EPFL

EMBEDDED SYSTEMS LABORATORY

Designing Accelerator-Centric Edge AI Architectures for Cyber-Physical Systems

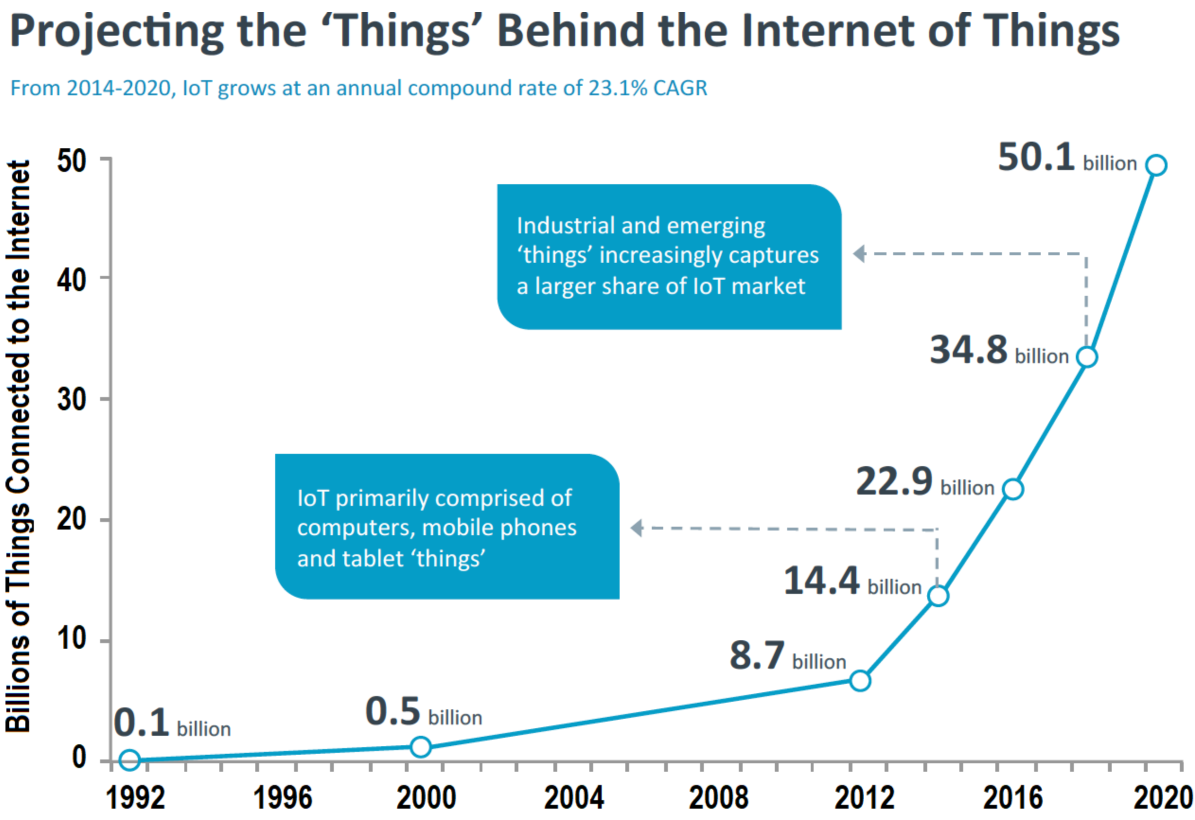
Dr. Jose Miranda

jose.mirandacalero@epfl.ch

Lecture - IEEE Cyber-Physical Systems Summer School 2024, September 17th, 2024

Internet of Things (IoT): Edge AI Systems

- IoT market constantly growing (20B devices by 2025)
- Artificial Intelligence (AI) and Machine Learning (ML) used for data analytics and classification
- Sustainable? Expected increase in energy consumption by **3x in 2040 in current trends** [IDC'22]



How to increase IoT efficiency? Edge AI systems

But adaptive (domain-specific and fast to build: co-design!)

1) Less data transmitted over energy-hungry comm. links
2) Faster response = Less latency
3) Increased system knowledge (e.g., medical systems)

The Sky is the Limit: Gen. AI at the Edge!

IoT/AI Market:
$76 billion by 2028
(TIRIAS RESEARCH)

20% Offload to Edge →

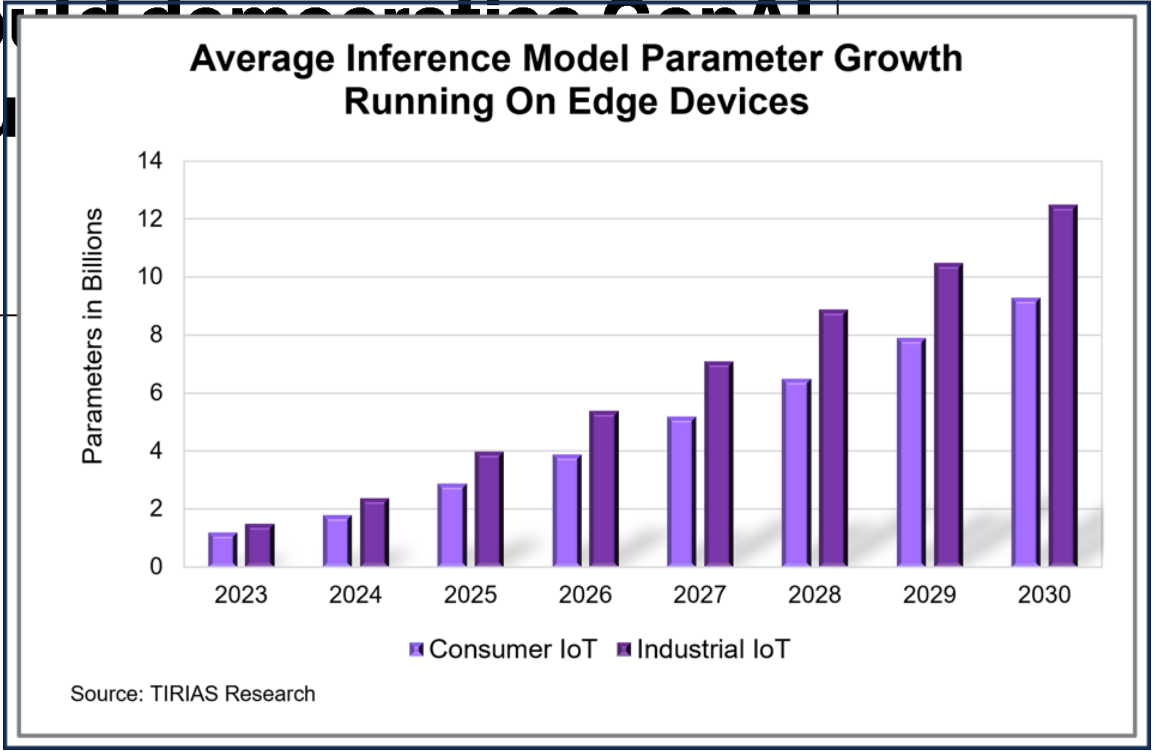
AI at the Edge: $15 billion
(savings of 800 MW)

WORLD ECONOMIC FORUM'24

EMERGING TECHNOLOGIES

On-device AI wo
and ensure inclu
the economy

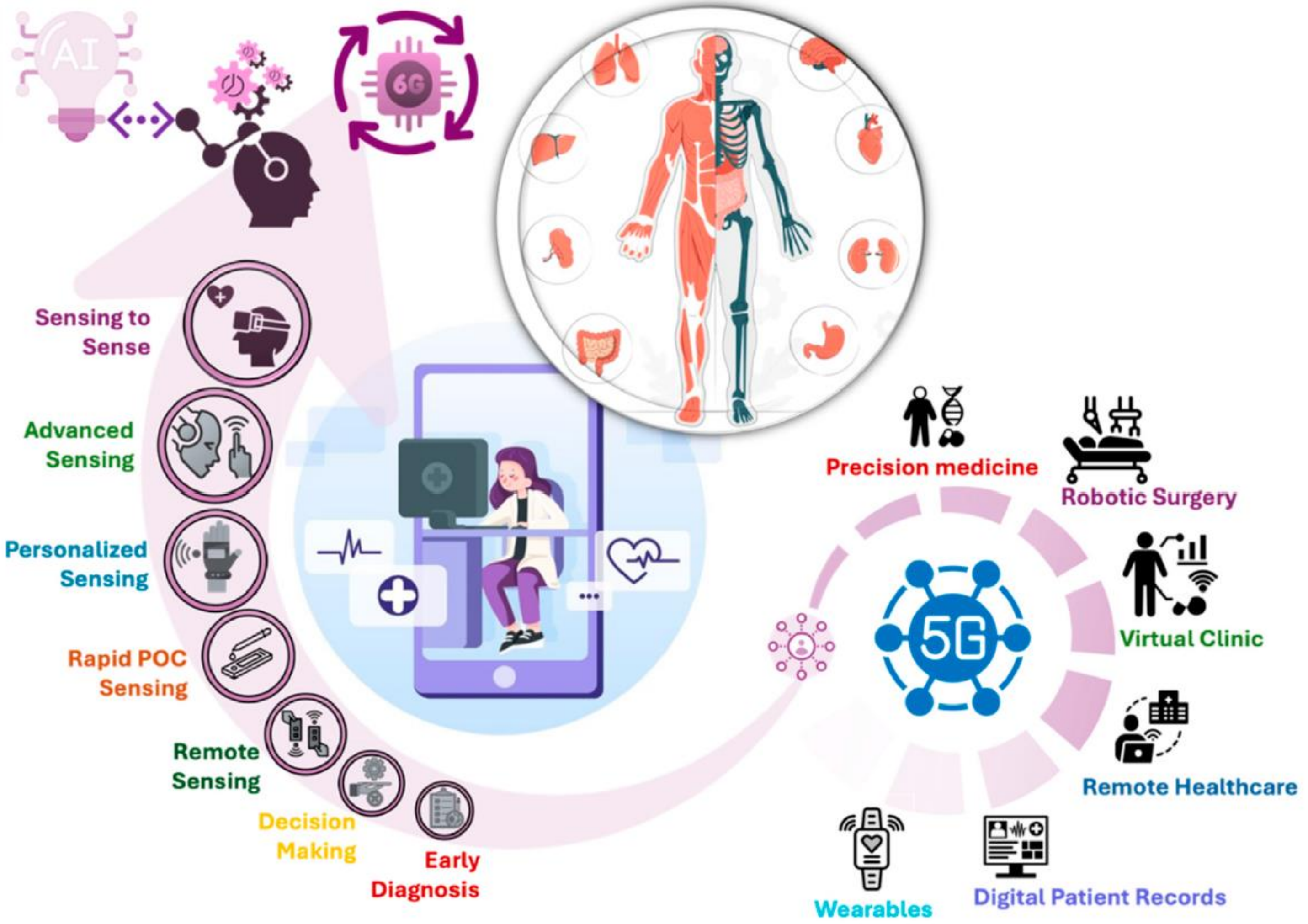
Jan 15, 2024



SoCs Performance improvements ≠ **Parameter growth**

The sixth sense is going into the edge!

