

**PULP PLATFORM**

Open Source Hardware, the way it should be!

# ***From Nano-Drones to Cars***

*A RISC-V Open Platform for Next-Generation Autonomous Vehicles*

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



European Research Council



**EuroHPC**  
Joint Undertaking



**ETH** zürich



<http://pulp-platform.org>



[@pulp\\_platform](https://twitter.com/pulp_platform)



[https://www.youtube.com/pulp\\_platform](https://www.youtube.com/pulp_platform)

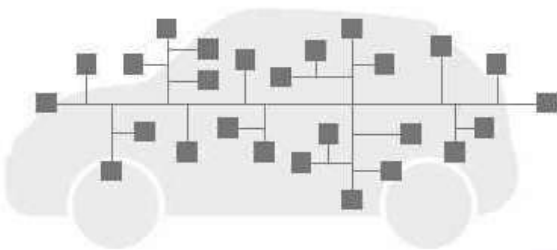
# Computing in Automotive

## MEGATREND – E/E ARCHITECTURE TRANSFORMATION

Today, 1500 chips/car → by 2030, rise to 3000

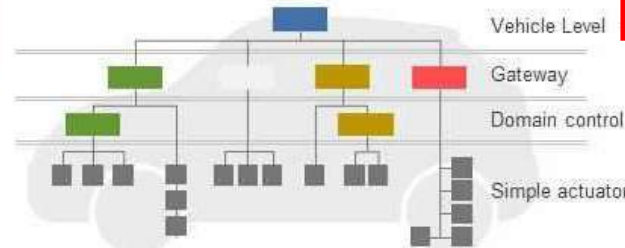
(EU working document – Chips Act for Europe)

TODAY - Distributed architecture



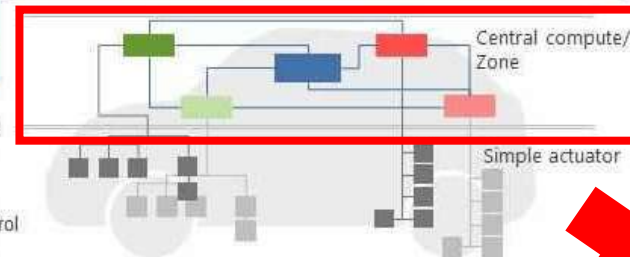
[renesas,bosh]

TOMORROW - Centralized architecture



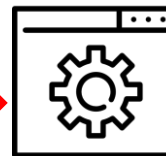
- Scalable & Easy to plug-in hardware
- Upgradable & Reusable software
- Safe, redundant and streamlined network

FUTURE - Zone architecture



**Cars need lots of computing**

**Computing**

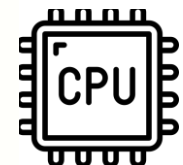


**SW**

```
int a = 1;  
int b = N;  
do  
  a b;  
  b - 1;  
} while (b != 0)
```

**ISA**

**HW**



# Open Source (Computing) Hardware

Hardware whose design is made publicly available so that anyone can study, modify, distribute, make, and sell the design or hardware based on that design

(source: [Open Source Hardware \(OSHW\) Statement of Principles 1.0](#))

Very wide definition – includes PCBs and “makers” stuff

I will focus on Open Source **Computing** Hardware (**OSCHW**)

# OSCHW Needs an Open Source ISA



**A modern, open, free ISA, extensible by construction**  
**Endorsed by 3000+ Members & large SW ecosystem**  
**An open ISA is a prerequisite!**



**Risc-V international moved to Zürich, CH**  
**for international neutrality, 1/3 of members from EU**



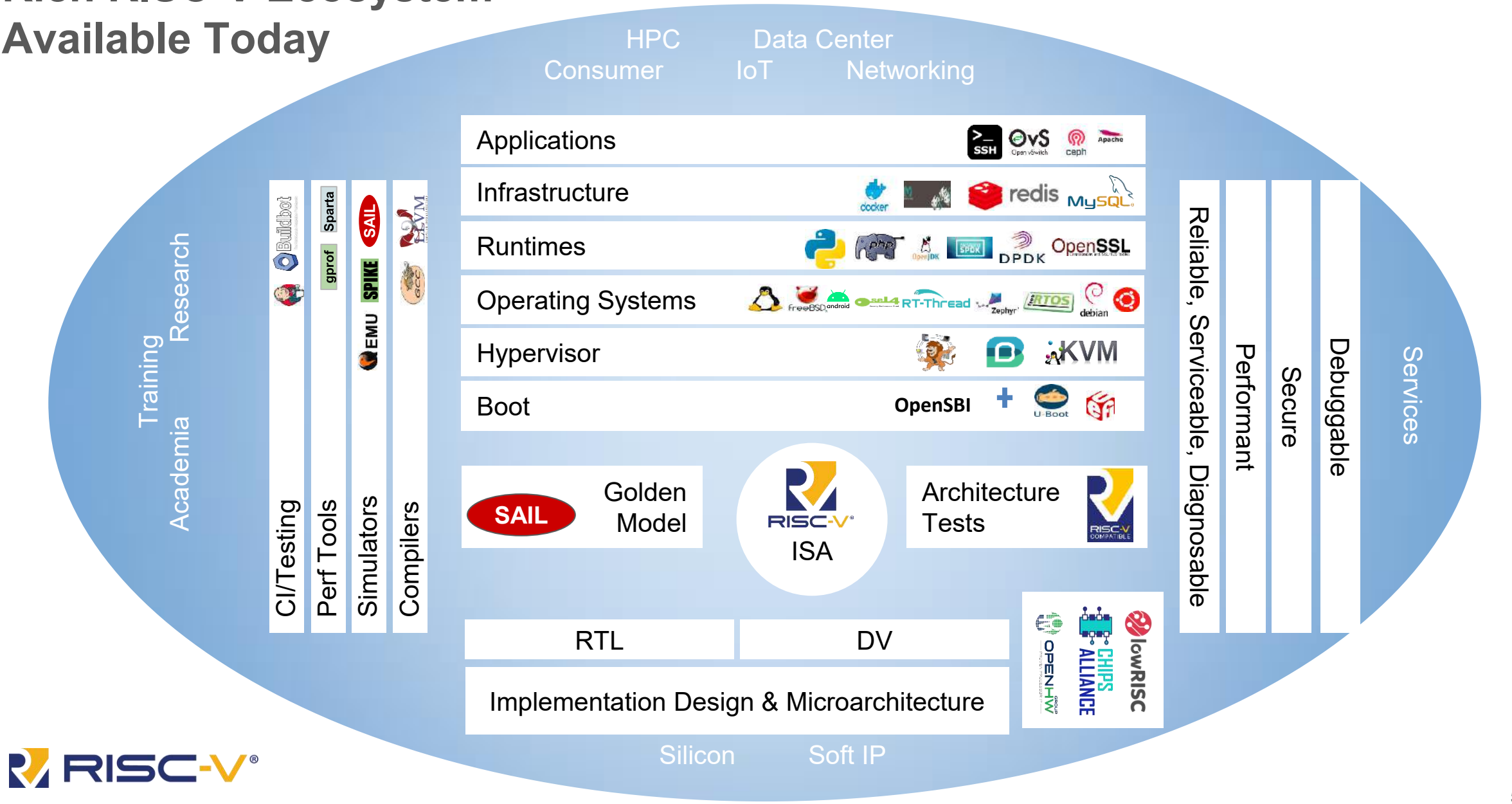
**ETH zürich**

wide in RV32I, 64 in RV64I, and 128 in RV128I (x0=0). RV64I/128I add 10 instructions for the wider formats. The RV1 base of <50 classic integer RISC instructions is required. Every 16-bit RVC instruction matches an existing 32-bit RV1 instruction. See risc.org.

width matches the widest precision, and a floating-point control and status register fcsr. Each larger address adds some instructions: 4 for RVM, 11 for RVA, and 6 each for RVF/DQ. Using regex notation, {} means set, so L{D|Q} is both LD and LQ. See risc.org. (8/21/15 revision)

# Rich RISC-V Ecosystem

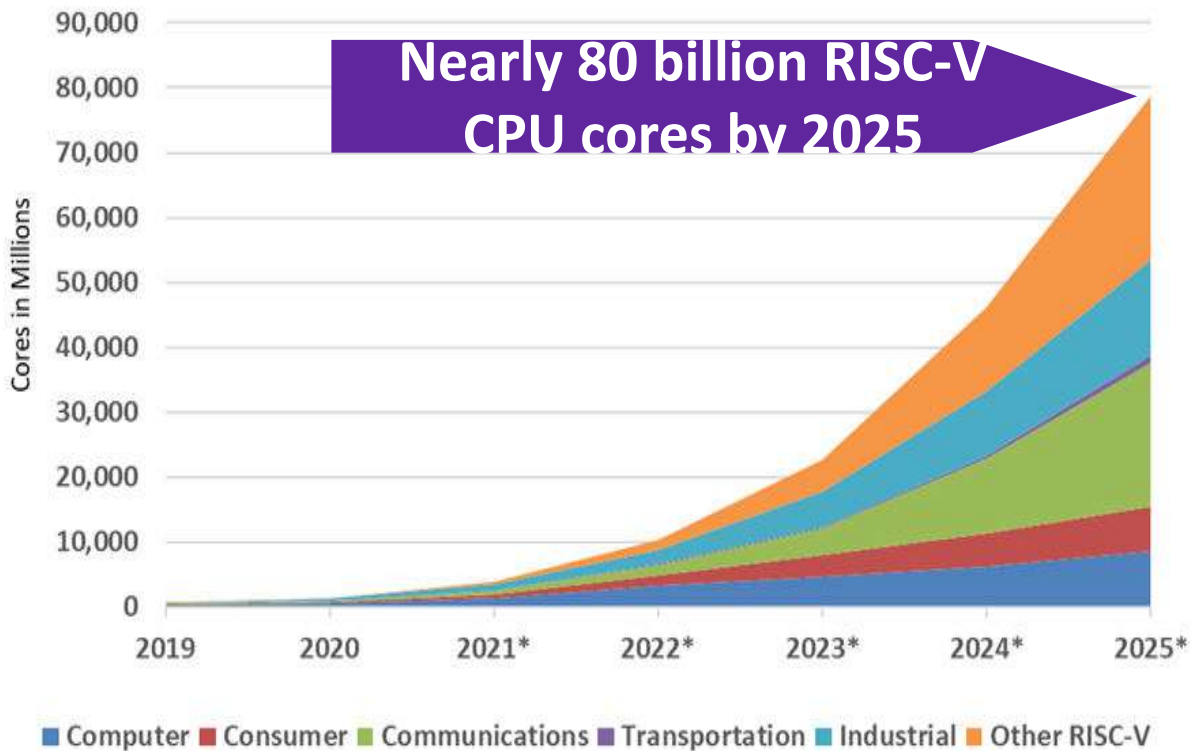
## Available Today





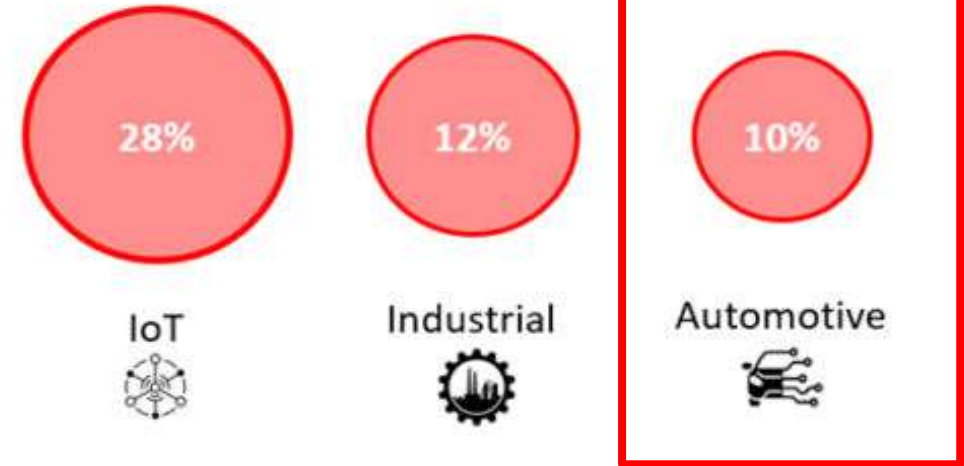
# RISC-V CPU Cores Market

Grows 114.9% CAGR 2025  
>14% of all CPU cores by 2025



Source: Semico Research Corp, March 2021

## RISC-V Penetration Rate by 2025



Source: Counterpoint Research, September 2021

“The rise of RISC-V cannot be ignored...  
RISC-V will shake up the \$8.6 Billion  
semiconductor IP market.”

-- William Li, Counterpoint Research

**MobileEye** EyeQ Ultra vision advanced driver assist systems chips for 176 trillion ops per second with 12 RISC-V CPU cores.

**Andes** ISO 26262 Functional Safety ASIL D Dev Process Certification for RISC-V embedded safety with Andes processors. Used in Renesas R9A02G020 MCU ASSP for motor control systems

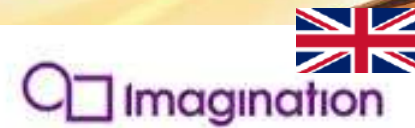
**Renesas and SiFive** partner on high-end automotive applications. SiFive licenses RISC-V core IP to Renesas (Automotive™ E6-A, X280-A, and S7-A for infotainment, cockpit, connectivity, ADAS)

**Imagination Technologies** GPU linked by a RISC-V core for ASIL-B level designs with ISO 26262 safety critical certification.

