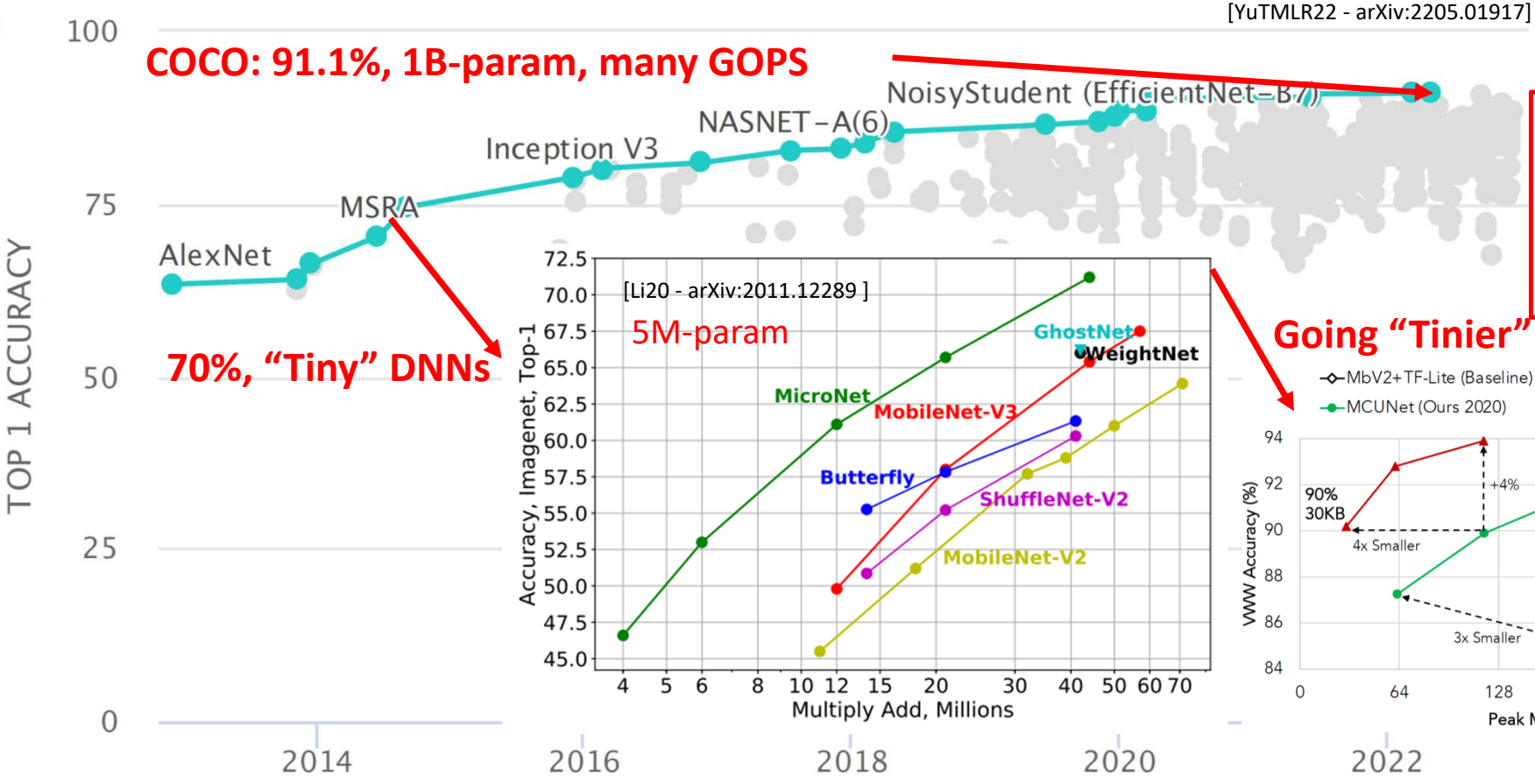
EU – India
Joint Researchers Workshop on Semiconductors
9 October 2024

The Open-Source PULP Platform for tinyML Heterogeneous Hardware Acceleration

Francesco Conti, Tenure-Track Assistant Professor
University of Bologna, Italy

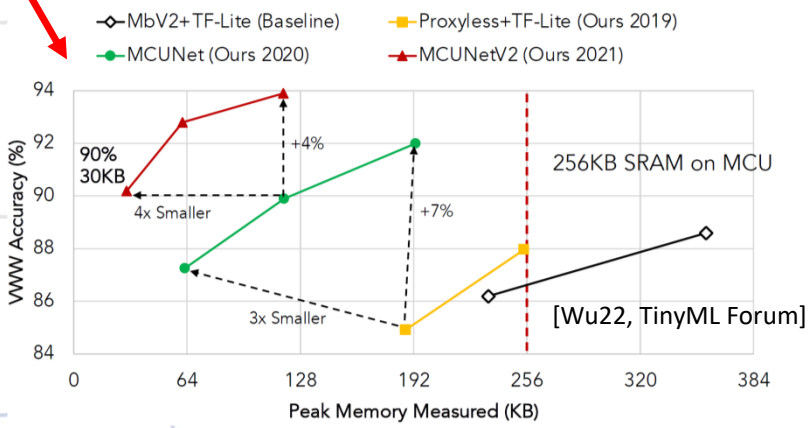
tinyML workloads

[YuTMLR22 - arXiv:2205.01917]



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

Going "Tinier"



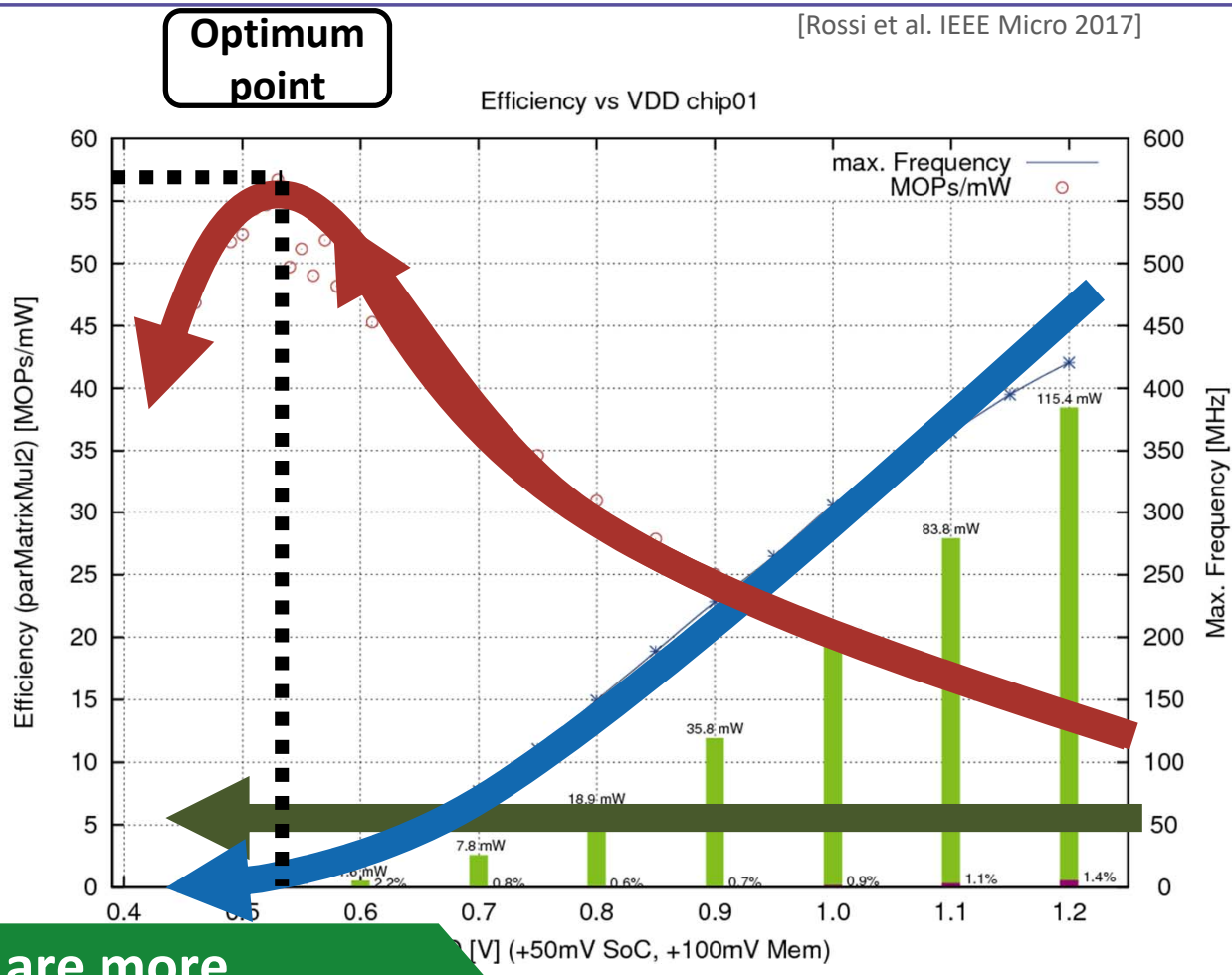
[Wu22, TinyML Forum]

