**PULP PLATFORM**
Open Source Hardware, the way it should be!

life.augmented | TechWeek 2024
Share, Inspire, Innovate!

# *Open Platforms for the Embodied AI era*

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

erc European Research Council

EuroHPC Joint Undertaking

Horizon Europe
THE NEXT EU RESEARCH & INNOVATION PROGRAMME (2021 – 2027)

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

KDT JU
KEY DIGITAL TECHNOLOGIES JOINT UNDERTAKING

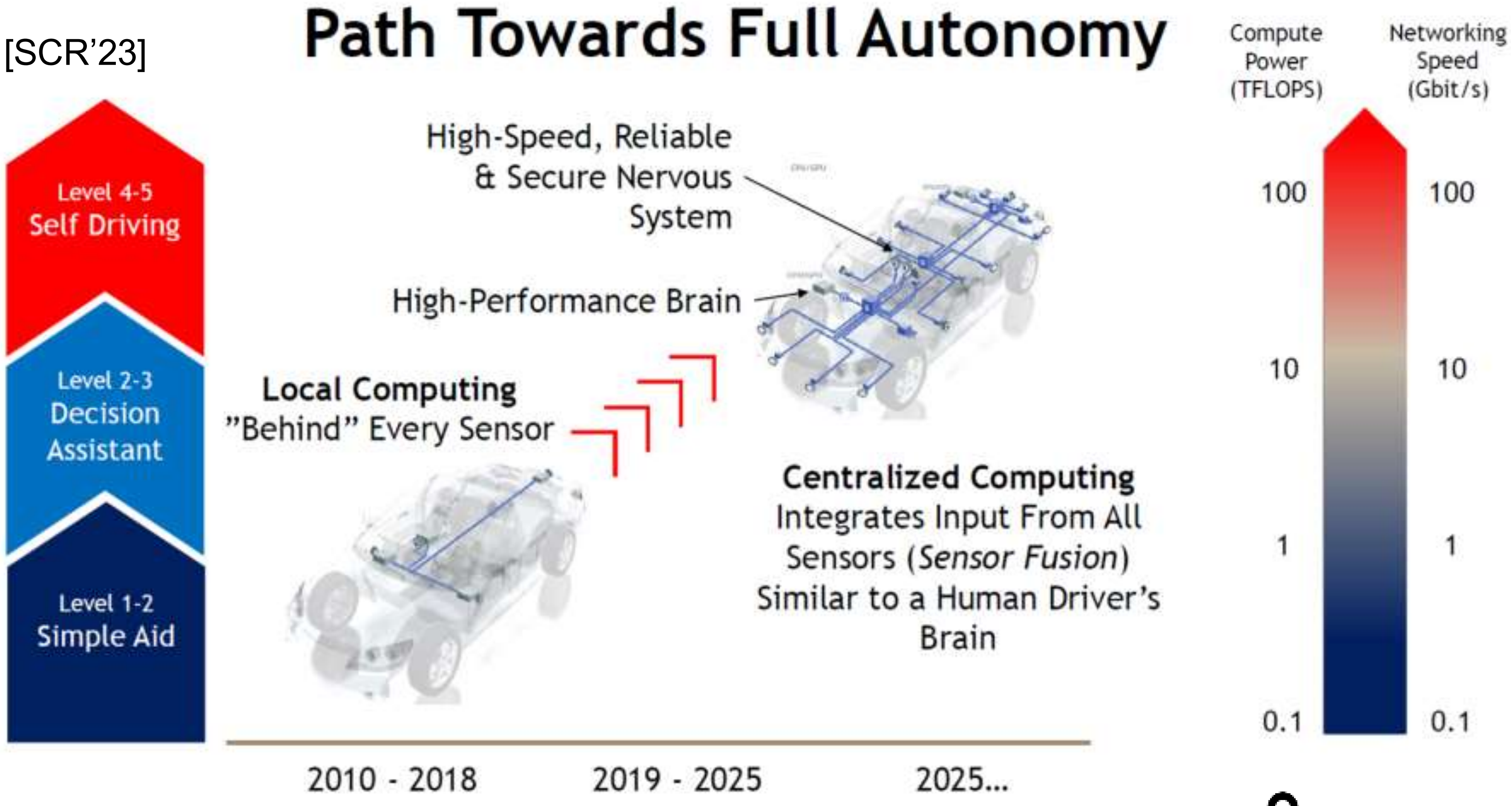
ETH zürich

ALMA MATER STUDIORUM UNIVERSITAS DI BOLOGNA A.D. 1088

http://pulp-platform.org @pulp_platform https://www.youtube.com/pulp_platform

# Embodied AI

[SCR'23]



## Path Towards Full Autonomy

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

**Efficient**

**On-car Computing PMAX < 1.5KW**

**Energy Efficiency**

$$\left(\frac{1}{\text{Power}\cdot\text{Time}}\right)$$

**10x/12Y by scaling vs. model complexity 10x/2Y**

**Safe** **Real-time** **Secure**

ETH zürich

# Start Small: Open Platform for Autonomous Nano-Drones

## Advanced autonomous drone

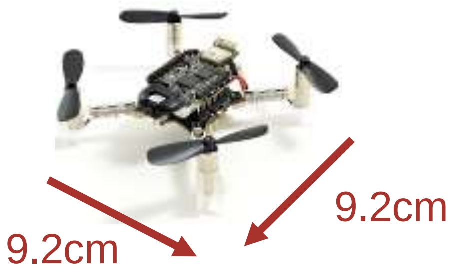
[1] A. Bachrach, “Skydio autonomy engine: Enabling the next generation of autonomous flight,” IEEE Hot Chips 33 Symposium (HCS), 2021

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

- 3D Mapping & Motion Planning
- Object recognition & Avoidance
- 0.06m2 & **800g of weight**
- Battery Capacity **5410mAh**

## Nano-drone

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



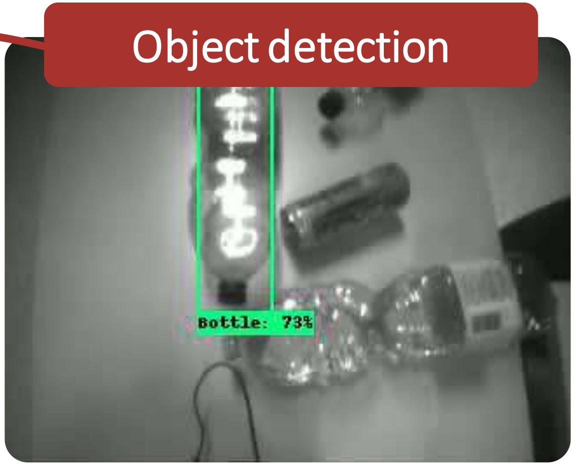
- Smaller form factor of 0.008m2
- Weight **27g (30X lighter)**
- Battery capacity **250mAh (20X smaller)**

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

ETH zürich

# Achieving True Autonomy on Nano-UAVs

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



Object detection

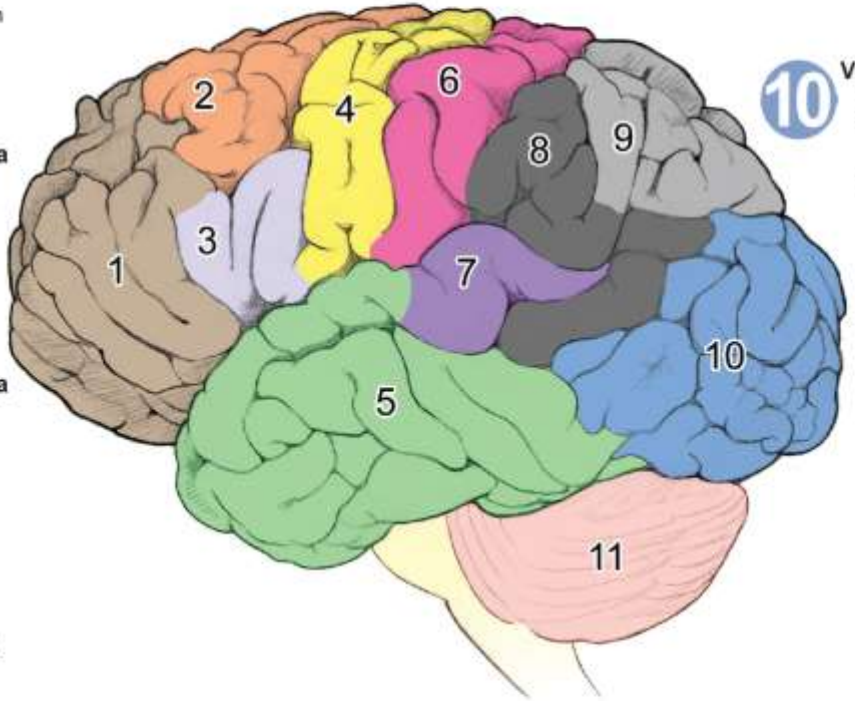
Obstacle avoidance & Navigation

Environment exploration

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

# Multiple Heterogeneous Accelerators

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



1 **Higher Mental Functions**
Concentration
Planning
Judgment
Emotional expression
Creativity
Inhibition - Ability to control self

2 **Motor Function Area**
Eye movement and placement of eyes

3 **Broca's Area**
Ability to talk
Ability to write

4 **Motor Function Area**
Ability to move muscles

5 **Association Area**
Short-term memory
Emotion

6 **Sensory Area**
Touching and feeling

7 **Auditory Area**
Hearing

8 **Wernicke's Area**
Written and spoken language understanding

9 **Somatosensory Association Area**
Understanding of weight, texture, temperature, etc. for recognizing and comprehending an object

10 **Visual Areas**
Sight
Ability to recognize pictures
Awareness of size and shape

