ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Enabling Sustainable AI

an Open Computing Platform Perspective

Luca Benini
lbenini@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

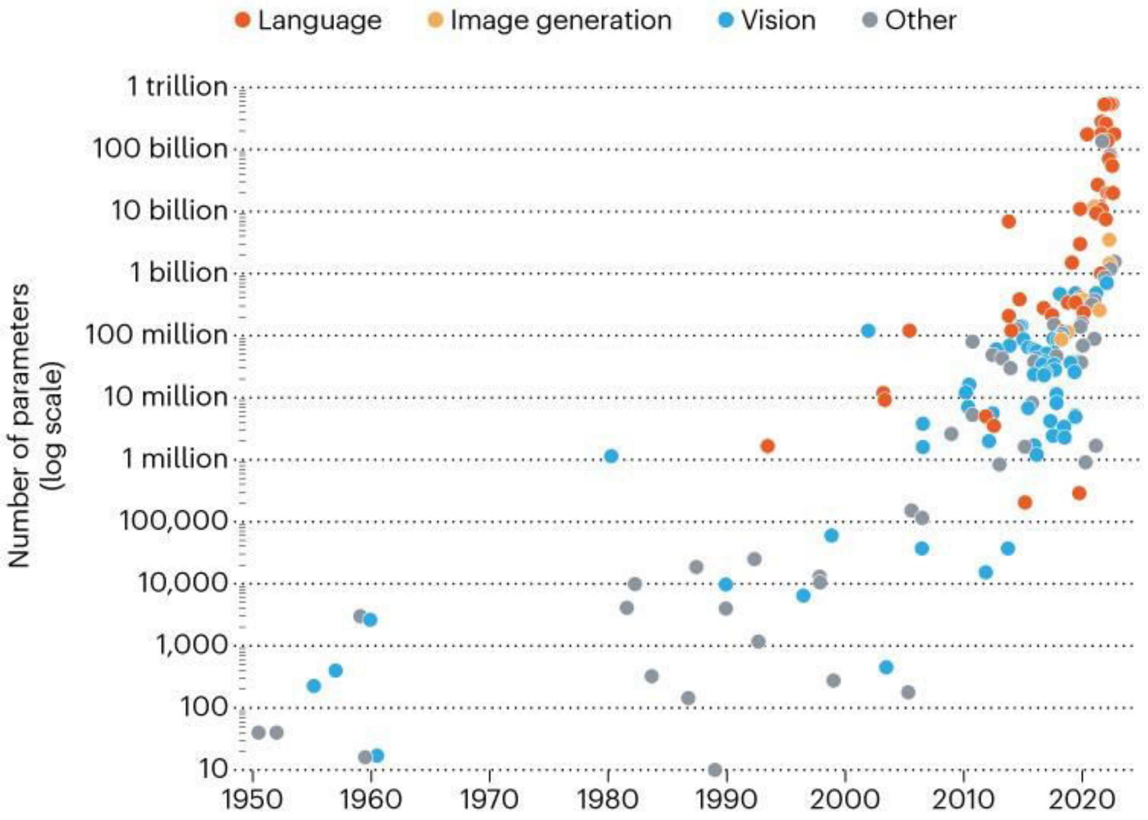
AI is Power Bound from Cloud to Edge

Datacenter PMAX < 150MW

On-car Computing PMAX < 1.5KW

THE DRIVE TO BIGGER AI MODELS

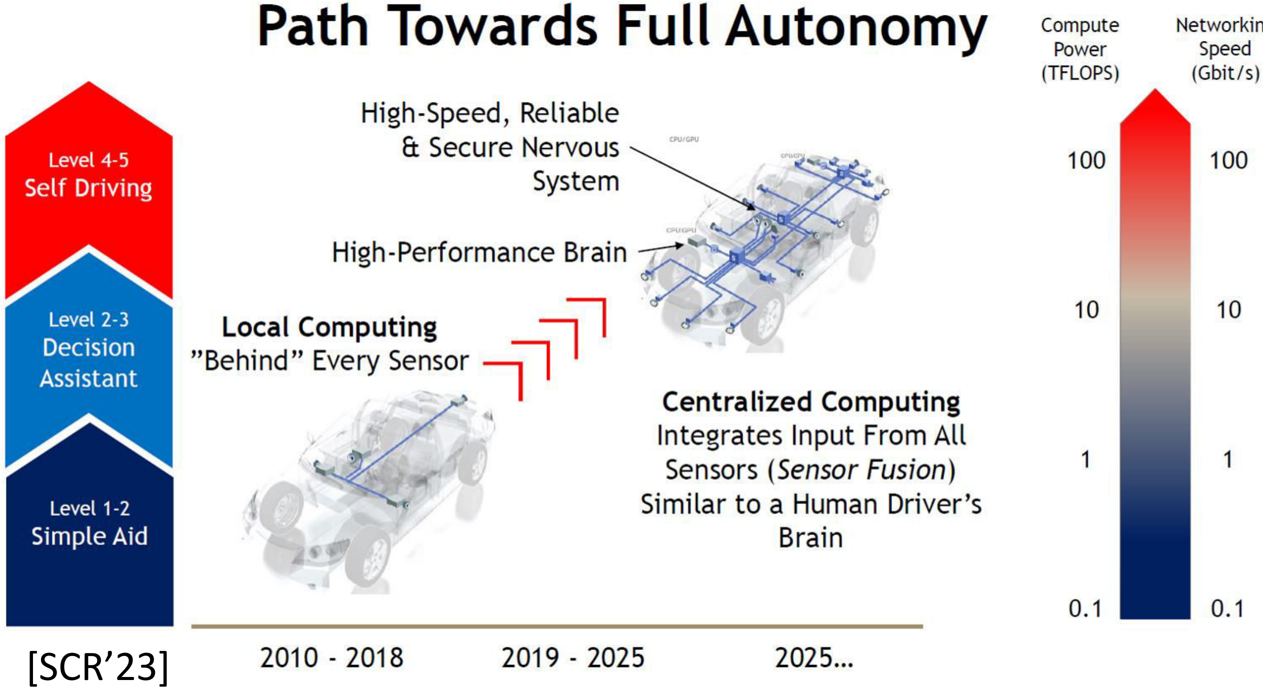
The scale of artificial-intelligence neural networks is growing exponentially, as measured by the models' parameters (roughly, the number of connections between their neurons)*.



[Nature'23]

10x every 2 years

Path Towards Full Autonomy



[SCR'23]