"Model zoo" fast evolution → need programmable solutions

Perf. + Eff. + Flexibility ← **Parallelism**

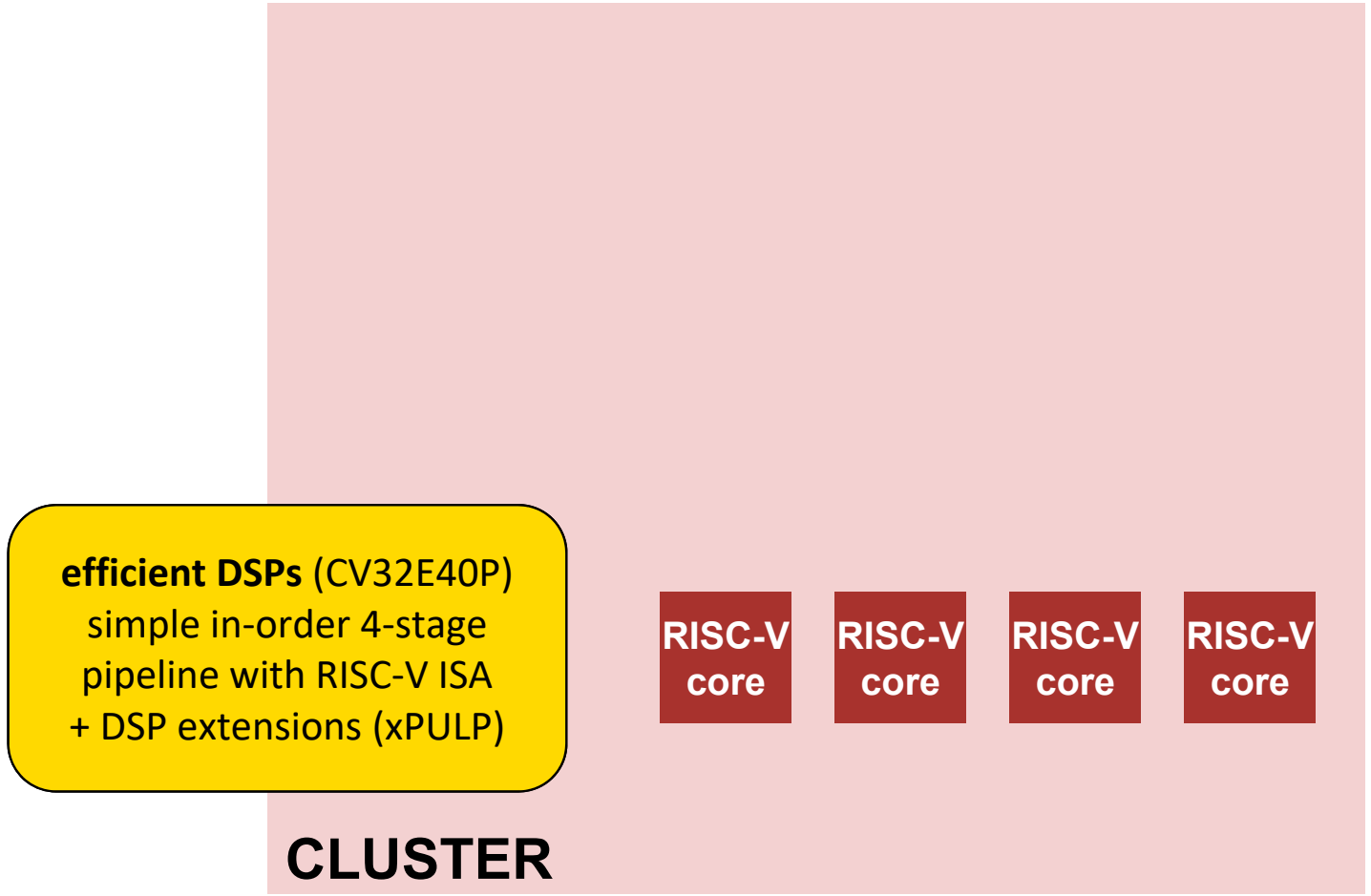
- **As VDD decreases, operating speed decreases as well.**
- **However efficiency increases→ more work done per Joule**
  - Until leakage effects start to dominate
- **More units in parallel**
  - Get performance up (if you can keep them busy)
  - Energy efficiency stays high!

[Rossi et al. IEEE Micro 2017]



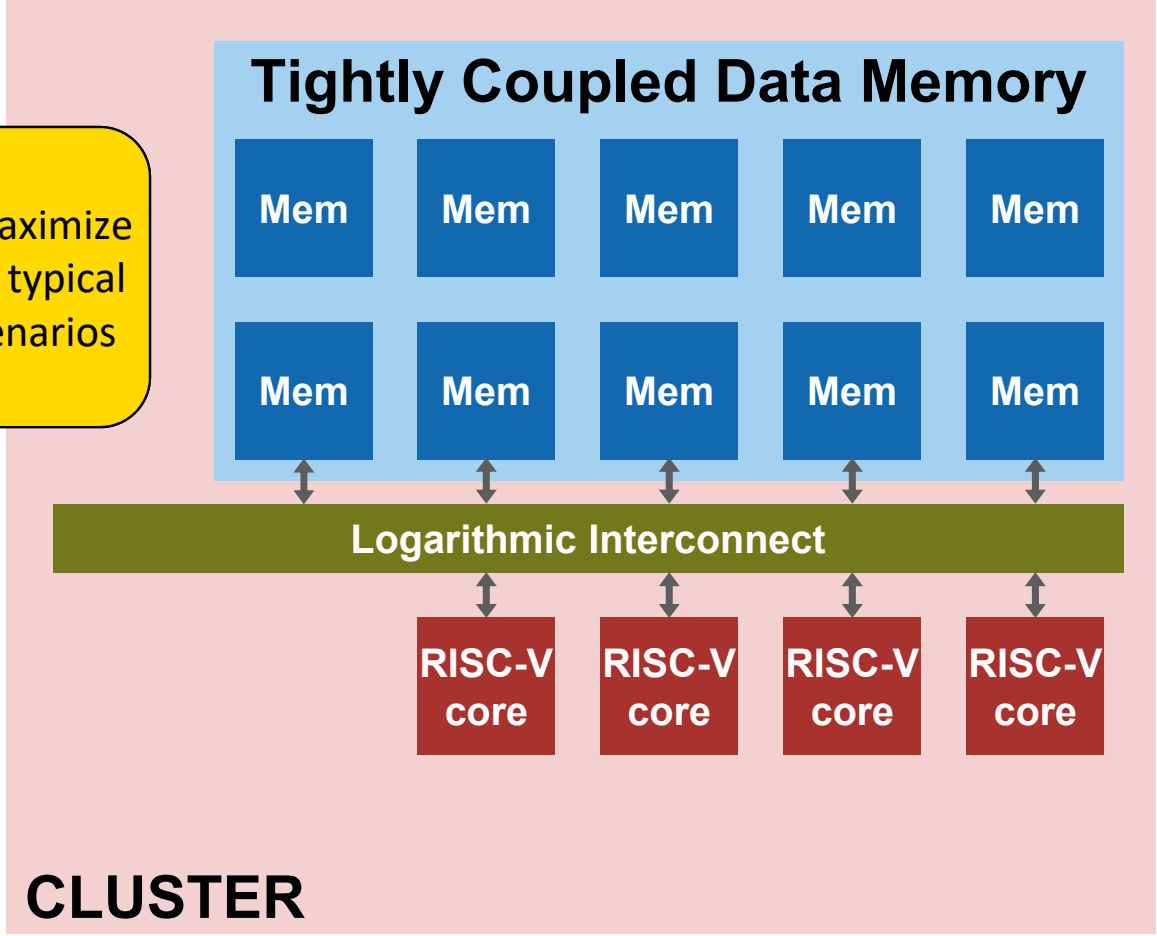
N cores running at moderate f, low Vdd are more energy efficient than a single core at N×f, high Vdd

