ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

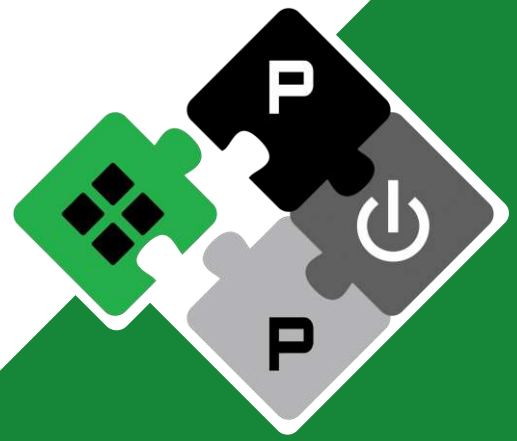
Toward End-to-End Open Platforms for the Embodied AI Era

Luca Benini

benny@stanford.edu

lbenini@iis.ee.ethz.ch

luca.benini@unibo.it



PULP Platform

Open Source Hardware, the way it should be!

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

Embodied AI: Artificial Intelligence Everywhere

Embodied AI!

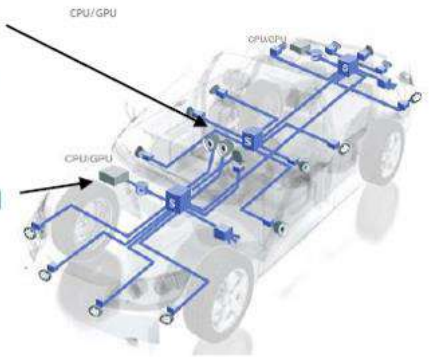
Level 4-5
Self Driving

Level 2-3
Decision Assistant

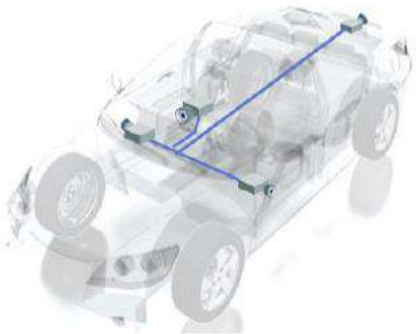
Level 1-2
Simple Aid

High-Speed, Reliable & Secure Nervous System

High-Performance Brain



Local Computing
"Behind" Every Sensor



Centralized Computing
Integrates Input From All Sensors (*Sensor Fusion*)
Similar to a Human Driver's Brain

2010 - 2018 | 2019 - 2025 | 2025...

Compute Power (TFLOPS)	Networking Speed (Gbit/s)
100	100
10	10
1	1
0.1	0.1

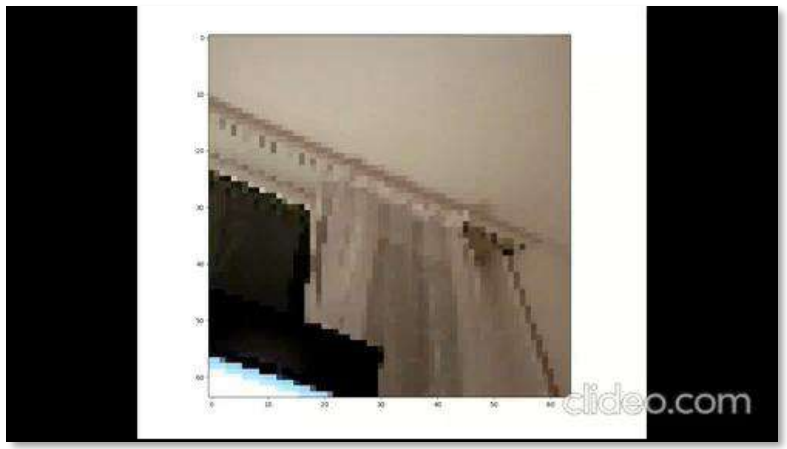
Efficient

On-car Computing
$P_{avg} < 1.5$ kW

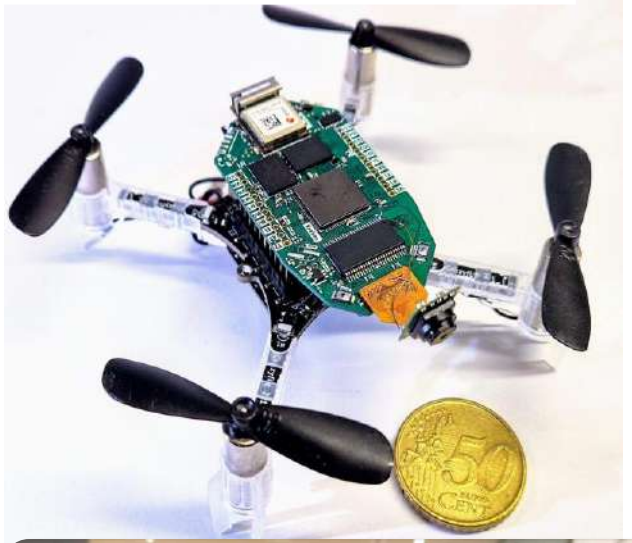
[SCR23]

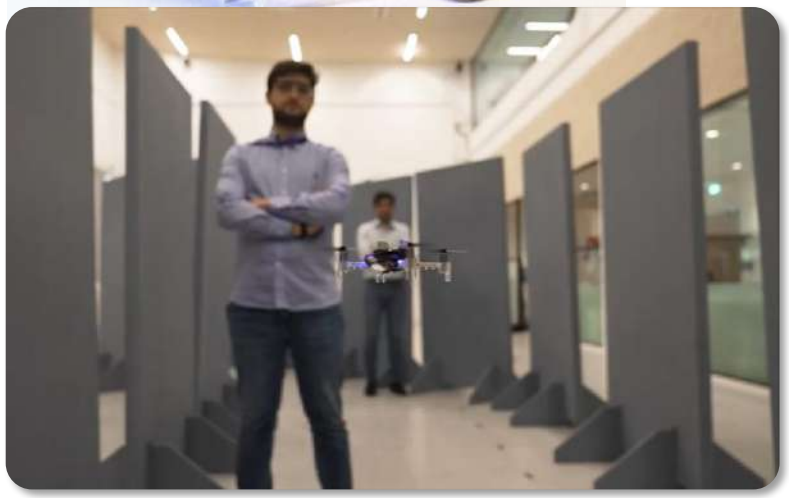
Embodied AI: Artificial Intelligence Everywhere

Smart Glasses



Nano-Drone





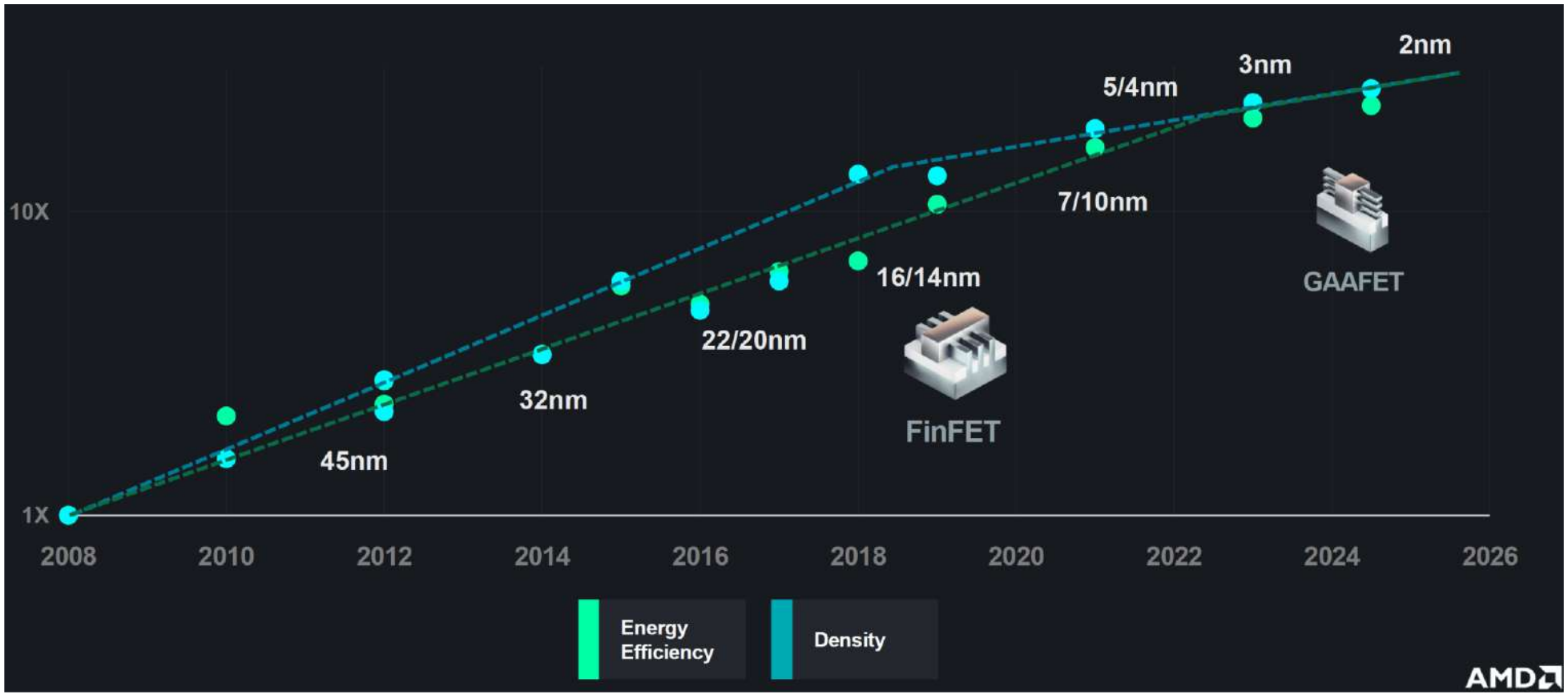
Efficient

On-car Computing
P_{avg} < 150 W

On-drone Computing
P_{avg} < 150 mW

On-glass Computing
P_{avg} < 1.50 mW

Embodied AI: Efficiency Challenge



[AMD HotChips24]

Efficient

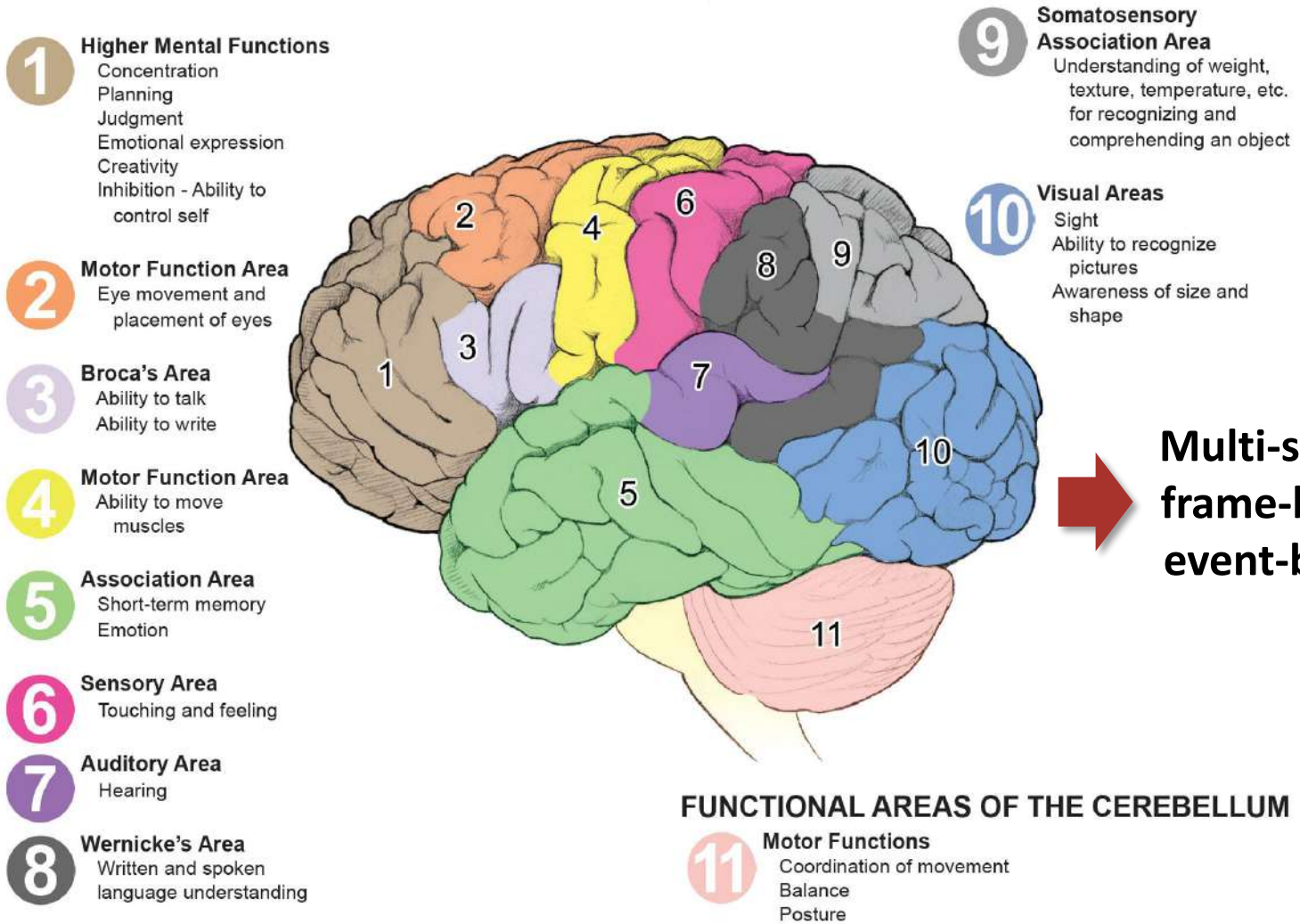
Model complexity
10× every ~2.5 years

Moore's Law
10x every 12 years!