**Codasip** Security and safety embedded by design in processor design automation tools



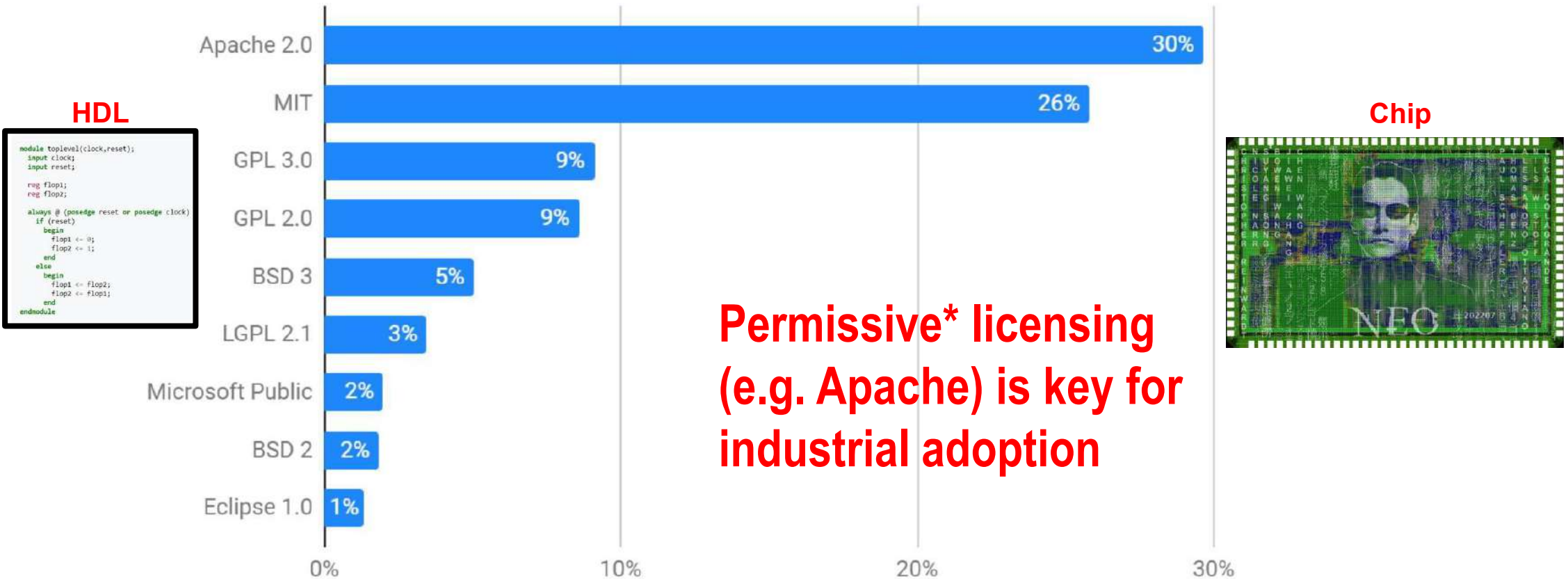
# Automotive



## What about OSCHW?

# Open ISA < OSCHW... But what is OSCHW, then?

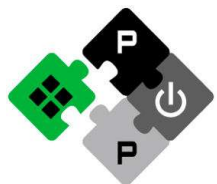
- The first stage of the silicon production pipeline → **HDL source code**
- Later stages contain closed IP of various actors → not open source by default



**Permissive\* licensing  
(e.g. Apache) is key for  
industrial adoption**



# Cores, and more! Many IPs for a Platform



## RISC-V Cores

RI5CY

32b

Ibex

32b

Snitch

32b

Ariane  
+ Ara  
64b

## Peripherals

JTAG

UART

DMA

SPI

I2S

GPIO

## Interconnect

Logarithmic interconnect

APB – Peripheral Bus

AXI4 – Interconnect

## Platforms

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

**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

RV

### Multi-core

- Open-PULP
- PULP-PM

R5

### Multi-cluster

- Hero
- Occamy

**IOT**

**HPC**

## Accelerators

HWCE  
(convolution)

Neurostream  
(ML)

HWCrypt  
(crypto)

PULPO  
(1<sup>st</sup> ord. opt)

# HWG: A Fast-Growing Precompetitive Industrial Ecosystem



**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) where HW and SW designers collaborate in the development of open-source cores, related IP, tools and SW such as the **Core-V** family
- **OpenHW Group** provides an infrastructure for hosting high quality open-source HW developments in line with industry best practices.



**80+ members!**



**ETH** zürich

# Start Small: Open Platform for Autonomous Nano-Drones

## Advanced autonomous drone

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



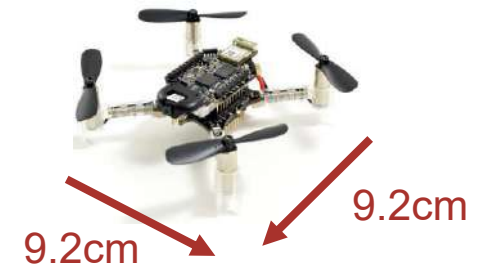
<https://www.skydio.com/skydio-2-plus>



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



## Nano-drone

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



- Smaller form factor of 0.008m<sup>2</sup>
- Weight **27g (30X lighter)** 
- Battery capacity **250mAh (20X smaller)** 

**Can we fit sufficient intelligence in a 30X smaller payload, 20X lower energy budget?**

# Achieving True Autonomy on Nano-UAVs

Multiple, complex, heterogeneous tasks at high speed and robustness **fully on board**

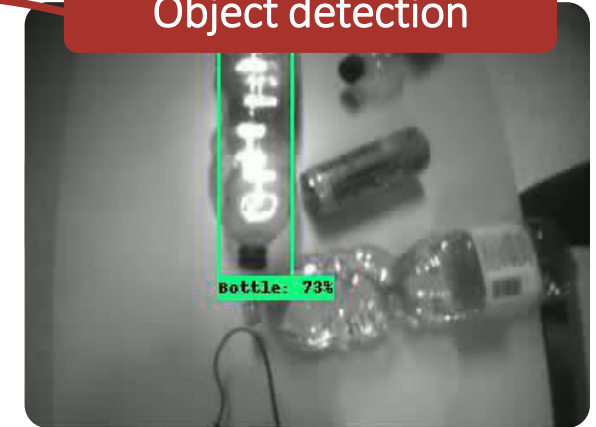
Obstacle avoidance & Navigation



Environment exploration



Object detection

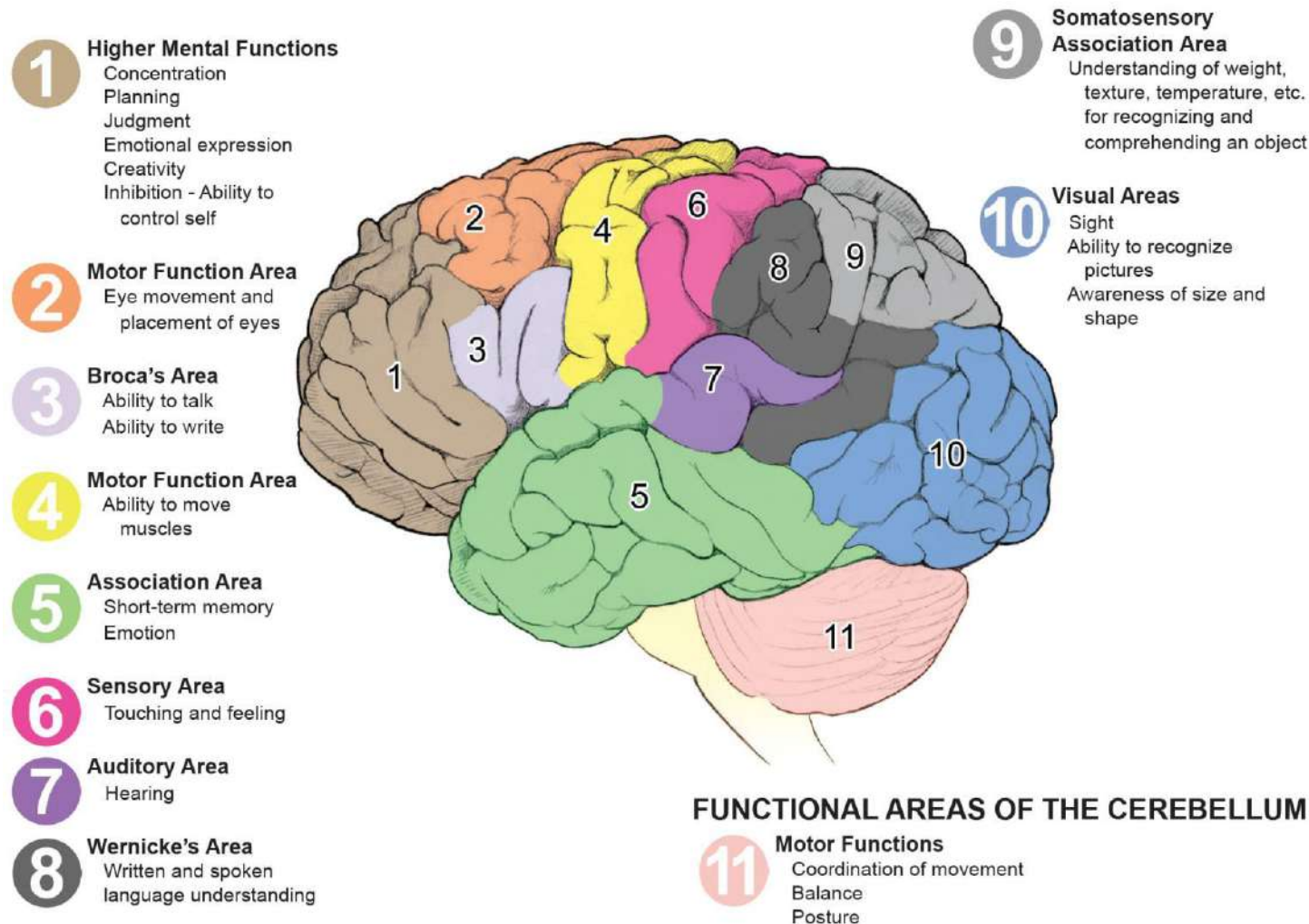


**Multi-GOPS workload at extreme efficiency  $\rightarrow P_{\max}$  100mW**



# Multiple Heterogeneous Accelerators

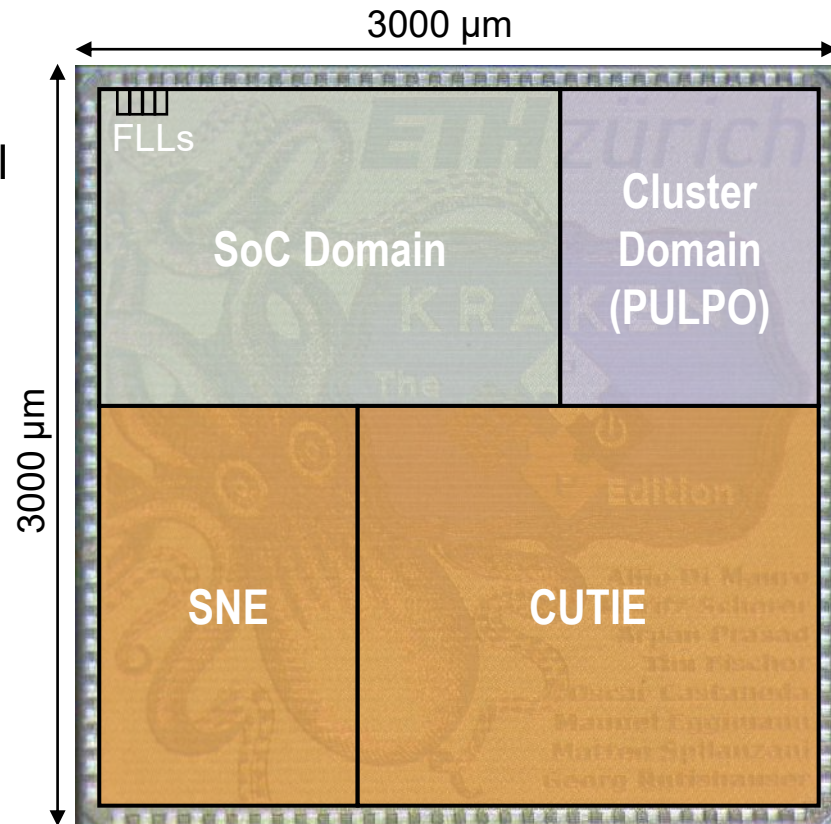
**Brain-inspired:** Multiple areas, different structure different function!



# Multiple Heterogeneous Accelerators

## The *Kraken*: an “Extreme Edge” Brain

- RISC-V Cluster (8 Cores + 1)
- CUTIE – dense ternary neural network accelerator
- SNE – energy-proportional spiking neural network accelerator



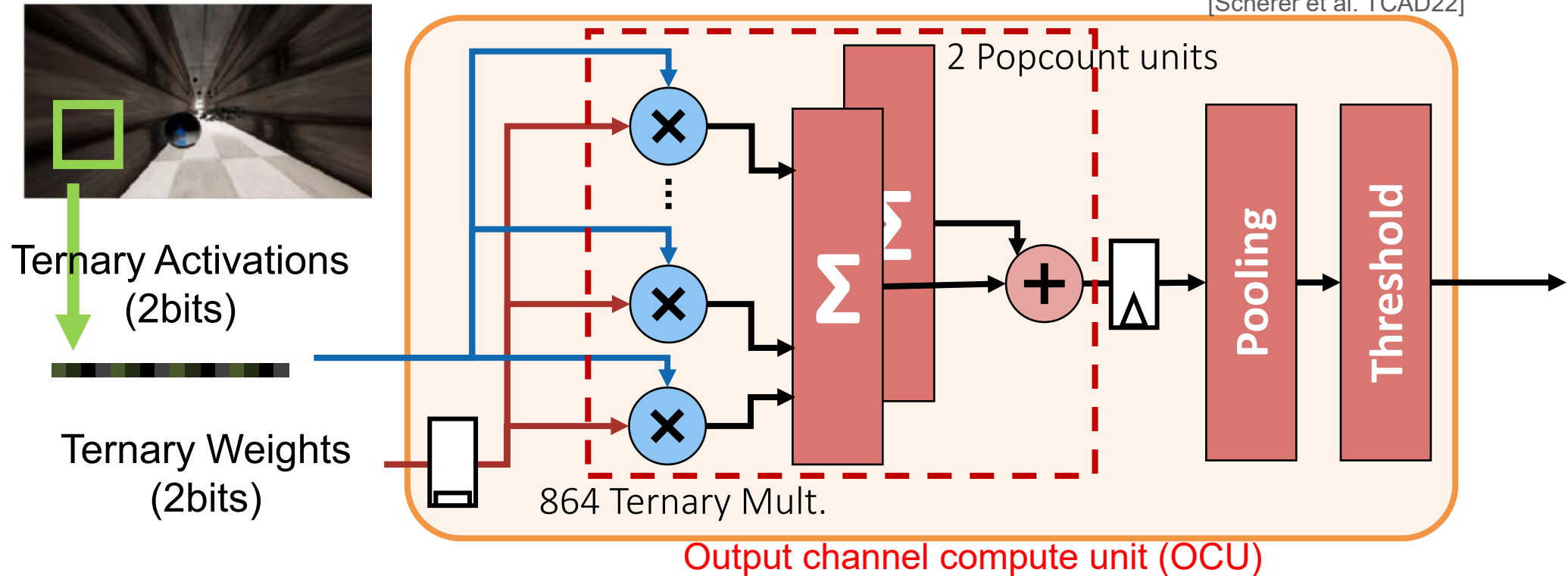
Technology	22 nm FDSOI
Chip Area	9 mm <sup>2</sup>
SRAM SoC	1 MB
SRAM Cluster	128 KB
VDD range	0.55 V - 0.8 V
Cluster Freq	~370MHz
SNE Freq	~250MHz
CUTIE Freq	~140MHz

[Di Mauro HotChips22]



# CUTIE: Minimize Switching Activity & Data Movement

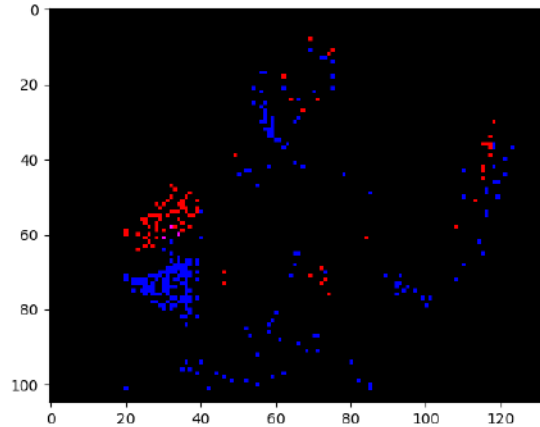
[Scherer et al. TCAD22]