The **PULP** paradigm

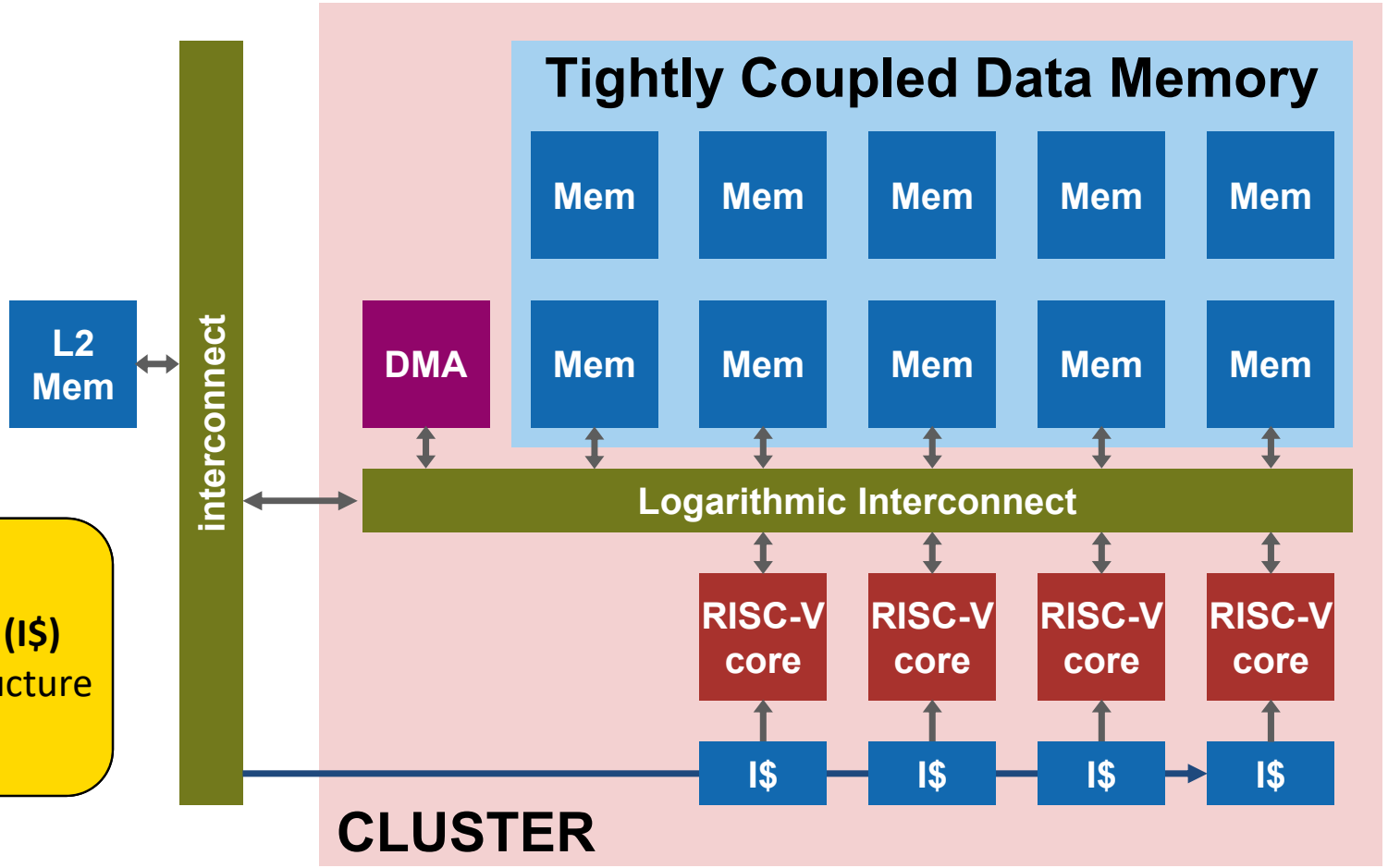


The **PULP** paradigm

bank interleaving to maximize available bandwidth in typical parallel computing scenarios

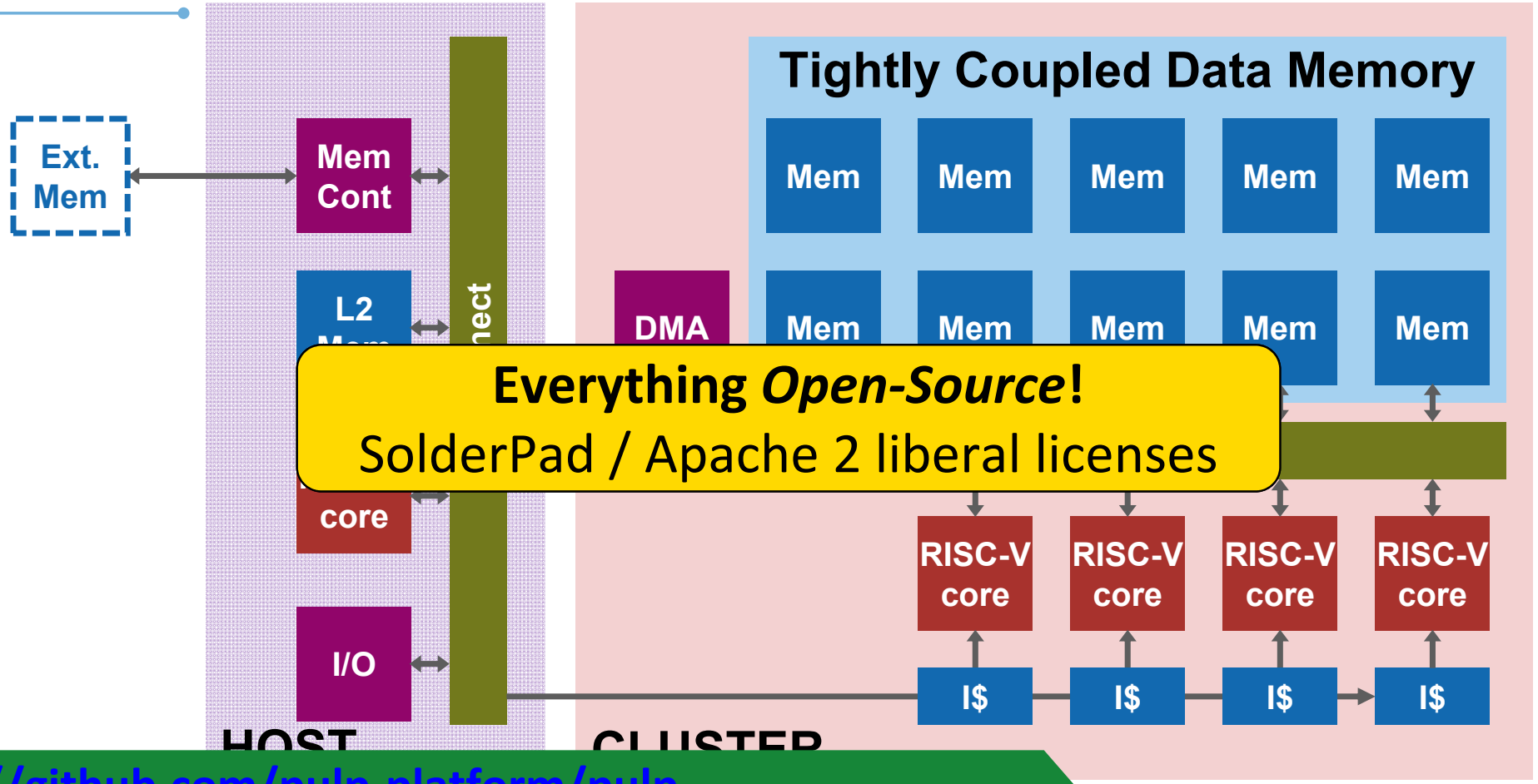


The **PULP** paradigm



Add **DMA** and **I-Cache (I$)** communication infrastructure

The **PULP** paradigm



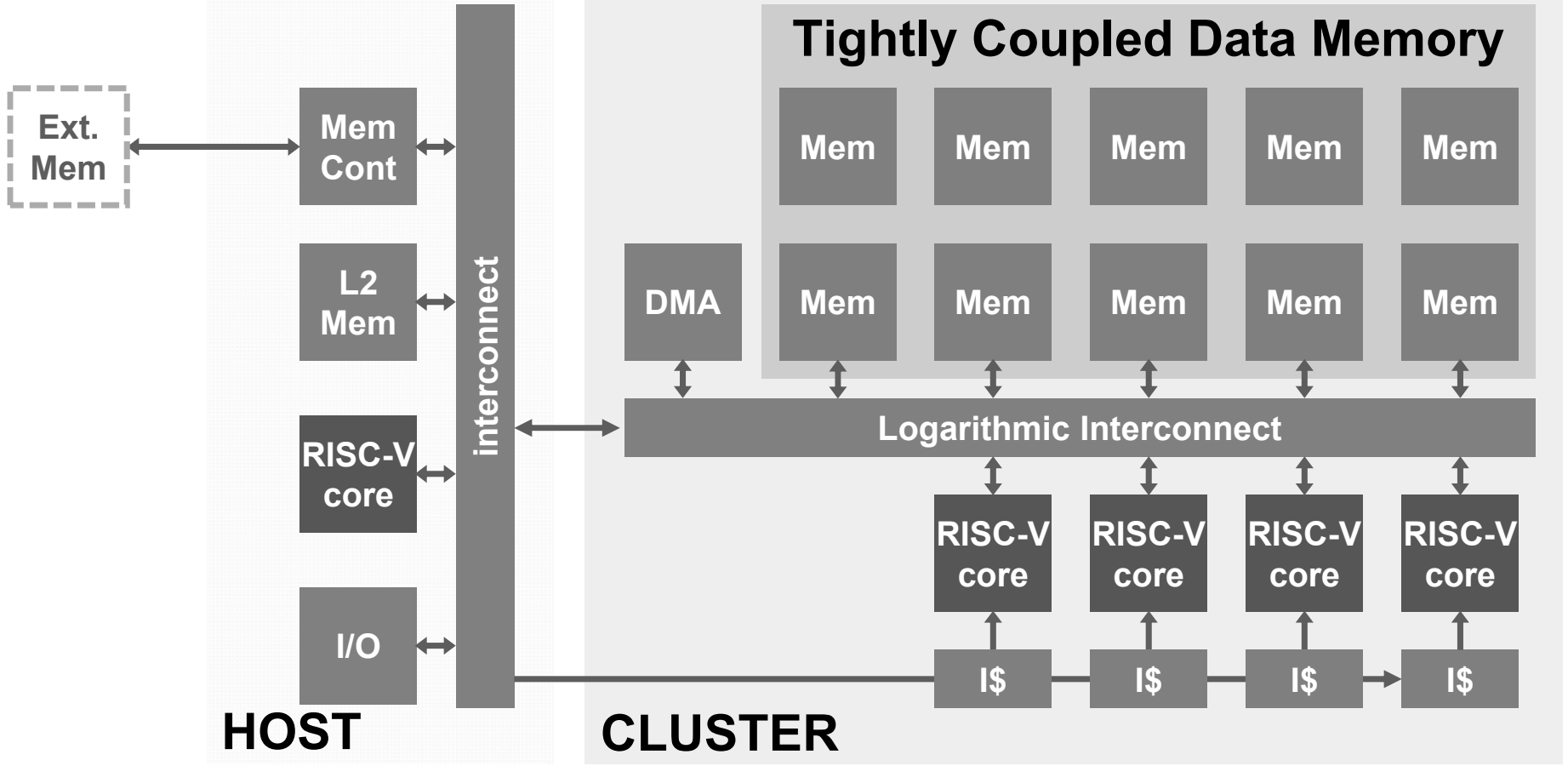
Everything *Open-Source*!