Algorithm, Architecture, Design are key!

Efficiency through Heterogeneity: Multi-Specialization

Brain-inspired: Multiple areas, different structure different function!



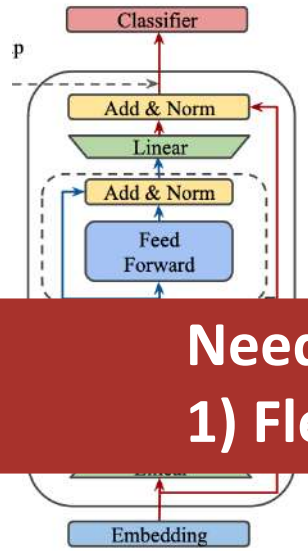
Multi-sensor
frame-based
event-based

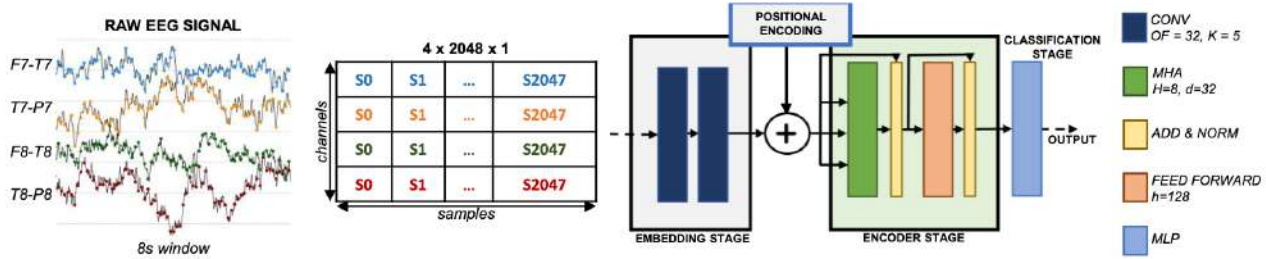
→ **Perception** → **Fusion**
Reasoning

Perceptive & Generative AI: A Fast-Evolving Model Zoo

[Z. Sun et al.] MobileBERT

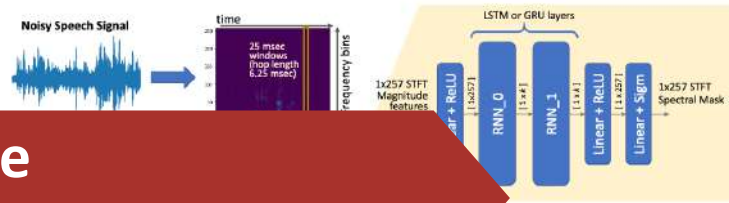
Encoder Transformer





[P. Busia et al.] EEGFormer

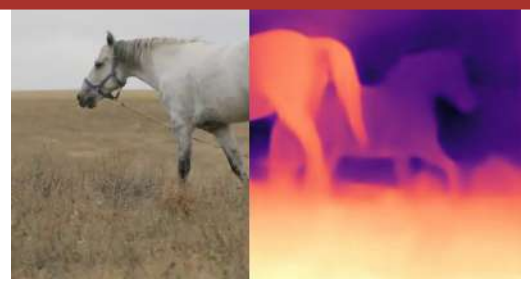
Encoder Transformer



Need to design embedded AI SoC's that provide
1) Flexibility 2) Performance 3) Energy Efficiency

yDenoiser

Recurrent NN



DINOv2: Learning Robust Visual Features without Supervision

[M. Oquab et al.]

Encoder Transformer

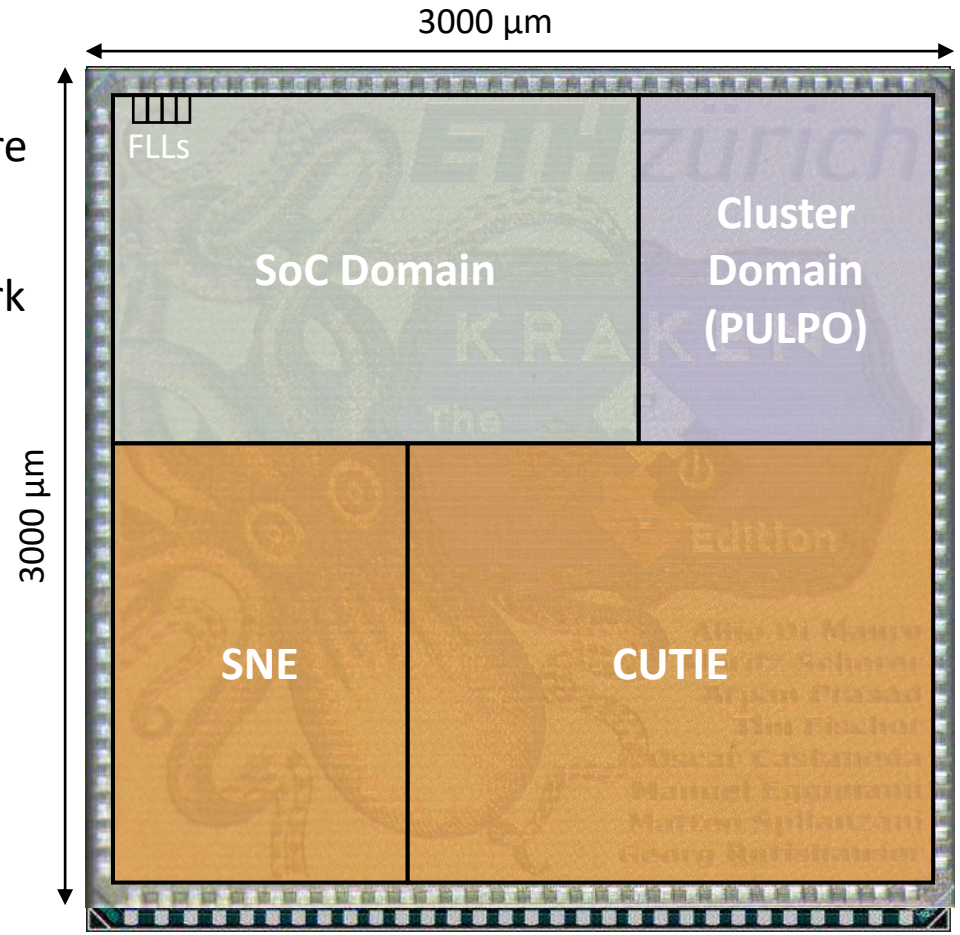
Auto-regressive Transformers?

Maybe...

Kraken: 22FDX SoC, Multiple Heterogeneous Accelerators

The *Kraken*: an "Extreme Edge" Brain

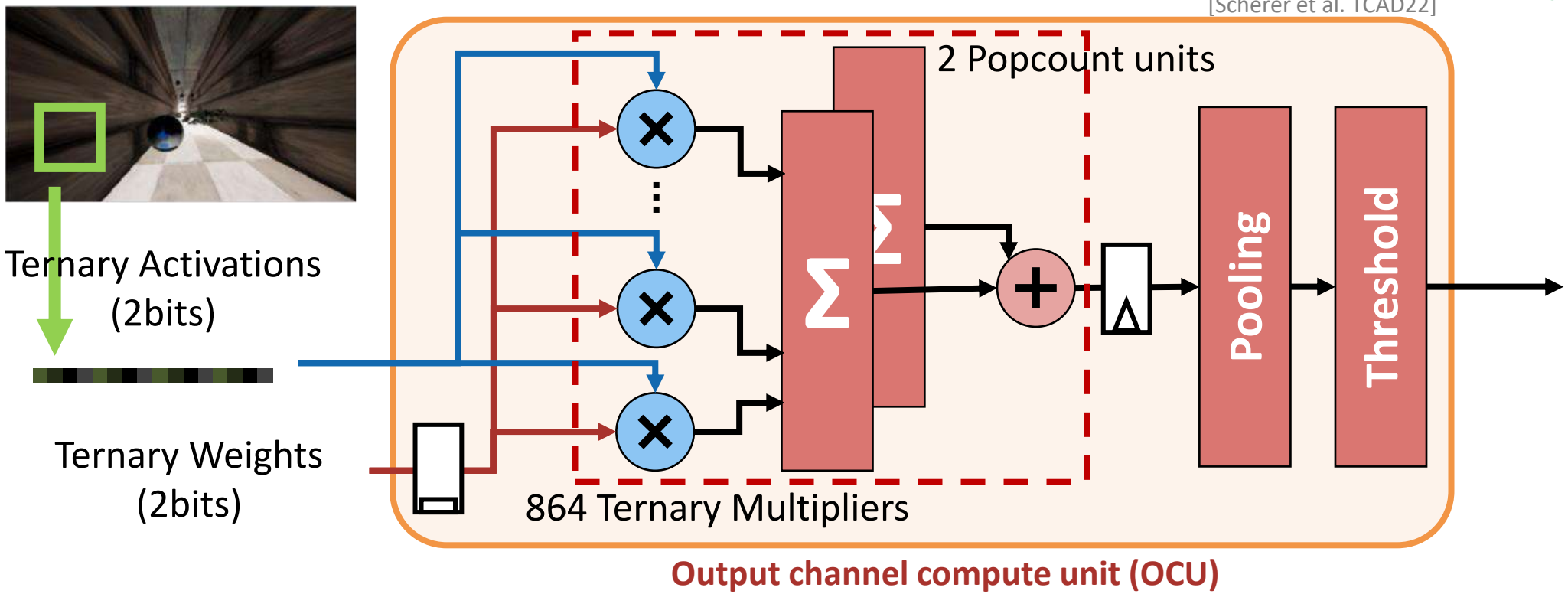
- **RISC-V Cluster**
  8 Compute cores +1 DMA core
- **CUTIE**
  Dense ternary-neural-network accelerator
- **SNE**
  Energy-proportional spiking-neural-network accelerator



Technology	22 nm FDSOI
Chip Area	9 mm^2
SRAM SoC	1 MiB
SRAM Cluster	128 KiB
VDD range	0.55 V - **0.8 V**
Cluster Freq	**~370 MHz**
SNE Freq	**~250 MHz**
CUTIE Freq	**~140 MHz**

CUTIE: Perception from Frame Sensors

[Scherer et al. TCAD22]

