ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

BRUGES
9-12 SEPTEMBER
ESSERC 2024

Circuits and Systems for Embodied AI

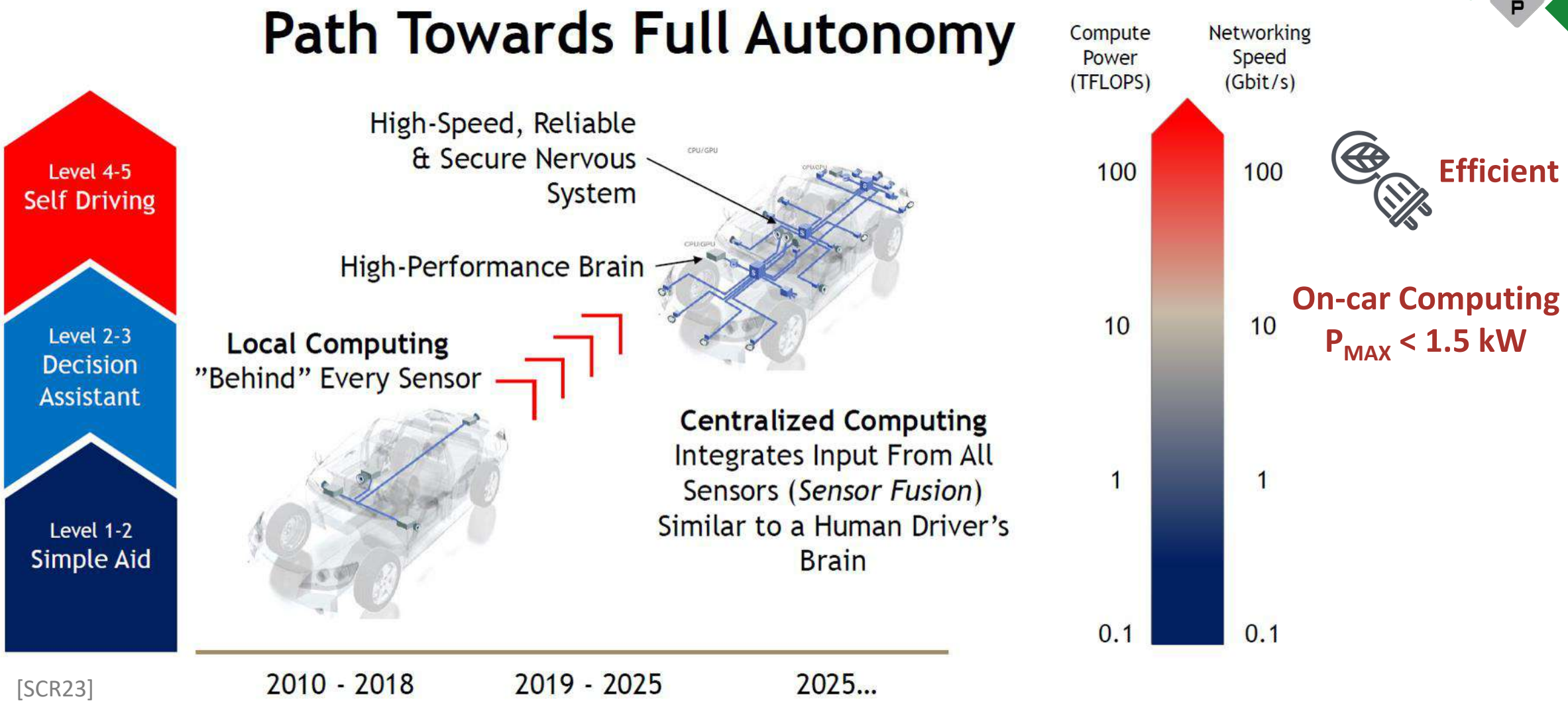
Exploring µJ Multi-Modal Perception for Nano-UAVs … and beyond

Luca Benini lbenini@iis.ee.ethz.ch
luca.benini@unibo.it

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

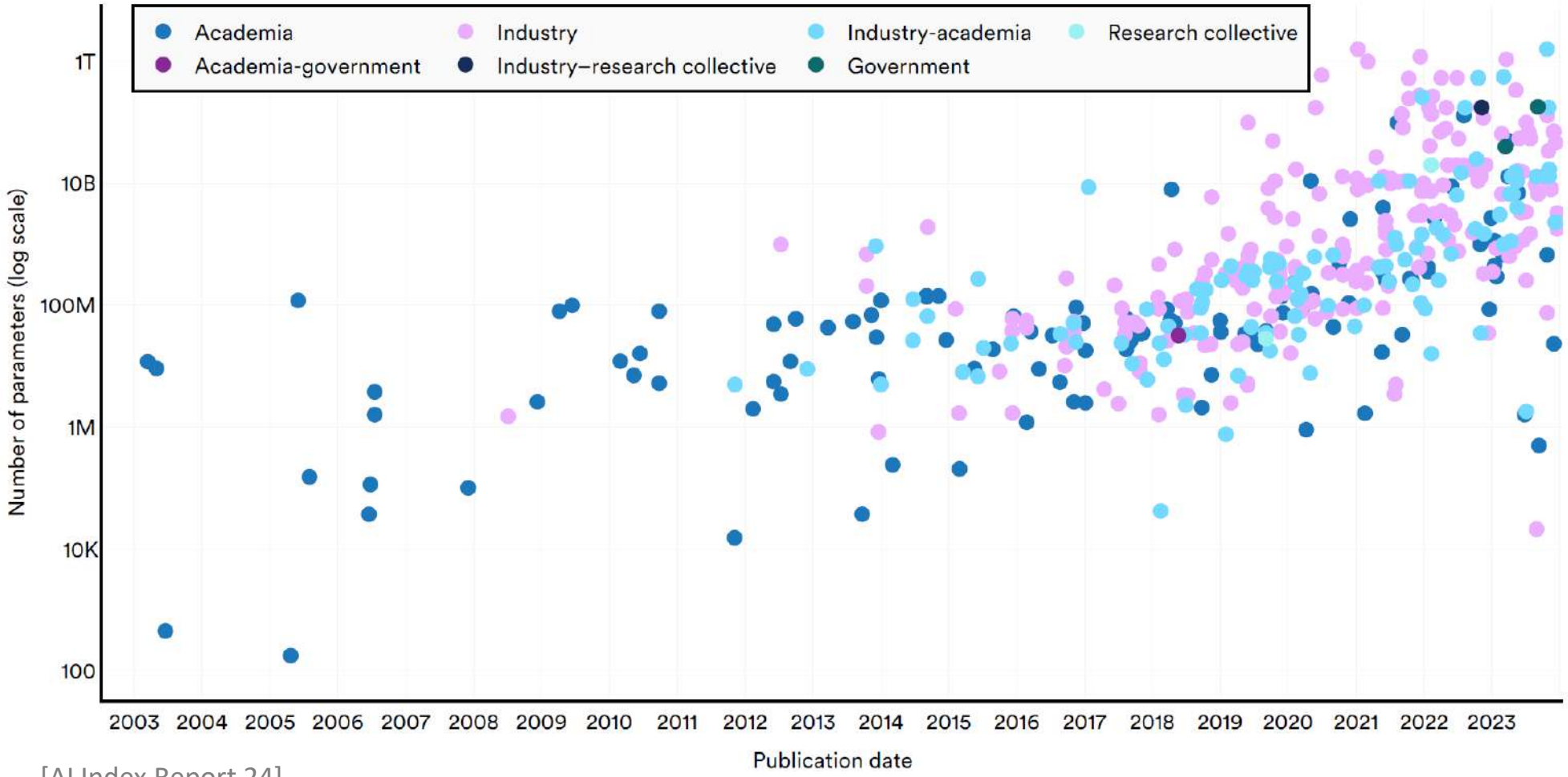
@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

Embodied AI



[SCR23]

Embodied AI



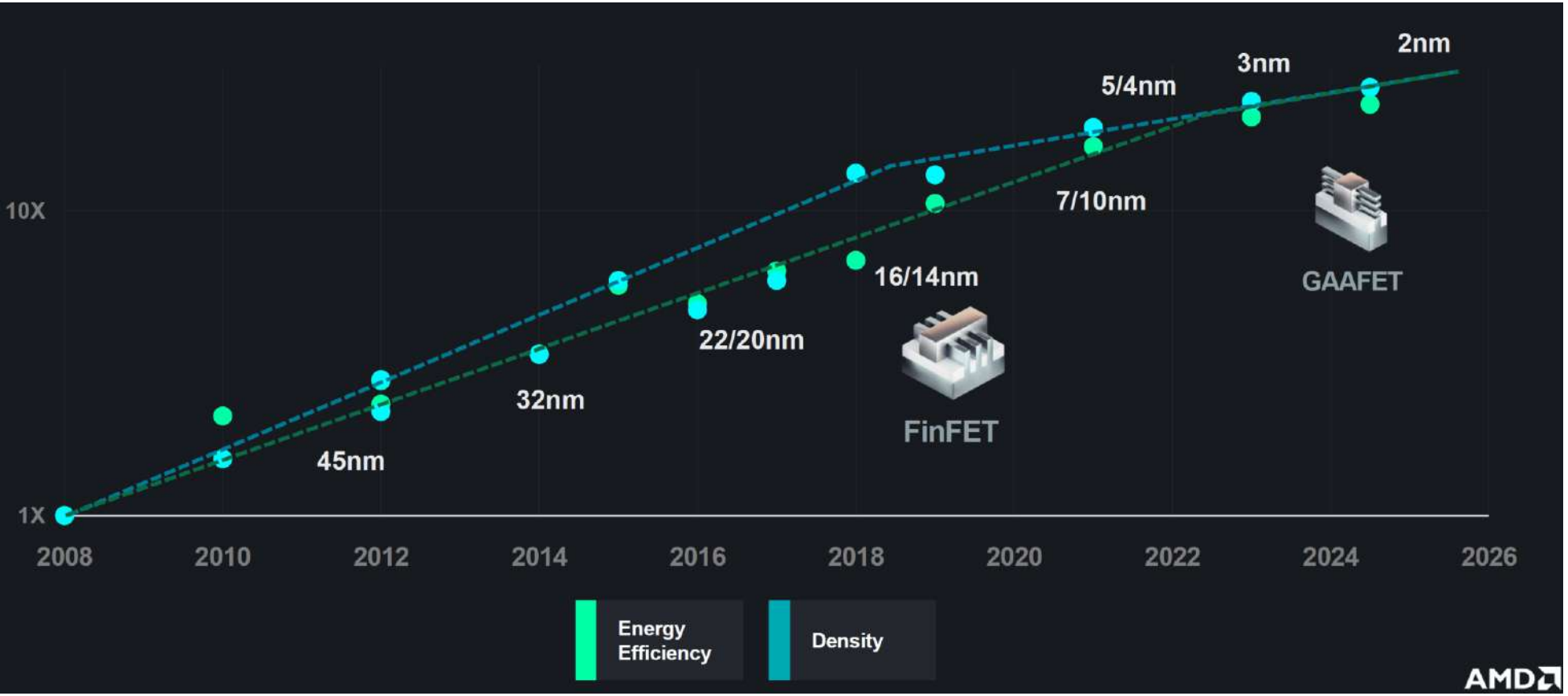
[AI Index Report 24]

Efficient

On-car Computing
P_{MAX} < 1.5 kW

Model complexity
10× every ~2.5 years

Embodied AI



[AMD HotChips24]

Efficient

On-car Computing
P_{MAX} < 1.5 kW

Model complexity
10× every ~2.5 years
Moore's Law
10x every 12 years!

Autonomous Nano-Drones

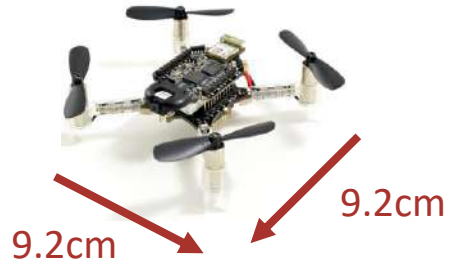
Advanced autonomous drone

A. Bachrach, "Skydio autonomy engine: Enabling the next generation of autonomous flight," IEEE Hot Chips 33 Symposium (HCS), 2021

https://www.skydio.com/skydio-2-plus

- 3D Mapping & Motion Planning
- Object recognition & Avoidance
- 0.06m^2 & **800g of weight**
- Battery Capacity **5410 mAh**

Nano-drone



https://www.bitcraze.io/products/crazyflie-2-1

- Smaller form factor of 0.008m^2
- Weight: **27 g (30× lighter)**
- Battery capacity: **250 mAh (20× smaller)**

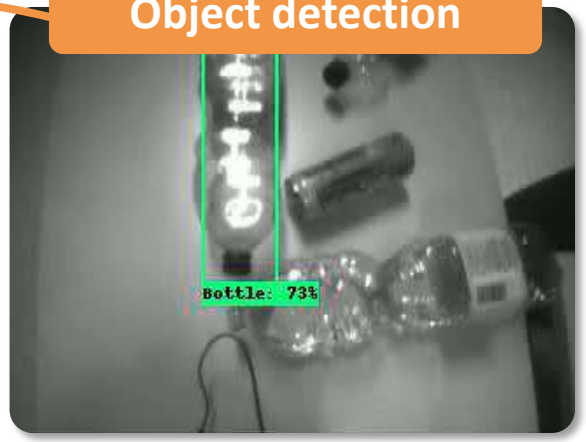
Intelligence in a 30× smaller payload, 20× lower energy budget?

Achieving True Autonomy on Nano-UAVs

Multiple,
complex,
heterogeneous

tasks at high speed and robustness
fully on board

Object detection



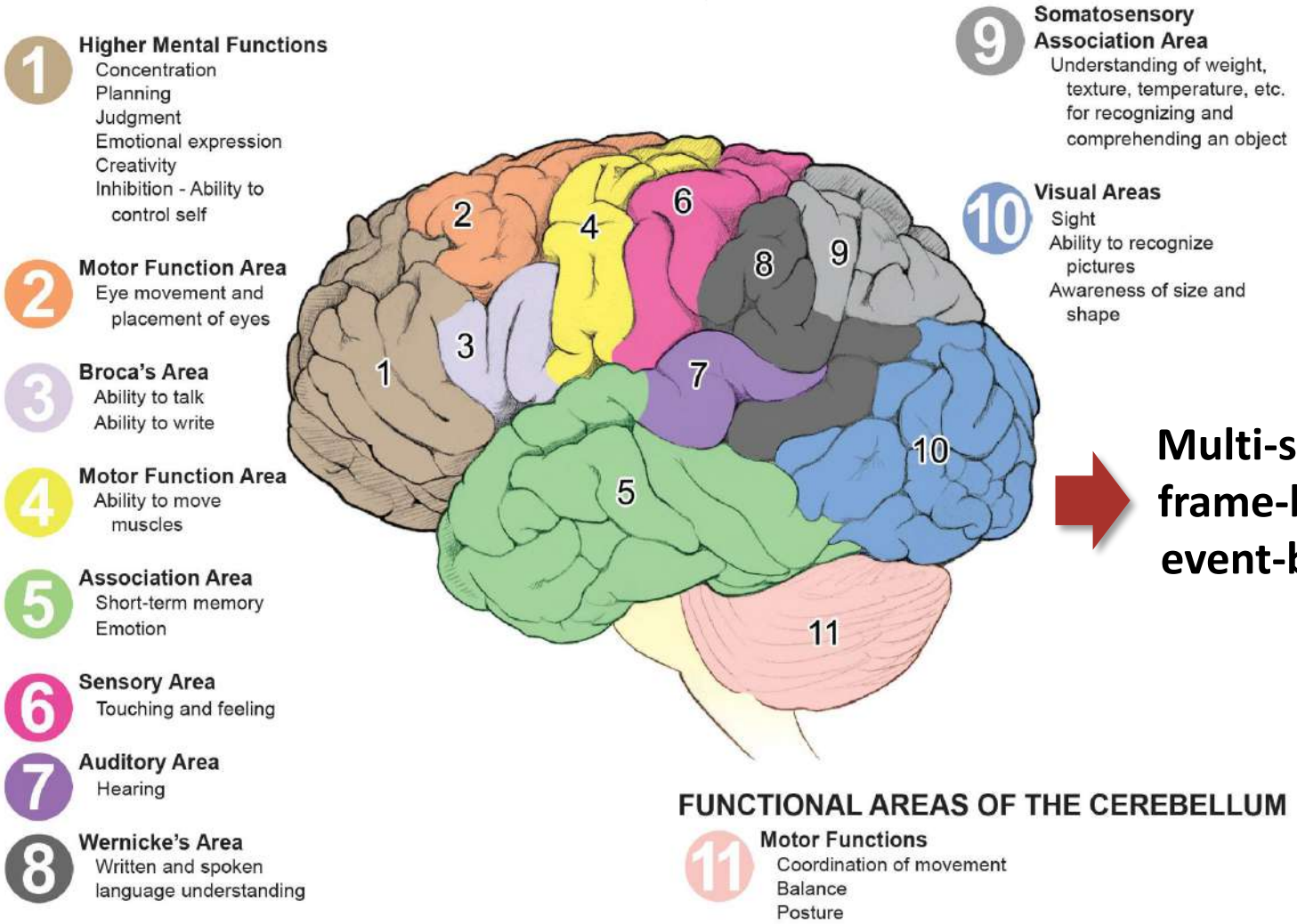
Obstacle avoidance & Navigation

Environment exploration

Multi-GOPS workload at extreme efficiency → P_{max} 100mW

Efficiency through Heterogeneity: Multi-Specialization

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



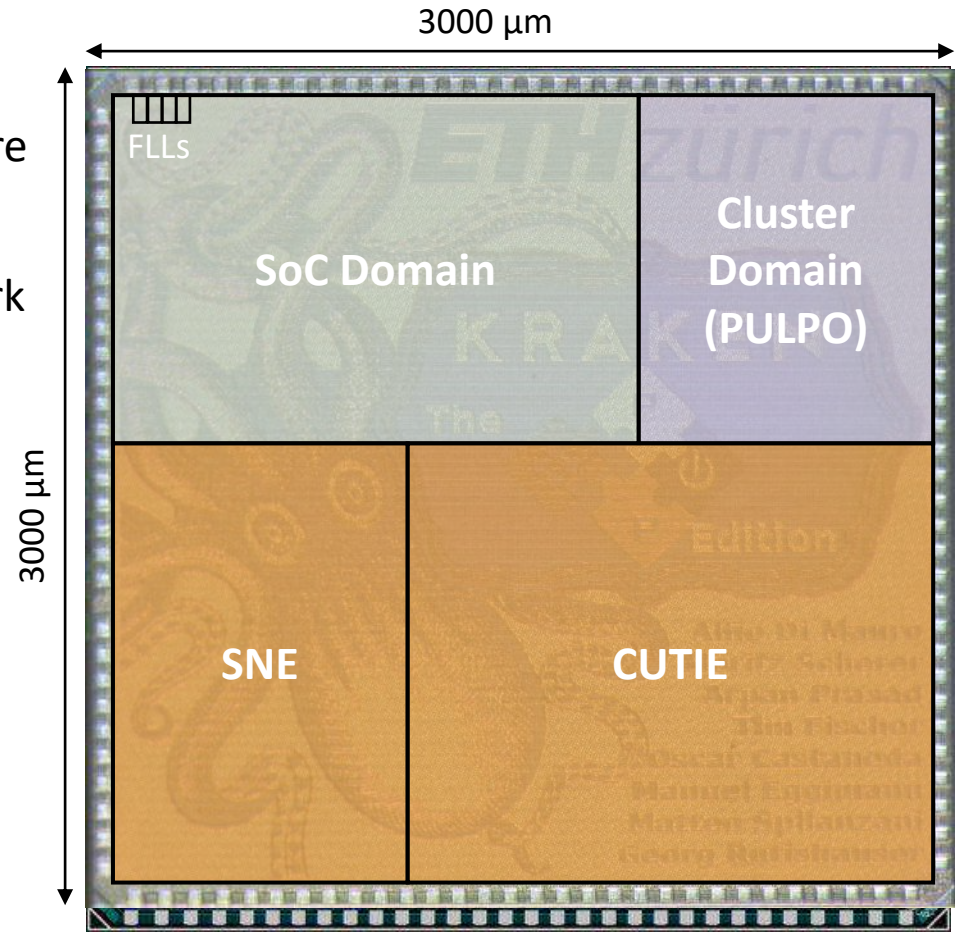
Multi-sensor
frame-based
event-based

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

Kraken: 22nm SoC, Multiple Heterogeneous Accelerators

The *Kraken*: an "Extreme Edge" Brain

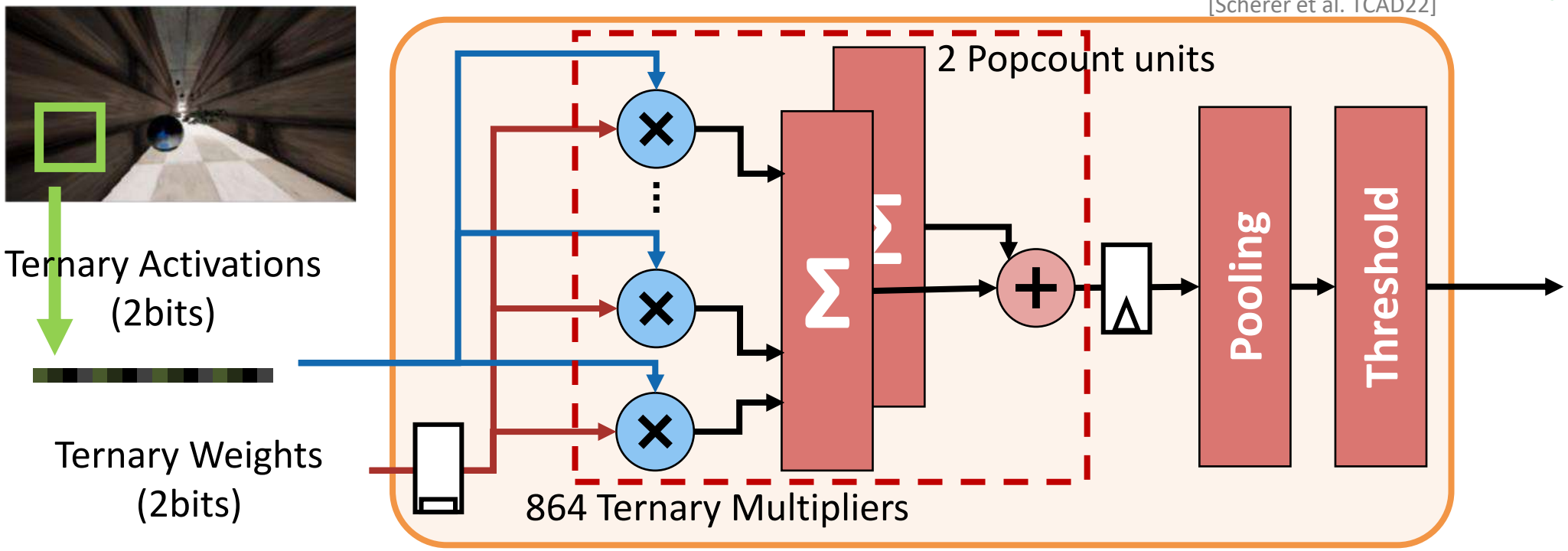
- RISC-V Cluster
  8 Compute cores +1 DMA core
- CUTIE
  Dense ternary-neural-network accelerator
- SNE
  Energy-proportional spiking-neural-network accelerator