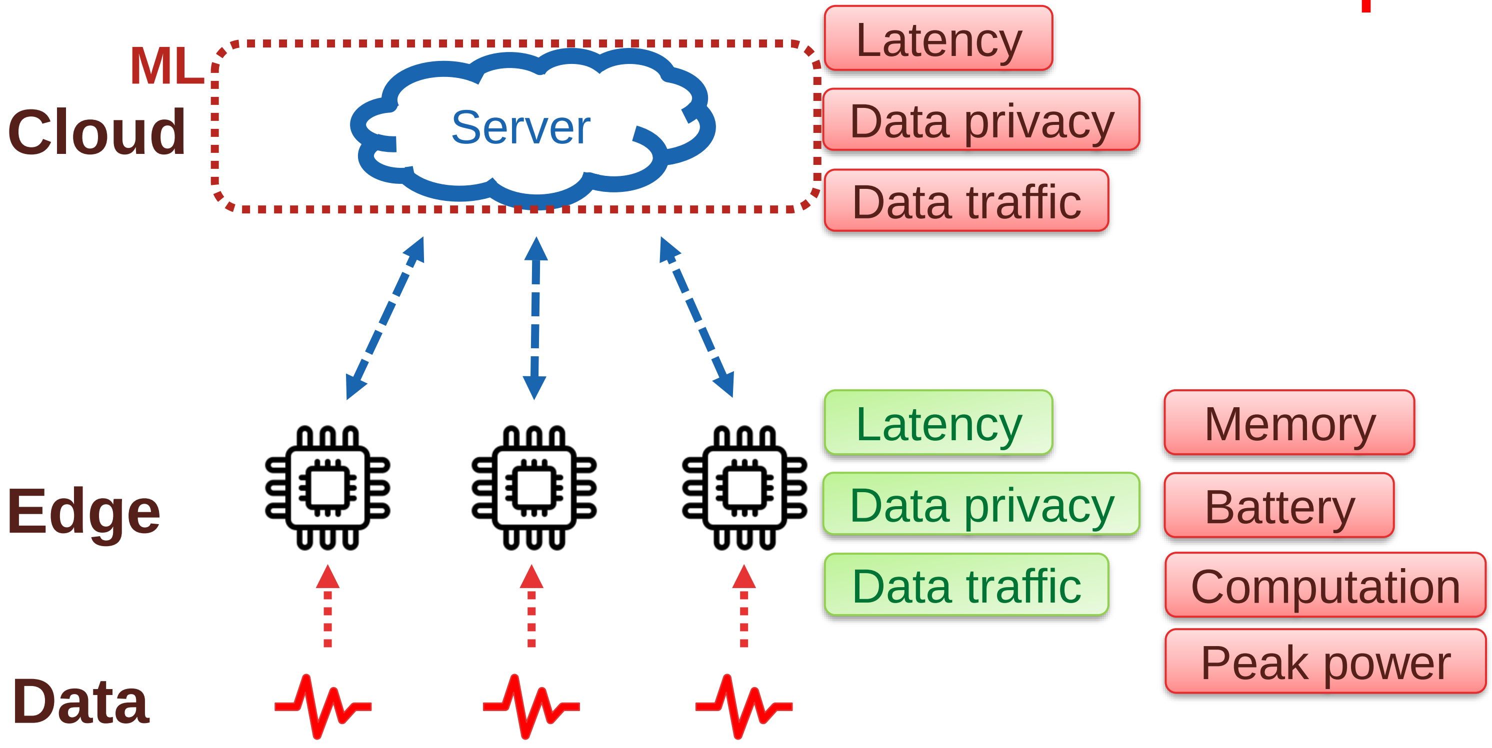
- From sensing to sense:
  - What is a sensor?
  - How do we integrate advanced computational approaches (AI, ML, DL)?
  - How do we make the sensors context-aware?
  - How do we achieve a truly responsive to real-time environments?
  - How do we personalize sensing?
  - How do we assure interoperability?



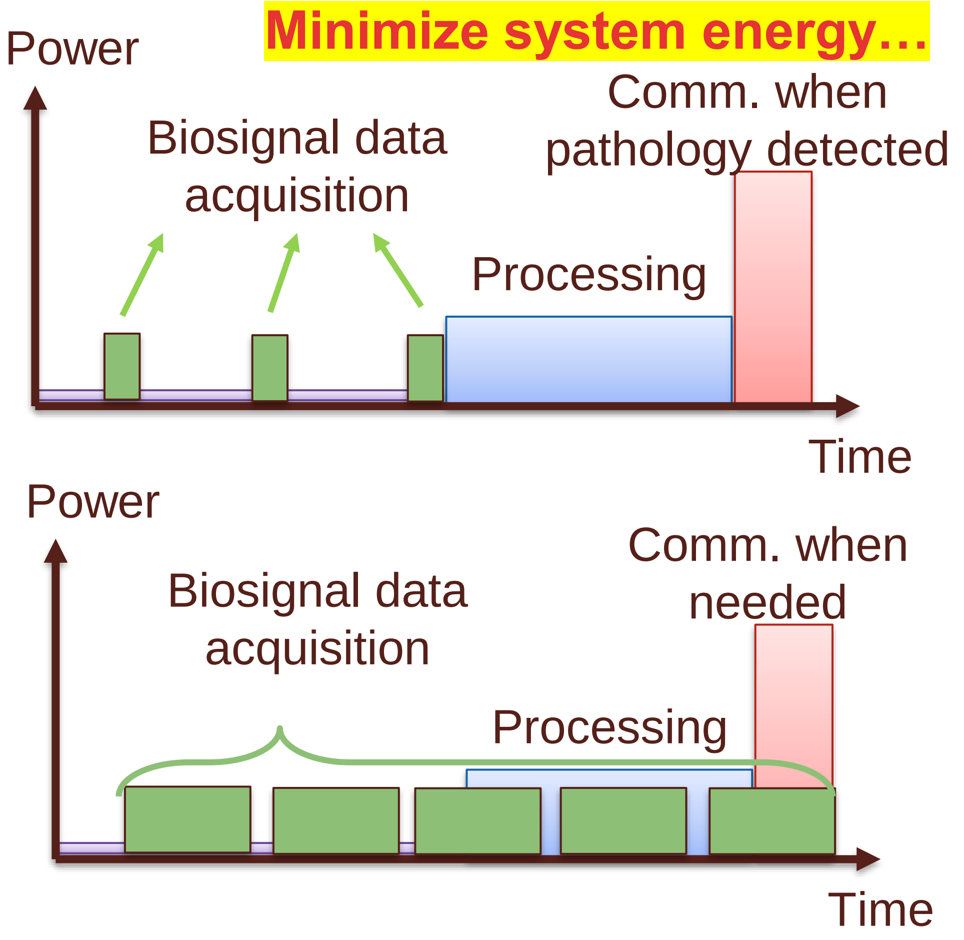
https://www.frontiersin.org/journals/nanotechnology/articles/10.3389/fnano.2024.1434014/full

Domain-Specific Edge AI to the Rescue: Key Properties for Medical Systems

State-of-the-art

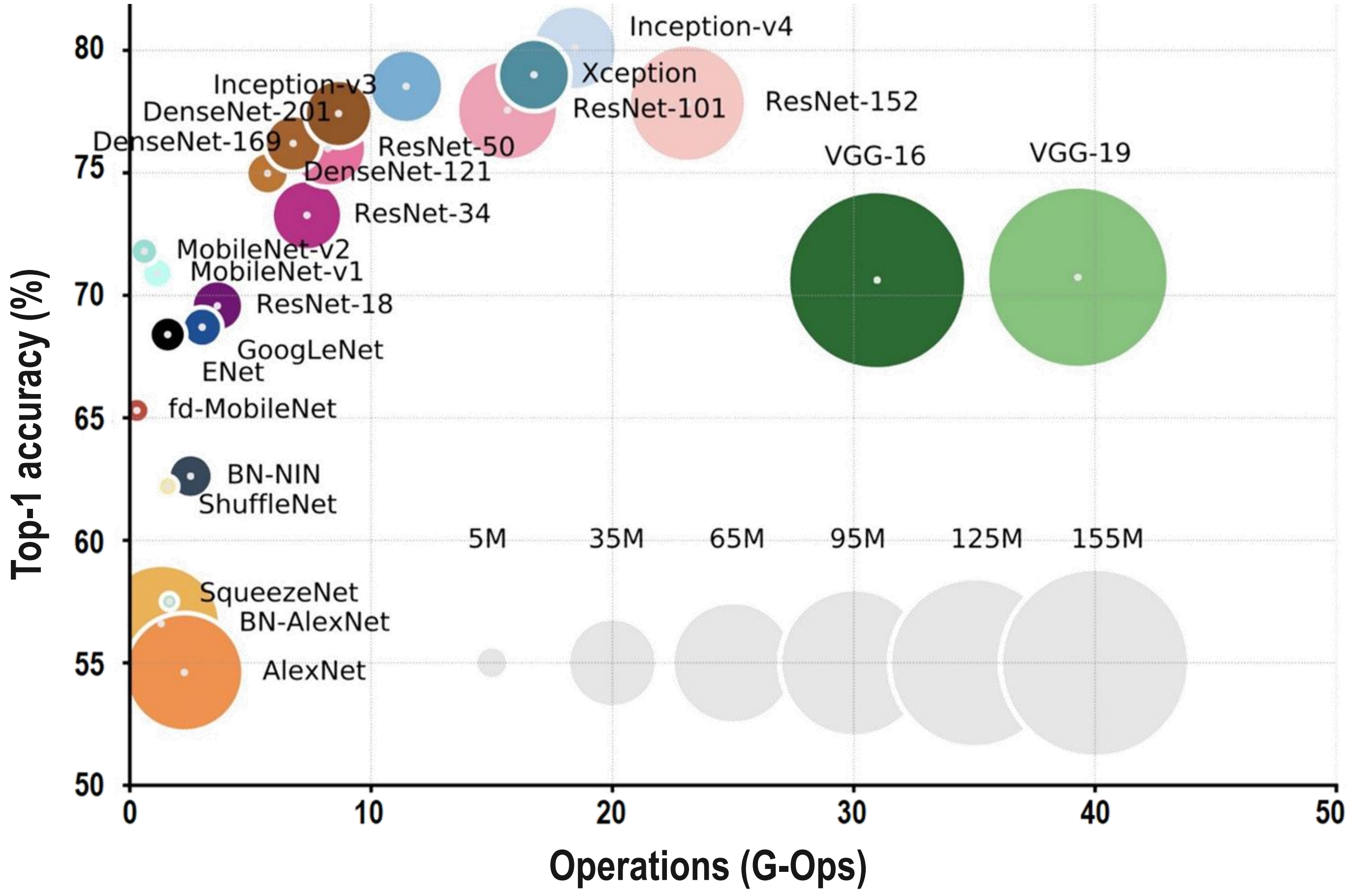


Medical IoT applications include clear phases for edge AI systems design



AI/ML Models Getting Very Complex: Co-Design Needed

- High accuracy achieved through:
  - Large models
  - Complex connections
  - Ensembles of CNN models