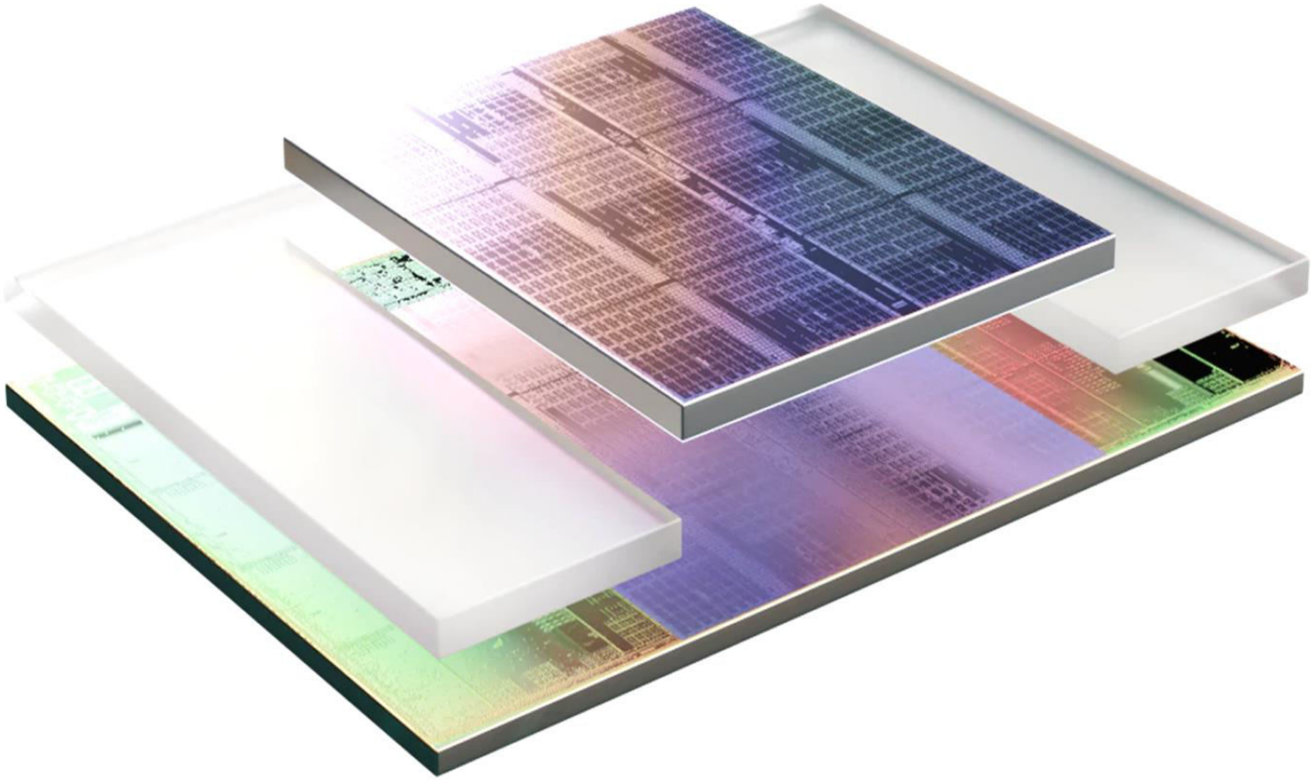
Energy Efficiency from Scaling

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

10x every 12 years...

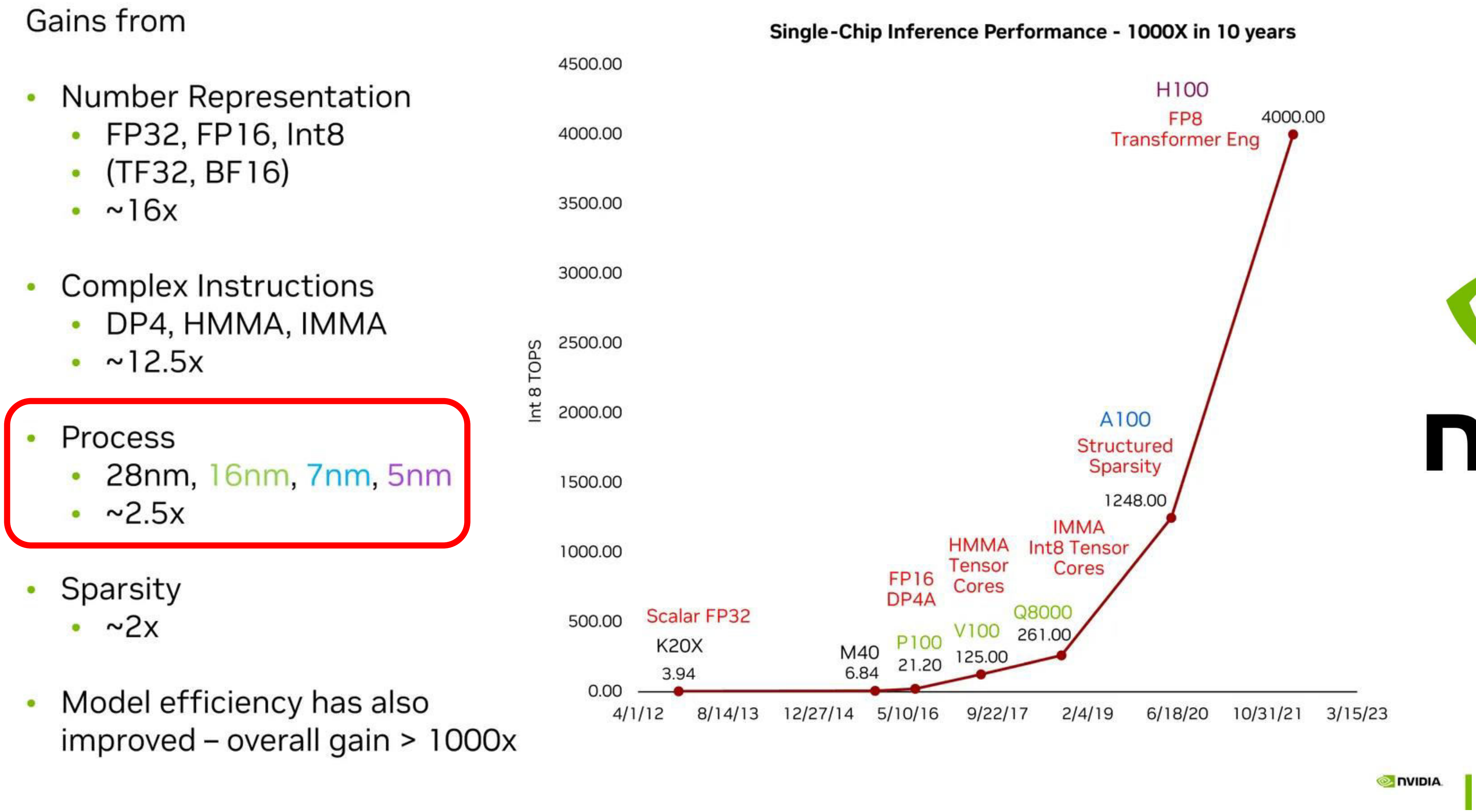
Energy-Efficient Computing: Core to Platform

- **Processing Element (PE)**
  - Efficient operations and local storage
- **Chip**
  - Data movement and storage hierarchy
- **Full platform**
  - IOs, main memory
  - Chiplets, 3D integration



The Renaissance of Design

Dally HotChips 2023



Gains from

- Number Representation
  - FP32, FP16, Int8
  - (TF32, BF16)
  - ~16x
- Complex Instructions
  - DP4, HMMA, IMMA
  - ~12.5x
- Process
  - 28nm, 16nm, 7nm, 5nm
  - ~2.5x
- Sparsity
  - ~2x
- Model efficiency has also improved – overall gain > 1000x

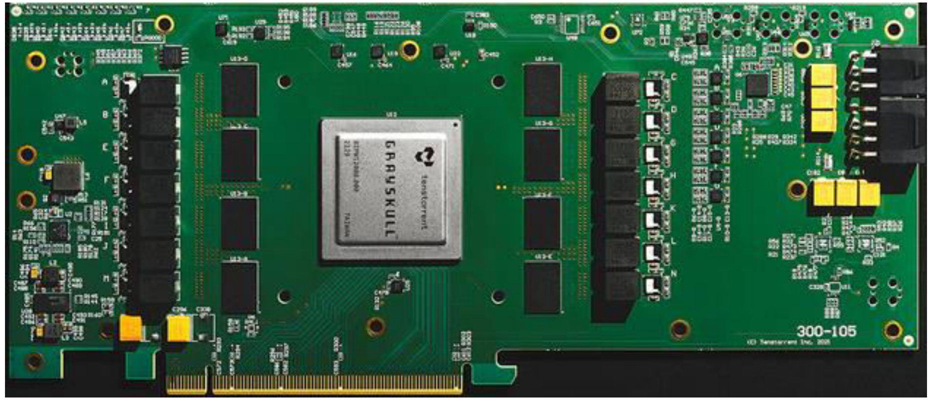
Single-Chip Inference Performance - 1000X in 10 years

