EPFL

EMBEDDED SYSTEMS LABORATORY

Neuro-Inspired Edge AI Architectures for Distributed Federated Learning

Dr. Denisa Constantinescu, Prof. David Atienza

Embedded Systems Laboratory (ESL) EPFL, Switzerland

denisa.constantinescu@epfl.ch

EMERALD: 1st Workshop on Distributed Computing with Emerging Hardware Technology, 21st June, Nantes 2024

Embedded Sytems Laboratory

Prof. David Atienza

Team

1professor

1senior scientist

3 engineers, admin

7 post-docs

30 PhD students

Outline

Motivation

- **Privacy Preserving Personalized Healthcare Monitoring**

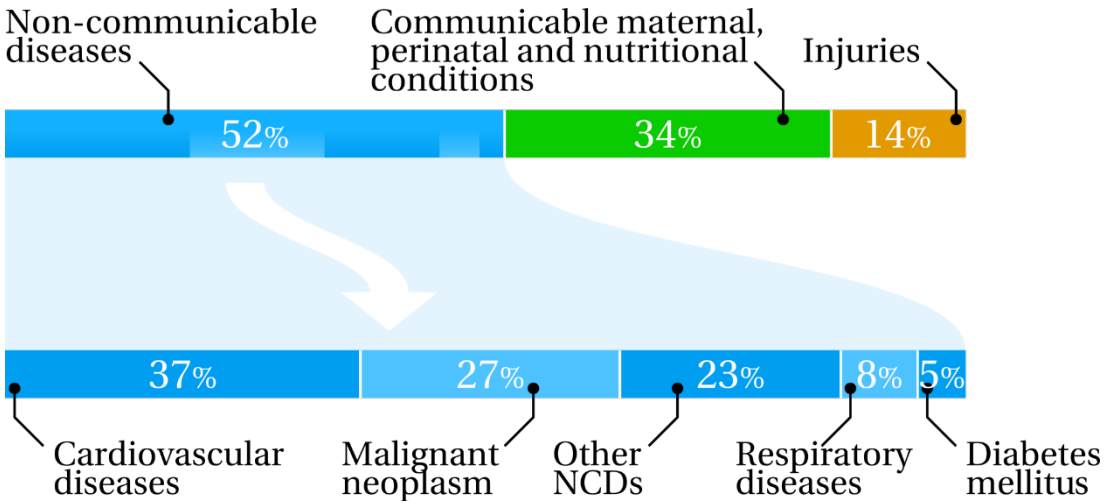
Enabling Technologies

- **SW:** Distributed Federated Learning
- **HW:** EdgeAI Architectures - Neuro-Inspired Accelerators for EdgeAI
- **System Level Co-Design:** Biosignal-Taylored EdgeAI

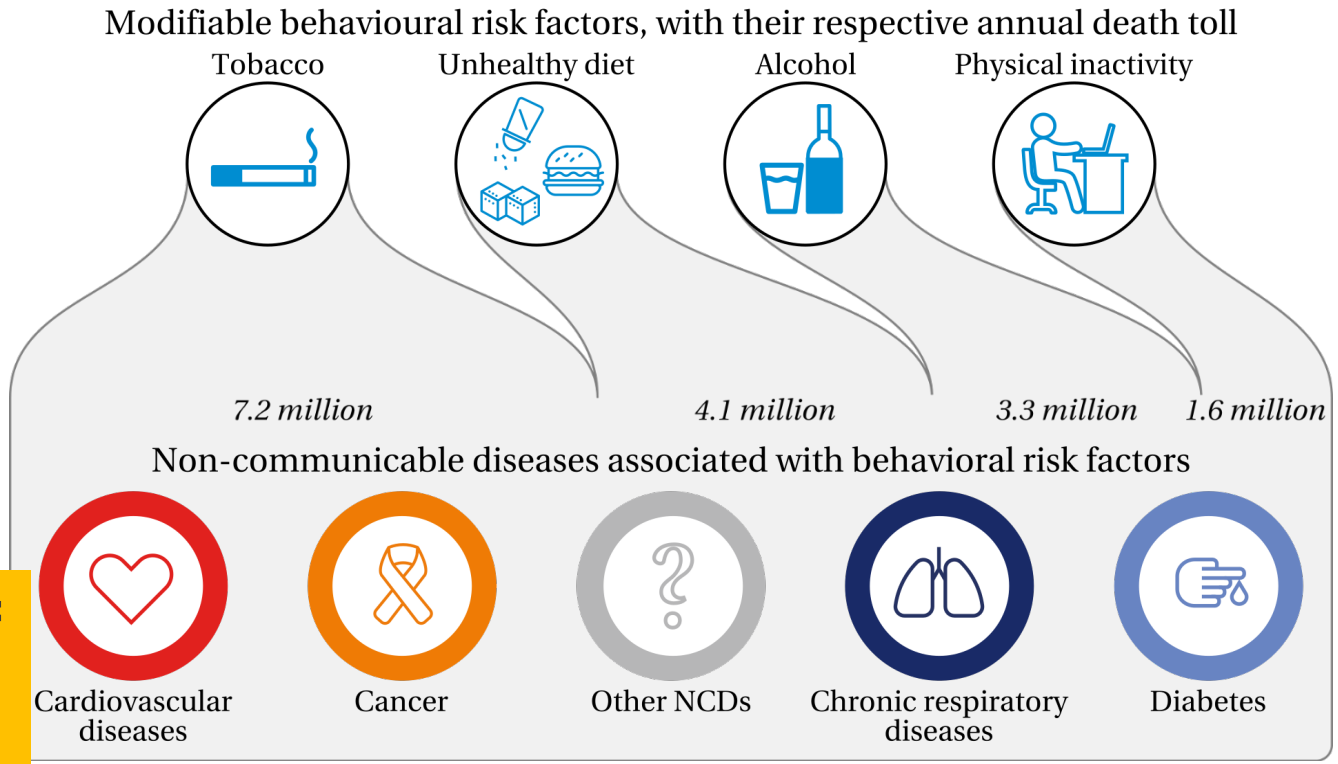
Discussion

Current Situation of **Healthcare**

Causes of death worldwide [3]



Behavioural risk factors [4]



Sustainable long-term monitoring of lifestyle habits with IoT (edge AI) wearables is key!

Each person is different!

- [3] World Health Organization. Noncommunicable diseases, 2018
- [4] World Health Organization. Noncommunicable diseases and mental health: challenges and solutions, 2014

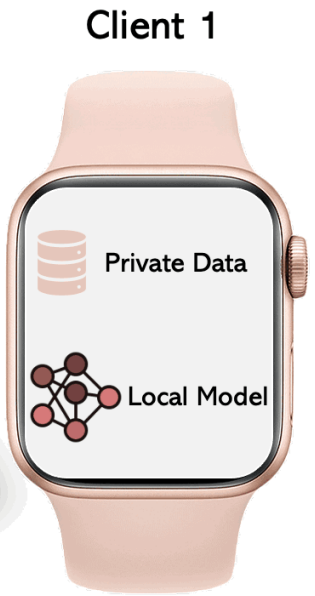
Federated Learning (FL) for Healthcare Monitoring

Originated from AI/ML community, and has limited connection with computing systems:

Von-Neumann architectures not good for this!
(1) Simple operations
(2) Lots of memory needed

Centralized approach

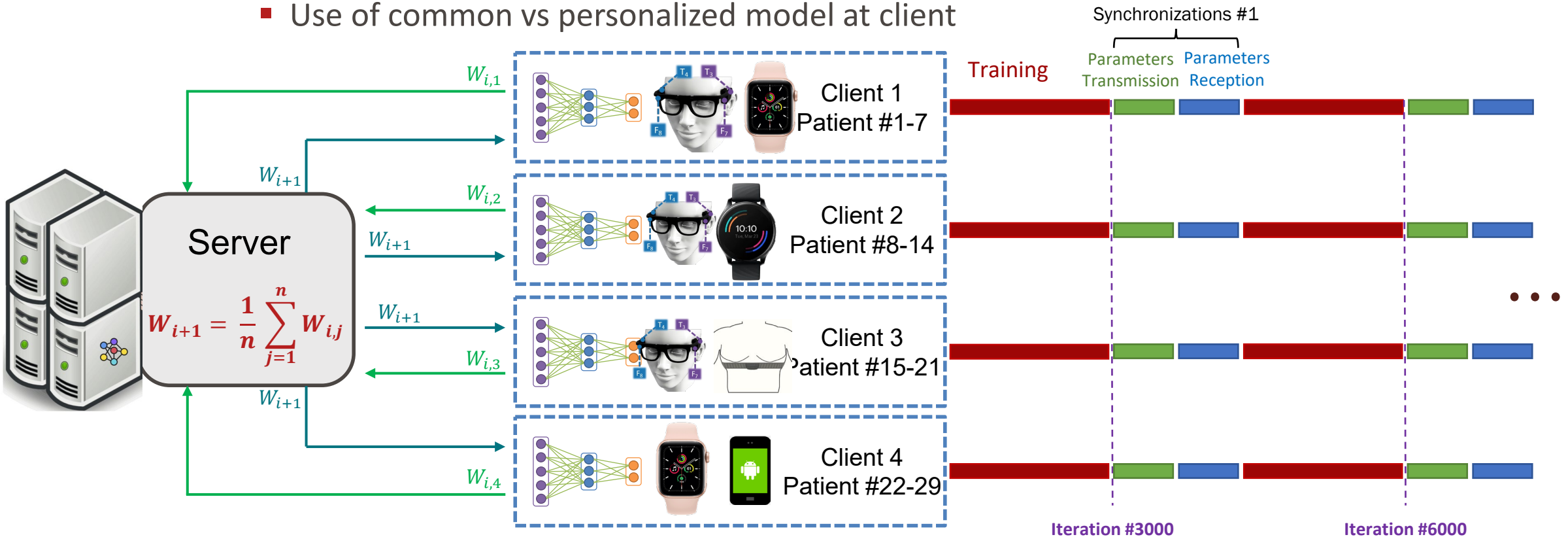
- ✗ Data Privacy Concerns
- ✗ High Energy Consumption
- ✗ Scalability Issues
- ✗ Reduced Personalization



Distributed Federated Learning on **Edge & IoT Devices**

Weights of ML/CNN models of each edge AI instance (client) shared after a certain number of local training iterations

- Explore trade-offs in transmission power vs. central coordination
- Use of common vs personalized model at client