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]



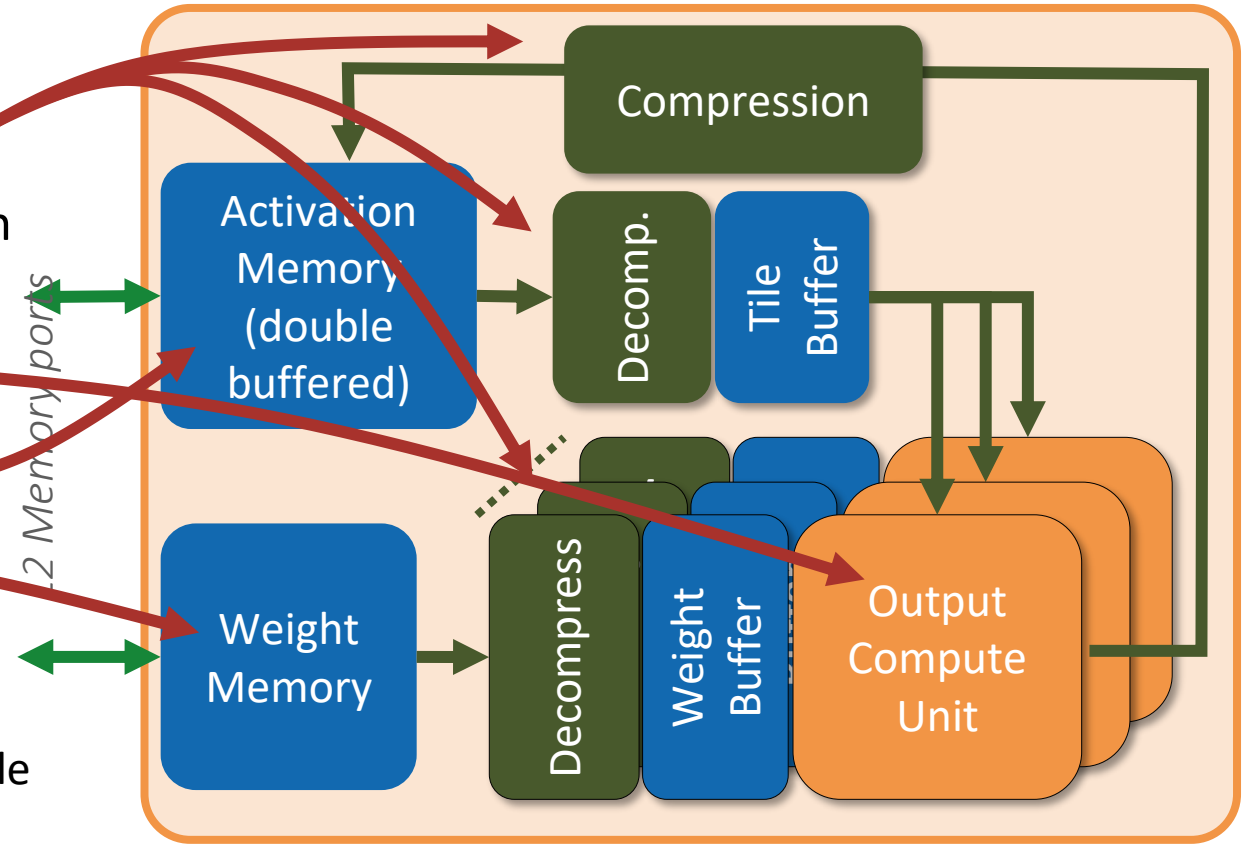
Output channel compute unit (OCU)

- **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

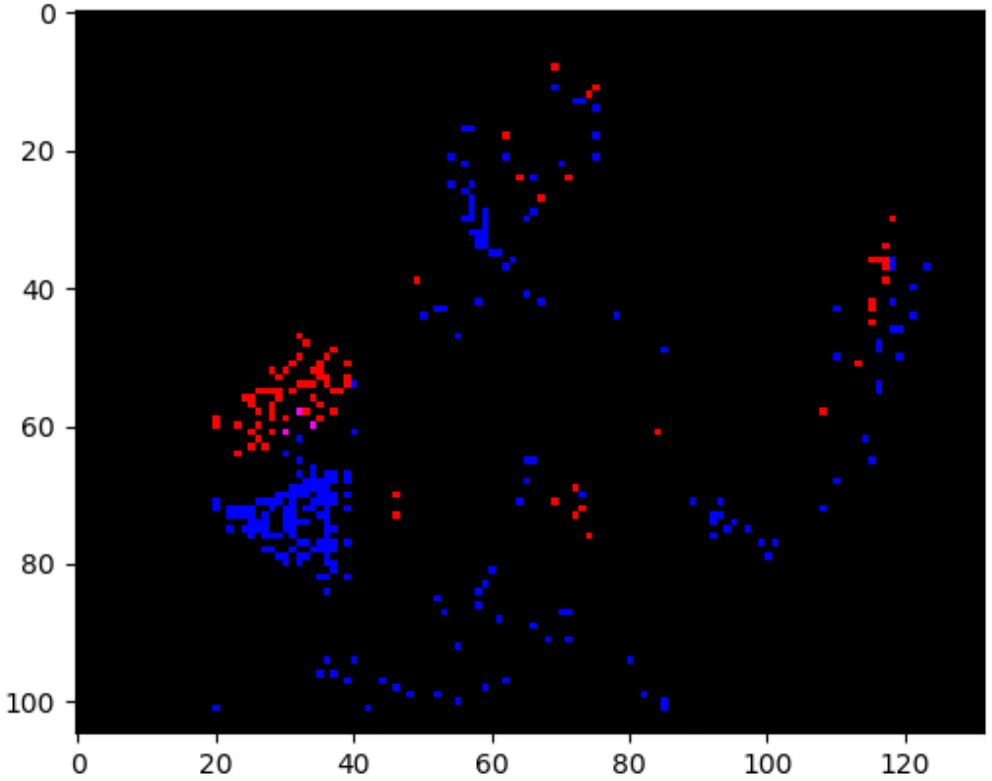
- Data in 1.6 bits (Ternary value) with On-the-fly Compression/Decompression
- Configuration in Kraken
  - 96 channels (Output compute units)
  - 3 × 3 kernels
  - 64 × 64 pixels feature maps (158 KiB)
  - 9 layers of weights (117 KiB)
- Lots of TMAC/cycle
  - 96 OCUs, 96 Input channels, 3 × 3 kernels: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

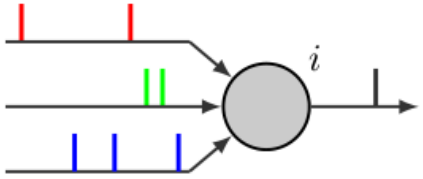


1fJ/MAC

SNE: Perception on Event Sensors

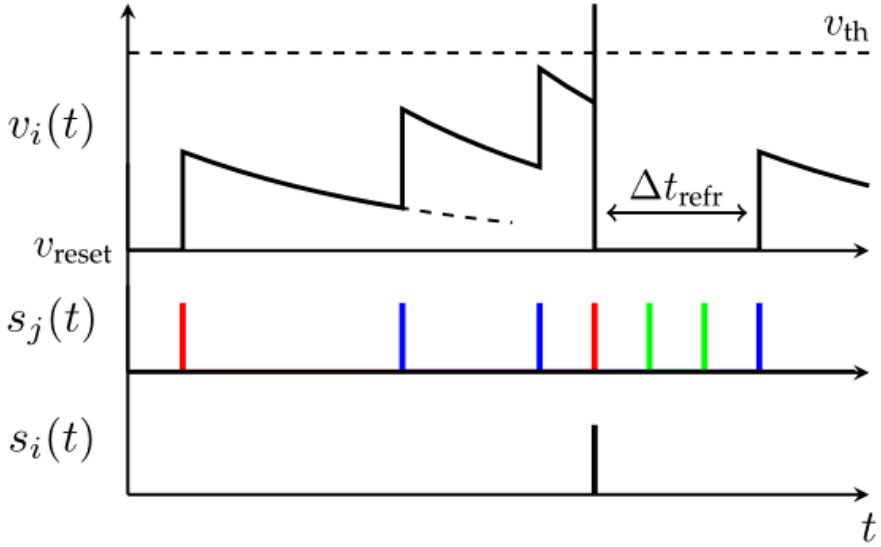
Event Sensors – DVS camera
Ultra-low latency
Energy- proportional interface





Spiking Neural Engine (SNE)

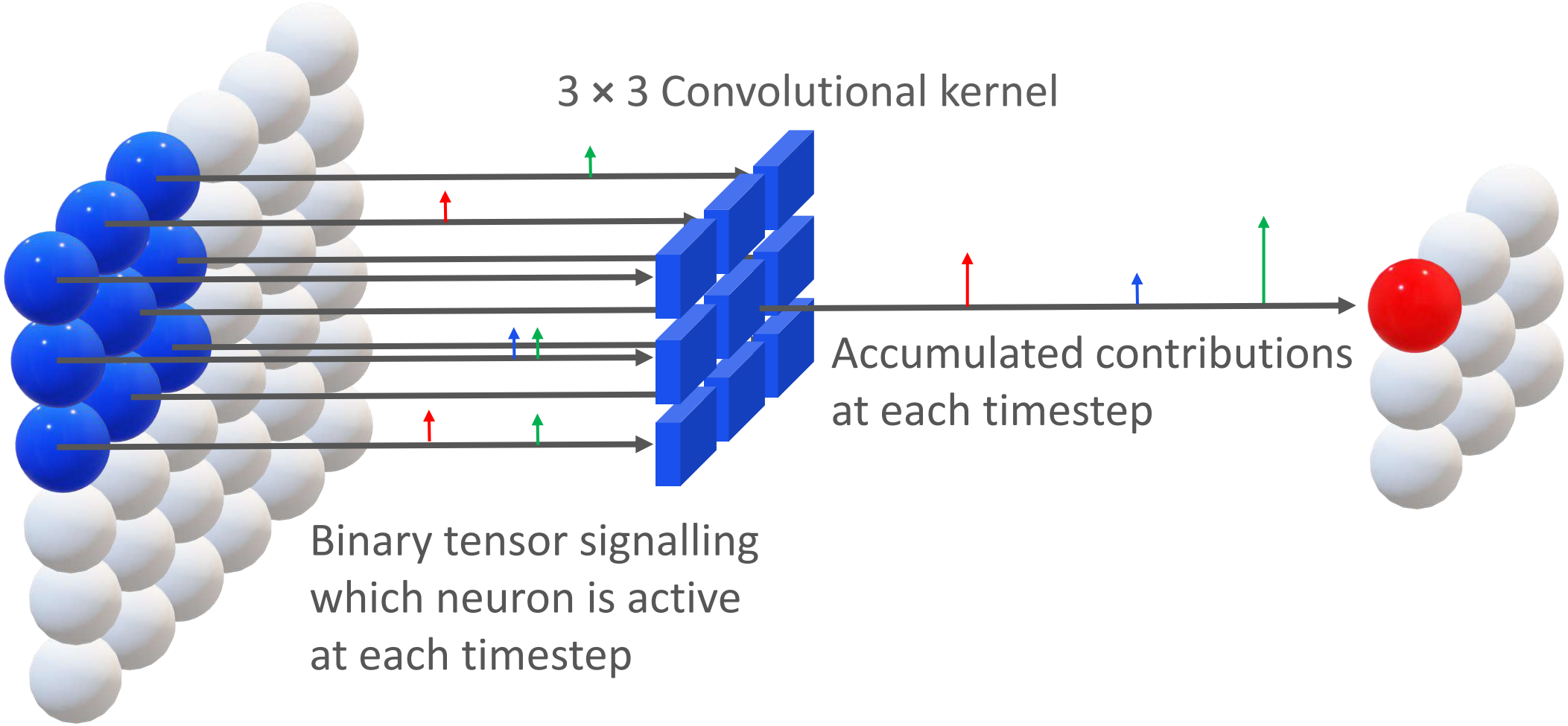
Leaky Integrate & Fire (LIF) neurons



[Di Mauro et al. DATE22]

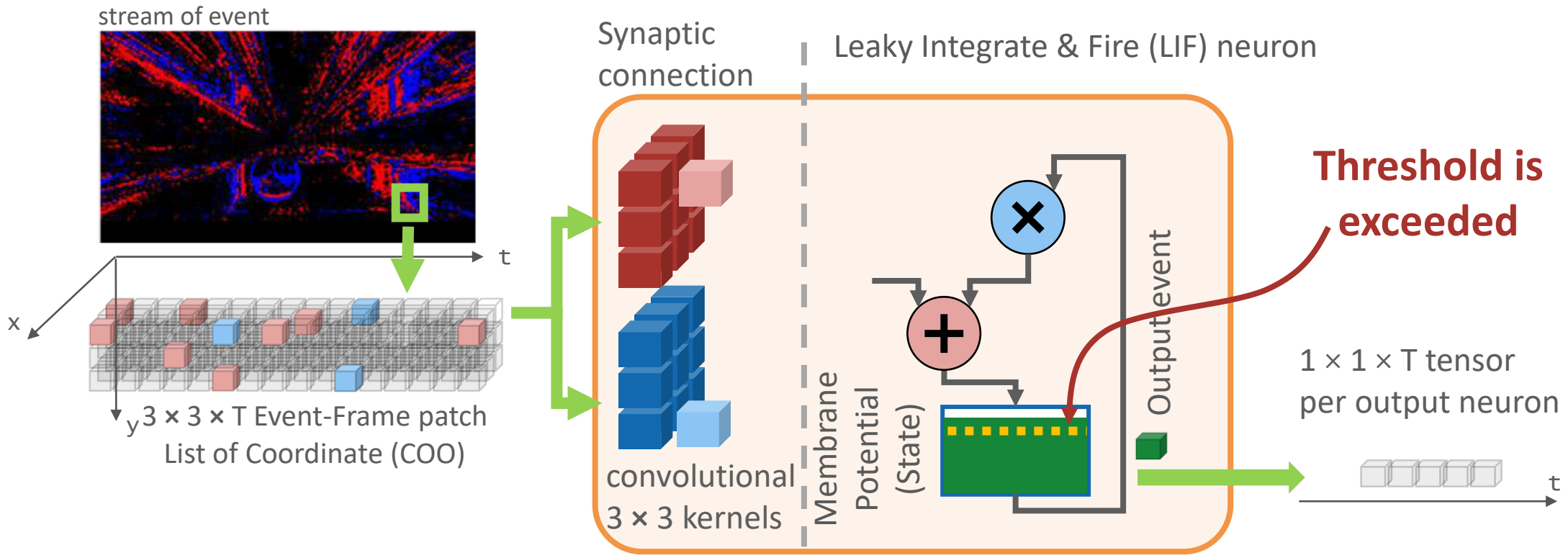
SNE works seamlessly with DVS (event-based) sensors

How does a convolution work in SNNs?



Perform operations only if a spike is present

Event consumption, and output spikes generation



SNN's neuron processing element (PE)

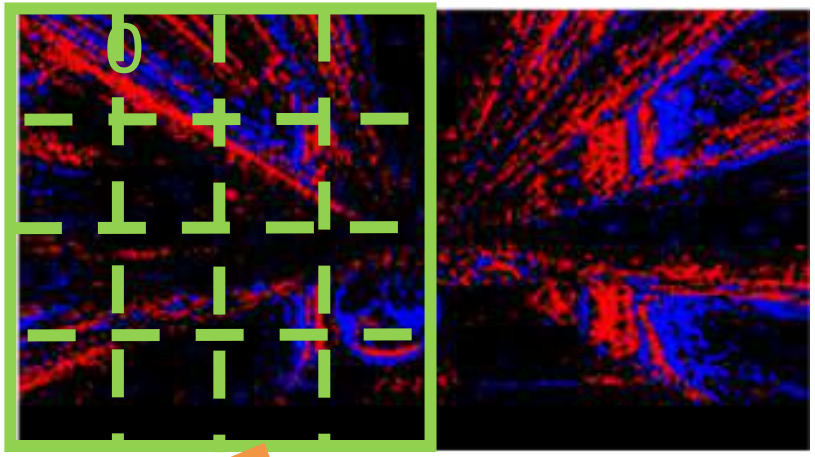
A more complex dynamic than conventional DNNs neurons:

- 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)