Efforts to use AI/ML on IoT nodes: edge AI systems

Software + Hardware Optimizations

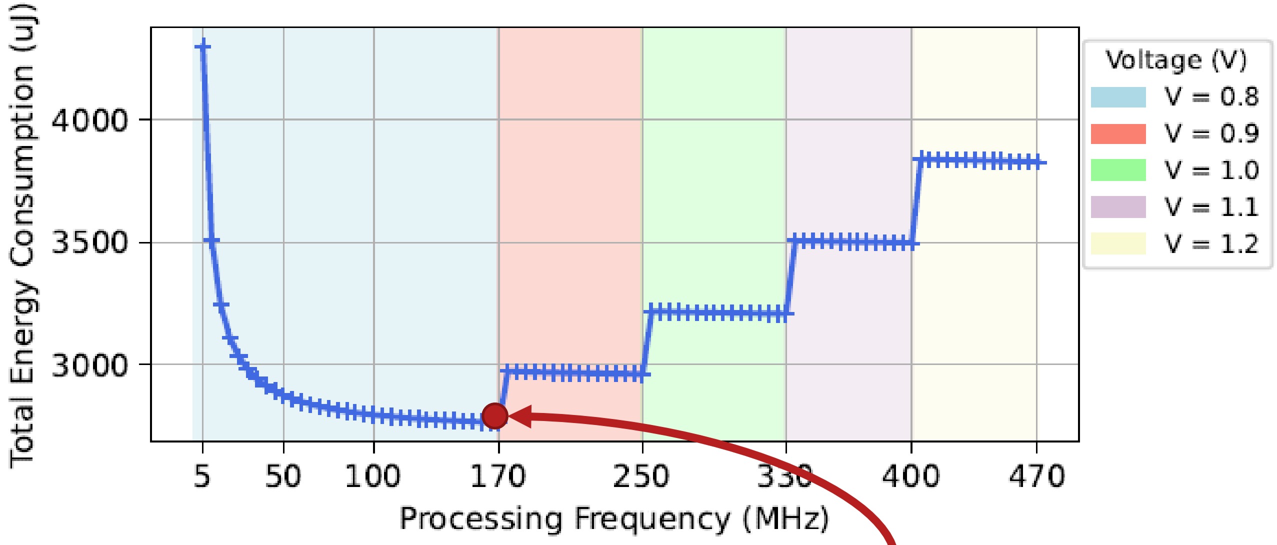
Computation Acceleration

Voltage and Freq scaling

IoT/AI applic. characteristics are key to design the final edge AI system!

Targeting Voltage and Frequency at the edge

Sometimes we can reduce energy by playing with both voltage and frequency of the system

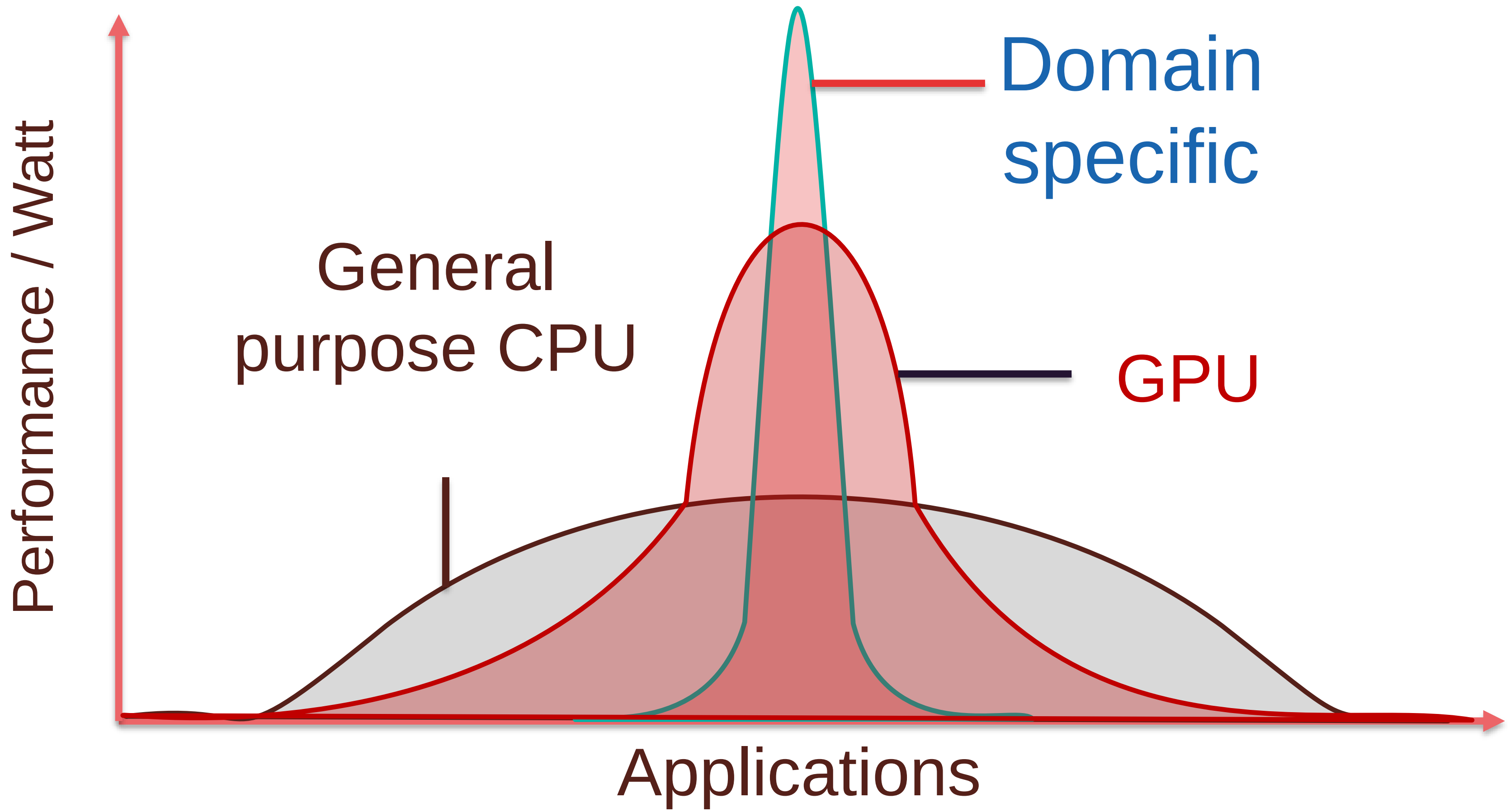


Real application example 12-Lead Heart Beat Classifier: When choosing the right point we can save up to 58% of the total energy consumption!

Warning!!

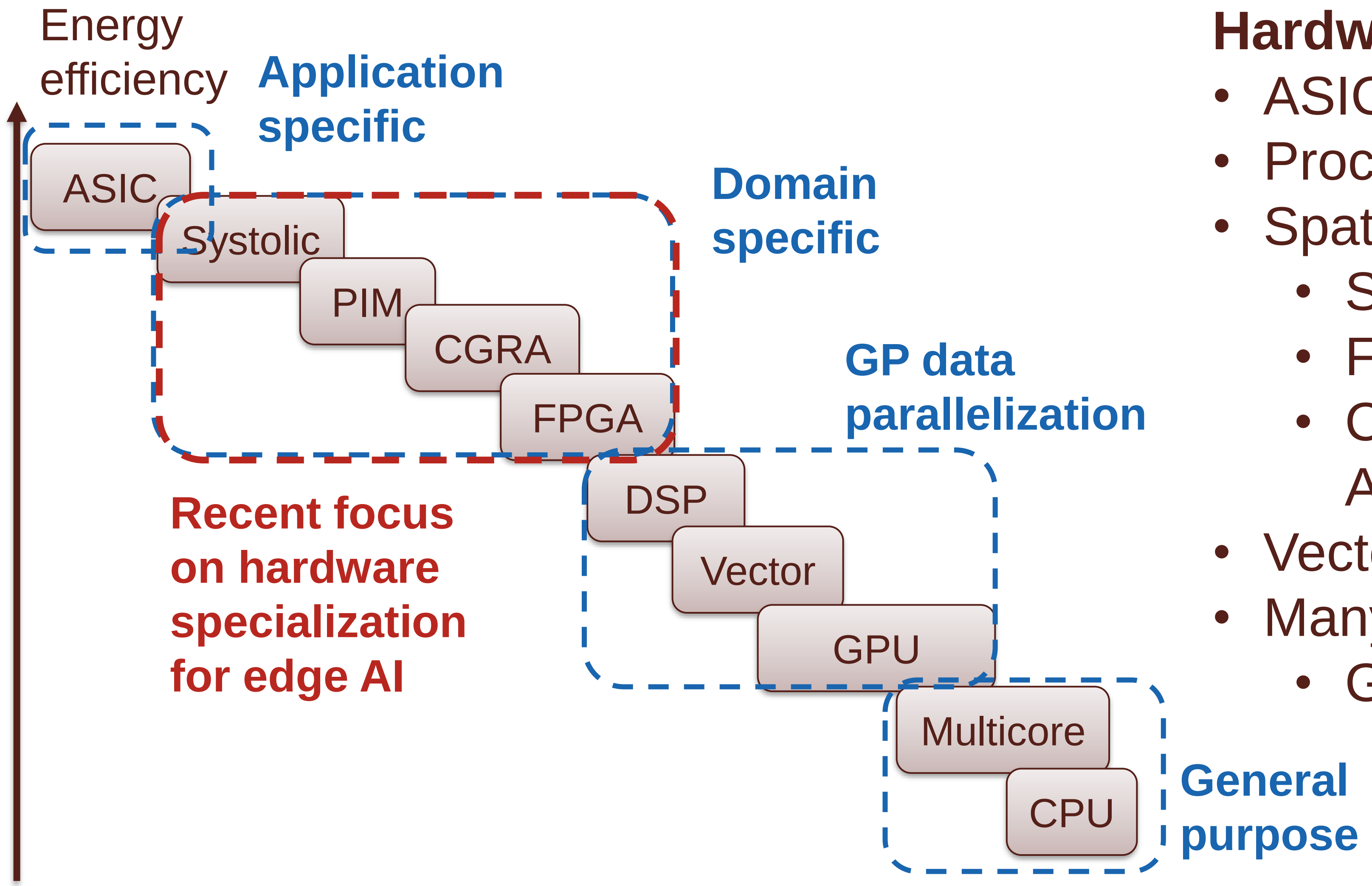
Usually, although this decision highly depends on the platform, for high-bandwidth applications the optimal frequency is determined by first selecting the lowest voltage that enables a frequency at which the system can meet its deadlines (i.e. maintaining a uniform frequency)

Targeting Computation acceleration at the edge



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

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

GenAI into the Edge is “Almost” there

NVIDIA

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.