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



tenstorrent

RISC-V is Accelerating

Chips JU

EuroHPC Joint Undertaking

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

MeitY
Government of India

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

RISC-V®

QUINTAURIS

BOSCH Infineon NORDIC SEMICONDUCTOR NXP Qualcomm ST

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

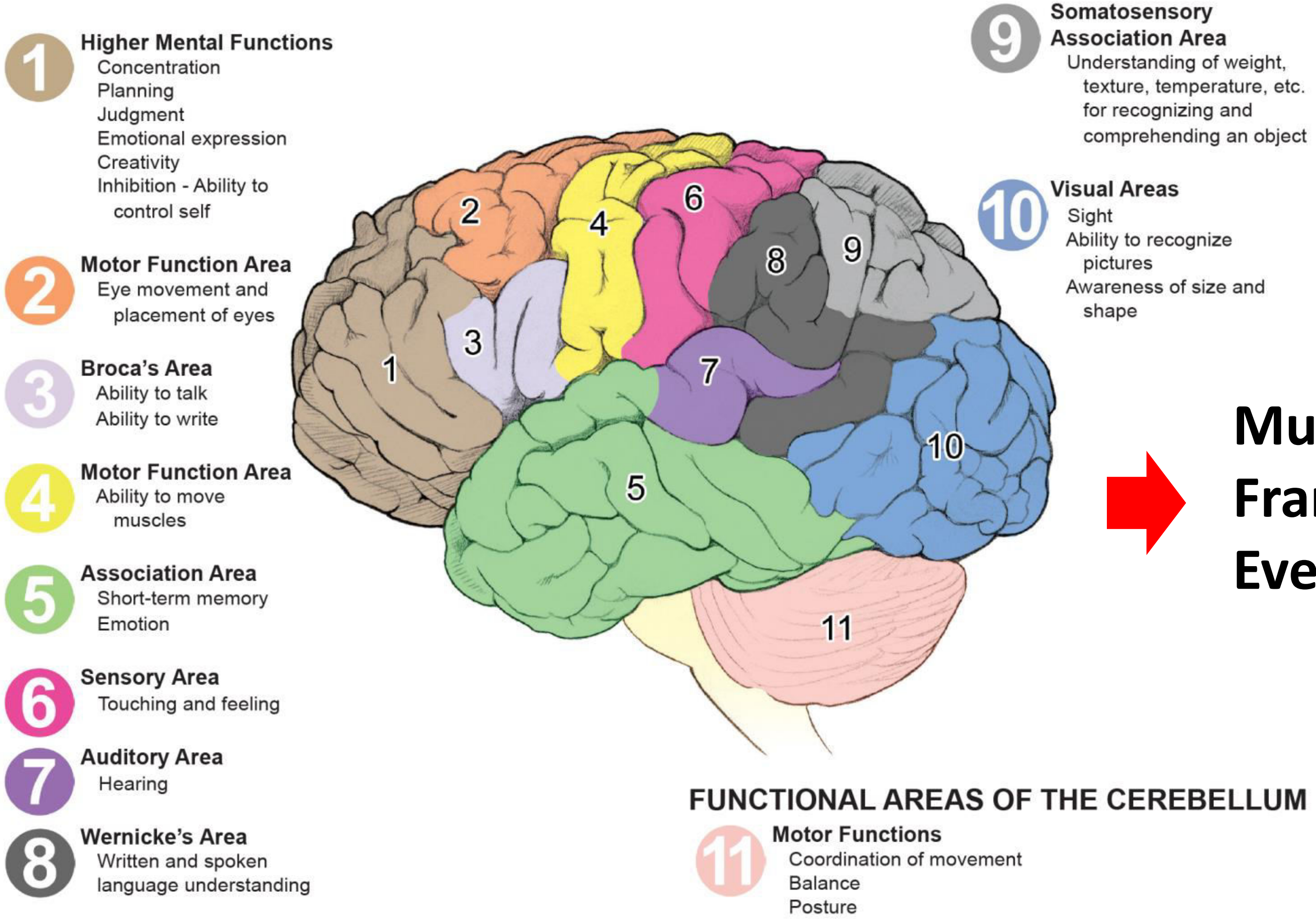
RISE
RISC-V Software Ecosystem

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

ANDES TECHNOLOGY Google Imagination intel MEDIATEK
NVIDIA Qualcomm Red Hat Rivos SAMSUNG
SiFive 達摩院 DAMO ACADEMY VENTANA

Heterogeneous Specialization for AI

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

➡ **Multi-sensory input**
Frame-based
Event based

➡ **Perception**
Fusion
Reasoning

Focus on Processing Element: Specialize

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

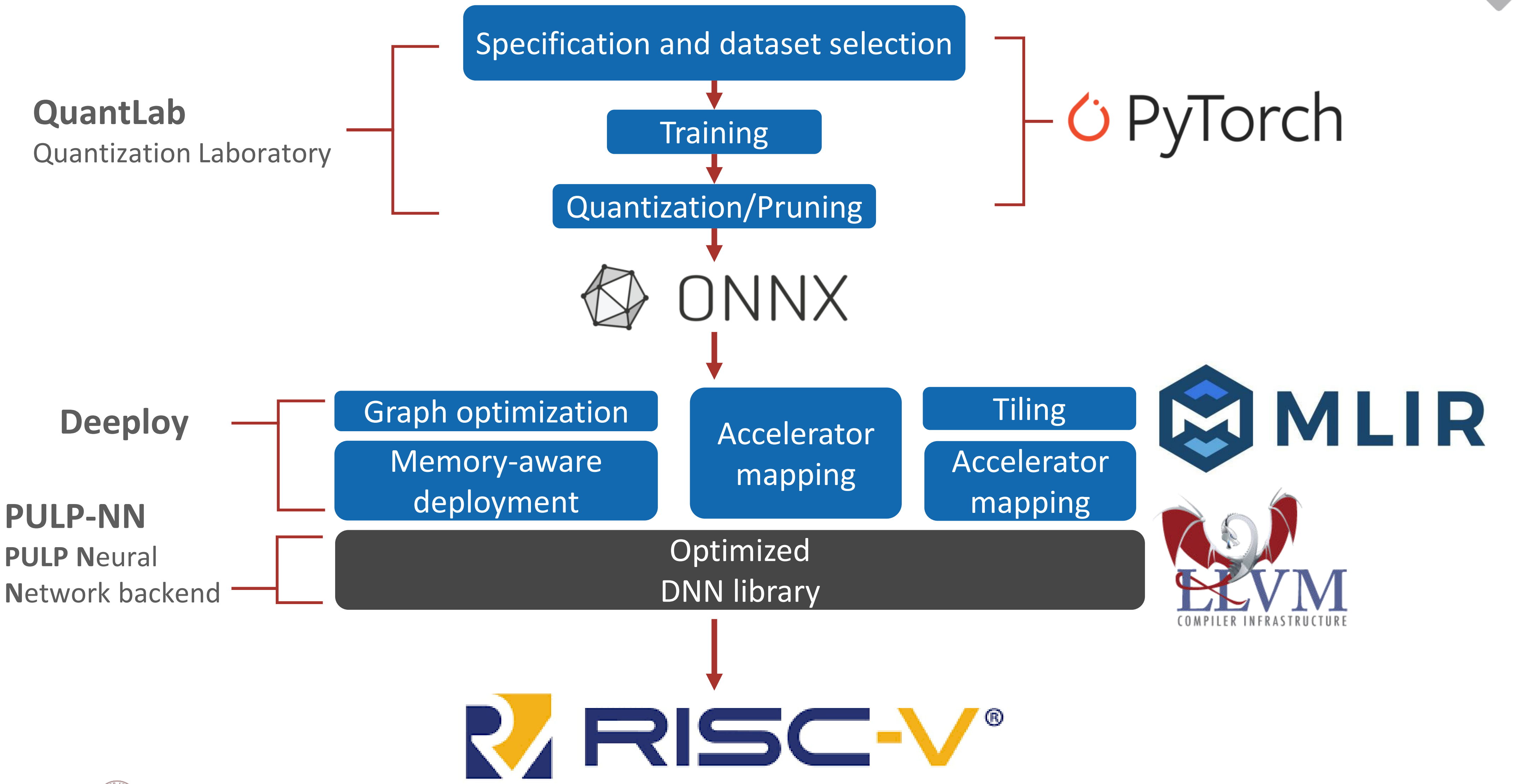
```
Init NN-RF (outside of the loop)
lp.setup
pv.nnsdotup.h  s0,ax1,9
pv.nnsdotsp.b  s1, aw2, 0
pv.nnsdotsp.b  s2, aw4, 2
pv.nnsdotsp.b  s3, aw3, 4
pv.nnsdotsp.b  s4, ax1, 14
end
```