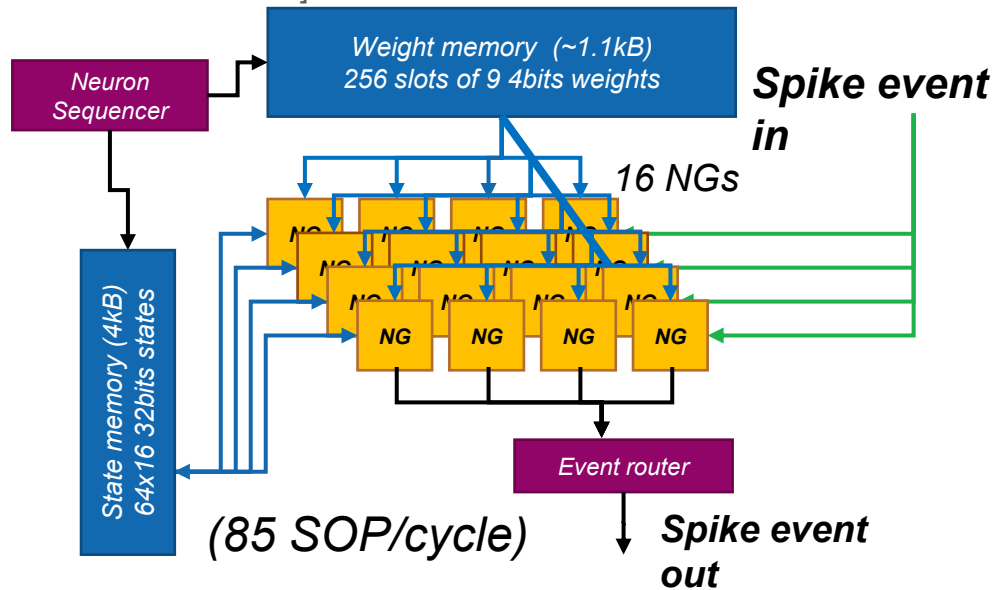
- $K \times K$  window on all input channels unrolled, cycle-by-cycle sliding
- Completely unrolled inner products one output activation per cycle!
- Zeros in weights and activations, spatial smoothness of activations reduce switching activity
- 96 OCUs, 96 Input channels,  $3 \times 3$  kernels:  $96 * 96 * 3 * 3 = 82'944$  TMAC/cycle

# Different Sensor Type, different Acceleration Engine

**Event Sensors:**  
**DVS**  
**Ultra-low latency**  
**Energy-**  
**proportional**  
**interface**

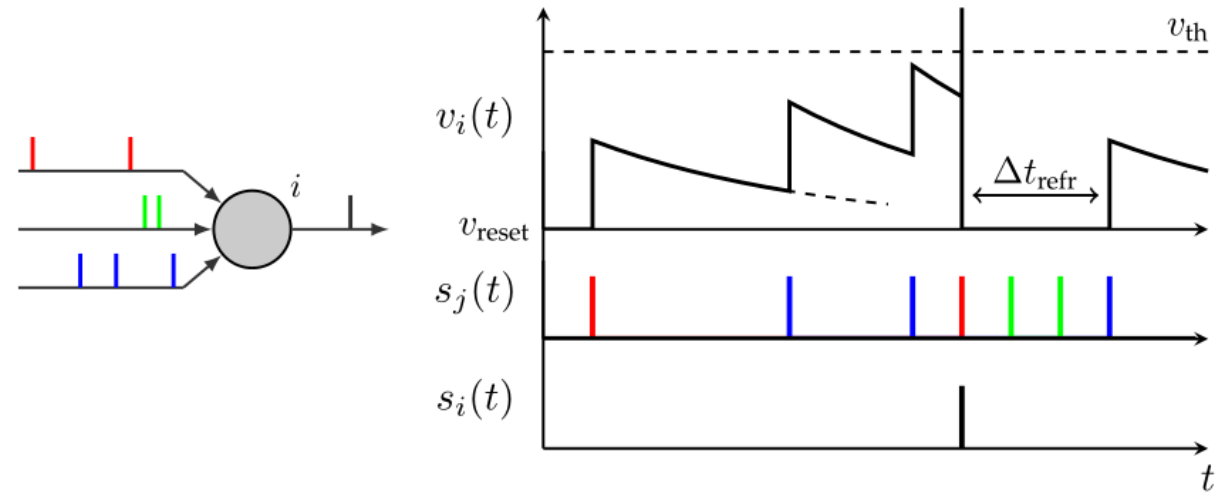


[Di Mauro et al. DATE22]



**Leaky Integrate & Fire (LIF) neurons**

**Spiking Neural Engine (SNE)**



**SNE works seamlessly with DVS (event-based) sensors**



# General Purpose PE: Domain-Specialized RV32 Core



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

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!

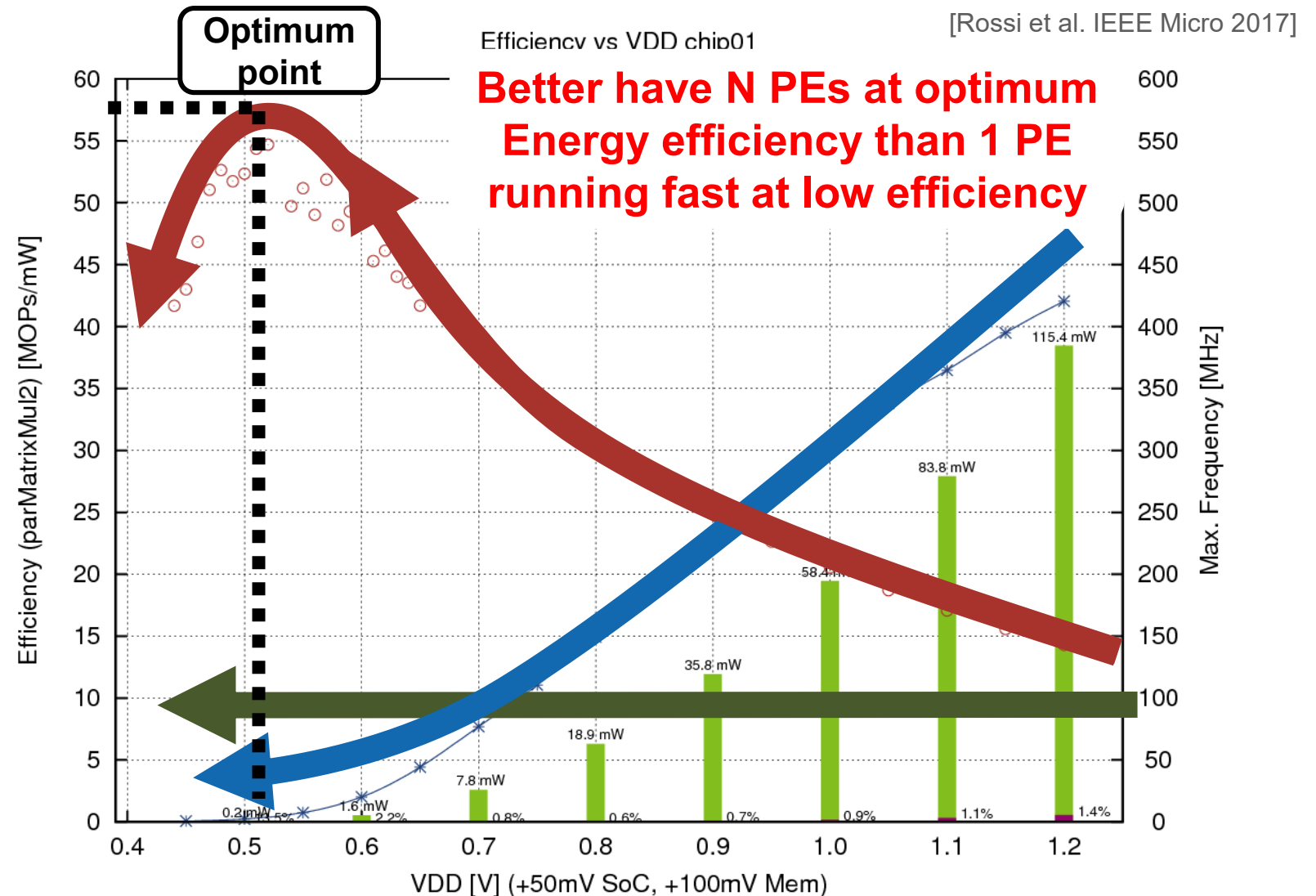
Specialization Cost: Power,Area: 1.5x↑ but Time 15x↓ → **E = PT 10x ↓**



# Parallel, Ultra-Low Power (PULP) PE Cluster

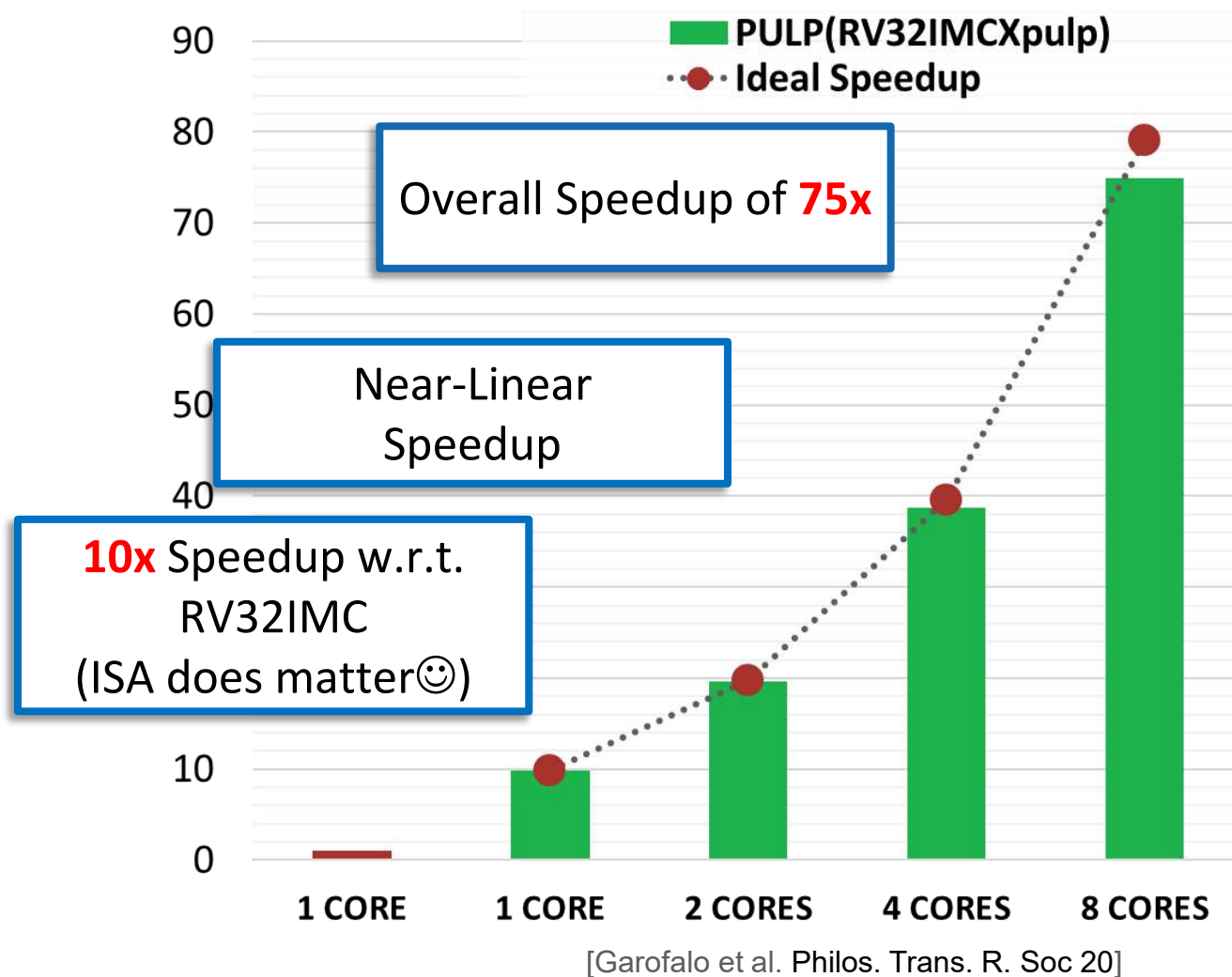
- As VDD decreases, operating speed decreases
- However efficiency increases → more work done per Joule
- Run parallel to get performance and efficiency!

**AI is parallel and scales  
More parallel with NN  
size**

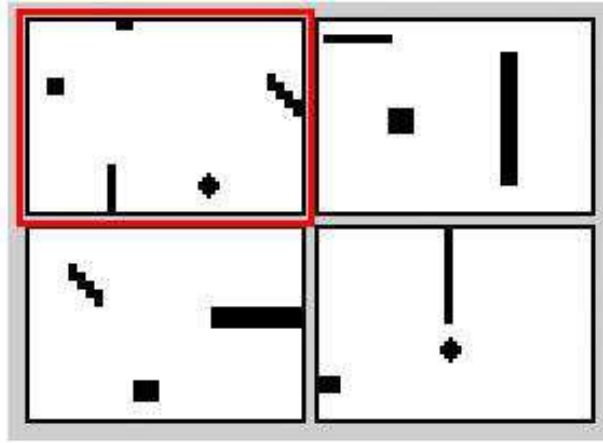


# Combining PE Specialization & Efficient Parallel Execution

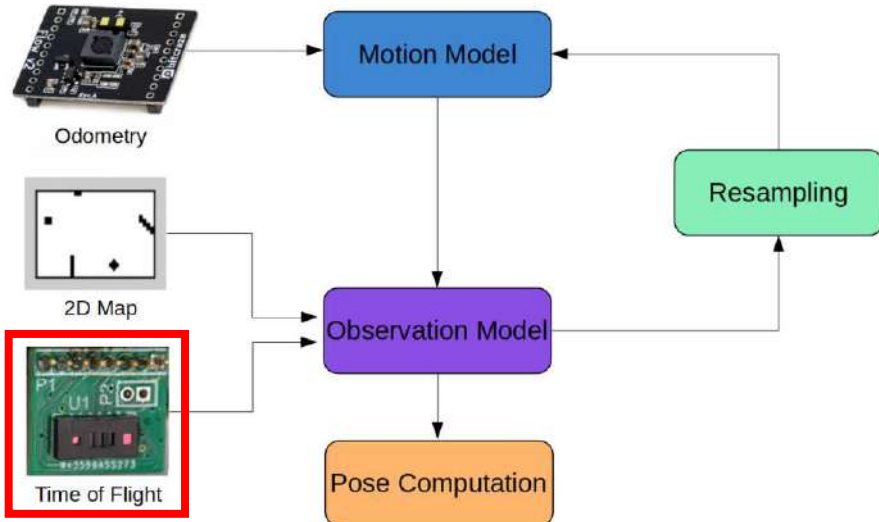
- 8-bit convolution
  - Open source DNN library
- **10x efficiency** from PE specialization
- Near-linear Parallel speedup
  - Scales well for regular workloads
- **75x performance (combined)**
- **7-8 GMACs, 1pJ/OP (1mW/GOPS)**
  - 250MHz
  - 4 MAC/Cycle (8bit)
  - 8 Cores