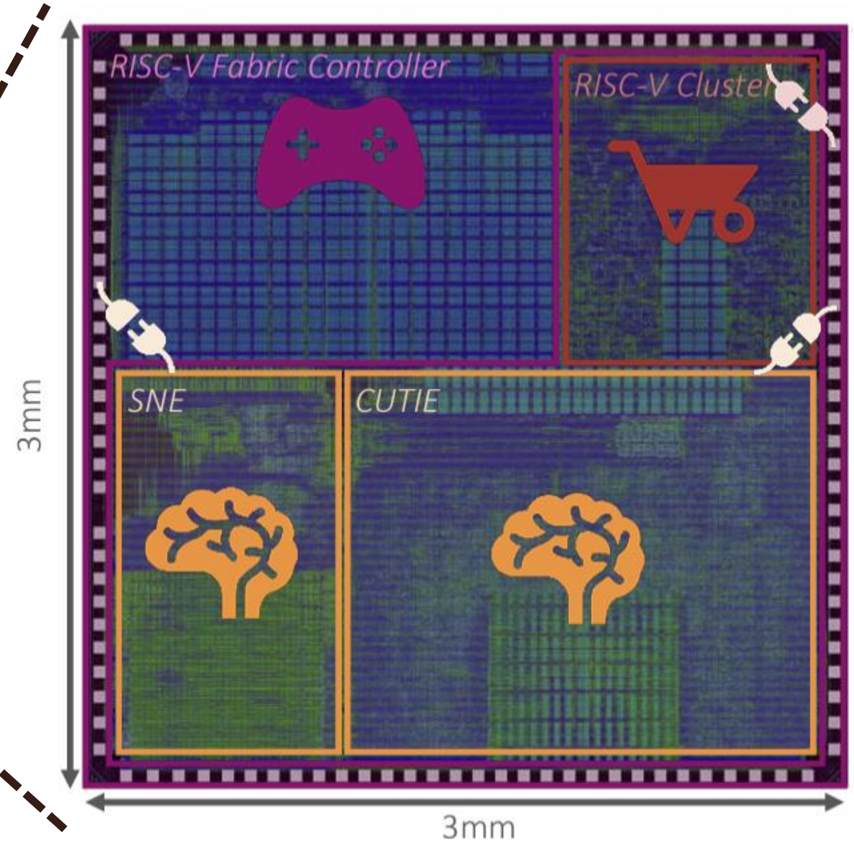
B. Keller *et al*., "A 17–95.6 TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization for Transformers in 5nm," *2022 IEEE VLSI Technology and Circuits*, Honolulu, HI, USA, 2022, pp. 16-17.

It provides energy-efficient inference with transformers (BERT):
95.6 TOPS/W – 1711 inferences/s/W – 0.7% Accuracy loss

But too high power for Medical edge AI systems…
(Nvidia statement: “just a few Watts”)

Need for domain-specific knowledge, technology alone does not work!

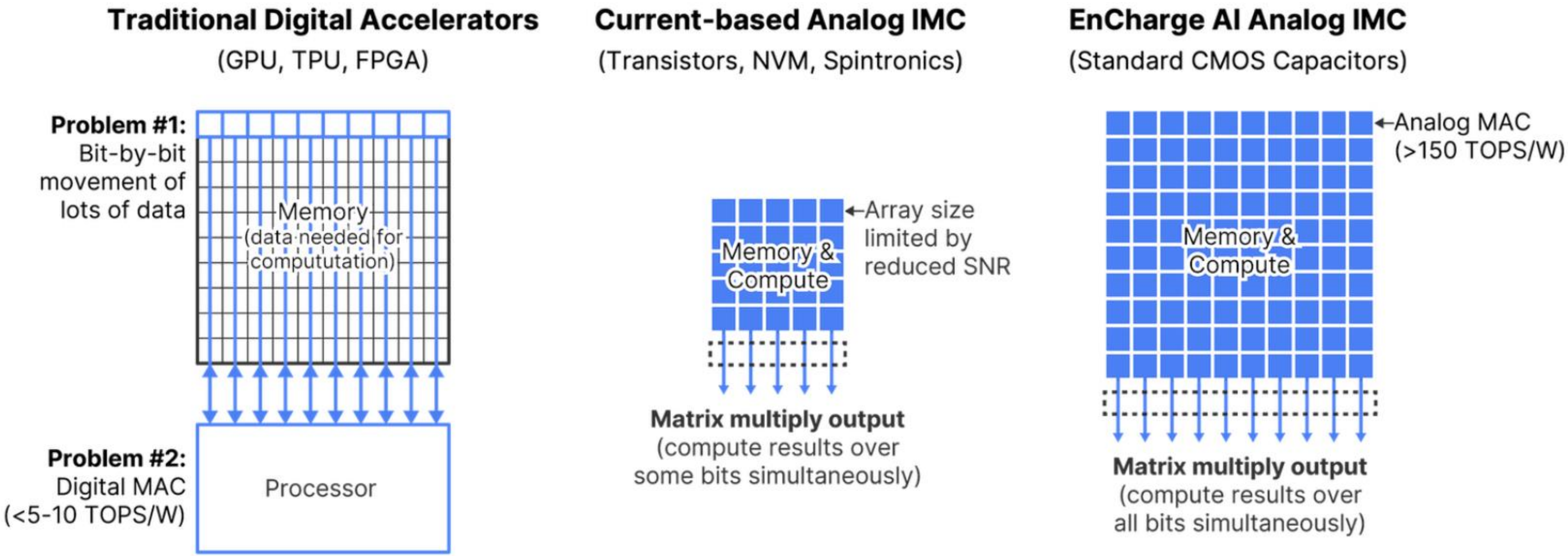
- RISC-V Cluster
- SNE – Spiking NN accelerator
- CUTIE – Ternary Neural Network
  - **> 1 PetaOps/s/W for Transformers**



Still too high power for edge AI in medical IoT, it does not use domain-knowledge (medical system co-design needed!)

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

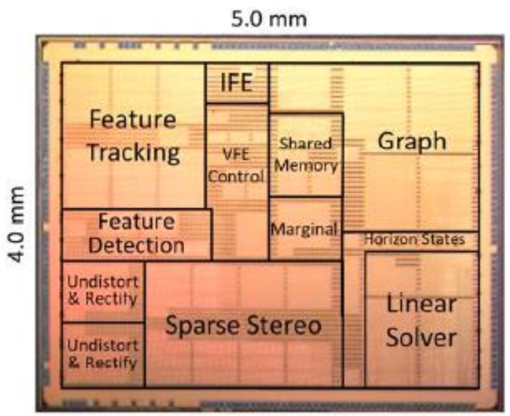
Why only Digital Accelerators? Analog can be Used To!



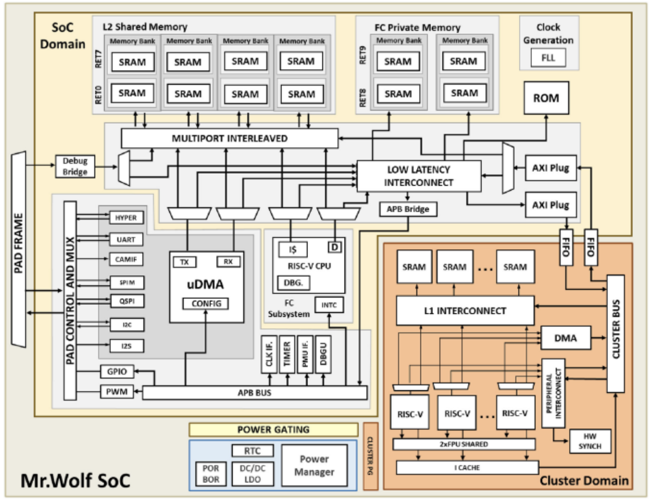
J. Klein with IBM: “““ALPINE: Analog In-Memory Acceleration with Tight Processor Integration for Deep Learning”, IEEE TC, 2022:
40% more energy efficient for complex NNs! Challenge: system-level interface!