RISC-V core

15x less instructions than Vanilla!

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

Fully Open-Source SW Stack for AI with RISC-V!



QuantLab
Quantization Laboratory

Specification and dataset selection

Training

Quantization/Pruning

PyTorch

ONNX

Deploy

Graph optimization

Memory-aware deployment

Accelerator mapping

Tiling

Accelerator mapping

MLIR

PULP-NN
PULP Neural
Network backend

Optimized DNN library

LLVM
COMPILER INFRASTRUCTURE

RISC-V®

Open-Source RISC-V Hardware: PULP

RISC-V Cores and Vector Units

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

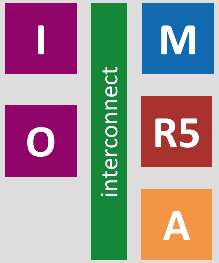
Peripherals

JTAG	SPI
UART	I2S
DMA	GPIO

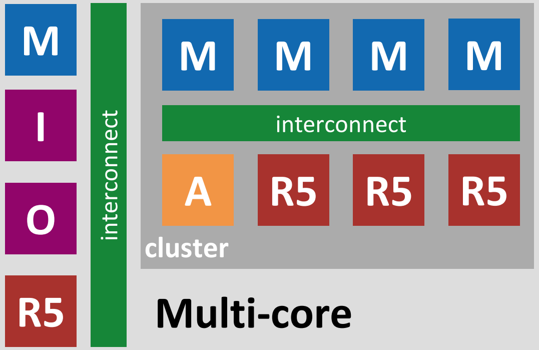
Interconnects

LIC	HCI
APB	FlooNoC
AXI4	

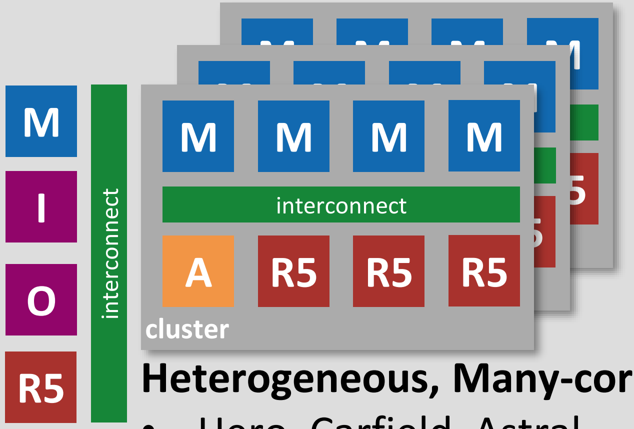
Platforms



Single core
- PULPino, PULPissimo
- Cheshire



Multi-core
- OpenPULP
- ControlPULP



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

IOT

HPC

Accelerators and ISA extensions

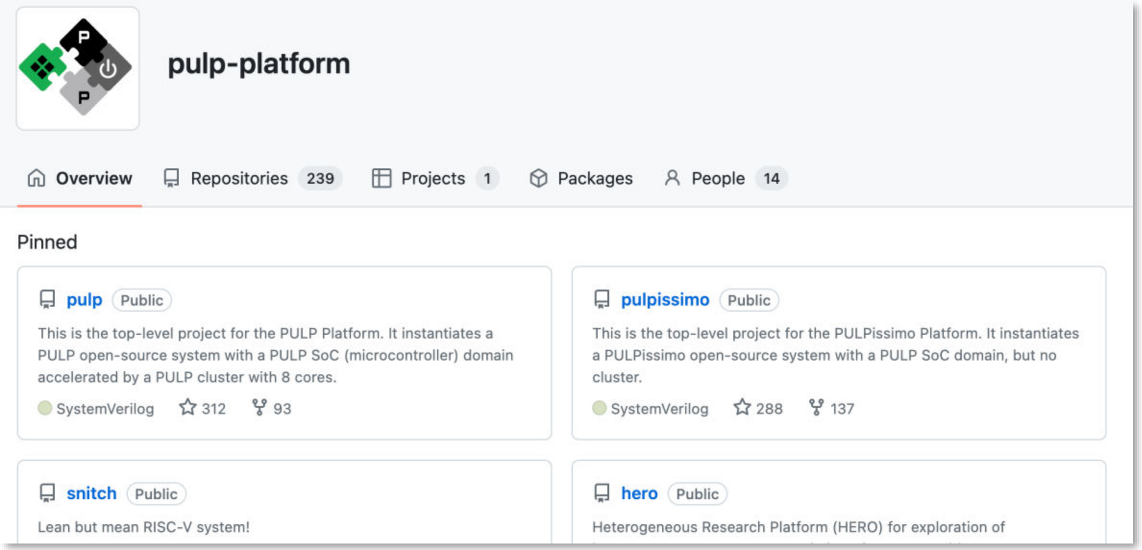
XpulpNN, XpulpTNN	ITA (Transformers)	RBE, NEUREKA (QNNs)	FFT (DSP)	REDMULE (FP-Tensor)

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.



pulp-platform

Pinned

pulp Public

This is the top-level project for the PULP Platform. It instantiates a PULP open-source system with a PULP SoC (microcontroller) domain accelerated by a PULP cluster with 8 cores.

SystemVerilog ☆ 312 93

pulpissimo Public

This is the top-level project for the PULPissimo Platform. It instantiates a PULPissimo open-source system with a PULP SoC domain, but no cluster.