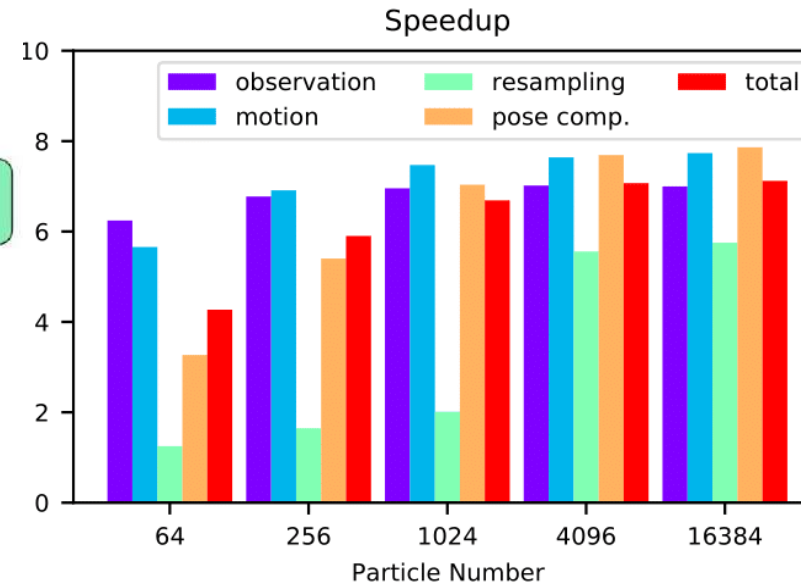
# Not only Perception: SLAM, Planning



## Particle filter-based



## Convergence + Low ATE for $N_{part} > 1024$ , 2ToF, FP16 acceptable



12MHz, 1Kpart. 13mW, 60msec  
400MHz, 1Kpart 61mW, 1msec  
400MHz 16Kpart 61mW, 30msec



# Advancing the SOA on all tasks

## RISC-V Cluster

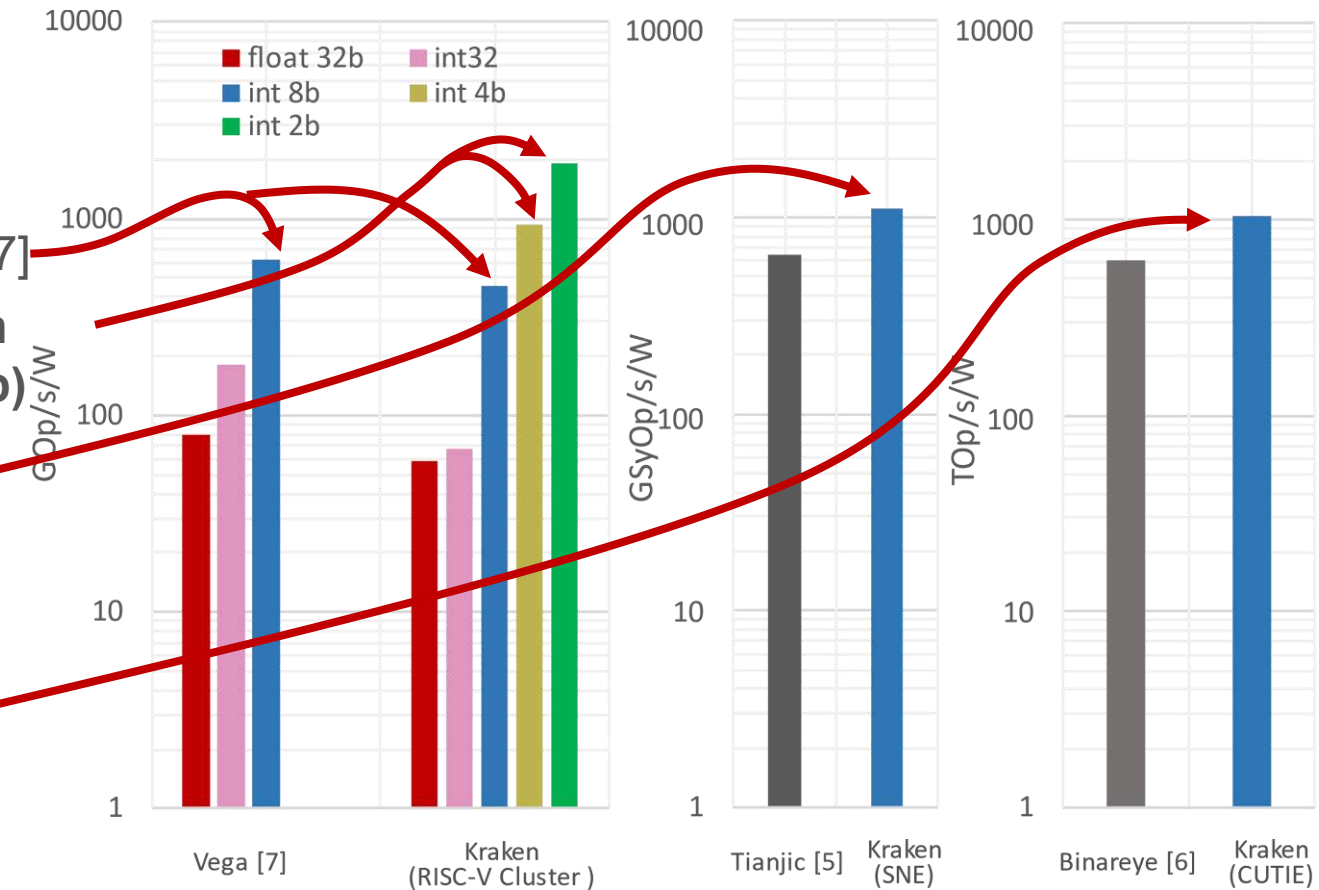
- Comparable 32bits-8bits SOA Energy efficiency to other PULPs [7]
- The highest energy efficiency on sub-byte SIMD operations (4b-2b)

## SNE

- 1.7X higher than SOA [5] energy/efficiency

## CUTIE

- 2X higher energy efficiency improvement over SOA [6]



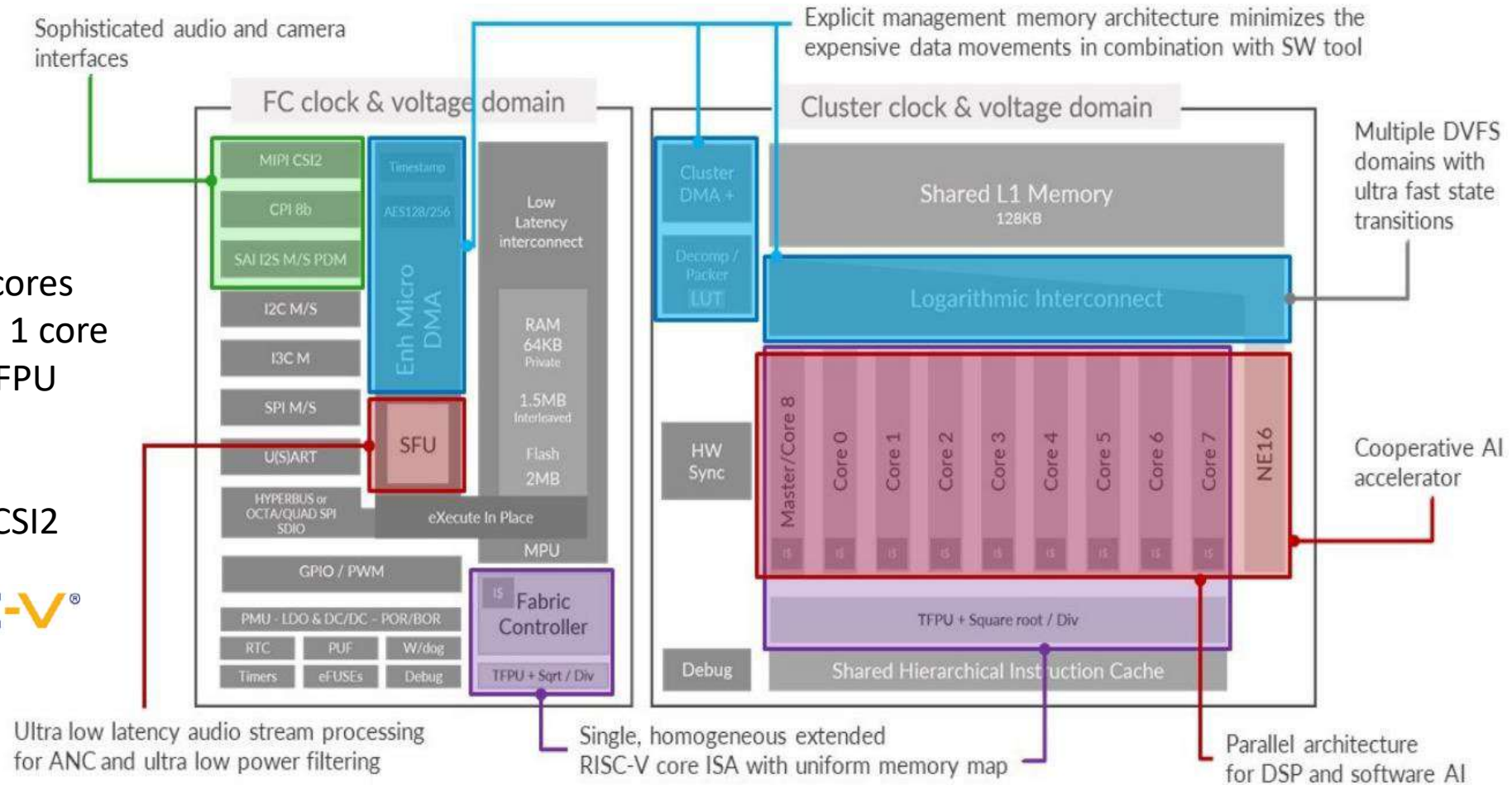
**CUTIE, SNE can work concurrently for SNN + TNN “fused” inference (never done so far)**

# From Pre-competitive to commercial: PULP → GAP9



## GAP9

- Cluster: 9 cores
- Fabric Ctrl: 1 core
- Hardware FPU
- L1: 128 KB
- L2: 1.5 MB
- Interface: CSI2



# 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**



# Precompetitive Partnership Buildup



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

**ETH** zürich



Universidade do Minho

**intel**®



**BOSCH**



life.augmented



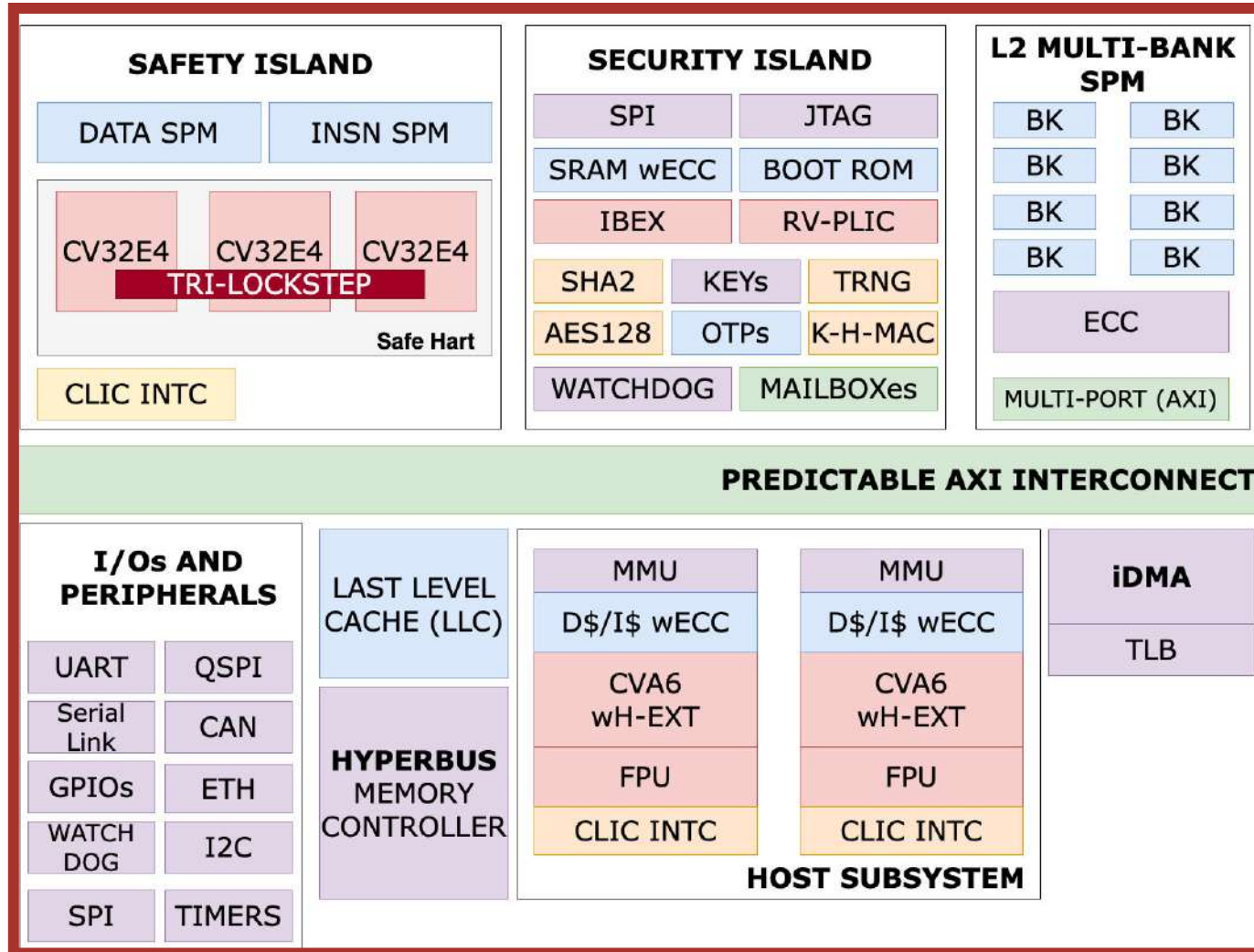
**ETH** zürich

- Project Leaders
- Digital Systems Design, PULP, Open-Source, RISC-V
- Processors/Ips/Interconnects/Interrupts/HW Acceleration
- SW stack, compilers, runtime and optimized routines
- Real-Time (RT) Systems and On/Off-Chip RT Communication
- Safe/Secure Cyber-Physical Systems
- Virtualization-assisted systems, OS, Hypervisors, RISC-V
- Security of Cyber-Physical Systems
- Intel16 FinFet technology (for the first prototype)
- ASIC design support and packaging
- Supporters: STMicroelectronics, BOSCH