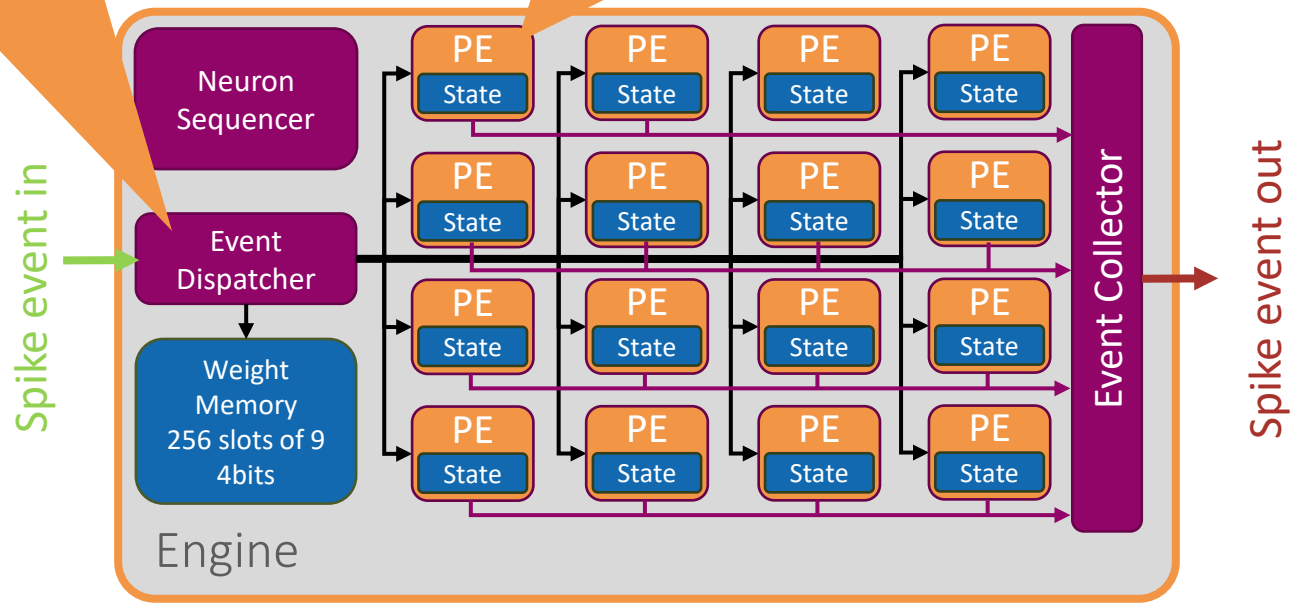
Single SNE Engine Architecture

Broadcast events to all PEs

Each PE filters only the events in its spatial region of interest



Tiled execution on an input patch



Achieves true energy-proportionality: 1 neuron update per cycle

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
```

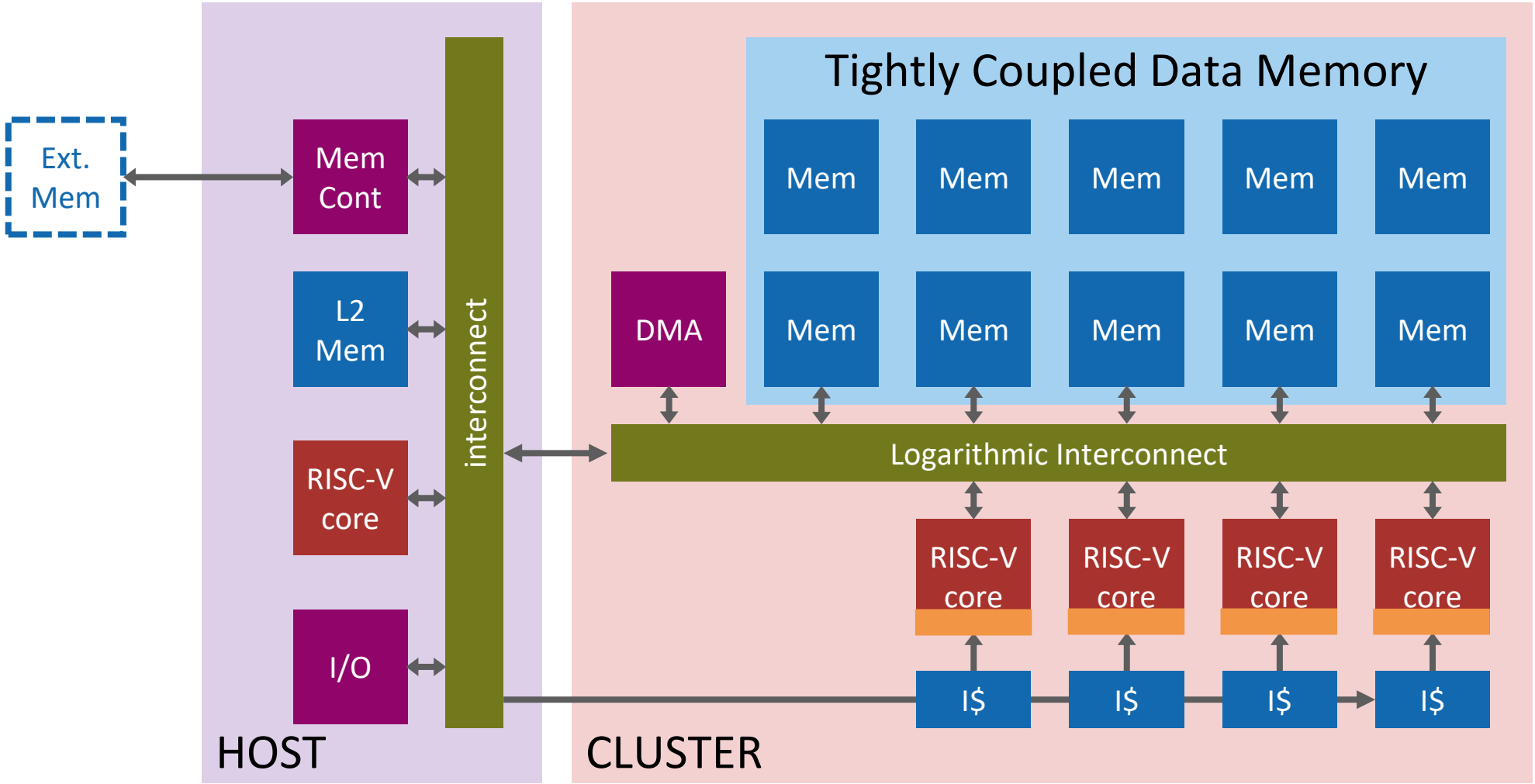
RISC-V core

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



Advancing the SOA on all tasks

RISC-V Cluster

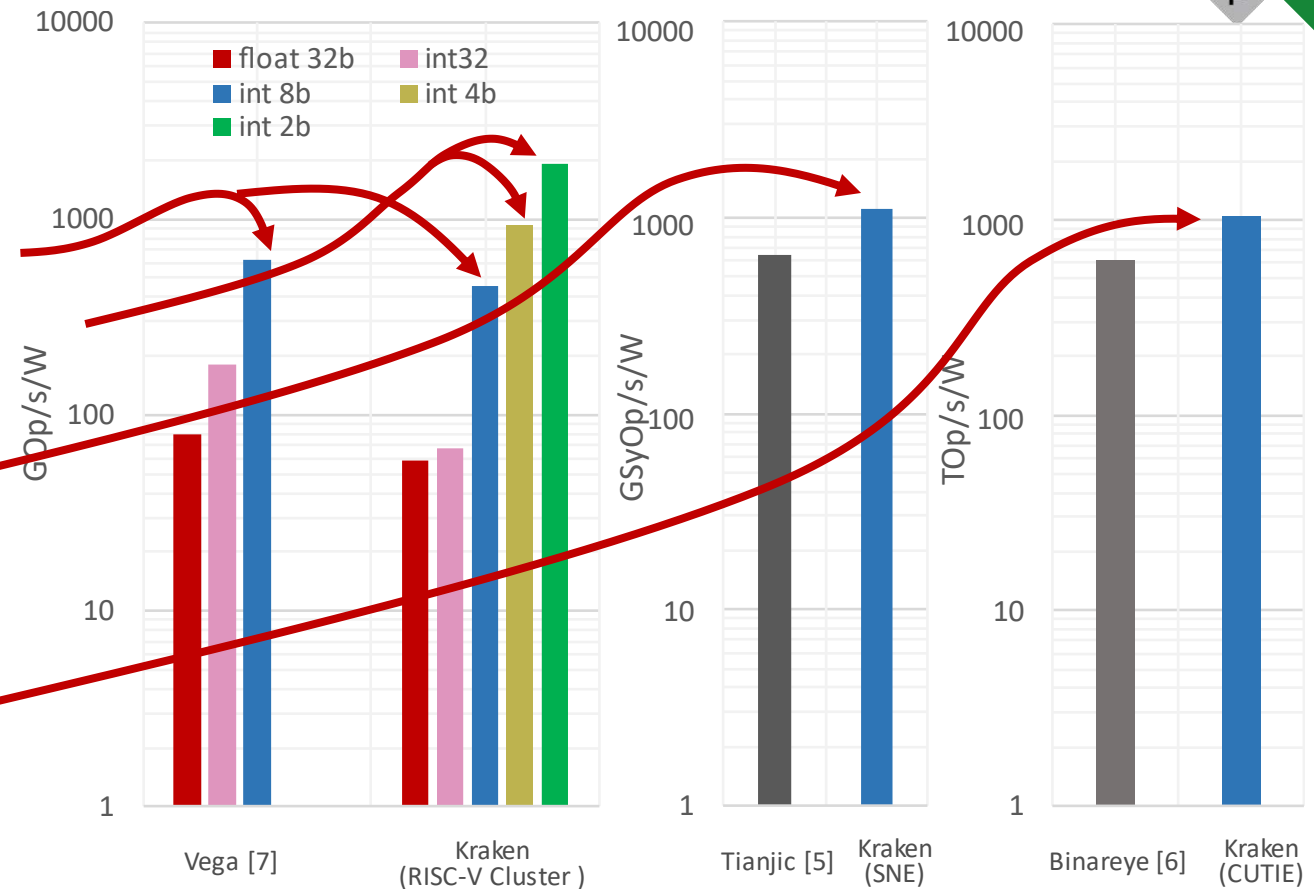
- Comparable 32bits-8bits SOA Energy efficiency to other PULPs
- The highest energy efficiency on sub-byte SIMD operations (4b-2b)

SNE

- 1.7× higher than SOA energy/efficiency

CUTIE

- 2× higher energy efficiency improvement over SOA



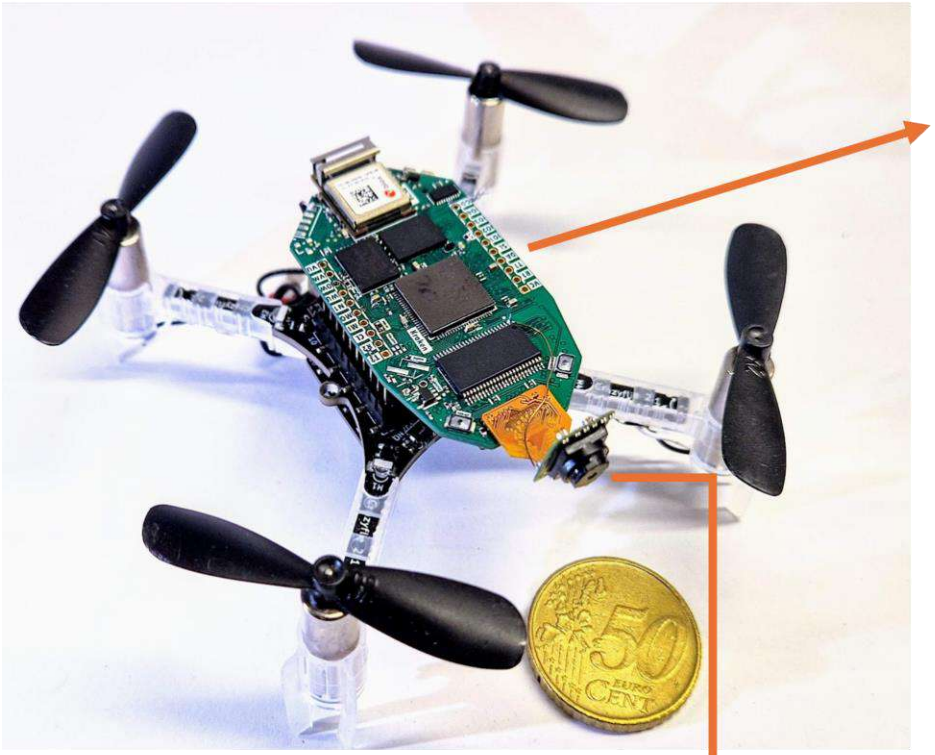
CUTIE, SNE work concurrently → SNN+TNN "fused" inference

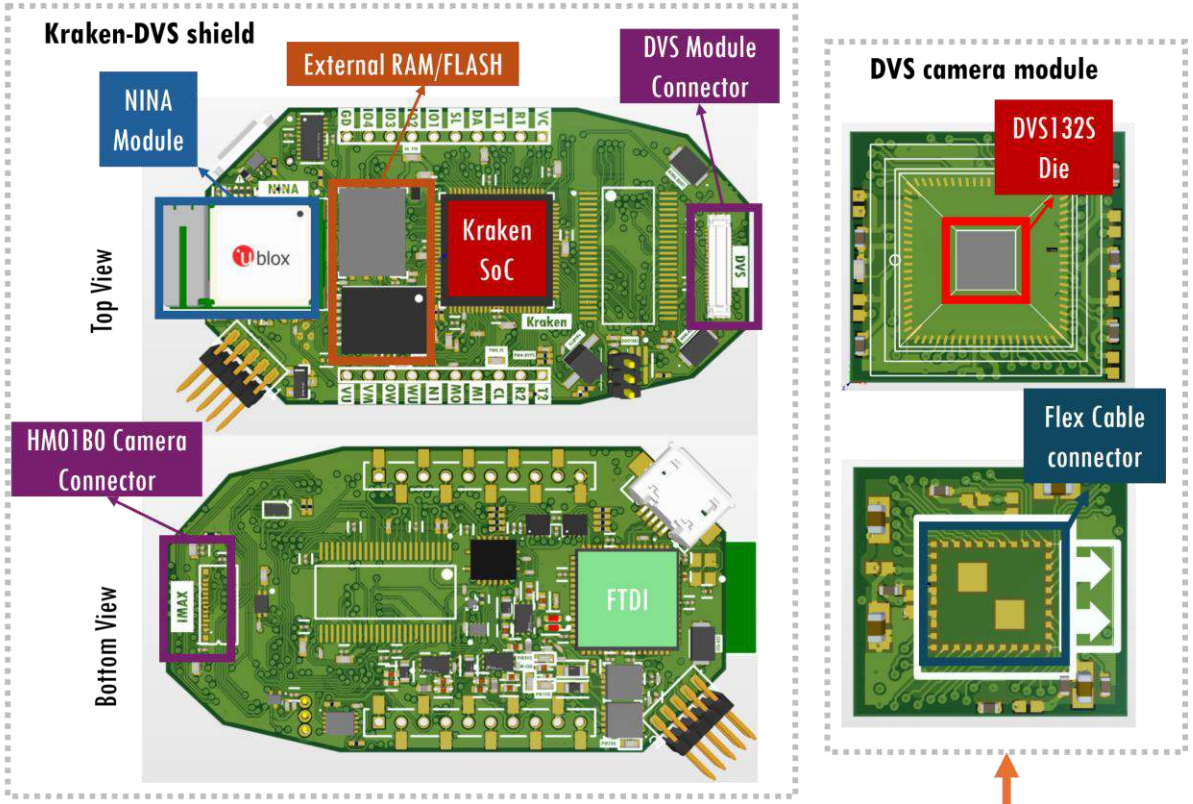
L. Deng et al., "Tianjic," JSSC 2020
B. Moons et al., "Binareye," CICC, 2018
D. Rossi et al., "Vega," JSSC 2022.

Kraken Shield and System Architecture

- 7g payload
- DVS and frame-based cameras → real-time multi-modal perception.
- Designed for integration into nano-UAV platforms

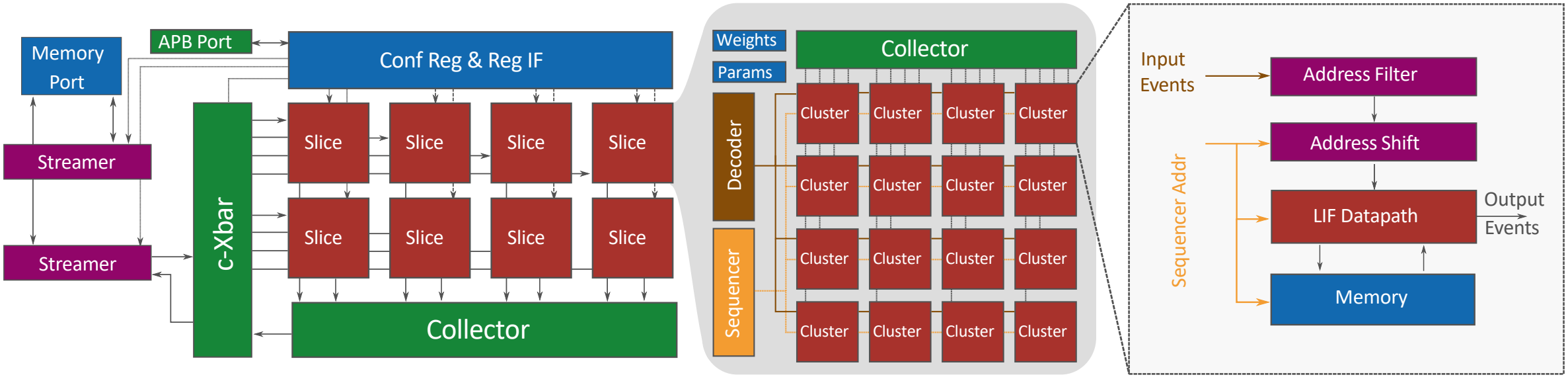
Kraken-DVS shield on the CF





Spiking Neural Networks for Depth Estimation

SNN → SCNNs for depth estimation.



Depth Estimation

1.02k inferences/s

Energy Efficiency

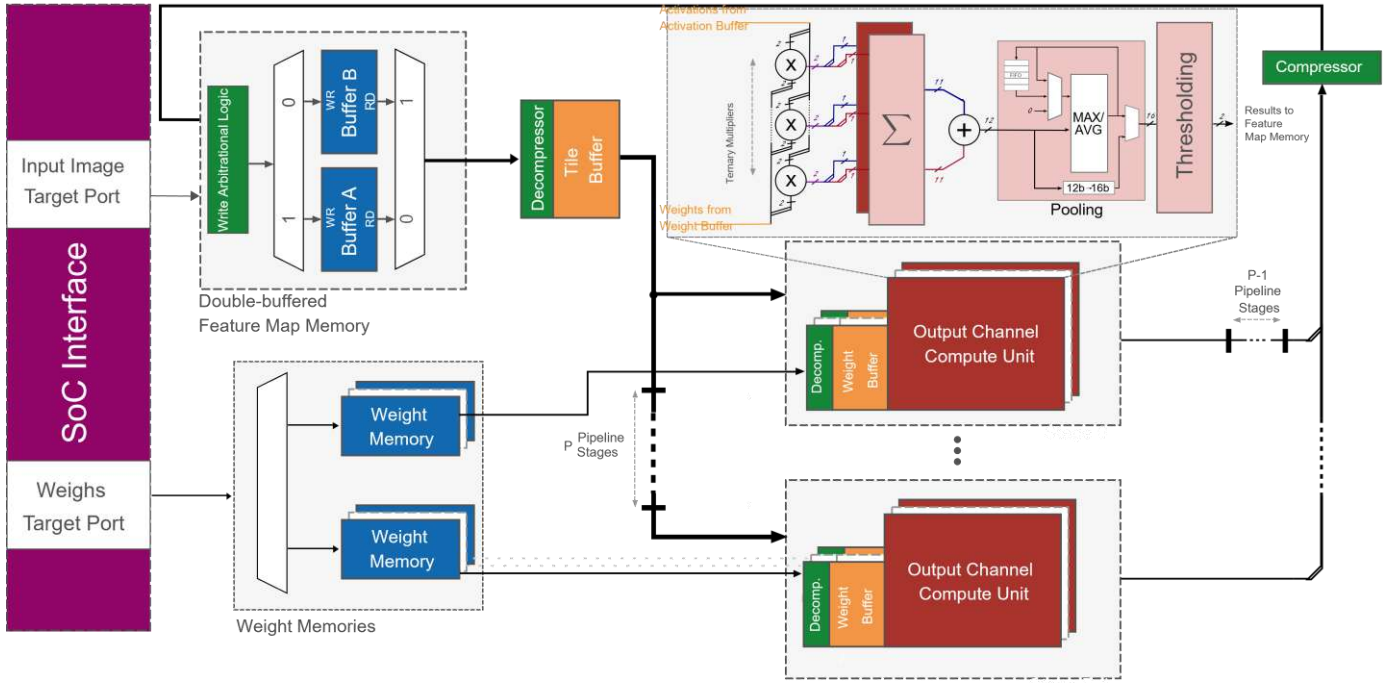
18 µJ per inference

Low Power

98mW @ (220MHz, 0.8V)

Ternary Neural Networks for Object Classification

CUTIE → TNN for object classification.



Object Classification

10k inferences/s

Energy Efficiency

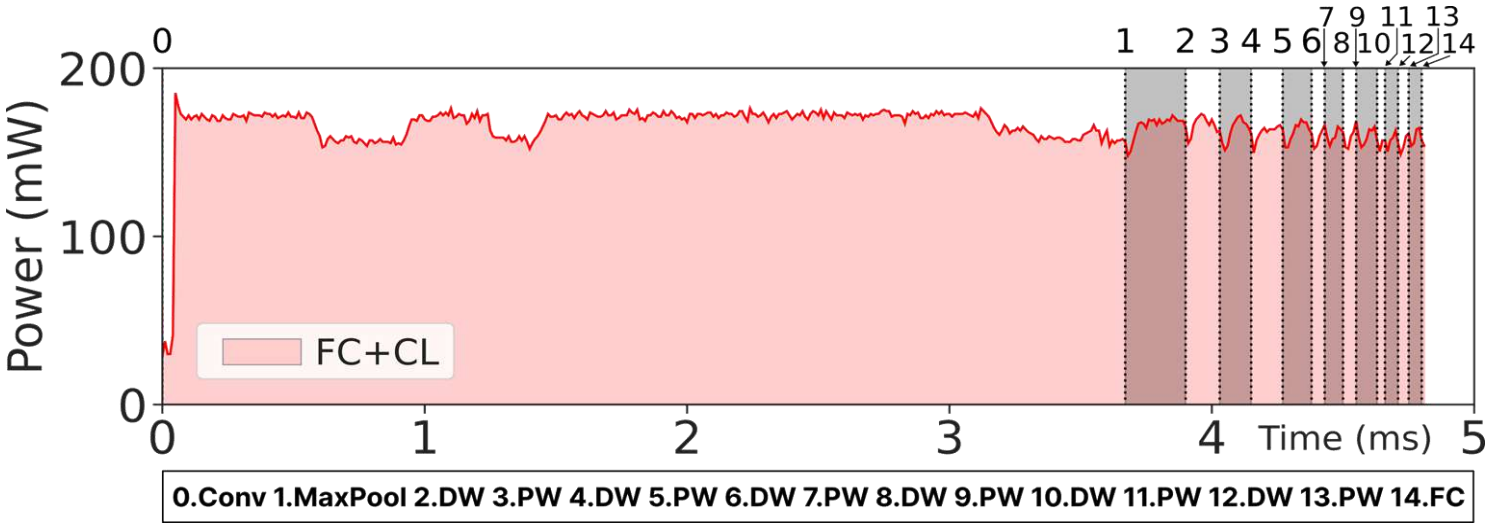
6 µJ per inference

Low Power

110mW @ (330MHz, 0.8V)

Kraken Power Consumption (all Included)

Combined power consumption of SNE, CUTIE, PULP cluster

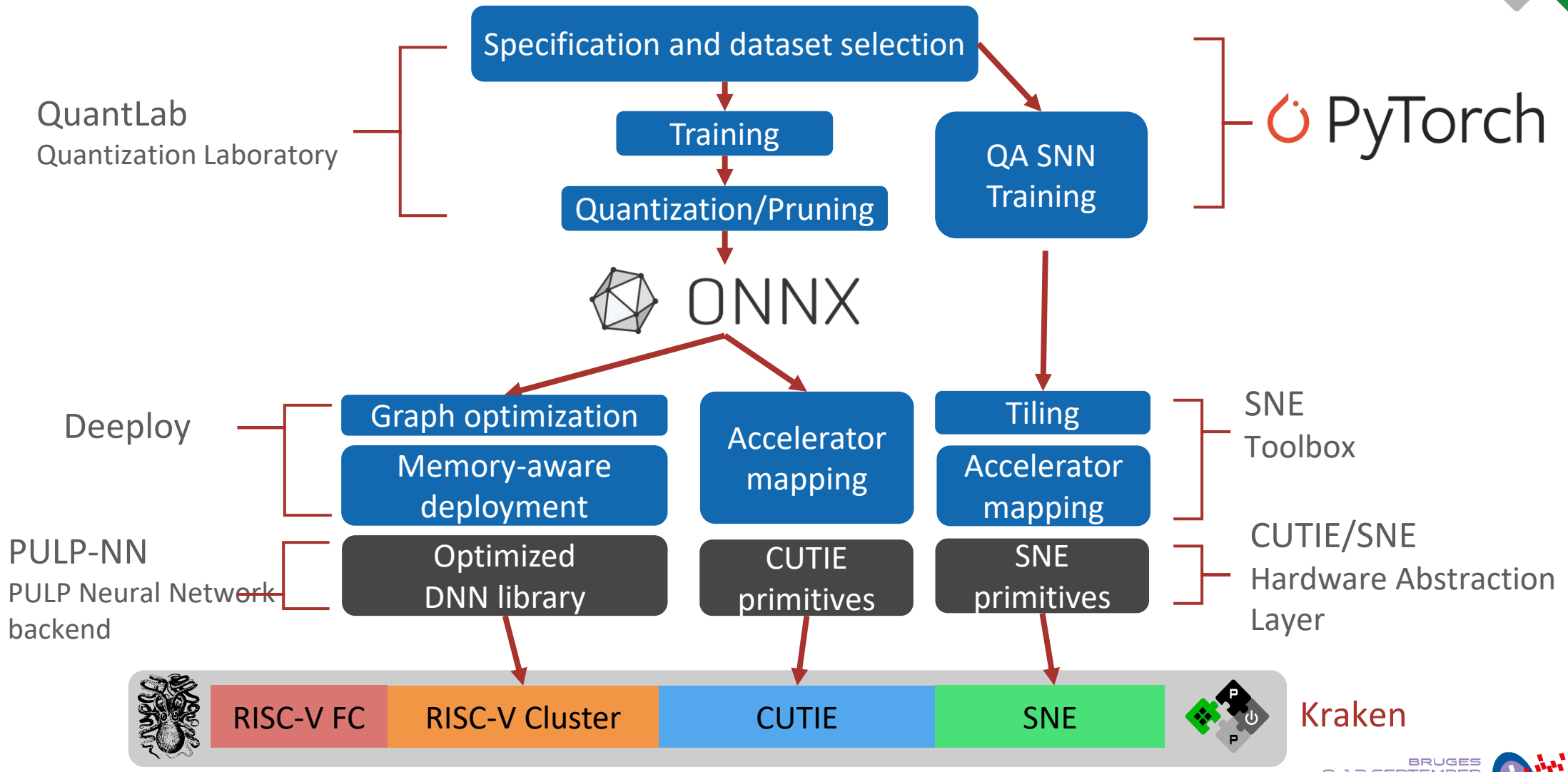


Model	Inference/s	µJ/inf	Power (mW)
SNE	1.02k	18	98
CUTIE	10k	6	110
PULP	221	750	165

Kraken power waveform executing Tiny-PULP-Dronet at FC@280 MHz, CL@300 MHz, Vdd@0.8 V

P=373mW, representing just 5% of the UAV's power budget

How to deploy applications to PULP/Kraken?



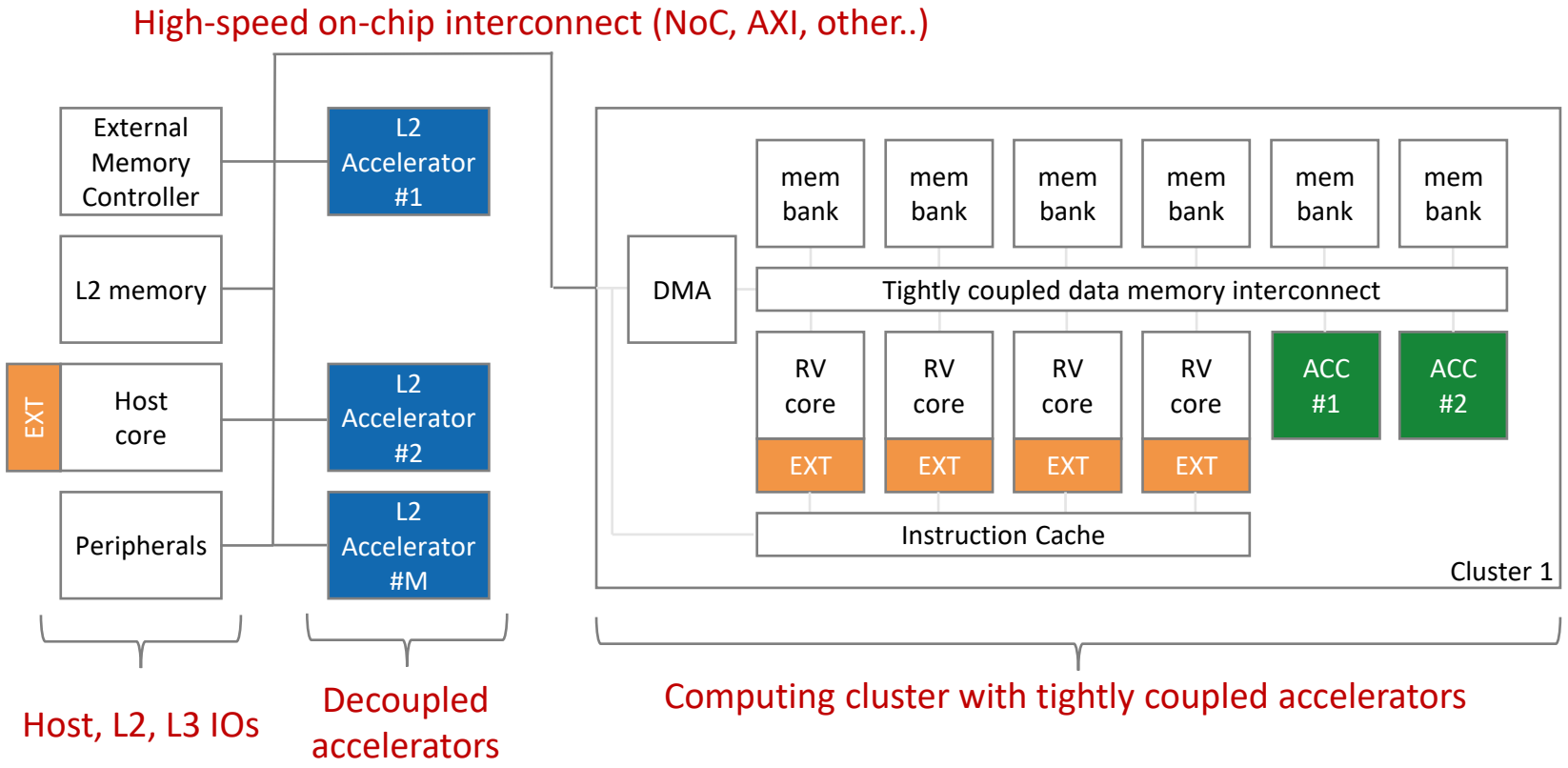
Heterogeneous, Multiscale Accelerated Computing

Multiple Scales of acceleration

Extensions to processor cores
- Explore new extensions
- Efficient implementations

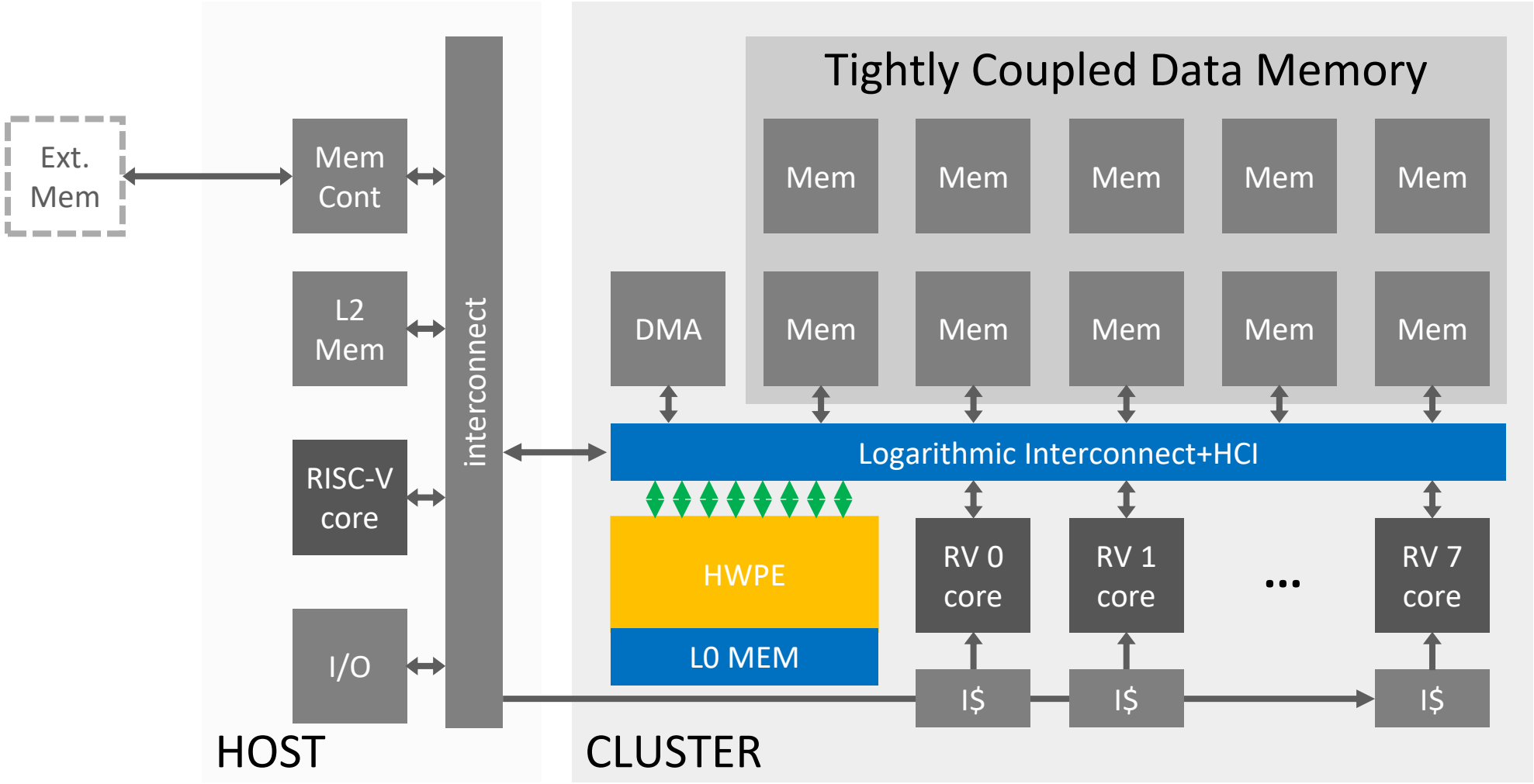
Shared-memory Accelerators
- Domain specific
- Local memory

Multiple Decoupled Accelerators
- Communication
- Synchronization



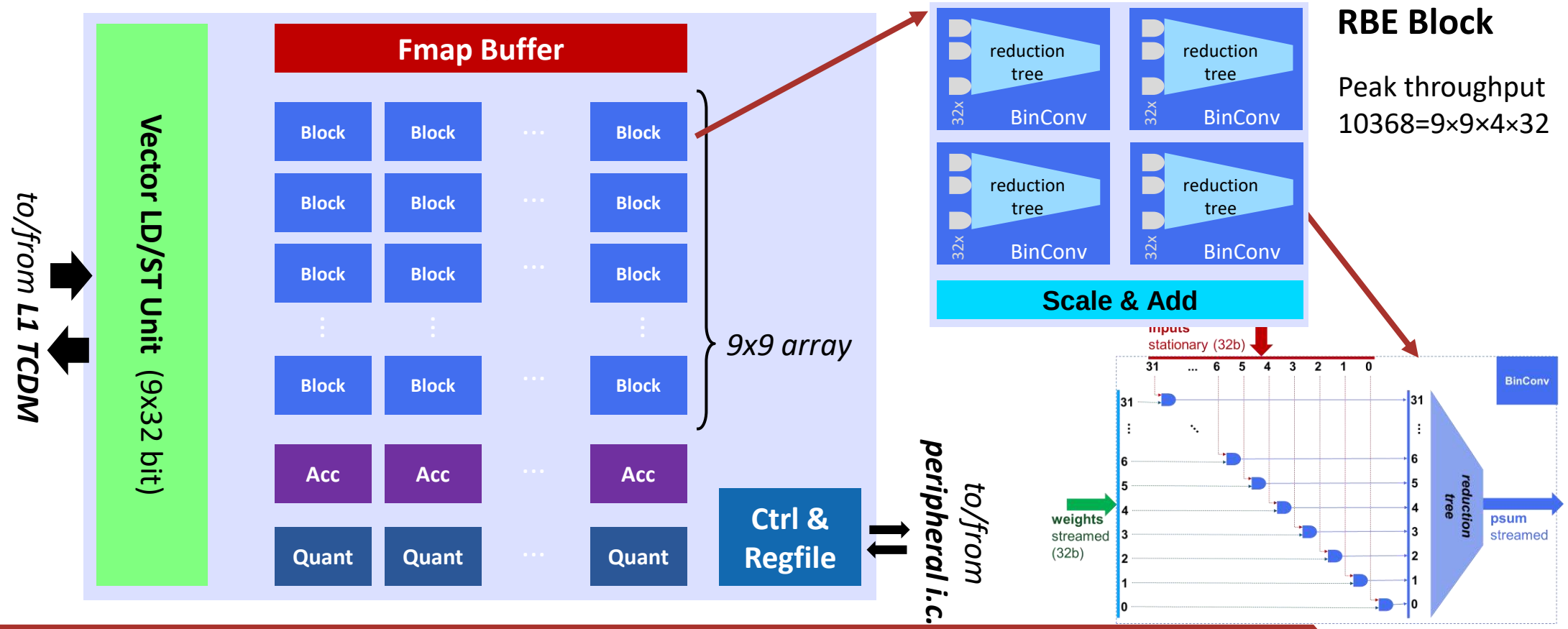
RISC-V is a key enabler → max agility, enabling SW build-up, without vendor lock-in

Tightly-coupled Accelerators



HWPE: Reconfigurable Binary Engine

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

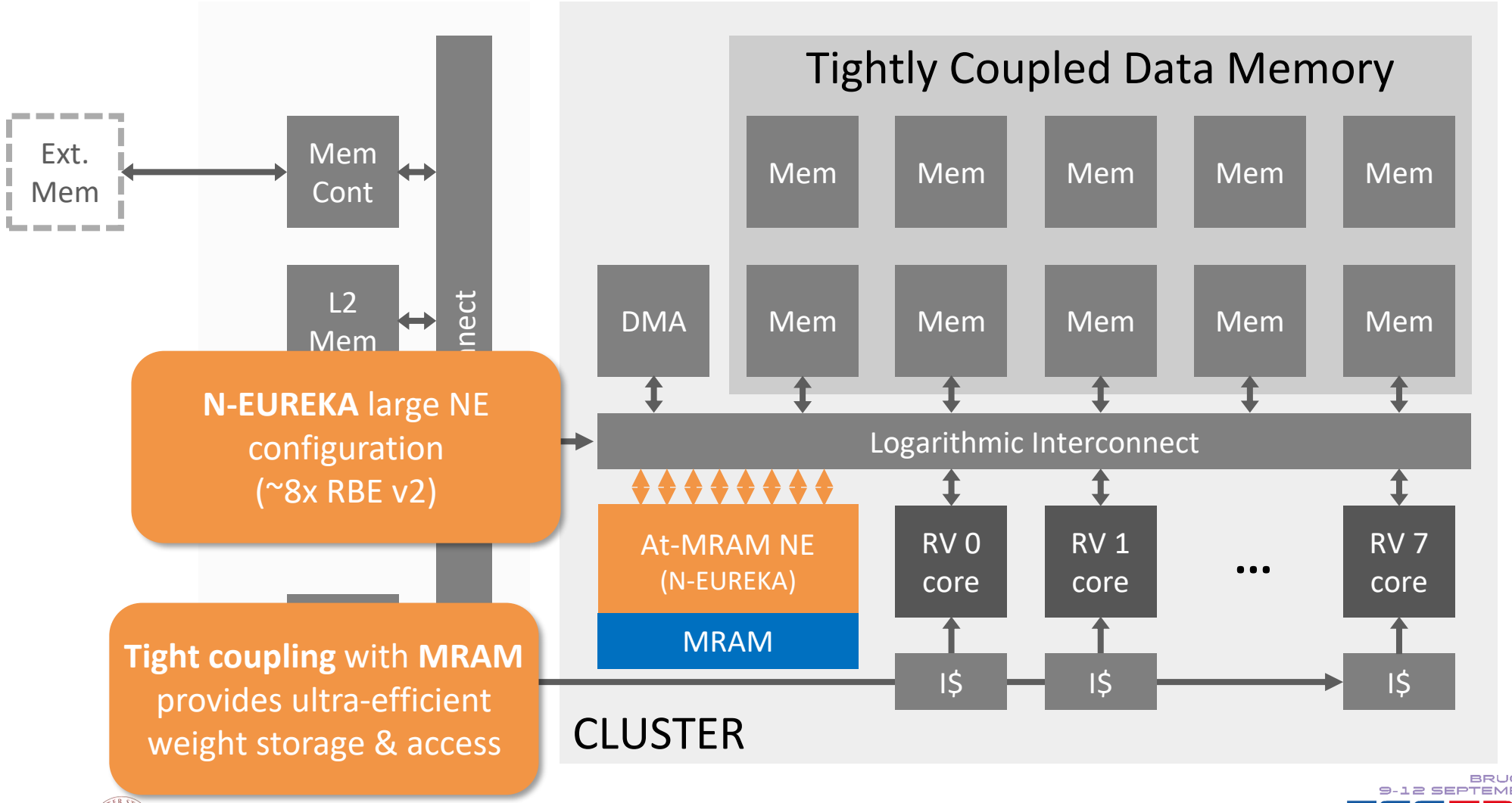


RBE Block

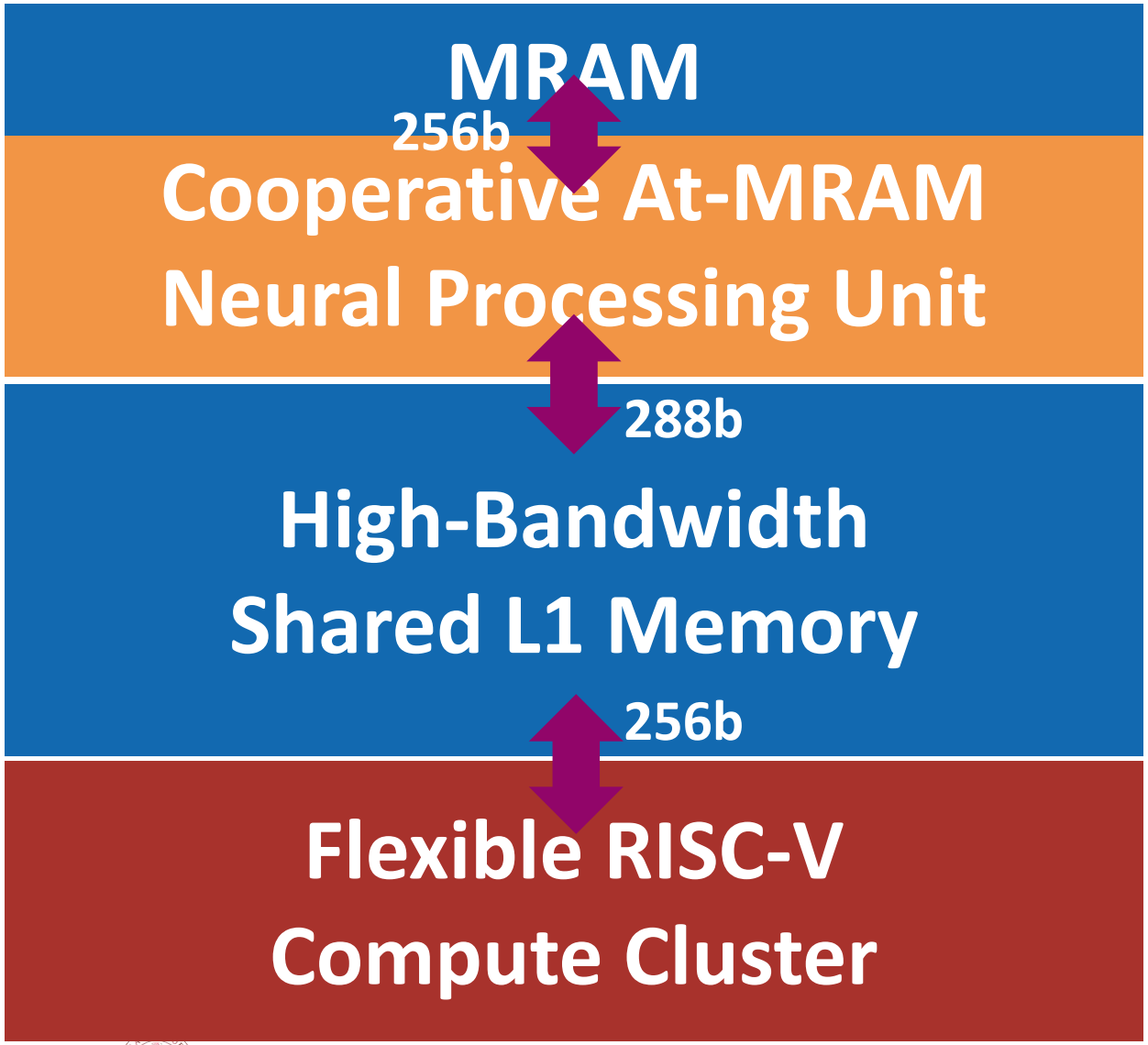
Peak throughput
10368=9×9×4×32

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

Siracusa: Higher performance cluster with N-Eureka

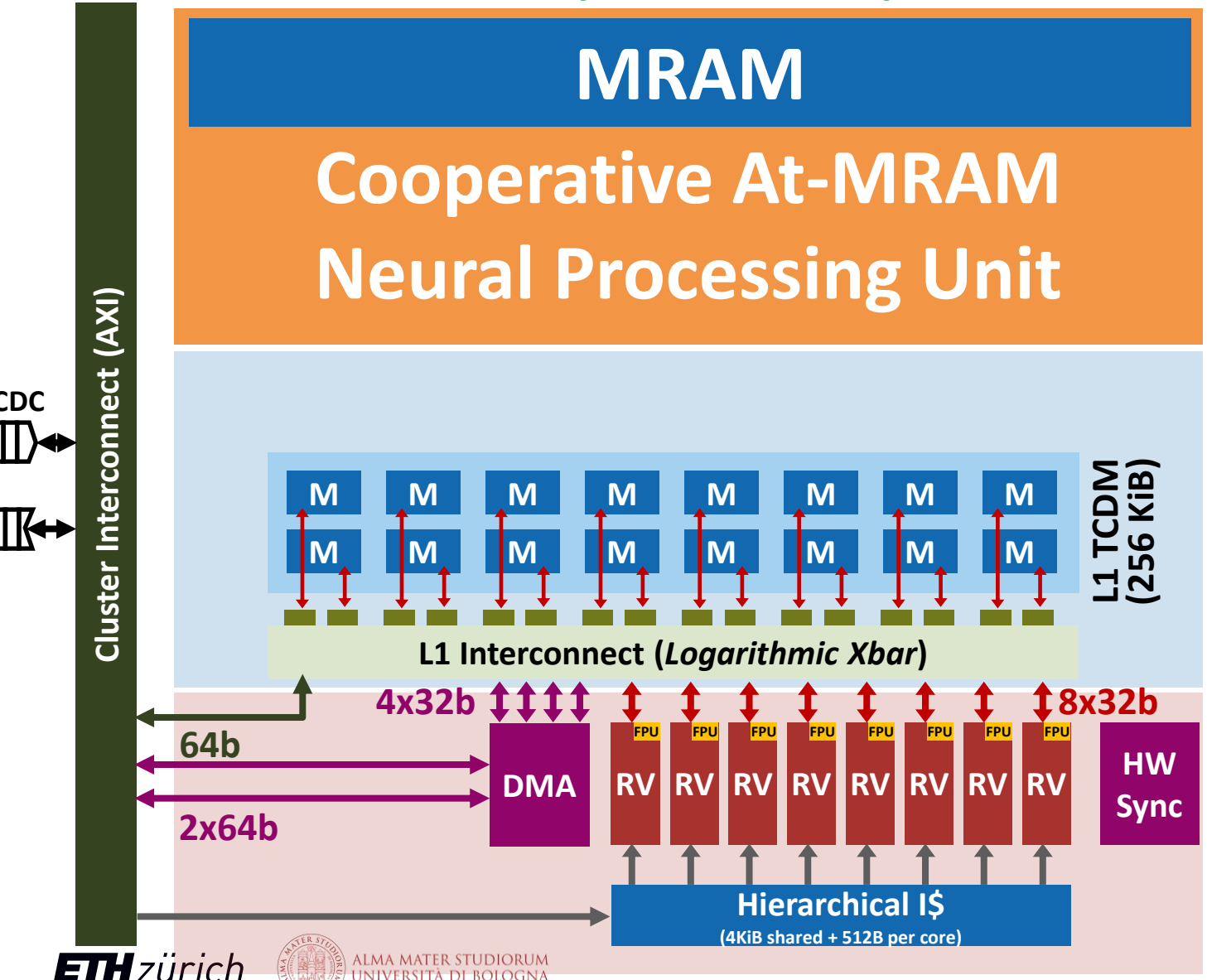


Siracusa: Memory Hierarchy and Dataflows