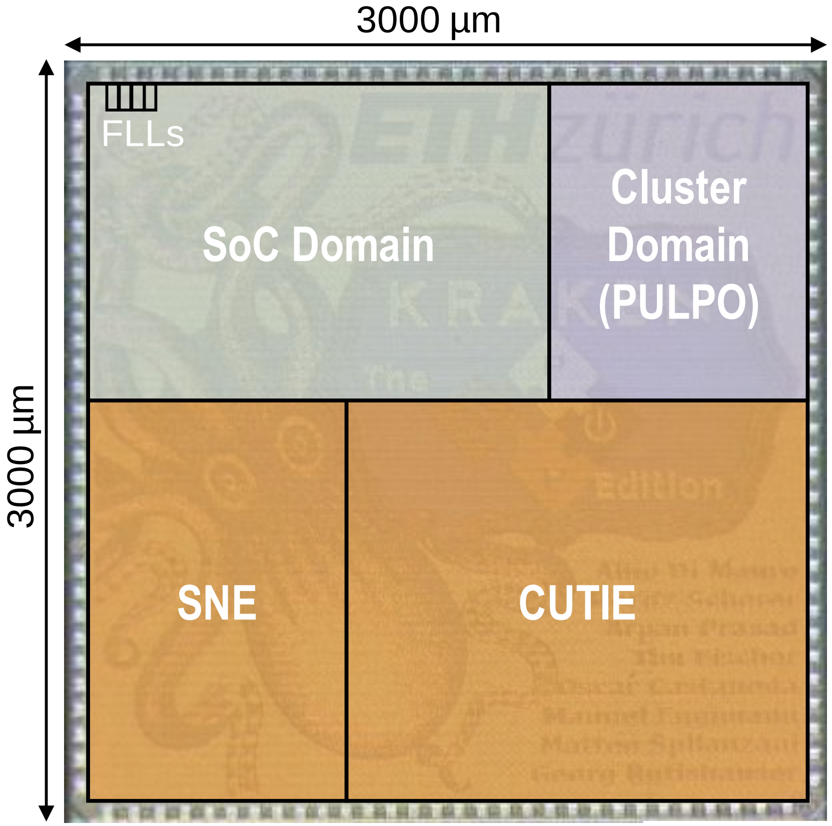
**FUNCTIONAL AREAS OF THE CEREBELLUM**

11 **Motor Functions**
Coordination of movement
Balance
Posture

# 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

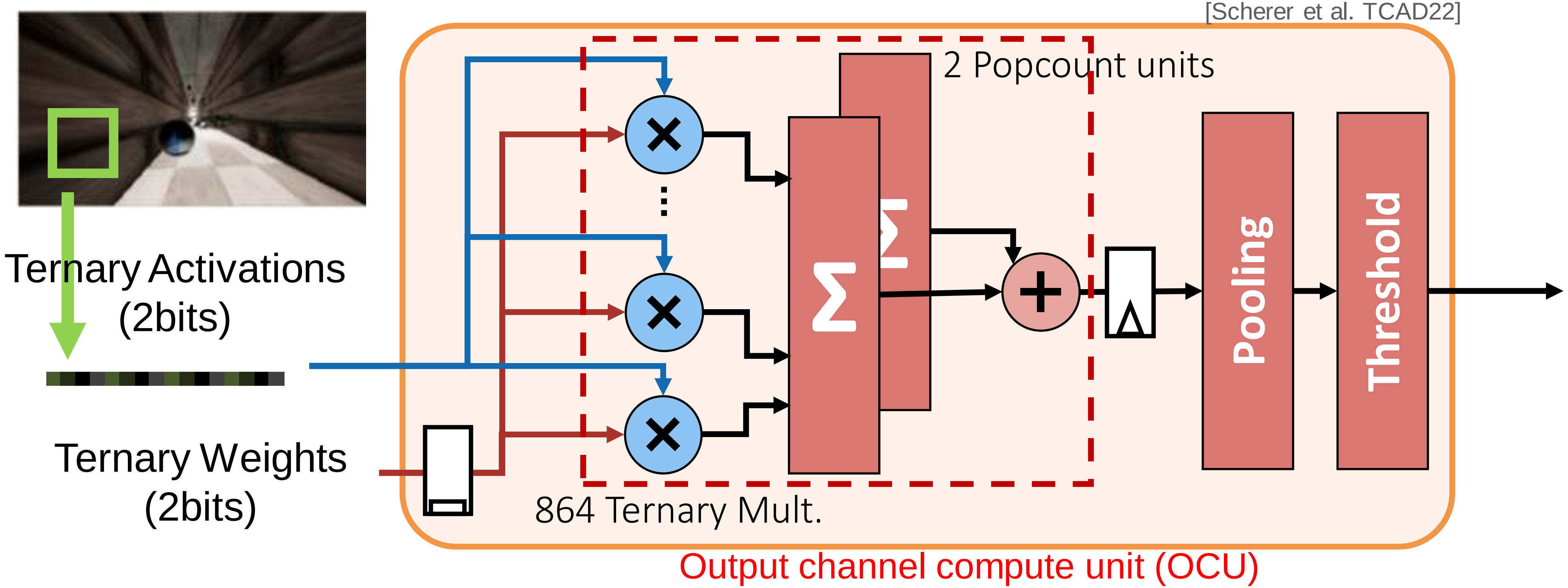


[Di Mauro HotChips22]

| Technology | 22 nm FDSOI |
|---|---|
| Chip Area | 9 mm$^2$ |
| 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** |

# CUTIE: Perception from Nyquist (Sampled) Sensors

[Scherer et al. TCAD22]

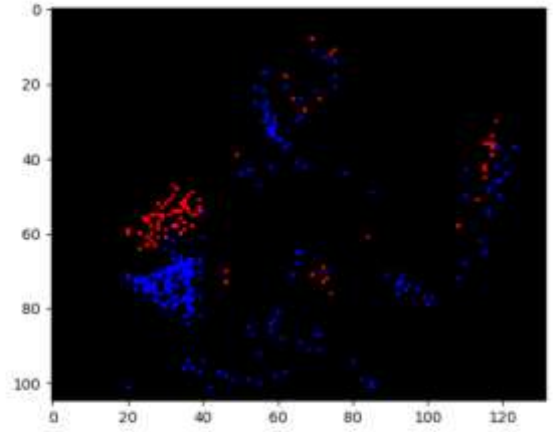


- **Completetely Unrolled Neural Inference Engine**: KxK window, all input channels, cycle-by-cycle sliding
- One OCU computes one output activation per cycle!
- Zeros in weights and activations, spatial smoothness of activations reduce switching activity
- 96 OCUs, 96 Input channels, 3x3 kernels: 96 * 96 * 3 * 3 **= 82'944 TMAC/cycle (~1fJ/MAC)**

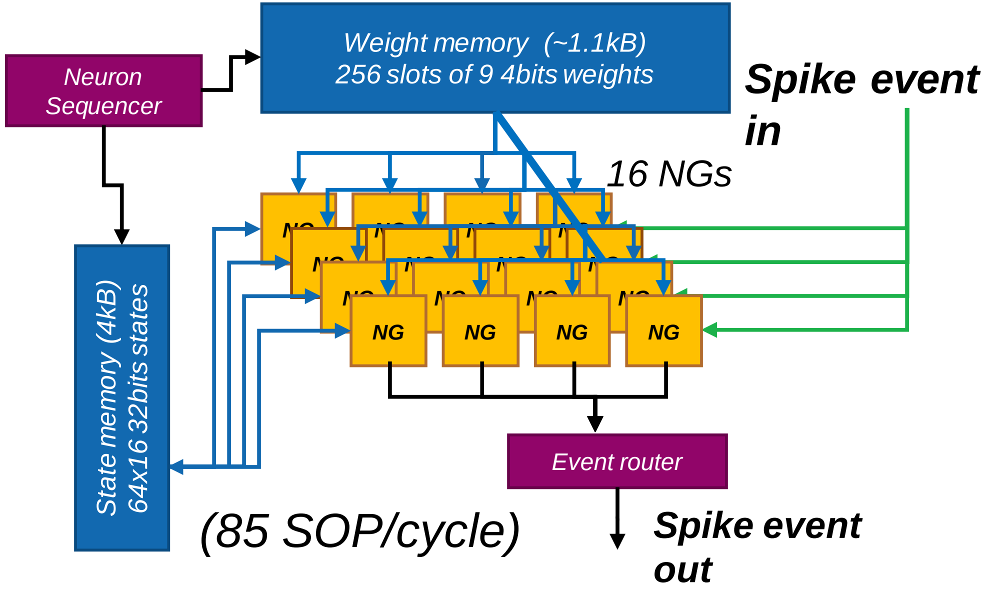
**Aggressive quantization and full specialization**

# SNE: Perception on Event Sensors

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

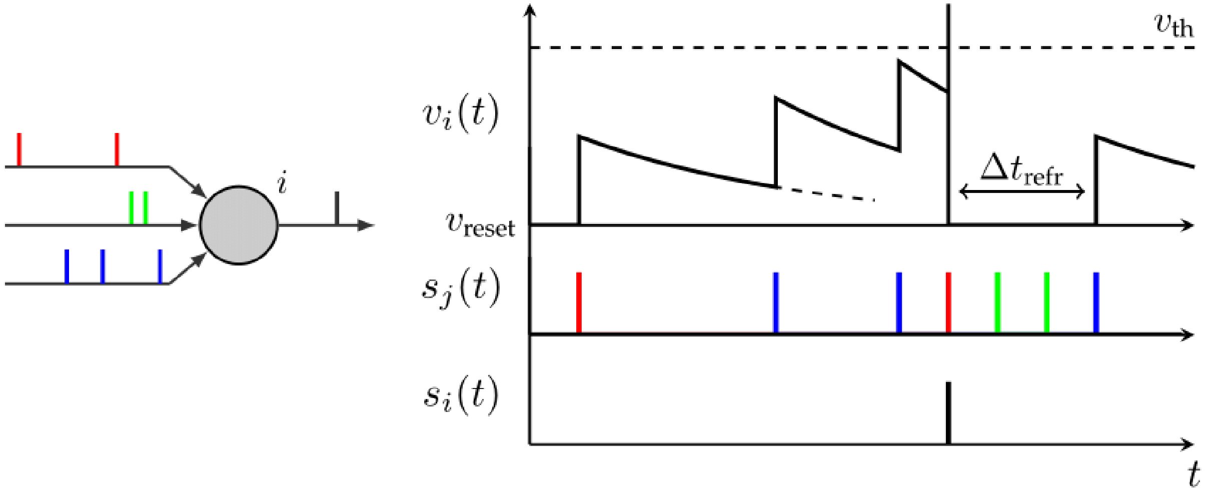


[Di Mauro et al. DATE22]



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

**Spiking Neural Engine (SNE)**



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

# General Purpose PE: Domain-Specialized RV32 Core

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

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!