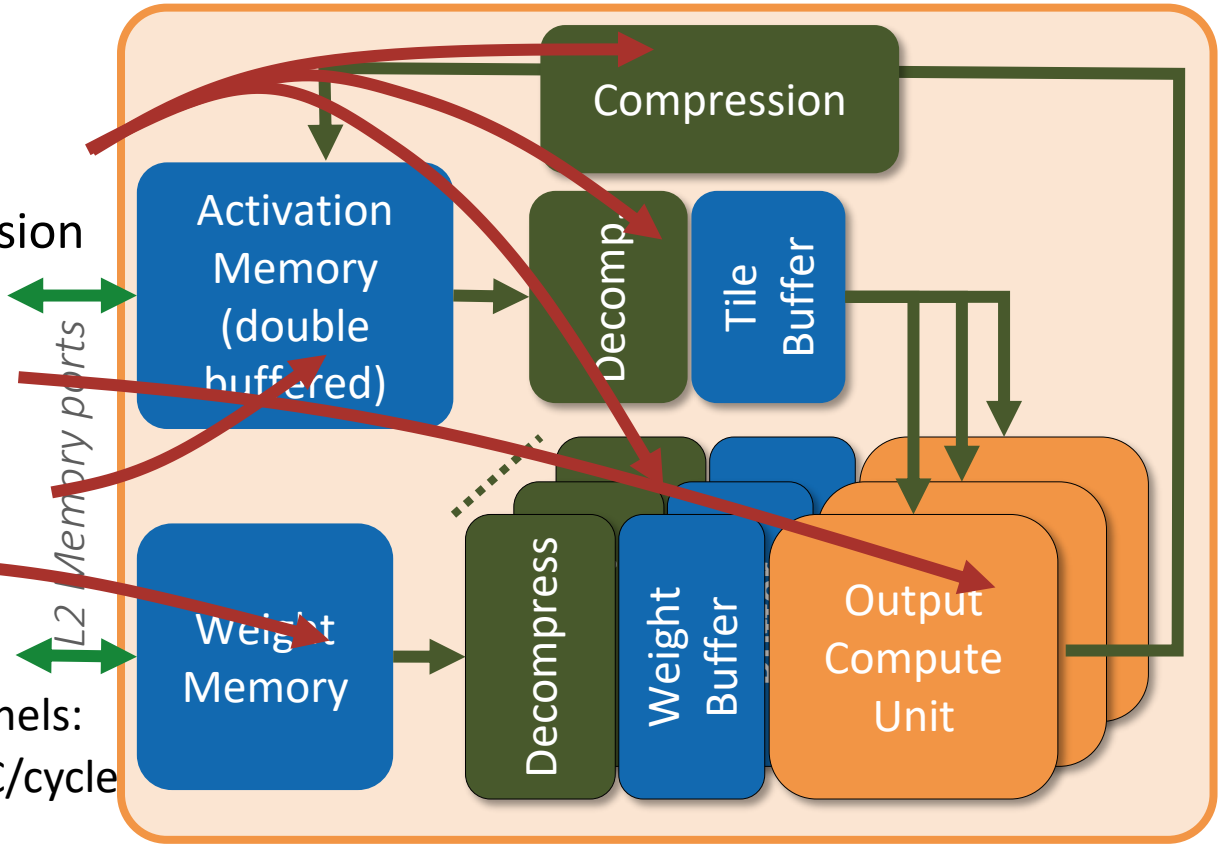


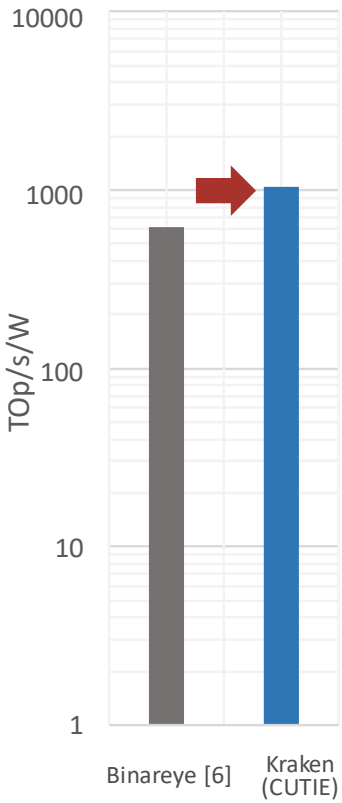
- **Completely Unrolled Ternary Neural Inference Engine**: K × K window, all input channels, cycle-by-cycle sliding
- One *Output Compute Unit* (OCU) computes one output activation per cycle!
- Zeros in weights and activations, spatial smoothness of activations reduce switching activity

Aggressive quantization and full specialization

Kraken`s CUTIE Implementation

- Data in 1.6 bits (Ternary value) with On-the-fly Compression/Decompression
- Configuration in Kraken
  - 96 channels (Output compute units)
  - 3 × 3 kernels
  - 64 × 64 pixels feature maps (158 KiB)
  - 9 layers of weights (117 KiB)
- Lots of TMAC/cycle
  - 96 OCUs, 96 Input channels, 3 × 3 kernels:
  - 96 × 96 × 3 × 3 = 82'944 Ternary-MAC/cycle

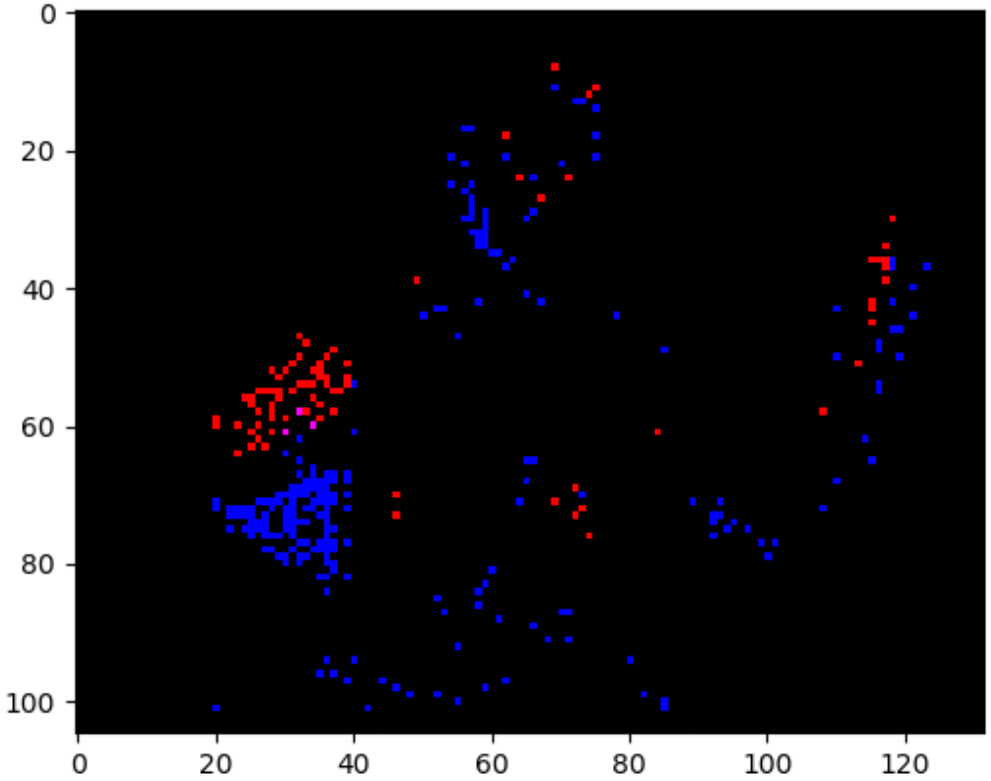




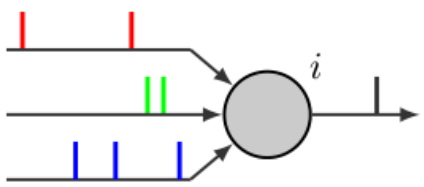
1fJ/MAC (1POP/s/W)
Ternary OPS

SNE: Perception on Event Sensors

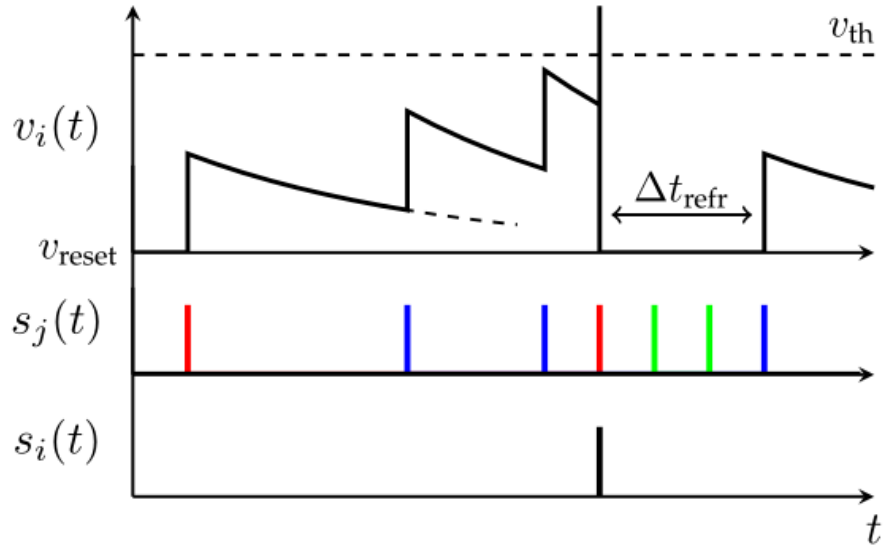
Event Sensors – DVS camera
Ultra-low latency
Energy- proportional interface



Spiking Neural Engine (SNE)



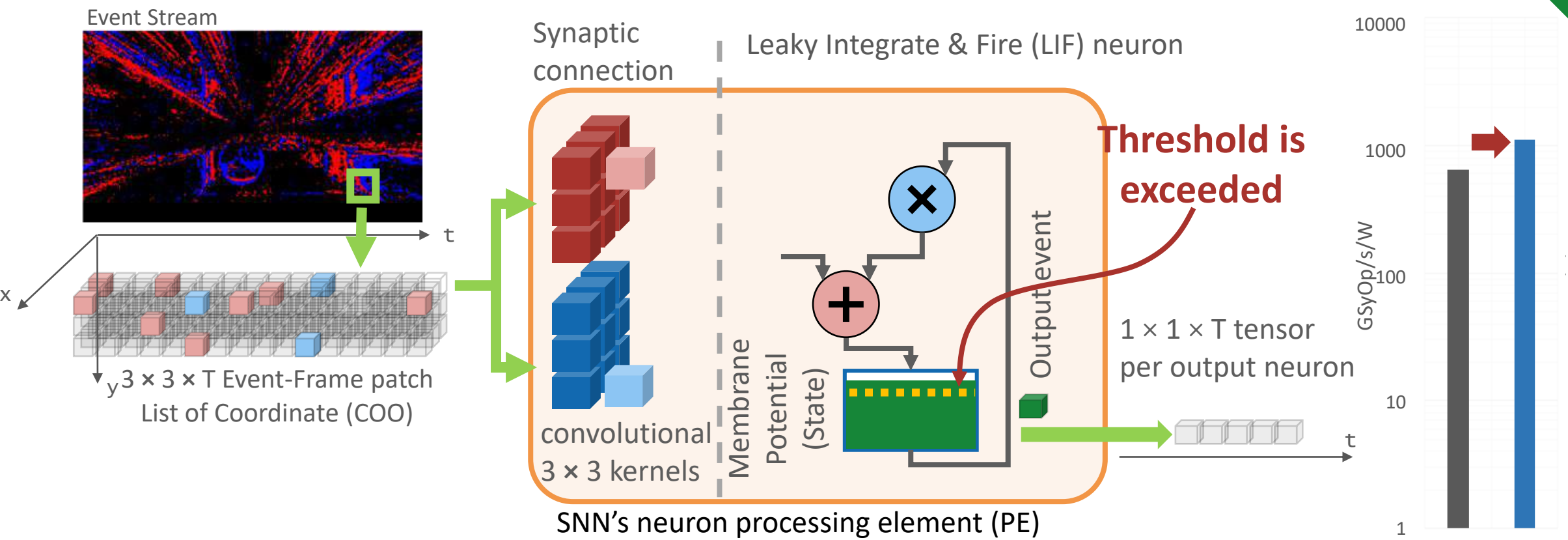
Leaky Integrate & Fire (LIF) neurons



[Di Mauro et al. DATE22]

SNE works seamlessly with DVS (event-based) sensors

Event consumption, and output spikes generation



A more complex dynamic than conventional DNNs neurons:

- Membrane Potential Accumulation/Activation 1× SynAcc = 1× 4b-ADD + 1× 8b-COMPARE
- Membrane Potential decay 1× SynDec = (1× 8b-MUL) + (1× 8b-MUL + 1× 8b-ADD)

1TSyOp/s/W

General Purpose: Domain-Specialized RV32 Core (PE)

RISC-V® Instruction set: open and extensible by construction (great!)

8-bit Convolution

Vanilla

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
```

RISC-V core

Specialized for AI → Mixed precision SIMD (16-2bit)

N/4