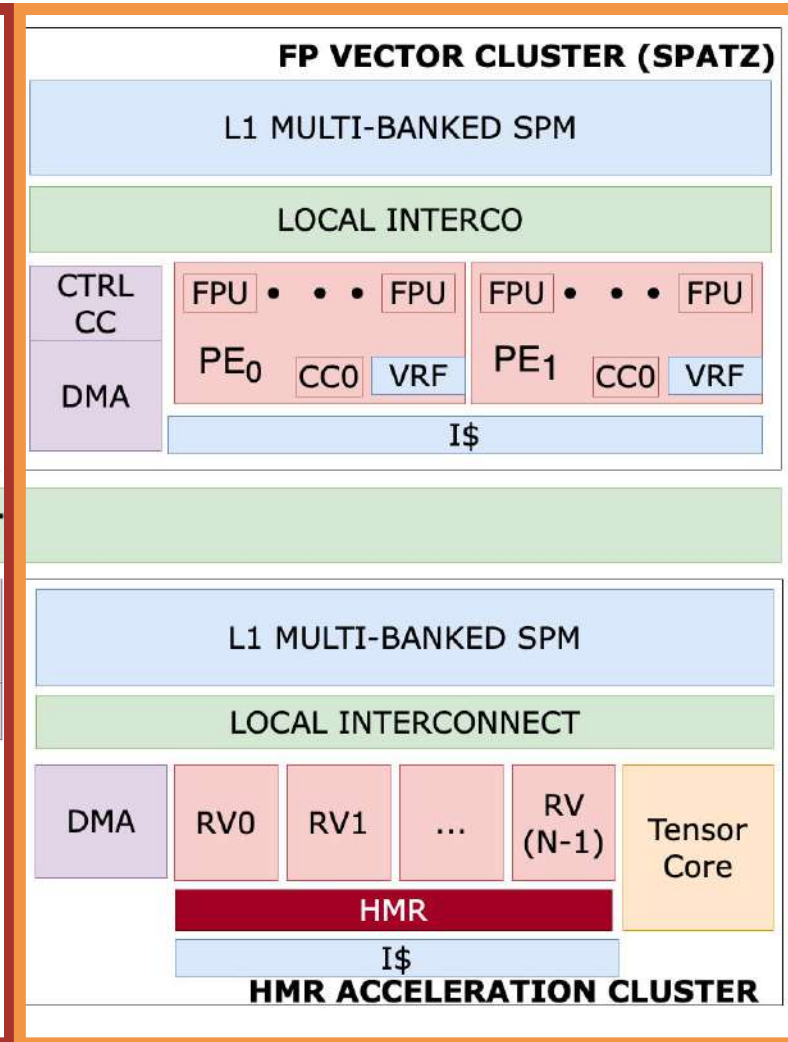


# Carfield: Efficiency + Safety, Security, Predictability

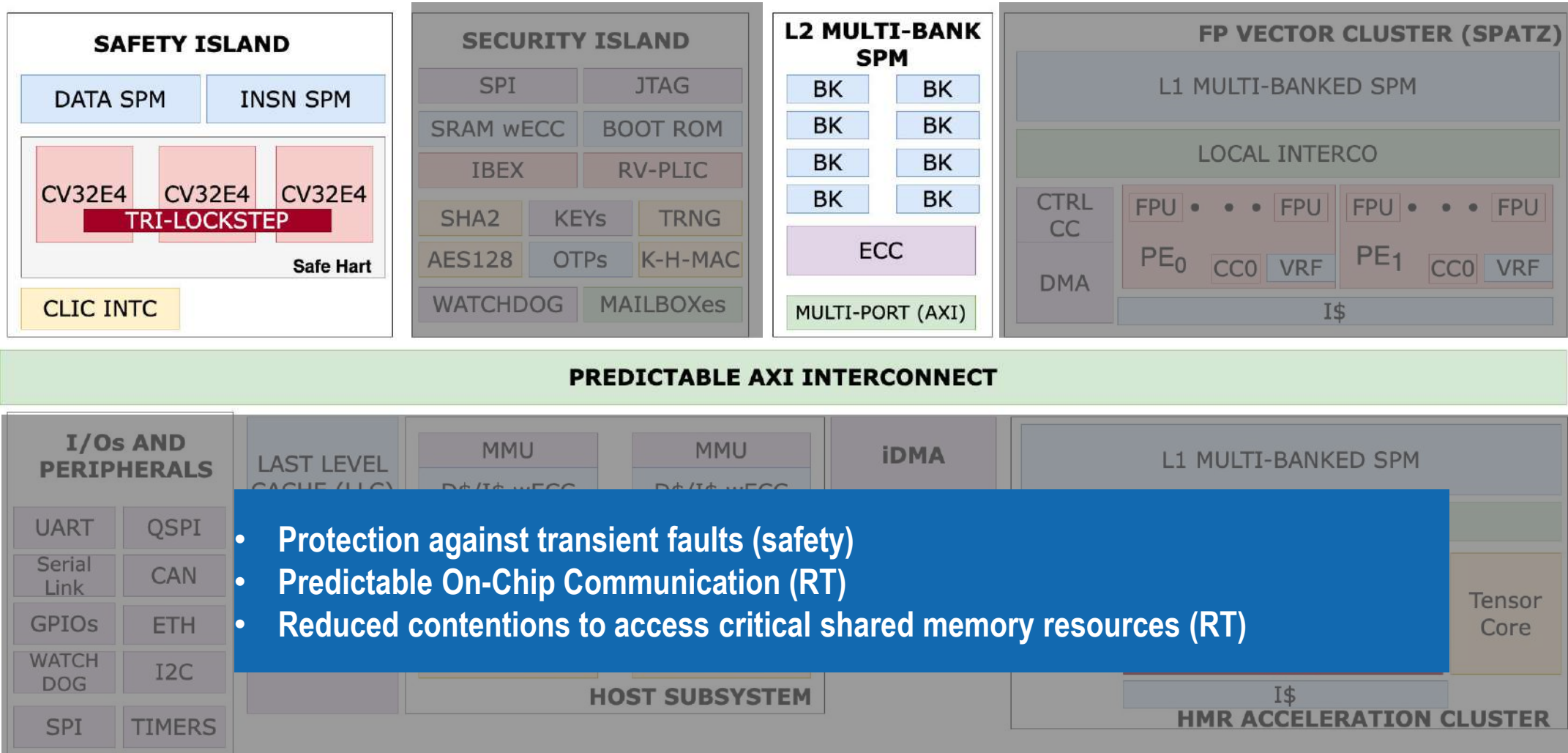
## Main Computing and I/O System



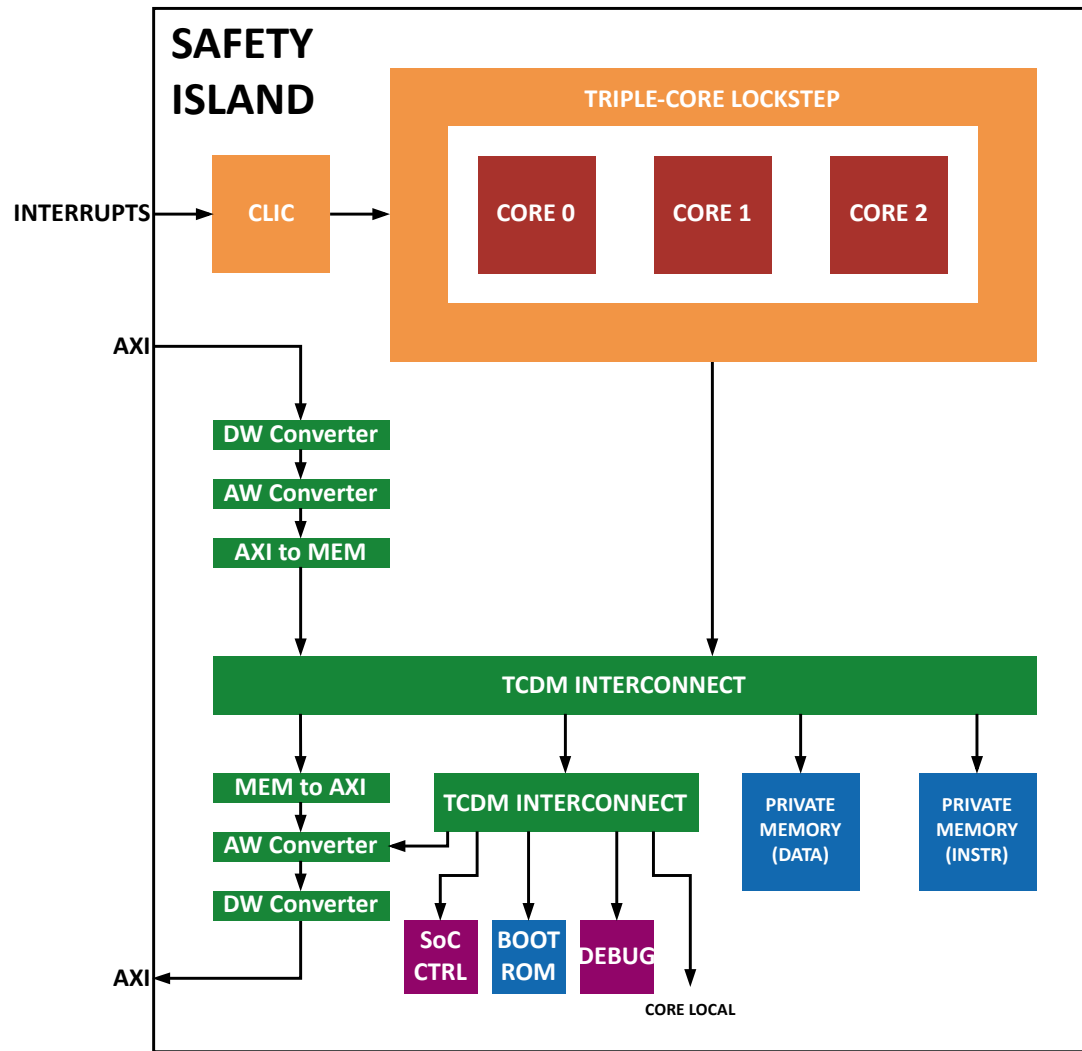
## Accelerators Domain



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

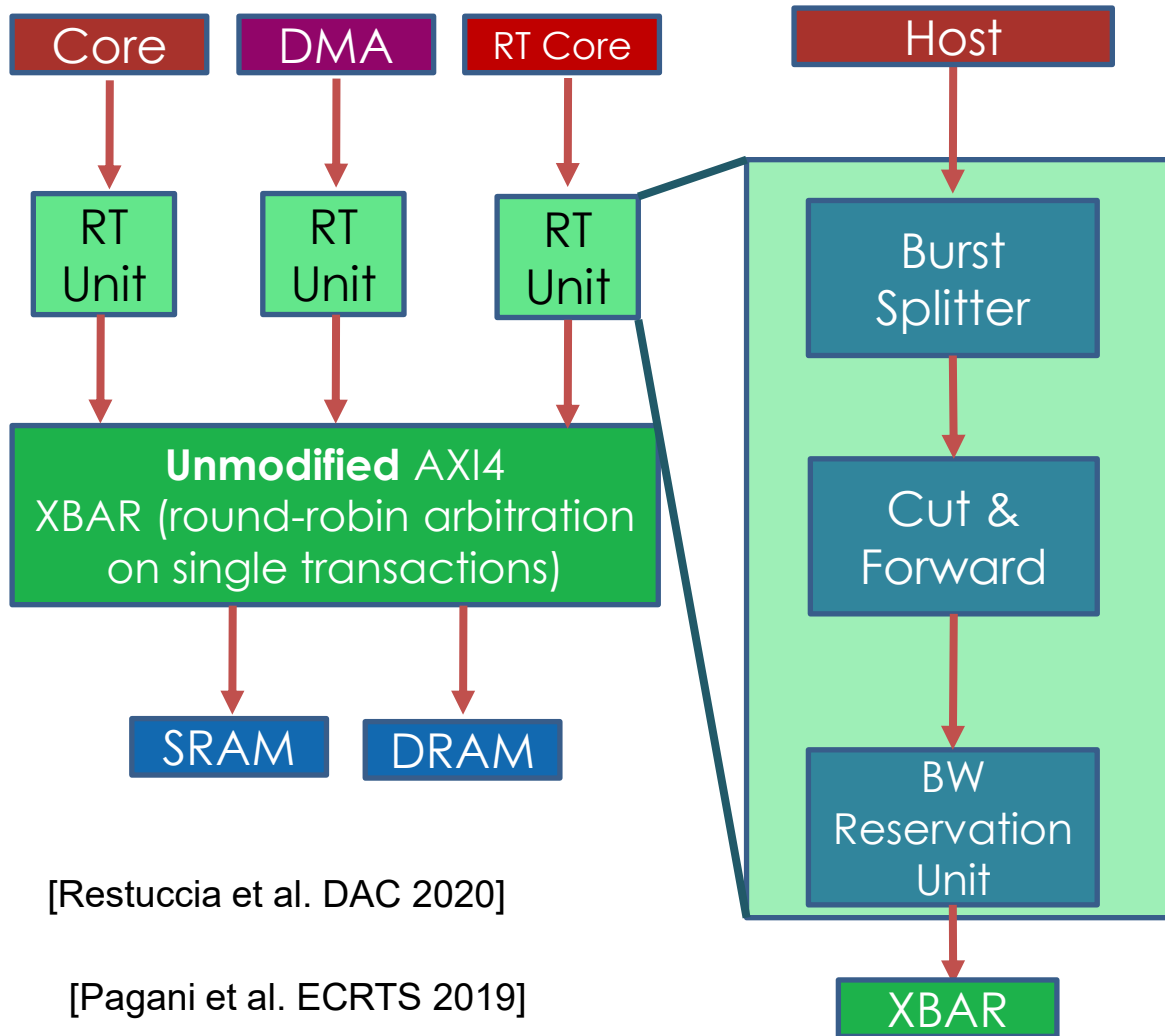


# The Safety Island



- 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)



- AXI4 inherently **unpredictable**
- **Minimally Intrusive Solution**
  - No huge buffering, limited additional logic
  - **Solution 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)