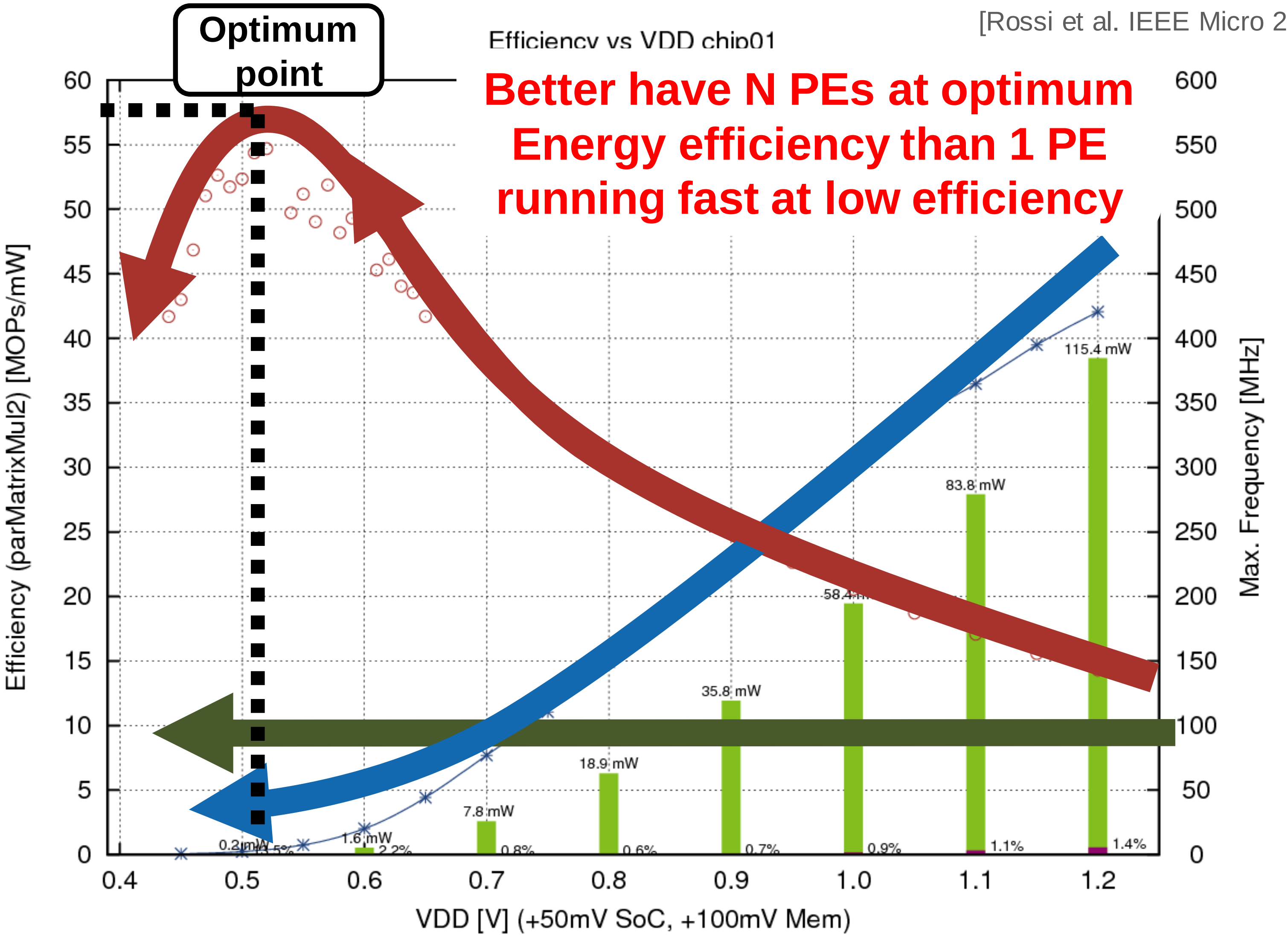
**Specialization Cost: Power,Area: 1.5x↑ but Time 15x↓ → E = PT 10x ↓**

ETH zürich

# 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 paralle with NN size**



[Rossi et al. IEEE Micro 2017]

Optimum point

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

# Advancing the SOA on all tasks

### RISC-V Cluster

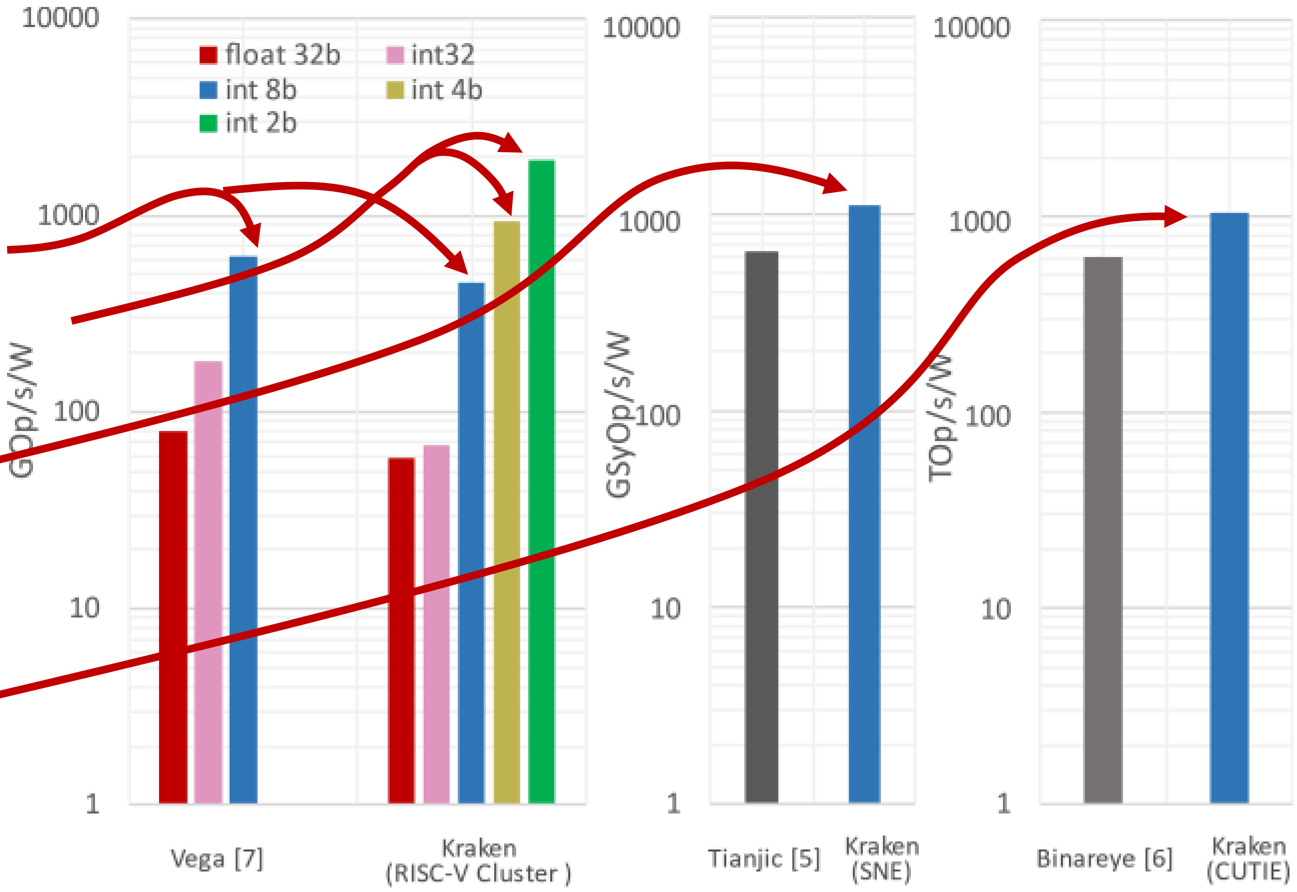
- Comparable 32bits-8bits SOA Energy efficiency to other PULPs [7]
- **The highest energy efficiency on sub-byte SIMD operations (4b-2b)**

### 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)**

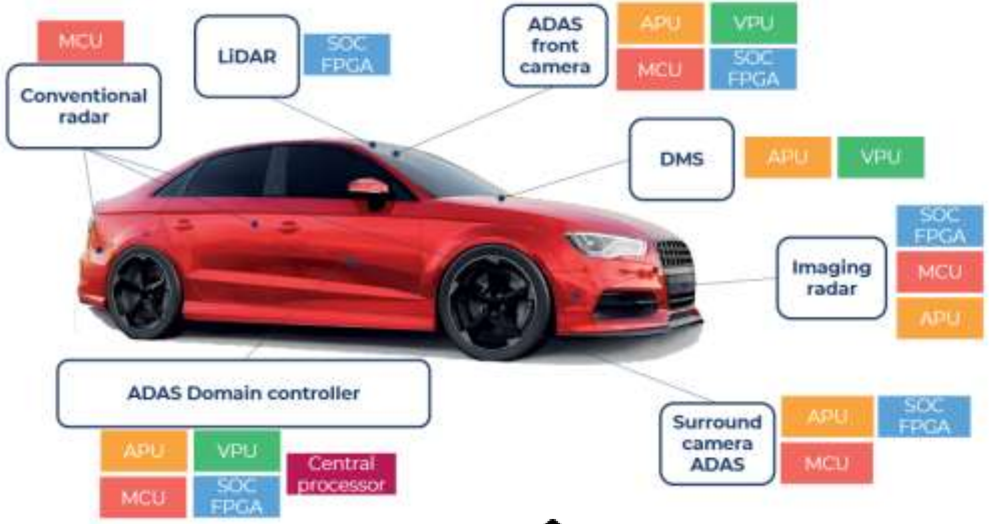
[5] L. Deng et al., “Tianjic,” JSSC 2020
[6] B. Moons et al., “Binareye,” CICC, 2018
[7] D. Rossi et al., “Vega,” JSSC 2022.