```
Init NN-RF (outside of the loop)
lp.setup
pv.nnsdotup.h  s0,ax1,9
pv.nnsdotsp.b  s1, aw2, 0
pv.nnsdotsp.b  s2, aw4, 2
pv.nnsdotsp.b  s3, aw3, 4
pv.nnsdotsp.b  s4, ax1, 14
end
```

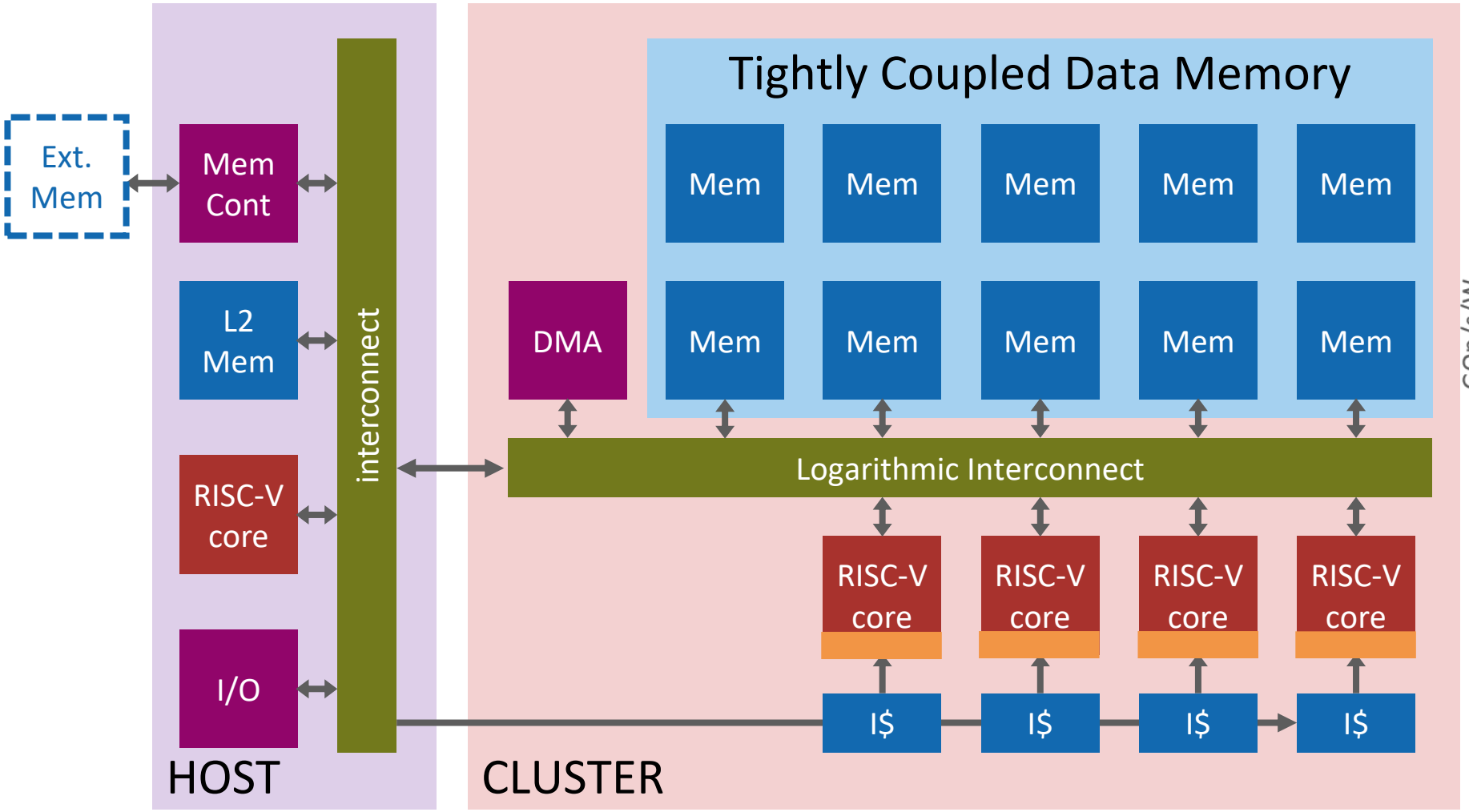
RISC-V core

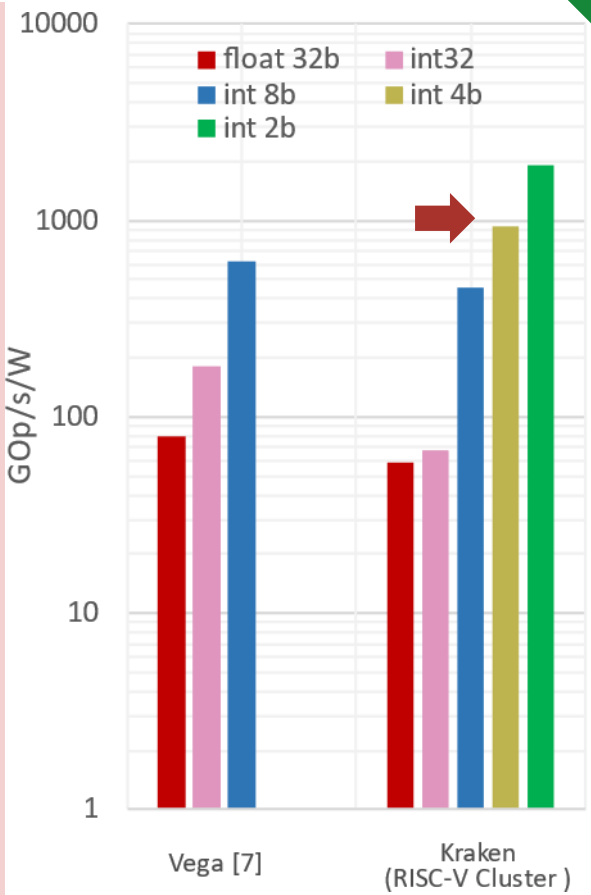
15x less instructions than Vanilla

90%+ ALU Utilization

Specialization Cost: Power, Area: 1.5×↑ Time 15×↓ → E = PT 10× ↓

PULP Paradigm: A PE cluster accelerates a host system





1TOP/s/W 2b/4b OPS

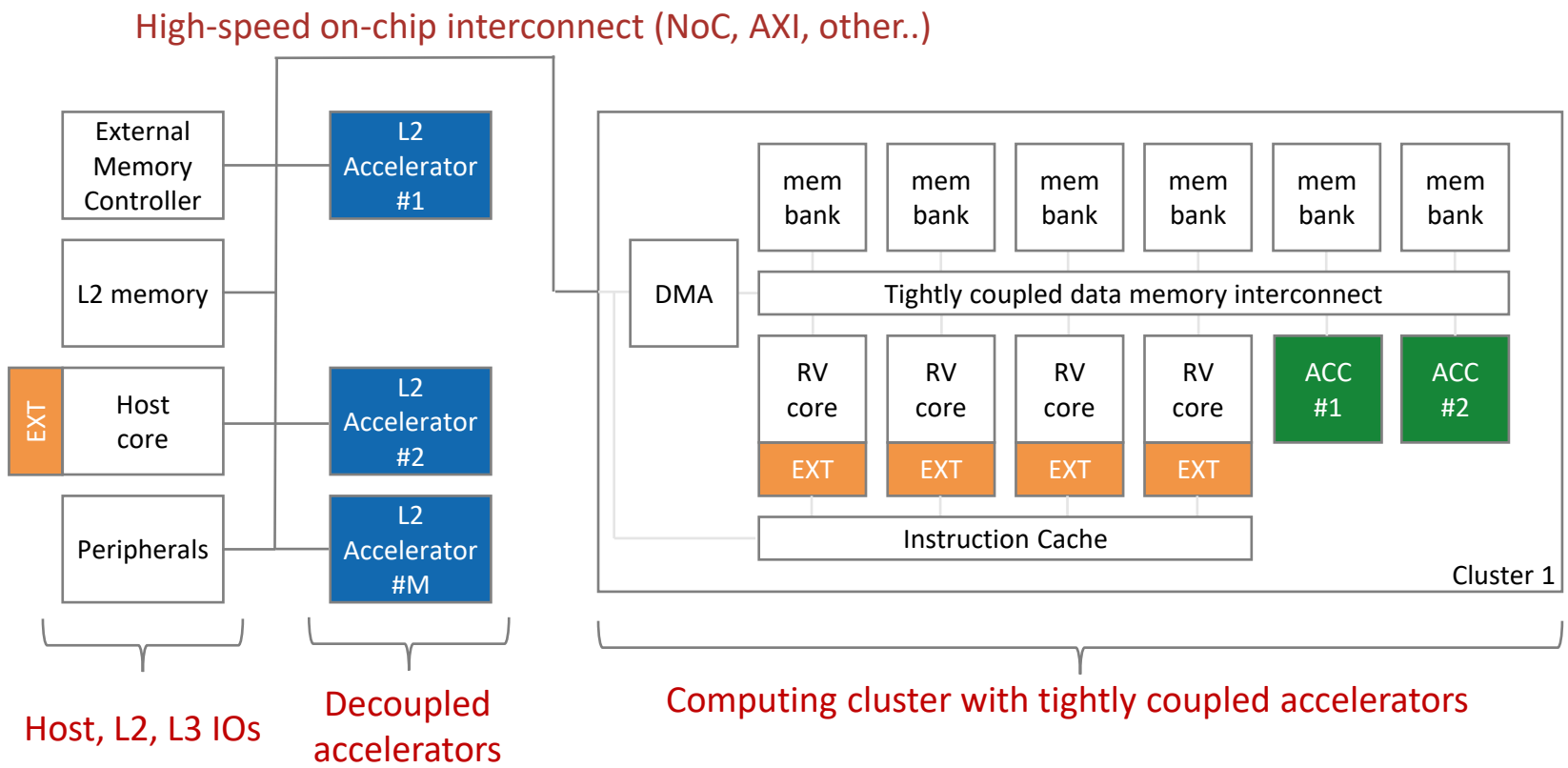
Heterogeneous, Multiscale Accelerated Computing

Multiple Scales of acceleration

Extensions to processor cores
- Explore new extensions
- Efficient implementations

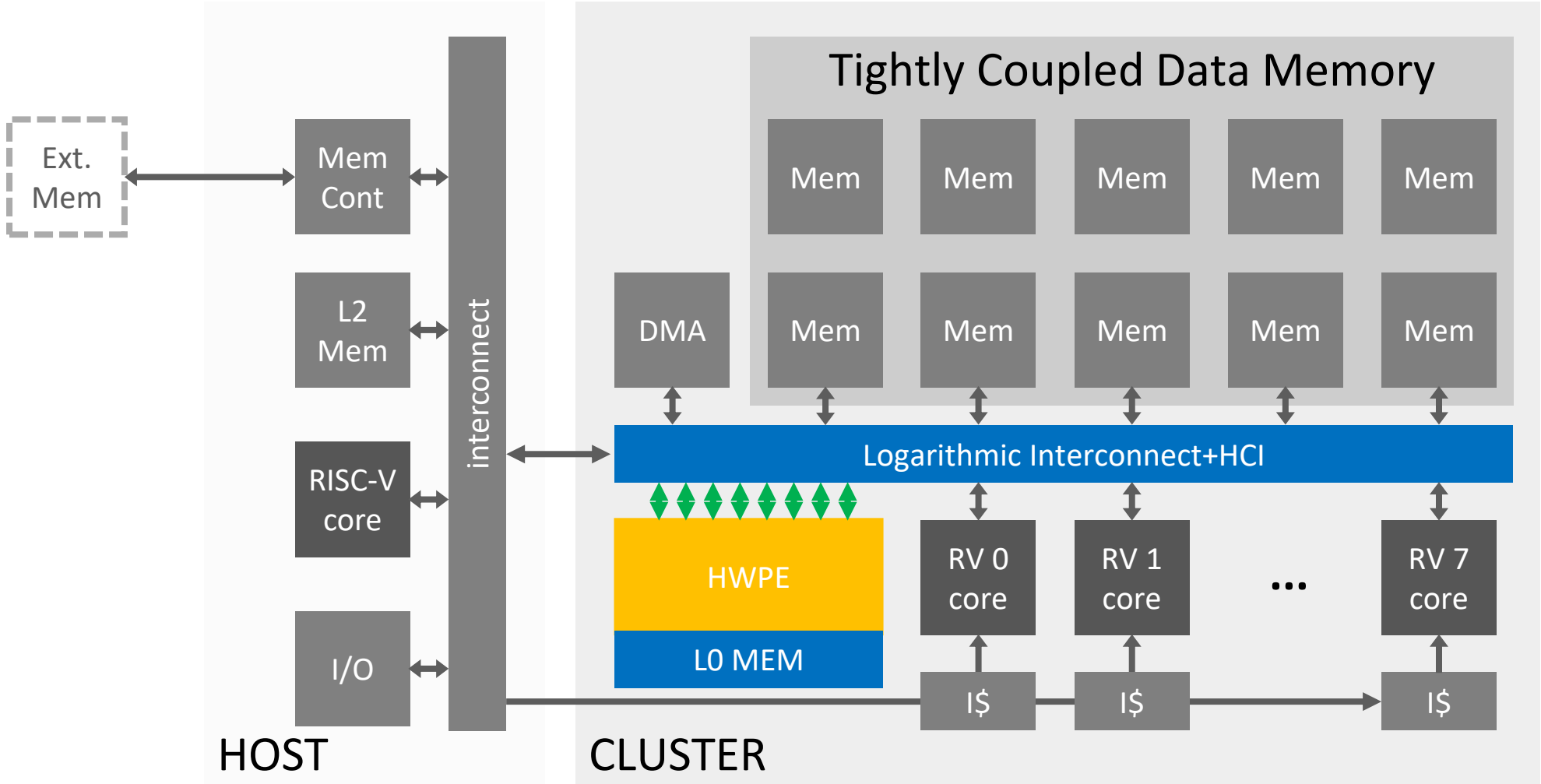
Shared-memory Accelerators
- Domain specific
- Local memory

Multiple Decoupled Accelerators
- Communication
- Synchronization



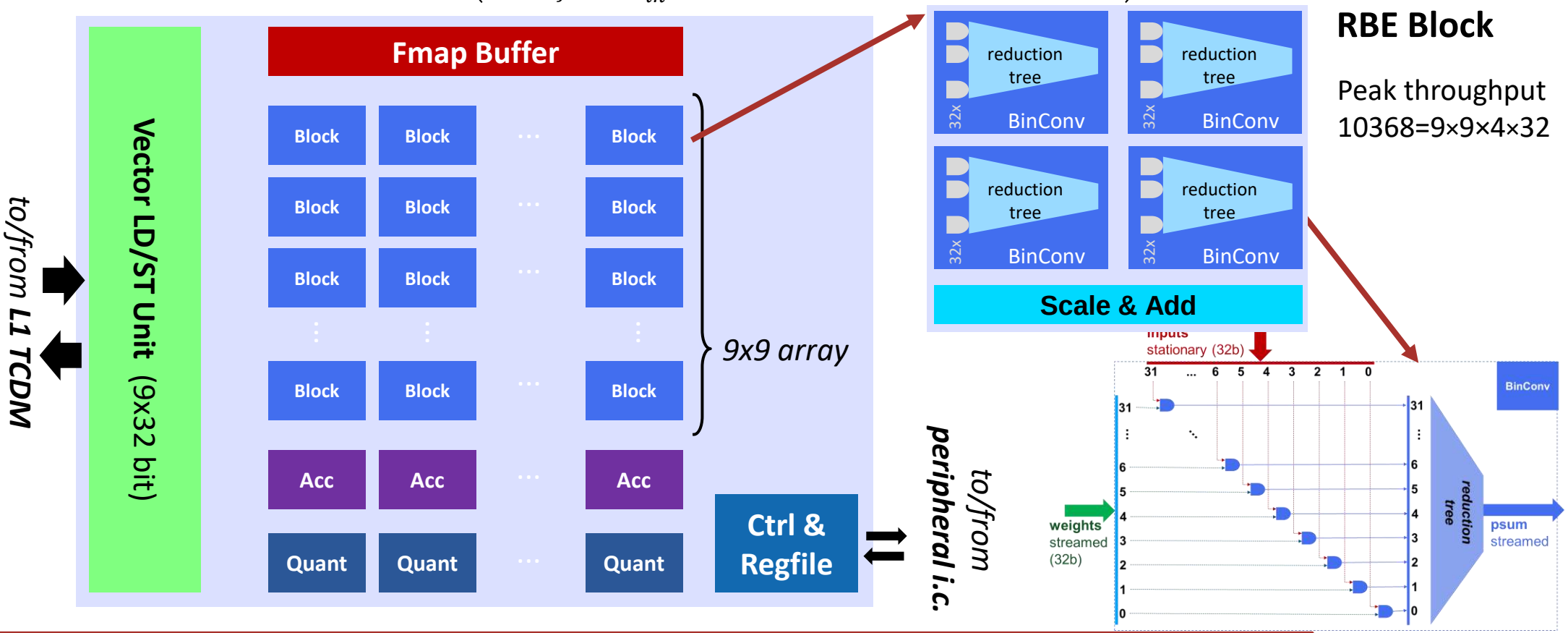
RISC-V is a key enabler → max agility, enabling SW build-up, without vendor lock-in

Tightly-coupled Accelerators



HWPE: Reconfigurable Binary Engine

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



Energy efficiency 10-20× (0.1pJ/OP) w.r.t. SW on cluster @same accuracy

Specialization in perspective

Using 22FDX tech, NT@0.6V, High utilization, minimal IO & overhead

Energy-Efficient RV Core → **20pJ (8bit)**

↓

ISA-based 10-20x →**1pJ (4bit)** → **XPULP**

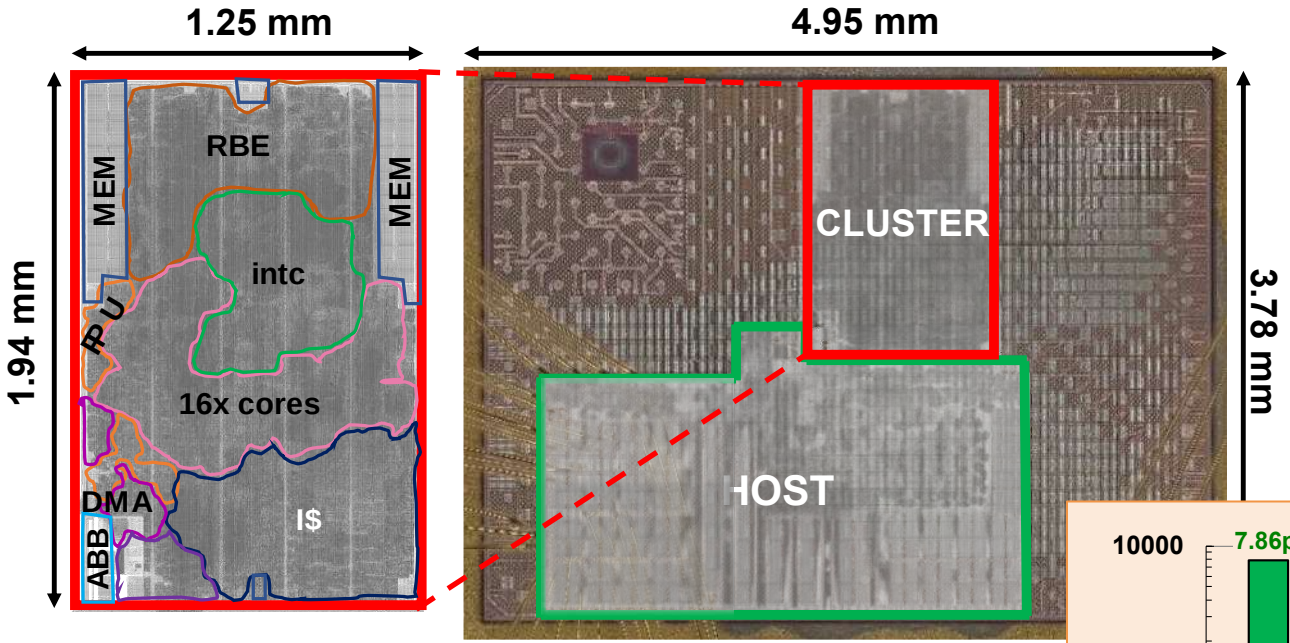
↓

Configurable DP 10-20x → **100fJ (4bit)** → **RBE**

↓