SolderPad / Apache 2 liberal licenses

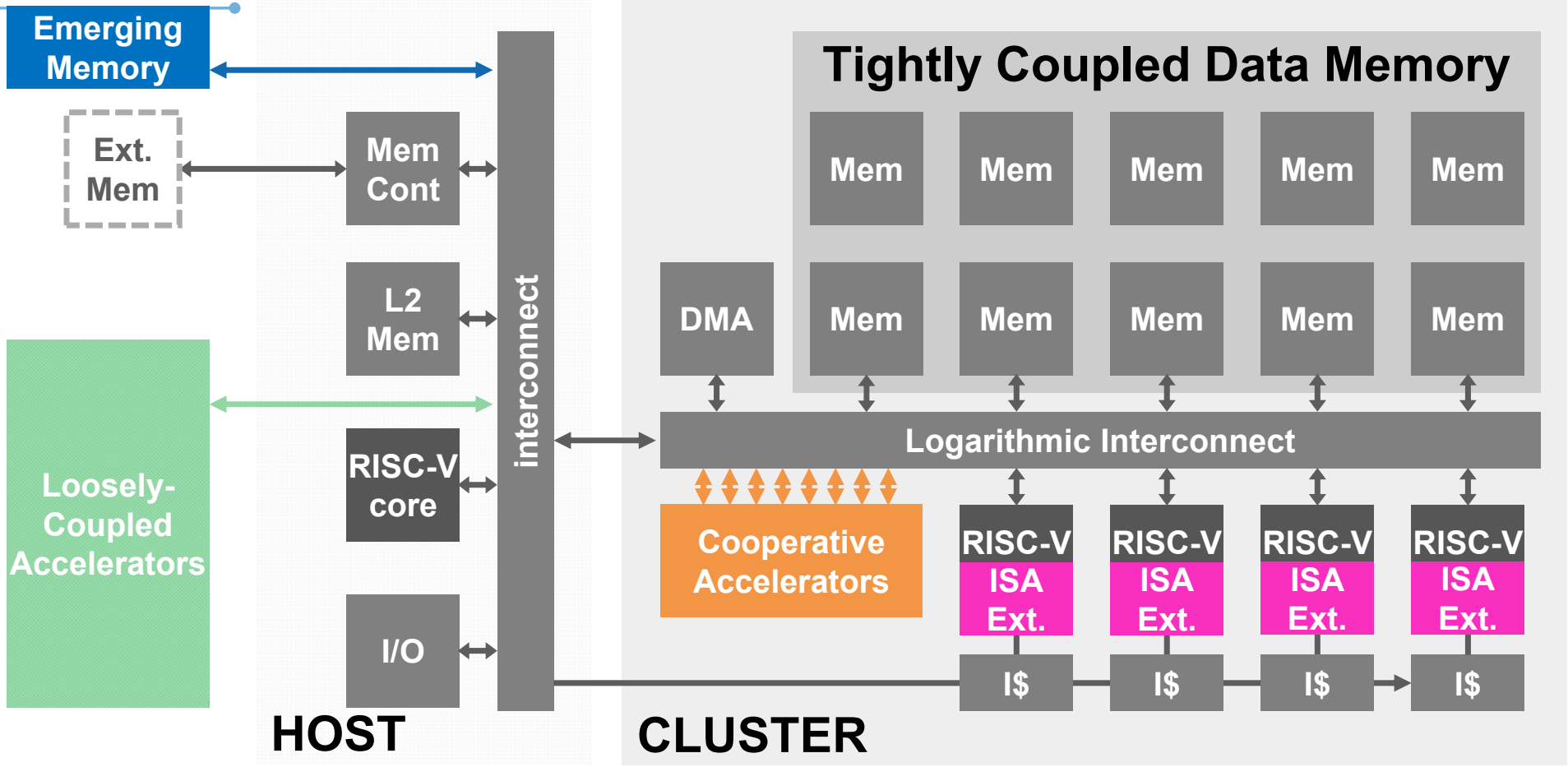
https://github.com/pulp-platform/pulp

https://pulp-platform.org

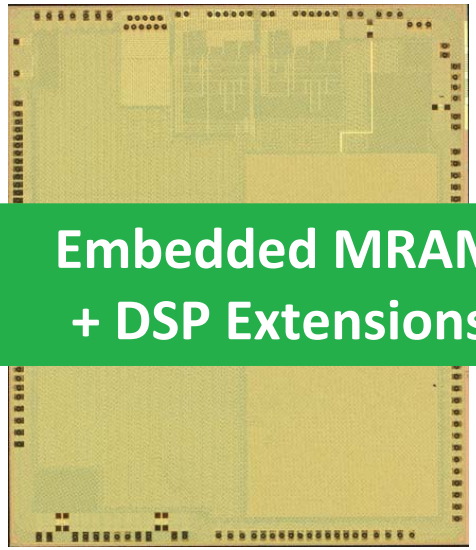
Heterogeneous computing on PULP



Heterogeneous computing on PULP



Heterogeneous computing on PULP



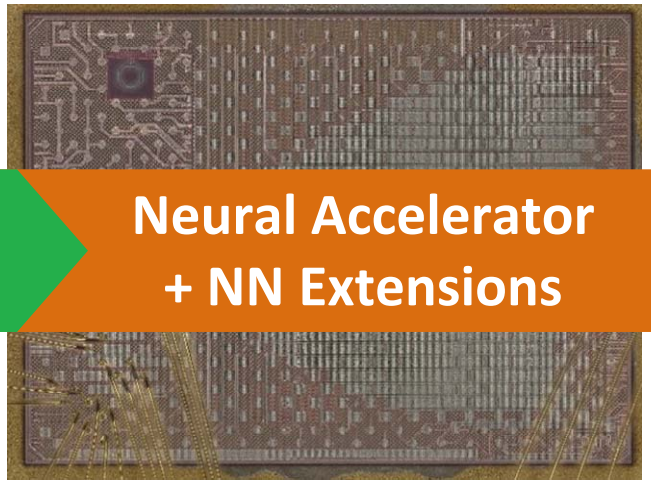
Embedded MRAM + DSP Extensions

Vega 22nm FDX

UNIBO, ETHZ, GreenWaves

[1] D. Rossi et al., ISSCC'21

[2] D. Rossi et al., JSSC'21



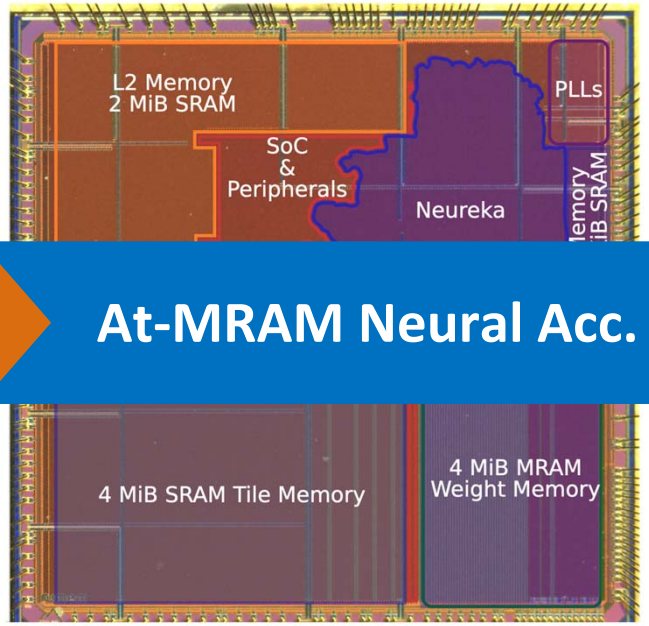
Neural Accelerator + NN Extensions

Marsellus 22nm FDX

UNIBO, ETHZ, Dolphin

[3] F. Conti et al., ISSCC'23

[4] F. Conti et al., JSSC'23



At-MRAM Neural Acc.