- Tight coupling between all units at high bandwidth and low latency
- Seamless cooperation between *hardware-accelerated* and *software-defined* functions

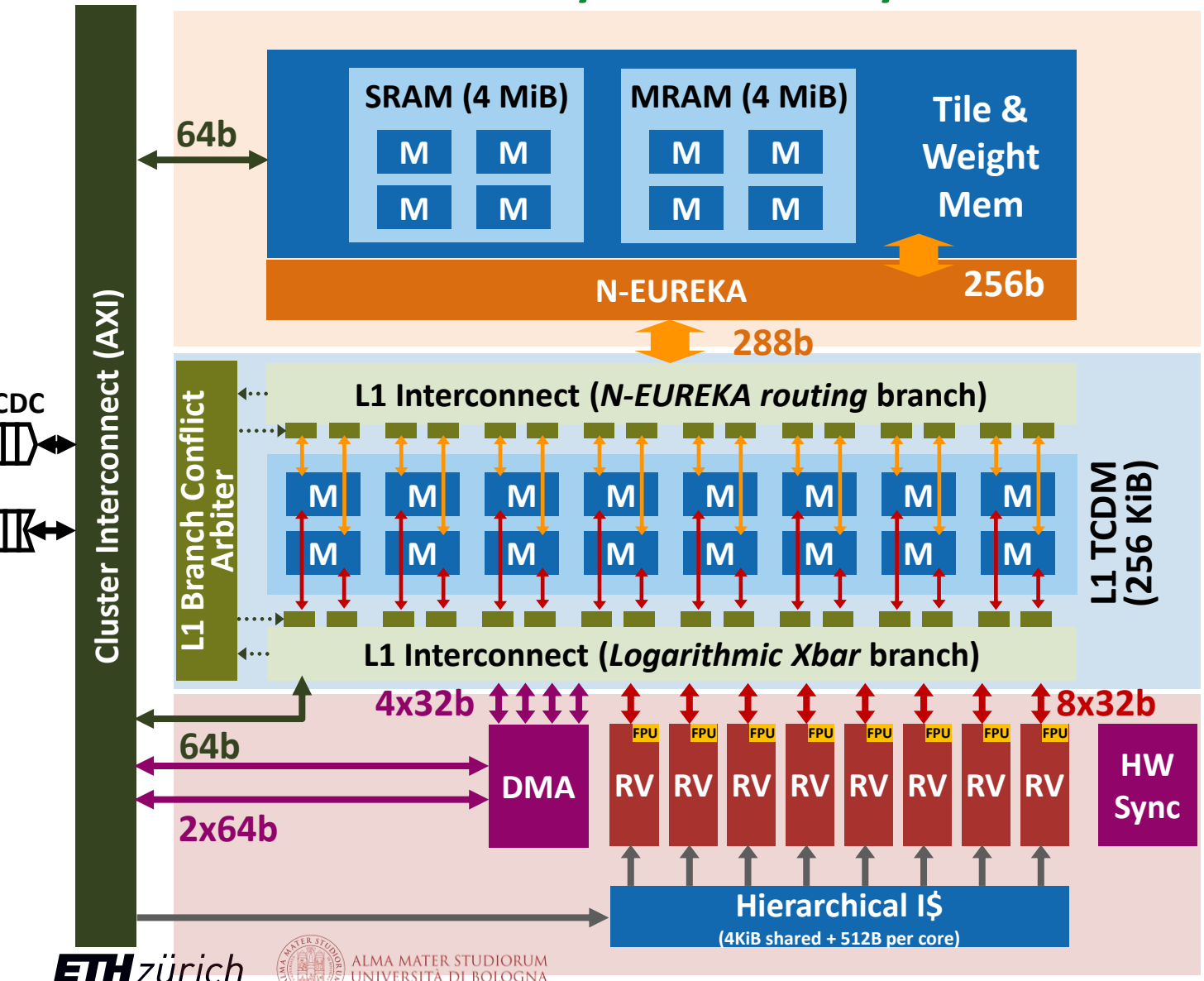
Siracusa: Memory Hierarchy and Dataflows



A "classic" PULP cluster with 8 RV32IMCFXpulpnn cores

- private multi-precision FPUs
- hierarchical instruction cache (4 KiB + 512B per core)
- Xpulpnn extensions for integer mixed-precision DSP + DNNs
- 256 KiB of Tightly-Coupled Data Memory (TCDM) divided in 16 word-interleaved SRAM banks
- L1 Logarithmic Xbar for single-cycle, high concurrency access

Siracusa: Memory Hierarchy and Dataflows



Boost memory energy efficiency

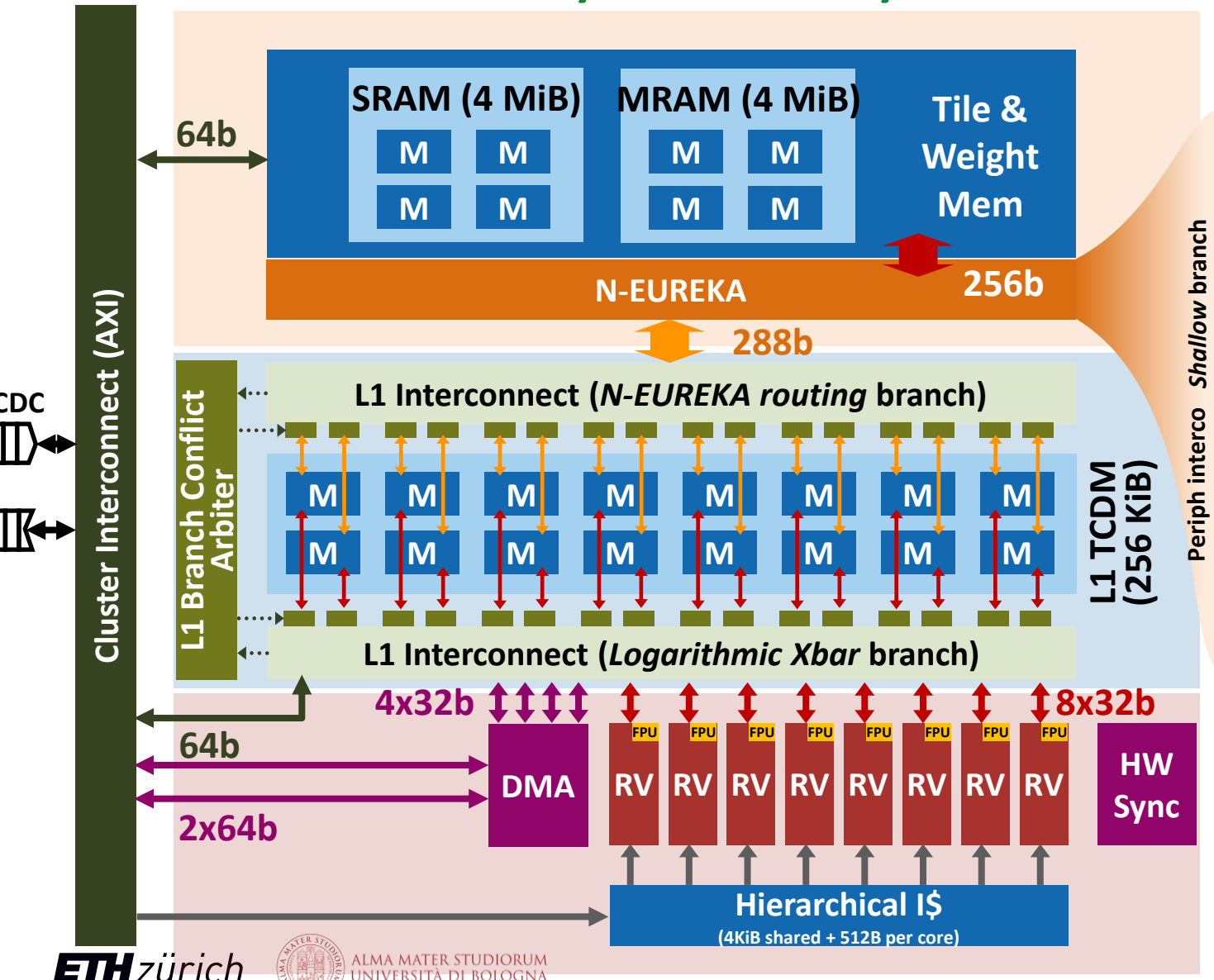
A large power-optimized on-chip memory for network weights → cluster-level weight stationarity

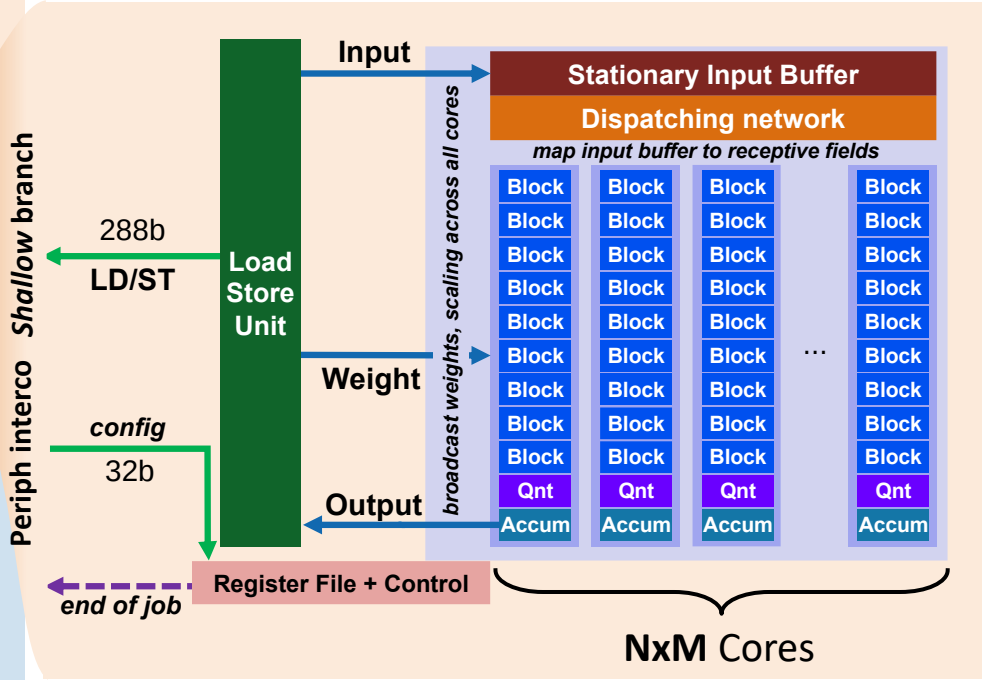
4x 1MiB SRAM banks (64b-wide)

4x 1MiB MRAM banks (64b-wide)

Paging support for transparent network reconfiguration with negligible increase in overall circuit area.

Siracusa: Memory Hierarchy and Dataflows





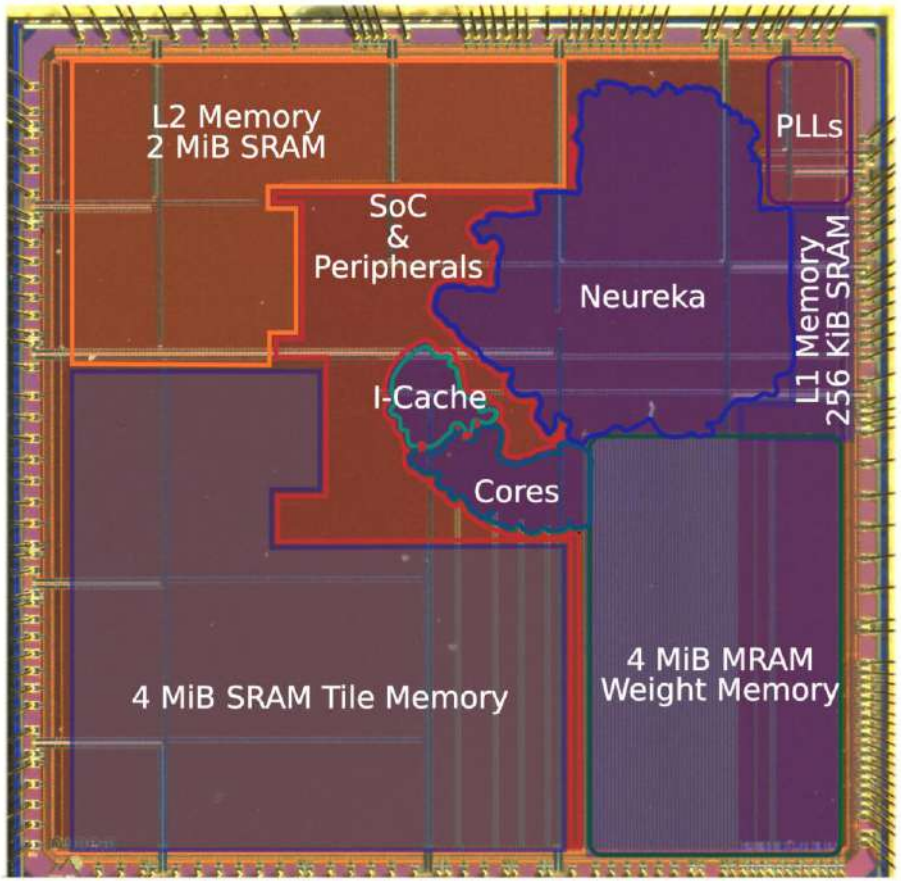
1 Core = receptive field of 1×1 pixel in/output across 32 out-channels

Output stationary, Input quasi-stationary

Parametric number of Cores (N×M out-pixel)

8b activations, 2-8b weights

Siracusa: 16nm SoC, Tightly Coupled at MRAM Accelerator



	Vega [1]	Diana [2]	Marsellus [3]	[4]	[5]	*Siracusa*
Technology	22nm FDX	22nm FDX	22nm FDX	40nm	22nm	16nm FinFET
Area	10mm²	10.24mm²	8.7mm²	25mm²	8.76mm²	16mm²
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/mm²	21.2 GOPS/mm²	47.4 GOPS/mm²	N/A	58.3 GOPS/mm²	**65.2 GOPS/mm²**

[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]

Specialization in perspective

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

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

↓

ISA-based 10-20x →**1-5pJ (8bit)** → **XPULP**

↓

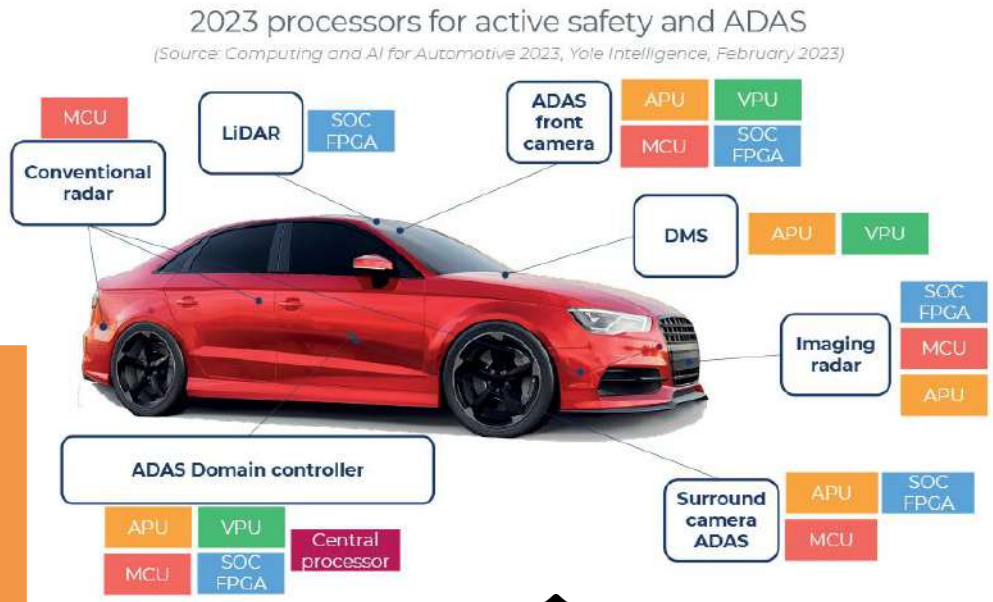
Configurable DP 10-20x → **20-100fJ (4bit)** → **RBE, NEUREKA**

↓

Highly specialized DP 10-20x →**1-5fJ (ternary)** → **CUTIE, SNN**

From Drones to Cars: Stepping up

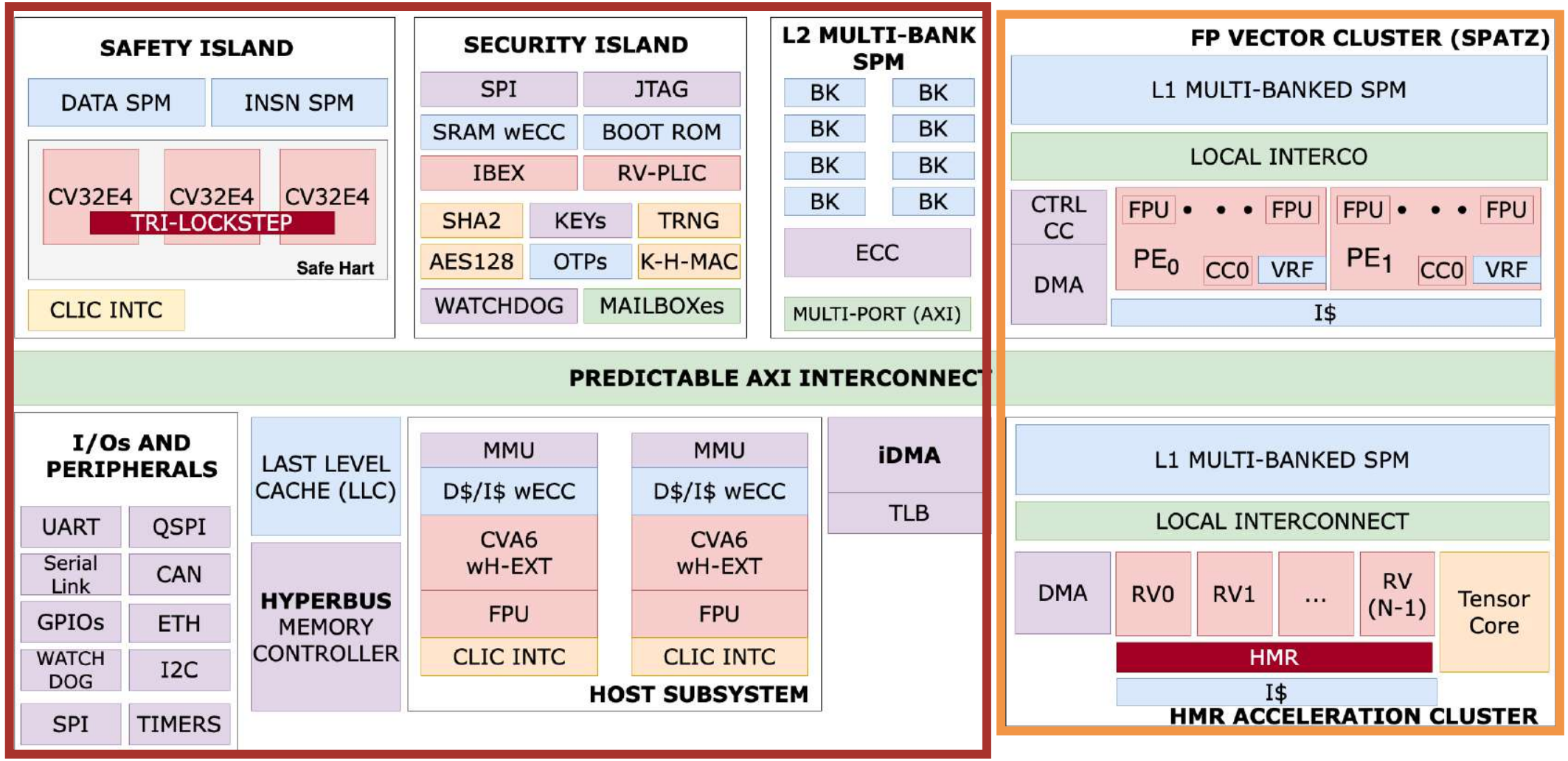
- **Microcontroller class of devices**
  - Infineon AURIX Family MCUs
  - **Control tasks, low-power sensor acquisition & data processing** Features: lockstepped **32-b HP TriCore CPU** , HW I/O monitor, dedicated accelerators
- **Powerful real-time architectures**
  - ST Stellar G Series (based on ARM Cortex-R cores)
  - **Domain controllers and zone-oriented ECUs**
  - Features: HW-based virtualization, Multi-core **Cortex-R52** (+NEON) cluster in split-lock, vast I/Os connectivity
- **Application class processors**
  - NXP i.MX 8 Family
  - **ADAS, Infotainment**
  - Features: Cortex-A53, **Cortex-A72,**HW Virtualization, **GPUs**



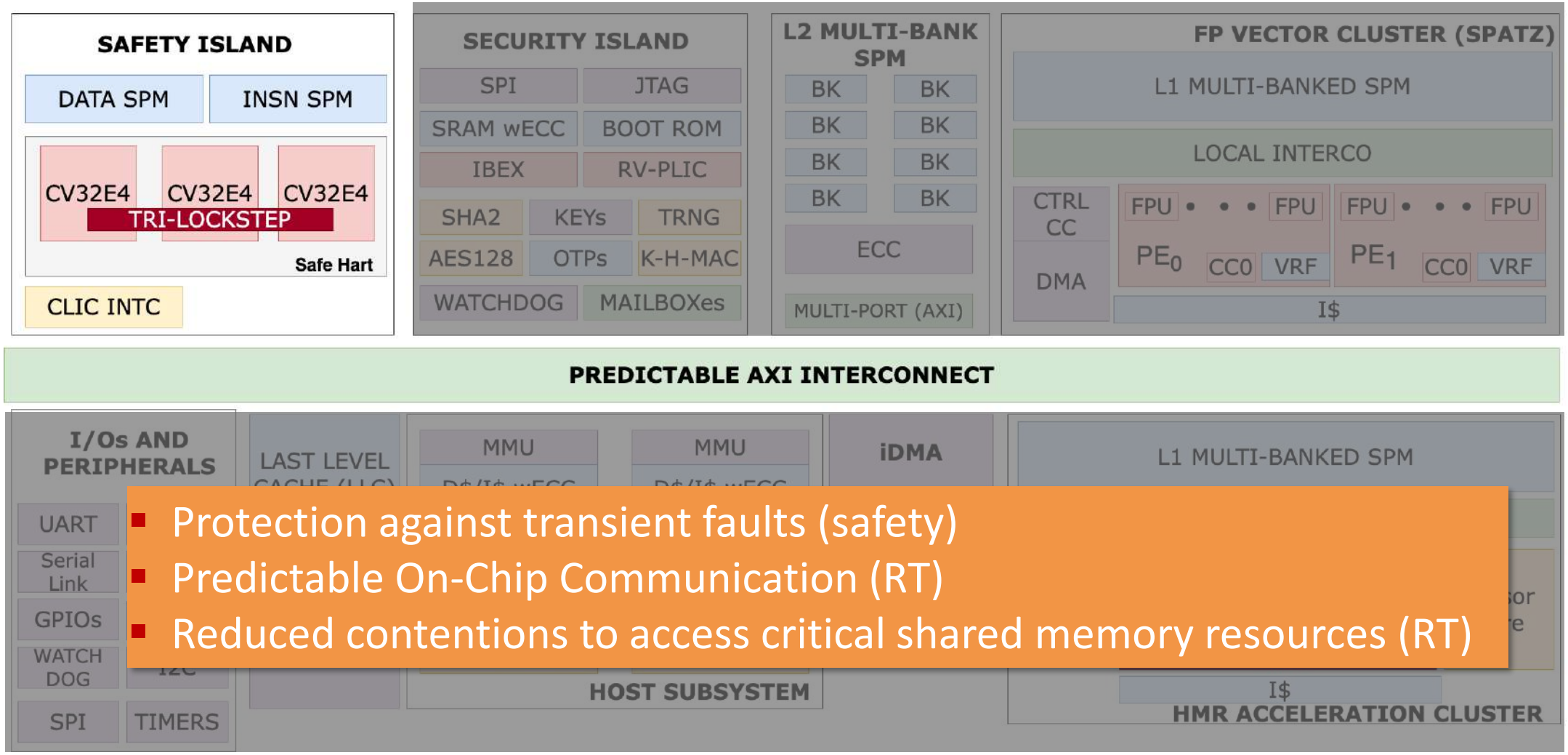
Carfield: 16nm SoC - Safety, Security, RT-Predictability

Main Computing and I/O System

Accelerators Domain

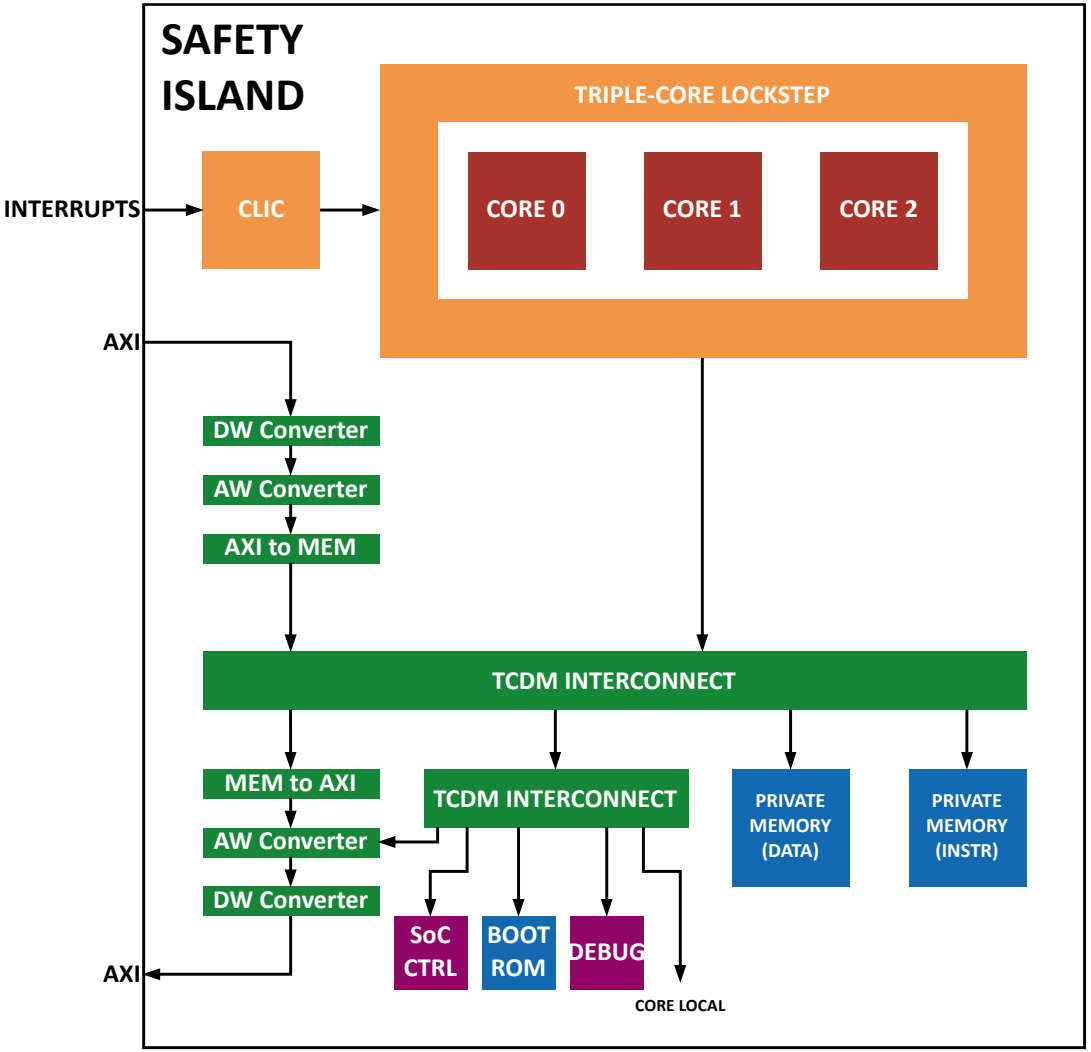


How Do We Handle Safety-Critical and Real-Time Tasks?



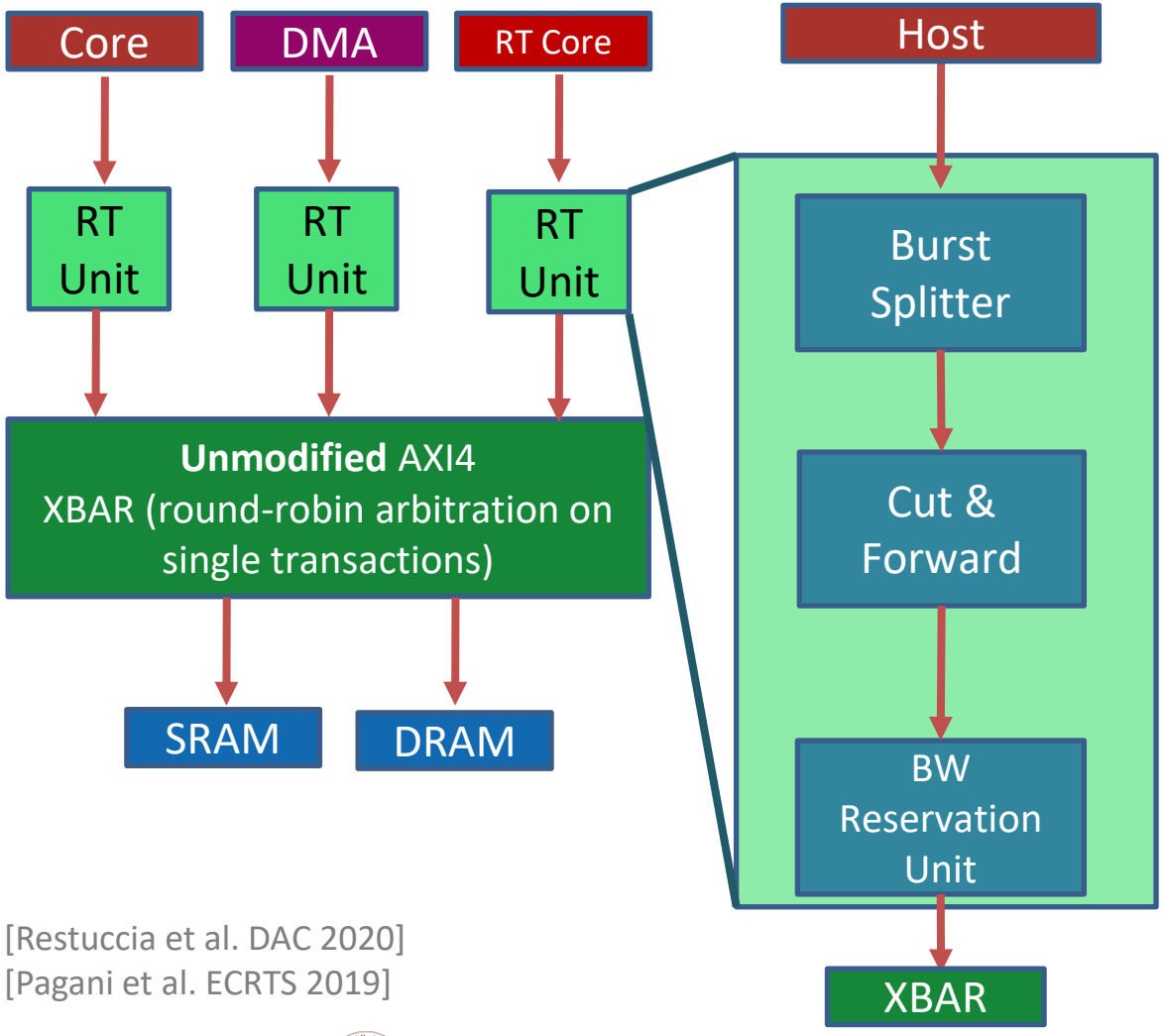
- Protection against transient faults (safety)
- Predictable On-Chip Communication (RT)