Highly specialized DP 100x →**1fJ (ternary)** → **CUTIE, SNN**

Marsellus: AI-IoT Heterogeneous SoC

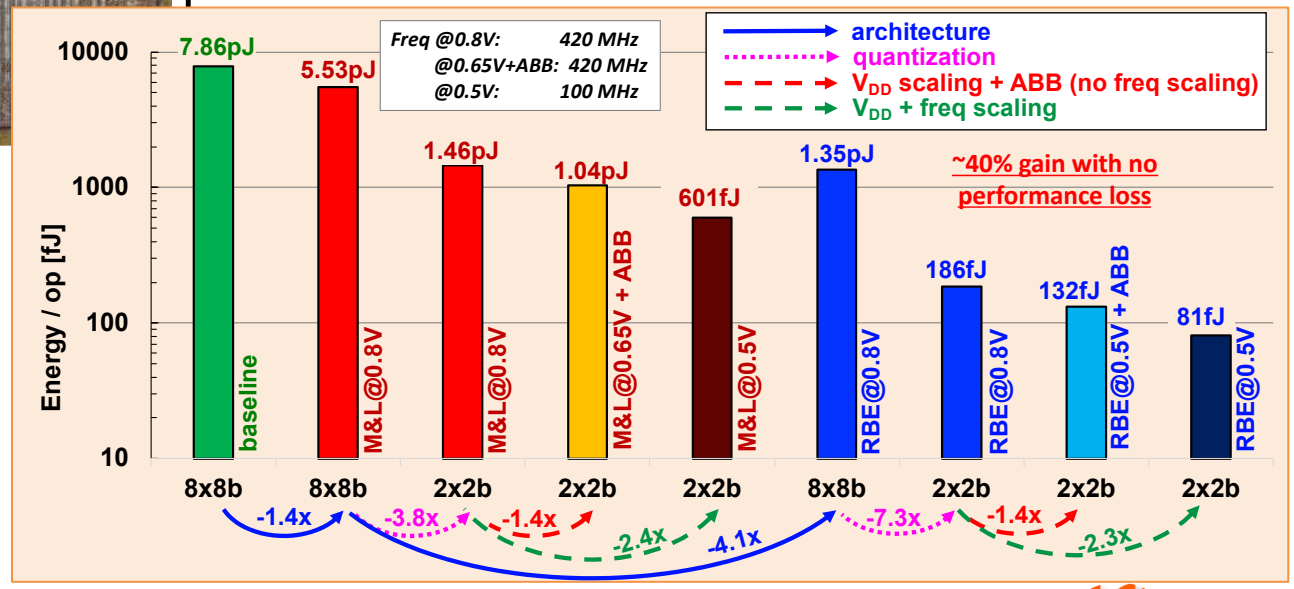


Combine:

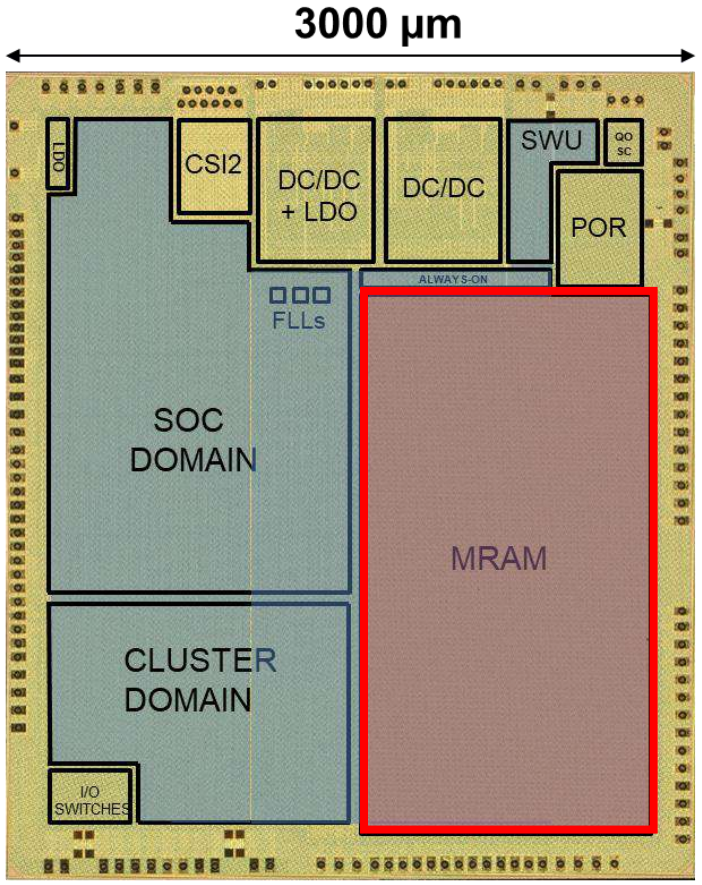
- Heterogeneous architecture
- Quantization
- V_{DD} scaling
- Adaptive Body Biasing

Prototype implemented in GF 22FDX

→ flip-well LVT & SLVT cells, 2.43mm² for CLUSTER

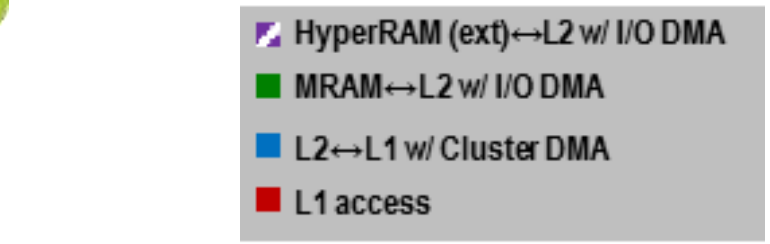


Vega: On-Chip NVMem for NN Weights

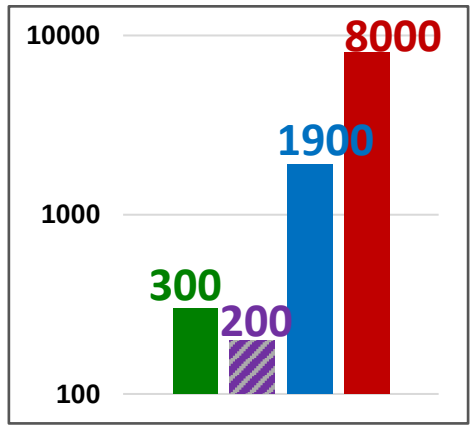


In cooperation with GREENWAVES TECHNOLOGIES

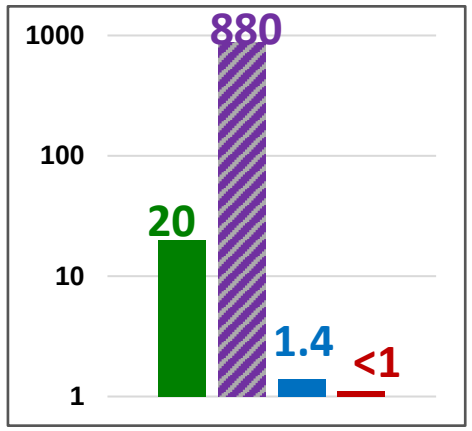
Technology	22nm FDSOI
Chip Area	12mm^2
SRAM	1.7 MB
MRAM	4 MB
VDD range	0.5V - 0.8V
VBB range	0V - 1.1V
Fr. Range	32 kHz - 450 MHz
Pow. Range	1.7 µW - 49.4 mW



Bandwidth [MB/s]

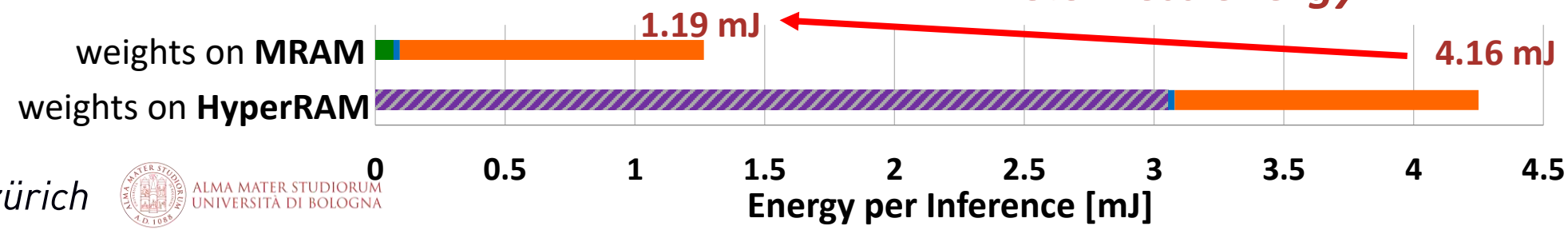


Energy per byte [pJ/B]



end-to-end on-chip computation

3.5x less energy



ETH zürich

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Not only academia: **GAP9** with **NE16**

GREENWAVES TECHNOLOGIES

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

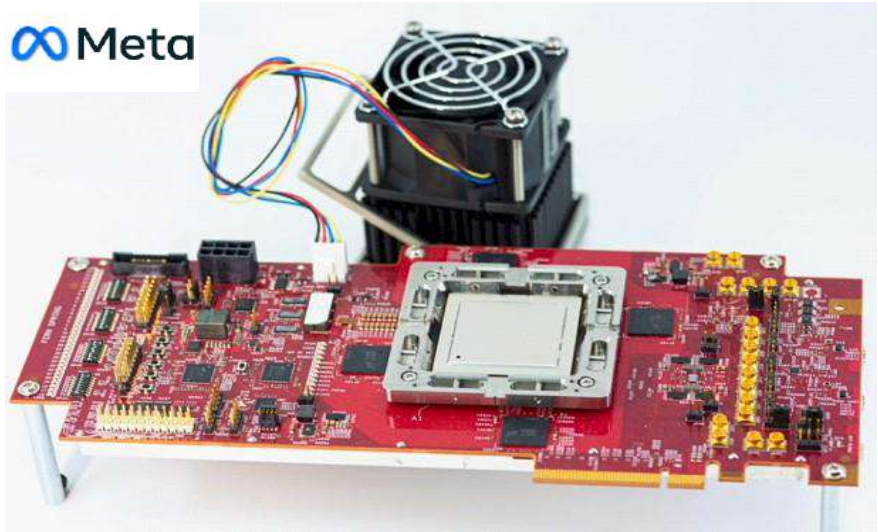
Submitter	Board Name	SoC Name	Processor(s) & Number	Accelerator(s) & Number	Software	Notes	Benchmark Results								
							Task	Visual Wake Words		Image Classification		Keyword Spotting		Anomaly Deteciton	
							Data	Visual Wake Words Dataset		CIFAR-10		Google Speech Commands		ToyADMOS (ToyCar)	
							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	17131.6	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									0.098	