# From Drones to Cars: Stepping up

- <u>**Microcontroller class of devices**</u>
  - 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
- <u>**Powerful real-time architectures**</u>
  - 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
- <u>**Application class processors**</u>
  - NXP i.MX 8 Family
  - **ADAS, Infotainment**
  - Features: Cortex-A53, **Cortex-A72,**HW Virtualization, **GPUs**

2023 processors for active safety and ADAS

*(Source: Computing and AI for Automotive 2023, Yole Intelligence, February 2023)*



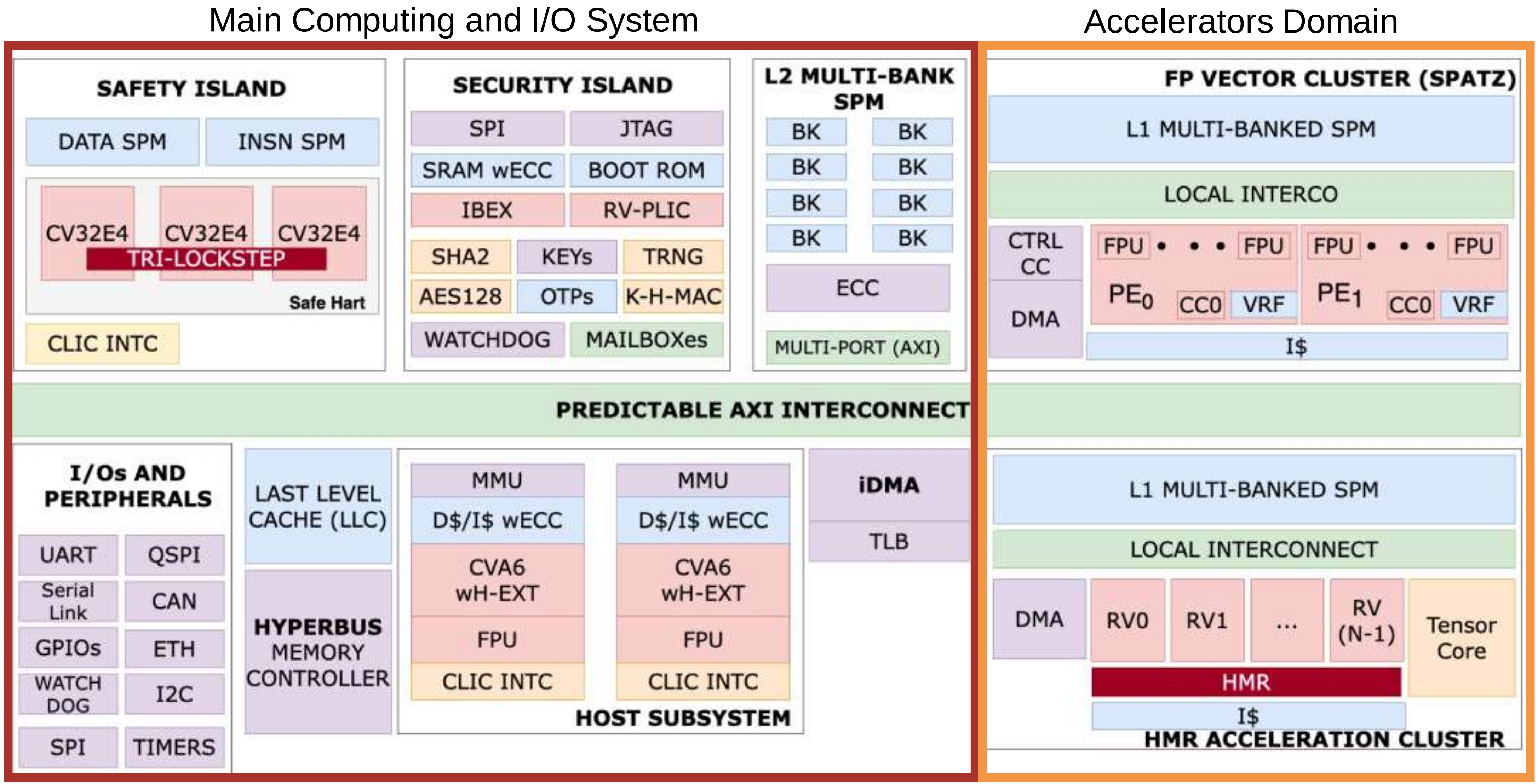
© Yole Intelligence 2023

Safe

Real-time

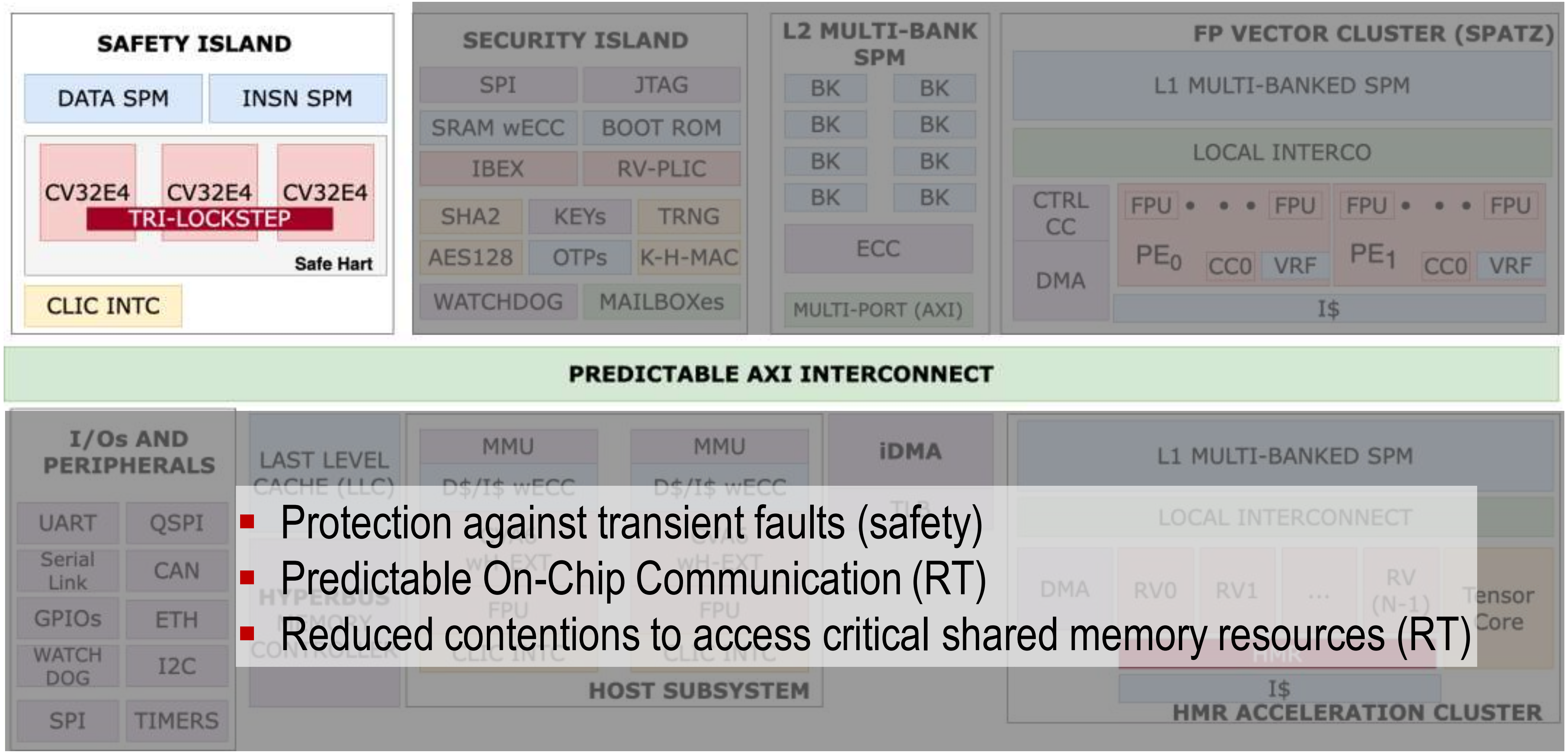
Secure

# Carfield: Efficiency + Safety, Security, RT-Predictability



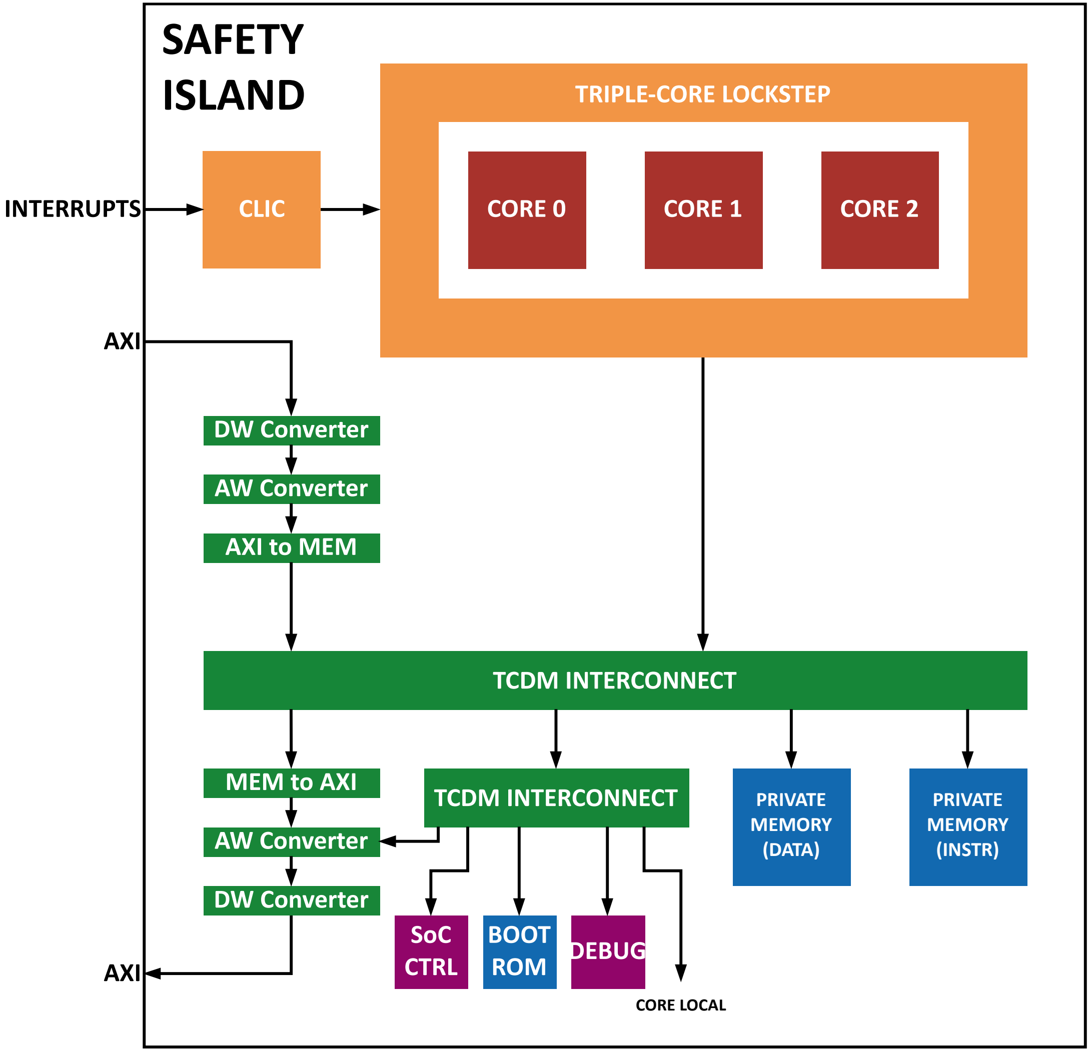
ETH zürich