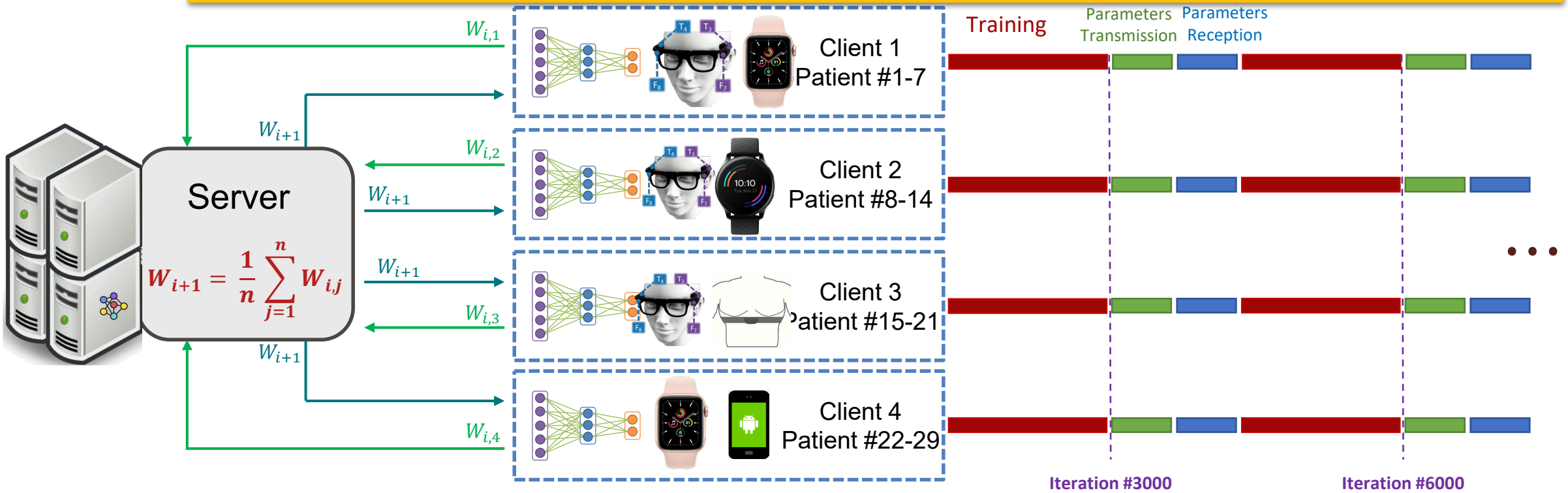


S. Baghersalimi, et al., "Real-Time Federated Learning for Epileptic Seizure Detection". IEEE Journal of Biomedical and Health Informatics (J-BHI), June 2021

Distributed Federated Learning on **Edge & IoT Devices**

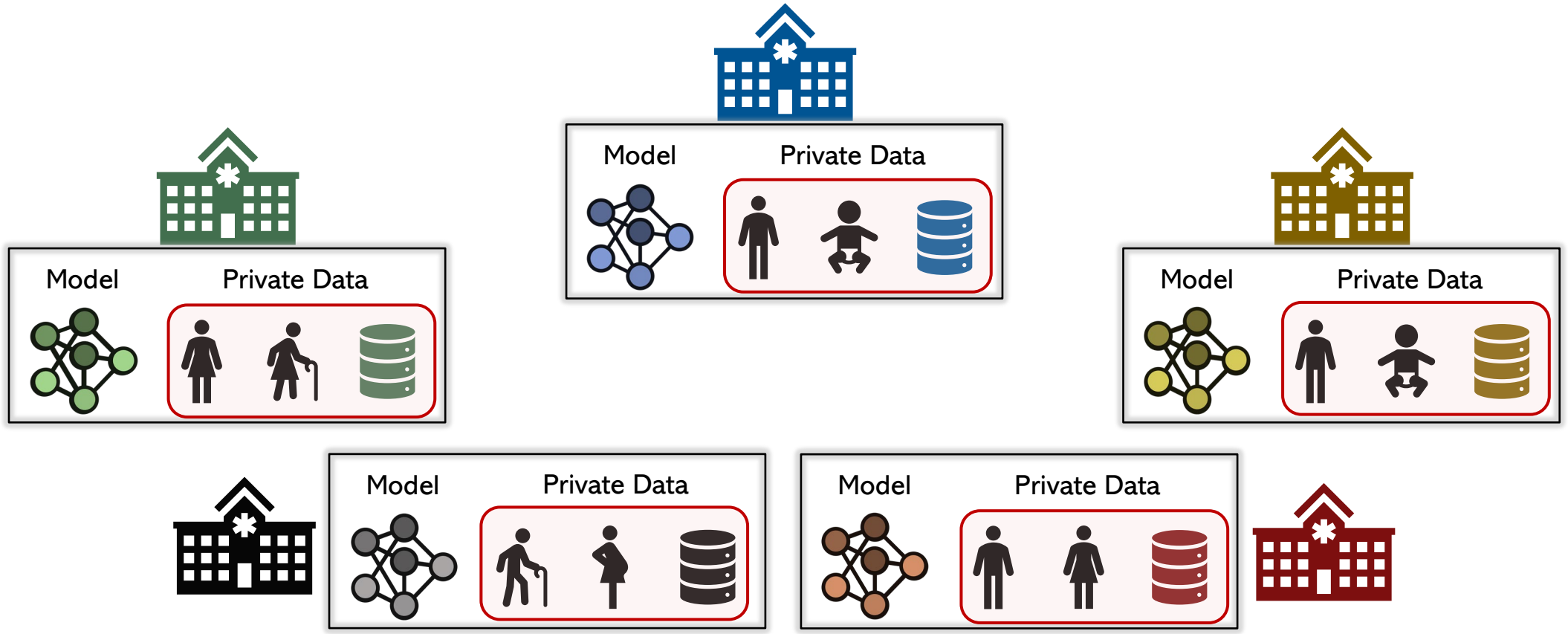
Weights of ML/CNN models of each edge AI instance (client) shared after a certain number of local training iterations

Use of FL for epilepsy monitoring can enable 95% seizures detection with personalized networks of new edge AI architectures!



S. Baghersalimi, et al., "Real-Time Federated Learning for Epileptic Seizure Detection". IEEE Journal of Biomedical and Health Informatics (J-BHI), June 2021

Privacy-Preserving Distributed FL for Healthcare Monitoring **in Hospitals**

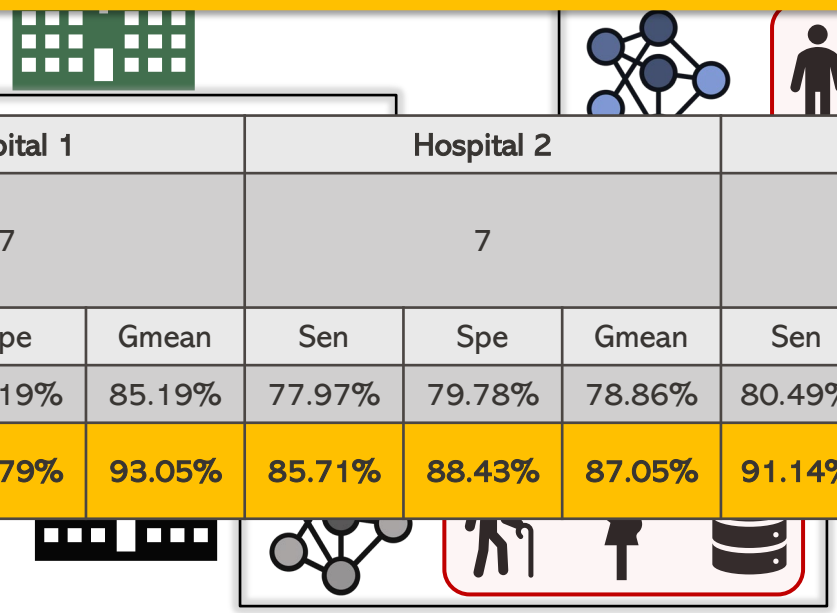


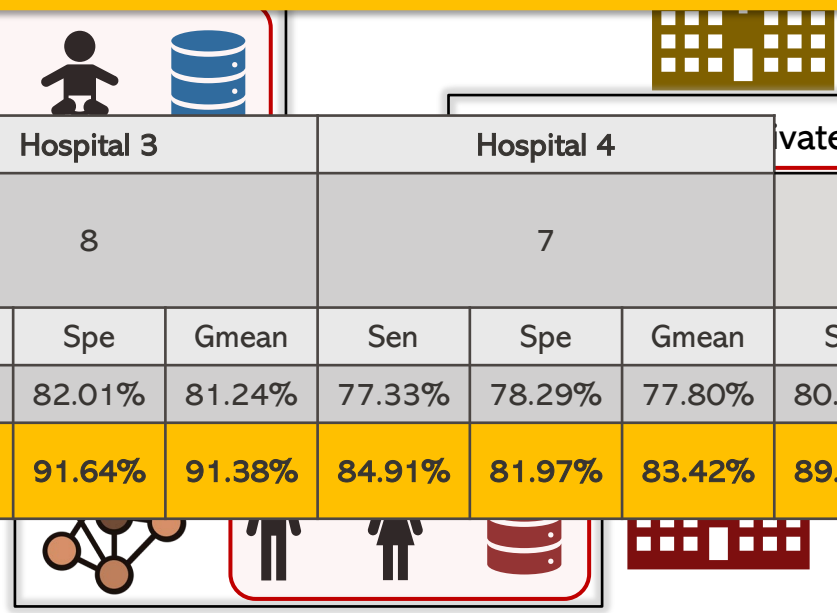
Baghersalimi, Saleh, et al. "Decentralized federated learning for epileptic seizures detection in low-power wearable systems." IEEE Transactions on Mobile Computing (2023).

Privacy-Preserving Distributed FL for Healthcare Monitoring **in Hospitals**

Serverless model w/ adaptive ensembling and knowledge distillation during training

- Tailored DNNs for each medical center, merging local and external models efficiently
- Suitable for large networks of hospitals with non-identically distributed patient data



