EPFL

EMBEDDED SYSTEMS LABORATORY

GenAI and the Transformation of Edge Computing: **Leveraging Heterogeneous Circuits for Innovation**

Jose Miranda

EPFL - Embedded Systems Laboratory

jose.mirandacalero@epfl.ch

03/27/2024

Outline

- **The Edge**
  - Current technological landscape
- **GenAI into the Edge**
  - Computation and Energy
  - Domain-specific accelerators and IMC
- **Deeply Heterogeneous SoCs**
  - Open-source frameworks
- **Overcoming Computational and Energy constraints**
- **Future directions**

The Edge: Current technological Landscape

Automation And Efficiency

Privacy-centric or Privacy by Design

EaaS

Environmental stepwardship

5G Adoption

Deployment Optimization Frameworks

Integration of GenAI into the Edge

GenAI into the Edge

$76 billion by 2028
(TIRIAS RESEARCH)

20% Offload to Edge

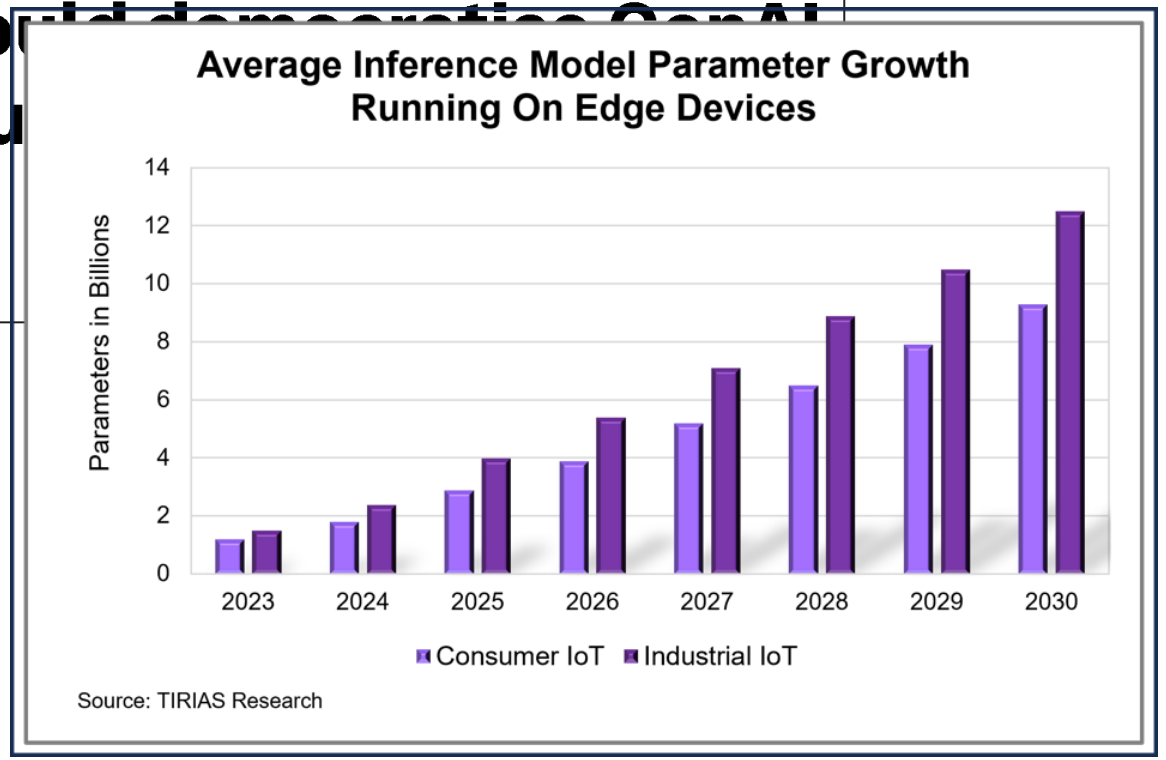
- $15 billion
- 800 MW

WORLD ECONOMIC FORUM'24

EMERGING TECHNOLOGIES

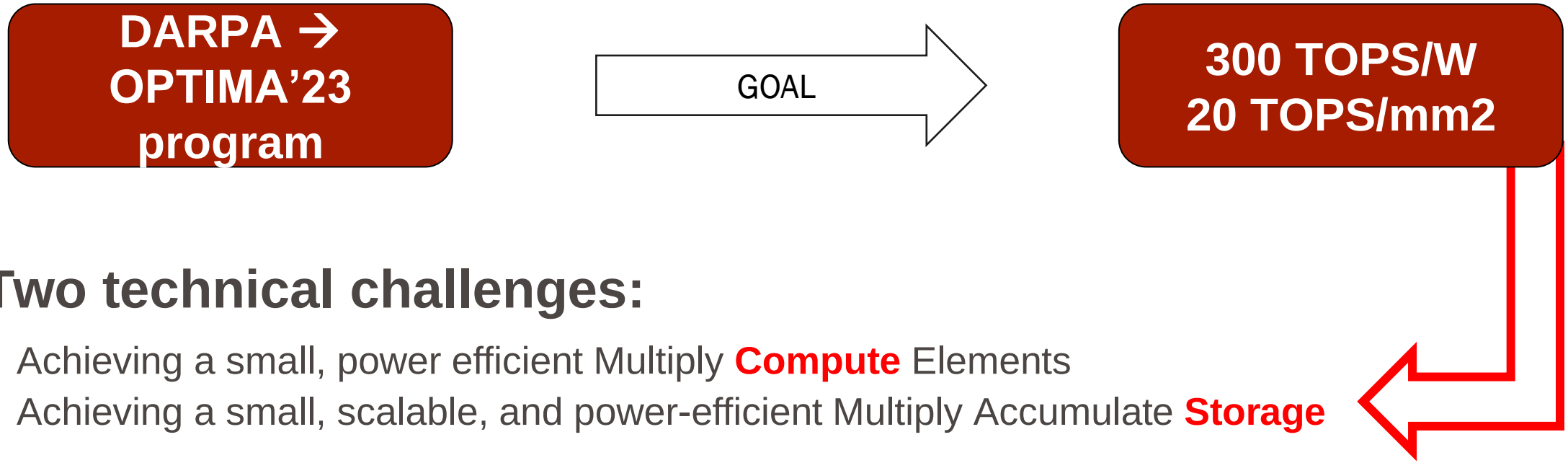
On-device AI wo
and ensure inclu
the economy

Jan 15, 2024



SoCs Performance improvements ≠ **Parameter growth**

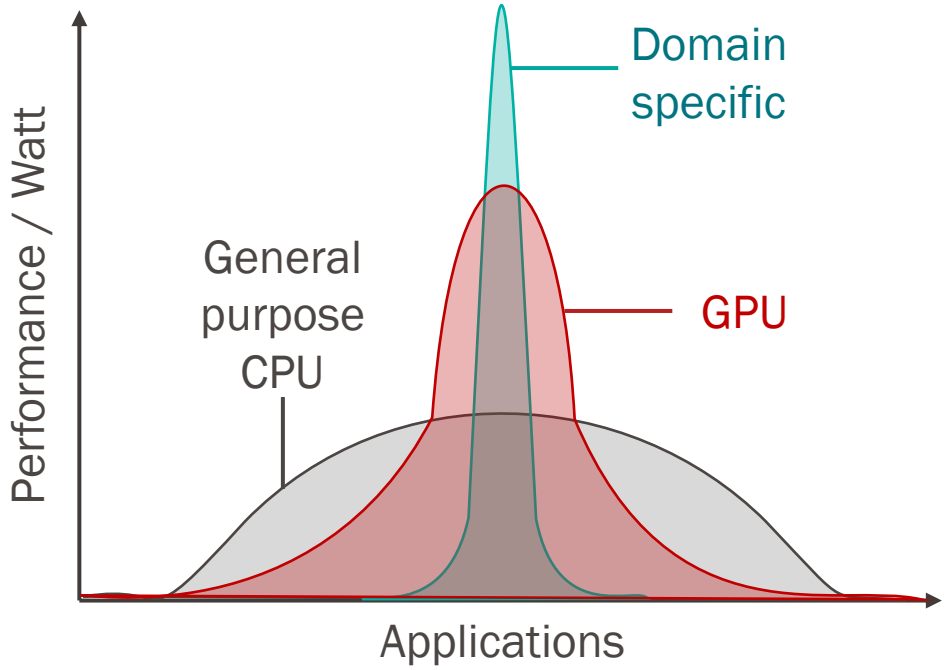
GenAI into the Edge: Computation and Energy