- Reduced contentions to access critical shared memory resources (RT)

Safety Island



- Safety-critical applications running on a RTOS
- **Three CV32E40 cores physically isolated operating in lockstep (single HART) and fast HW/SW recovery** from faults
- **ECC protected scratchpad memories** for instructions and data
- **Fast and Flexible Interrupts Handling** through RISC-V compliant CLIC controller
- AXI-4 port for in/out communication

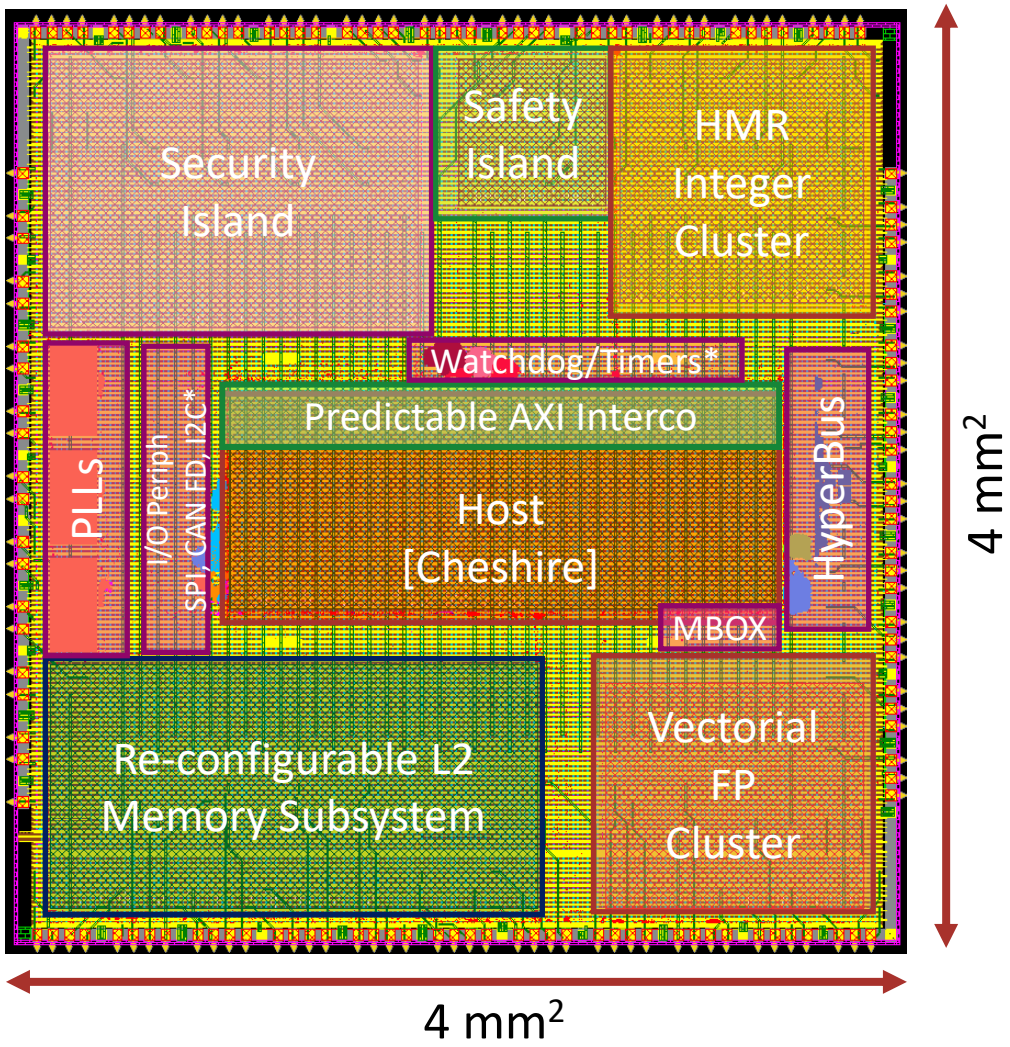
Predictable On-Chip Communication (AXI RT)



[Restuccia et al. DAC 2020]
[Pagani et al. ECRTS 2019]

- **AXI4 inherently unpredictable**
- **Minimally Intrusive Solution**
  - No huge buffering, limited additional logic
  - **Verified in systematic worst-case real-time analysis**
- **AXI Burst Splitter**
  - **Equalizes length of transactions** to avoid unfair BW distribution in round-robin scheme
- **AXI Cut & Forward**
  - Configurable **chunking unit** to avoid long transaction delays influencing access time to the XBAR
- **AXI Bandwidth Reservation Unit**
  - Predictably enforces a given **max nr of transactions per time period** (to each master)
  - **Per-address-range credit-based** mechanism
  - Periodically **refreshed** (or by user)

Carfield SoC Flooplan – Tested in August 2024

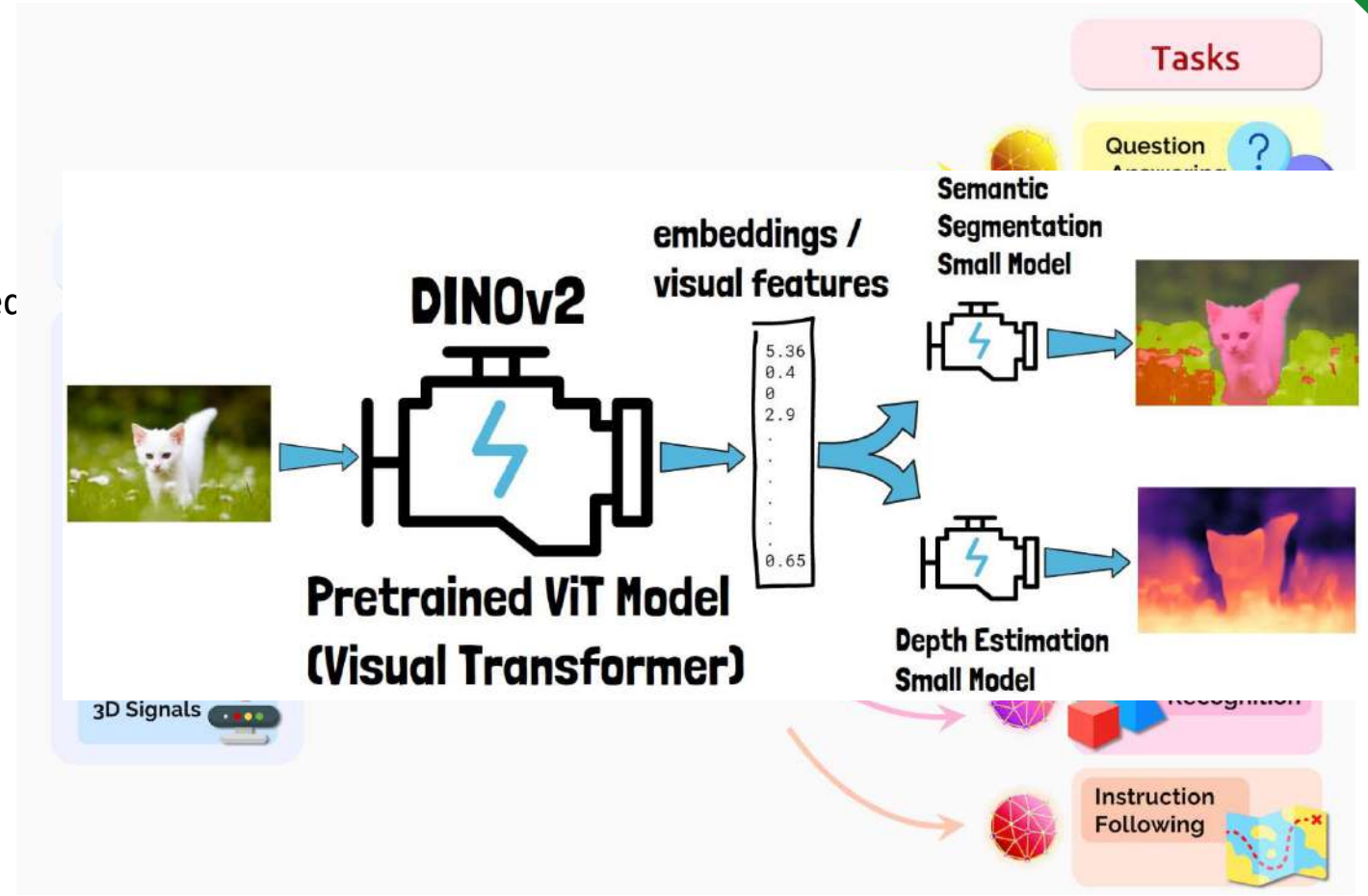


Modules marked with (*) are not in scale

- **Host**
  - Dual-Core 64-bit RISC-V processor; **2.45 mm²**; 600 MHz;
- **Security Island**
  - Low-power secure monitor; **1.94 mm²** ; 100 MHz;
- **Safety Island**
  - **0.42 mm²**; 500 MHz
- **Re-configurable L2 Memory Subsystem**
  - 1MB; **2.33 mm²**; 500 MHz
- **HMR Integer Cluster**
  - **1.17 mm²**; 500 MHz;
- **RVV FP Cluster**
  - **1.14 mm²**; 600 MHz;
- **Hyperbus**
  - 2 PHY, 2 Chips; 200 MHz; Max BW **400 MB/s**

Generative AI: The era of Foundation Models

- **Versatility and Multi-modality**
  - Natural language processing, computer vision, robotics, biology, …
- **Self-supervision, Fine-tuning**
  - Self-supervised training on large-scale unlabelec dataset
  - Fine-tune (few layers) on specific tasks with smaller labeled datasets.
- **Zero-shot specialization**
  - Prompt engineering for new tasks
- **Transformer Baseline**
  - Many variations
  - Ultra-fast evolution

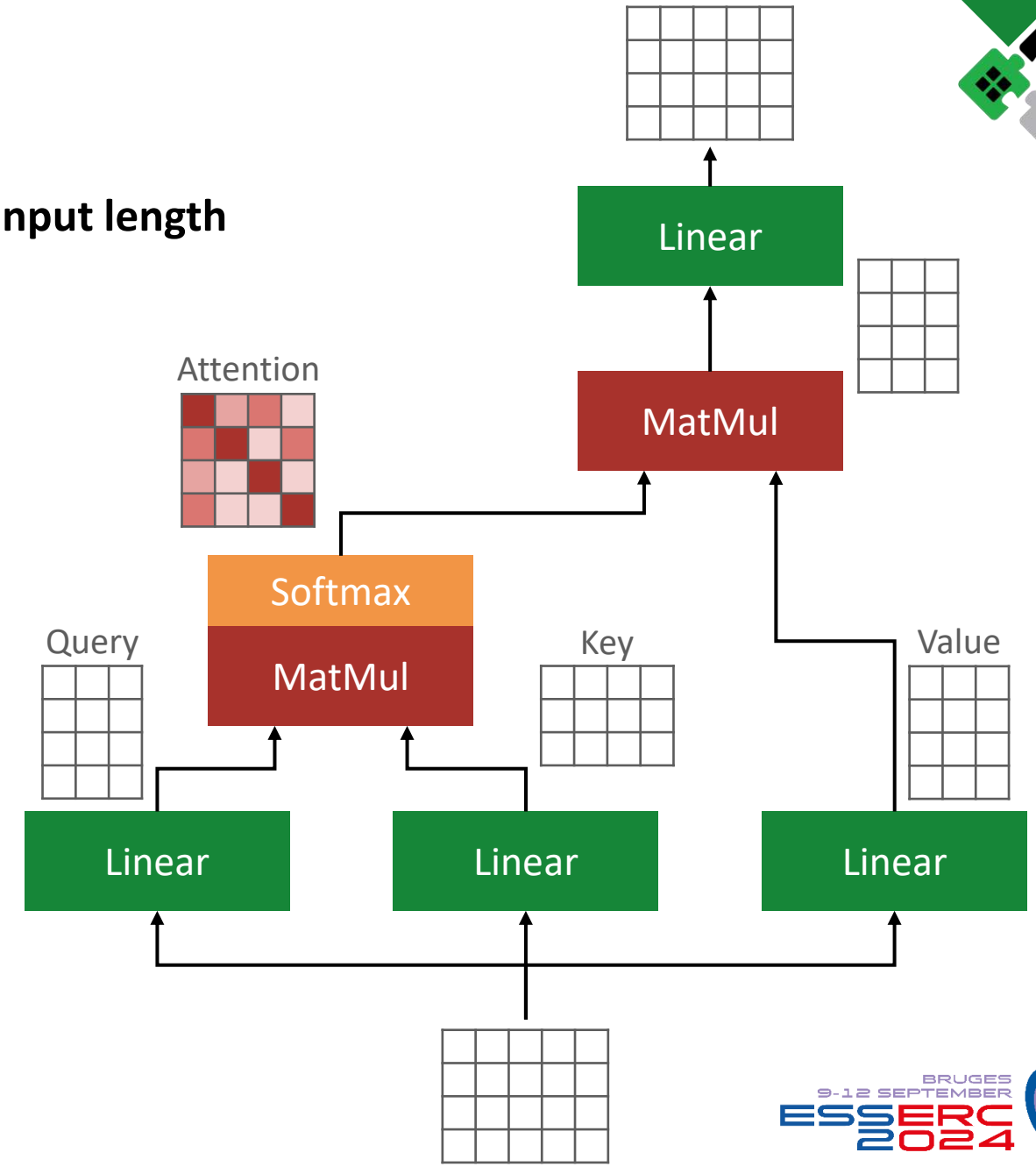


Bommasani, Rishi, et al. "**On the Opportunities and Risks of Foundation Models**." *Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI).*

Challenges in *Attention*

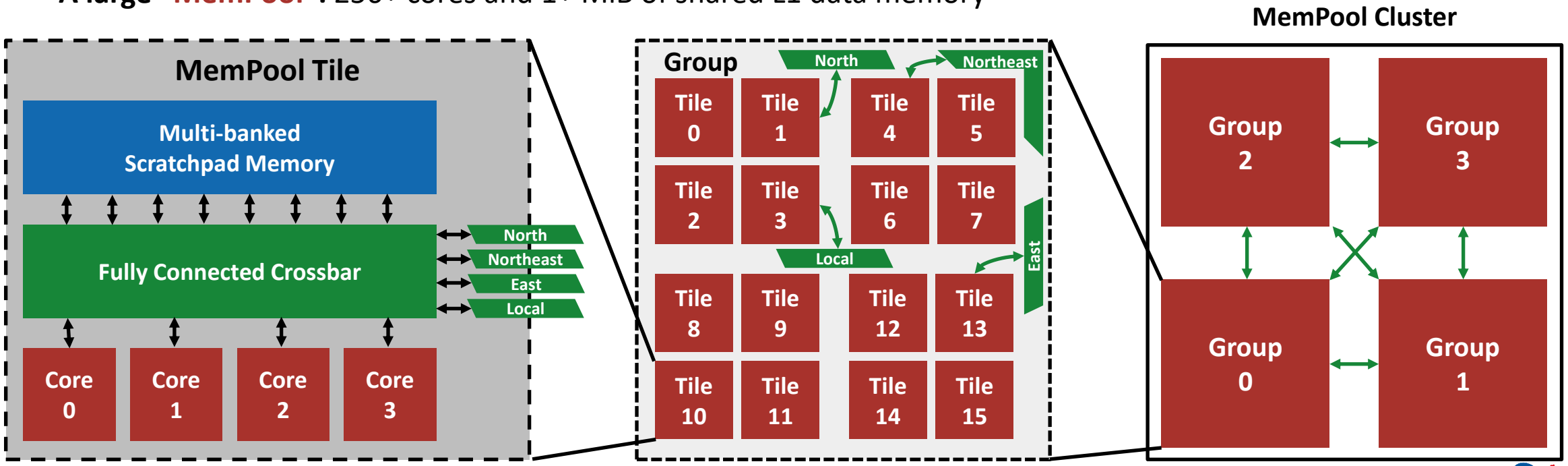
- **Attention matrix is a square matrix of order input length**
  - Computational complexity
  - Memory requirements
- **MatMul & Softmax dominate**

$$\mathrm{Softmax}(\mathbf{x})_i = \frac{e^{x_i - \max(\mathbf{x})}}{\sum_j^n e^{x_j - \max(\mathbf{x})}}$$



Matmul Benefits from Large Shared-L1 clusters

- **Why?**
  - Better global latency tolerance if $L1_{size} > 2\times L2_{latency} \times L2_{bandwidth}$ (Little's law + double buffer)
  - Smaller data partitioning overhead
  - Larger Compute/Boundary bandwidth ratio: N^3/N^2 for MMUL grows linearly with N!
