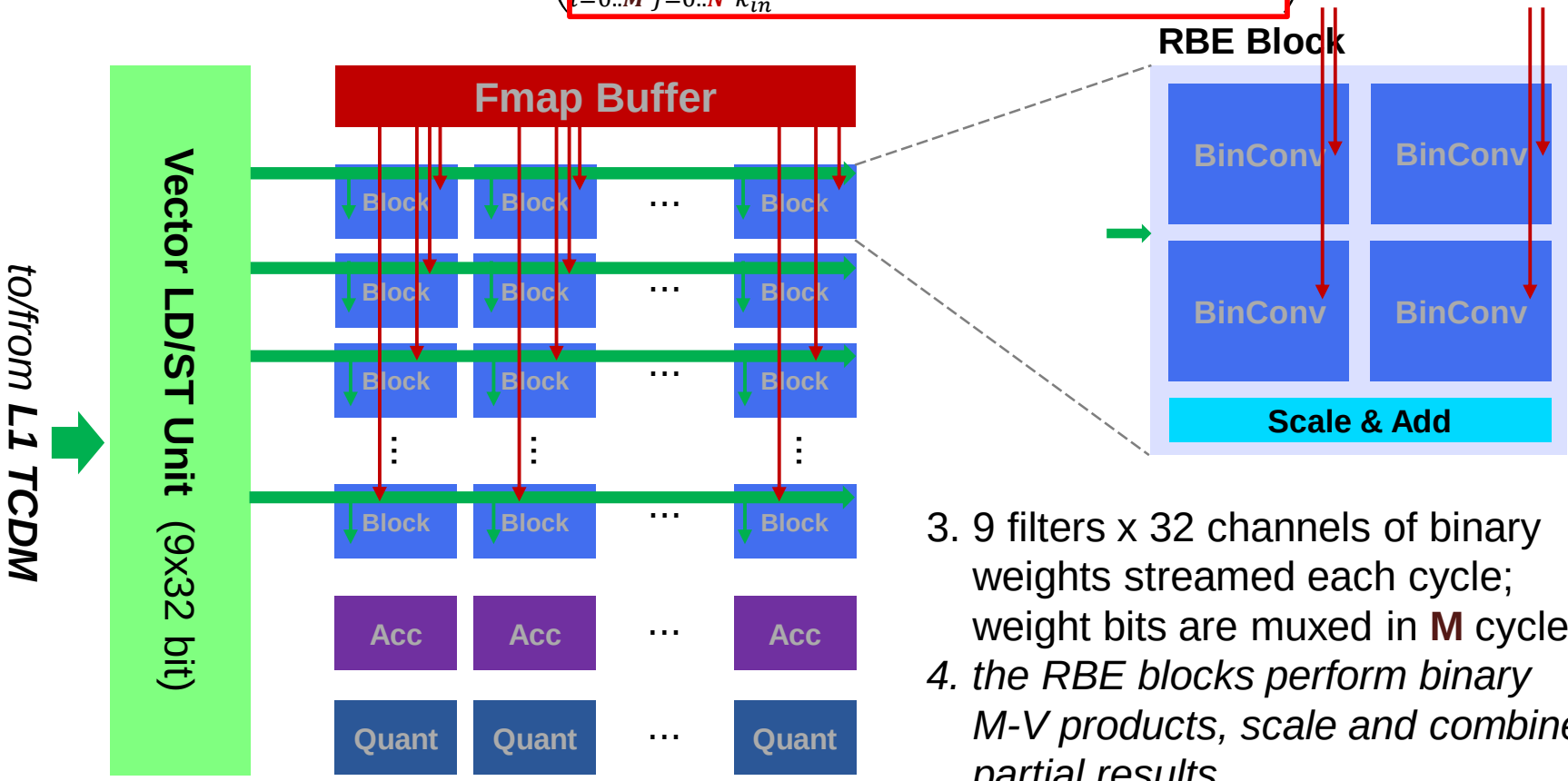
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$



3. *9 filters x 32 channels of binary weights streamed each cycle; weight bits are muxed in **M** cycles*

Reconfigurable Binary Engine

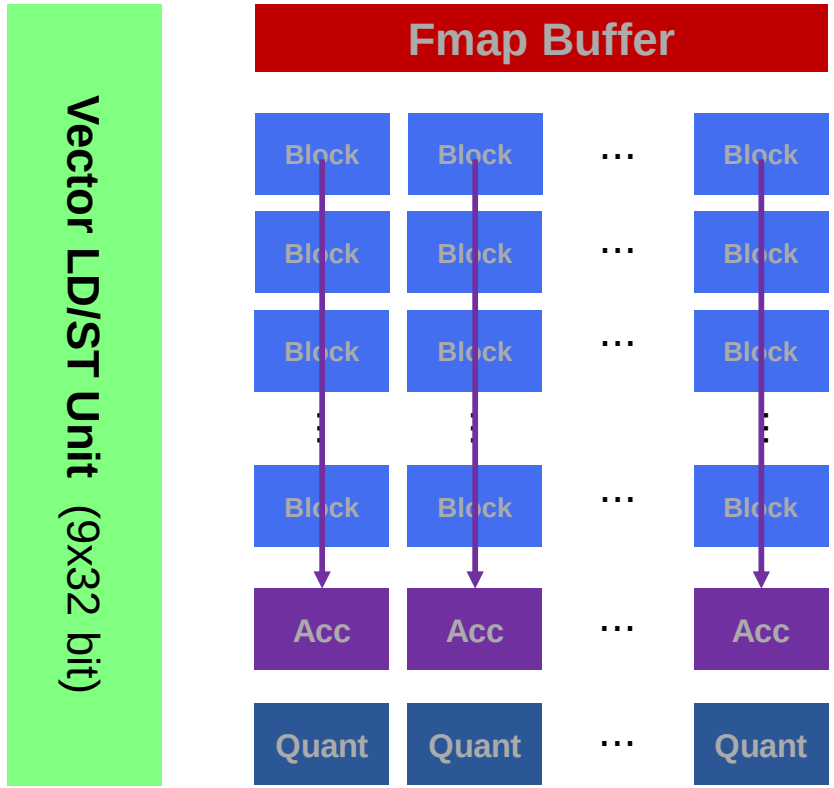
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$



3. 9 filters x 32 channels of binary weights streamed each cycle; weight bits are muxed in **M** cycles
4. *the RBE blocks perform binary M-V products, scale and combine partial results*