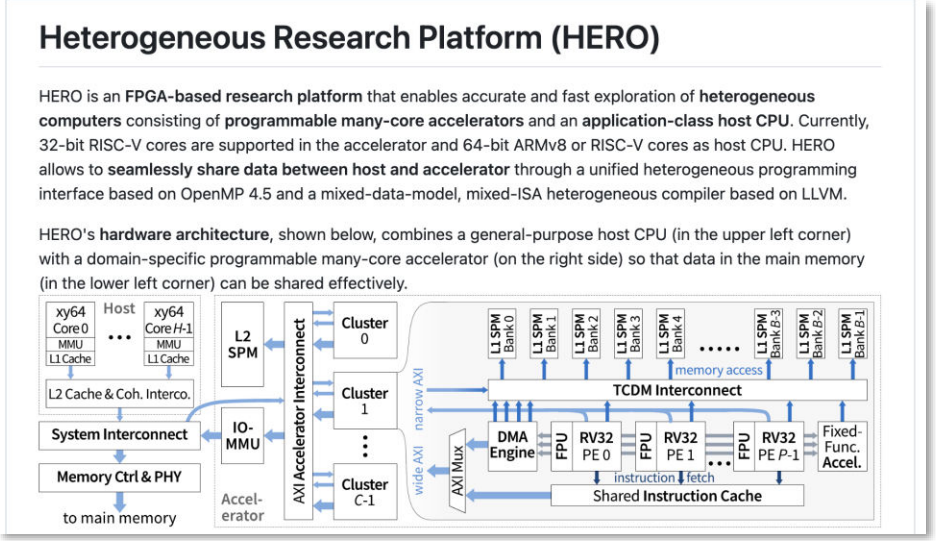
SystemVerilog ☆ 288 137

snitch Public

Lean but mean RISC-V system!

hero Public

Heterogeneous Research Platform (HERO) for exploration of

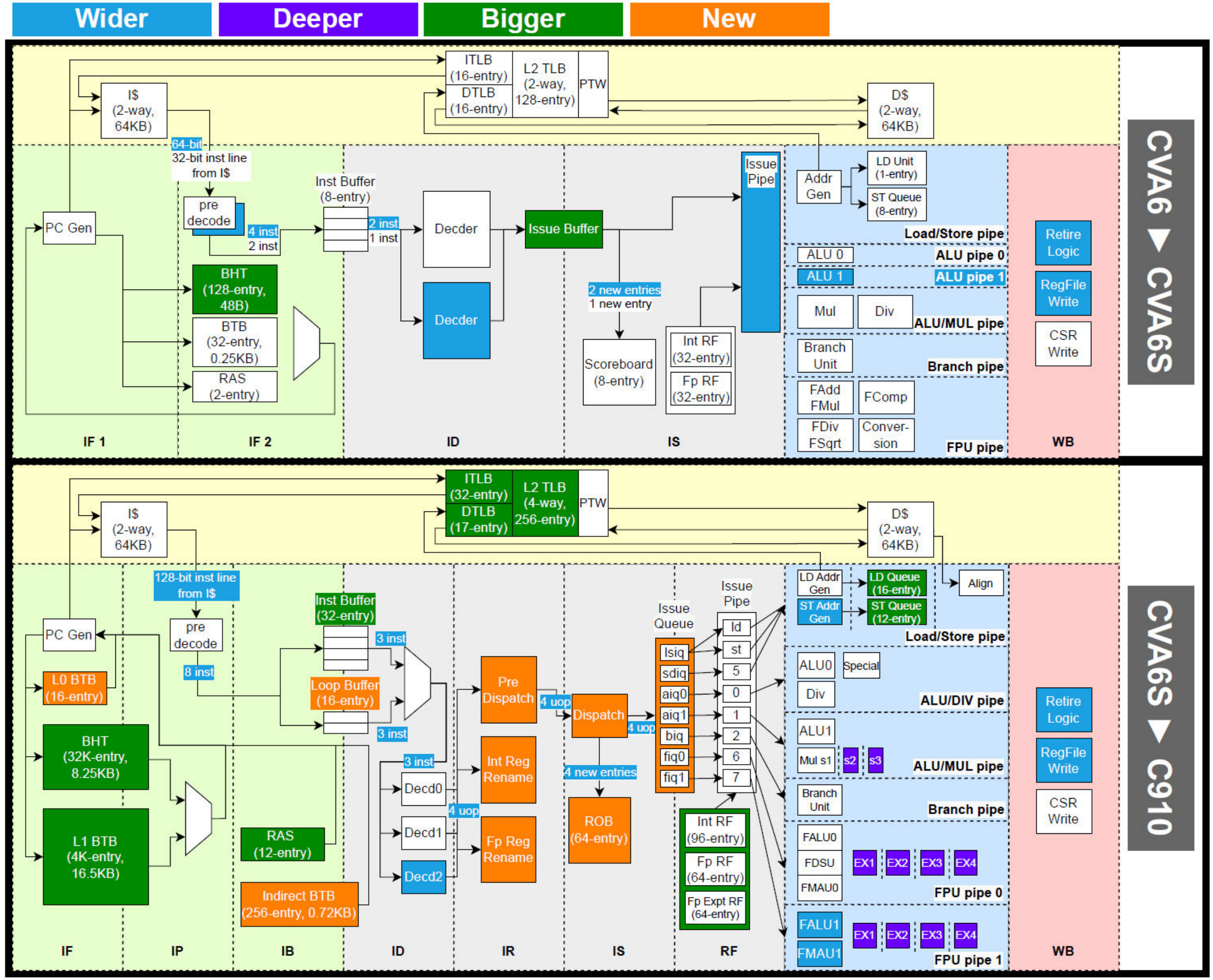


Heterogeneous Research Platform (HERO)

HERO is an **FPGA-based research platform** that enables accurate and fast exploration of **heterogeneous computers** consisting of **programmable many-core accelerators** and an **application-class host CPU**. Currently, 32-bit RISC-V cores are supported in the accelerator and 64-bit ARMv8 or RISC-V cores as host CPU. HERO allows to **seamlessly share data between host and accelerator** through a unified heterogeneous programming interface based on OpenMP 4.5 and a mixed-data-model, mixed-ISA heterogeneous compiler based on LLVM.

HERO's **hardware architecture**, shown below, combines a general-purpose host CPU (in the upper left corner) with a domain-specific programmable many-core accelerator (on the right side) so that data in the main memory (in the lower left corner) can be shared effectively.

Open RISC-V PEs: In-order → Superscalar → OoO



CVA6

OPENHW FOUNDATION PROVEN PROCESSOR IP

CVA6S+

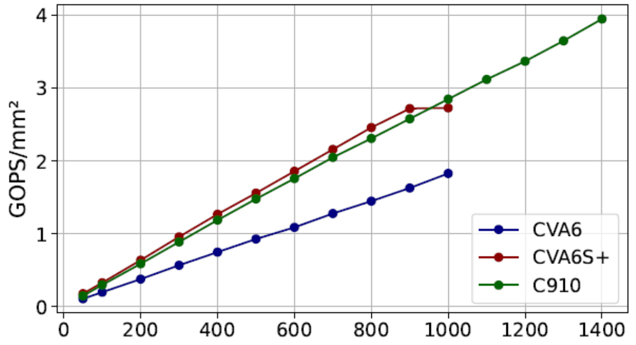
life.augmented

cea

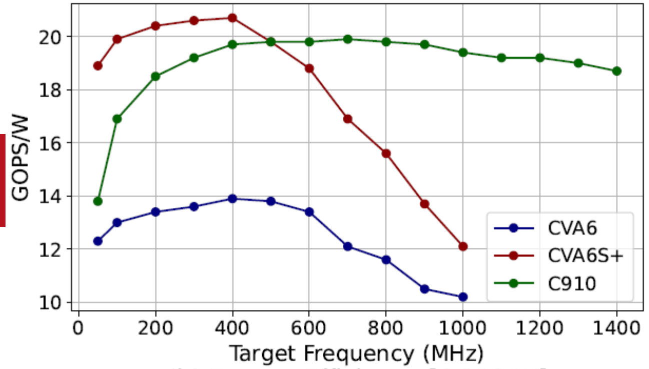
THALES

XuanTie C910

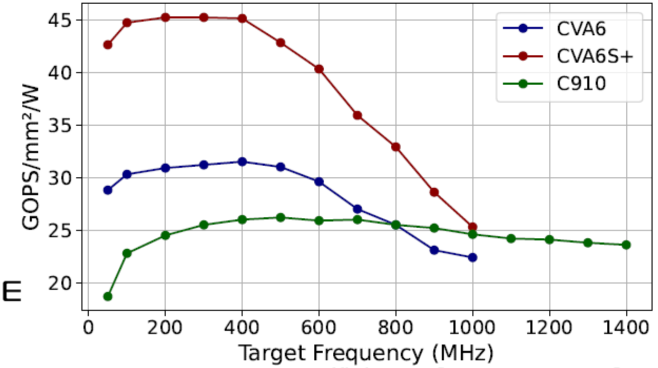
達摩院 DAMO ACADEMY | XUANTIE



(a) Area Efficiency [GOPS/mm²]