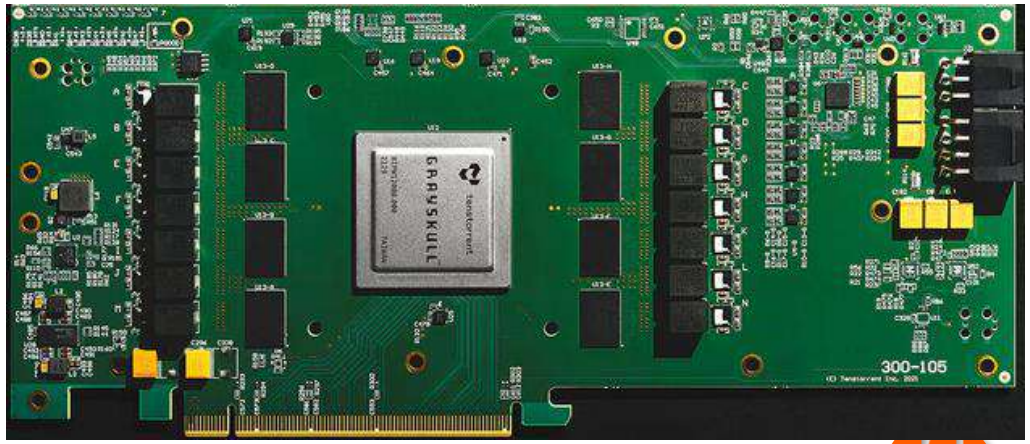
AI Innovation beyond “NVIDIA Gravity” is Challenging!

- **It’s the software → flexibility, fast evolution!**
- **Need an open standard to counter a monopoly**

RISC-V®

RISC-V: The Free and Open RISC Instruction Set Architecture





RISC-V is Accelerating

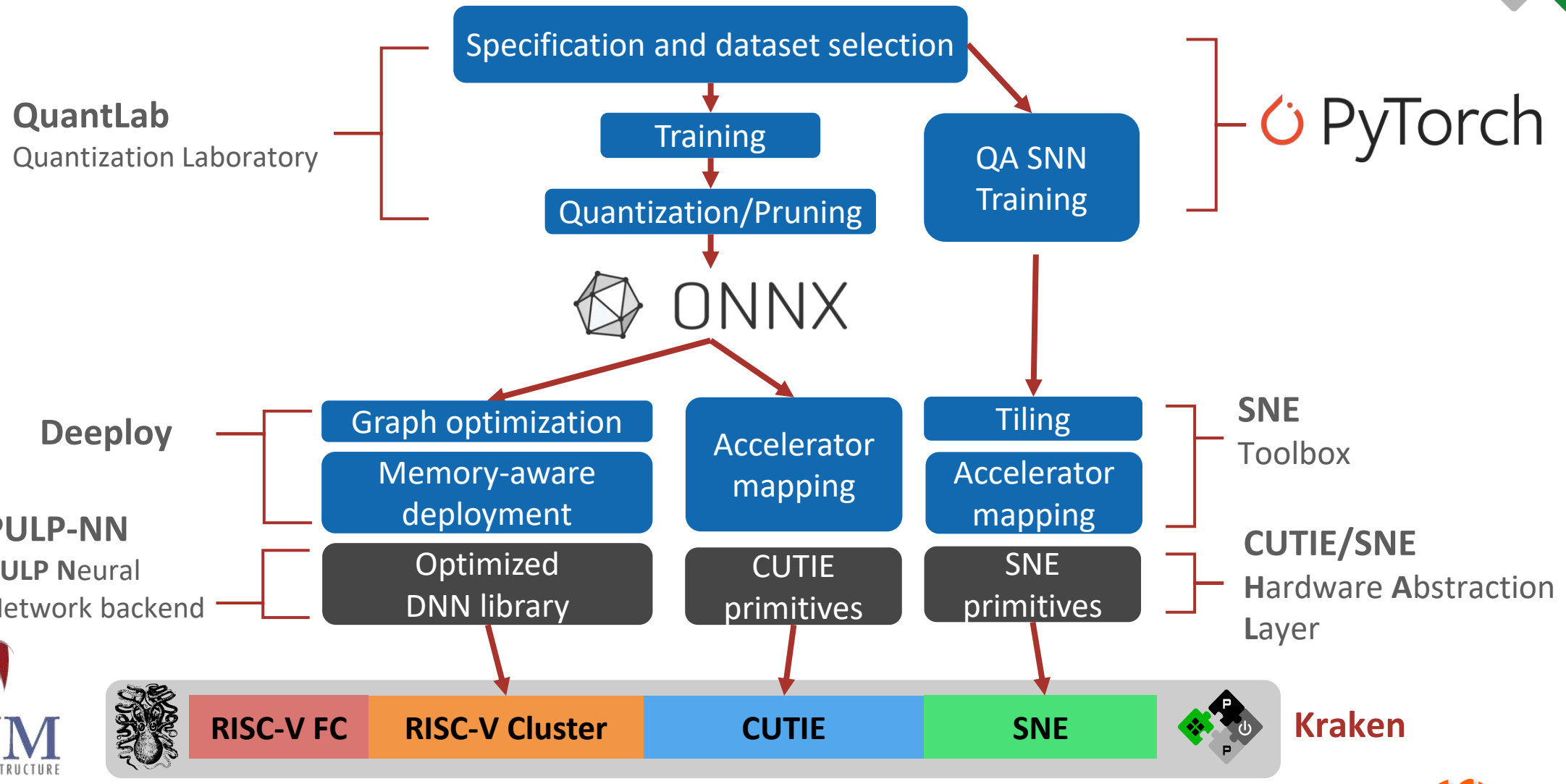
EuroHPC 200+M€ for RV HPC (DARE FPA)
Chips (KDT) 300+M€ for RV Automotive

India Ministry for Electronics & Information Technology launched Digital India RISC-V (DIR-V) program for commercial SHAKTI & VEGA silicon.

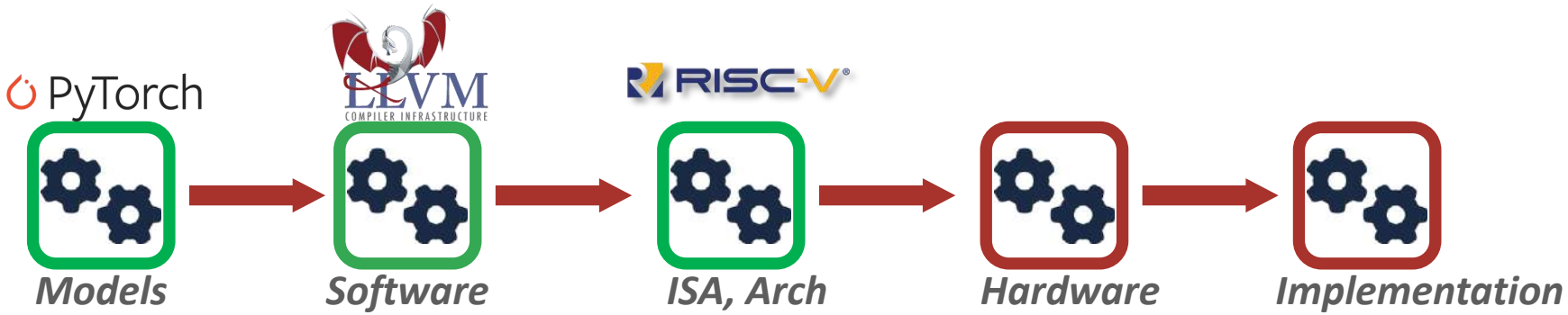
Industry Leaders Launch RISE to Accelerate the Development of Open Source Software for RISC-V

Six chip giants to drive **RISC-V application in automotive**, enhance industry resilience

Fully Open-Source Deployment Flow!



Open SW & HW Embodied AI Platform?



$$Cost = IP_{€} + EDA_{€} + Si_{€}$$

Curtailing IP€: Open-Source Hardware

RISC-V Cores and Vector Units

RI5CY *CV32E*	Zero R *Ibex*	Snitch	Spatz	Ariane *CVA6*	ARA
RV32	RV32	RV32	RVV	RV64	RVV

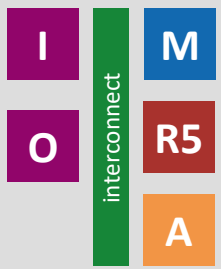
Peripherals

JTAG, SPI, UART, I2S, DMA, GPIO

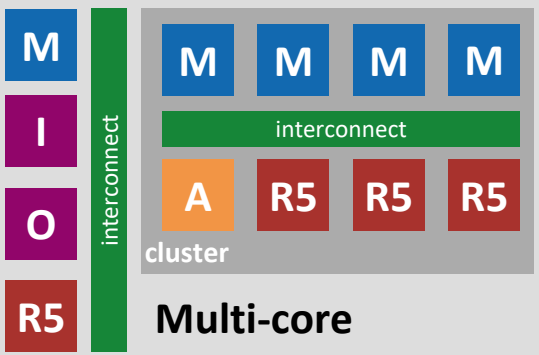
Interconnects

LIC, HCI, APB, FlooNoC, AXI4

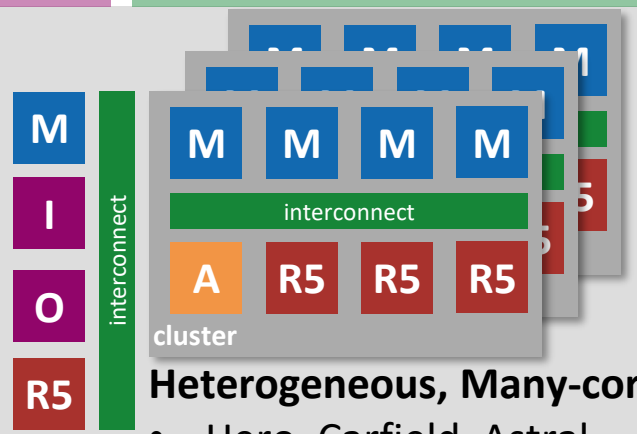
Platforms



Single core
- PULPino, PULPissimo
- Cheshire



Multi-core
- OpenPULP
- ControlPULP



Heterogeneous, Many-core
- Hero, Carfield, Astral
- Occamy, Mempool