- **Two technical challenges:**
  - Achieving a small, power efficient Multiply **Compute** Elements
  - Achieving a small, scalable, and power-efficient Multiply Accumulate **Storage**

E(memory access) >> E(computations)

GenAI into the Edge: **Domain-Specific Accelerators** and IMC



Mark Papermaster: “Advancing EDA Through the Power of AI and High-performance Computing”, DAC59 Keynote, 2022

GenAI into the Edge: **Domain-Specific Accelerators** and IMC

NVIDIA

One of the best DSA providing energy efficient inference with transformers (BERT)
95.6 TOPS/W – 1711 inferences/s/W – 0.7% ACC loss

Table 2: Comparison to prior work.

	[3]	[7]	[8]	[9]	This work		
Process Technology	7nm	28nm	5nm	7nm	5nm		
Area (mm²)	19.6	1.9	5.46	3.04	0.153		
Supply Voltage (V)	0.55 – 0.75	0.6 – 0.9	0.55 – 0.9	0.58 – 0.83	0.46 – 1.05		
Frequency (MHz)	1000 – 1600	100 – 470	332 – 1196	290 – 880	152 – 1760		
On-Chip SRAM (KB)	8192	206	3072	2176	141		
Data Formats	INT2/4, FP8/16/32	INT8	INT8, INT16	INT8/16, FP16	INT4	INT4 VSQ	INT8
Performance (TOPS)	102.4 (4b, 0.75V)	1.43 (8b, 0.9V)	14.7 (8b, 0.9V)	3.6 (8b, 0.83V)	3.6 (1.05V)	3.6 (1.05V)	1.8 (1.05V)
Energy Efficiency (TOPS/W)	16.5* (4b, 0.55V)	17.5* (8b, 0.6V)	13.6* (8b, 0.6V)	6.8* (8b, 0.58V)	**91.1**† (0.46V)	**95.6**† (0.46V)	**39.1**† (0.46V)
Area Efficiency (TOPS/mm²)	5.22 (4b, 0.75V)	0.75 (8b, 0.9V)	2.69 (8b, 0.9V)	1.2 (8b, 0.83V)	**23.3** (1.05V)	**23.3** (1.05V)	**11.7** (1.05V)

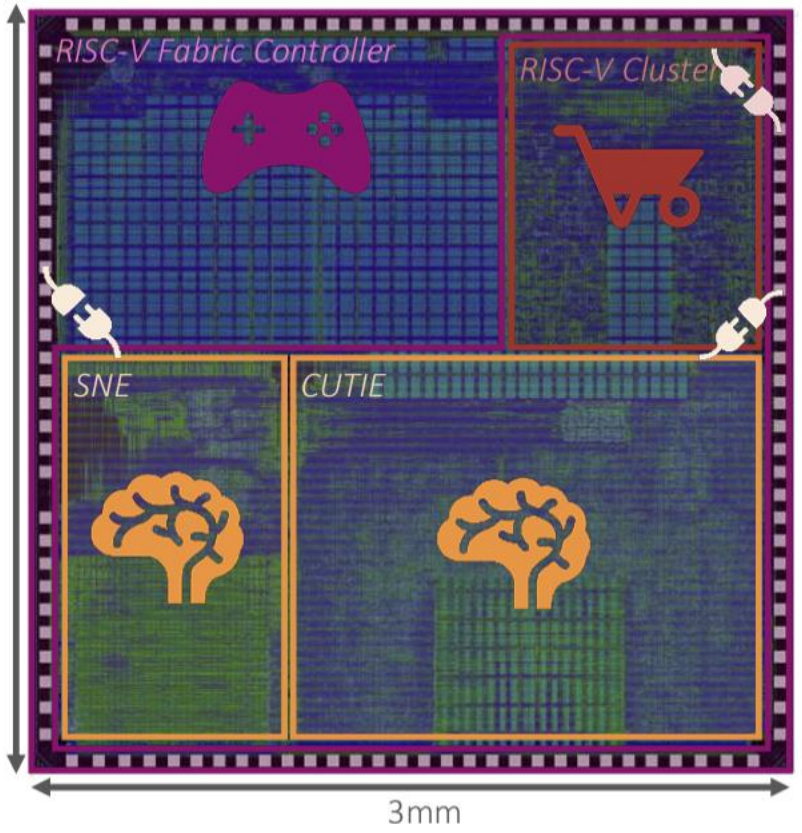
* Input densities not reported. † Measured with 50% non-zero input densities. Includes estimated leakage power.