Reconfigurable Binary Engine

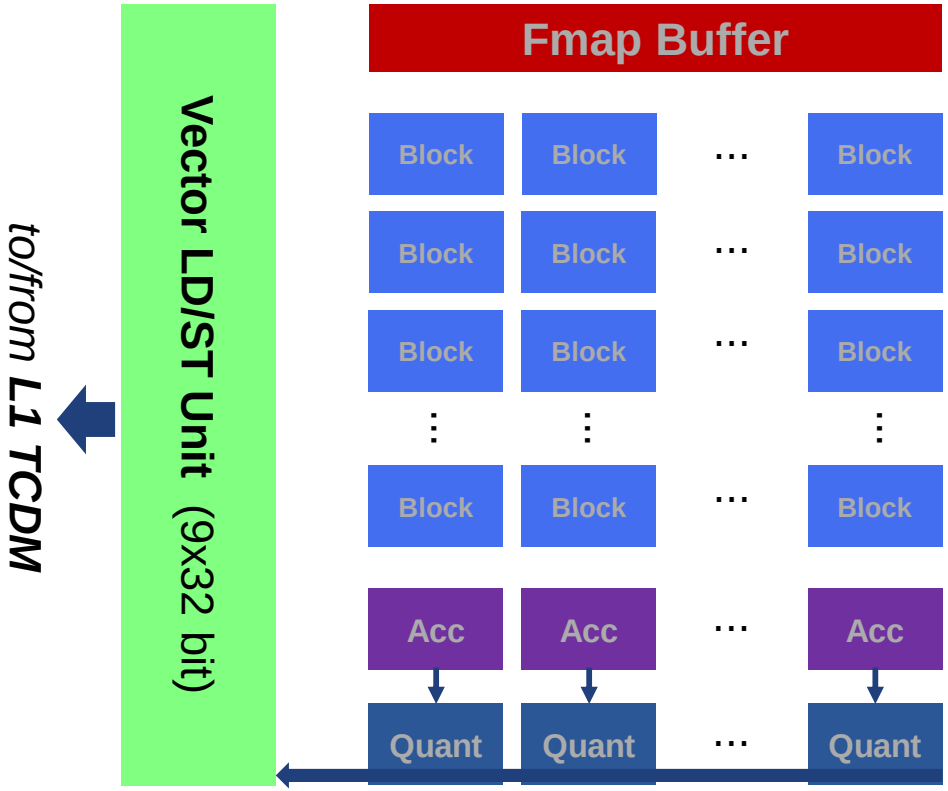
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$



5. *results are added column-wise and accumulated in 32-bit accumulator banks (9x32x32 bits) for many iterations until full accumulation*

Reconfigurable Binary Engine

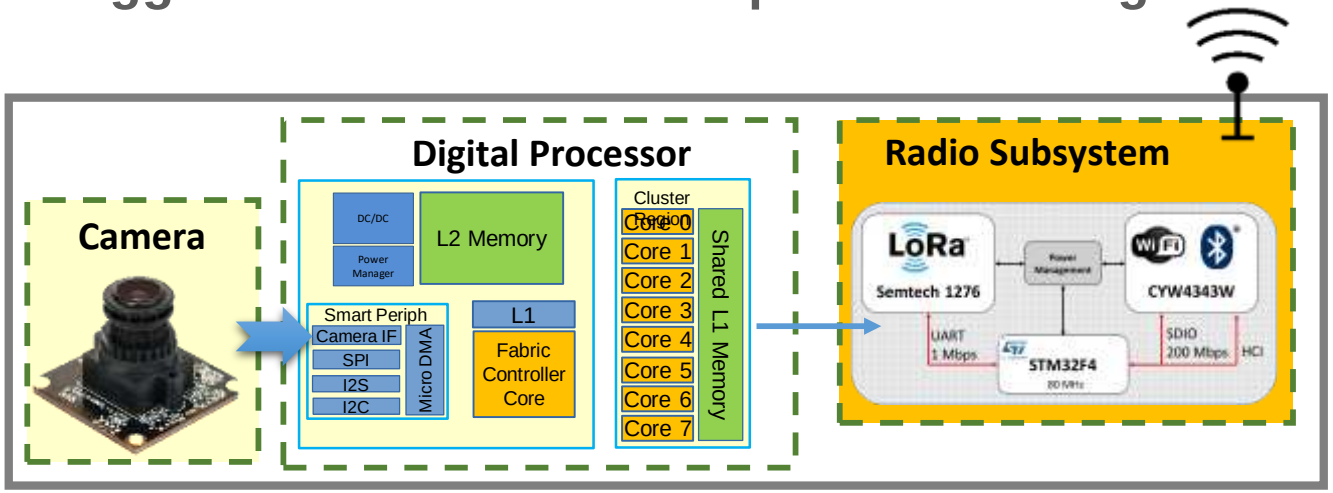
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$

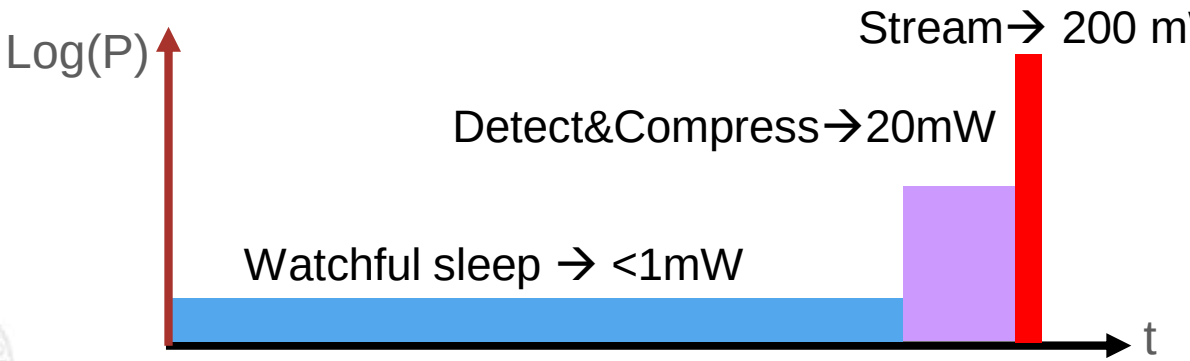


5. results are added column-wise and accumulated in 32-bit accumulator banks (9x32x32 bits) for many iterations until full accumulation
6. *after full accumulation, the accumulator values are **quantized** and **streamed out***