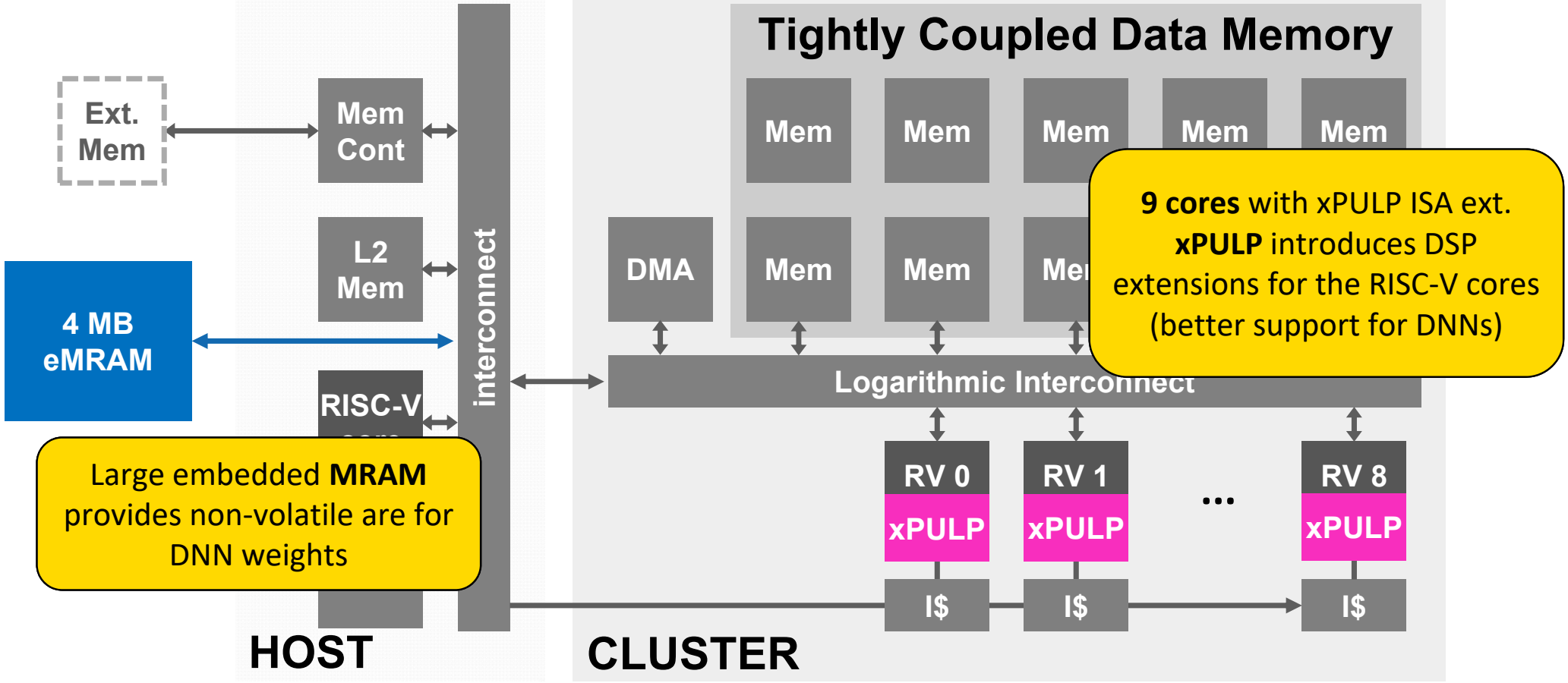
Siracusa 16nm FinFET

ETHZ, UNIBO, Meta

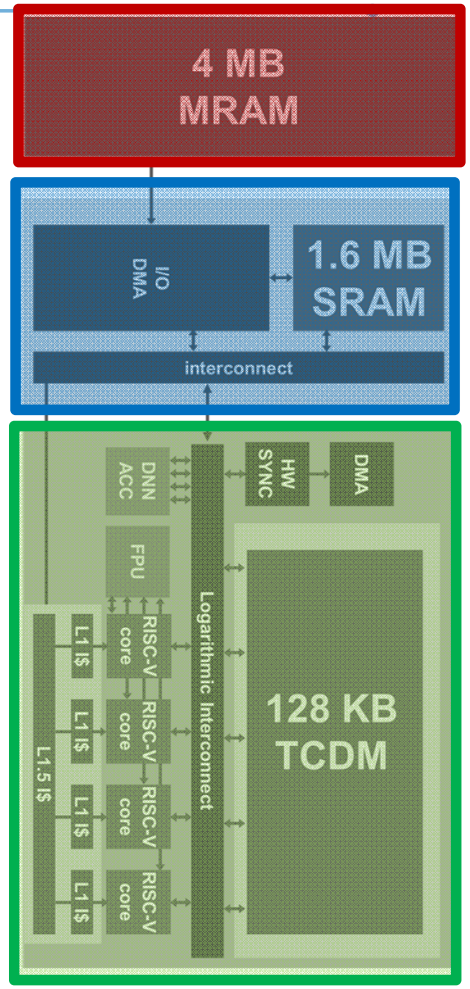
[5] M. Scherer et al., ESSCIRC'23

[6] A. Prasad et al., JSSC'24

Vega: extreme-edge AI MCU



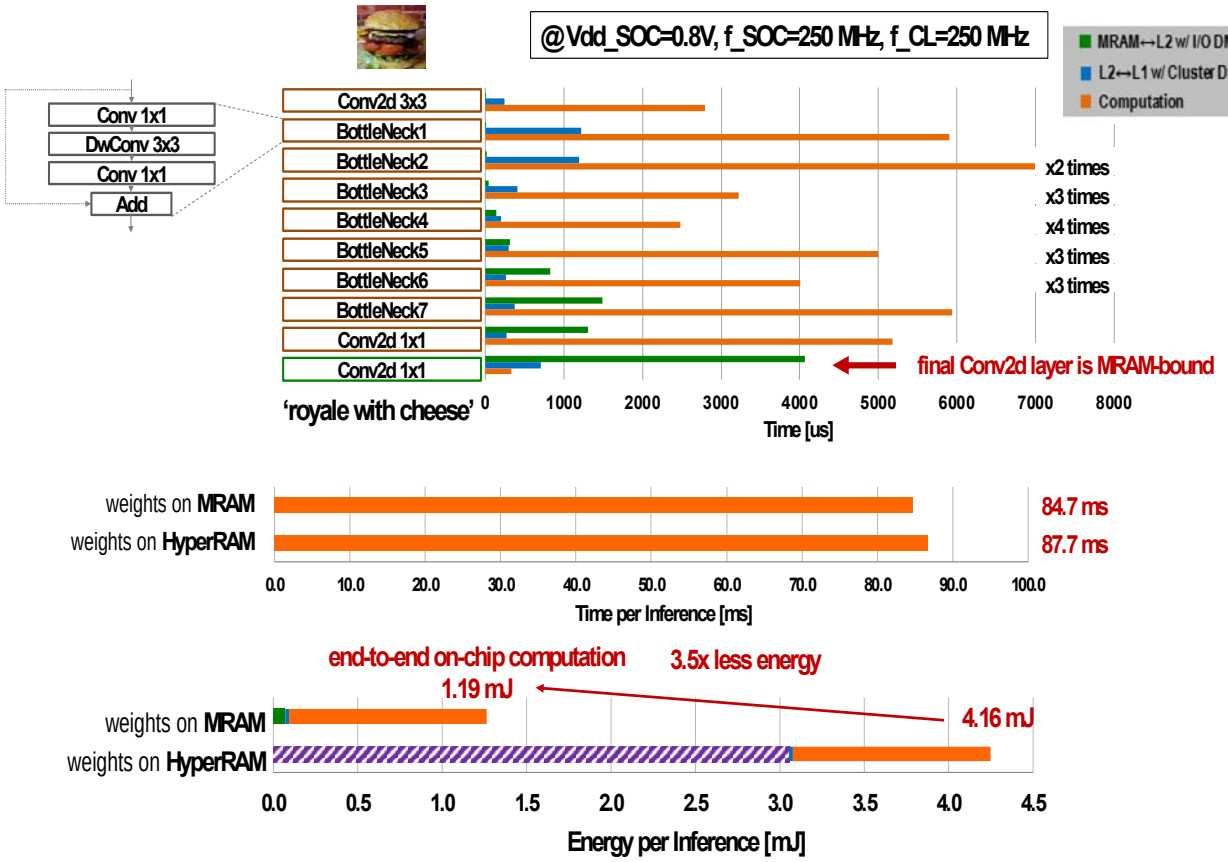
Vega: DNNs on MRAM/SRAM



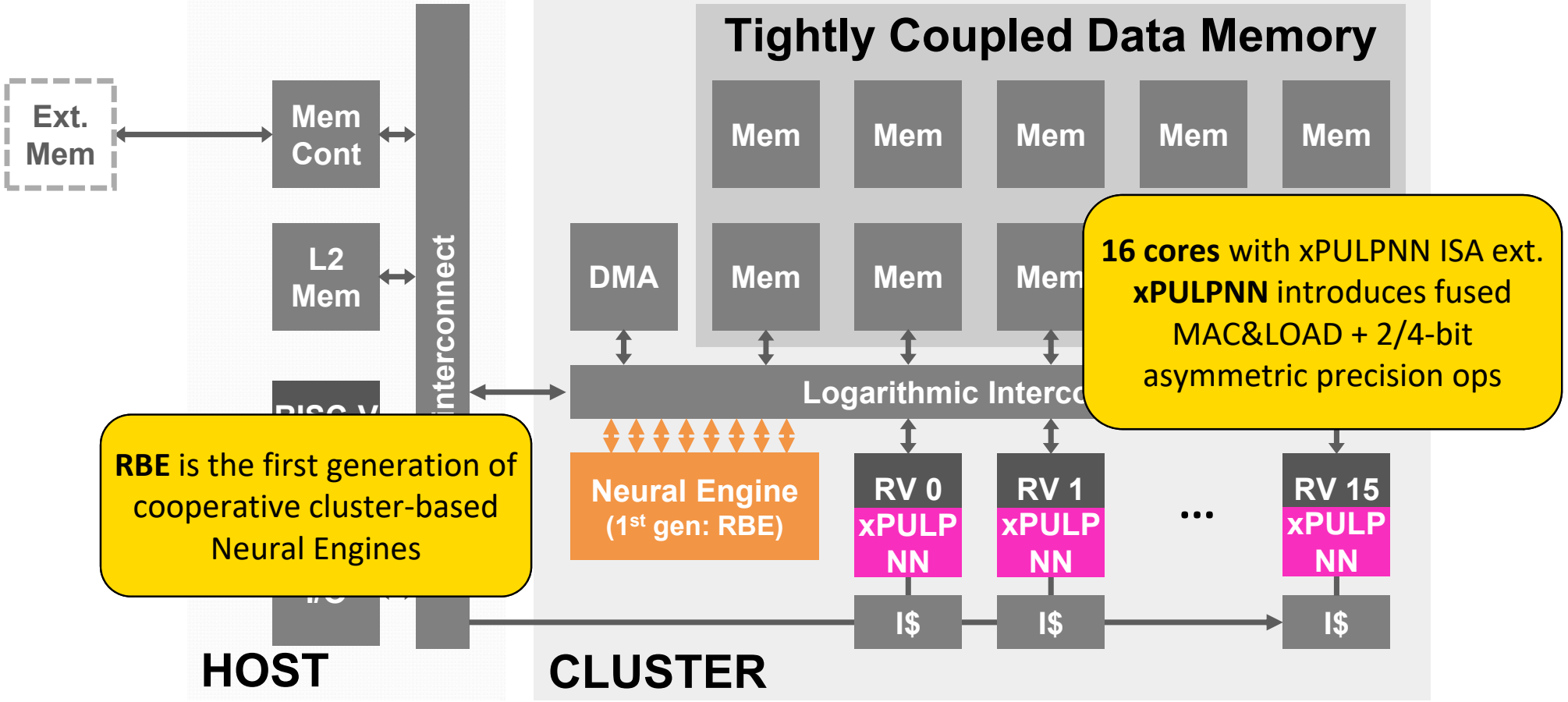
I/O DMA
MRAM / L2 tiling
(Weights Only)