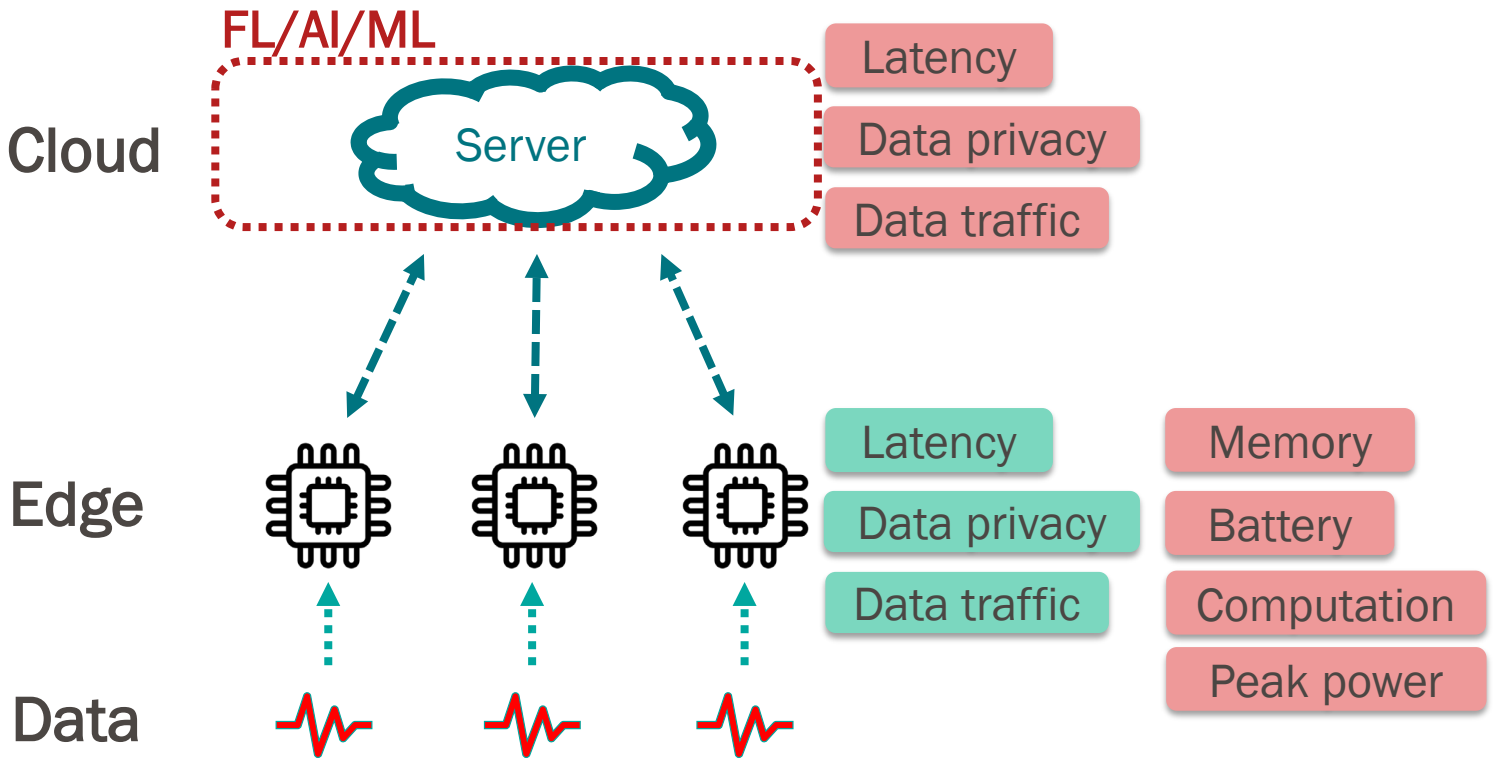


EPILEPSIAE		Hospital 1			Hospital 2			Hospital 3			Hospital 4			Average		
	Number of patients	7			7			8			7					
		Sen	Spe	Gmean	Sen	Spe	Gmean	Sen	Spe	Gmean	Sen	Spe	Gmean	Sen	Spe	Gmean
	FedAvg	84.22%	86.19%	85.19%	77.97%	79.78%	78.86%	80.49%	82.01%	81.24%	77.33%	78.29%	77.80%	80.00%	81.56%	80.77%
	Ensemble Learning	**94.34%**	**91.79%**	**93.05%**	**85.71%**	**88.43%**	**87.05%**	**91.14%**	**91.64%**	**91.38%**	**84.91%**	**81.97%**	**83.42%**	**89.02%**	**88.45%**	**88.72%**

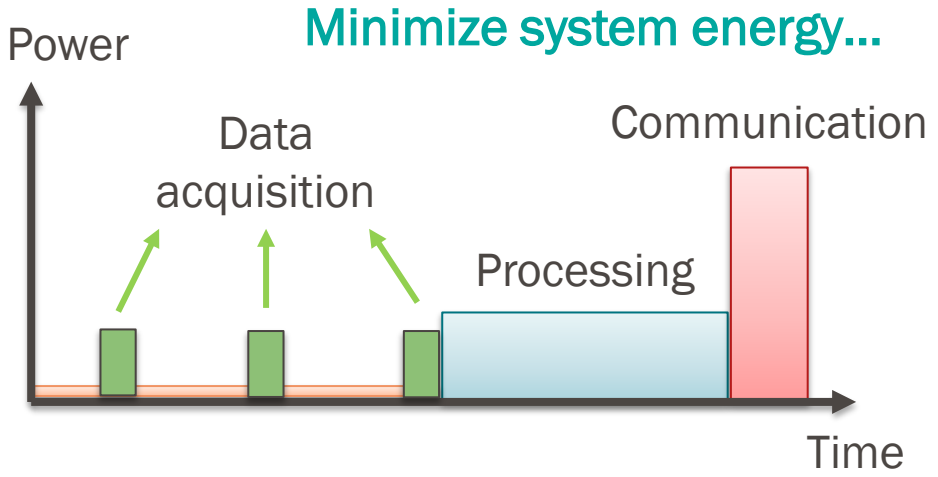
Baghersalimi, Saleh, et al. "Decentralized federated learning for epileptic seizures detection in low-power wearable systems." IEEE Transactions on Mobile Computing (2023).

Key Properties for **Edge AI Architectures** for FL

State-of-the-art



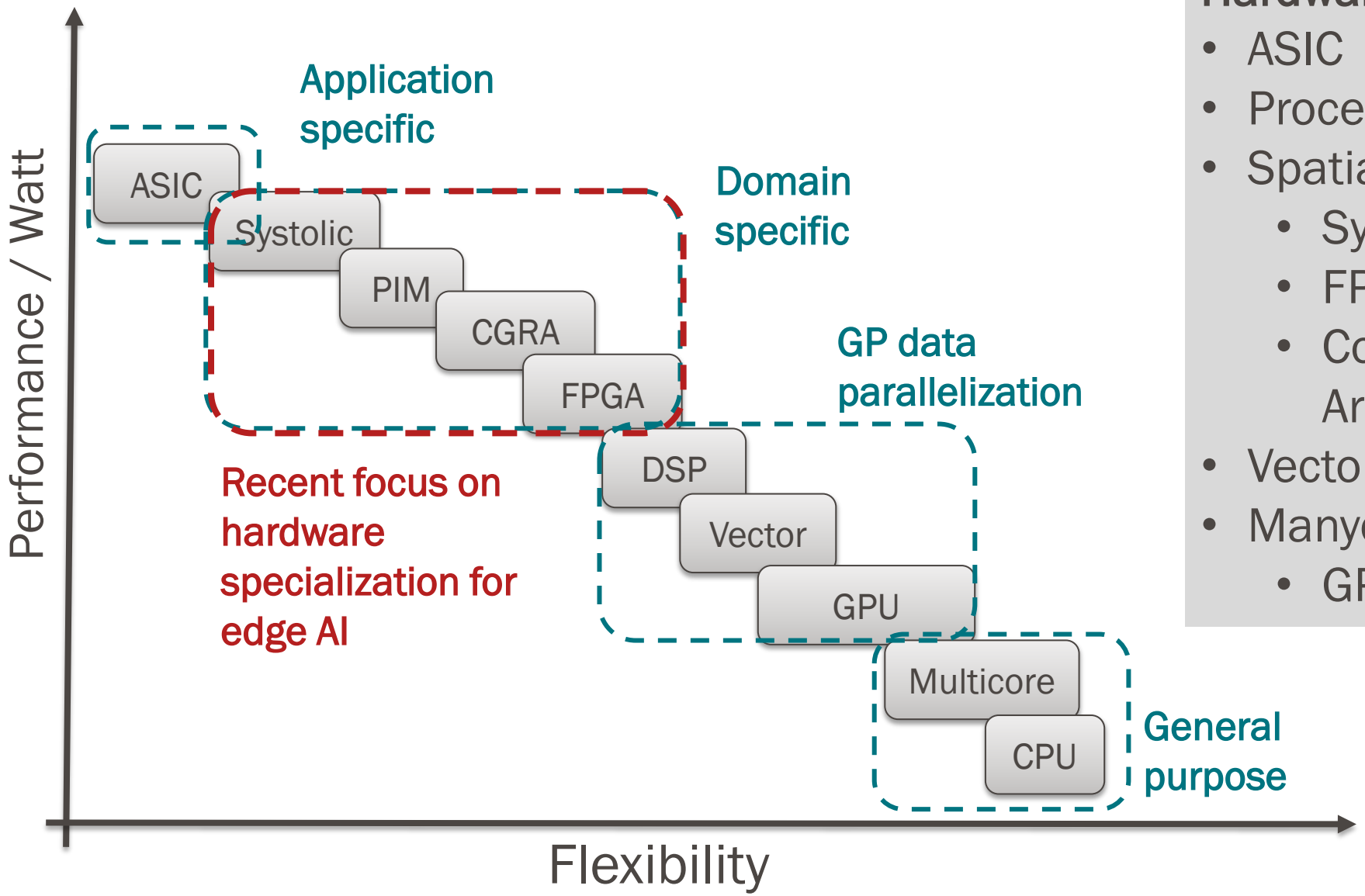
IoT applications include clear phases for edge AI systems design



Co-design needed:
one edge AI architecture solution does not fit all!

But high potential for even larger models: Generative AI @ edge?

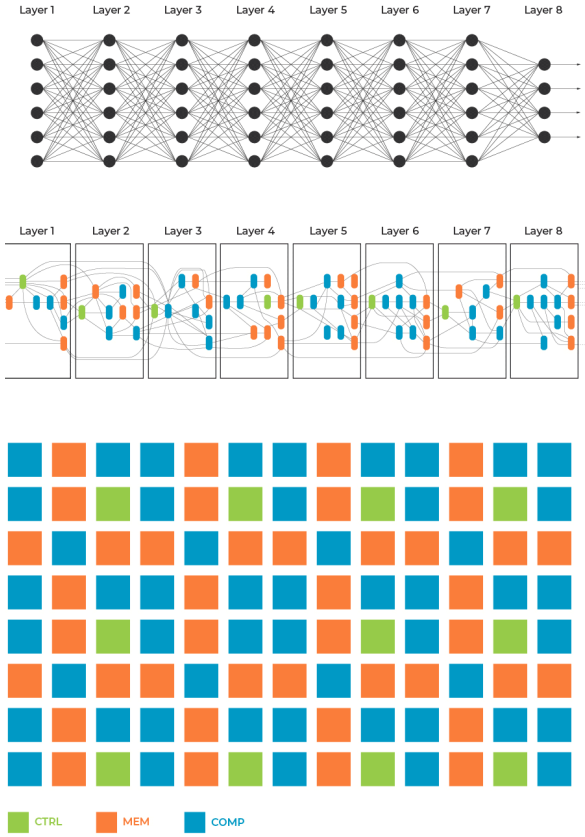
Many **Edge AI Architectures**

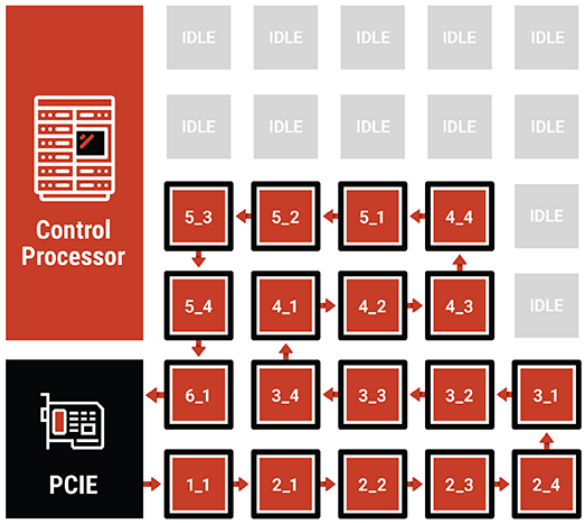


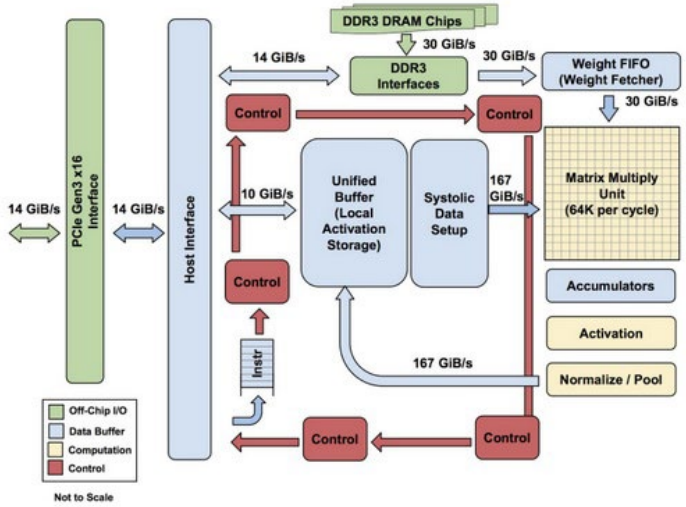
Hardware accelerators

- ASIC
- Process-in-memory (PIM)
- Spatial accelerator
  - Systolic array
  - FPGA
  - Coarse Grained Reconfig. Arrays (CGRA)
- Vector machine
- Manycore
  - GPU