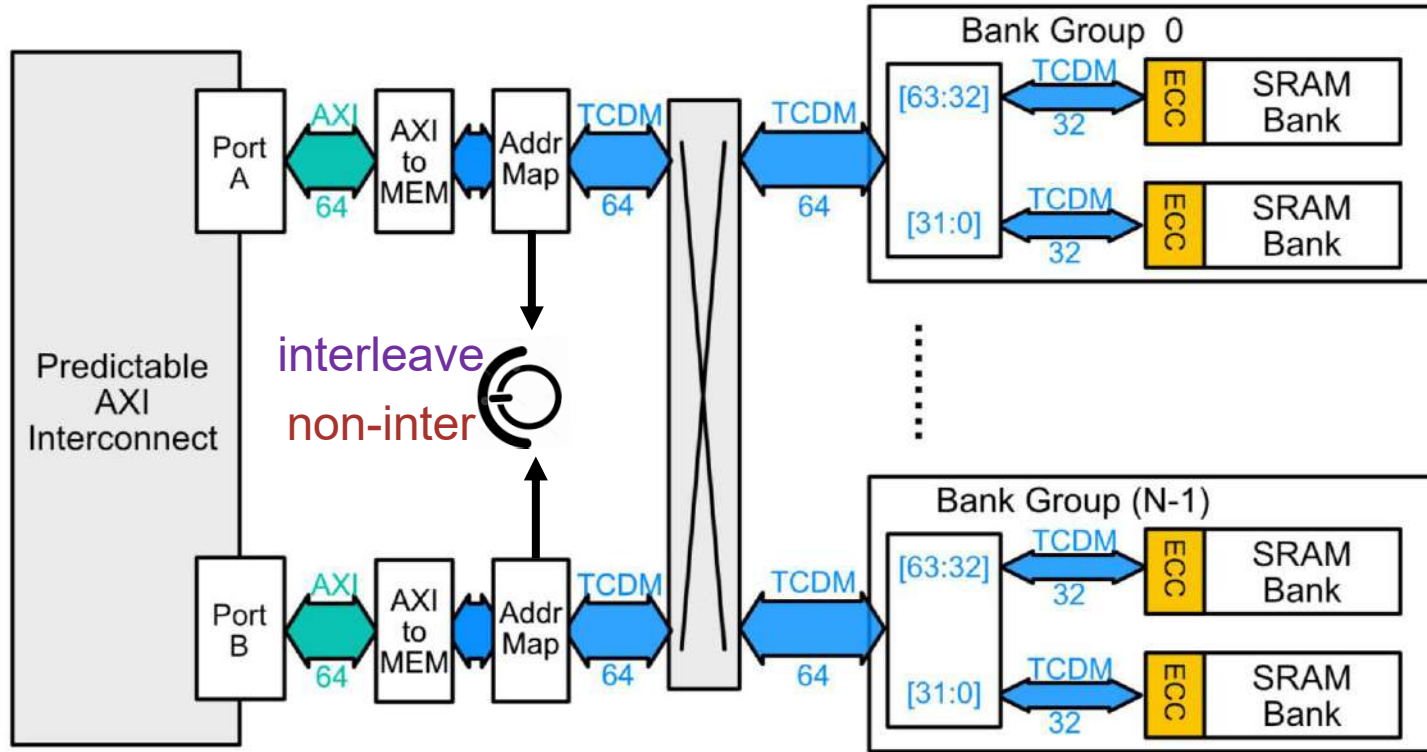
[Restuccia et al. DAC 2020]

[Pagani et al. ECRTS 2019]

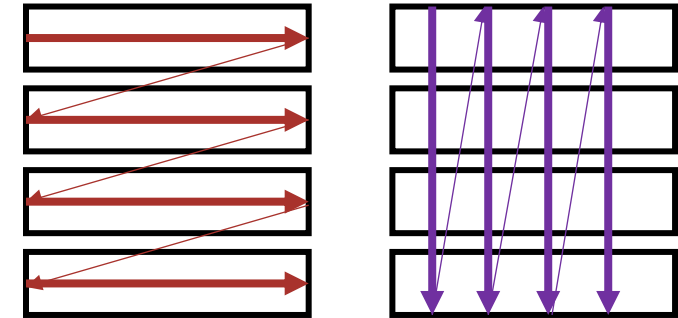
# Contention-Free Shared L2 Scratchpad Memory

## 1. Dual-AXI-Port L2 Mem Subsystem

Multi-banked L2 SPM accessible from two different AXI ports



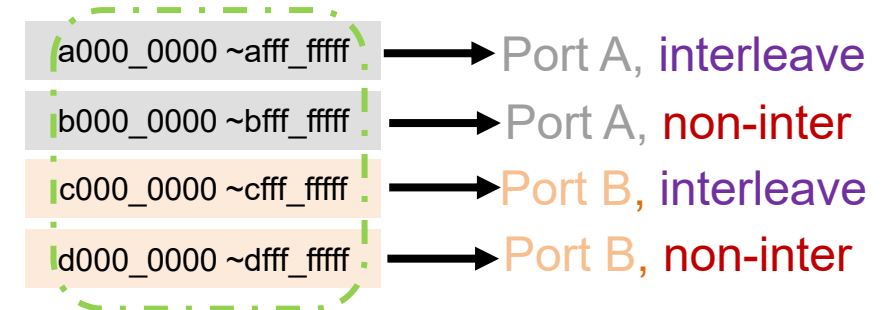
## 2. Two Address Mapping Modes



Non-interleaved

Interleaved

## 3. Dynamic Address Mapping by Address spaces, eg:



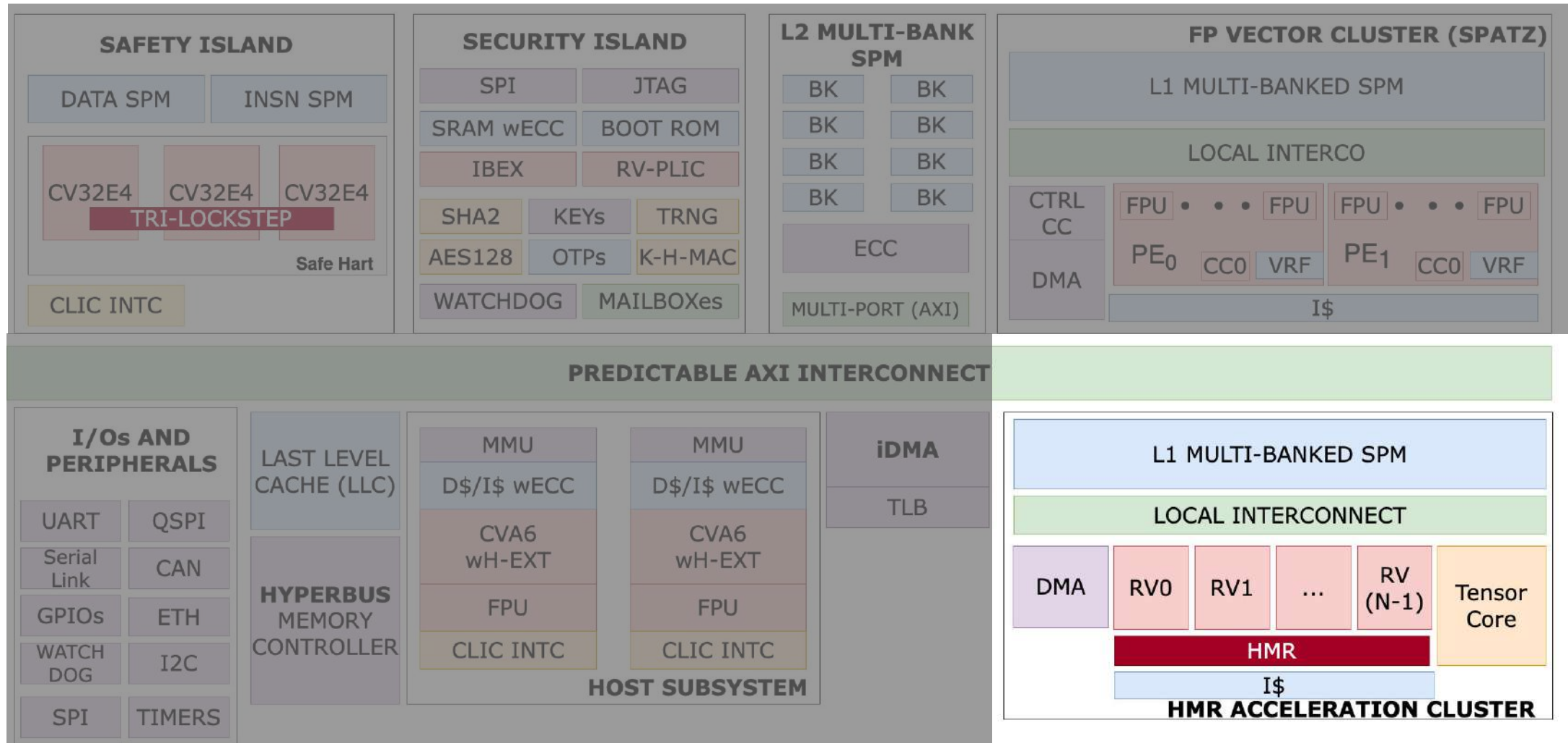
Point to the same L2 physical Mem space

## 4. We determine in SW which port and which mode to use

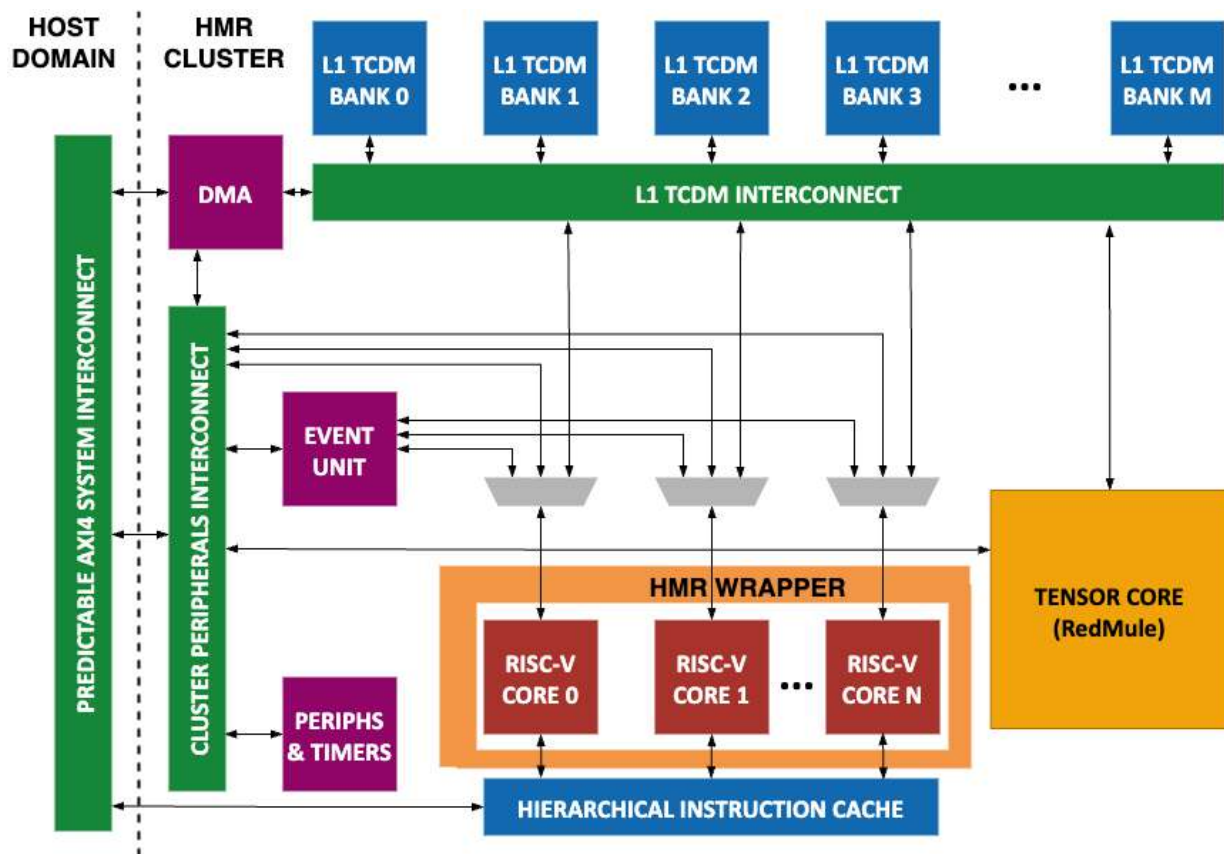
By using different address space!



# The HMR Acceleration Cluster



# The HMR Cluster for DNN-Oriented INT/FP Workloads



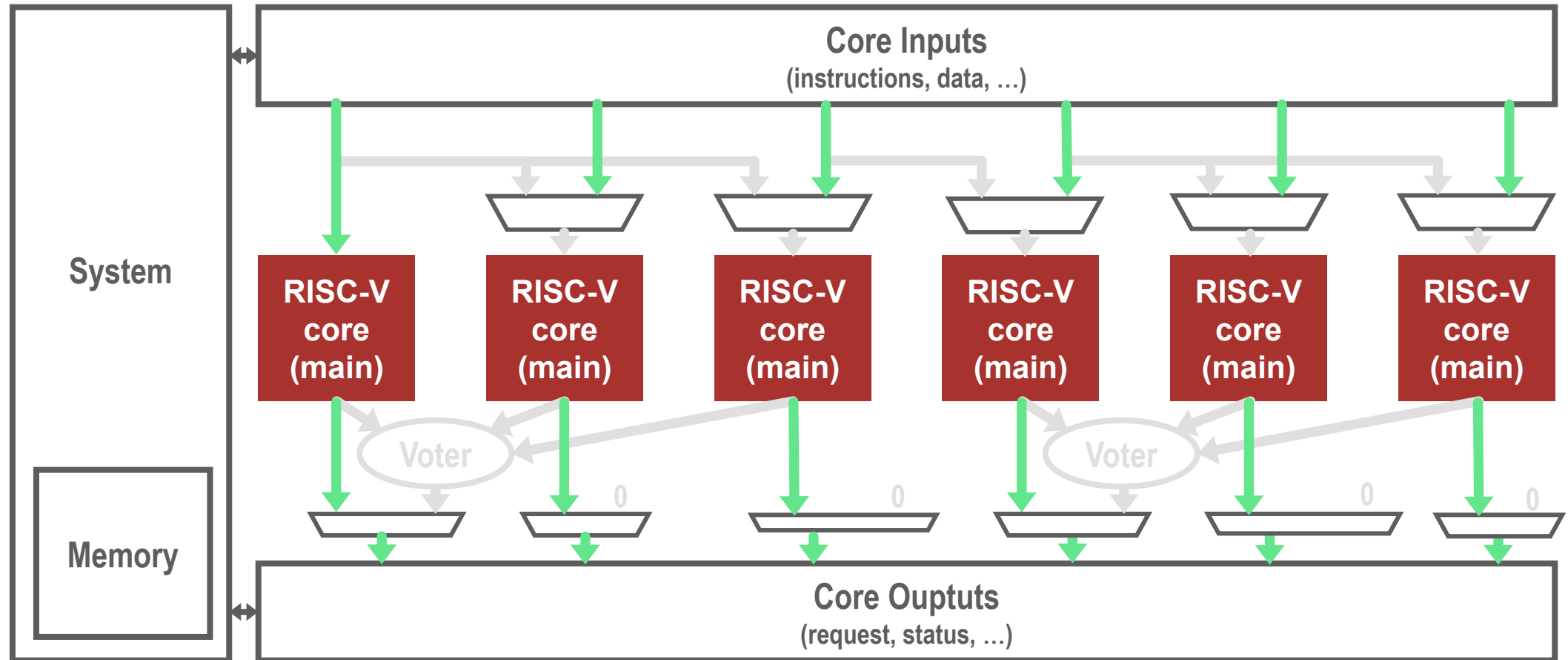
- 12x 32-bit RISC-V cores with support for DSP/QNN ISA Extensions
- Single-Cycle Multi-Banked Tightly-Coupled Data Memory (Scratchpad)
- Hardware Synchronizer
- DMA Controller for Explicit Memory Management
- L1-coupled **TensorCore** (RedMule)
- **Runtime-configurable Dual/Triple core redundancy mode** + hw/sw-based quick recovery mechanism