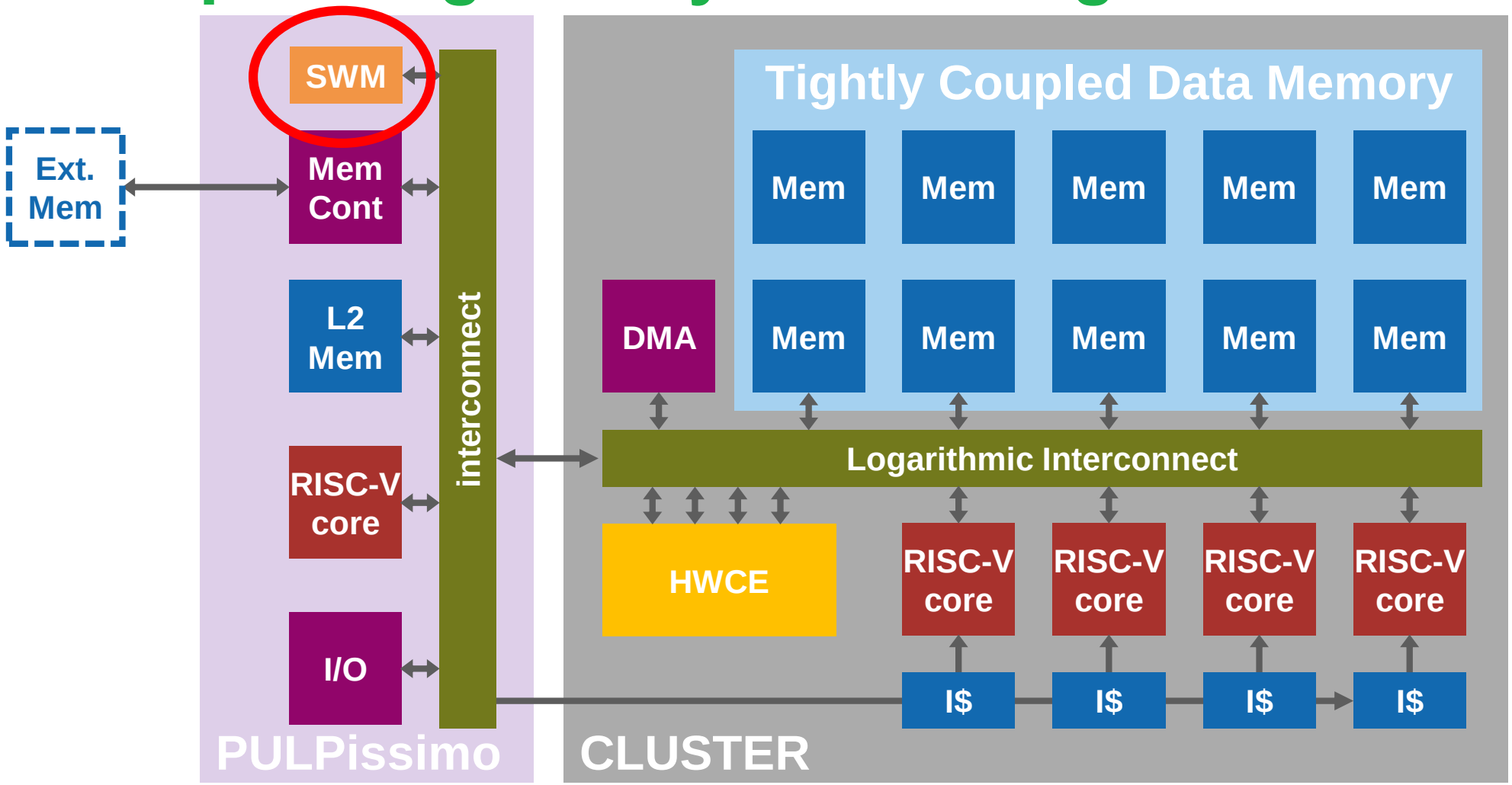
What about µW «sleep»?

Small always-on network →
Triggers alarm and video capture/streaming for cloud-based forensics