Application Specific **Edge AI Accelerators**







Hailo:
26 TOPS at 3 TOPS/W

Mythic:
35 TOPS at 4 TOPS/W

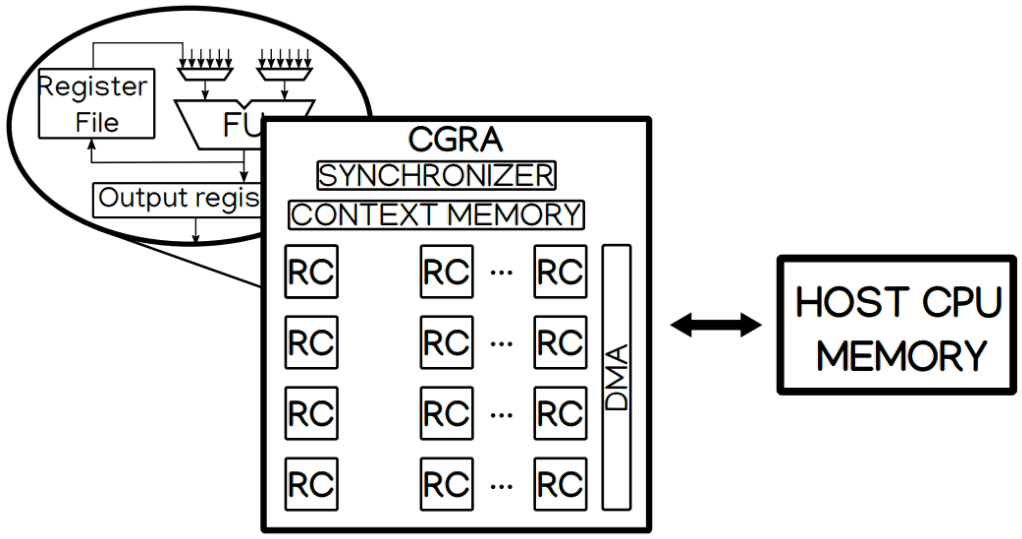
Edge TPU:
4 TOPS at 2 TOPS/W

Image Credit: Hailo, Mythic, Google

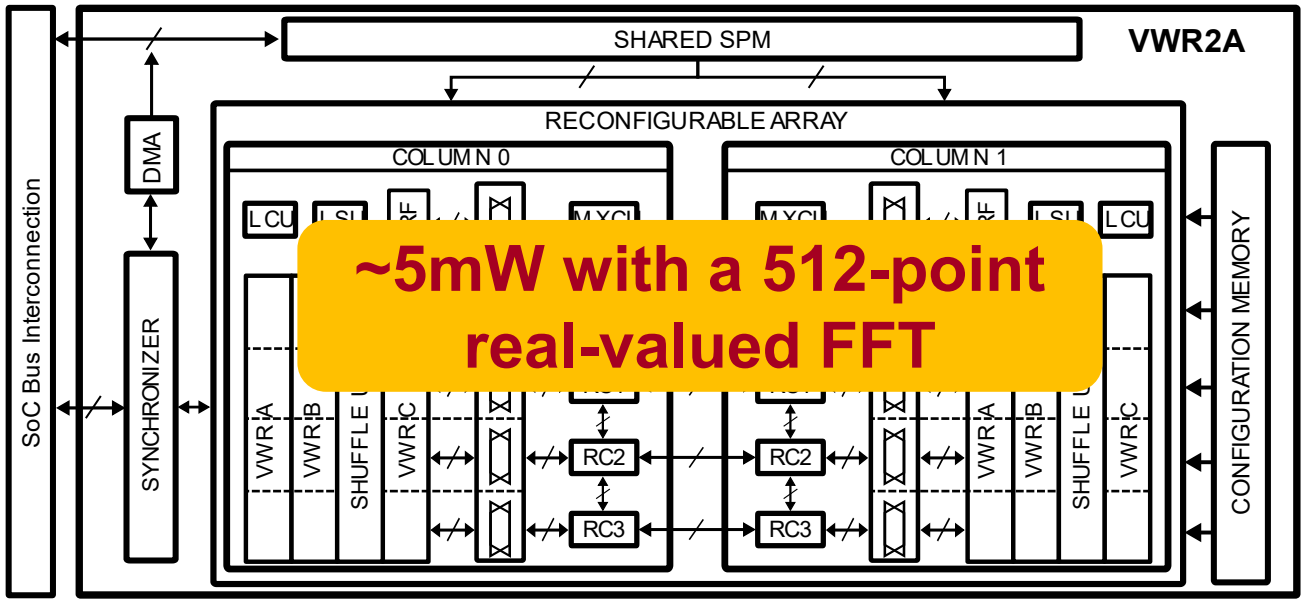
Ultra-Low Power Accelerators

SAT- MAPit Assembler [1] (USI-Lugano)

- *(2017) HEAL-WEAR: an Ultra-Low Power Heterogeneous System for Bio-Signal Analysis*
- *(2022) VWR2A: A Very-Wide-Register Reconfigurable-Array Architecture for Low-Power Embedded Devices*



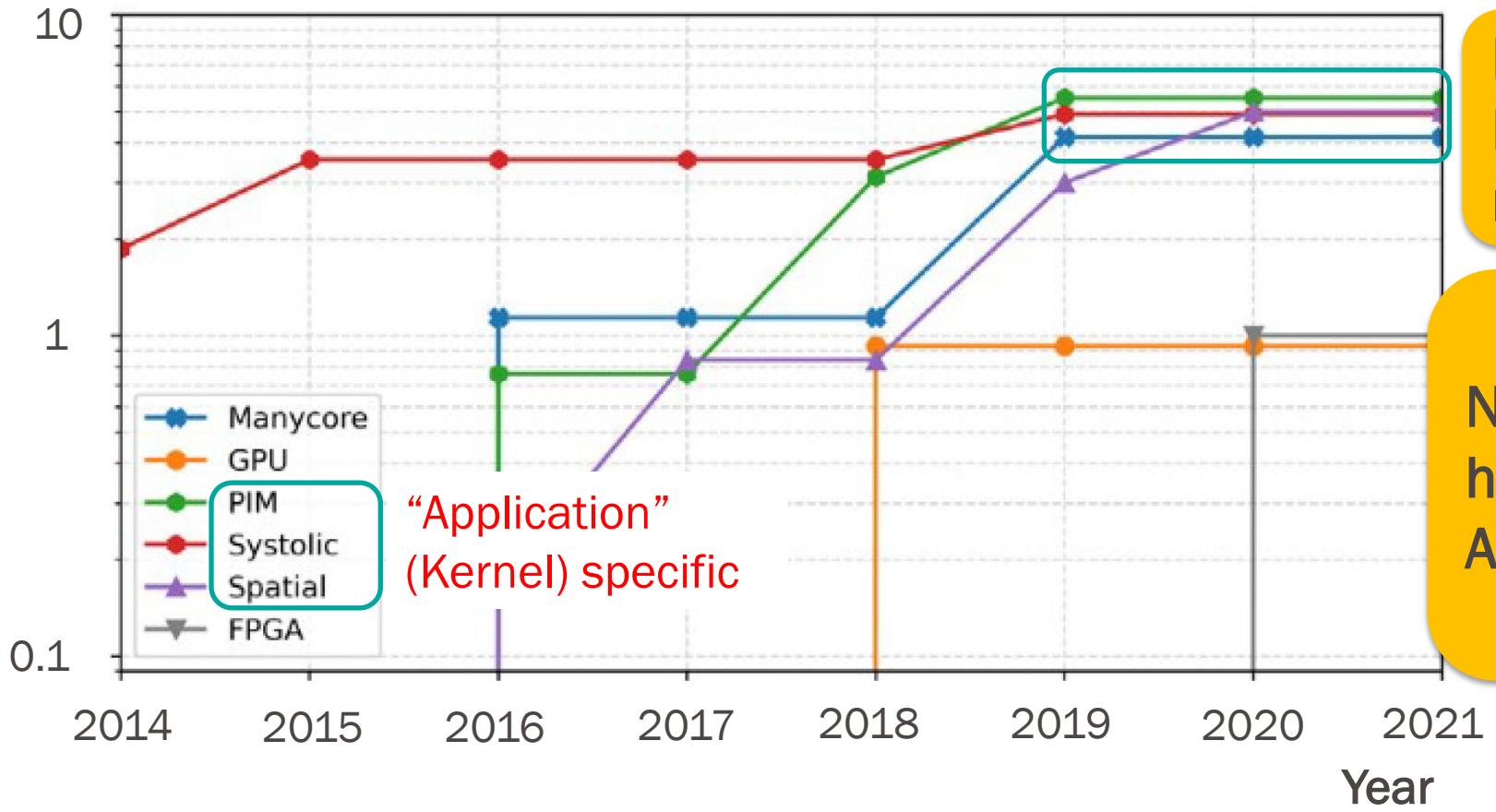
Heal-wear OpenEdgeCGRA
Optimized processing elements



VWR2A DSIP
Optimized memory hierarchy
(in collab. with IMEC)

[1] Tirelli, Cristian, Lorenzo Ferretti, and Laura Pozzi. "SAT-MapIt: A SAT-based Modulo Scheduling Mapper for Coarse Grain Reconfigurable Architectures." *Design, Automation and Test in Europe Conference Exhibition (DATE)*. 2023.

“Application”-Specific Edge AI Reaching a Limit…



Efficiency Plateau reached!

Need for heterogeneous AI/ML systems!