IOT → **HPC**

Accelerators and ISA extensions

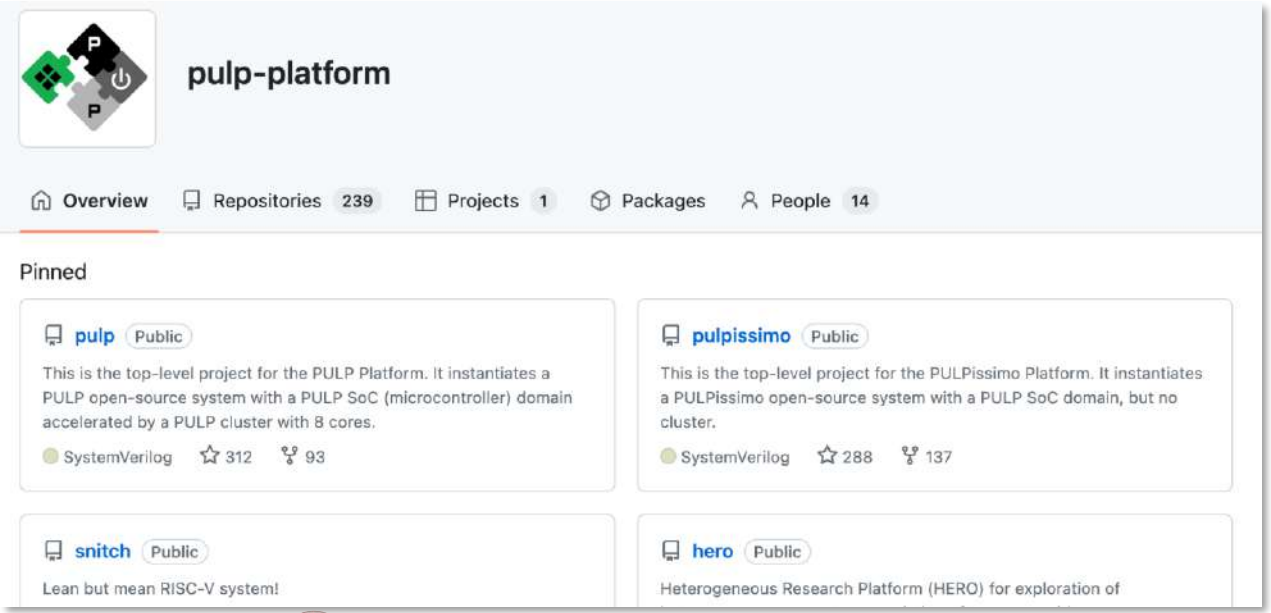
XpulpNN, XpulpTNN	ITA (Transformers)	RBE, NEUREKA (QNNs)	FFT (DSP)	REDMULE (FP-Tensor)

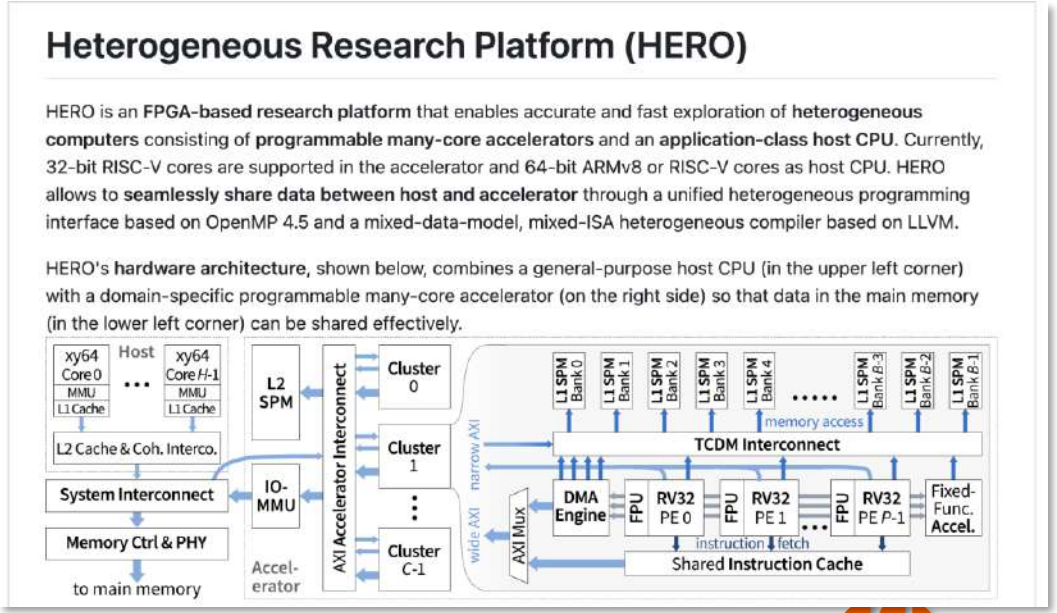
We make everything (we can) available openly

- **All our development is on GitHub using a permissive license**
  - HDL source code, testbenches, software development kit, virtual platform

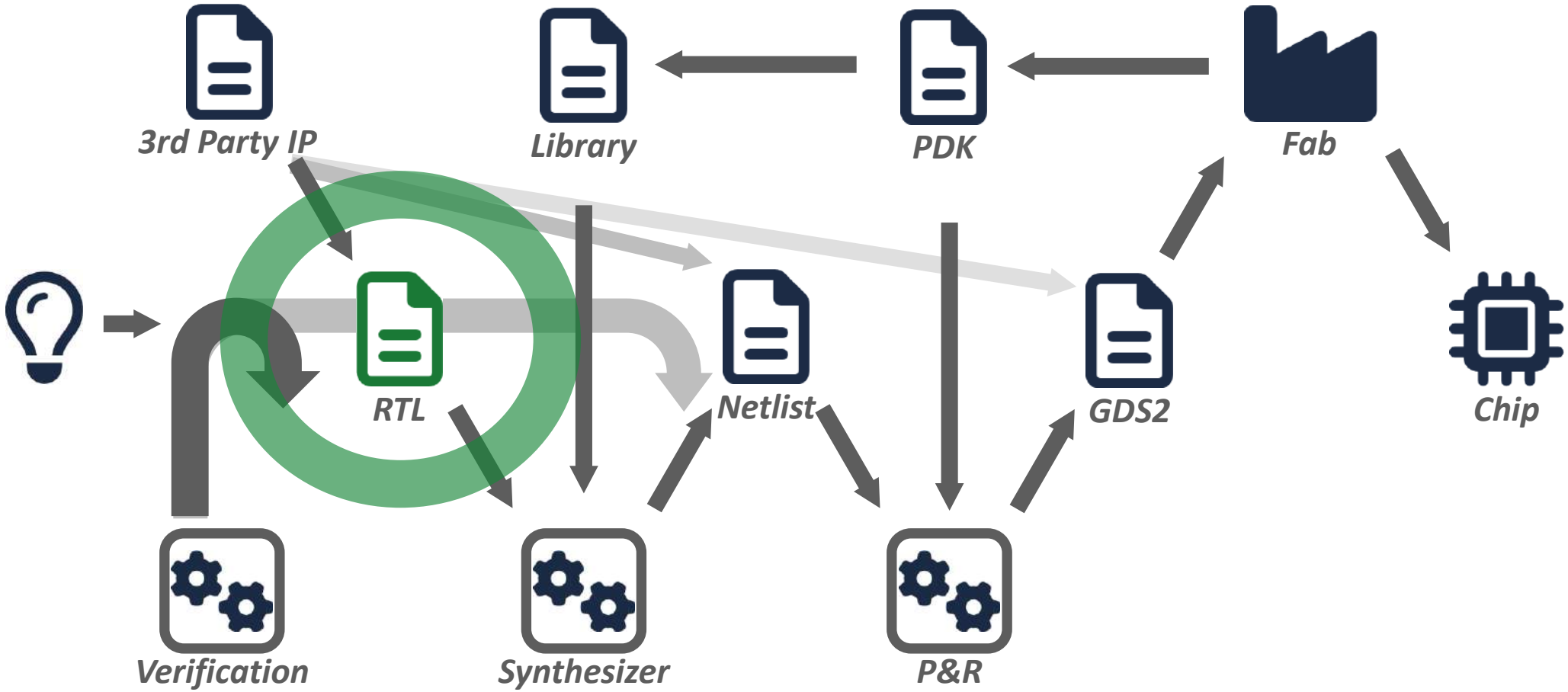
https://github.com/pulp-platform

- Allows anyone to use, change, and make products without restrictions.

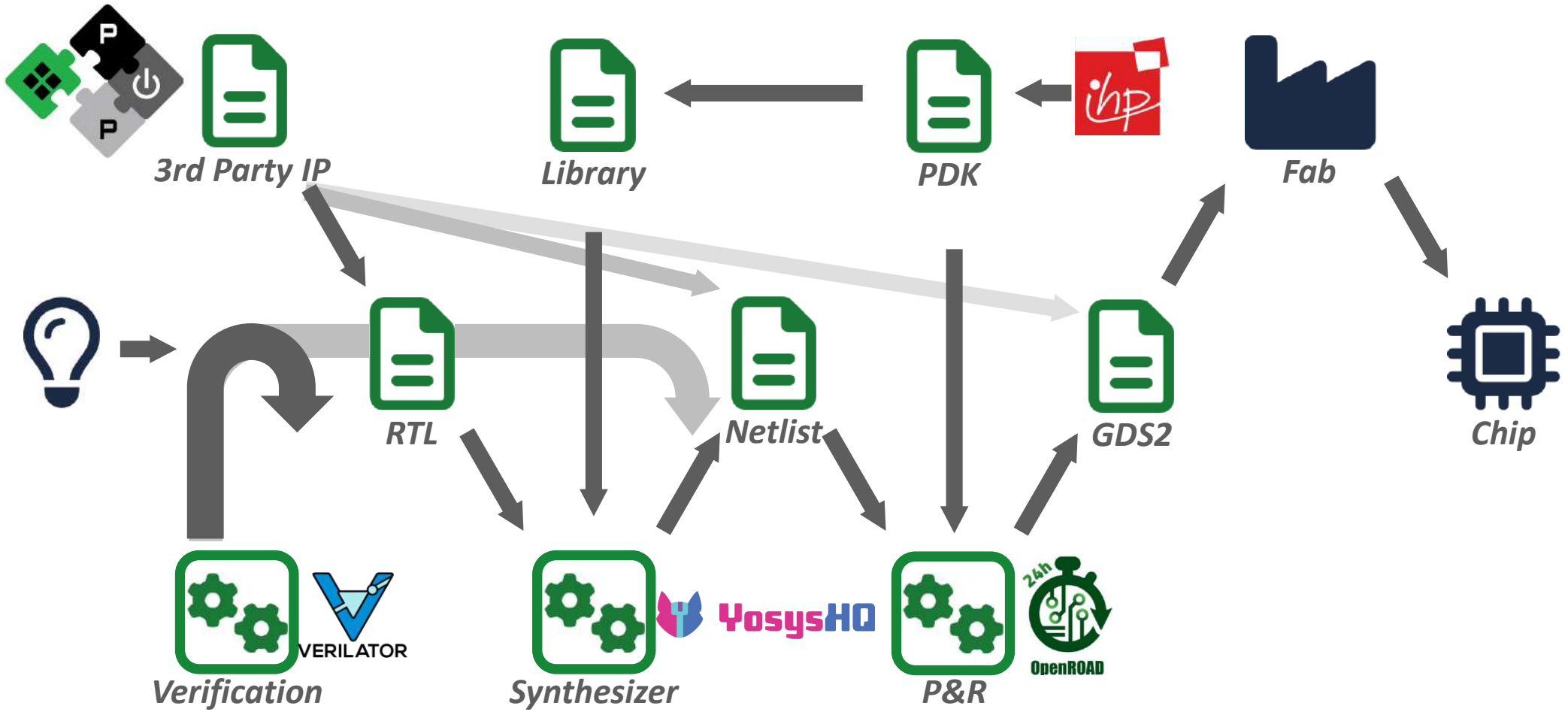




Curtailing EDA$_€$: Open-Source Implementation?



End-to-end Open-Source Digital IC Design is Possible Today!

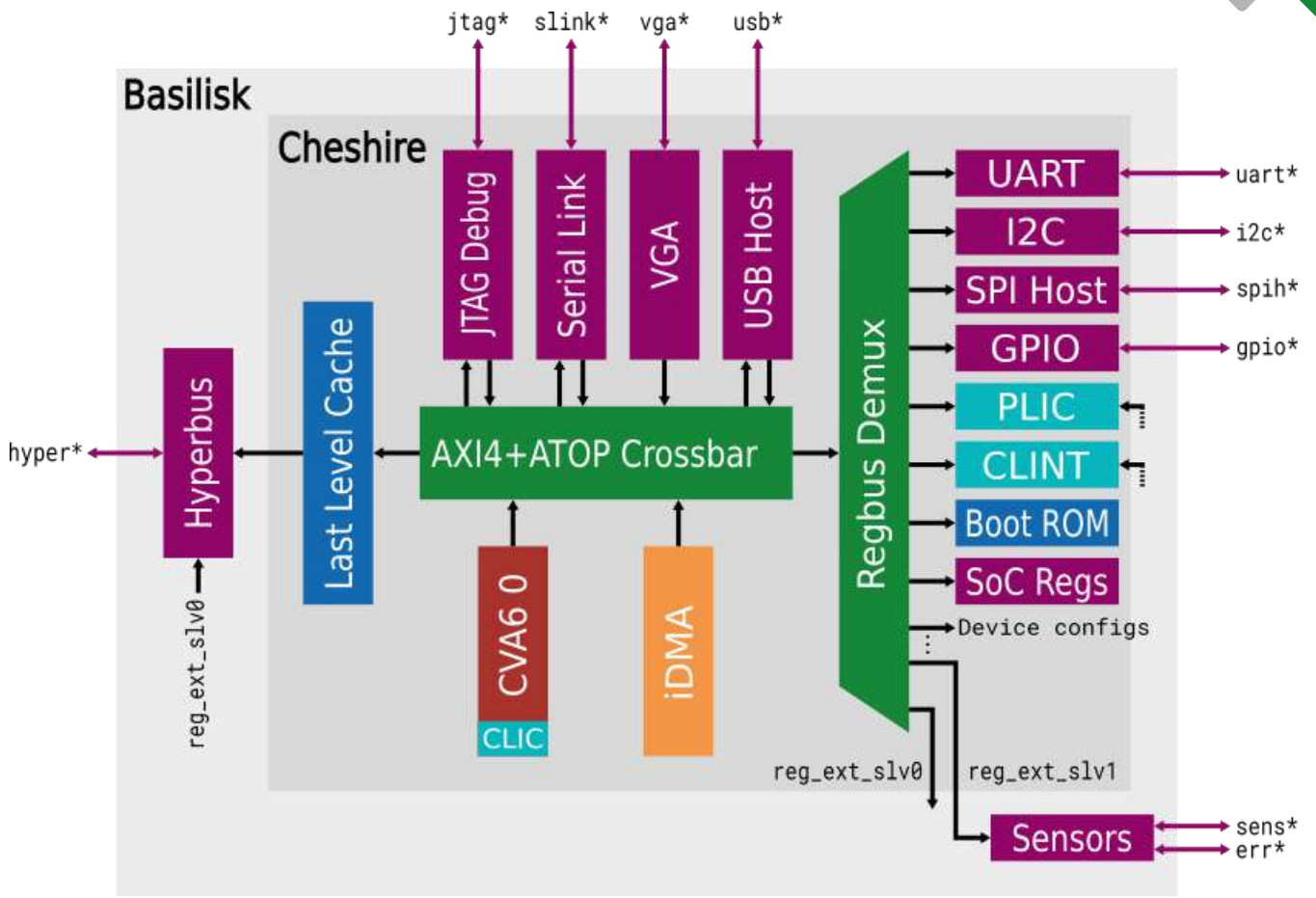


Basilisk: Open **RTL**, Open **EDA**, Open **PDK**

- **Designed in IHP 130nm OpenPDK**
  - 6.25mm x 5.50mm
  - 60MHz
    - 1.08 MGE logic, 60% density
    - 24 SRAM macros (114 KiB)
- **CVA6 based SoC**
  - Runs and boots Linux
- **Active collaboration with**

Basilisk SoC: Cheshire Platform

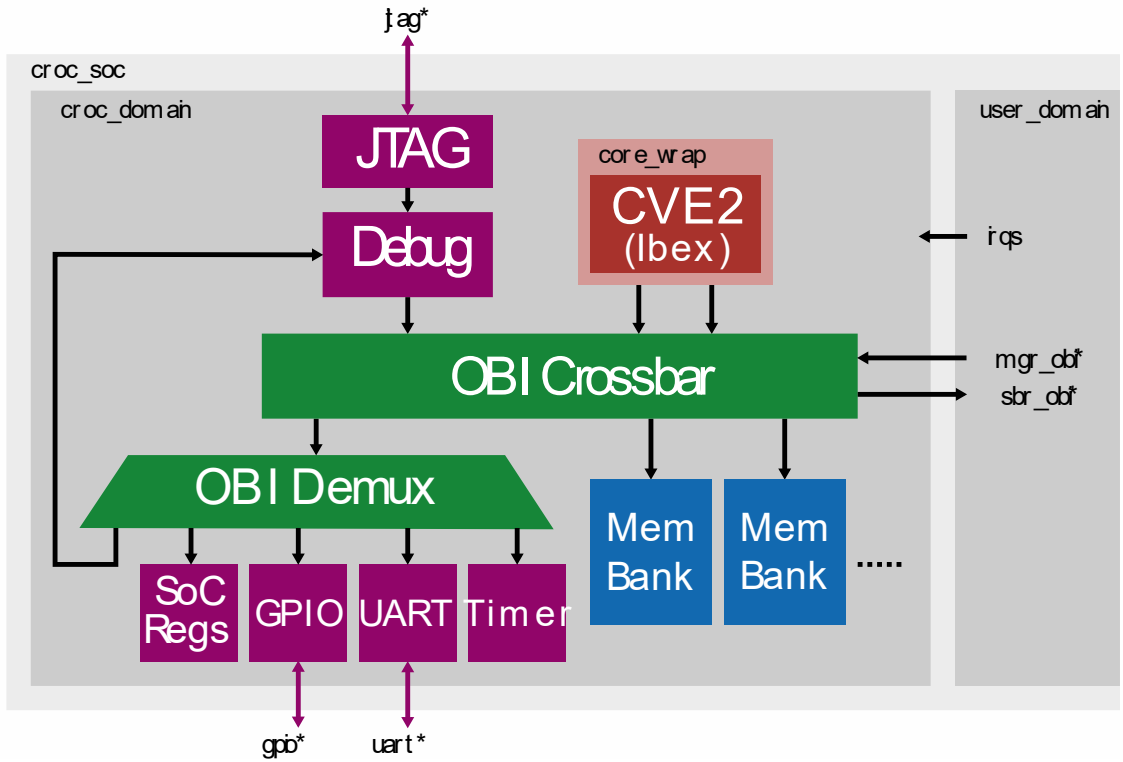
- **Multi-million gate design**
- **64-bit RISC-V Core**
  - Complete Linux-capable SoC
  - Simple “Raspberry Pi”
- **Rich Peripherals**
  - Includes an open USB 1.1 host
- **Open-source DRAM interface**
  - Digital-only interface
- **Silicon-proven**
  - Multiple tapeouts with commercial EDA

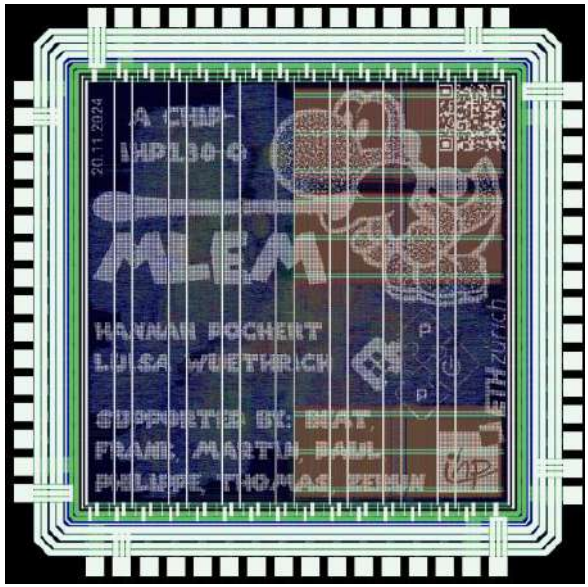


github.com/pulp-platform/cheshire-ihp130-o