- **A large "MemPool":** 256+ cores and 1+ MiB of shared L1 data memory



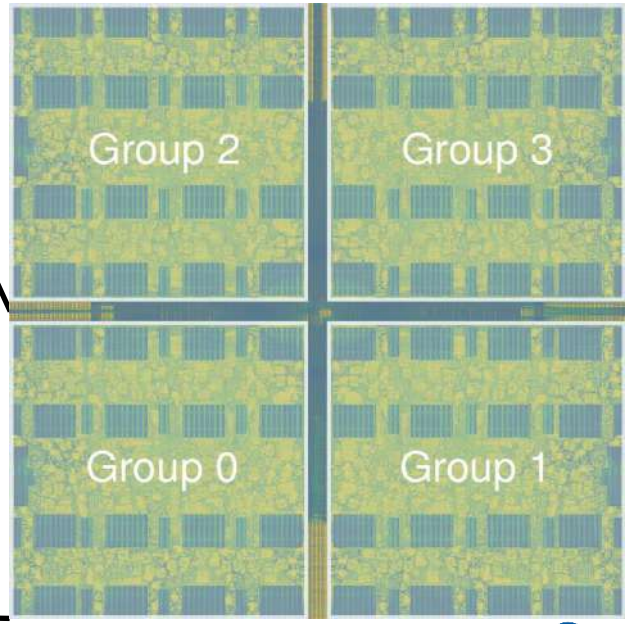
MemPool: A physical-aware design

- **A Scalable Manycore Architecture with Low-Latency Shared L1 Memory**
  - 256+ cores
  - 1+ MiB of shared L1 data memory
  - ≤ 8 cycle latency (Snitch can handle it)
- **Hierarchical design**
- **Implemented in GF22**
  - Targeting 500 MHz (SS/0.72V/125°C)
  - Reaching 600 MHz (TT/0.80V/25°C)
  - Targeting iso-frequency with PULP
- **Cluster area of 13 mm²**
  - 5 mm diagonal
  - Round trip in 5 cycles

MemPool
&
Terapool

MemPool Group

MemPool Cluster



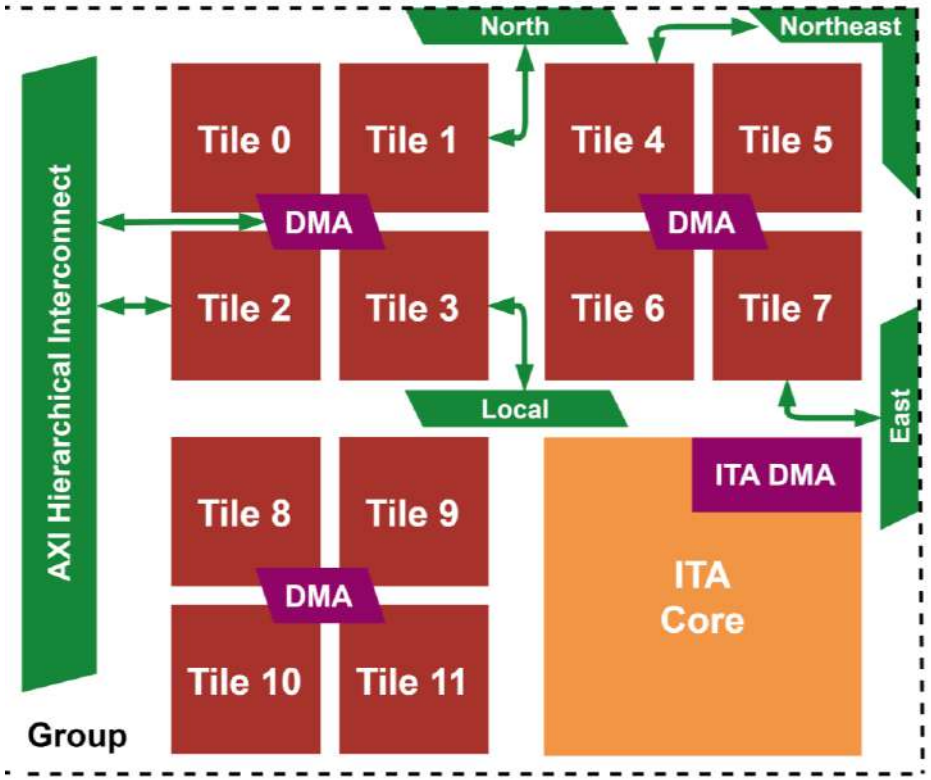
MemPool + Integer Transformer Accelerator (ITA)

Executing Transformer Networks

- Attention operation dominated by MatMul
- Flexible programmable accelerated architecture
  - 192 Snitch cores split into 48 tiles
  - 4 ITA cores to accelerate 8-bit attention operation
  - Automatic mapping of attention operation to ITA in Deeploy

Collaborative Execution

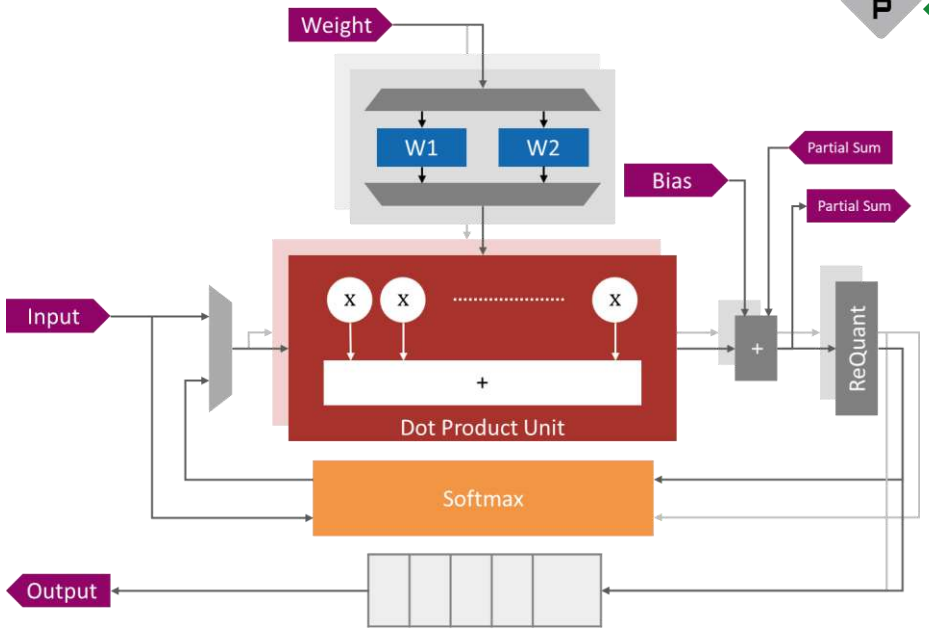
- Support convolutions and “exotic” operators on cores
- MatMul and Softmax accelerated with ITA
- Cores prepare activations for the next attention head
- Final head accumulation computed in cores



MemPool + Integer Transformer Accelerator (ITA)

Integer Attention Accelerator

- 8-bit inputs, weights & outputs
- Builtin data marshaling & pipelined operation
- Streaming partial Softmax **adding no additional latency**
- Fused Q× K^T, Softmax and A × V computation
- Support for hardware-aware Softmax approximation in QuantLib



Dot Product Units	Q	K	V	Q.K^T	A.V	Output
Softmax				DA / DI	EN	

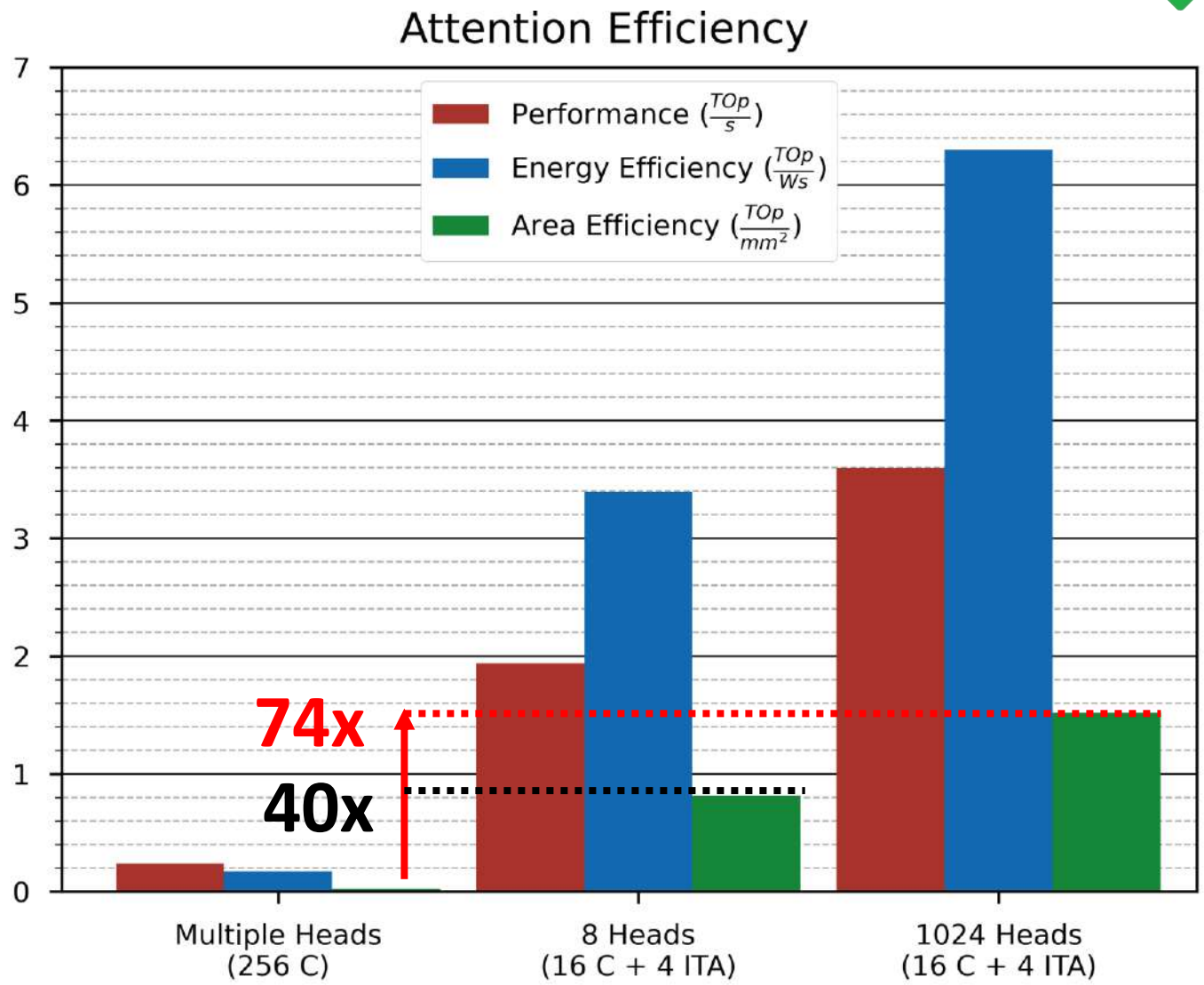
$$e^{a_i - a_{max_{n+1}}} = e^{a_i - a_{max_n}} \cdot e^{a_{max_n} - a_{max_{n+1}}}$$

Offloading Attention Operation to ITA

Performance
increase of **15x**

Energy Efficiency
increase of **36x**

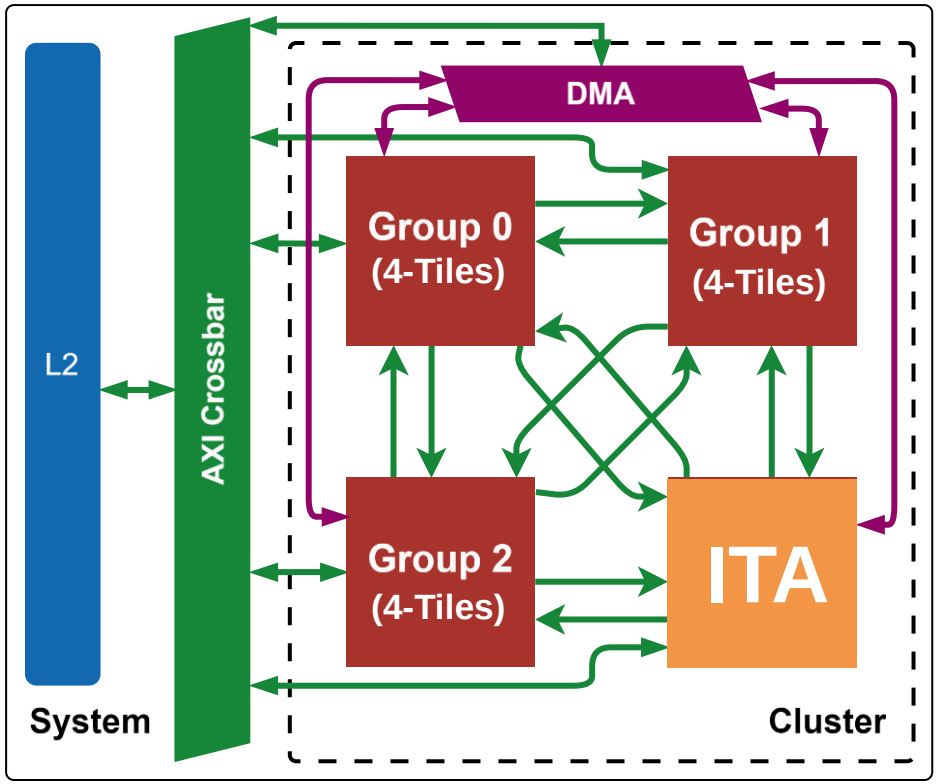
Area Efficiency
increase of **74x**



Heartstream: 12nm SoC – MemPool on Silicon

64 cores, 256kiB L1, peak 1.6MOPs @8b (TT-25°C-0.8V)

Boosting dot product, matmul + Softmax @8bit → ITA



Closing Thoughts

Embodied Gen-AI

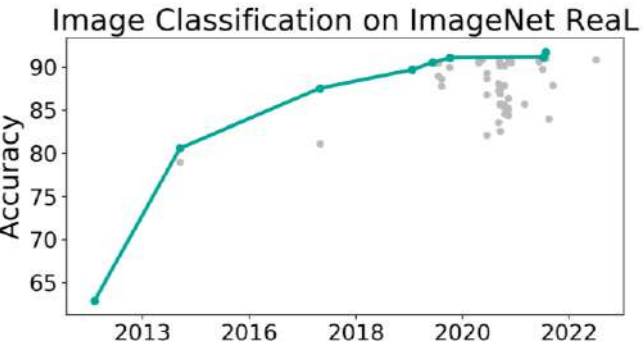
Prompted by @ashraf osman, AWS

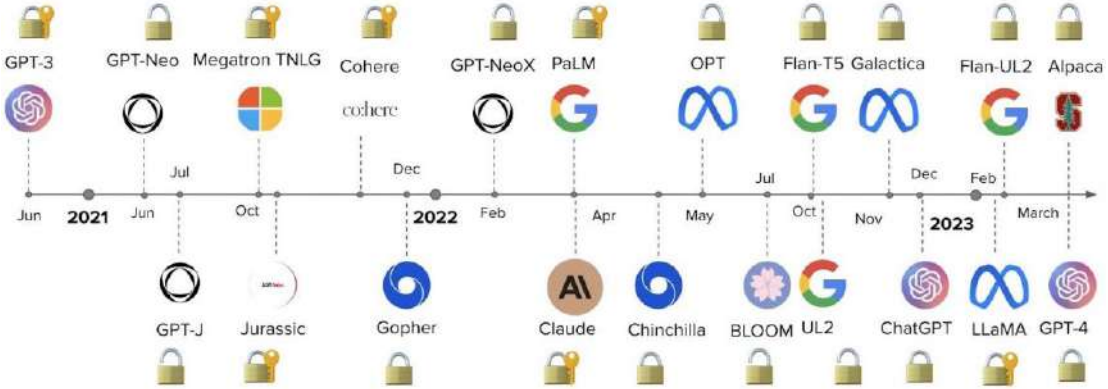
Gen-AI products @CES24

"A more complete picture is emerging of LLMs not as a chatbot, but the kernel process of a new Operating System"

Interactive, embodied intelligence: low-latency, edge inference

Perception → Gen.AI → Embodied Gen.AI





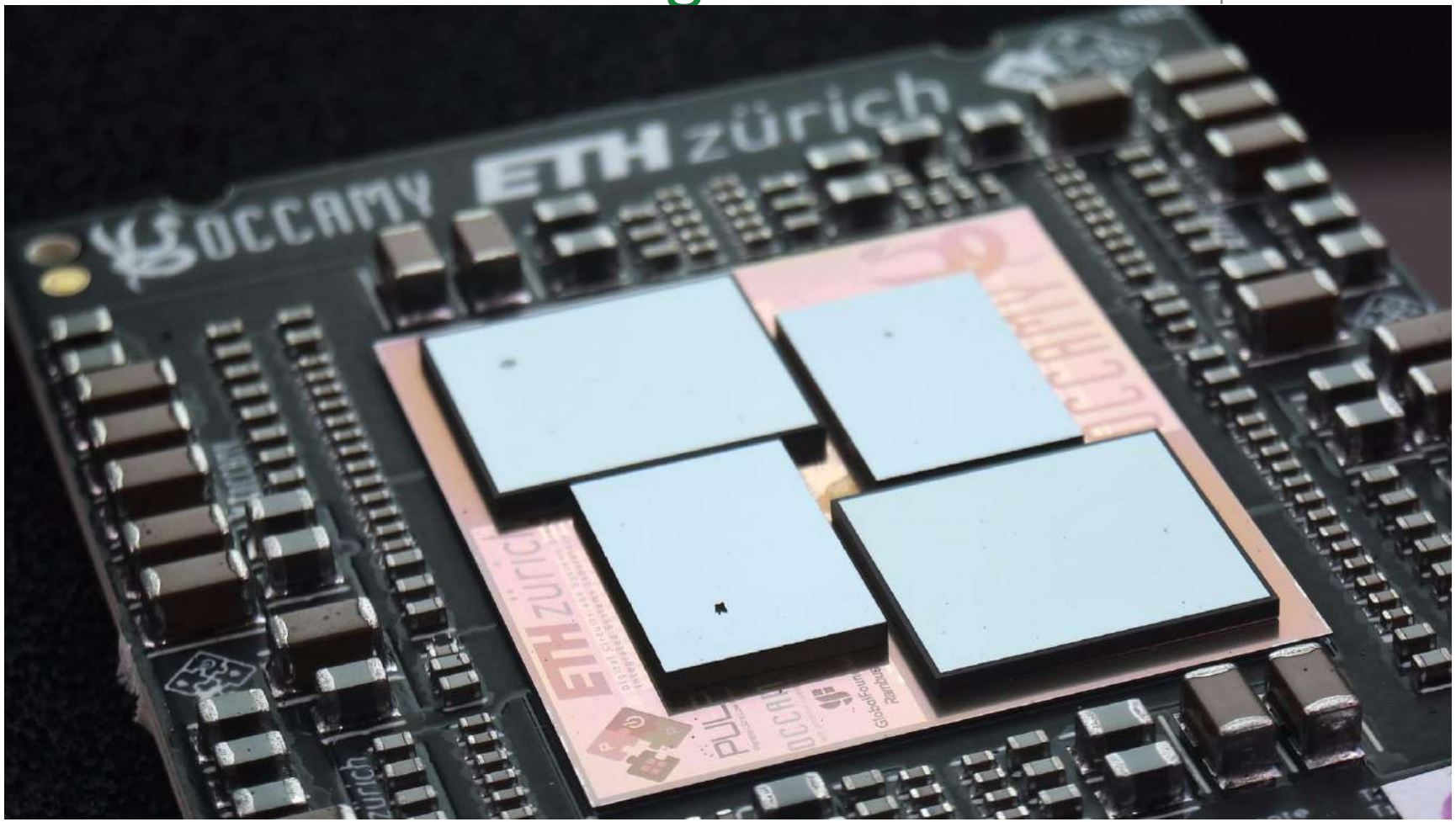
Precise

Interactive, creative

Efficient, RT-safe, secure

Embodied Gen.AI Challenge

OpenAI'23 arXiv:2303.08774



Challenge accepted; we are already on the right path, working on next gen circuits

PULP
Parallel Ultra Low Power

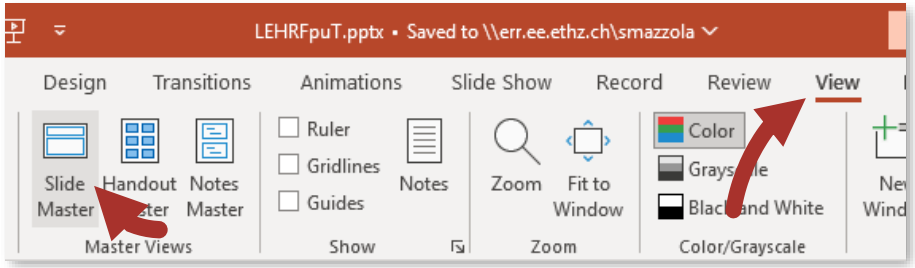
Thank You!

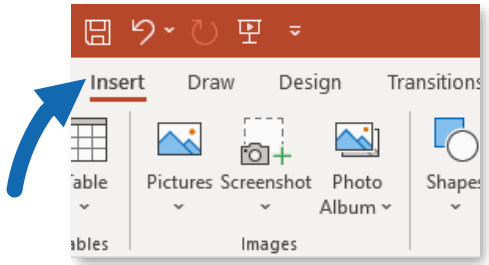
pulp-platform.org

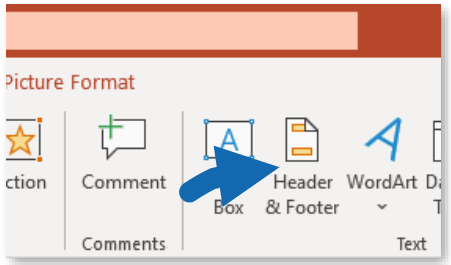
@pulp_platform

Customize template to your presentation

- **You can change the logo under** `View` → `Master`
  - Use the logo of your project
  - Leave it empty (or with Bianca) if you do not have an additional logo