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



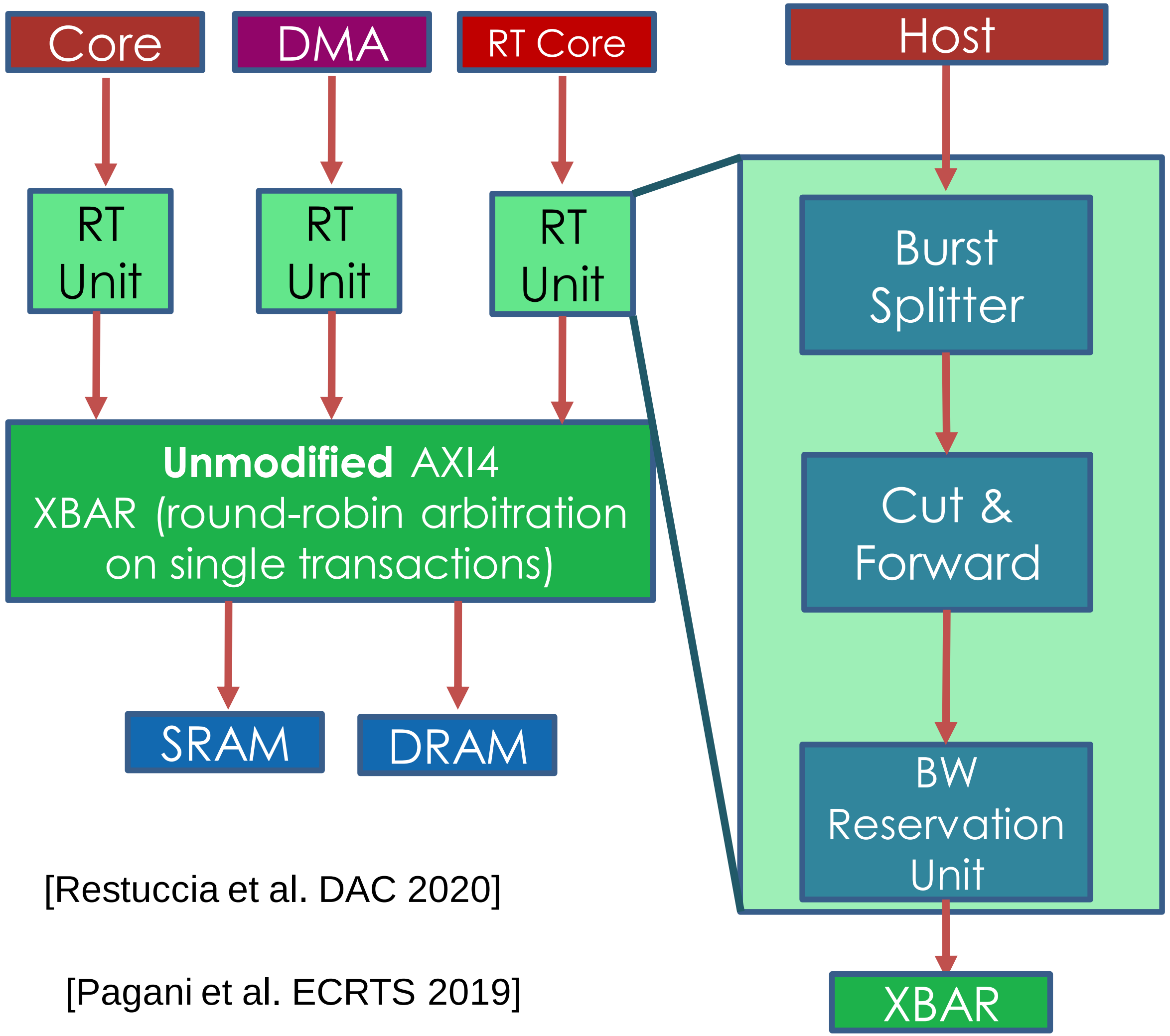
- Protection against transient faults (safety)
- Predictable On-Chip Communication (RT)
- Reduced contentions to access critical shared memory resources (RT)

# 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)

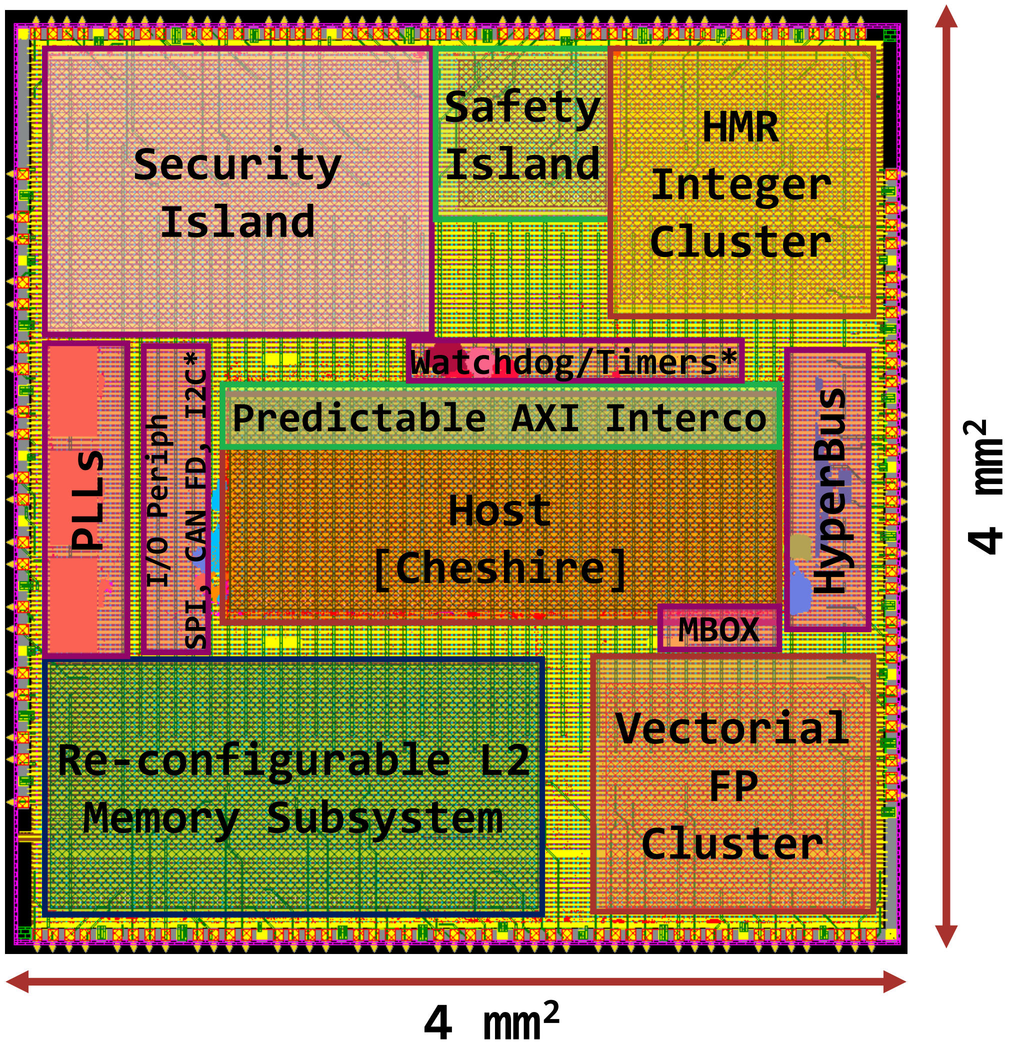


[Restuccia et al. DAC 2020]

[Pagani et al. ECRTS 2019]

- **AXI4 inherently unpredictable**
- **Minimally Intrusive Solution**
  - No huge buffering, limited additional logic
  - **Solution verified in systematic worst-case real-time analysis**
- **AXI Burst Splitter**
  - **Equalizes length of transaction**s 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)

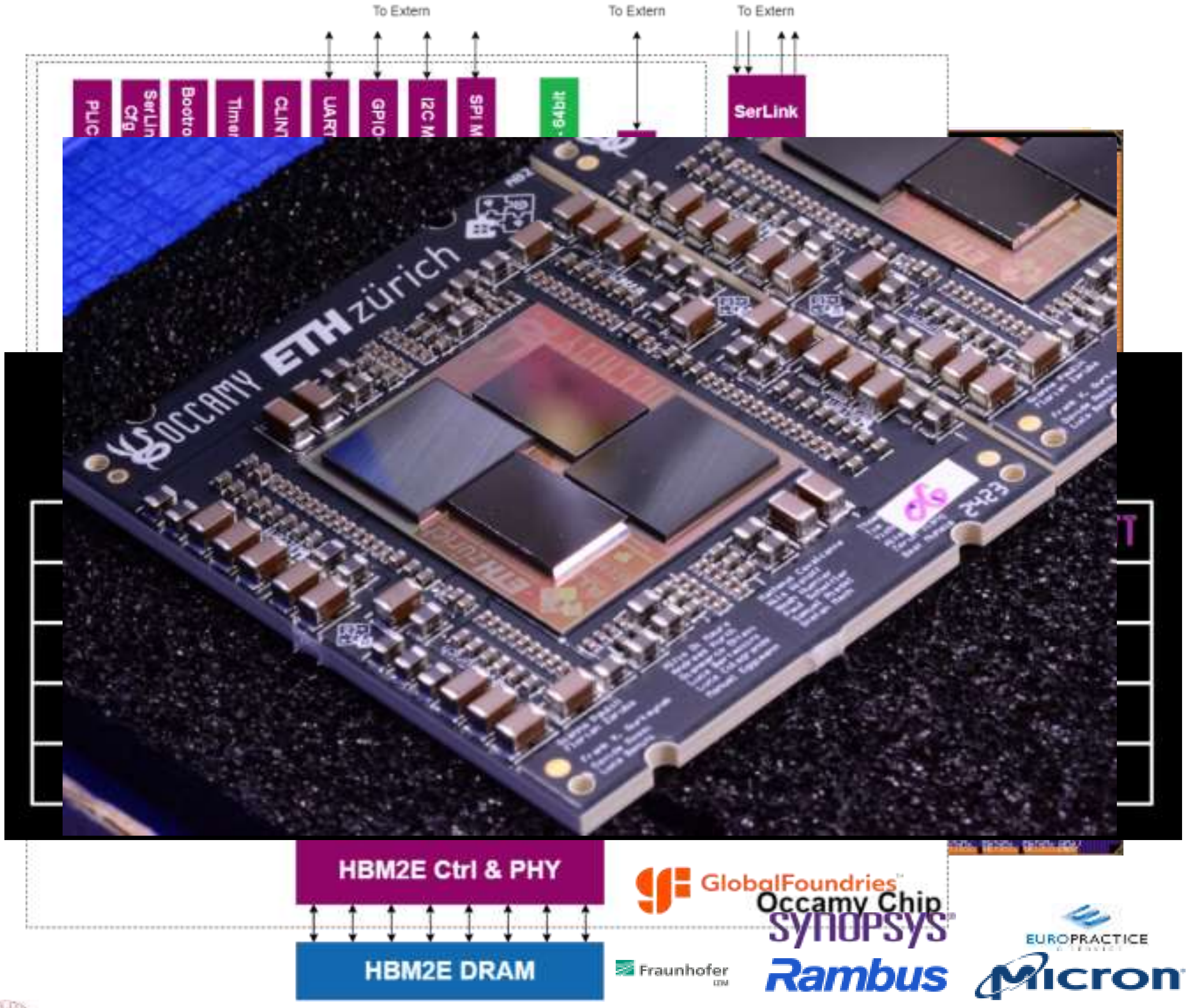
# Carfield SoC Flooplan – Taped out 11/2023