[Rogenmoser et al., arXiv, 2023]

[Tortorella et al., arXiv, 2023]

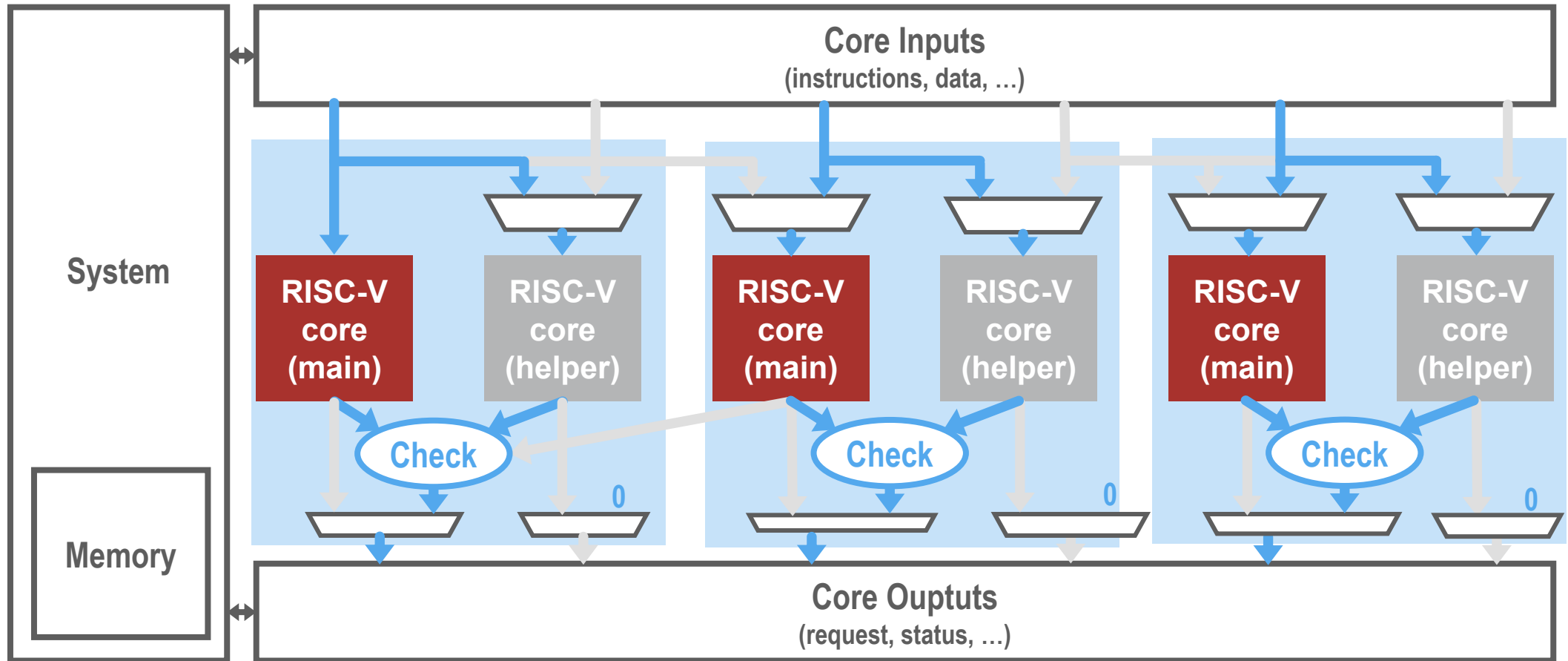
# Hybrid Modular Redundancy (HMR): Reconfigurable

**Independent Mode:** high performance, no reliability



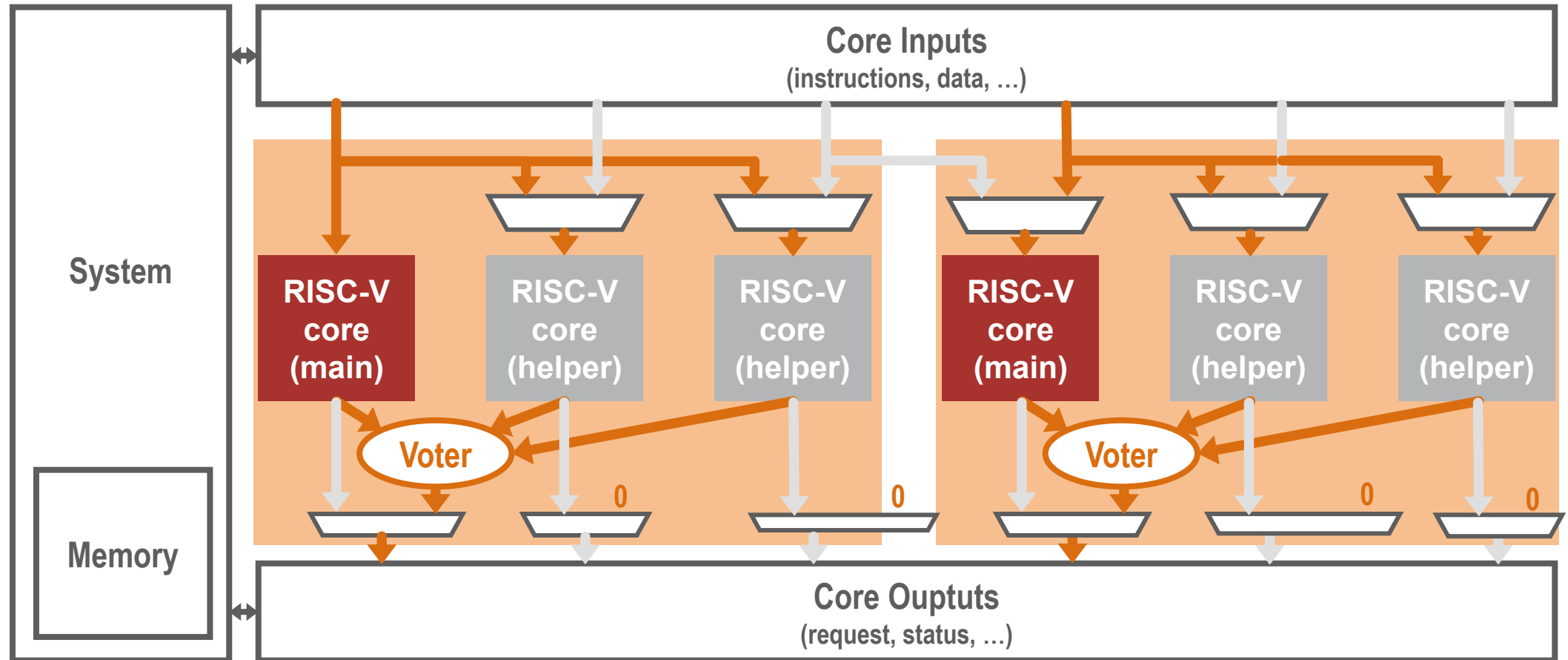
# Hybrid Modular Redundancy (HMR): Reconfigurable

**DMR Mode:** good performance, good reliability, slow recovery



# Hybrid Modular Redundancy (HMR): Reconfigurable

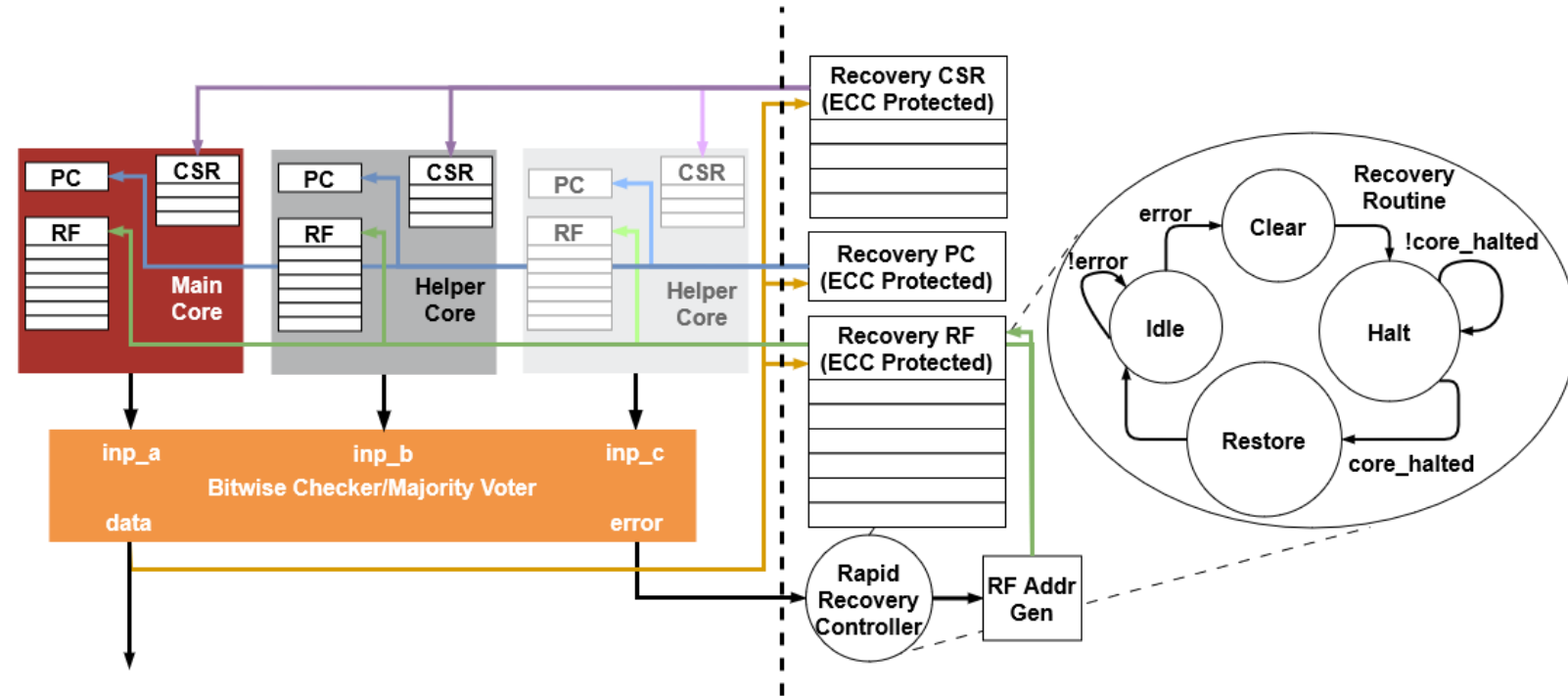
**TMR Mode:** low performance, high reliability, quick recovery





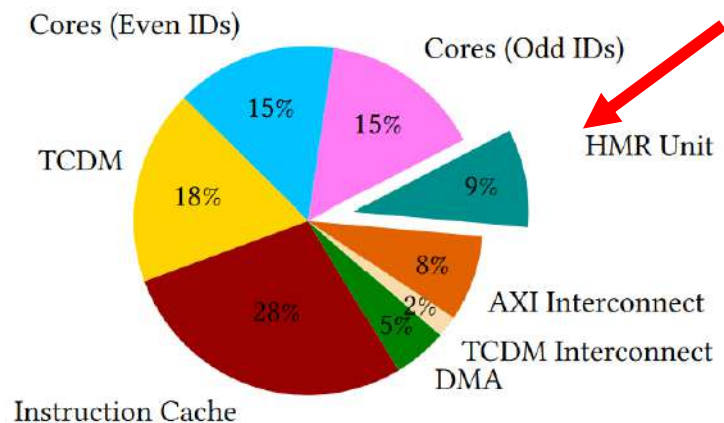
# Rapid Recovery: shared hardware extension

- Cycle-by-cycle backup of the cores state in ECC-protected Status Registers
- Quick recovery procedure (24 cycles!)
- Shared logic between TMR and DMR modes

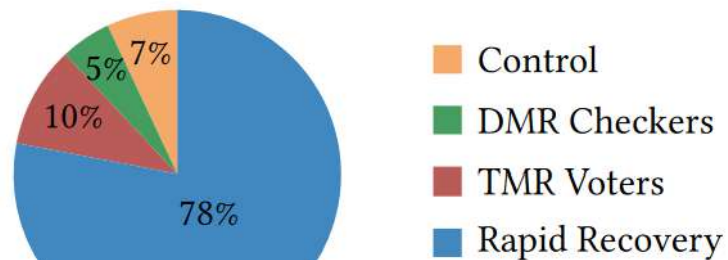


# HMR, yes... but at which cost?

Cluster Area breakdown with HMR Unit



HMR Unit Area Breakdown



Area Overhead of HMR Configurations

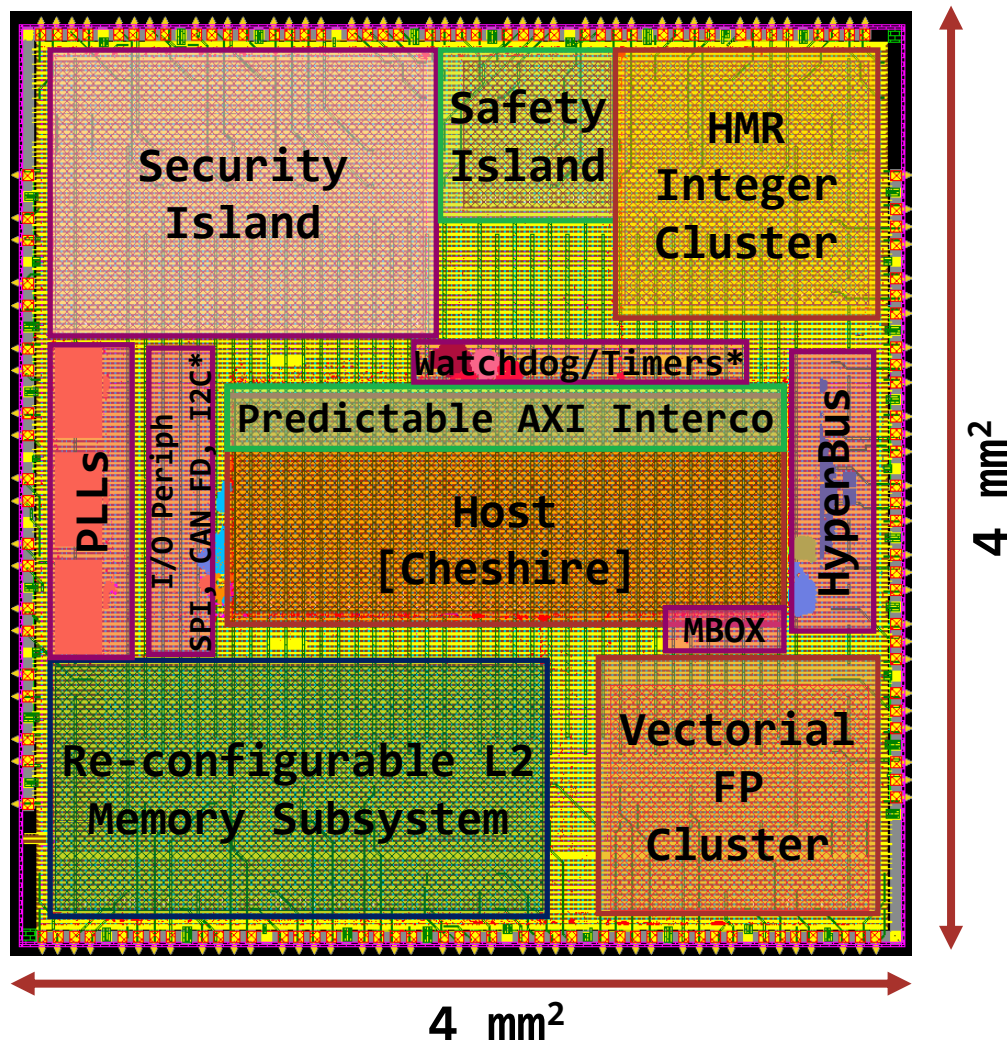
PULP Cluster Area [mm <sup>2</sup> ]	Overhead
Baseline	-
DMR	0.3%
TMR	0.7%
HMR	1.3%
With Rapid Recovery	
DMR	8.4%
TMR	8.8%
HMR	9.4%