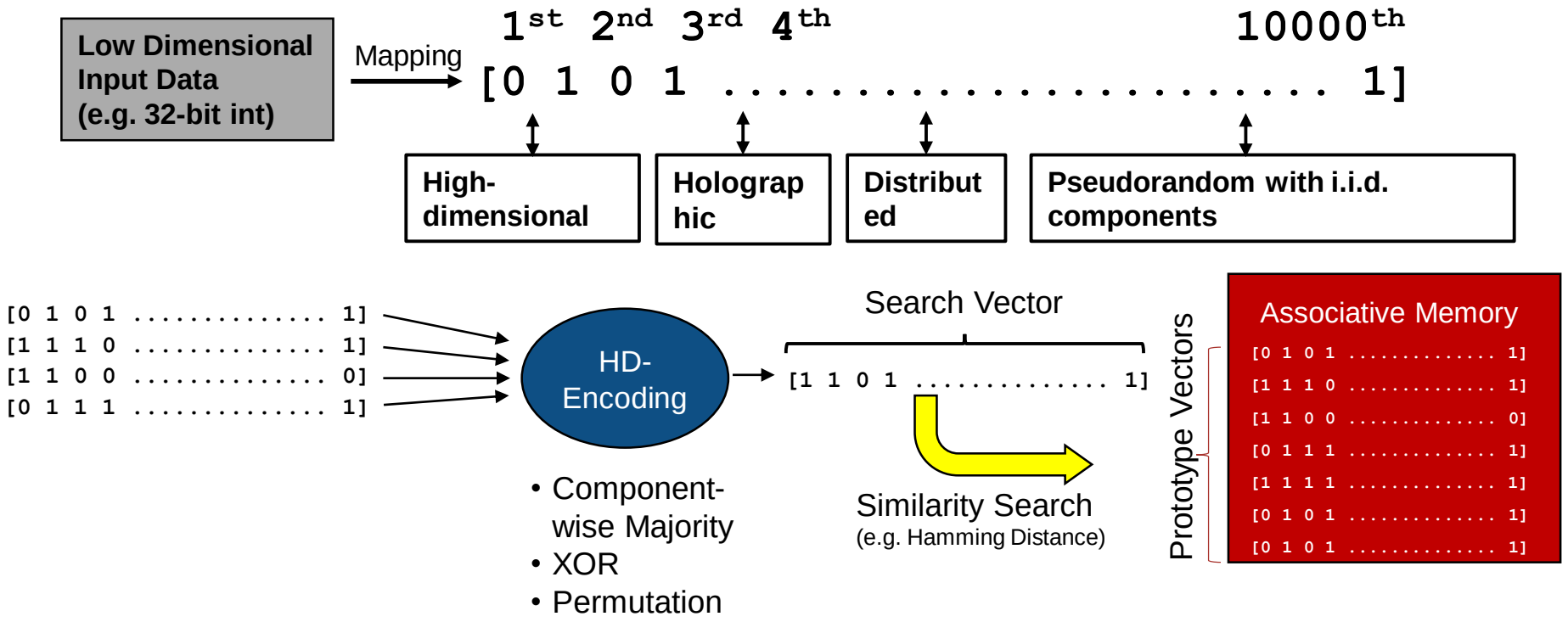


Need µW-range always-on Intelligence



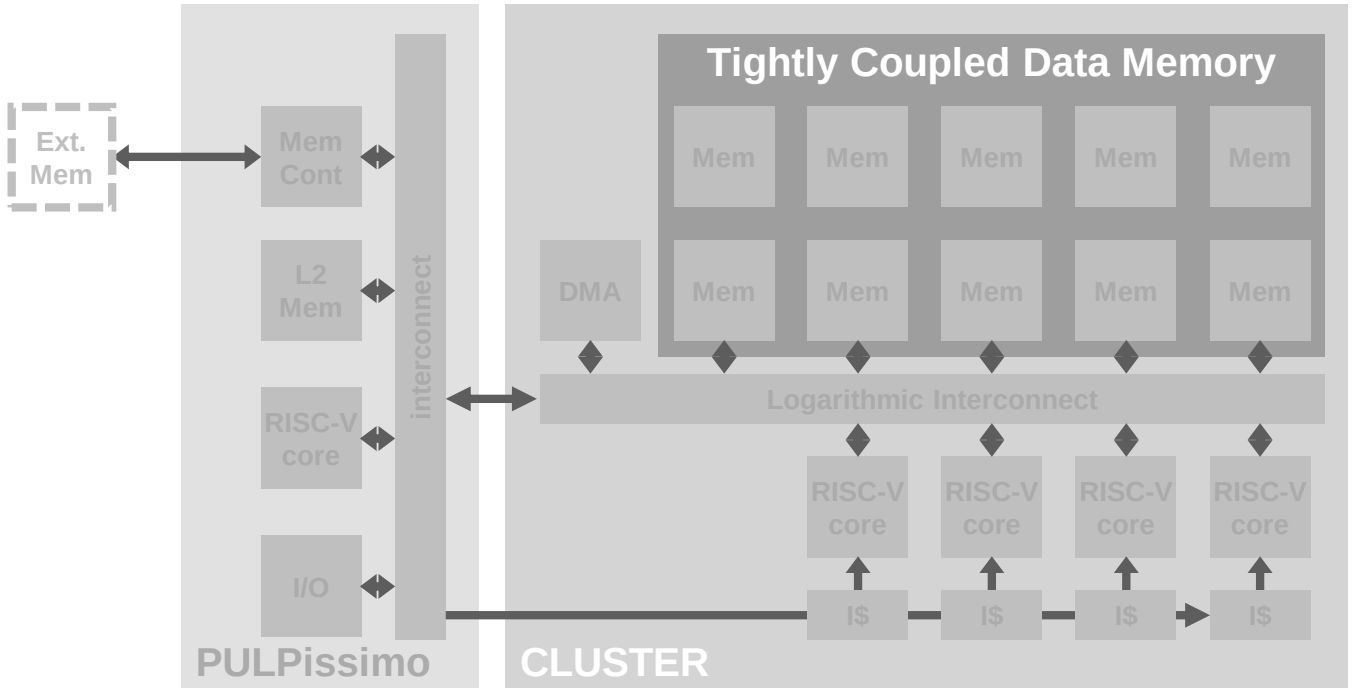
Smart Wakeup Module

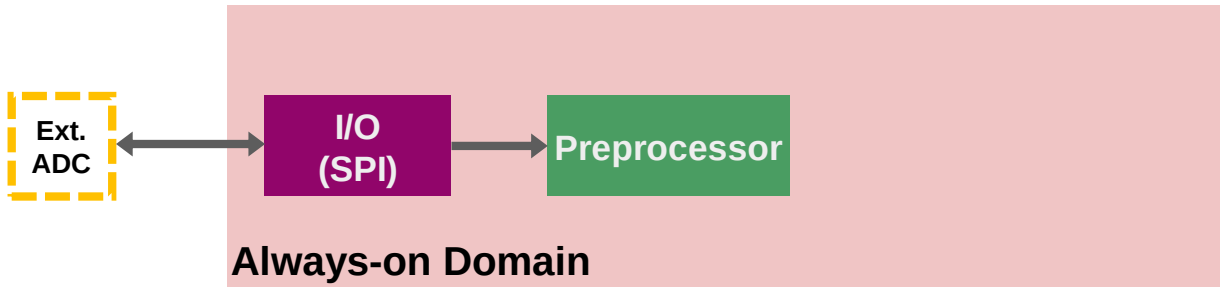
Not Only CNNs: Hyper-Dimensional Computing



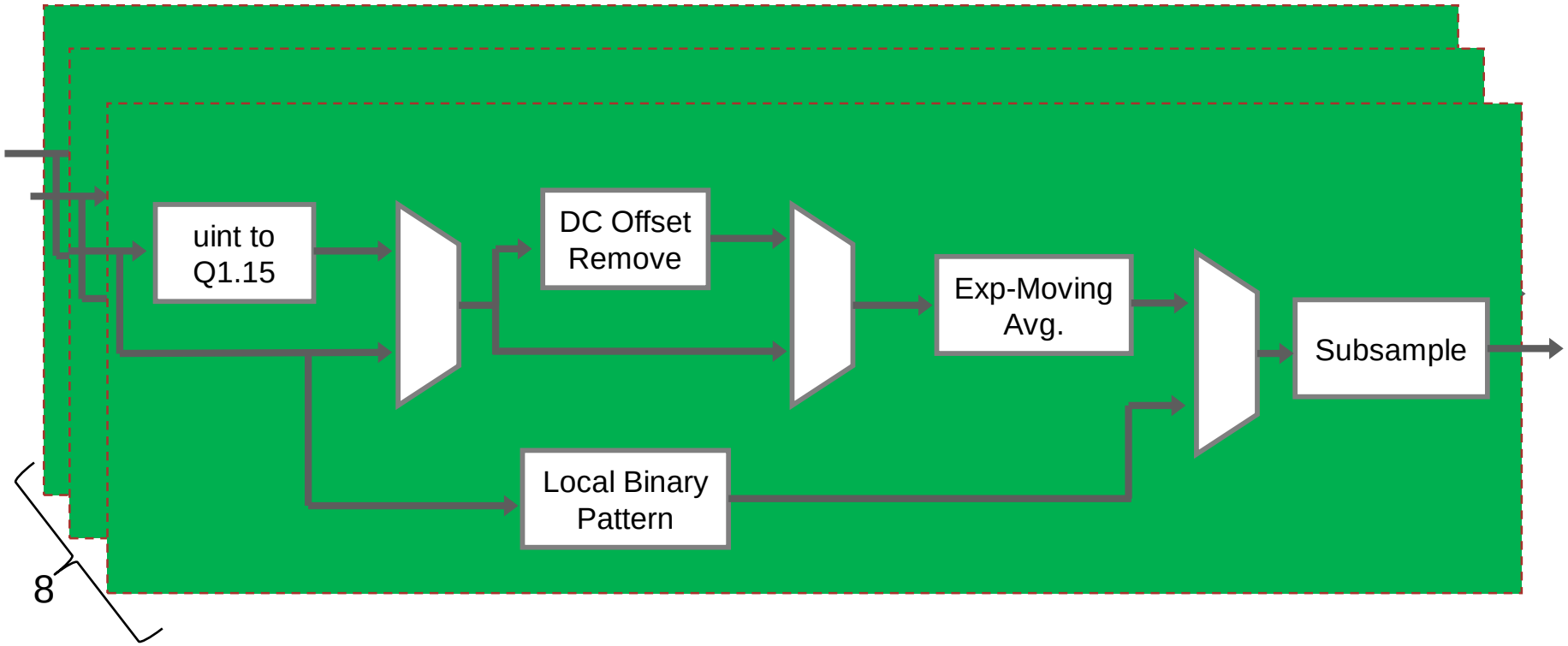
Highly parallel, fault-tolerand binary operators, assoc-min-distance search ➡ Merge storage & computation i.e. **In-memory computing**