B. Keller *et al*., "A 17–95.6 TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization for Transformers in 5nm," *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, HI, USA, 2022, pp. 16-17.

GenAI into the Edge: **Domain-Specific Accelerators** and IMC

Heteregeneous integrations of domain-specific accelerators: KRAKEN by PULP

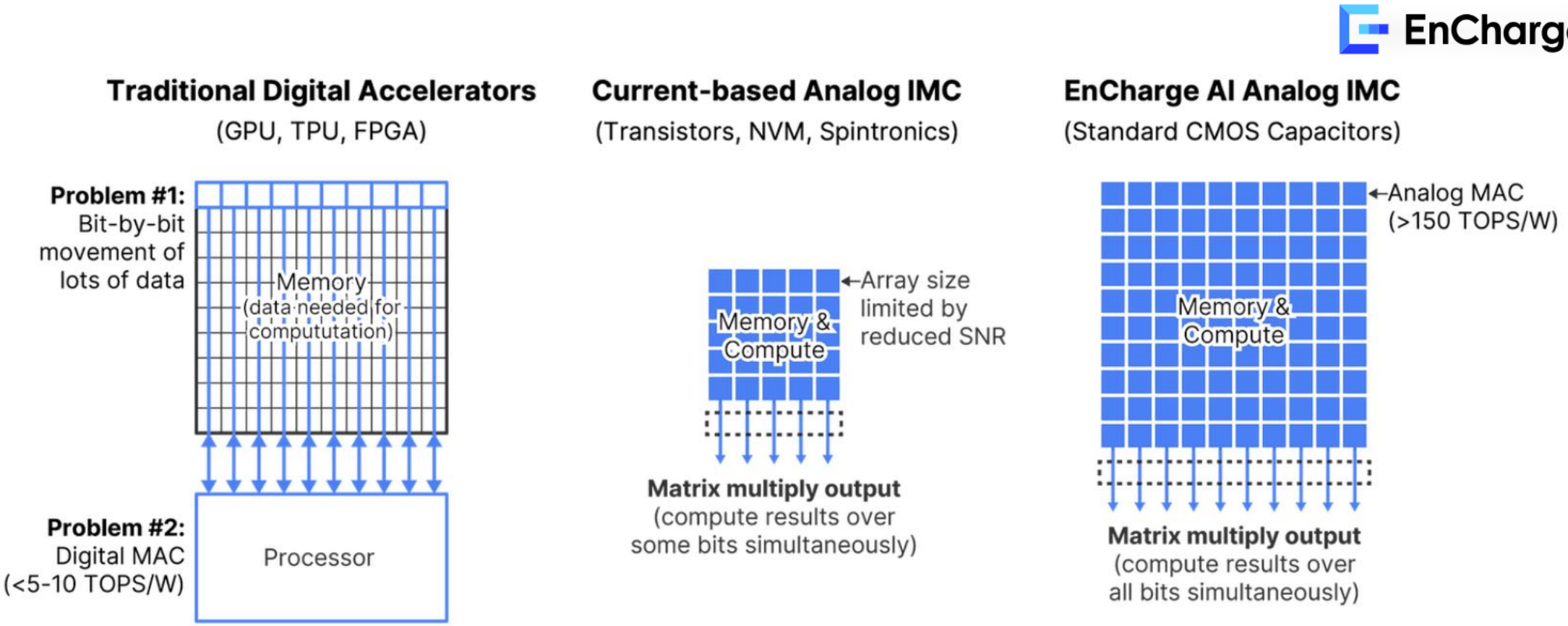
- RISC-V Cluster
- SNE – Spiking NN accelerator
- CUTIE – Ternary Neural Network
  - > 1 PetaOps/s/W



M. Scherer et al., "A 1036 TOp/s/W, 12.2 mW, 2.72 µJ/Inference All Digital TNN Accelerator in 22 nm FDX Technology for TinyML Applications," 2022 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS), 2022