Architectures for conventional ML: DNN, CNN, RNN

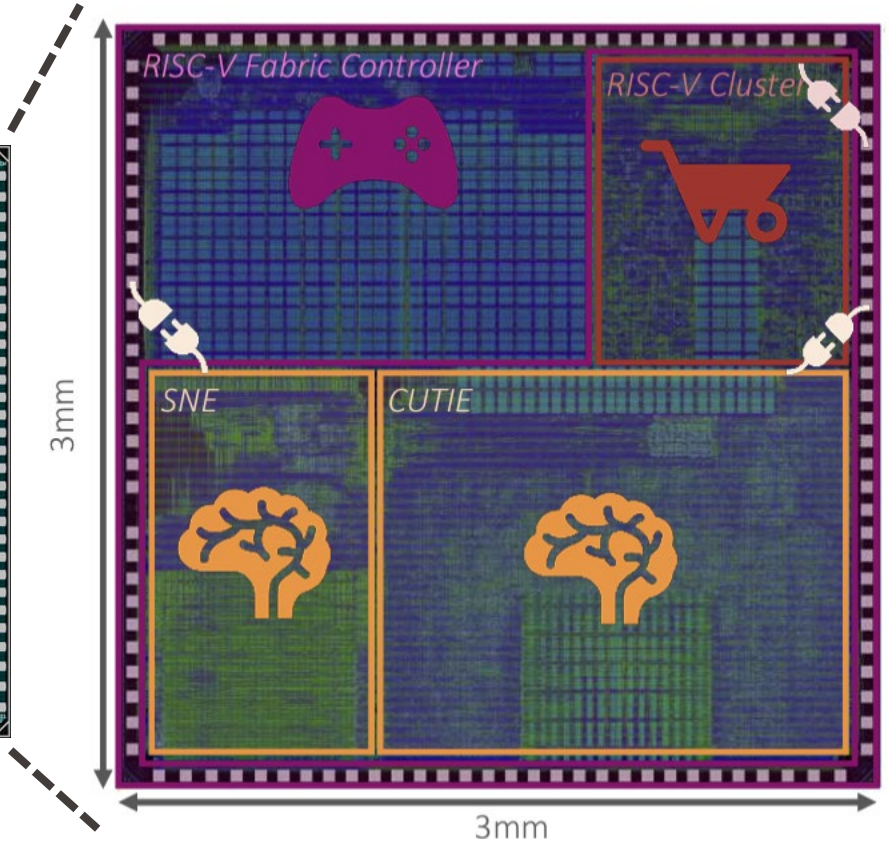
B. Peccerillo, et. al. , “A survey on hardware accelerators: Taxonomy, trends, challenges, and perspectives,” Journal of Systems Architecture, vol. 129, p. 102561, 2022.

New trend: Simple core + domain-specific accelerators (with true system codesign), need for exploration frameworks!

Heterogeneous platforms for EdgeAI: KRAKEN

Multi-Core + Domain-Specific Accel.

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



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 COOL CHIPS, 2022

Still **too high power for wearables** and it does **not consider the complete system** (true co-design approach needed!)

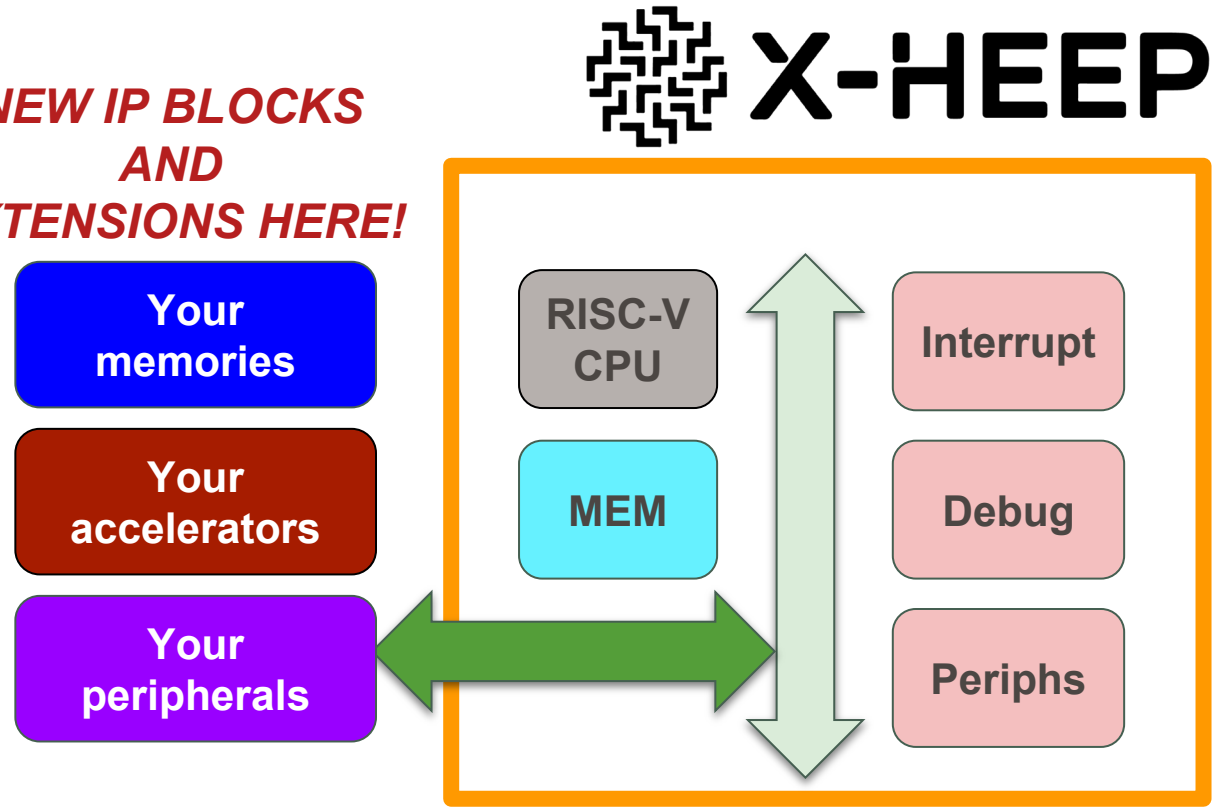
How to design **Deeply Heterogeneous SoCs?**

X-HEEP: eXtensible Heterogeneous Energy Efficient Platform

Open edge AI hardware framework for AI accelerators with IP/royalty-free designs!

X-HEEP provides the **basic blocks**, and we can make **research** around it

www.epfl.ch/labs/esl/research/2d-3d-system-on-chip/x-heep



This model encourages reutilization, long-term life, and collaboration between companies and academic institutions

Davide P. Schiavone, et al. "X-HEEP: An Open-Source, Configurable and Extendible RISC-V Microcontroller.“, RISC-V Annual Conference – Europe (2023).

How to design **Deeply Heterogeneous SoCs?**

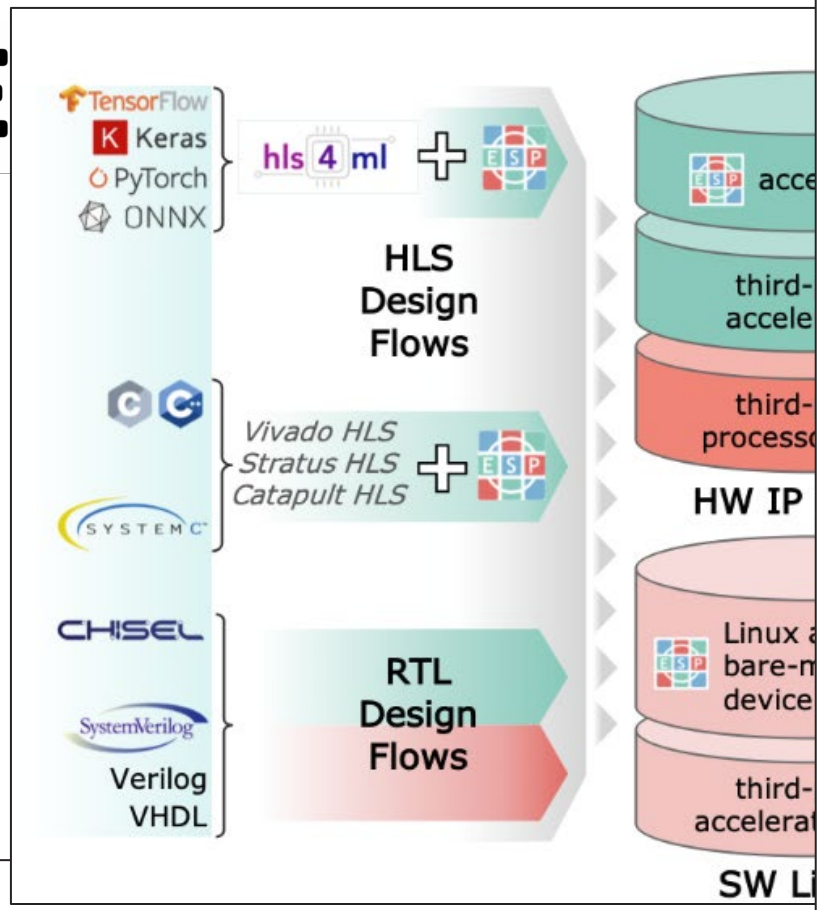
New Open-Source Frameworks

X-HEEP

Configurability

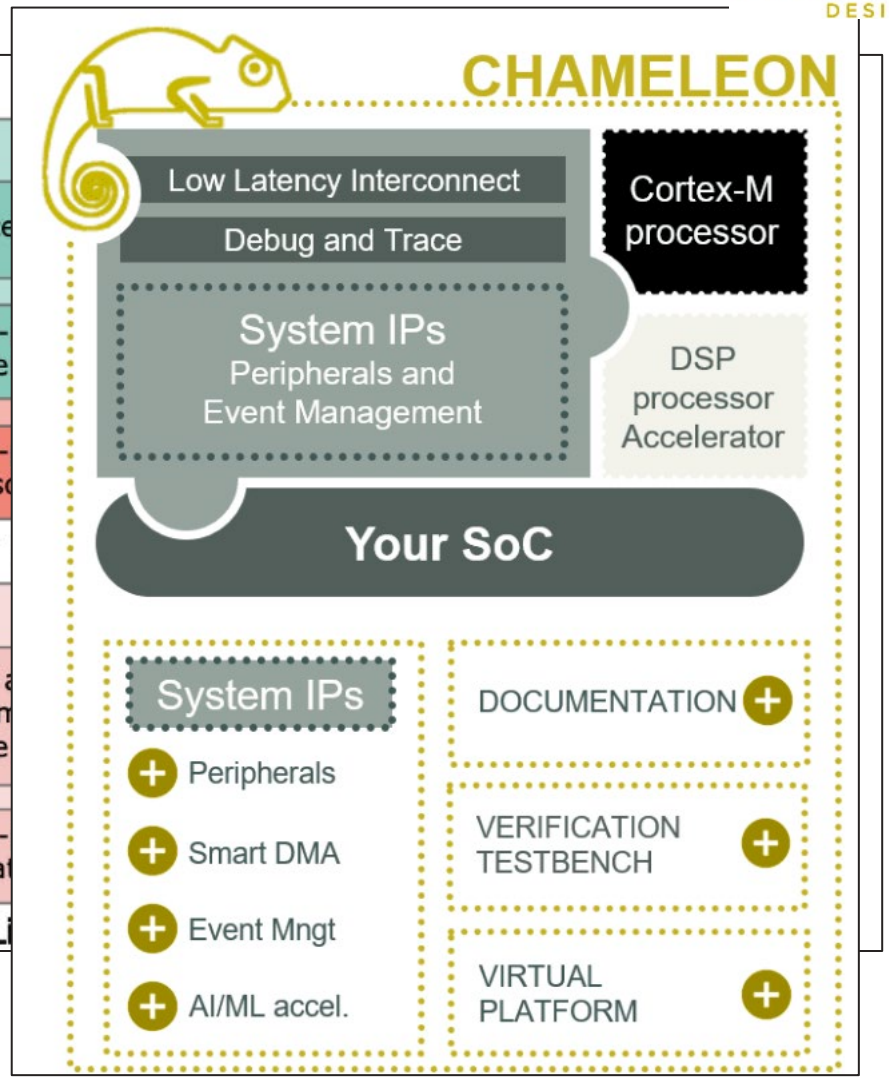
1. RISC-V core
2. Coprocessor interface
3. Peripherals
4. Interrupt controller
5. Accelerator interface
6. Power manager
7. Bus topology
8. Number of banks