Cluster DMA
L2 / L1 tiling
(Weights + Activations)

MobileNet v2

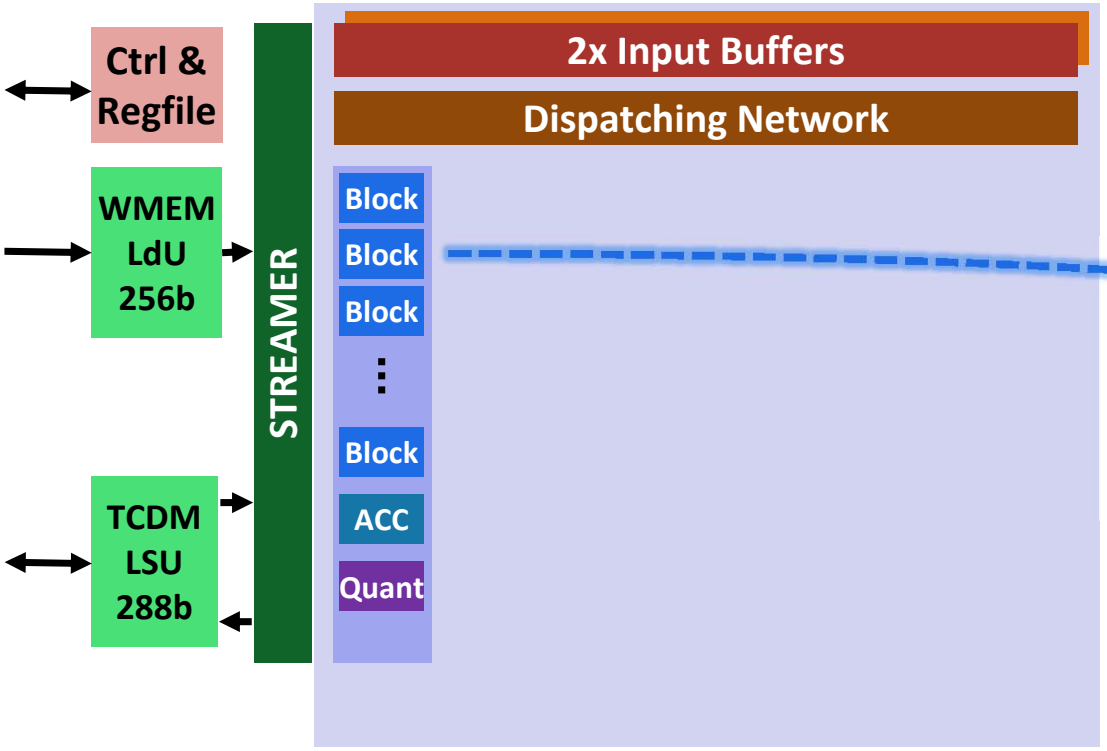


Marsellus: AI-IoT Heterogeneous SoC

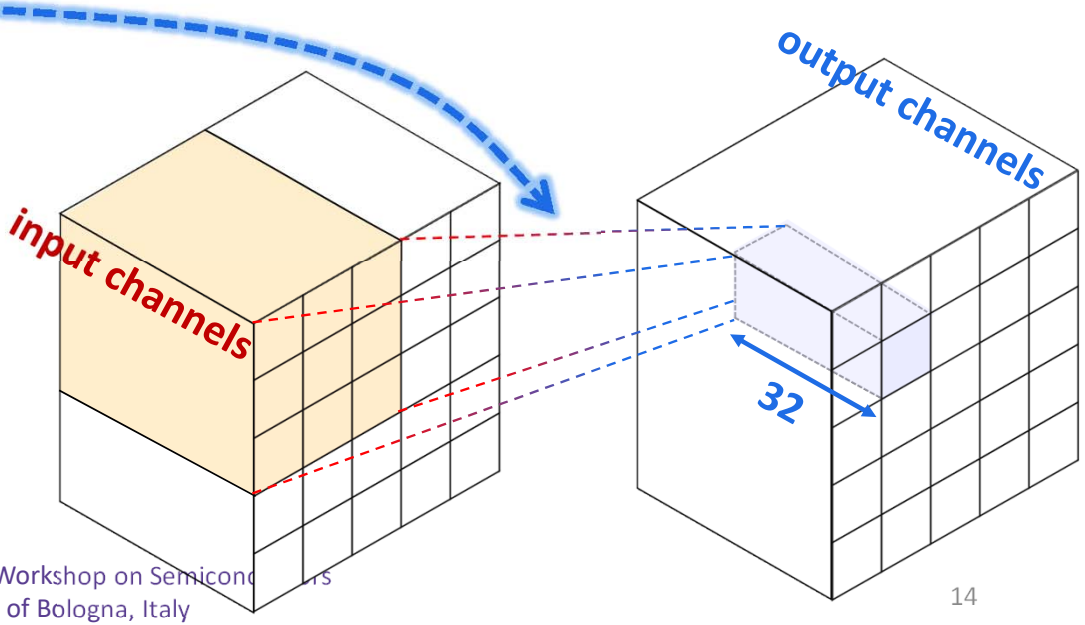


The Neural Engine family

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

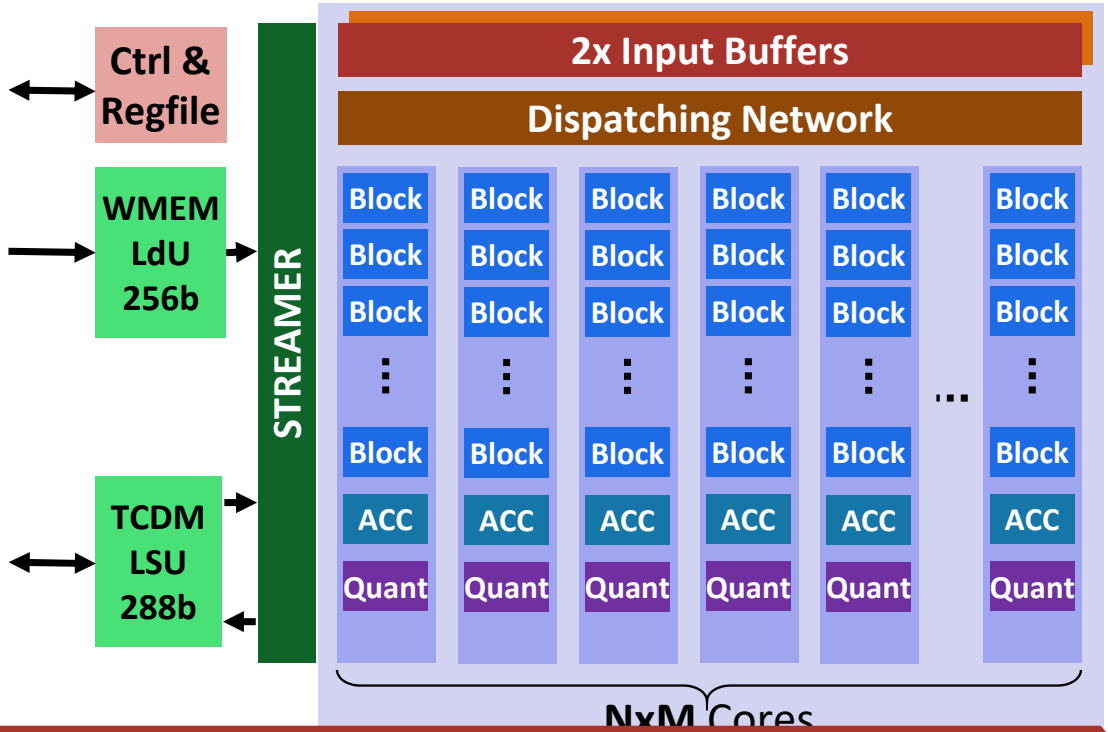


- **Partially bit-serial dataflow for CONV3x3, PW1x1, DWCONV3x3**
  - 3x3, 1x1 and 3x3 depthwise mode
  - Act. 2-8b (1st) / 8b (NE16/N-EUREKA), Weights 2-8b
- **Core – receptive field of 1 output px across 32 output chans → more cores, larger output "tile"**