GenAI into the Edge: Domain-Specific Accelerators and **IMC**

EnCharge AI



https://www.enchargeai.com/technology

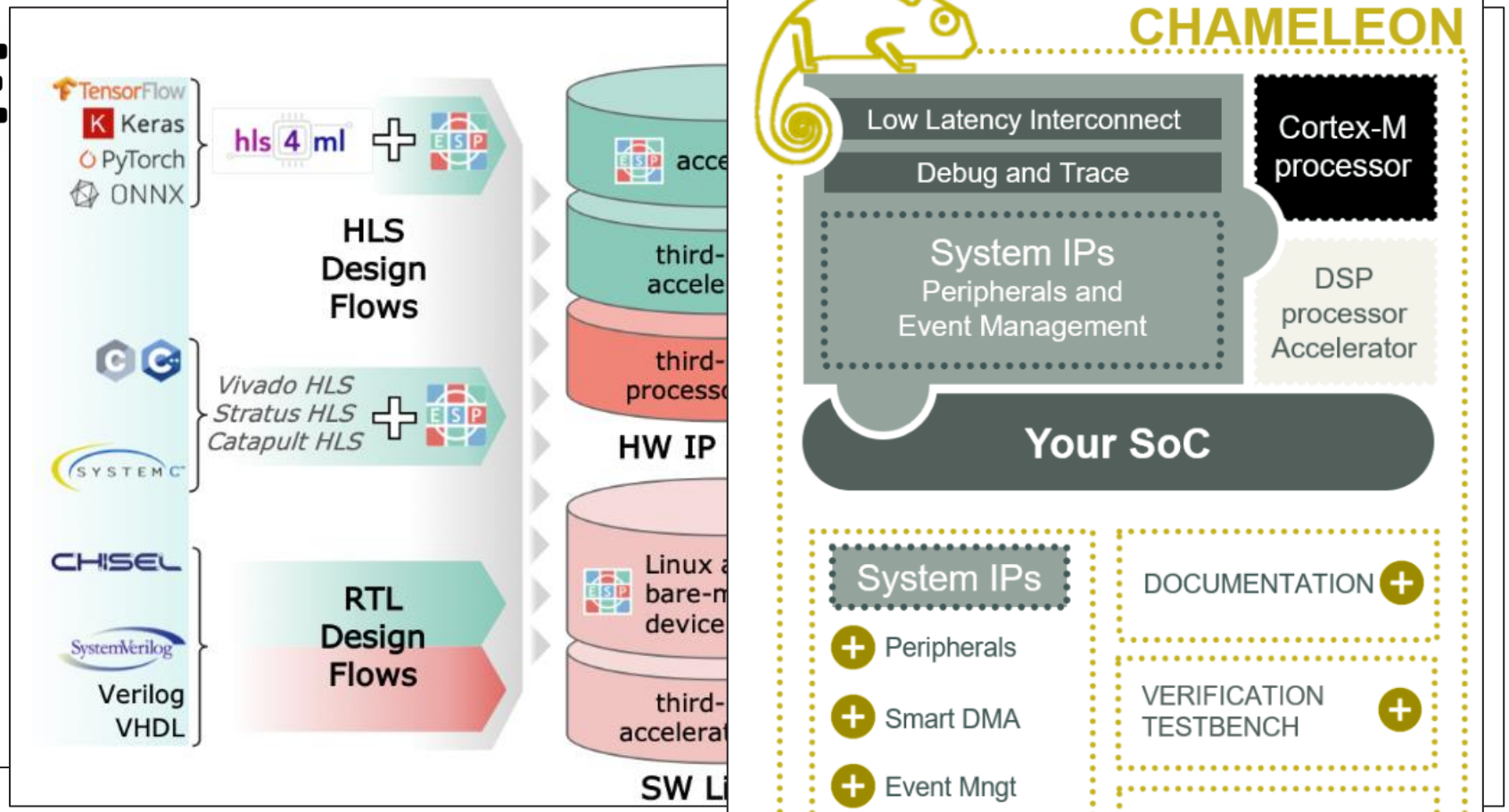
Deeply Heterogeneous SoCs: Open source Frameworks