x-heep.epfl.ch



www.esp.cs.columbia.edu

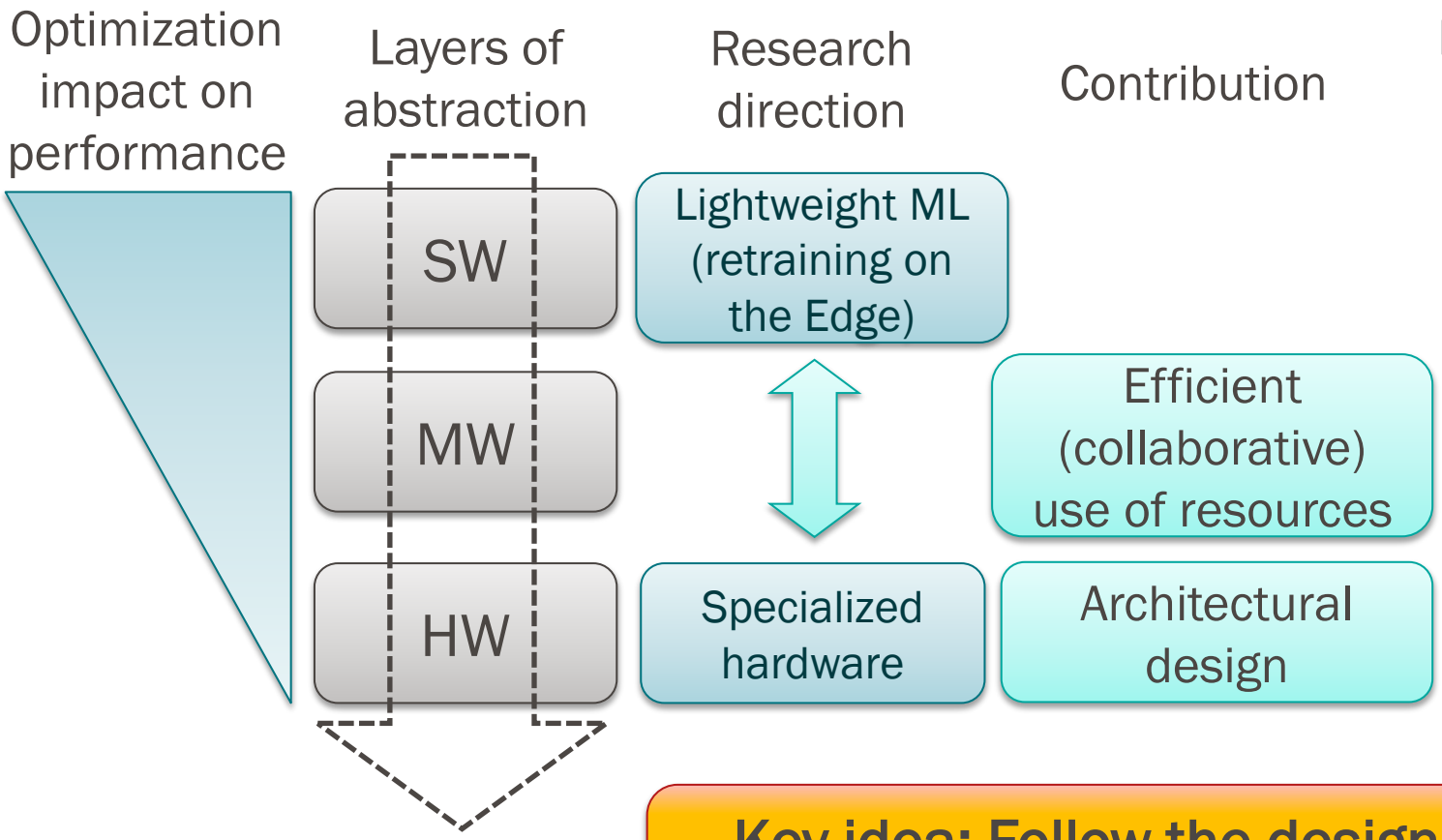
DOLPHIN DESIGN



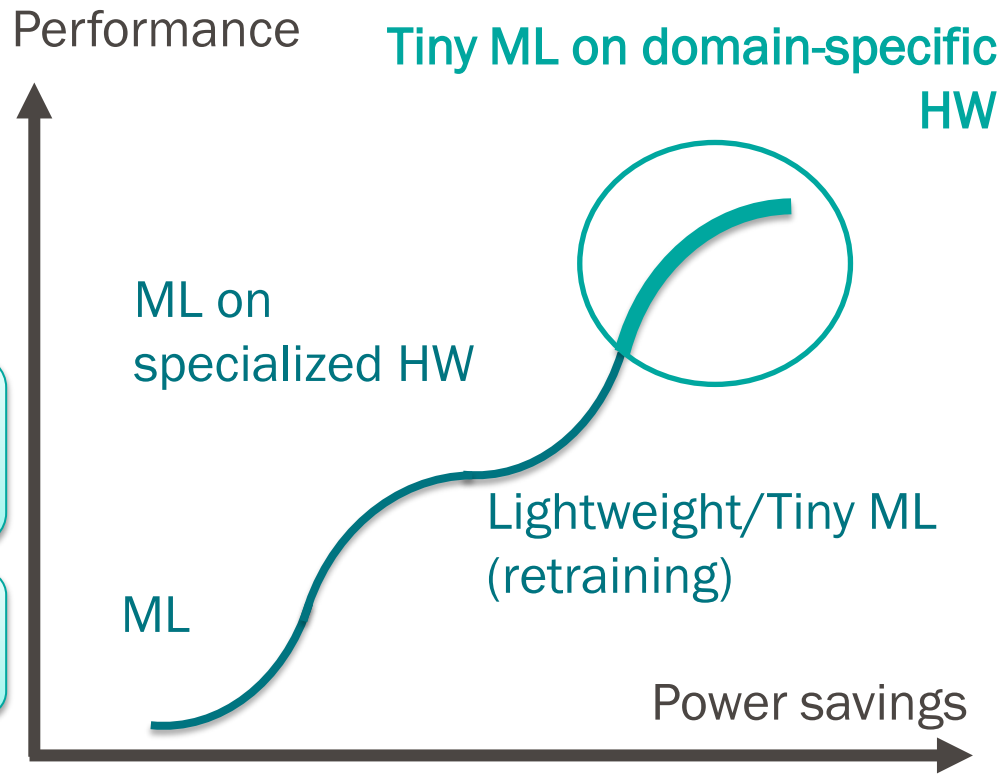
www.dolphin-design.fr/chameleon-mcu-subsystem