Mlem is our 2nd end-to-end open SoC: Croc Platform

- **Scalable ULP design**
- **32-bit RISC-V Core**
  - Complete Linux-capable SoC
  - Simple “Raspberry Pi”
- **Rich Peripherals**
- **Ready for Acceleration**
  - Digital-only interface
- **Silicon-proven**
  - Tapeouts with open & commercial EDA

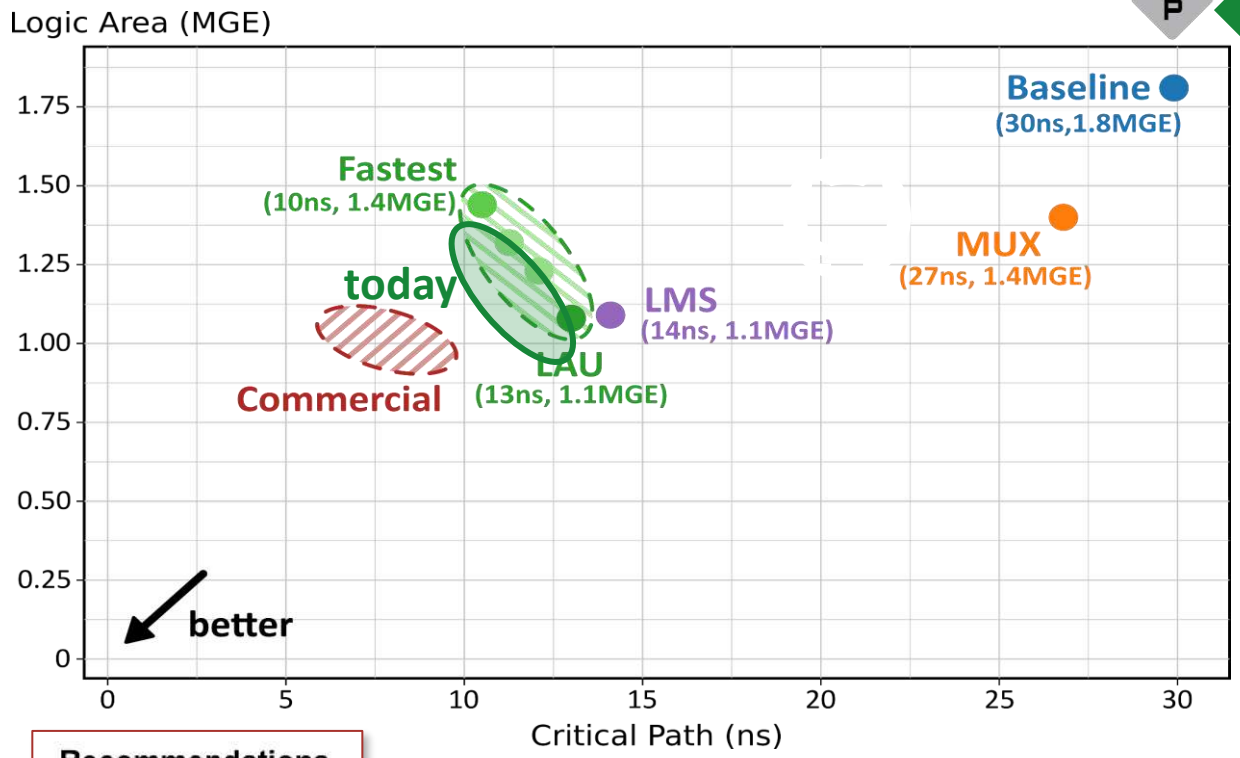




github.com/pulp-platform/croc

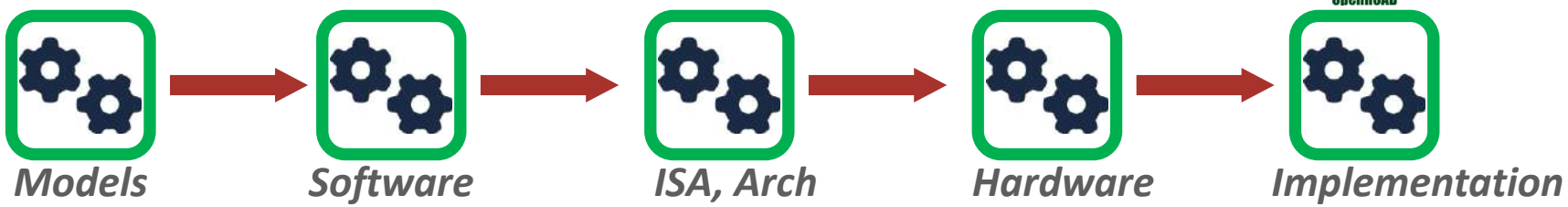
Open-source vs. Commercial EDA – Reality Check

- **Baseline (2023)→ June 2024**
  - SV-to-Verilog chain @ **<2min** runtime
  - Yosys synthesis:
    → **1.1 MGE** (**1.6×**) @ **77 MHz** (**2.3×**)
    → **1.4×** less runtime, **2.4×** less peak RAM
  - OpenROAD P&R: tuning
    → **-12%** die area, **+10%** core utilization
- **Improvements June-October**
  - Yosys-slang replaces SV2V
    - **1.6×** less runtime, **10×** less peak RAM
    - **-10%** logic area (preliminary)



Open EDA is maturing really fast!

Does it Make Sense for a Foundry?



Cost=IP$_€$+EDA$_€$+SI$_€$

Yes!

1. Silicon cost remains as the bottom line
2. Openness Facilitates Ecosystem build-up
3. Eases life-cycle (training, audit, certification, support)
4. Great to boost return (€) on a mature node
5. Hybrid models are always possible

But…

Need a mature node for **Energy-Efficient Digital (FDX22 ☺)**

Embodied AI is the Perfect Target Market for End-to-end Open Platforms!

pulp-platform.org

Thank You!

Thank you

youtube.com/pulp_platform

pulp-platform.org

@pulp_platform