(b) Energy Efficiency [GOPS/W]



(c) Area-Energy Efficiency [GOPS/mm²/W]

ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Heterogeneous, Multiscale Accelerated Computing

Multiple Scales of acceleration

Extensions to processor cores

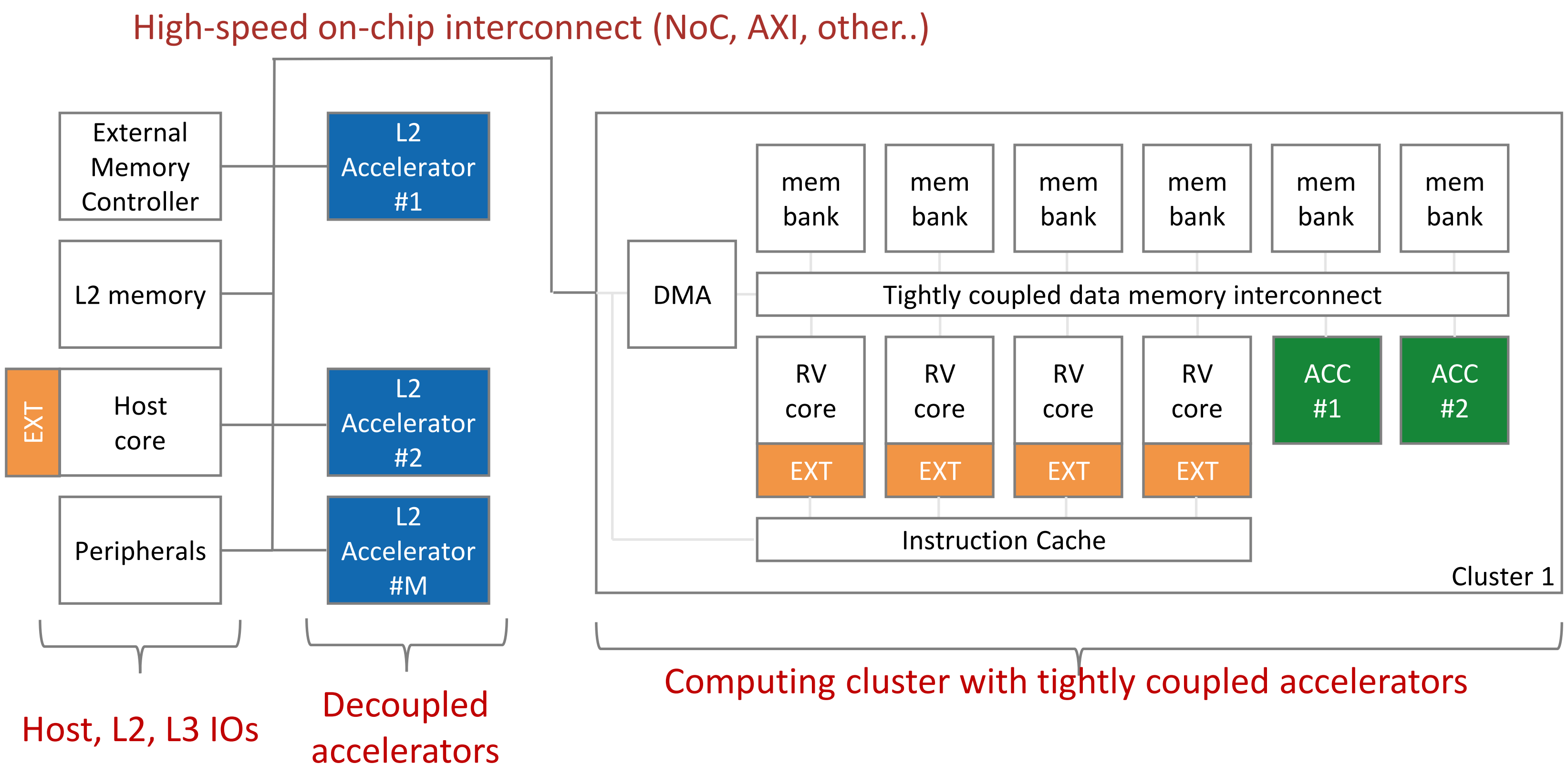
- Explore new extensions
- Efficient implementations