The Neural Engine family

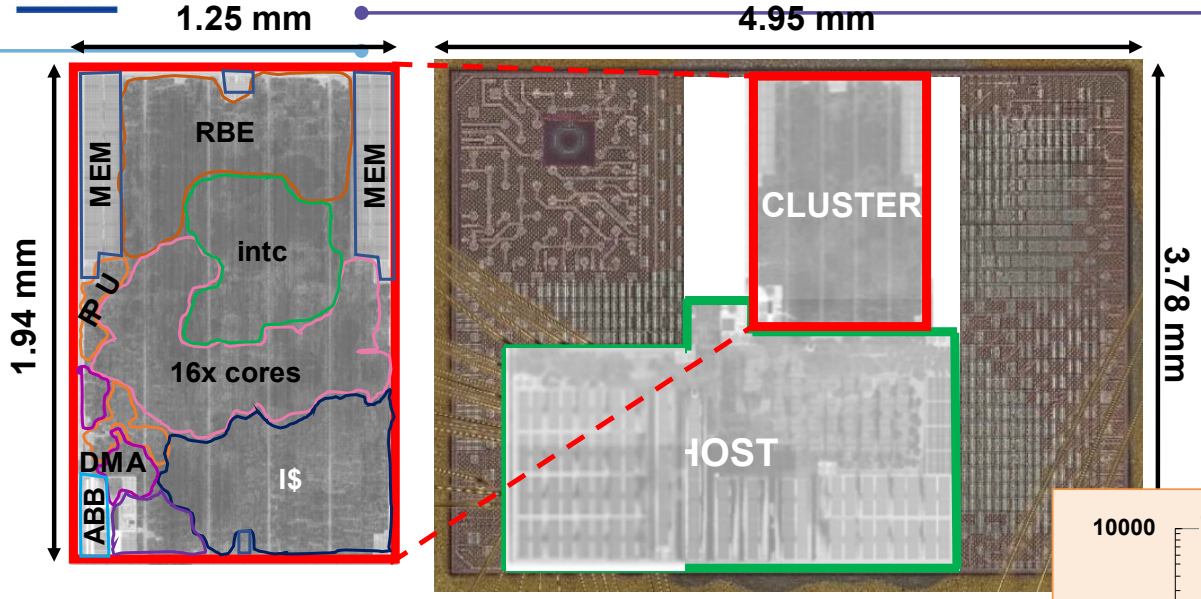
$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..Wbit}\sum_{k_{in}} 2^i(\mathbf{W_{bin}}(k_{out},k_{in}) \otimes \mathbf{x}(k_{in}))\right)$$



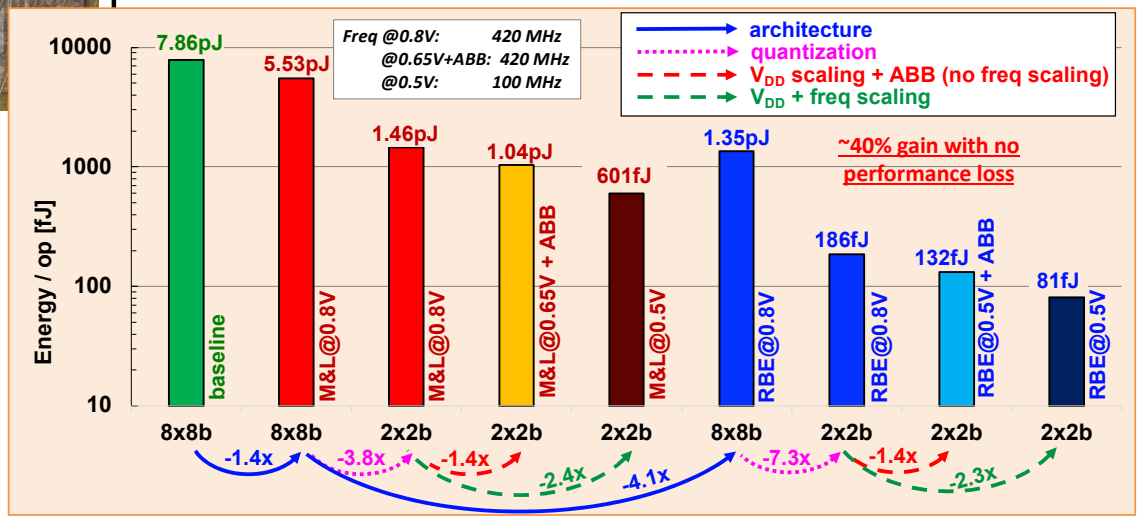
- **Partially bit-serial dataflow for CONV3x3, PW1x1, DWCONV3x3**
  - 3x3, 1x1 and 3x3 depthwise mode
  - Act. 2-8b (1st) / 8b (NE16/N-EUREKA), Weights 2-8b
- **Core – receptive field of 1 output px across 32 output chans → more cores, larger output "tile"**
- **"Two and a half" generations**
  - **Gen 1 (Marsellus): Reconfigurable Binary Engine (RBE)** – fully bit serial, full configurability of size in activations & weights
  - **Gen 2: NE16** – 8b activations, many more supported layers
  - **Gen 2.5 (Siracusa): N-EUREKA** – parametric size, embedded weight memory

1st gen https://github.com/pulp-platform/rbe
2nd gen https://github.com/pulp-platform/ne16
2.5th gen https://github.com/pulp-platform/neureka

Marsellus: AI-IoT Heterogeneous SoC



Combine heterogeneous architecture, quantization, Vdd scaling and Adaptive Body Biasing



Prototype implemented in GF 22FDX

- flip-well LVT & SLVT cells, 2.43mm^2 for CLUSTER

1st gen RBE (9-cores)

Not only academia! GAP9

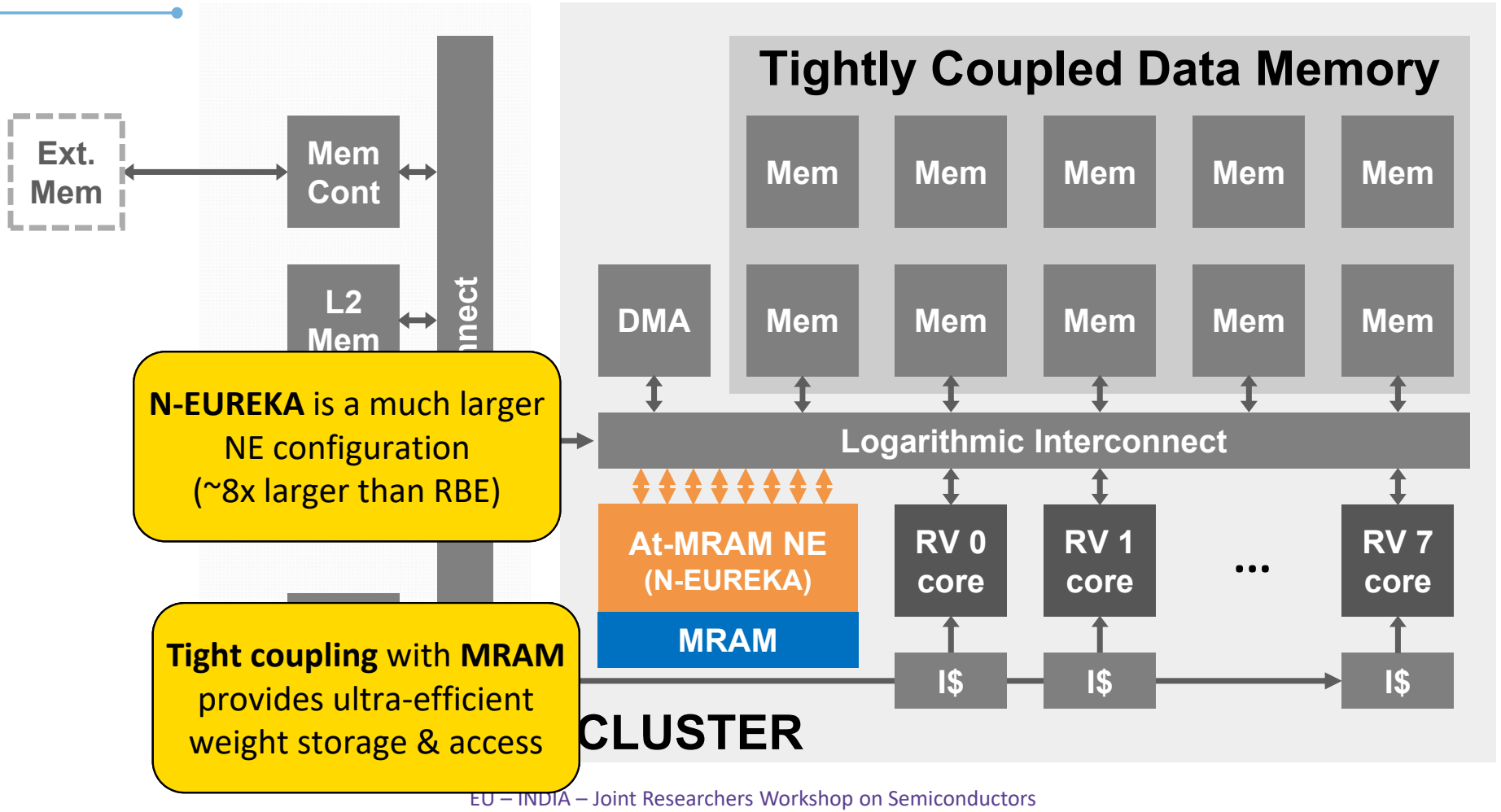
- **Best-in-class in latency and energy efficiency in MLPerf Tiny 1.0!**