HMR Unit Recovery and Switching Mode Latency

	DMR	TMR	DMR Rapid Recovery	TMR Rapid Recovery
Recovery Latency [cycles]	Application dependant	363	24	24
Mode Switching [cycles]	703	598	603	515

[Rogenmoser et al., arXiv, 2023]

# Carfield SoC Flooplan – Taped out 11/2023



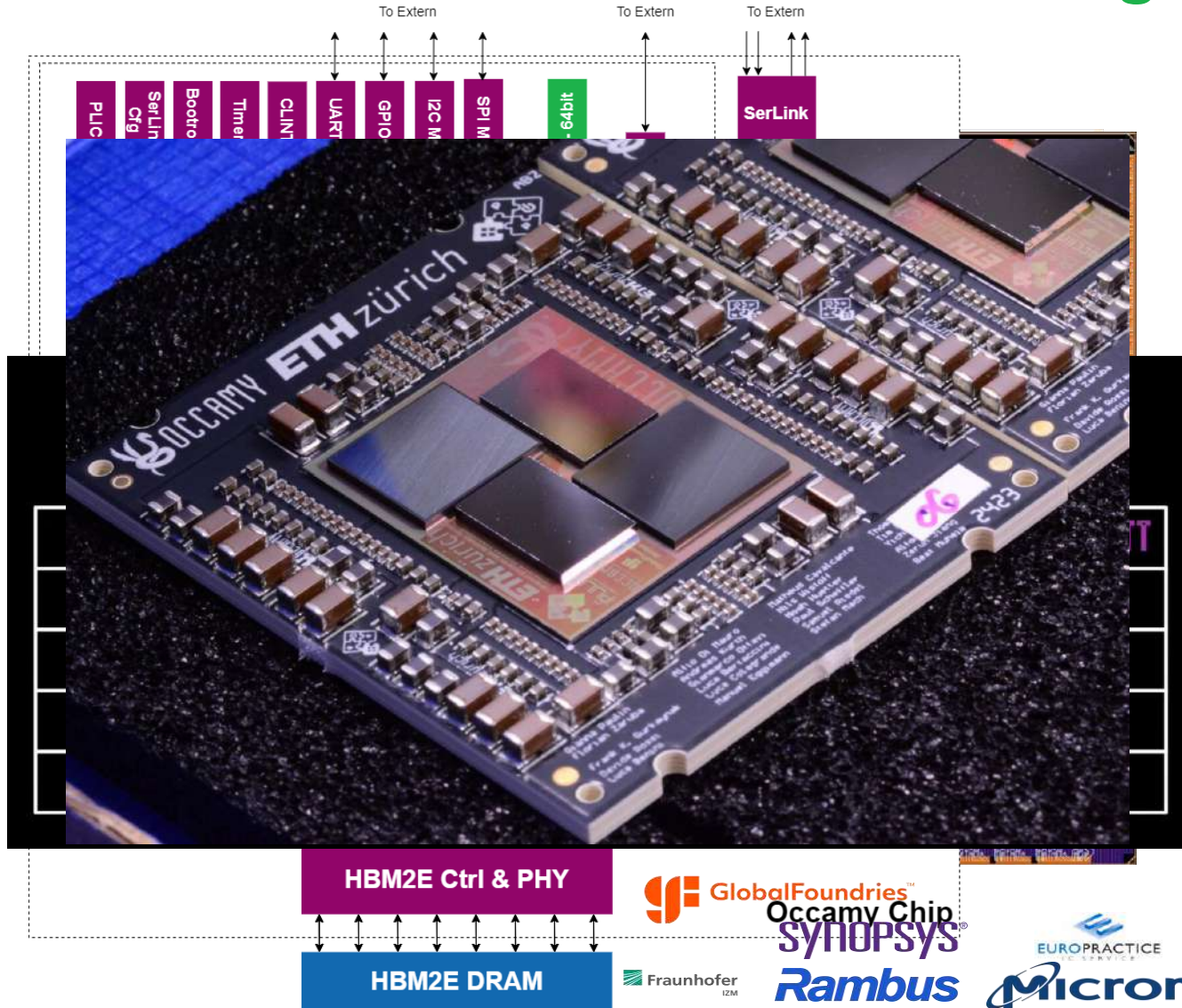
- **Host [Cheshire]**
  - Dual-Core 64-bit RISC-V processor; **2.45 mm<sup>2</sup>**; 600 MHz;
- **Security Island**
  - Low-power secure monitor; **1.94 mm<sup>2</sup>** ; 100 MHz;
- **Safety Island**
  - **0.42 mm<sup>2</sup>**; 500 MHz
- **Re-configurable L2 Memory Subsystem**
  - 1MB; **2.33 mm<sup>2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

Frequency bound by RAMs (limited availability in Intel offering for Universities)

Modules marked with (\*) are not in scale



# Moonshot: Toward Self-Driving Cars

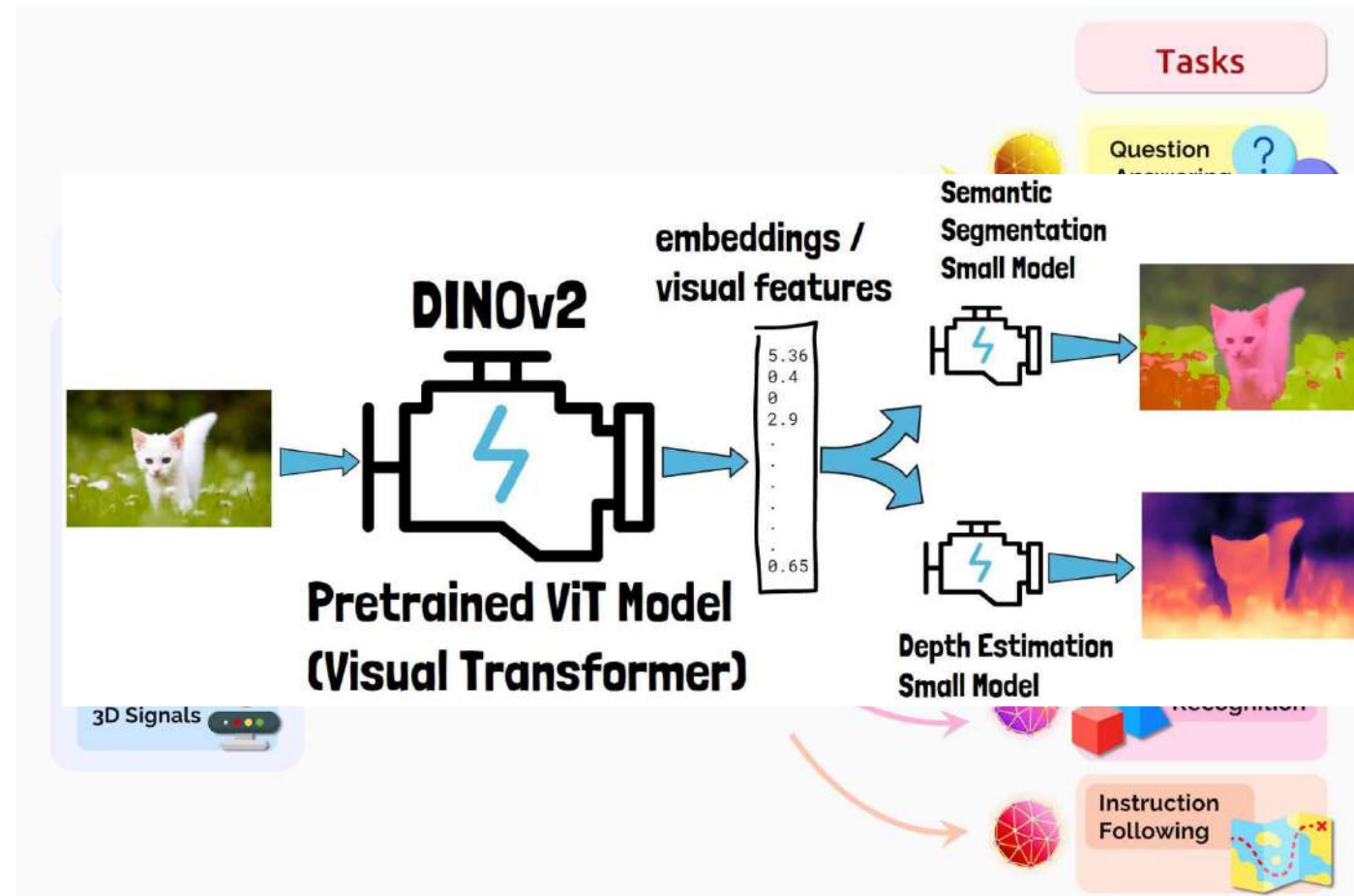


- GF12, target **1GHz** (typ)
- 2 AXI NoCs (multi-hierarchy)
  - 64-bit
  - 512-bit with “interleaved” mode
- Peripherals
- Linux-capable manager core CVA6
- 6 Quadrants: 216 cores/chiplet
  - 4 cluster / quadrant:
    - 8 compute + 1 DMA core / cluster
    - 1 multi-format FPU / core (FP64,x2 32, x4 16/alt, x8 8/alt)
- 8-channel HBM2e (8GB) **512GB/s**
- D2D link (Wide, Narrow) **70+2GB/s**
- System-level DMA
- SPM (2MB wide, 512KB narrow)

**Peak 384 GDPflop/s per chiplet**

# What's Next? The era of Foundation Models

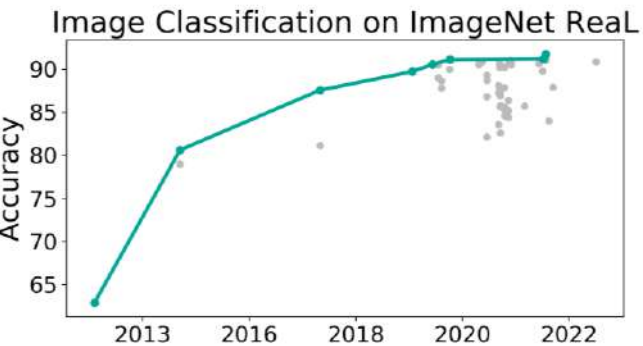
- Versatility and Multi-modality
  - Natural language processing, computer vision, robotics, biology, ...
- Homogenization of models
  - **Transformers as foundation models**
- Self-supervision, Fine-tuning
  - Self-supervised training on large-scale unlabeled dataset
  - Fine-tune (few layers) on specific tasks with smaller labeled datasets.
- Zero-shot specialization
  - Prompt engineering for new tasks



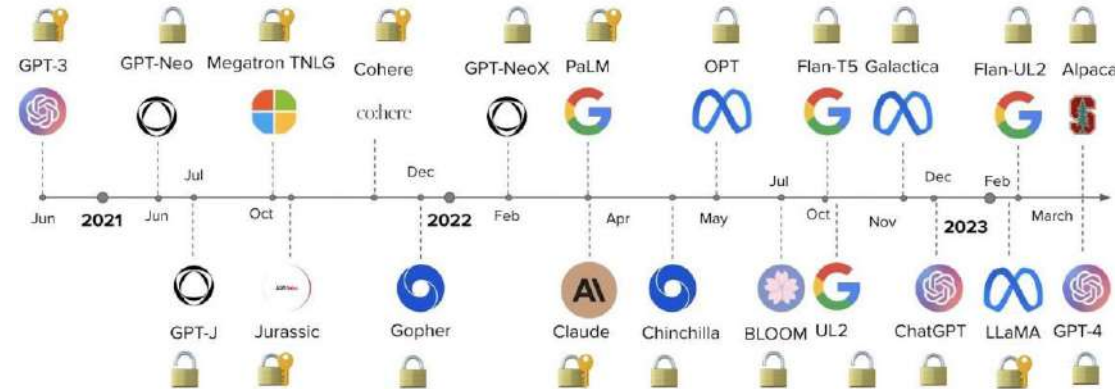
Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models." Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI).



# Perceptive → Generative → Embodied AI



**Precise**



**Interactive, creative**



**Efficient, RT-safe, secure**

# Final Thoughts

## ■ Open Source Computing HW is Happening

- RISC-V Momentum is growing fast
- Stronger Ecosystems emerge
- RISC-V automotive products are in production (e.g. renesas)

## ■ Europe is **NOT** late

- Strong support from Public Authorities
- Key EU OEMs (NXP, Infineon, STM) are active
- Academic world is strongly committed

## ■ Nurture the **industrial OSCHW** ecosystem

- Automotive electronics is a key priority area
- **Avoid fragmentation**
- **Create a stronger vertical ecosystem in the automotive value chain**







# PULP

Parallel Ultra Low Power

Luca Benini, Alessandro Capotondi, Alessandro Ottaviano, Alessio Burrello, Alfio Di Mauro, Andrea Borghesi, Andrea Cossettini, Andreas Kurth, Angelo Garofalo, Antonio Pullini, Arpan Prasad, Bjoern Forsberg, Corrado Bonfanti, Cristian Cioflan, Daniele Palossi, Davide Rossi, Fabio Montagna, Florian Glaser, Florian Zaruba, Francesco Conti, Georg Rutishauser, Germain Haugou, Gianna Paulin, Giuseppe Tagliavini, Hanna Müller, Luca Bertaccini, Luca Valente, Manuel Eggimann, Manuele Rusci, Marco Guermandi, Matheus Cavalcante, Matteo Perotti, Matteo Spallanzani, Michael Rogenmoser, Moritz Scherer, Moritz Schneider, Nazareno Bruschi, Nils Wistoff, Pasquale Davide Schiavone, Paul Scheffler, Philipp Mayer, Robert Balas, Samuel Riedel, Segio Mazzola, Sergei Vostrikov, Simone Benatti, Stefan Mach, Thomas Benz, Thorir Ingolfsson, Tim Fischer, Victor Javier Kartsch Morinigo, Vlad Niculescu, Xiaying Wang, Yichao Zhang, Frank K. Gürkaynak, all our past collaborators **and many more that we forgot to mention**



<http://pulp-platform.org>



@pulp\_platform