Shared-memory Accelerators

- Domain specific
- Local memory

Multiple Decoupled Accelerators

- Communication
- Synchronization



RISC-V is a key enabler → agility, enabling SW build-up, no vendor lock-in

Kraken: 22FDX SoC, Multiple Heterogeneous Accelerators

The *Kraken*: an “Extreme Edge” Brain

- **RISC-V Cluster**

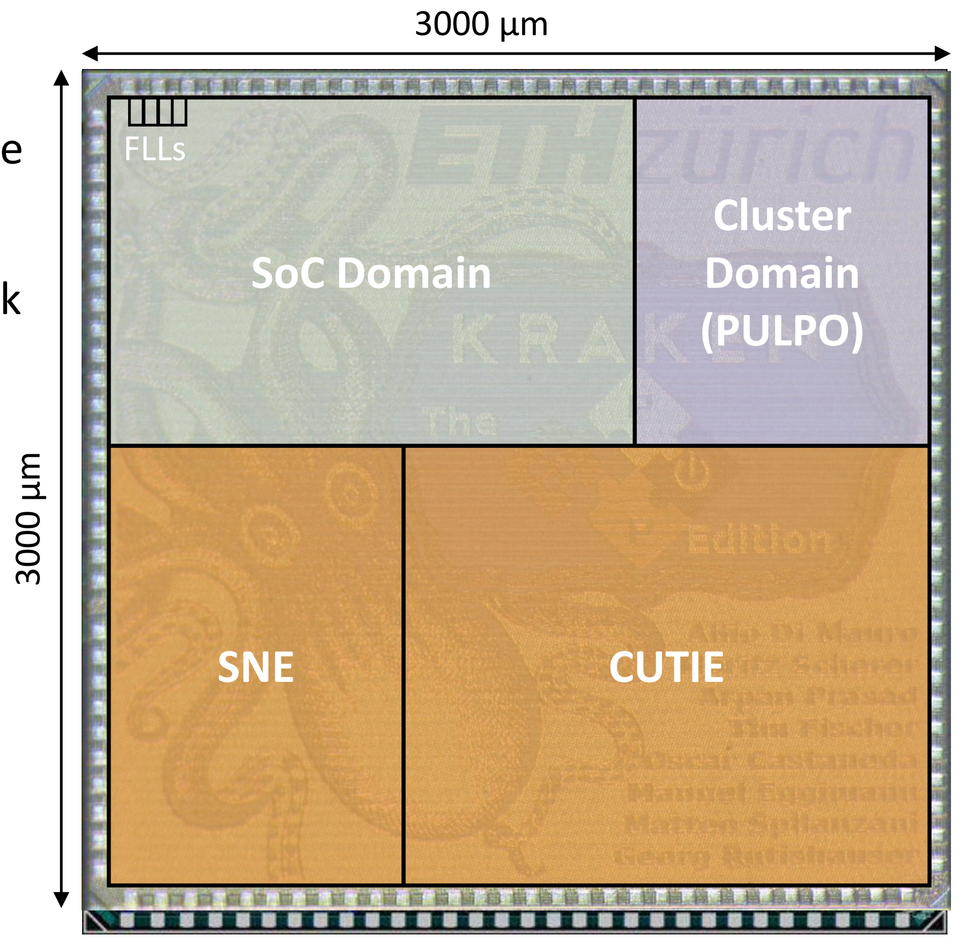
  8 Compute cores +1 DMA core

- **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**

Design is King

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

Energy-Efficient RV Core → **10pJ (8bit)**

⬇

PE specialization 10-20x →**1pJ (8bit)** ➡ **x10**

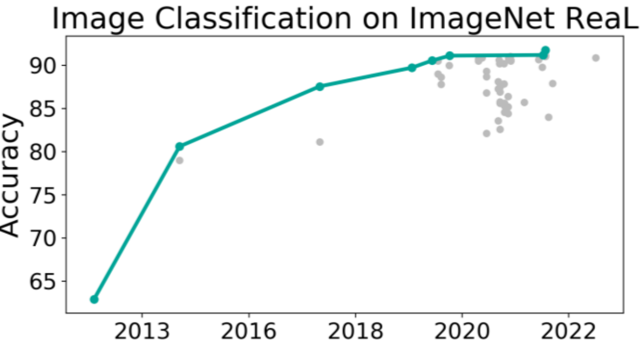
⬇

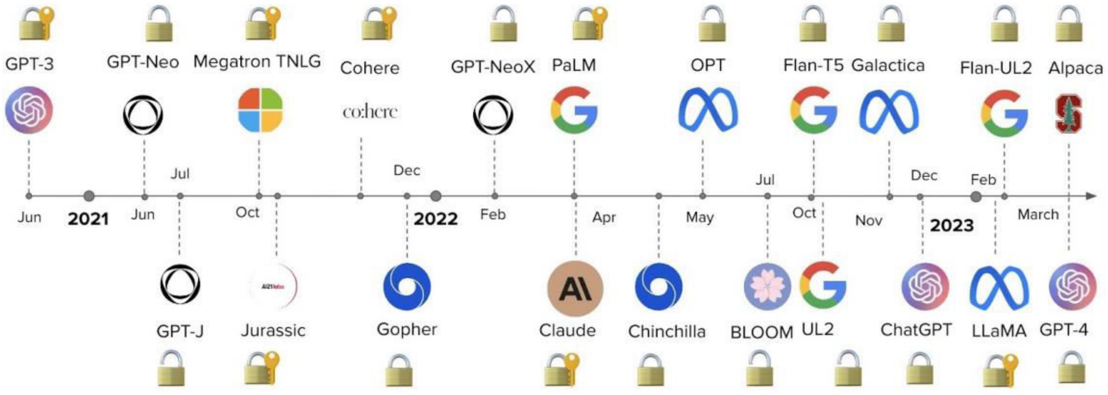
Configurable TE 10-20x → **100fJ (4bit)** ➡ **x10 (x100)**