Submitter	Board Name	SoC Name	Processor(s) & Number	Accelerator(s) & Number	Software	Notes	Benchmark Results							
							Task	Visual Wake Words		Image Classification		Keyword Spotting		Anomaly Deteciton	
							Data	Visual Wake Words Dataset		CIFAR-10		Google Speech Commands		ToyADMOS (ToyCar)	
							Model	MobileNetV1 (0.25x)		ResNet-V1		DSCNN		FC AutoEncoder	
							Accuracy	80% (top 1)		85% (top 1)		90% (top 1)		0.85 (AUC)	
							Units	Latency in ms	Energy in uJ	Latency in ms	Energy in uJ	Latency in ms	Energy in uJ	Latency in ms	Energy in uJ
Greenwaves Technologies	GAP9 EVK	GAP9	RISC-V Core (1+9)	NE16 (1)	GreenWaves GAPFlow	GAP9 (370MHZ, 0.8Vcore)		1.13	58.4	0.62	40.4	0.48	26.7	0.18	7.29
Greenwaves Technologies	GAP9 EVK	GAP9	RISC-V Core (1+9)	NE16 (1)	GreenWaves GAPFlow	GAP9 (240MHZ, 0.65Vcore)		1.73	40.8	0.95	27.7	0.73	18.6	0.27	5.25
OctoML	NRF5340DK	nRF5340	Arm® Cortex®-M33		microTVM using CMSIS-NN backend	128MHz		232.0		316.1		76.1		6.27	
OctoML	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		microTVM using CMSIS-NN backend	120MHz, 1.8Vbat		301.2	15531.4	389.5	20236.3	99.8	5230.3	8.60	443.2
OctoML	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		microTVM using native codegen	120MHz, 1.8Vbat		336.5	7131.6	389.2	21342.3	144.0	7950.5	11.7	633.7
Plumerai	B_U585I_IOT02A	STM32U585	Arm® Cortex®-M33		Plumerai Inference Engine 2022.09	160MHz		107.0		107.1		35.4		4.90	
Plumerai	CY8CPROTO-062-4343w	PSoC 62 MCU	Arm® Cortex®-M4		Plumerai Inference Engine 2022.09	150MHz		192.5		193.1		61.4		6.70	
Plumerai	DISCO-F746NG	STM32F746	Arm® Cortex®-M7		Plumerai Inference Engine 2022.09	216MHz		57.0		64.8		19.1		2.30	
Plumerai	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		Plumerai Inference Engine 2022.09	120MHz		208.6		173.2		71.7		5.60	
Silicon Labs	xG24-DK2601B	EFR32MG24	Arm® Cortex®-M33	Silicon Labs MVP(1)	TensorFlowLite for Microcontrollers, CMSIS-NN, Silicon Labs Gecko SDK			111.6	1139.2	120.9	1234.7	36.3	401.9	5.43	47.3
STMicroelectronics	NUCLEO-H7A3ZI-Q	STM32H7A3ZIT6Q	Arm® Cortex®-M7		X-CUBE-AI v7.3.0	280MHz, 3.3Vbat		50.7	7978.5	54.3	8707.3	16.8	2721.8	1.82	266.5
STMicroelectronics	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		X-CUBE-AI v7.3.0	120MHz, 1.8Vbat		230.5	10066.6	226.9	10681.6	75.1	3371.7	7.57	323.0
STMicroelectronics	NUCLEO-U575ZI-Q	STM32U575ZIT6Q	Arm® Cortex®-M33		X-CUBE-AI v7.3.0	160MHz, 1.8Vbat		133.4	3364.5	139.7	3642.0	44.2	1138.5	4.84	119.1
Syntiant	NDP9120-EVL	NDP120	M0 + HiFi	Syntiant Core 2 (98MH	Syntiant TDK	Syntiant Core 2 (98MHz, 1		4.10	97.2	5.12	139.4	1.48	43.8		
Syntiant	NDP9120-EVL	NDP120	M0 + HiFi	Syntiant Core 2 (30MH	Syntiant TDK	Syntiant Core 2 (30MHz, 0		12.7	71.7	16.0	101.8	4.37	31.5		
Qualcomm Innovation Center	Next Generation Snapdragon Mobile Platform HDK	Next Generation Snapdragon Mobile Platform	Qualcomm Kryo CPU(1)	Qualcomm Sensing Hub(1)	Qualcomm AI Stack										

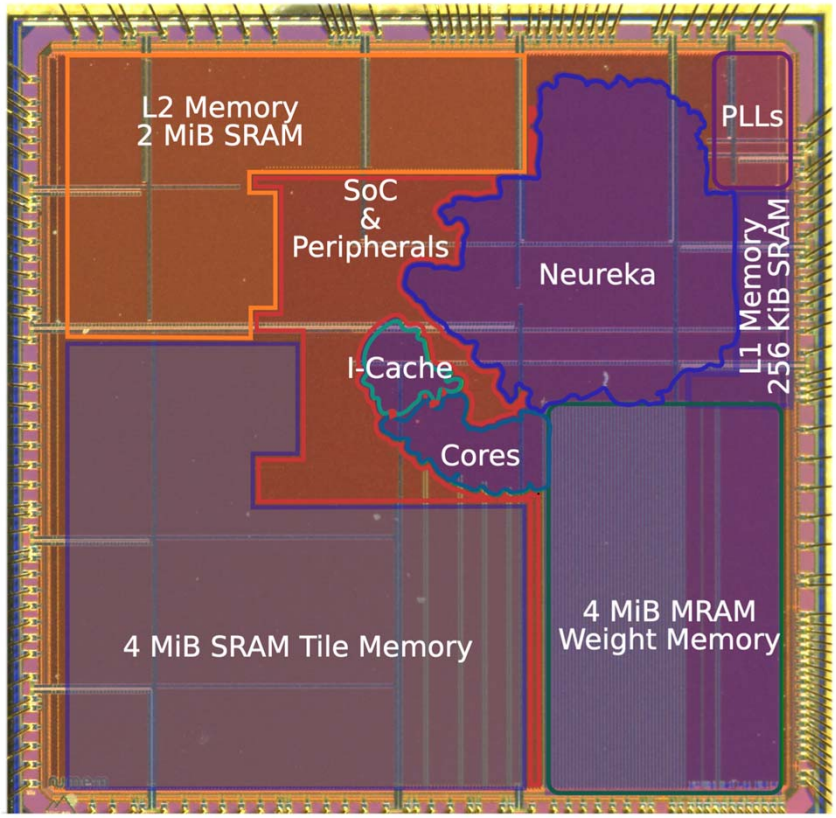
Latency 1.73ms (against ~29ms on GAP8)
Energy 41uJ (against 1450uJ on GAP8)

2nd gen NE16 (9-cores)

Siracusa: towards XR computing



Siracusa: towards XR computing



	Vega [1]	Diana [2]	Marsellus [3]	[4]	[5]	*Siracusa*
Technology	22nm FDX	22nm FDX	22nm FDX	40nm	22nm	16nm FinFET
Area	10mm^2	10.24mm^2	8.7mm^2	25mm^2	8.76mm^2	16mm^2
On-chip mem	1728 KB SRAM 4 MB MRAM (L3)	896 KB SRAM	1152 KB SRAM	768 KB	1428 KB	6400 KB SRAM 4 MB MRAM (L1)
Peak Perf 8b	32.2 GOPS	140 GOPS	90 GOPS	N/A	146 GOPS	**698 GOPS**
Peak Eff 8b	1.3 TOPS/W	2.07 TOPS/W	1.8 TOPS/W	0.94 TOPS/W	0.7 TOPS/W	**2.68 TOPS/W**
Peak Eff (WxAb)	1.3 TOPS/W	4.1TOPS/W (2x2b) **600 TOPS/W (analog)**	12.4 TOPS/W (2x2b)	60.6 TOPS/W (1x1b)	0.7 TOPS/W	8.84 TOPS/W (2x8b)
Area Eff	3.2 GOPS/mm^2	21.2 GOPS/mm^2	47.4 GOPS/mm^2	N/A	58.3 GOPS/mm^2	**65.2 GOPS/mm^2**

[1] D. Rossi et al., JSSC'21
[2] P. Houshmand et al., JSSC'23
[3] F. Conti et al., JSSC'23
[4] M. Chang et al., ISSCC'22
[5] Q. Zhang et al., VLSI Symposium'22

Balance efficiency, peak performance, area efficiency without compromises in precision

N-EUREKA 36-cores configuration

[A. Prasad et al., "Siracusa: a 16nm Heterogeneous RISC-V SoC for Extended Reality with At-MRAM Neural Engine," IEEE Journal of Solid-State Circuits (accepted)]

2.5th gen N-EUREKA (36-cores + MRAM)
https://arxiv.org/abs/2312.14750

Conclusion

- **A taste of the PULP experience with 3 embedded AI SoC's...**
  - **Vega**: full PULP concept for embedded AI, embedded NVM
  - **Marsellus**: more architectural heterogeneity, 1st gen NE
  - **Siracusa**: at-MRAM tightly-coupled computing, "high-performance" architecture

- **Fully European design**
  - SwitzerlandCH + Italy IT
- **Open-source hardware 💡:**
  https://pulp-platform.org

ICOS International Cooperation On Semiconductors

European Commission

This project has received funding from the European Union's Horizon Europe research and innovation programme under GA N° 101092562

EU – INDIA – Joint Researchers Workshop on Semiconductors

www.icos-semiconductors.eu