EPFL

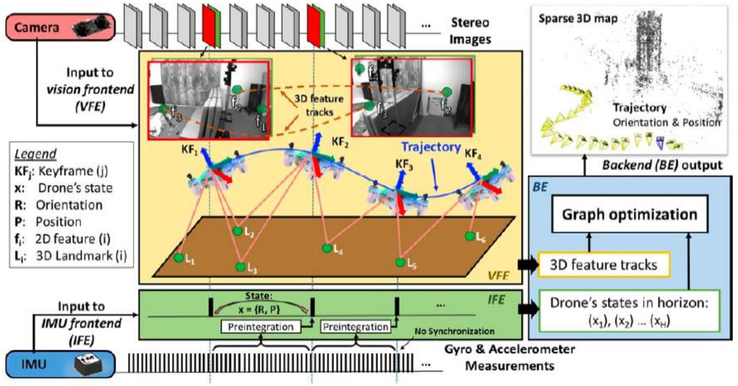
Applic.-Specific Medical Edge AI Systems: Digital + Analog Co-Design



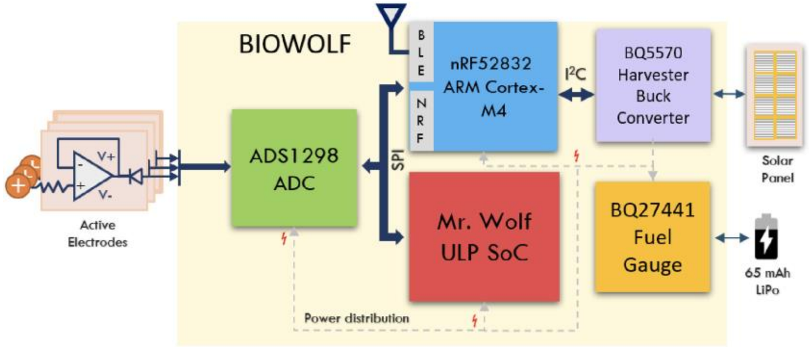
Navion: Visual-Inertial Odometry (VIO) Accelerator
24mW at 65nm

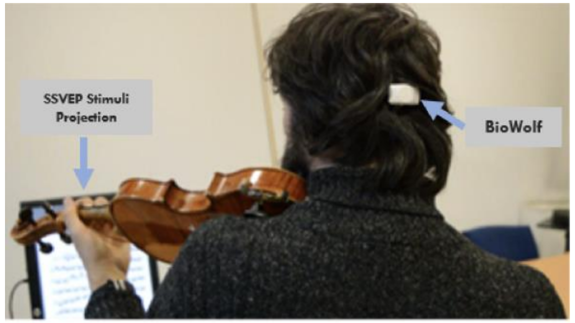


BioWolf: Brain-Computer Interface Platform
6.3mW providing 38hrs with 65mAh battery



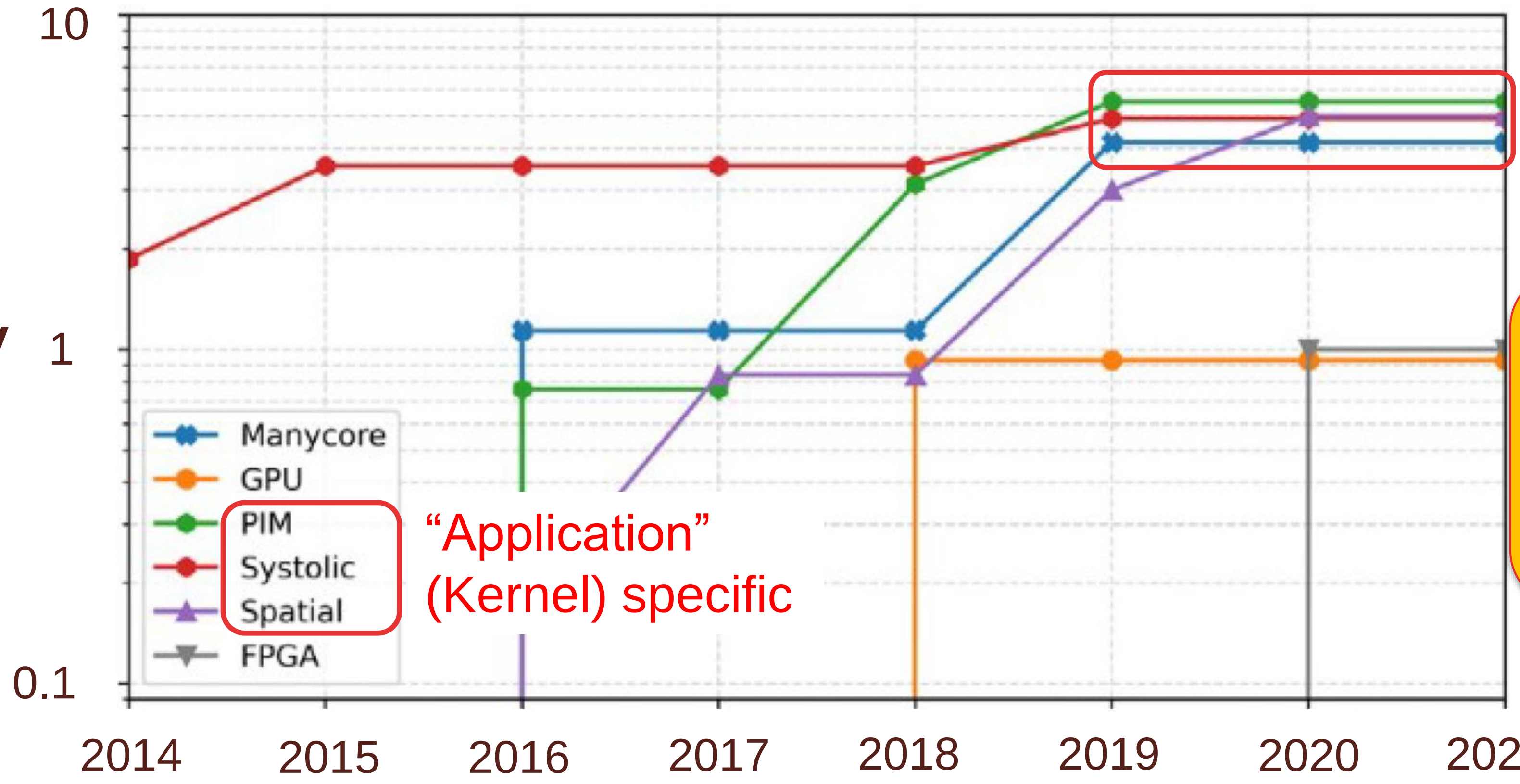
Amr Suleiman et al. JSSC'19





Victor Kartsch et al. TBioCAS'19

But App.-Specific Edge AI Reaching a Limit…



Efficiency Plateau reached!

Need for heterogeneous edge AI arch. in medical domain!

Architectures for conventional ML: DNN, CNN, RNN

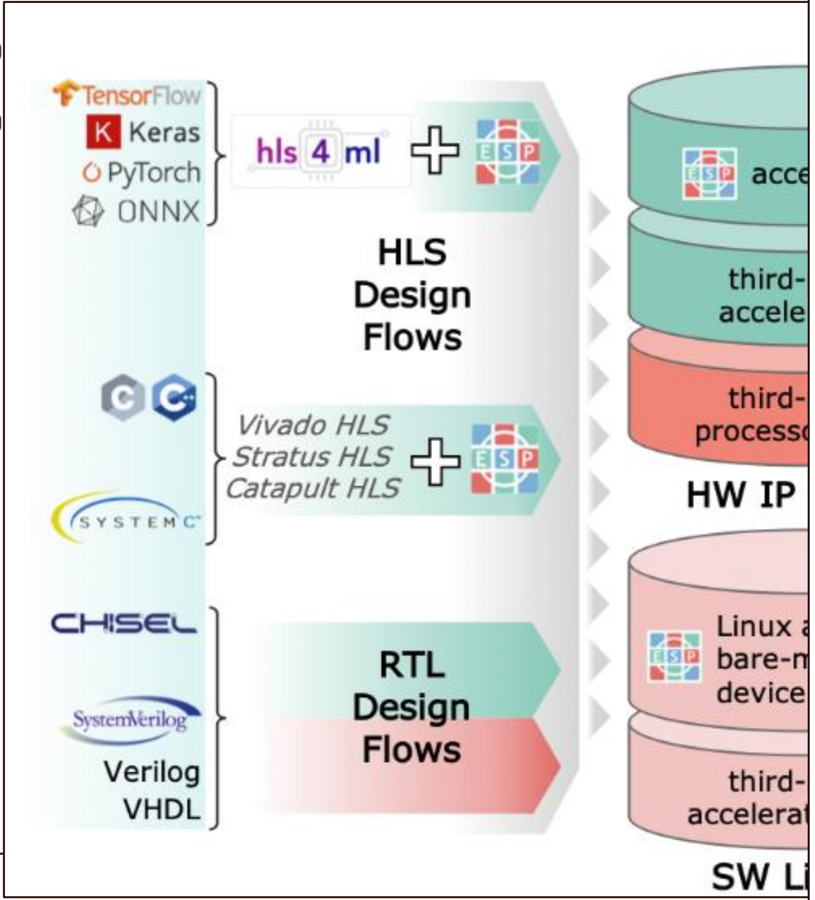
<u>**New trend:**</u> Simple core and different domain-specific accelerators together (with system codesign = need for **open and fast system exploration frameworks!**)