Exploiting **Domain Knowledge** for **System Co-Design**

FL & ML Deployment

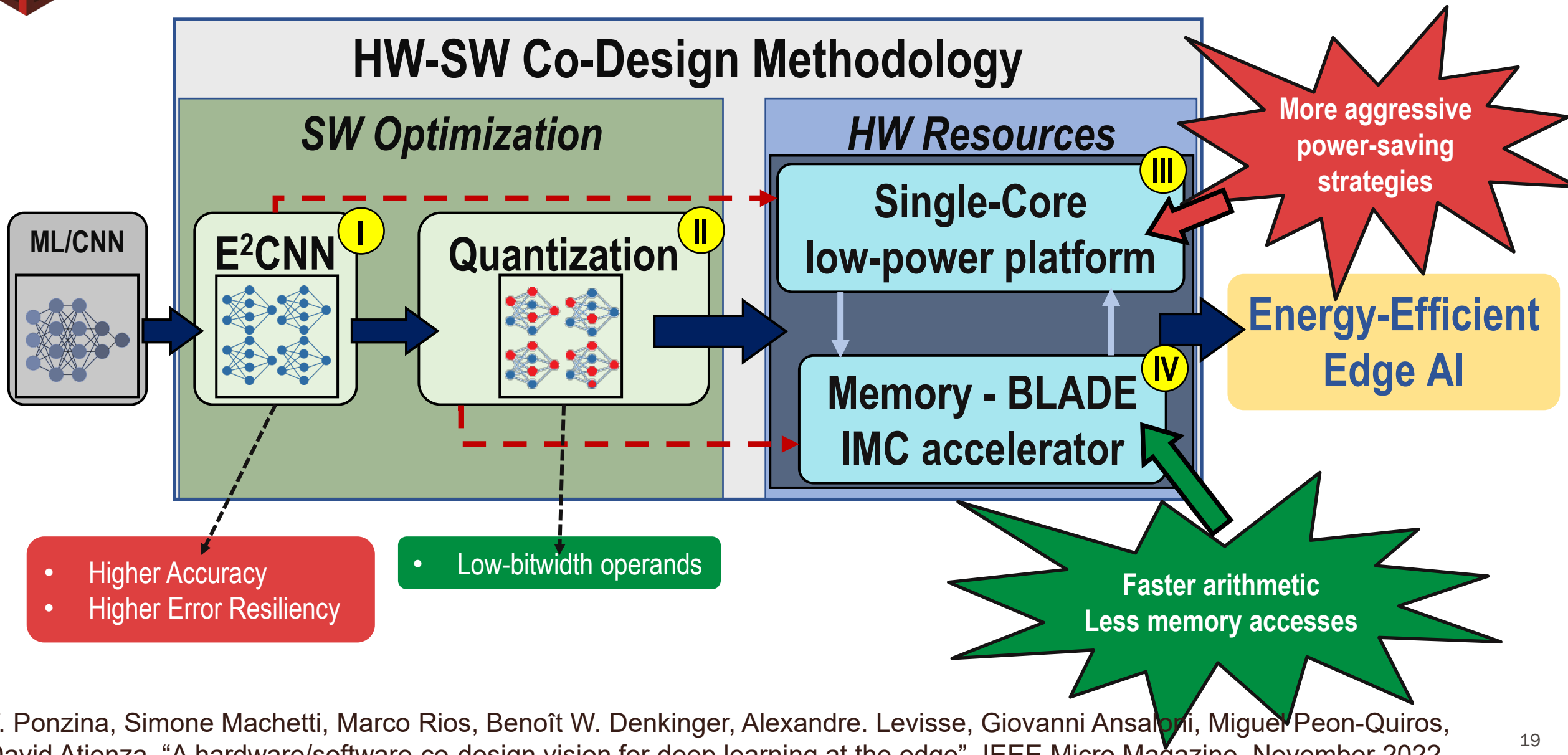


Domain-Specific Exploration



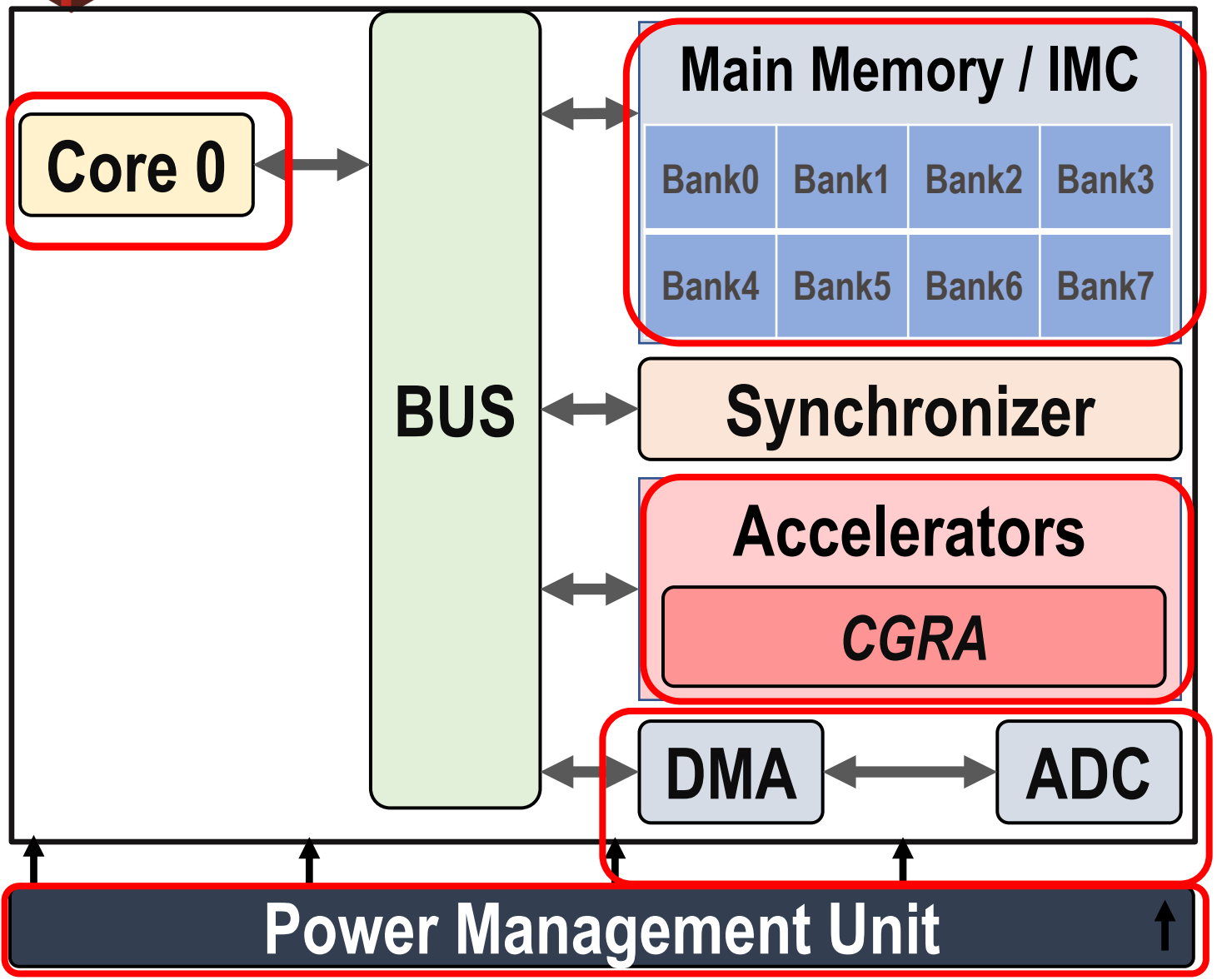
<u>Key idea:</u> Follow the design concepts of biological systems!
Let's use medical example for edge AI

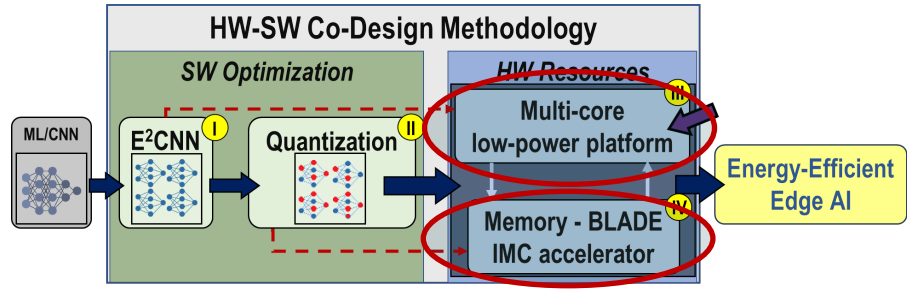
Medical Usecase: System Co-Design for Edge AI on FL Era



F. Ponzina, Simone Machetti, Marco Rios, Benoît W. Denkinger, Alexandre. Levisse, Giovanni Ansaloni, Miguel Peon-Quiros, David Atienza, "A hardware/software co-design vision for deep learning at the edge", IEEE Micro Magazine, November 2022

X-HEEP for Healthcare: **HEEPocrates**

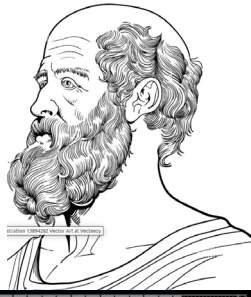




- Single-core architecture
  - Control of accelerators flow (parallel execution)
- Independent memory banks
  - Switch-off unnecessary banks
- Coarse-Grained Reconfigurable Accelerator (CGRA) and In-Memory Computing (IMC)
  - CGRA: compute-intense kernels (irregular flow)
  - IMC: Simple ML ops with regular comp. flow
- Power Management Unit
  - Voltage/frequency over-scaling
  - ADC (event-based adaptive sampling)

https://www.epfl.ch/labs/esl/research/2d-3d-system-on-chip/x-heep/

HEEPocrates: first Open-Source Brain-Inspired Edge AI Architecture

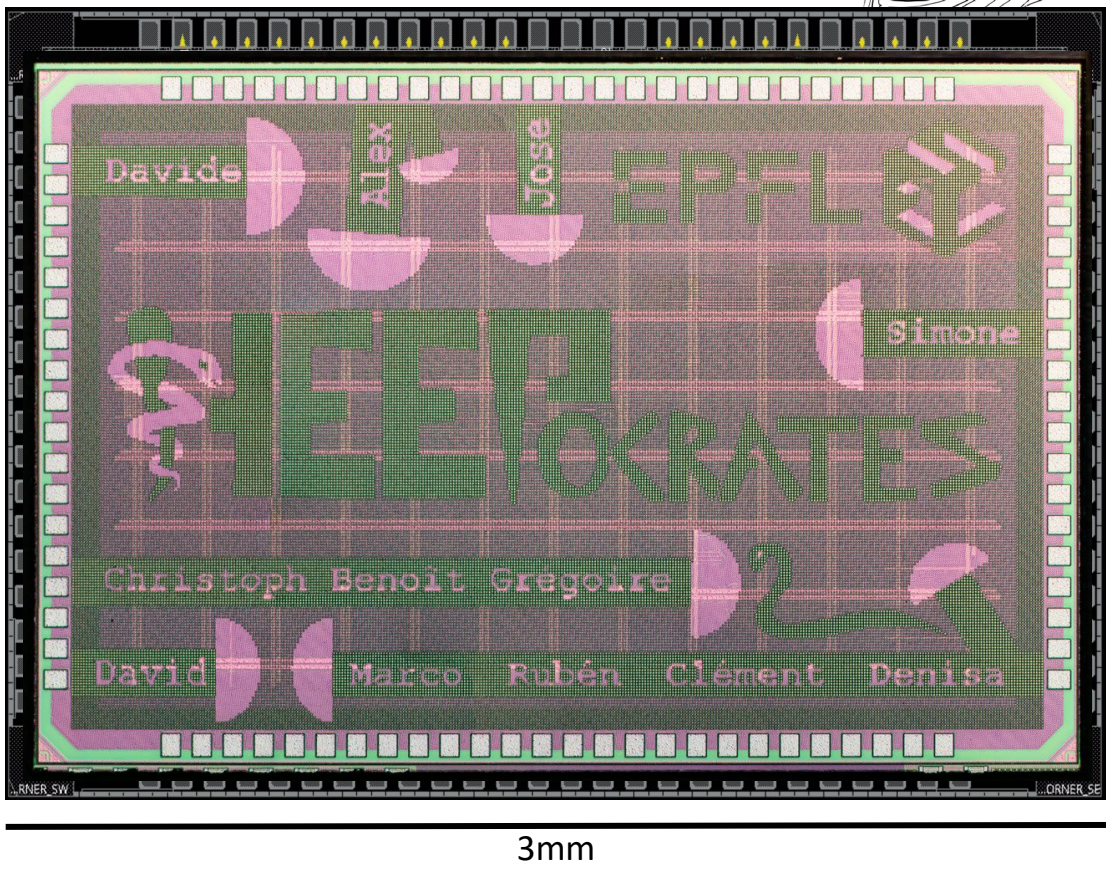


CPU: Core-V RISC-V [1]
- Ibex

Bus: AMBA AXI interfaces

Memory: 8 banks, 64KB each

ASIC implementation, 65nm TSMC
- Area: 6mm²
- Max frequency: 250MHz
- Power consumption: **28uW/MHz**



Extensions enable ACCELERATORS:
1. Coarse-Grained Reconfig. Array (CGRA)
2. In-memory (bit-line) computing

Complete design done in 6 months

www.epfl.ch/labs/esl/research/2d-3d-system-on-chip/x-heep/

OpenHW group github: github.com/openhwgroup

System Level Co-Design

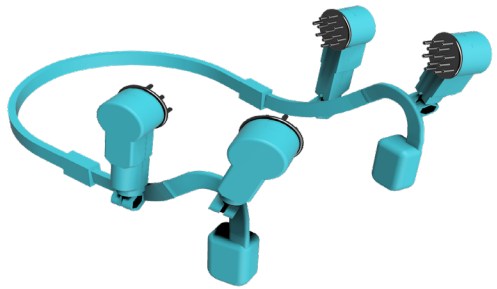
e-Glass: A system for real-time seizure monitoring

Comparison:
24 vs 4 electrodes

e-Glass first prototype.