Modules marked with (*) are not in scale

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

# Toward Self-Driving Cars

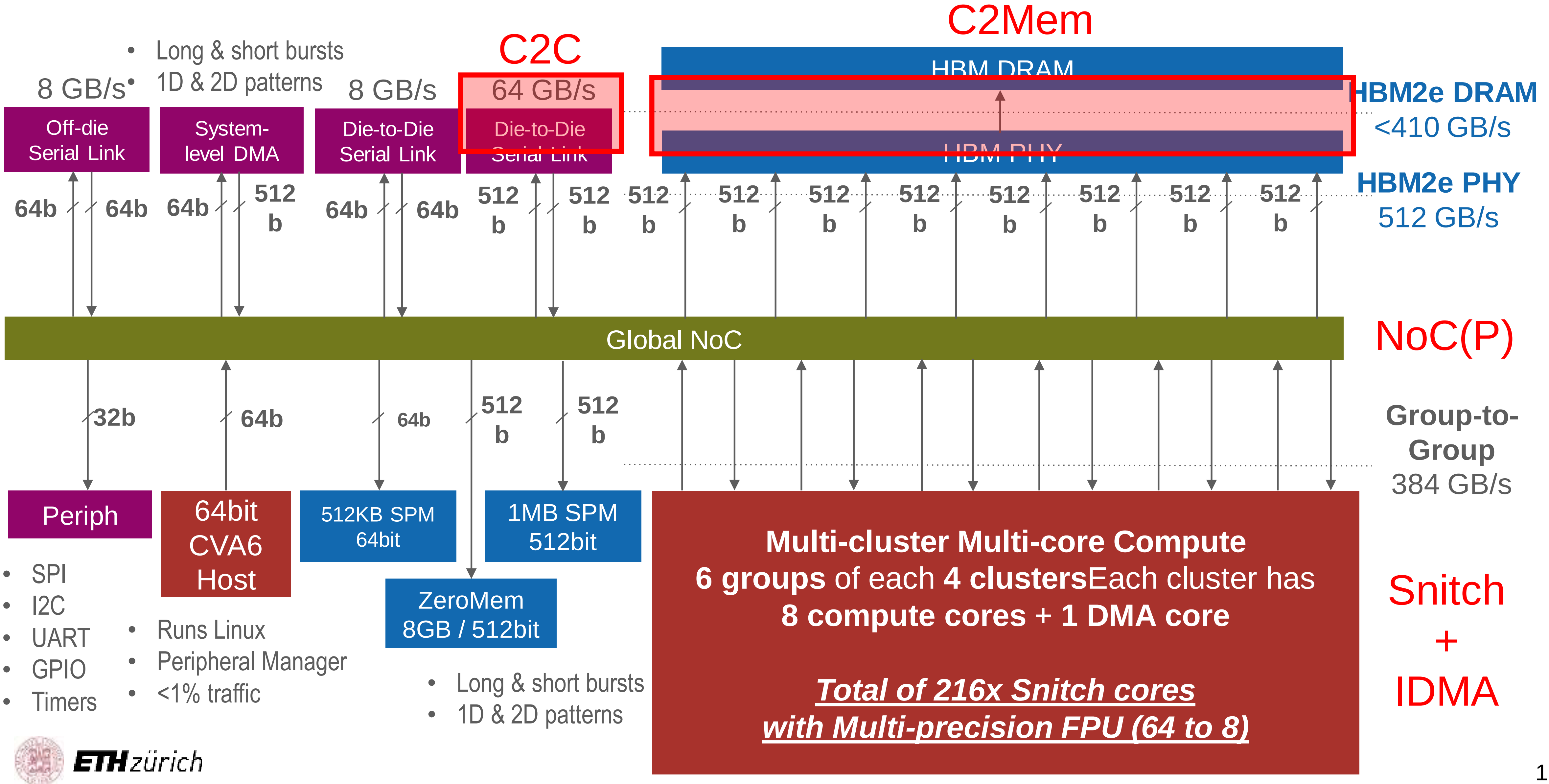


Occamy Chip

- 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

# Occamy: RISC-V goes HPC Chiplet!



C2Mem

C2C

- Long & short bursts
- 1D & 2D patterns

8 GB/s

8 GB/s

64 GB/s

Off-die Serial Link

System-level DMA

Die-to-Die Serial Link

Die-to-Die Serial Link

HBM DRAM

HBM PHY

**HBM2e DRAM**
<410 GB/s

**HBM2e PHY**
512 GB/s

64b 64b 64b 512 b 64b 64b 512 b 512 b 512 b 512 b 512 b 512 b 512 b 512 b 512 b 512 b

Global NoC

NoC(P)

32b 64b 64b 512 b 512 b

**Group-to-Group**
384 GB/s

Periph

64bit CVA6 Host

512KB SPM 64bit

1MB SPM 512bit

ZeroMem 8GB / 512bit

- SPI
- I2C
- UART
- GPIO
- Timers

- Runs Linux
- Peripheral Manager
- <1% traffic

- Long & short bursts
- 1D & 2D patterns

**Multi-cluster Multi-core Compute**
**6 groups** of each **4 clusters**Each cluster has **8 compute cores + 1 DMA core**

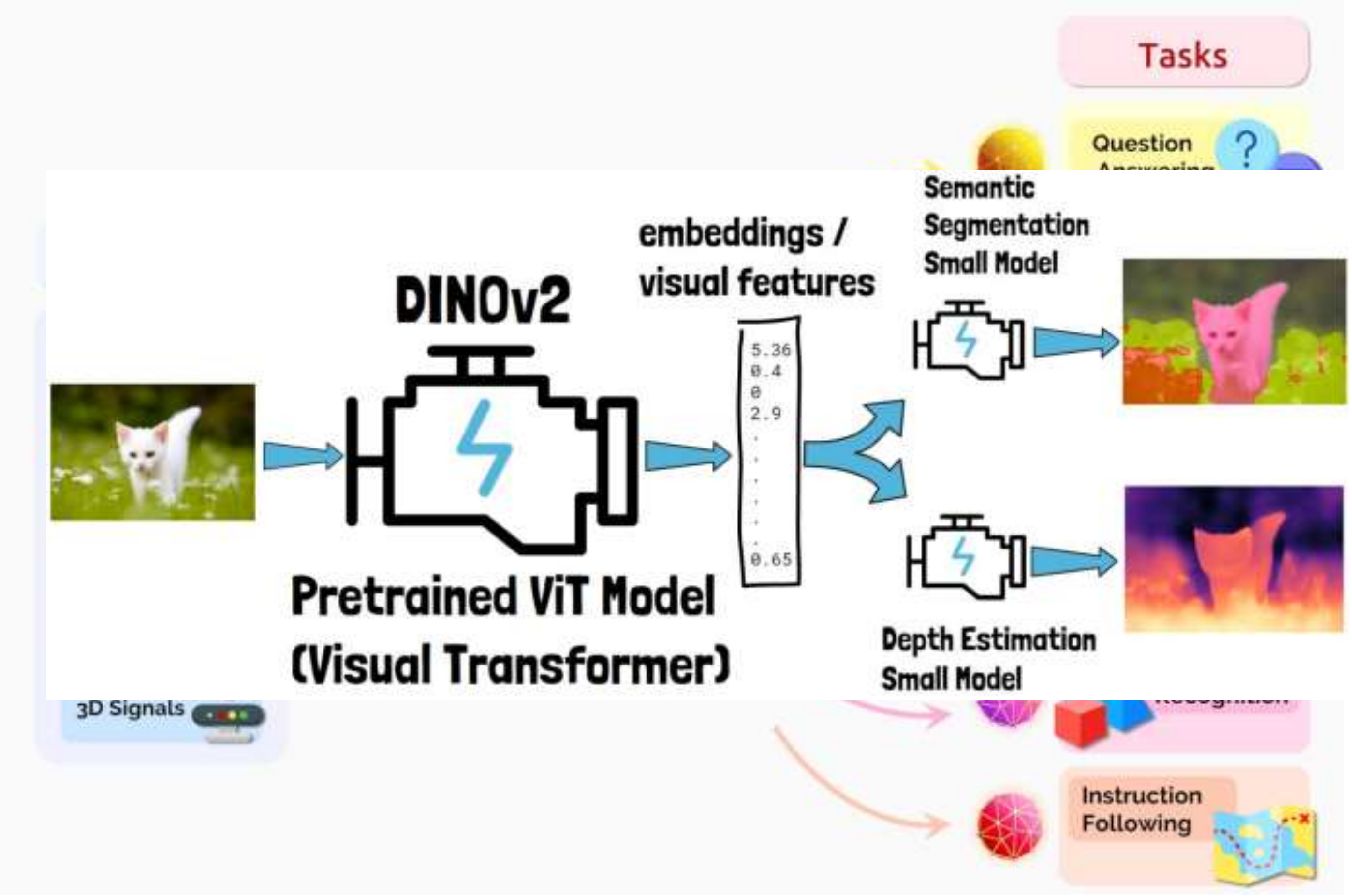
***Total of 216x Snitch cores***
***with Multi-precision FPU (64 to 8)***