X-HEEP

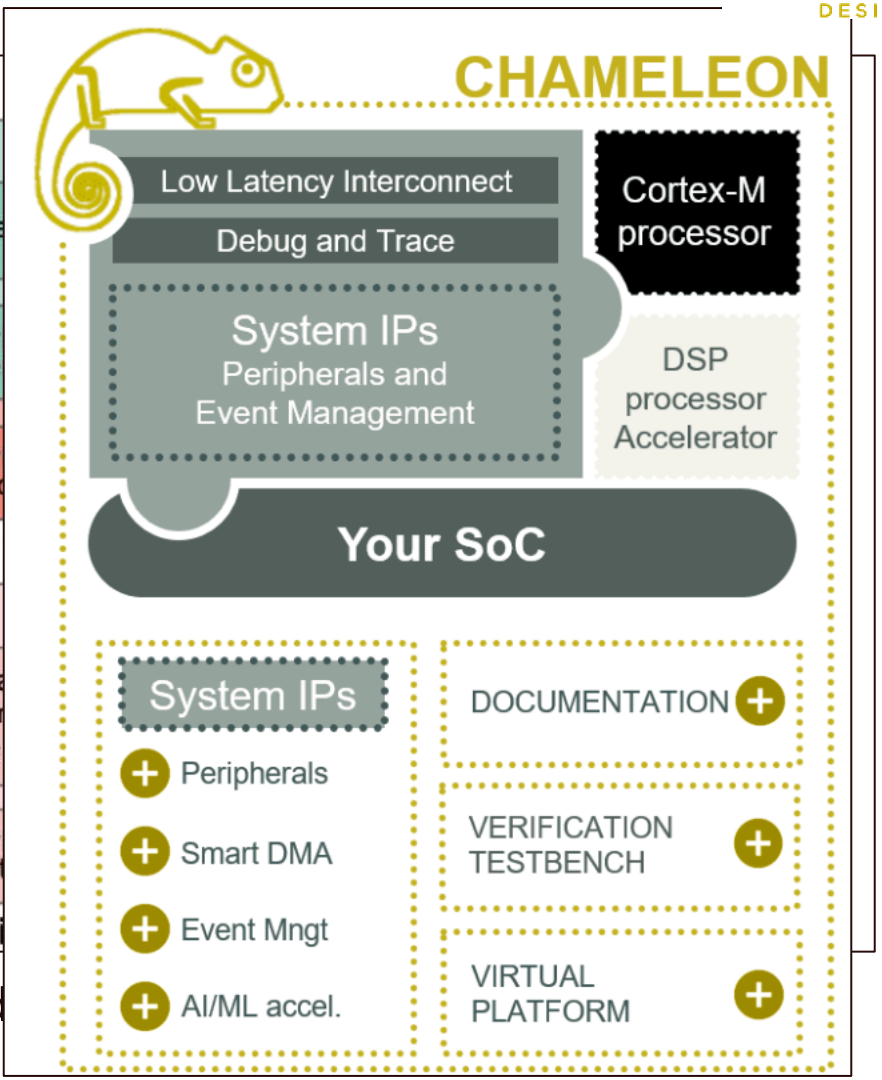
Configurability

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.ed



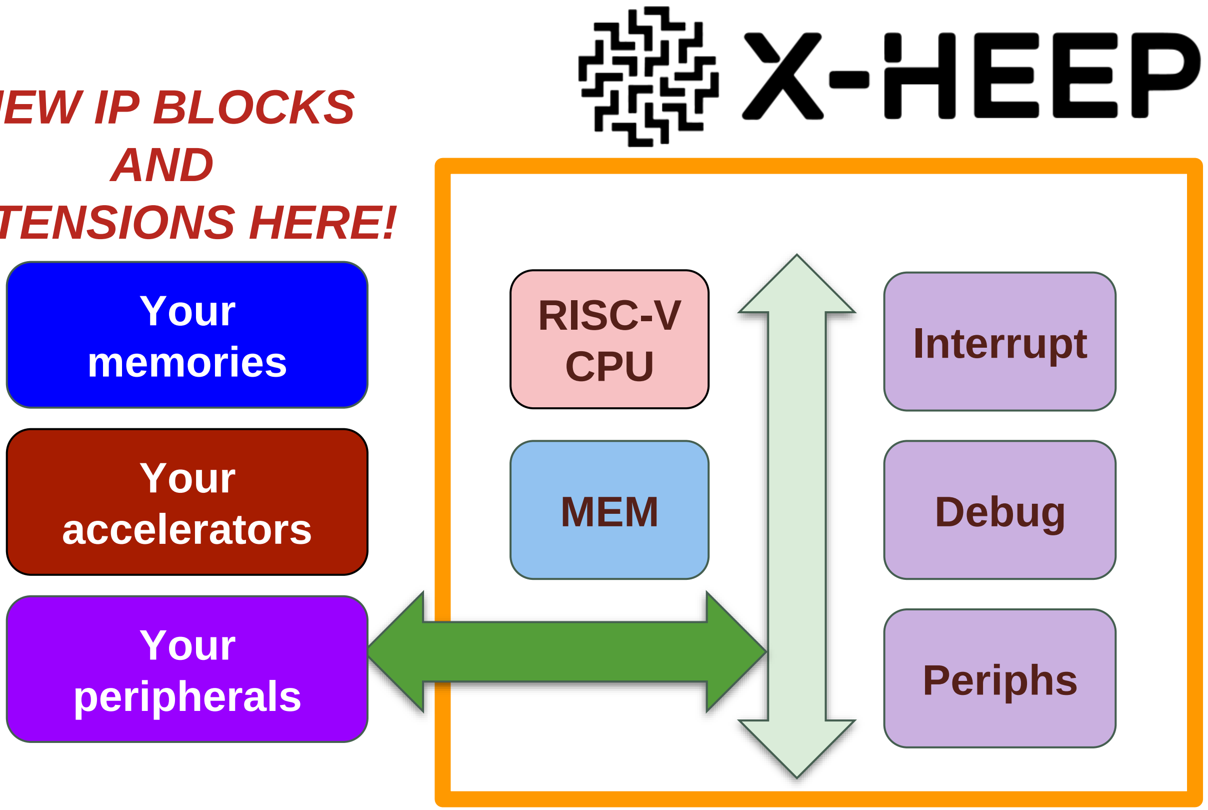
https://www.dolphin-design.fr/chameleon-mcu-subsystem/

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

X-HEEP PROVIDES THE BASIC BLOCKS, AND WE CAN MAKE THE RESEARCH AROUND IT

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

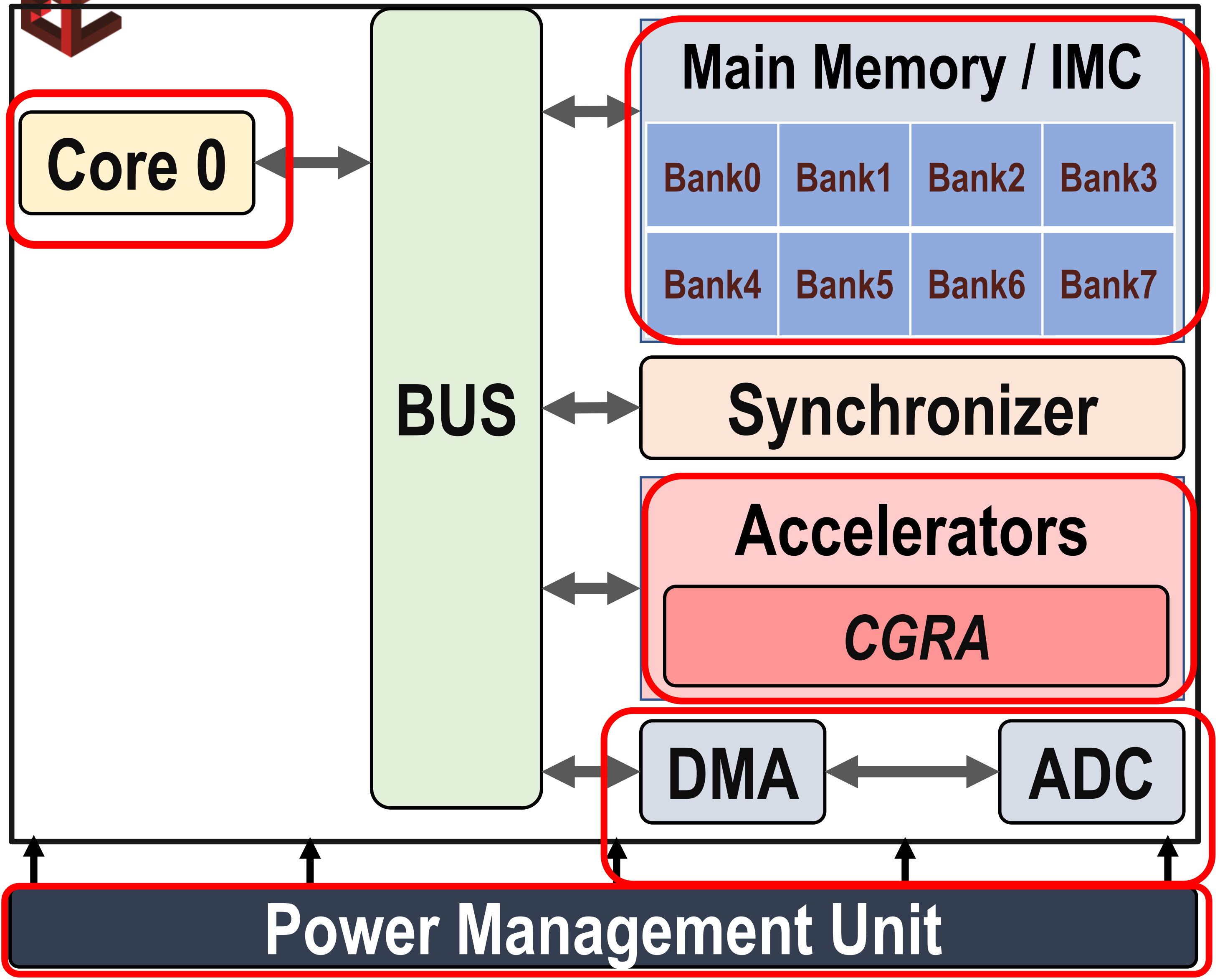
NEW IP BLOCKS AND EXTENSIONS HERE!



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).

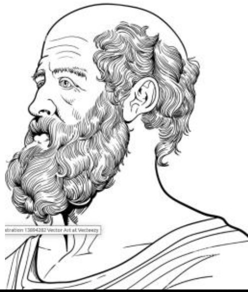
X-HEEP for Healthcare: HEEPocrates



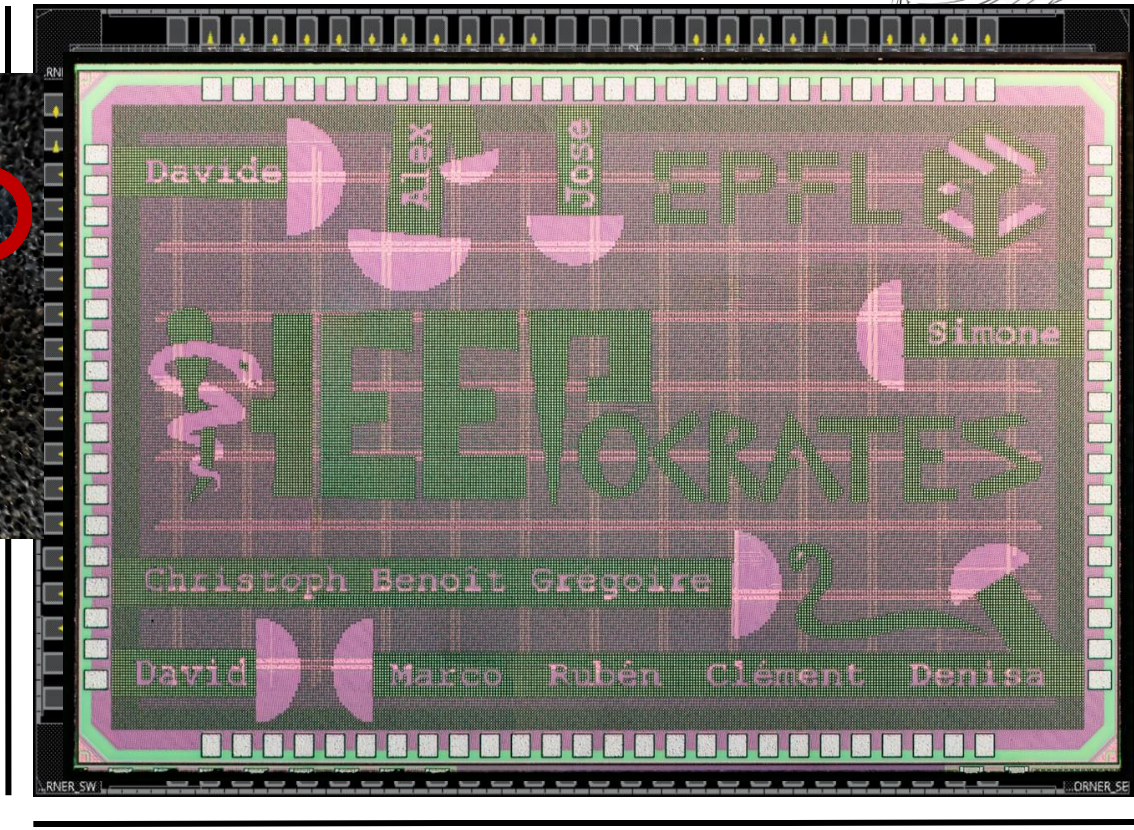
- Single-core architecture
  - Control of accelerators flow (parallel execution)
- Independent memory banks
  - Switch-off unnecessary banks
- Coarse-Grained reconfigurable accelerator (CGRA) and in-memory computing (IMC)
  - CGRA: compute-intense kernels (irregular flow)
  - IMC: Simple ML ops with 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, 256KB total
- ASIC implementation, 65nm TSMC
  - Area: 6mm^2
  - Frequency: 32KHz/ 470MHz
  - Power: **27.7mW@170MHz, 0.8V**
    48.1mW@470MHz, 1.2V
- Extensions enable ACCELERATORS:
  1. **Coarse-Grained Reconfig. Array (CGRA)**
  2. **In-memory (bit-line) computing**