Glasses embodiment **minimize social stigma**
(new version: bone conducting headset)

EpiPhone



Sensors:

- EEG:
  - 24-bits
  - 3 channels
  - Soft-dry electrodes
- Accelerometer (3-axial) /Gyroscope

Interfaces:

- Bluetooth 4.2
- USB 2.0

Processing – Generation 3:

- HEEPocrates – Ultra-low power edge AI
- Onboard memory: 64 MB (up to 7 days of recording of EEG signals)

Battery powered: up to 24h monitoring.

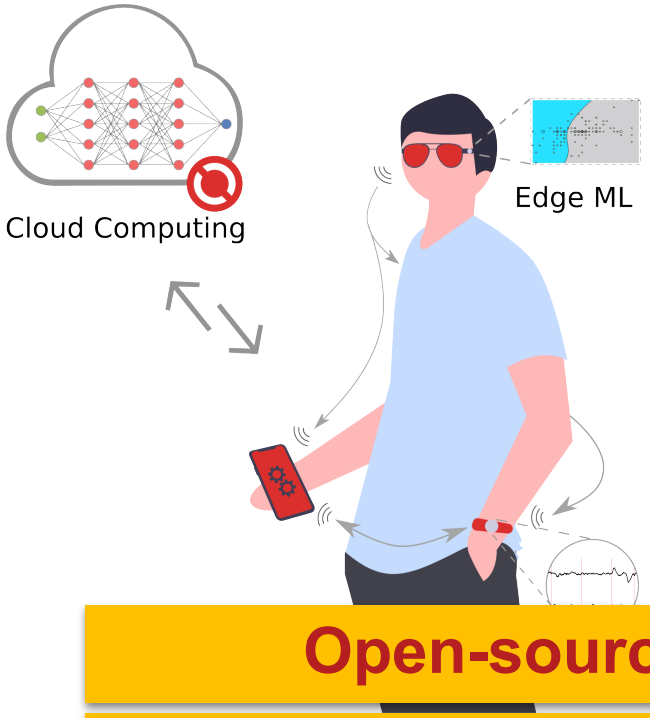
D. Sopic, A. Aminifar, and D. Atienza, "e-Glass: A Wearable System for Real-Time Detection of Epileptic Seizures," in 2018 IEEE International Symposium on Circuits and Systems (ISCAS). Florence, Italy: IEEE, may 2018, pp. 1–5.

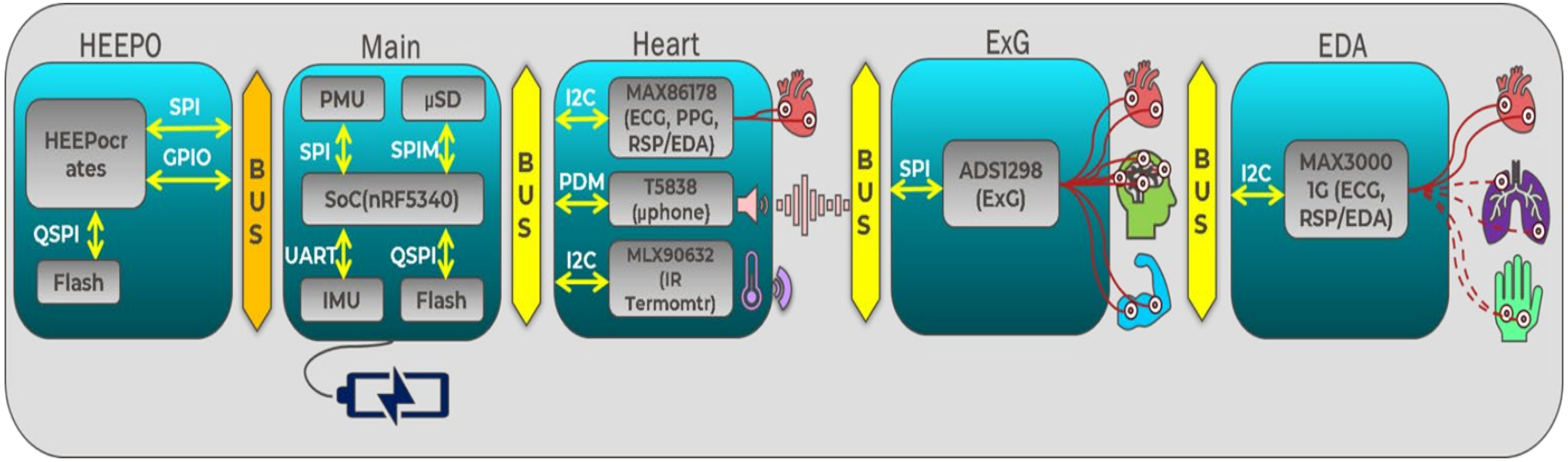
Enabling System Level Co-Design R&D

VersaSens: Multi-Parametric Plug&Play EdgeAI System

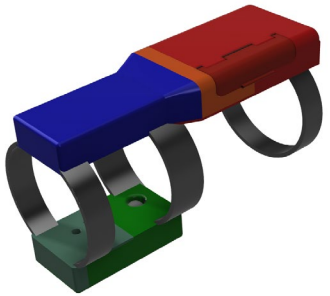
Plug&Play your edge AI devices

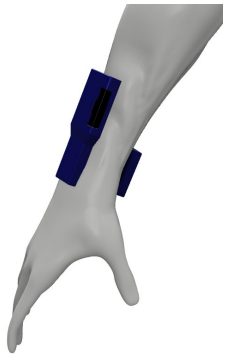
2) Adapt-Expand
3) Connect in any position





www.epfl.ch/labs/esl/research/smart-wearables/versasens/





Open-source, available to all!

It can reach 85% accuracy in epilepsy, and we can use FL!

Discussion

FL/AI/ML with IoT

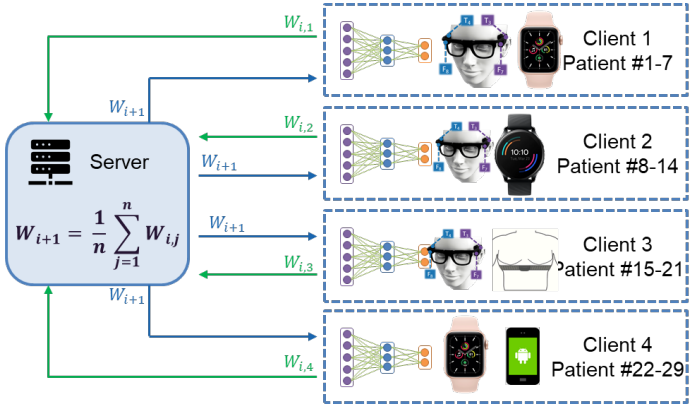
- New computing systems
- Different than Von Neumann, true low power needs!

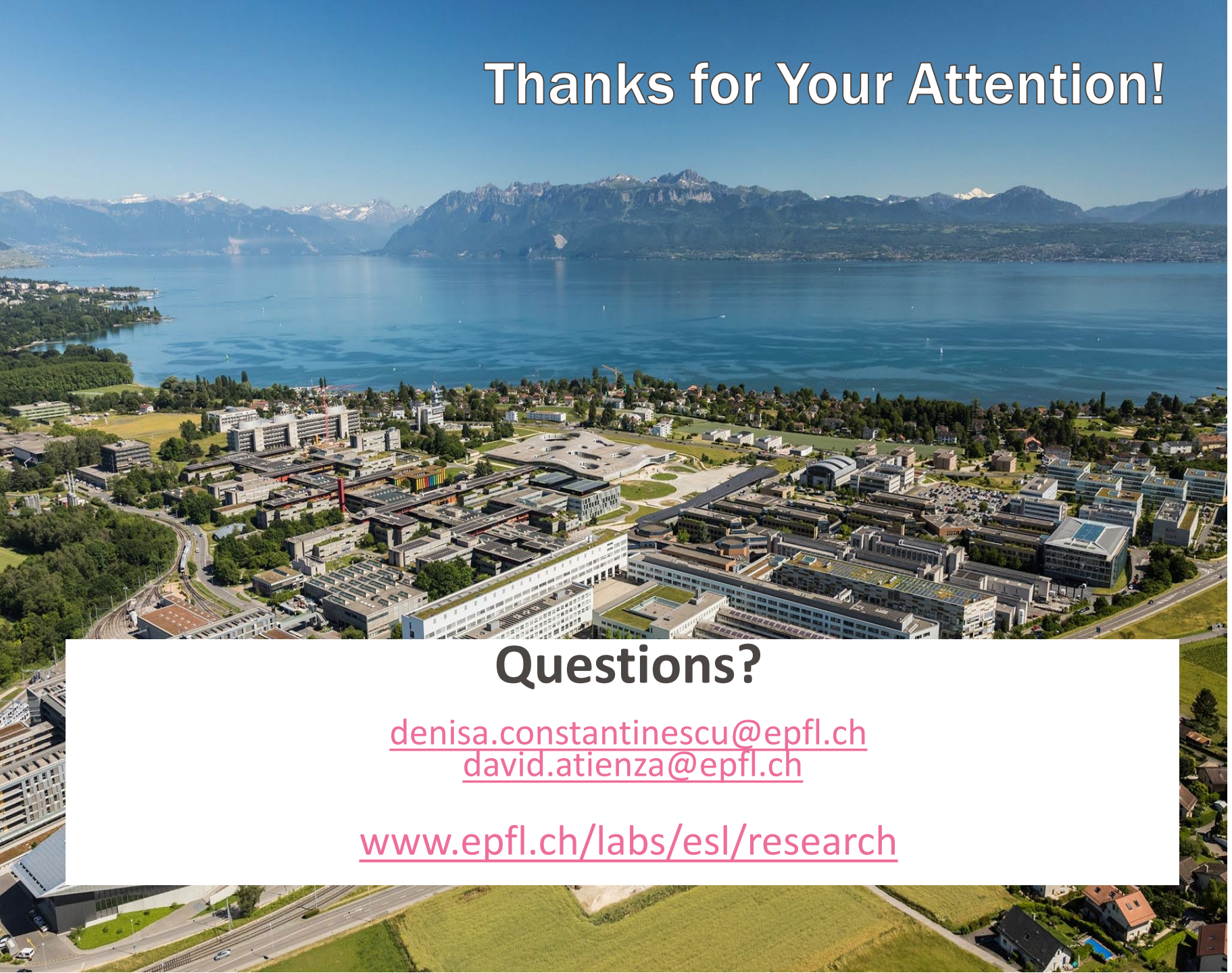
Democratization of IoT chip design: X-HEEP, PULP, etc.

- New system-level flows: combining accelerators with ensembles of NNs

Neuro-inspired co-design: next-gen. edge AI systems

- Enables efficient use of FL for efficient edge AI training: results competitive with not so advanced technologies!





Thanks for Your Attention!

Questions?

denisa.constantinescu@epfl.ch
david.atienza@epfl.ch

www.epfl.ch/labs/esl/research

Acknowledgements

Dr. Saleh Baghersalimi,
Taraneh Aminosharieh,
Rubén Rodríguez Álvarez,
Dr. Flavio Ponzina,
Simone Machetti,
Dr. Benoît Denkinger,
Dr. Christoph Müller,
Dr. Davide Schiavone,
Dr. Miguel Peón-Quirós,
Dr. Giovanni Ansaloni

And thanks to