Snitch + IDMA

ETH zürich

# What’s Next? The era  of Foundation Models

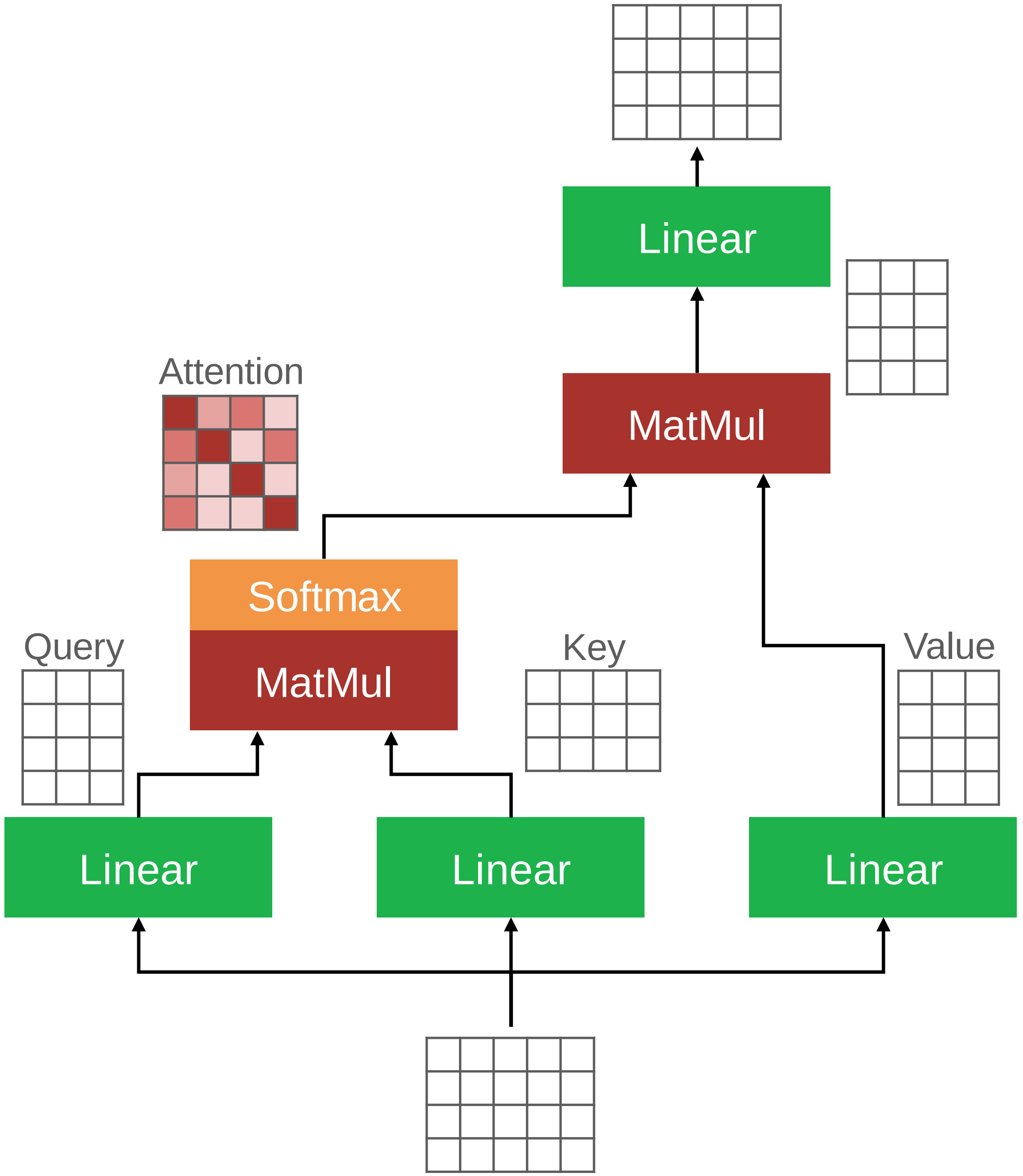
- Versatility and Multi-modality
  - Natural language processing, computer vision, robotics, biology, …
- Homogenization of models
  - **Transformers as *foundation models***
- **Self-supervision, Fine-tuning**
  - Self-supervised training on large-scale unlabeled dataset
  - Fine-tune (few layers) on 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)*.

# Challenges in *Attention*

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

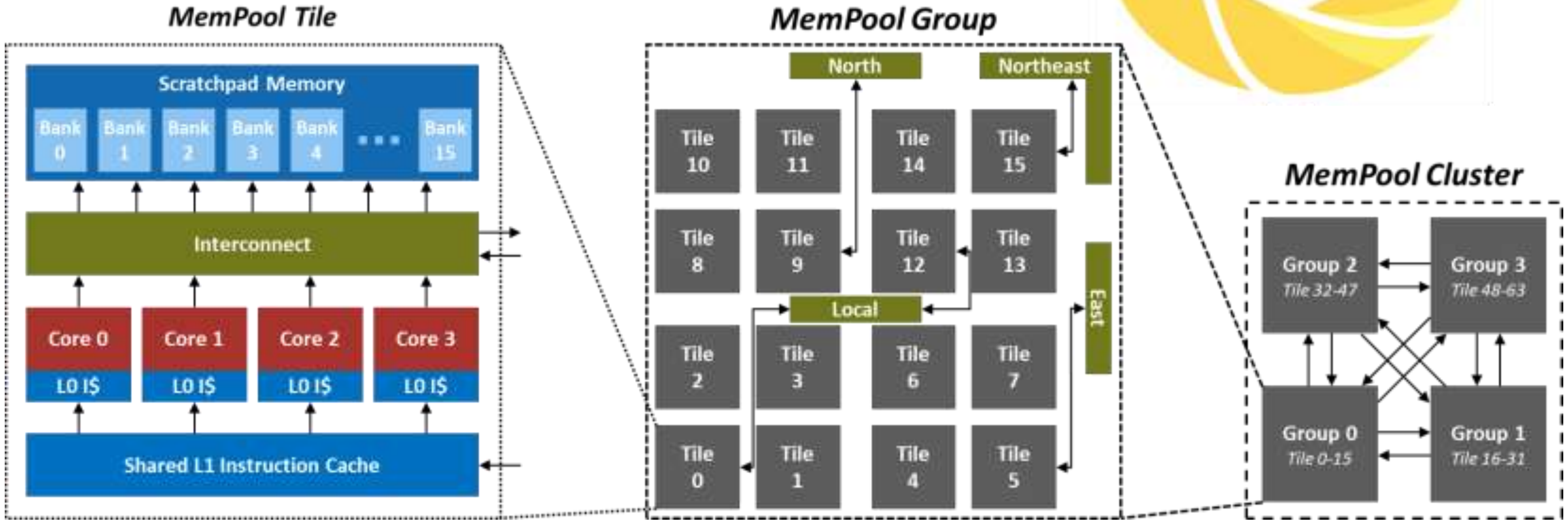


# Matmul Benefits from Large Shared-L1 clusters

- **Why?**
  - Better global latency tolerance if $L1_{size} > 2*L2_{latency}*L2_{bandwidth}$ (Little's law + double buffer)
  - Smaller data partitioning overhead
  - Larger Compute/Boundary bandwidth ratio: $N^3/N^2$ for MMUL grows linearly with N!

MemPool & Terapool

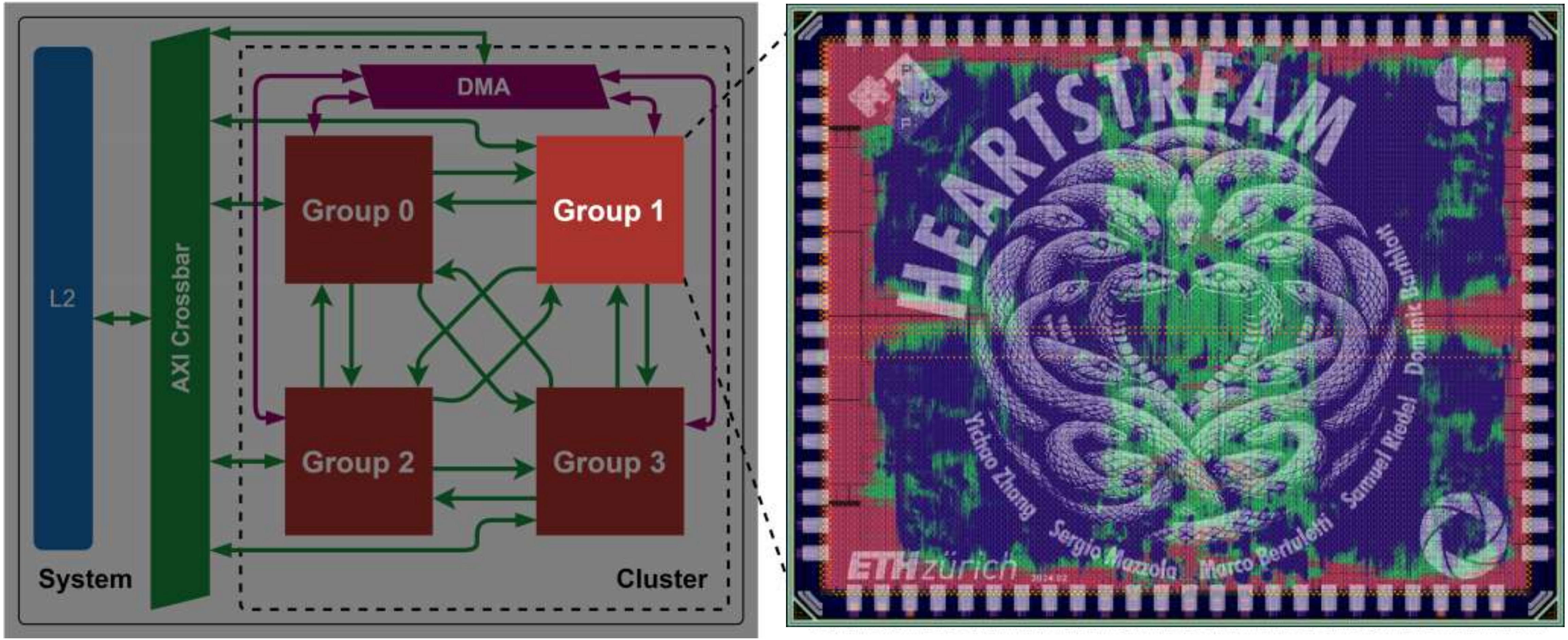
- **A large "MemPool"**
  - 256+ cores
  - 1+ MiB of shared L1 data memory
  - ≤ 10 cycle latency (Snitch can handle it)
- **Physical-aware design**
  - WC Frequency > 700+Mhz
  - Targeting iso-frequency with small cluster