3mm

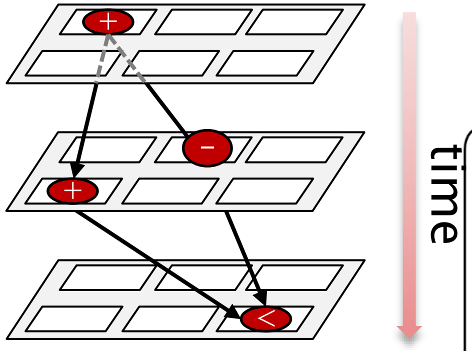
Complete design done in 5 months (6 people)

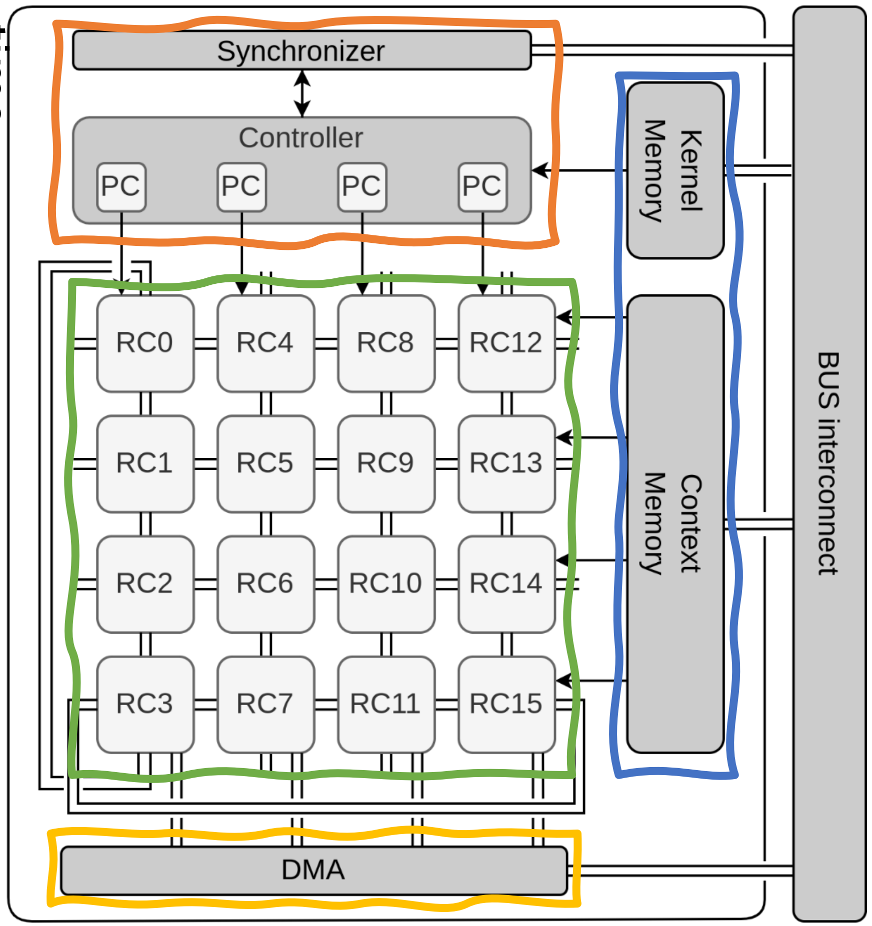
https://www.epfl.ch/labs/esl/research/2d-3d-system-on-chip/x-heep/

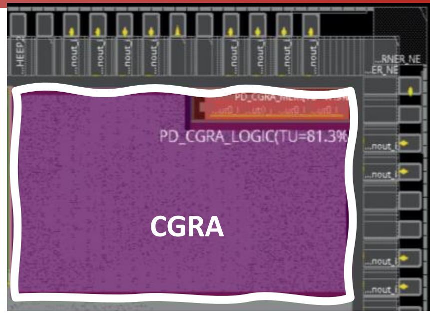
[1] OpenHW group github: https://github.com/openhwgroup

CGRA implementation

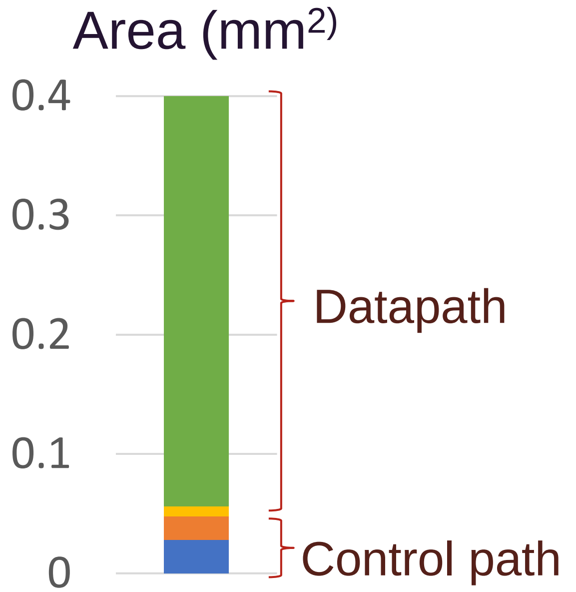
- 2D Mesh of ALUS
  - Spatio-temporal kernel mapping
  - 16 reconfig. cells
  - 4 indep. columns







1. Synchronizer and Controller
   - orchestrates execution
2. Datapath
   - ALUs and register files
3. DMA port per column
   - input/output to/from memory
4. Context/Kernel memory (2KB)
   - stores CGRA configurations



- Datapath occupies >80% of area

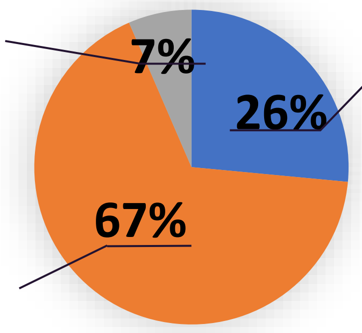
Loris Duch, Soumya Basu, Rubén Braojos, Giovanni Ansaloni, Laura Pozzi, David Atienza, "HEAL-WEAR: An Ultra-Low Power Heterogeneous System for Bio-Signal Analysis" IEEE Transactions on Circuits and Systems: Part I (TCAS-I), September 2017.

BLADE: Bit-line array IMC architecture

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.

→ Shift, add, negation implement MAC
- 2.2GHz
- 64KB SRAM
- 28nm TSMC

Area: 1240 µm²



Energy

16-bits word		
Read	Write	IMC
376pJ	**414pJ**	**381pJ**

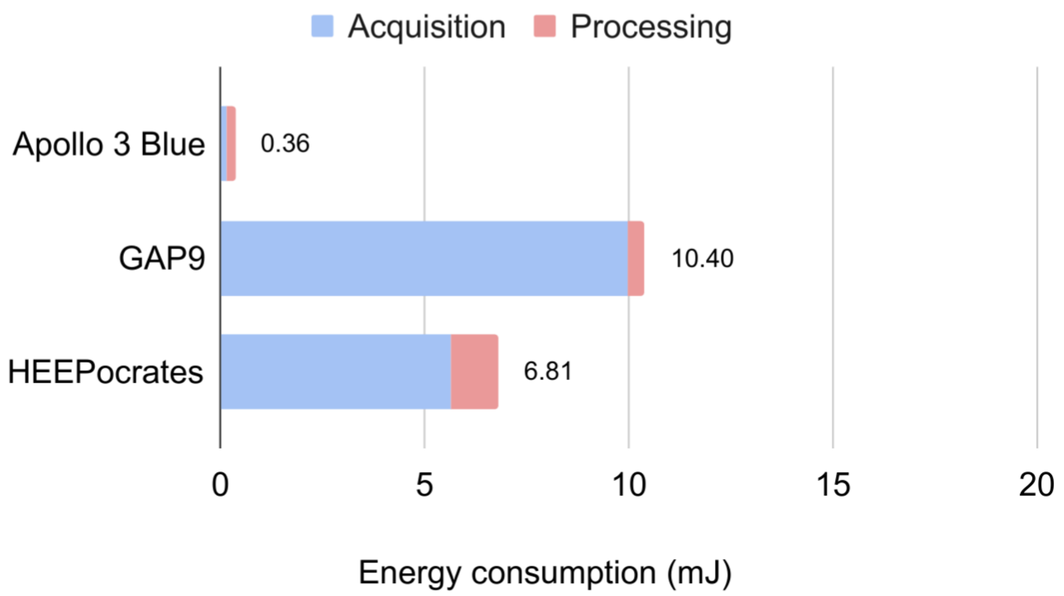
IMC operations cost slightly more than memory read operations

But 3x performance gain for convolutional layers!