HD-Based smart Wake-Up Module

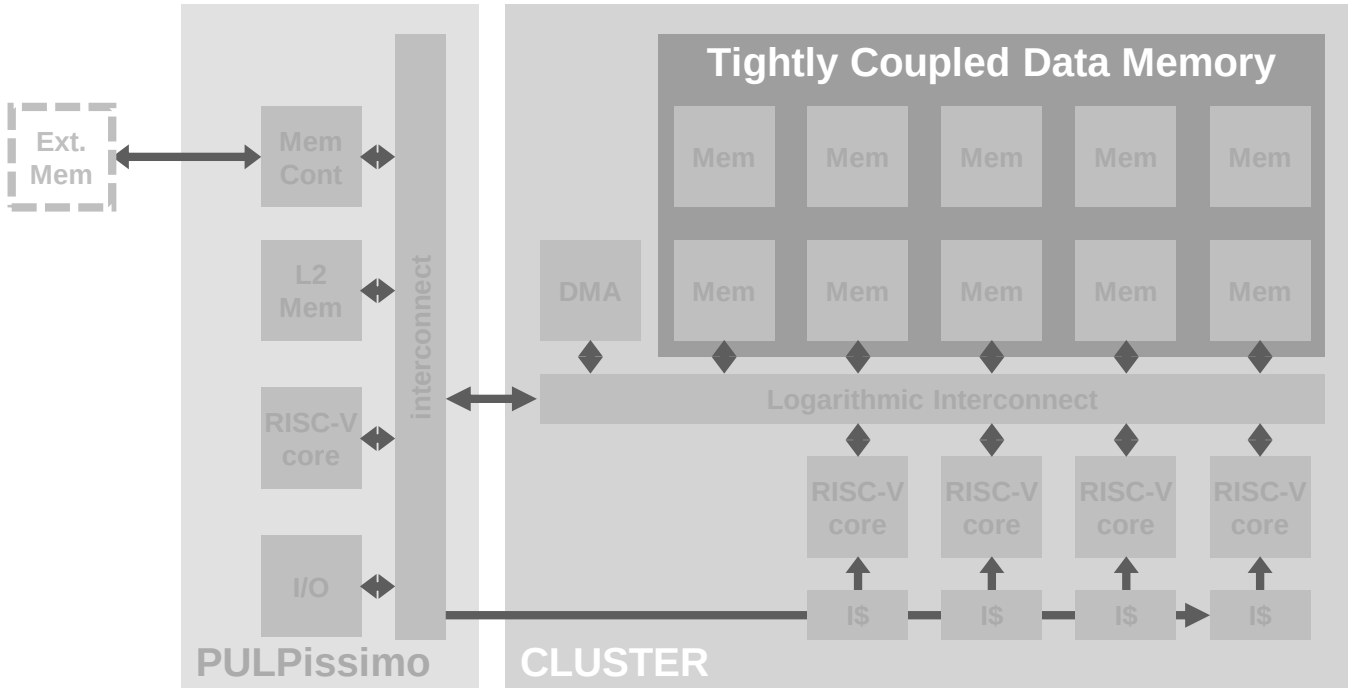


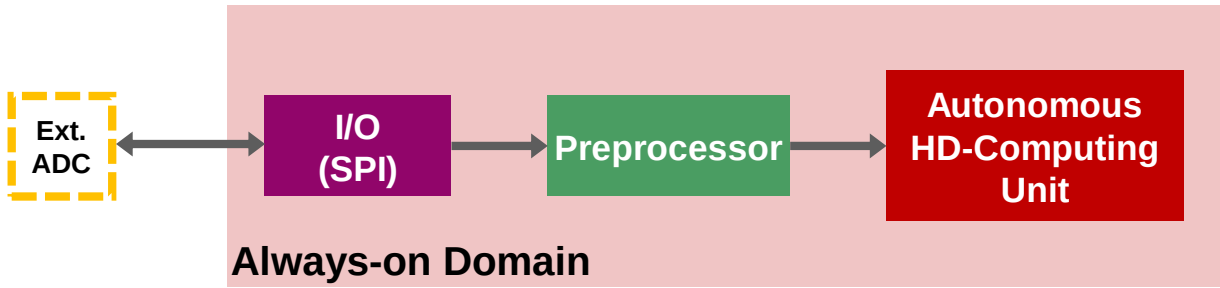


Preprocessor

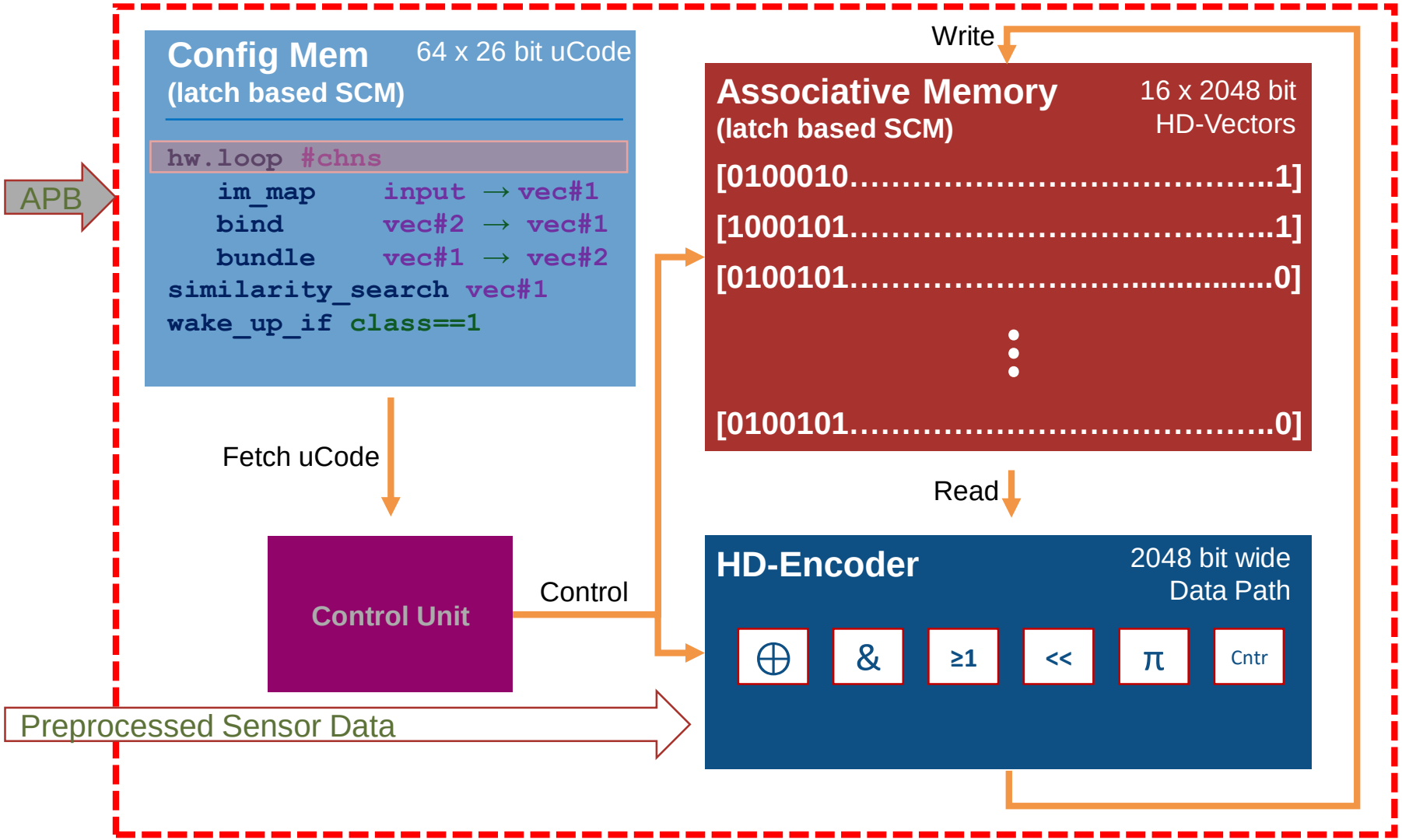


HD-Based smart Wake-Up Module