X-HEEP

Configurability

1. RISC-V core
2. Coprocessor interface
3. Peripherals
4. Interrupt controller
5. Accelerator interface
6. Power manager
7. Bus topology
8. Number of banks

https://x-heep.epfl.ch/



https://www.esp.cs.columbia.edu/

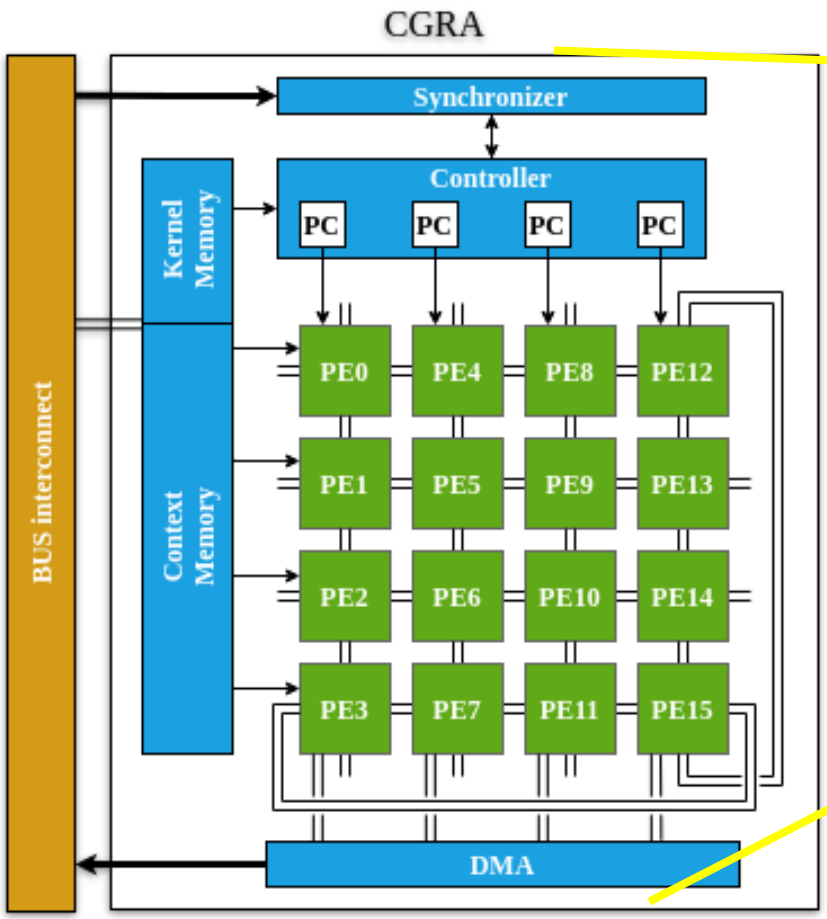
DOLPHIN DESIGN

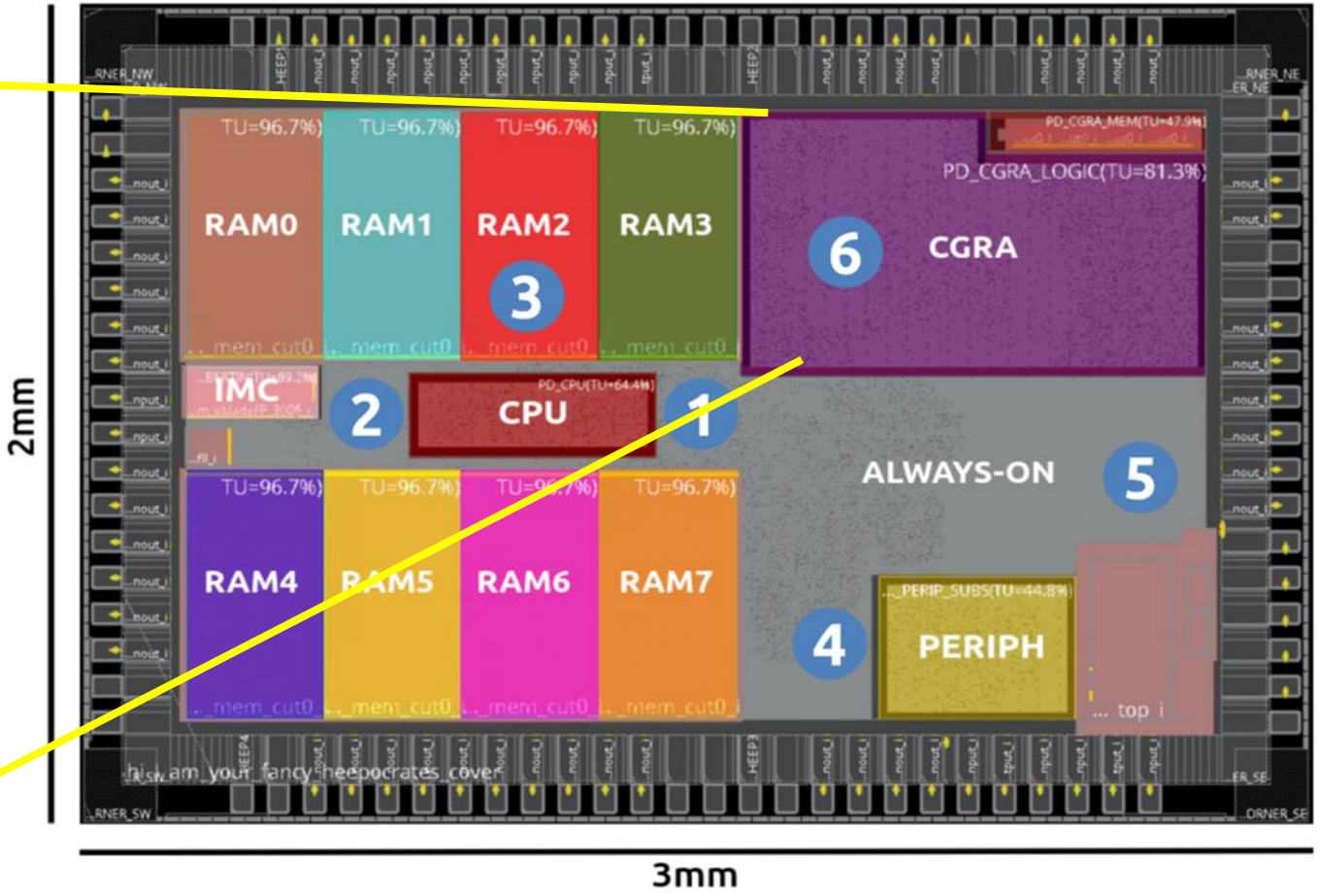
CHAMELEON

https://www.dolphin-design.fr/chameleon-mcu-subsystem/