- **Please adapt the footer**
  - Use `Insert` → `Header & Footer` to insert date and conference/presentation name
  - Page number will show up on the right
- **You can slightly shift left/right the footers** (in Master view) **to adjust to logo size/shape**

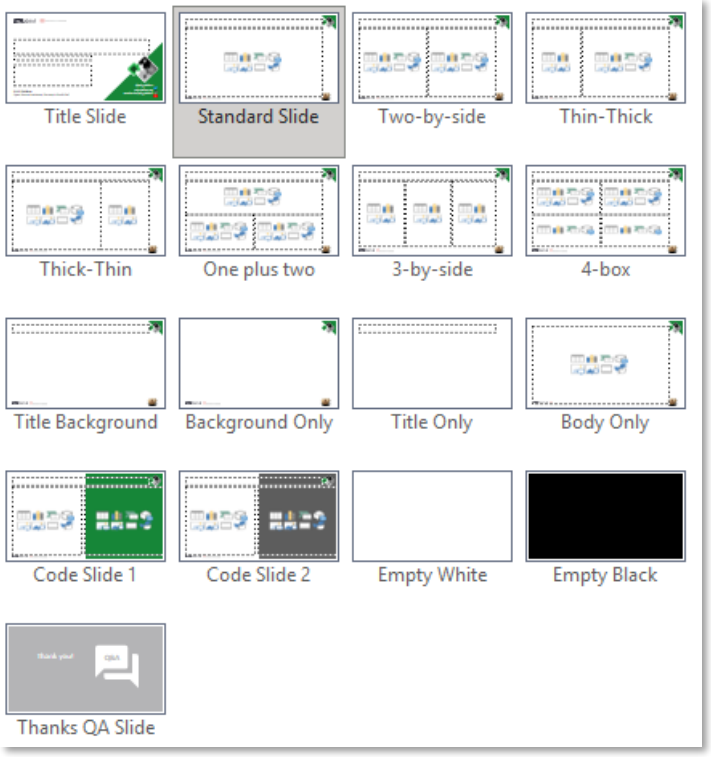






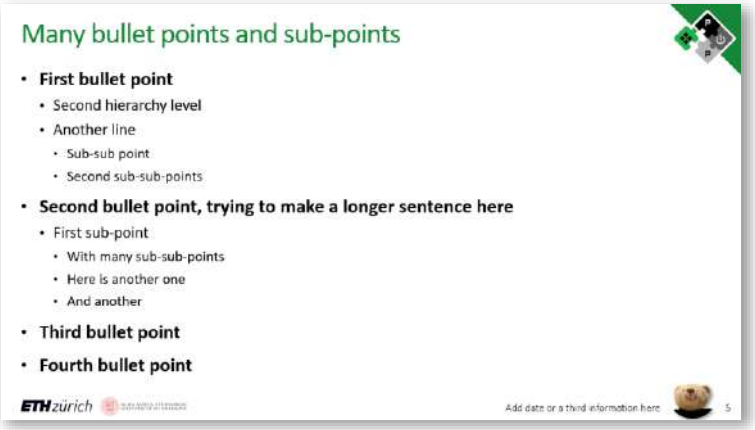
Default slide layouts

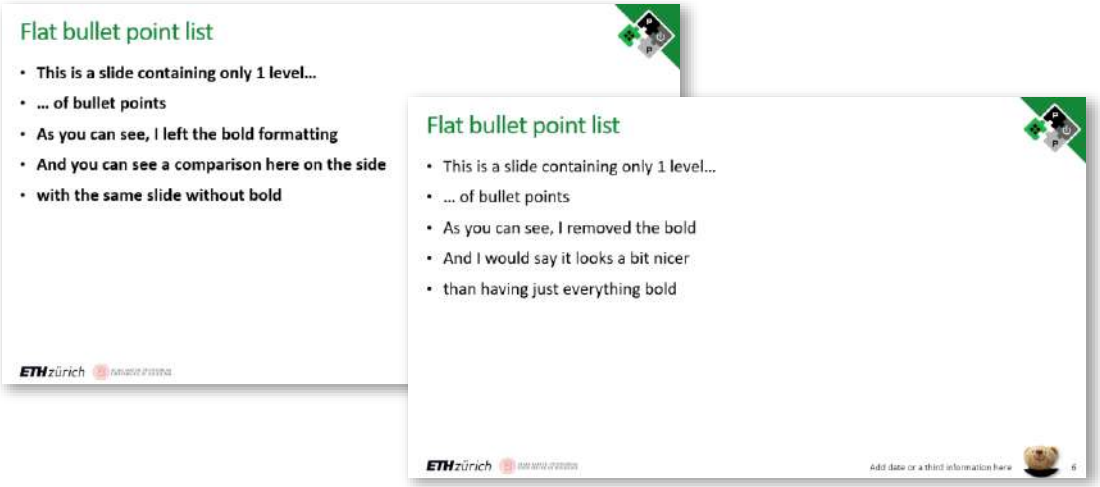
- **This is a PowerPoint template**
  - Other tools do not support all its features
- **Use the standard layouts as much as possible**
  - There are several default page styles
    - If you copy from other presentations, in most cases it will add new styles. If possible, change it to the existing styles, which makes life easy when sharing again later on.
  - To change template for a slide
    - Right click on your slide → Layout → Select preferred layout



Default fonts

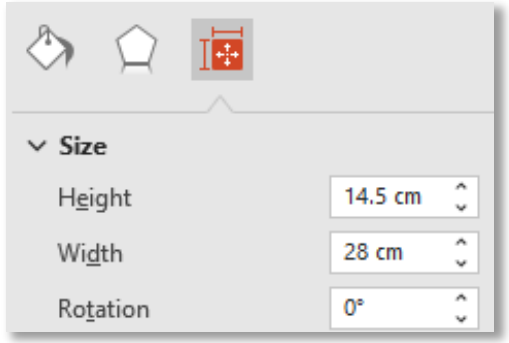
- **Default (body) font is Calibri, size 24pt**
  - Do not go smaller than 18pt in the body
  - You can use Calibri Light if you need some lighter writings
- **Default bullet point lists**
  - First hierarchy level is **bold** by default
  - This helps making the slide more readable when there is a lot of text
  - If you don't have sub-points, it is nicer to un-bold





The slide is 16:9 format and is 32cm x 18cm

- You can use shape properties (right click → `Size and Position…`) to enter sizes of boxes and images
- Default sizes are nice and even, to make alignment easy



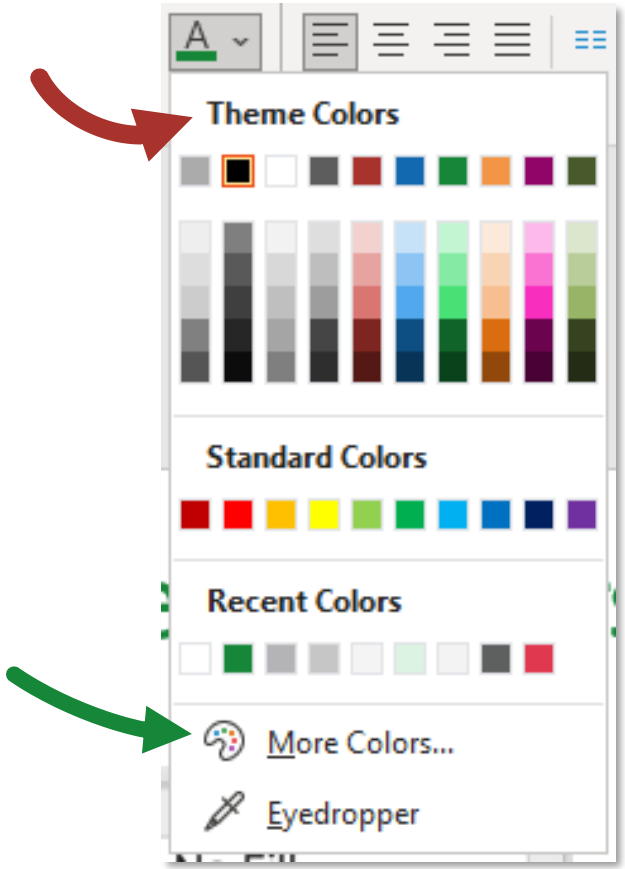
14.5 cm

28 cm

32 cm

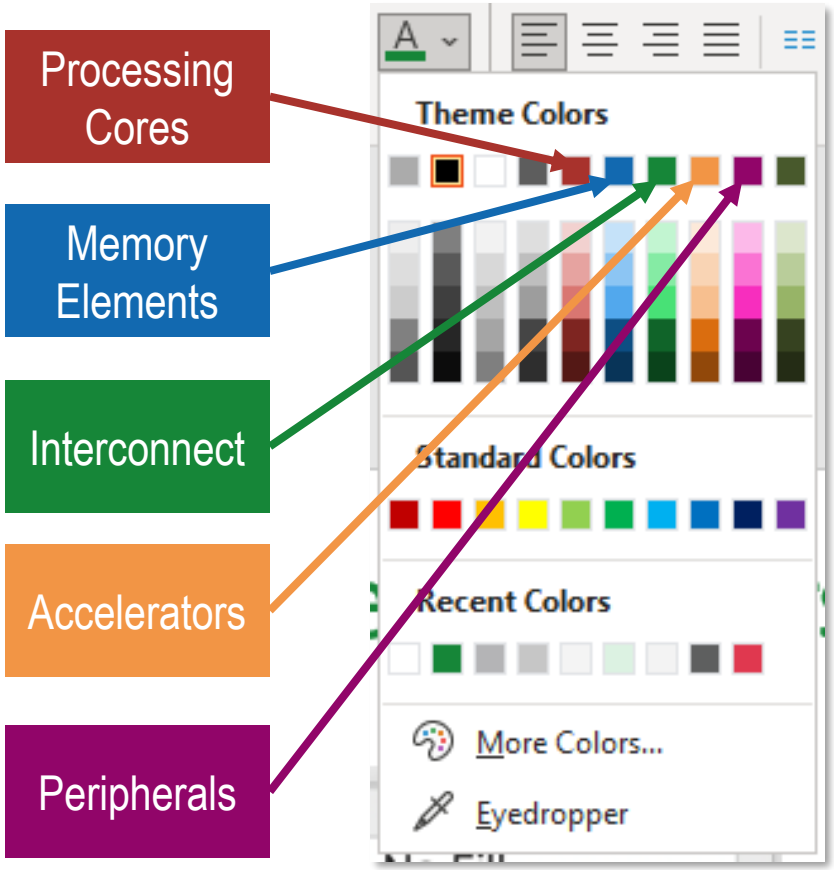
Try to limit to palette colors

- **If we make slight modifications to colors** (i.e., to help color blind people) **your slides will automatically get adjusted**
  - This is generally good, also helps with consistency
  - Problem is when the template colors differ significantly between different presentations, then using absolute colors creates less confusion
  - I suggest sticking to the template colors
- **In general, always pay attention to contrast**
  - Always keep in mind your slides will show on a projector (much lower contrast than a monitor)

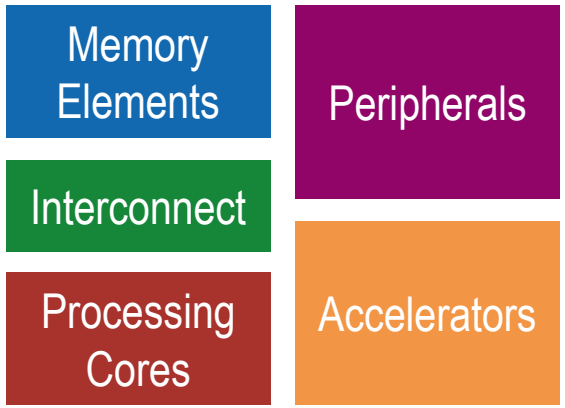


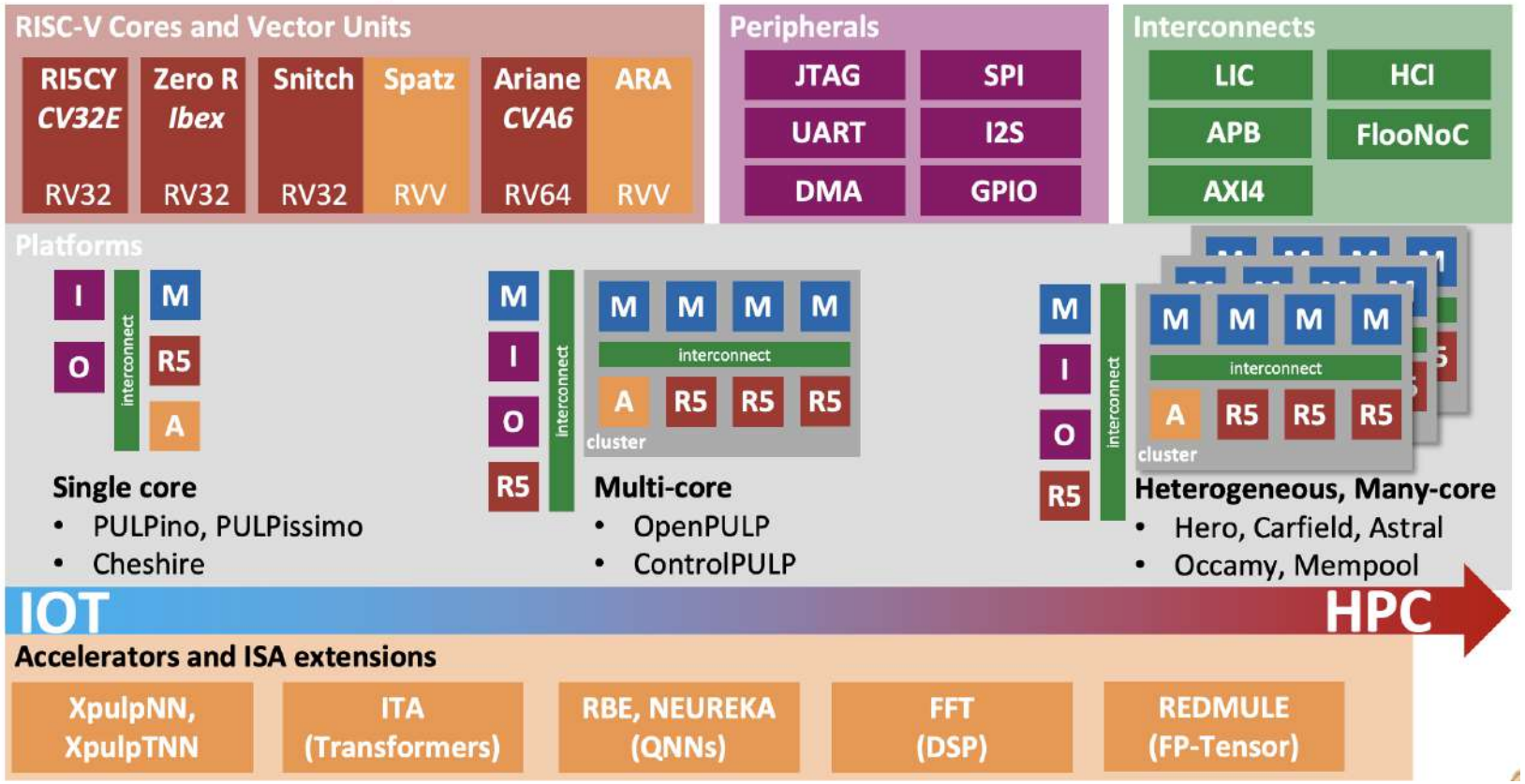
Our standard colors for architectural diagrams

- **All are template colors**
- **No outlines necessary**
- **Standard boxes should allow you to add text**
  - Default is not to autofit



Our standard colors for architectural diagrams



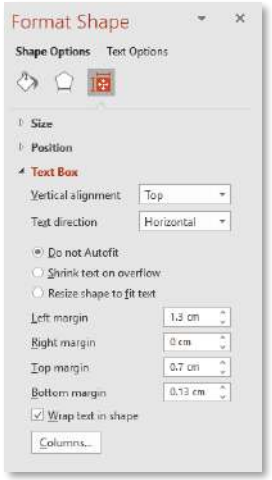


Useful examples

- **In the following slides you can find useful examples of**
  - Slides to show your code snippets
  - Highlight boxes and other useful objects compliant with the template style
    - They are for free :D Take as many as you want
    - Directly copy-paste in your presentation

Example slide for code

- **Suggestion**: keep the title short so that it fits in the white half of the slide
- **Use the**
  - lighter
  - shades
  - of the color palette
  - to highlight **keywords** in the code
- **You can also play with the margins to make it look nicer, based on your snippet**