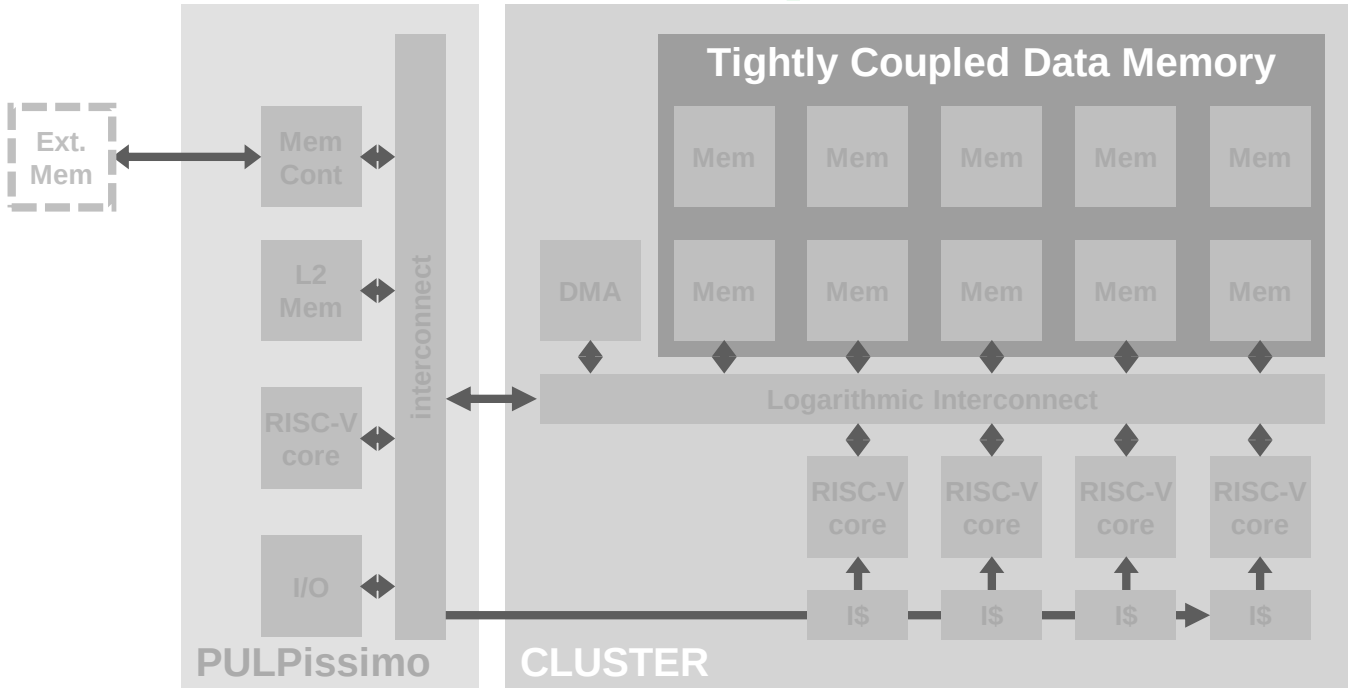


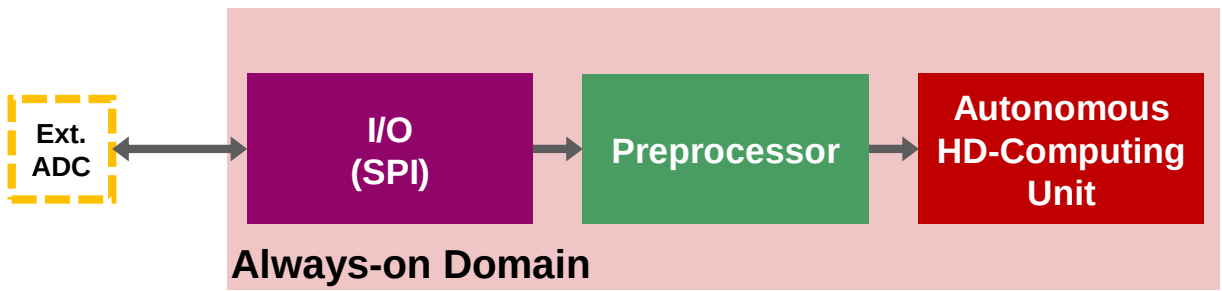


HD-Based smart Wake-Up Module

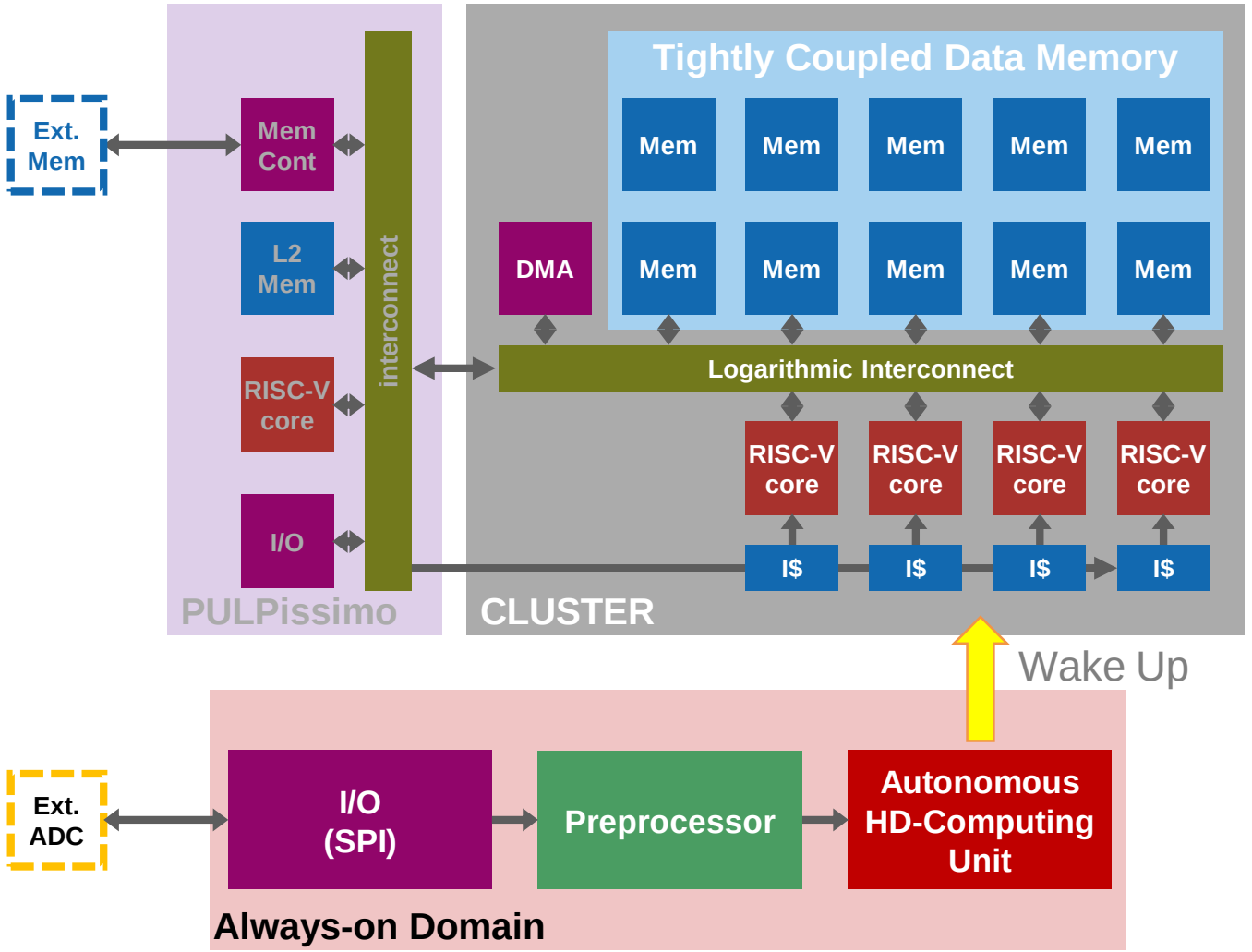


HD-Based smart Wake-Up Module

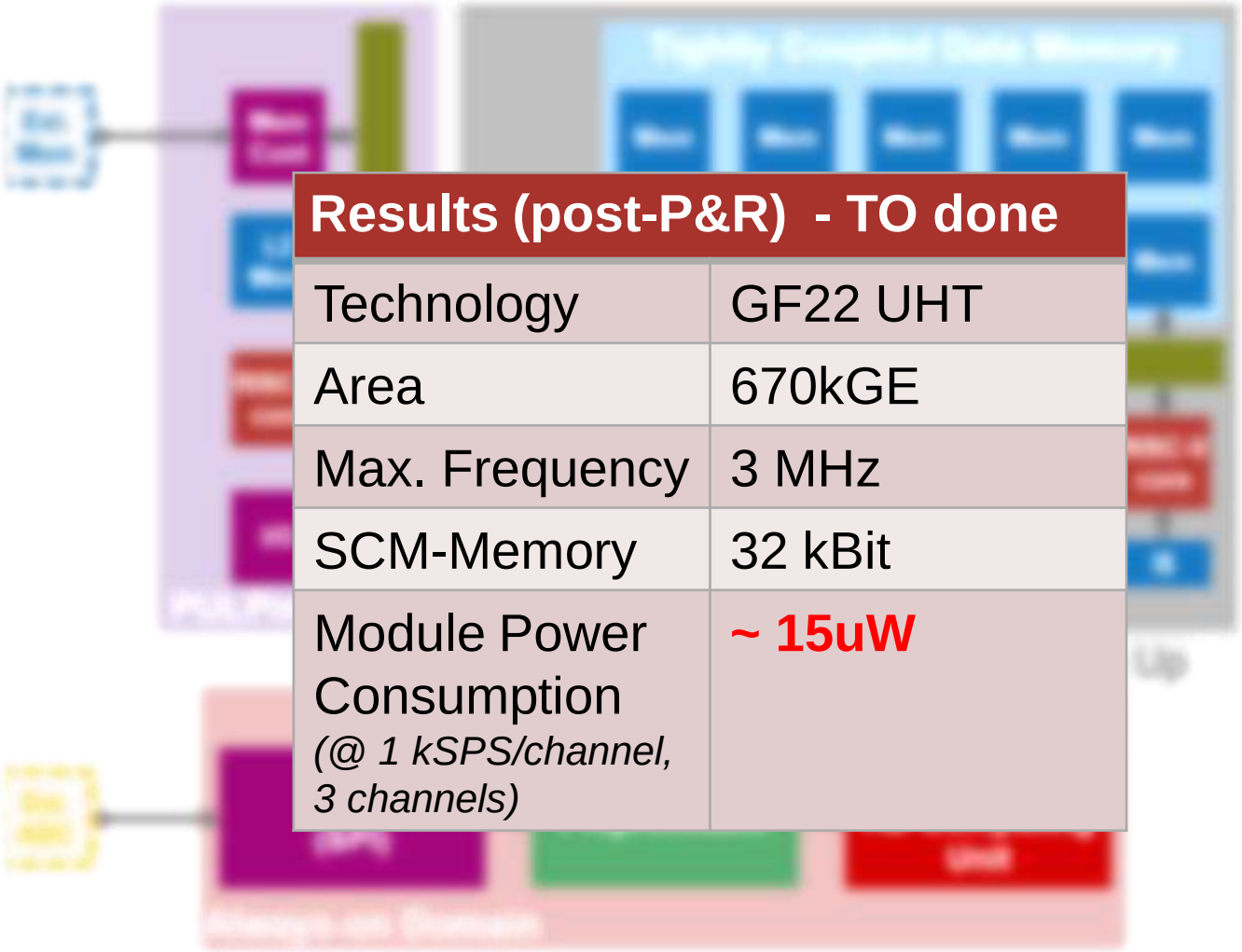




HD-Based smart Wake-Up Module



HD-Based smart Wake-Up Module



Results (post-P&R) - TO done	
Technology	GF22 UHT
Area	670kGE
Max. Frequency	3 MHz
SCM-Memory	32 kBit
Module Power Consumption *(@ 1 kSPS/channel, 3 channels)*	**~ 15uW**