Overcoming computational and energy constraints

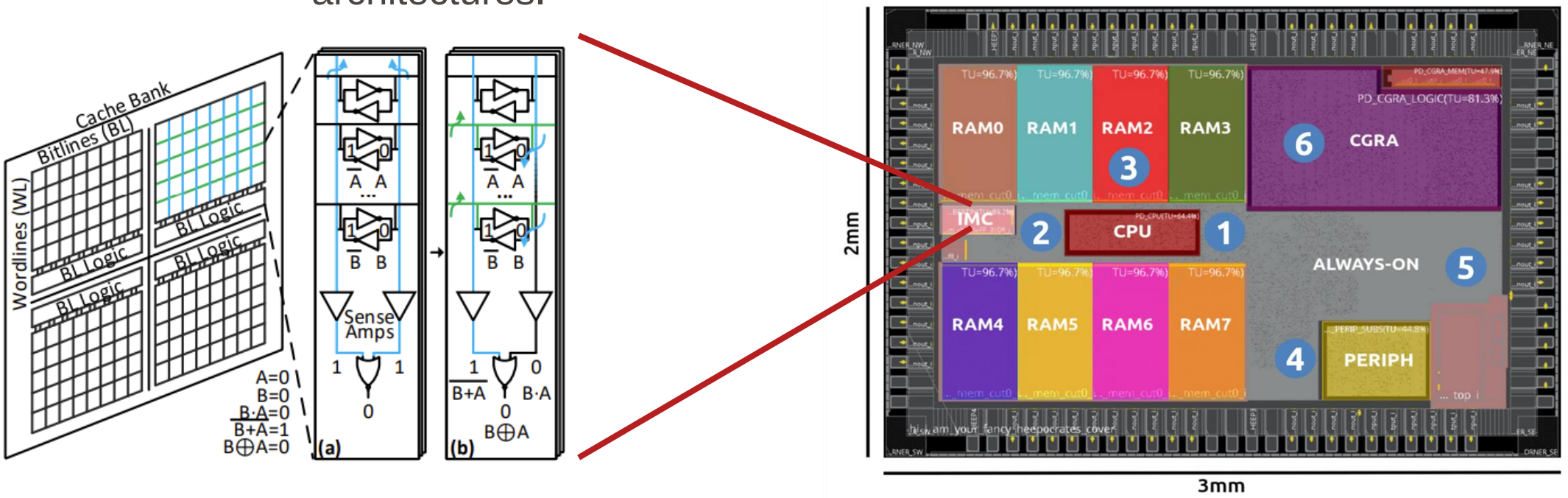
HEEPocrates tape-out





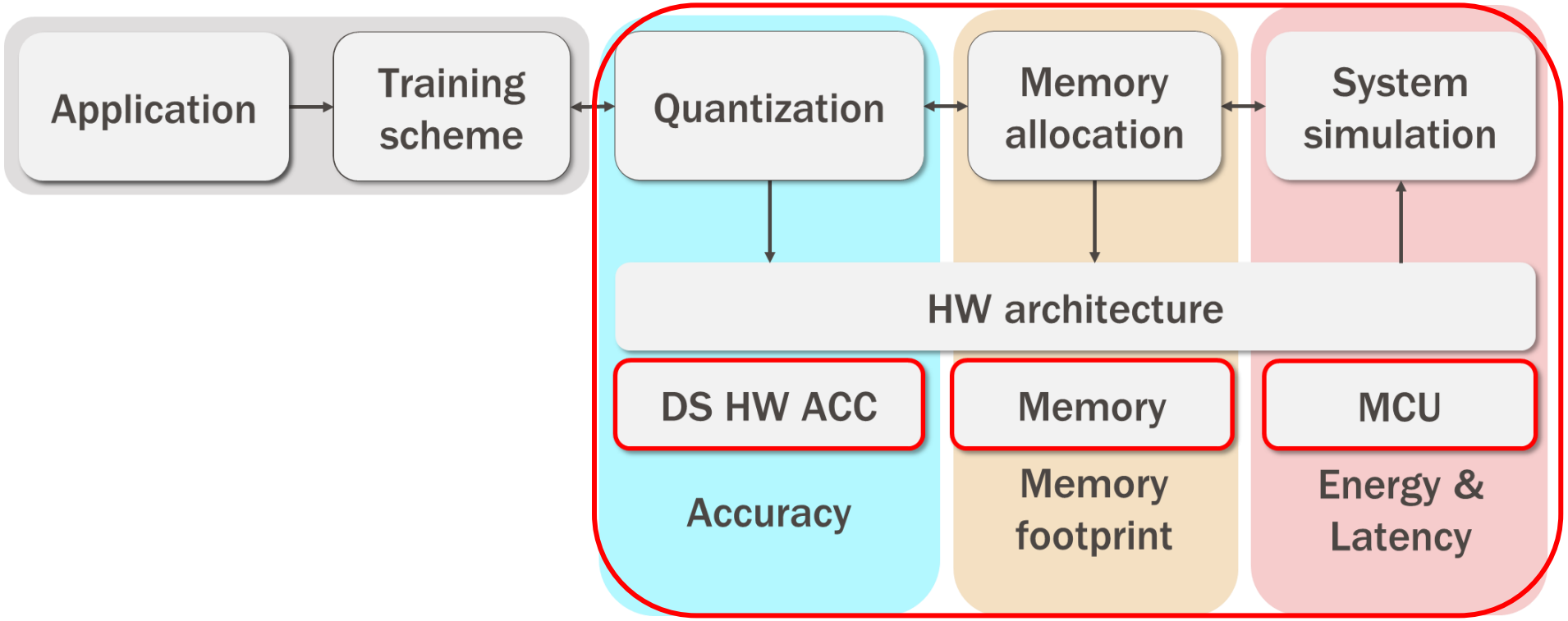
Overcoming computational and energy constraints

- **in-SRAM computing**
  - Exploiting bit-line computing → SIMD enabler
  - Blade bit-line computing
    - **New IMC architecture developed at ESL**: BLADE is an in-SRAM computing architecture that utilizes local word-line groups to perform computations at a frequency 2.8x higher than state-of-the-art in-SRAM computing architectures.



Future directions:

- **Need for next-generation end-to-end and HW-aware deployment frameworks for deeply heterogeneous processor and SoC architectures**



Thank you!

Jose Miranda

EPFL - Embedded Systems Laboratory
jose.mirandacalero@epfl.ch