**Butterfly Multi-stage Interconnect 0.3req/core/cycle, 5 cycles**

# MemPool + Integer Transformer Accelerator (ITA)

## Boosting dot product & matmul



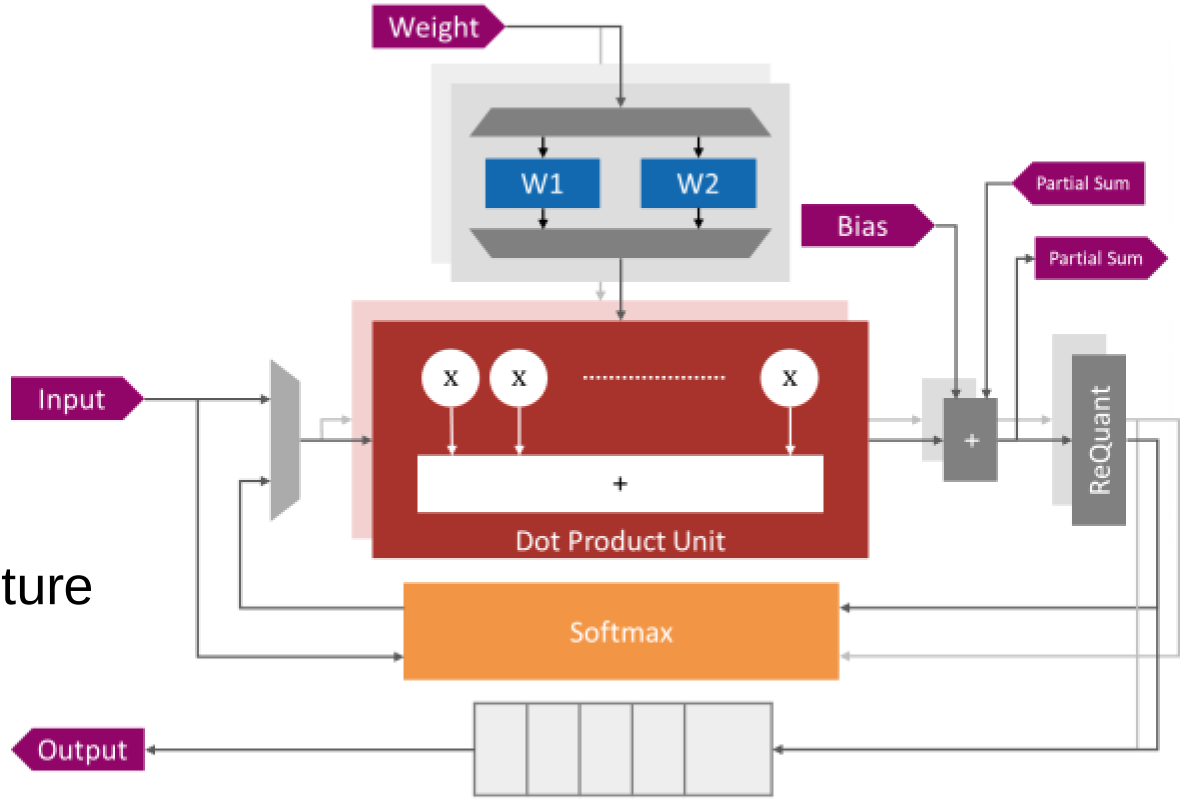
# MemPool + Integer Transformer Accelerator (ITA)

## Transformer Accelerator

- INT8 operand precision
- Builtin data marshaling & pipelined operation
- Streaming Softmax (support in QuantLib)
- Last layer for MH-Attention, the head accumulation computed in cores

## High-performance multi-core system

- Flexible, programmable snitch-based architecture
  - 192 cores split into 48 tiles
- Support convolutions and “exotic” operators

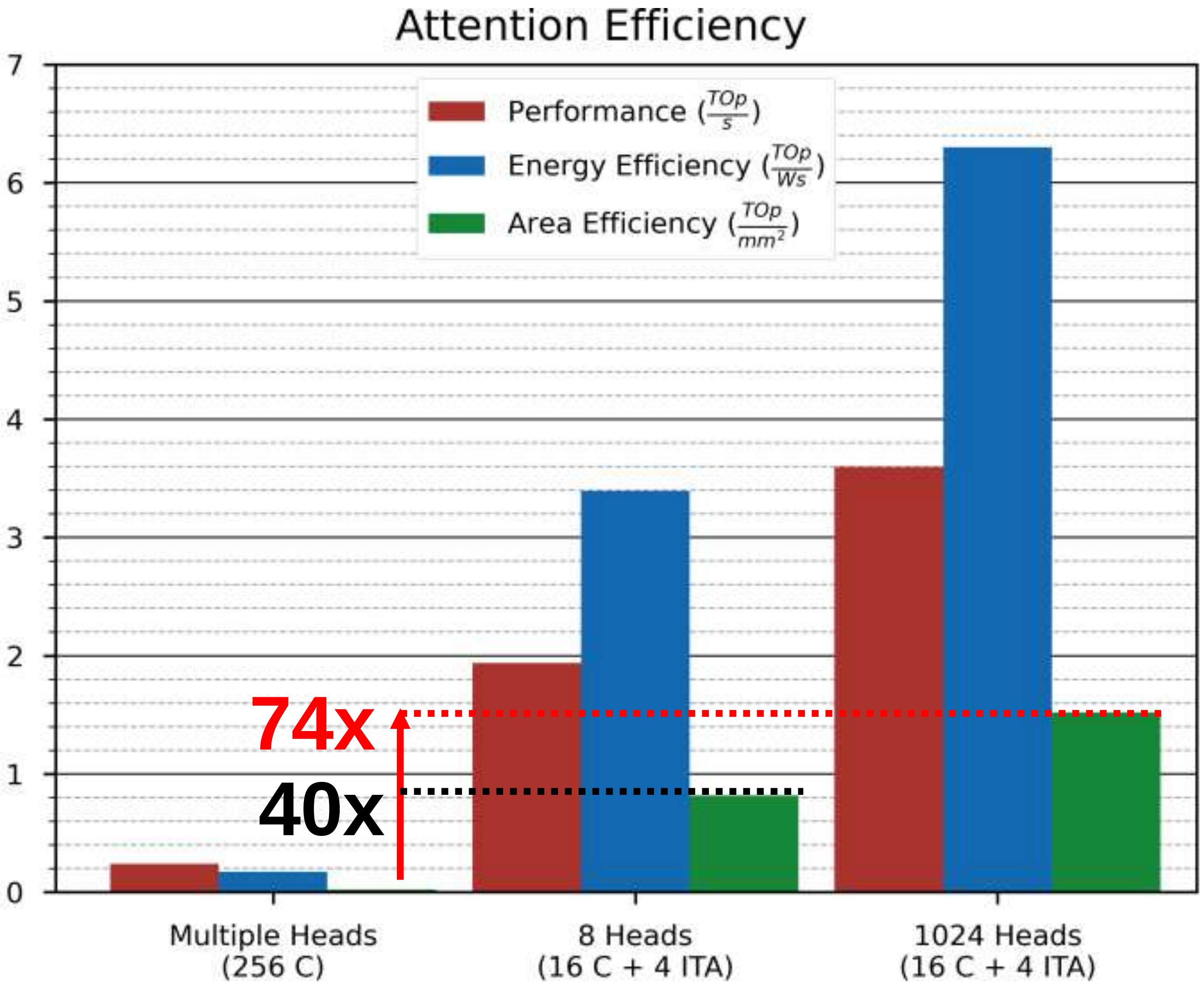


# Offloading Attention Operation to ITA

**Performance** increase of **15x**

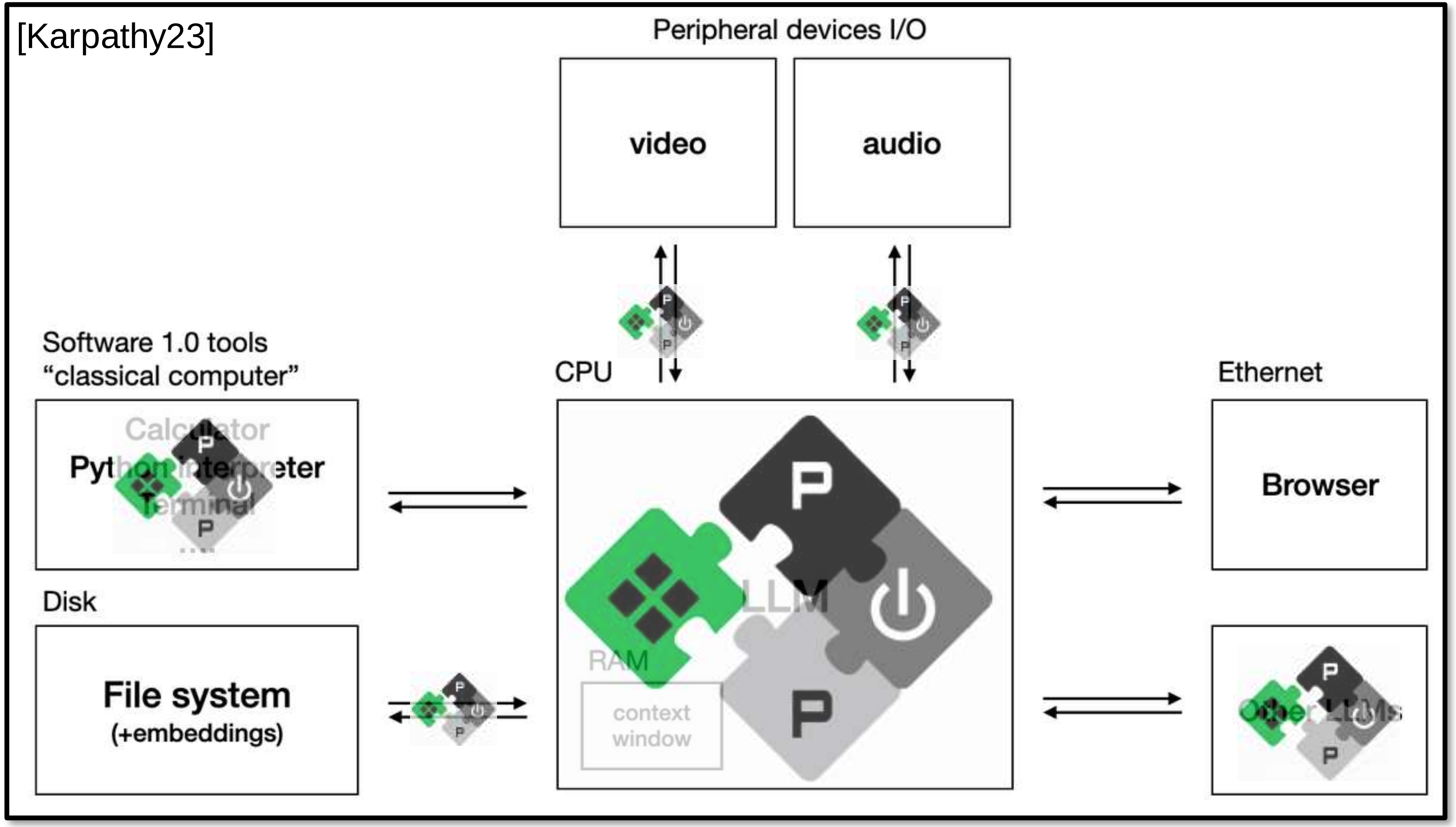
**Energy Efficiency** increase of **36x**

**Area Efficiency** increase of **74x**



ETHzürich

# Embodied AI vision:  Transformers everywhere?



Efficient

Safe

Real-time

Secure

# Thank You!