Integrating Accelerators into PULP – The big picture

RISC-V Cores

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

Peripherals

JTAG	SPI
UART	I2S
DMA	GPIO

Interconnect

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

Platforms

Tens of active users, many use-cases
HW, SW specialization, verification, documentation, training
Cannot be sustained by one University, or two…

Single Core
- PULPino
- PULPissimo

Multi-core
- Fulmine
- Mr. Wolf

Multi-cluster
- Hero
- Open Piton

IOT → HPC

Accelerators

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

Academic open source → Industrial open source

OPENHW GROUP™ PROVEN PROCESSOR IP

Rick O'Connor (OpenHW CEO, former RISC-V foundation director)

- **OpenHW Group** is a not-for-profit, global organization (EU,NA,Asia) driven by its members and individual contributors where HW and SW designers collaborate in the development of open-source cores, related IP, tools and SW such as the **Core-V** family of cores.
- **OpenHW Group** provides an infrastructure for hosting high quality open-source HW developments in line with industry best practices.

PULP Parallel Ultra Low Power

RI5CY, ARIANE

CORE-V™

Alibaba Group, ASHLING, axiomise, bluespec, Barcelona Supercomputing Center, BTA Design Services, CMC Microsystems, DATVM, ECSPEC, EMBECOSM, ETH zürich, FUTUREWEI, GREENWAVES, HENSOLDT, HUAWEI, IAR SYSTEMS, imperas, metrics, Minres, MYTHIC, NXP, onespin, OPERSYS, Polytechnique Montréal, PRAESUM COMMUNICATIONS, SECURE THINGZ, SILICON LABS, Symbiotic EDA, THALES, ultraSoC, uOttawa, University of Toronto, The University of Utah, verifAI, VeriSilicon

A vertical, application-focused open approach

opentitan

- OpenTitan is the first open source silicon project building a transparent, high-quality reference design for silicon root of trust (RoT) chips.
- Founding Partners

ETH zürich Giesecke+Devrient Creating Confidence Google

lowRISC nuvoTon Western Digital.

Feel the momentum!

Ibex RISC-V core, flash interface, communications ports, cryptography accelerators, and more.

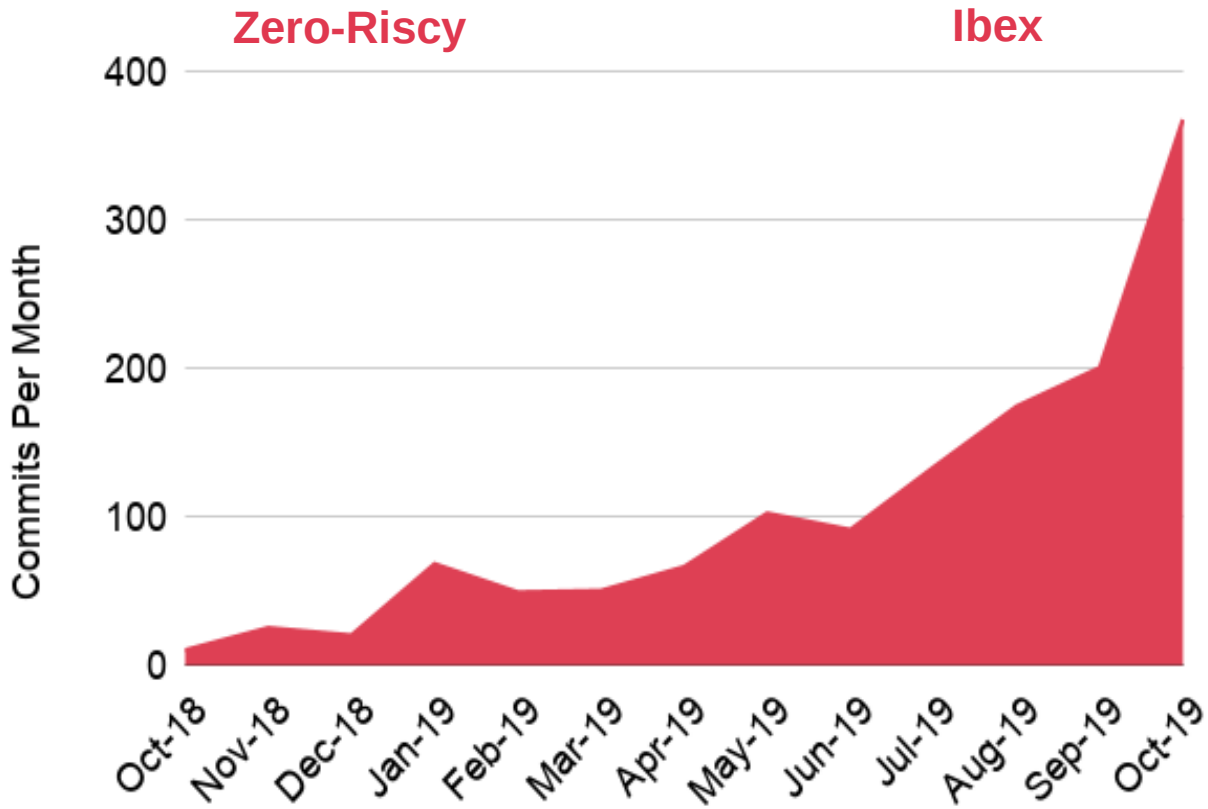
Vibrant repository

35+ Contributors

1300+ Contributions

470 GitHub Issues

PULP Parallel Ultra Low Power → lowRISC



HPC Vertical: The European Processor Initiative

European Processor Initiative

Europe Needs its own Processors

- Processors now control almost every aspect of our lives
- Security (back doors etc.)
- Possible future restrictions on exports to EU due to increasing protectionism
- A competitive EU supply chain for HPC technologies will create jobs and growth in Europe
- Sovereignty (data, economical, embargo)

- High Performance General Purpose Processor for HPC
- **High-performance RISC-V based accelerator**
- Computing platform for autonomous cars
- Will also target the AI, Big Data and other markets in order to be economically sustainable

Closing thoughts… the open HW revolution

For science … fundamental "research infrastructure"

- Community building: sharing of ideas, artefacts
- Fantastic tool for dissemination (more citations ☺)!
- Reduce "getting up to speed" overhead for partners
- Enables fair and well controlled benchmarking

For Business … it is truly disruptive

- Reduces the NRE cost for silicon design
- 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

For society …long term sustained benefits

- More innovation, growth, jobs
- Personalized silicon vision "Moore-for-all"
- More Secure, safe, auditable HW

DARPA

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

USA's
electronics resurgence initiative

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