⬇

Perception Accelerator 10-20x →**10fJ (events)** ➡ **x10 (x1000)**

Perceptive → Generative → Embodied AI



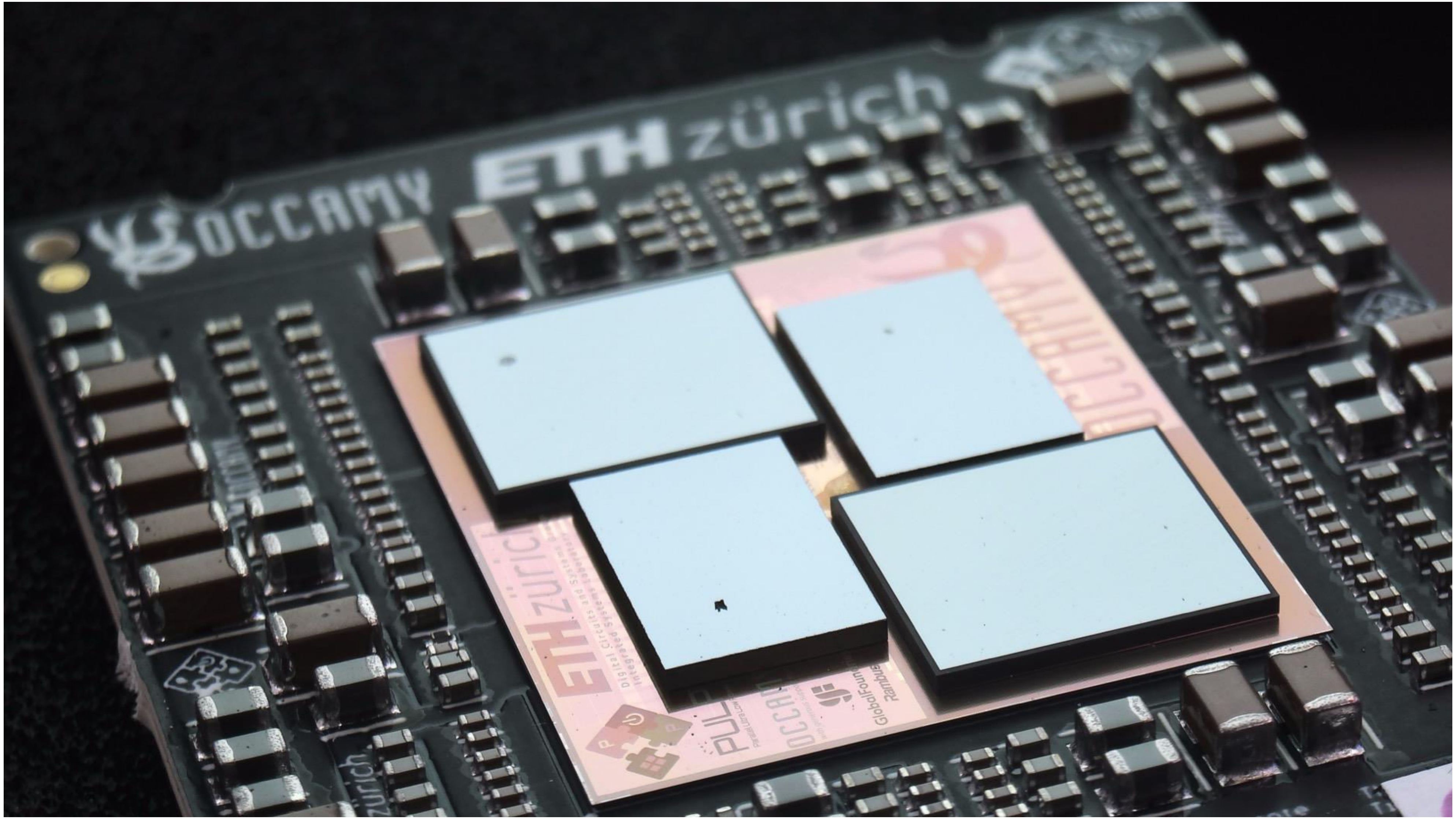


Precise

Interactive, creative

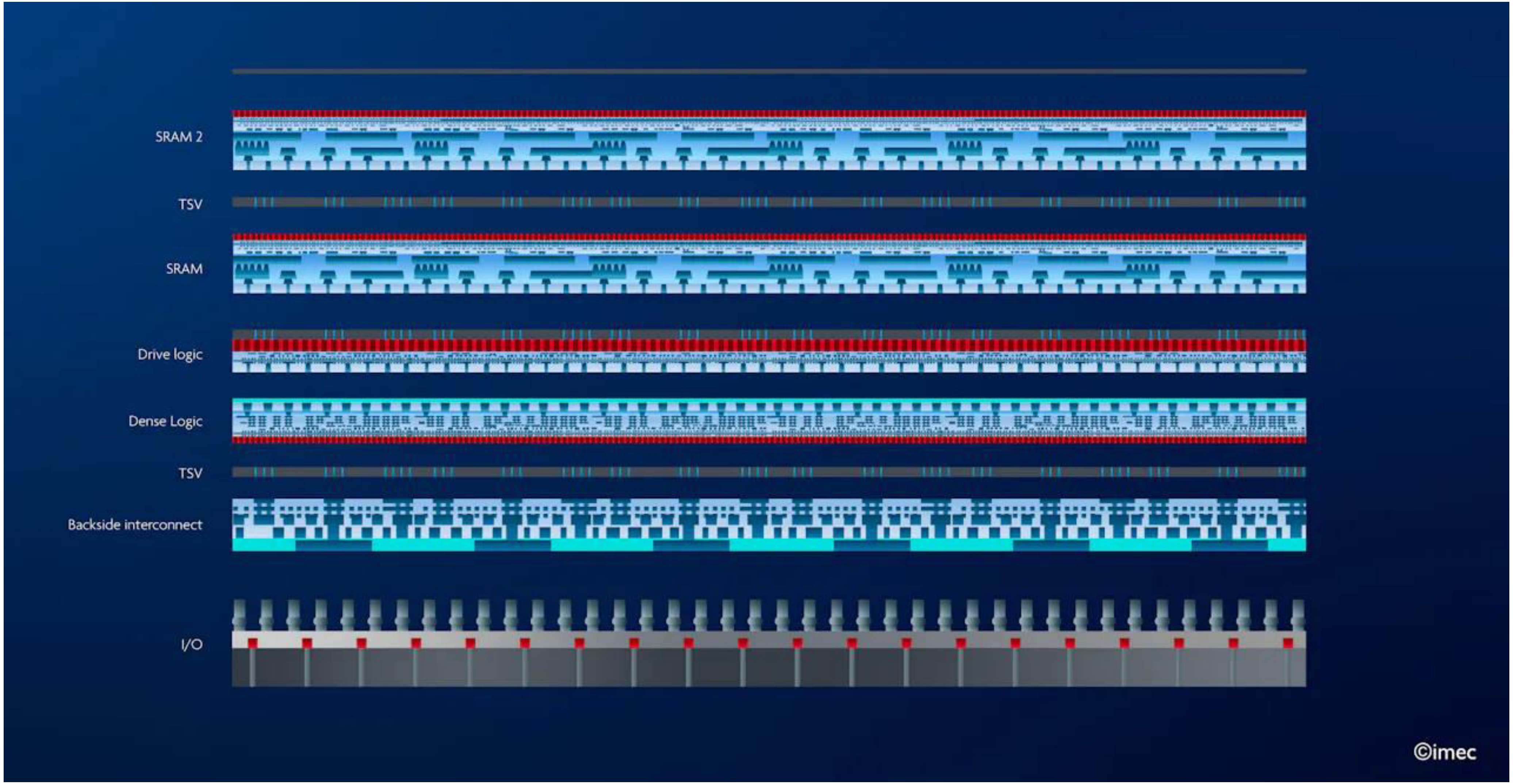
Efficient,
RT-safe,
secure

Chiplet Scaling: RISC-V gen.AI Platform for Agents



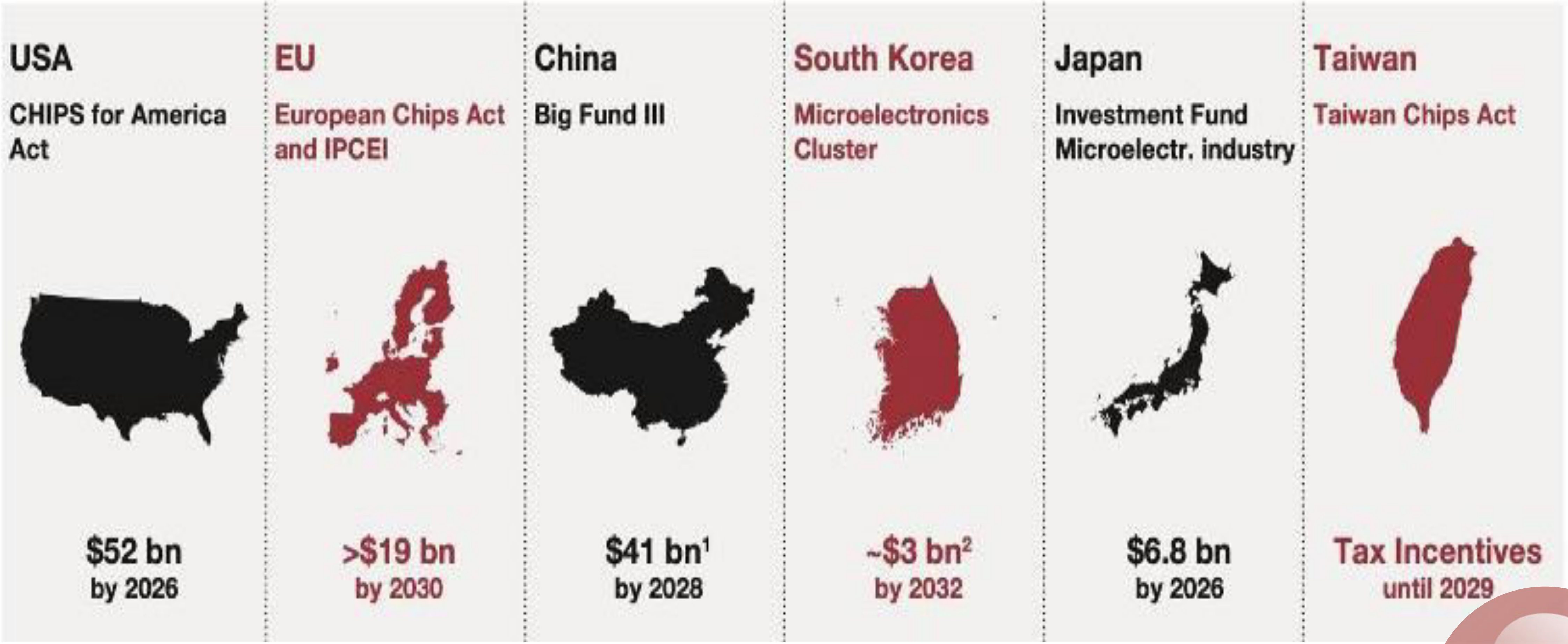
What’s next?

What’s next? CMOS 2.0!



Thank You!

The Race is On: Chips Acts all around us



USA	EU	China	South Korea	Japan	Taiwan
CHIPS for America Act	European Chips Act and IPCEI	Big Fund III	Microelectronics Cluster	Investment Fund Microelectr. industry	Taiwan Chips Act
$52 bn by 2026	>$19 bn by 2030	$41 bn[1] by 2028	~$3 bn[2] by 2032	$6.8 bn by 2026	Tax Incentives until 2029

Source: Strategy & analysis (LinkedIn)

Computing for AI is the key driver!