```
module snitch (
  input logic clk_i,
  input logic rst_i,
  input logic [31:0] h_i,
  /// Interrupts
  input itrpts_t irq_i,
  /// Other I/O...
);
  // Module content
endmodule
```

Example slide for code

- If you like it more, you can also use this green layout (in case you don't plan to use a lot of code highlighting)

```
module snitch (
  input logic clk_i,
  input logic rst_i,
  input logic [31:0] h_i,
  /// Interrupts
  input itrpts_t irq_i,
  /// Other I/O...
);
  // Module content
endmodule
```

```
suggestion begin
  code highlighting not \
  really nice here
  /// Too little contrast
end
```

Callouts

Example of **callout** usage: can experiment with/without shadow

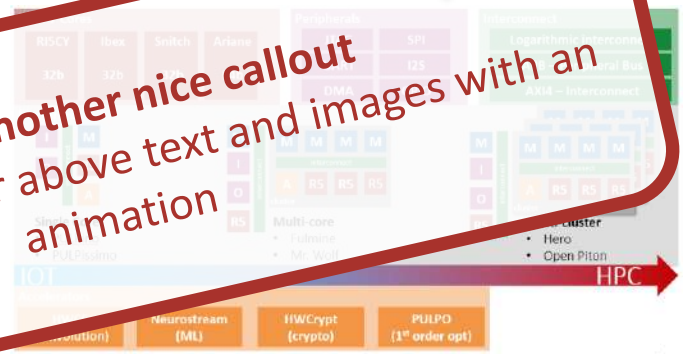
This slide provides **10x** performance improvement in presentations development

Another example

Highlight box to highlight e.g. your **contributions**

- Lorem ipsum dolor sit amet, consectetur adipiscin
- do eius
- Ut enim ad minim veniam, quis
- Duis aute irure dolor in
- cil um dolore eu fugiat nulla

This is another nice callout that can also appear above text and images with an animation



Source: https://pulp-platform.org

References

Duranton, Marc, et al. "**HiPEAC Vision 2021**: high performance embedded architecture and compilation." (2021).

Larger references

Benini, Luca, et al. "**A survey of design techniques for system-level dynamic power management**." *IEEE transactions on very large scale integration (VLSI) systems* 8, no. 3 (2000): 299-316.

Example of **highlight box** and **reference box** usage

Source: https://iis-nextcloud.ee.ethz.ch/example-figure-source-caption

Image source caption

[Burrello et al. TCOMP21]

"A survey of design techniques for system-level dynamic power management." *TVLSI2000.08*

Non-invasive references

Benini, Luca, et al. "**A survey of design techniques for system-level dynamic power management**."
IEEE transactions on very large scale integration (VLSI) systems 8, no. 3 (2000): 299-316.

Links

github.com/pulp-platform/snitch

github.com/pulp-platform/snitch

Play around with **size**, palette **colors** and **position** to adapt to your slide

github.com/pulp-platform/snitch

iis-nextcloud.ee.ethz.ch/f/1403905

iis-digital > presentations > templates > pulp_2022.potx

github.com/pulp-platform/snitch

https://docs.google.com/document/d/1EGpF9aboL5q40O287sNyed0ZmOmiQY61MY-BhL-bphs

The PULP Story

RISC-V Cores and Vector Units

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

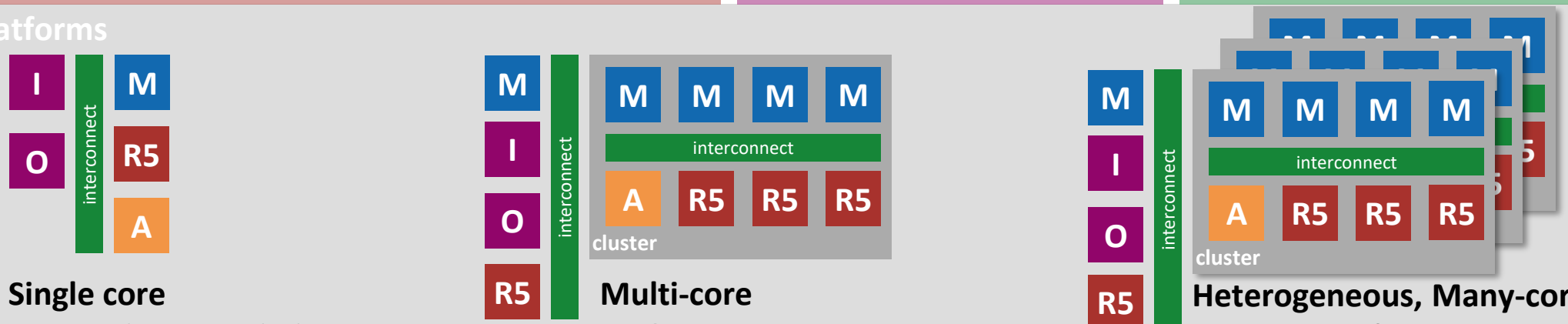
Peripherals

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)

Sergio Mazzola smazzola@iis.ee.ethz.ch
Frank K. Gürkaynak kgf@iis.ee.ethz.ch

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Università di Bologna
Viale del Risorgimento 2
Bologna, Italy

ETH zürich
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

WIP (don't use this ones)

10x Speedup w.r.t.
RV32IMC
(ISA does matter☺)

Better to have N× PEs running at optimum Energy than 1 PE running fast at low Energy efficiency

14.5x less instructions
at an extra 3% area cost
(~600GEs)

~15x latency and energy reduction for a barrier

TinyML challenge

AI capabilities in the power envelope of an MCU: 10-mW peak (1mW avg)

CHANGELOG

- **v1.0 (06.2022, smazzola)**
  - First version of the new template
- **v1.1 (07.2022, smazzola)**
  - Fix position of logo, footer, page number
  - Change palette green (position #7) from #24AF4B to #168638
  - Add slide layouts for code snippets
  - Add pre-made objects (callouts, references, links) ready to copy-paste
  - Various improvements to template instructions
- **v1.2 (03.2023, smazzola)**
  - Change default body font to Calibri (in place of Calibri Light)
  - Make level 0 of bullet point lists bold
  - Extend footer's length 
  - Substitute Samuel L. Jackson default logo with Bianca
  - Enhance closing slide with author names and contact points
  - Unify all slide types (standard, code, blank) under one single Master Slide
  - Re-organize slide for default elements to copy-paste, add new callouts and links
  - Various improvements to template instructions
- **v1.3 (03.2024, kgf)**
  - Adapted the PULP diagram
  - Tried to adapt some of the defaults
- **v1.4 (03.2024, fconti)**
  - Updated UNIBO logo
- **v1.5 (05.2024, fconti, kgf, lbenini)**
  - Updated IP picture