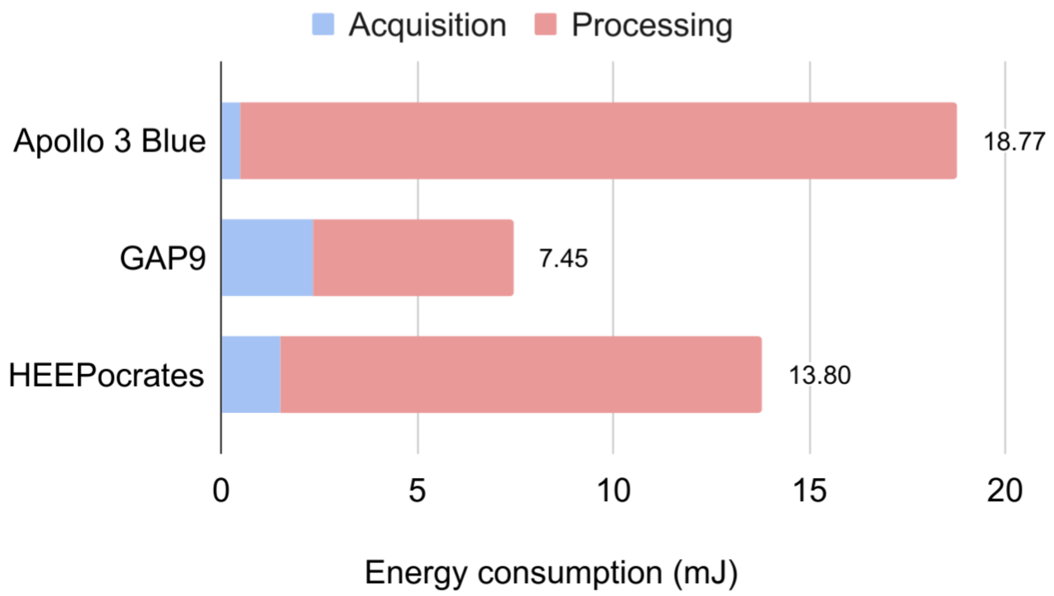
Rios, Marco, et al. "Bit-Line Computing for CNN Accelerators Co-Design in Edge AI Inference." *IEEE Transactions on Emerging Topics in Computing* (2023).

- Energy consumption: competitive vs. systems in newer tech.

ECG Heartbeat Classifier



EEG Seizure Detection CNN



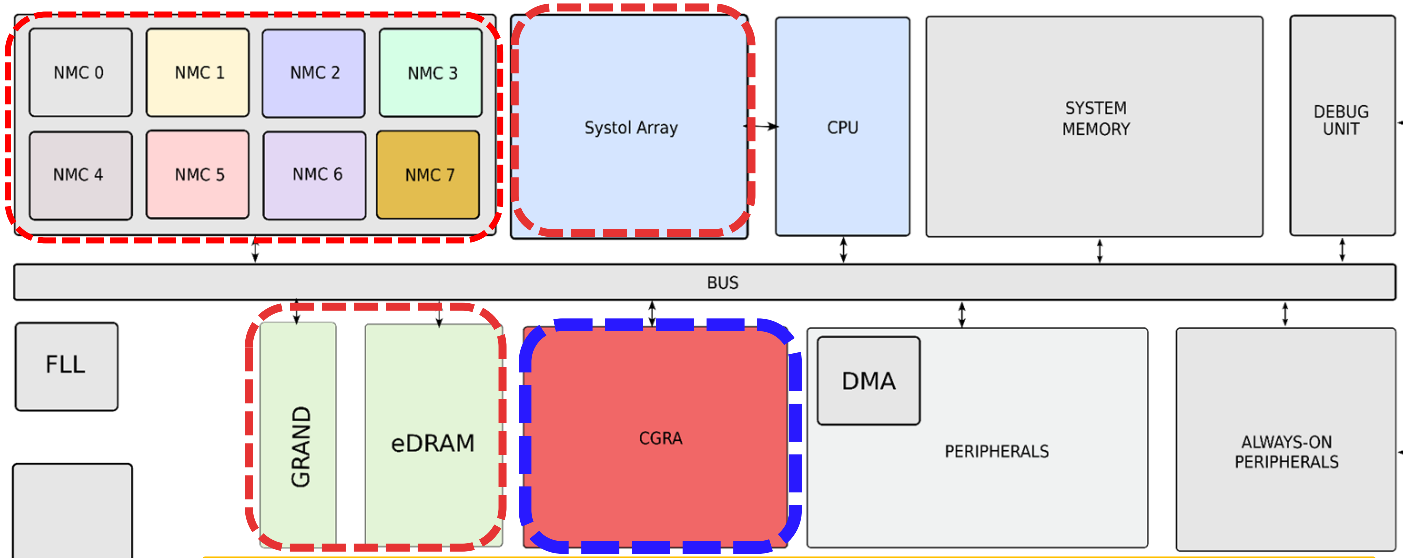
Competitive and flexible Open-Source Edge AI systems for medical domain! So, what's next? Use in medical applications!

But a new iteration of HW-SW co-design with learned lessons: evolution step in our neuro-inspired medical edge AI systems!

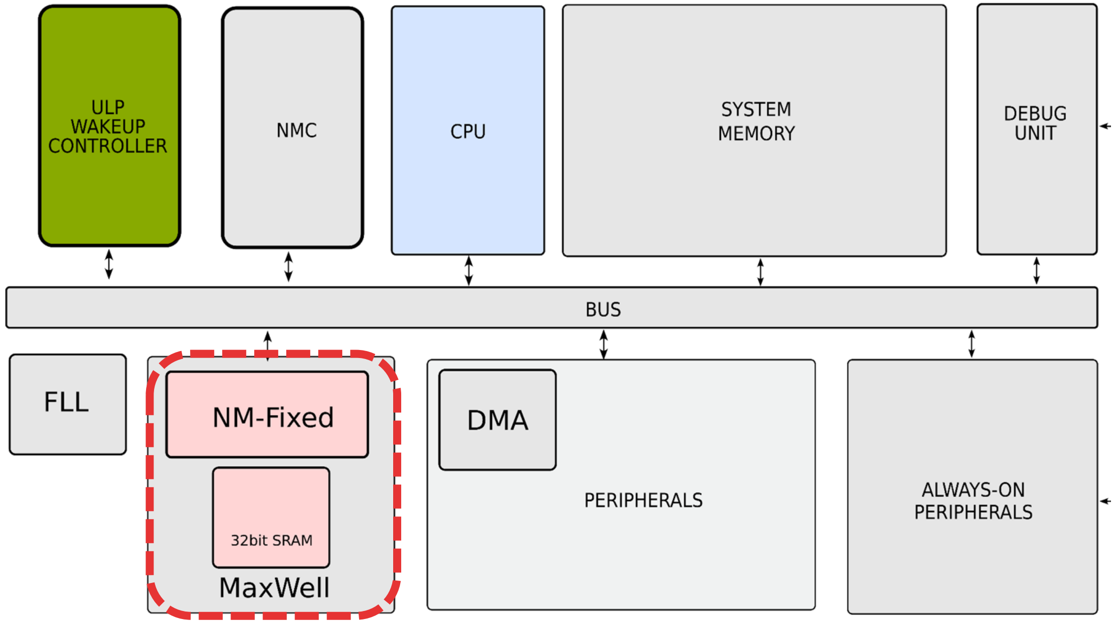
- Q4, 2024

IMC has become near-mem. comput. (NMC)

Heepatia (16nm)



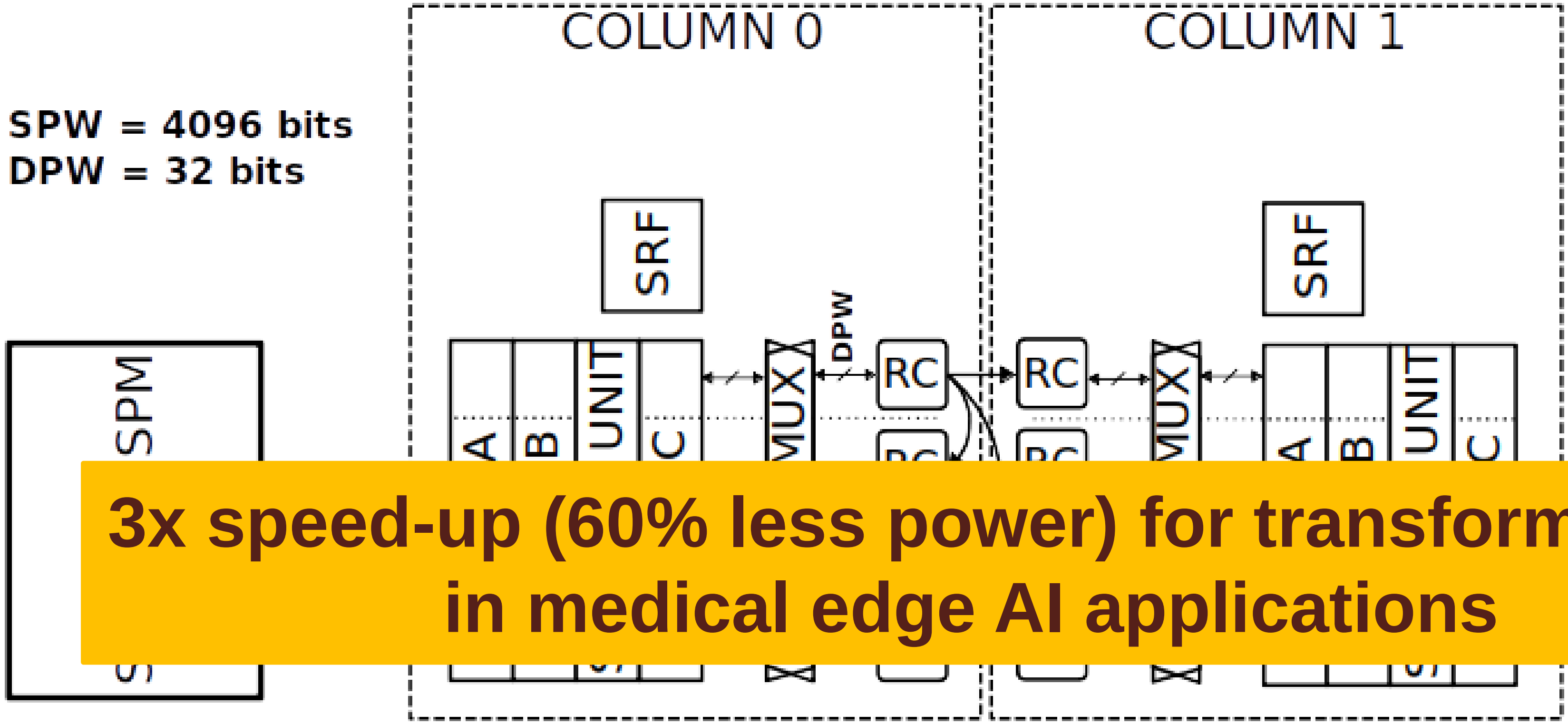
Heepnosis (22nm)



Our CGRA accel. has "evolved"

Choosing between NMC vs. IMC computing is a key research topic for different edge AI domains!

- Wider and more efficient memory hierarchy
  - Load Store Unit (LSU)
  - Loop Control Unit (LCU)
  - MultipleXer-Control Unit (MXCU)

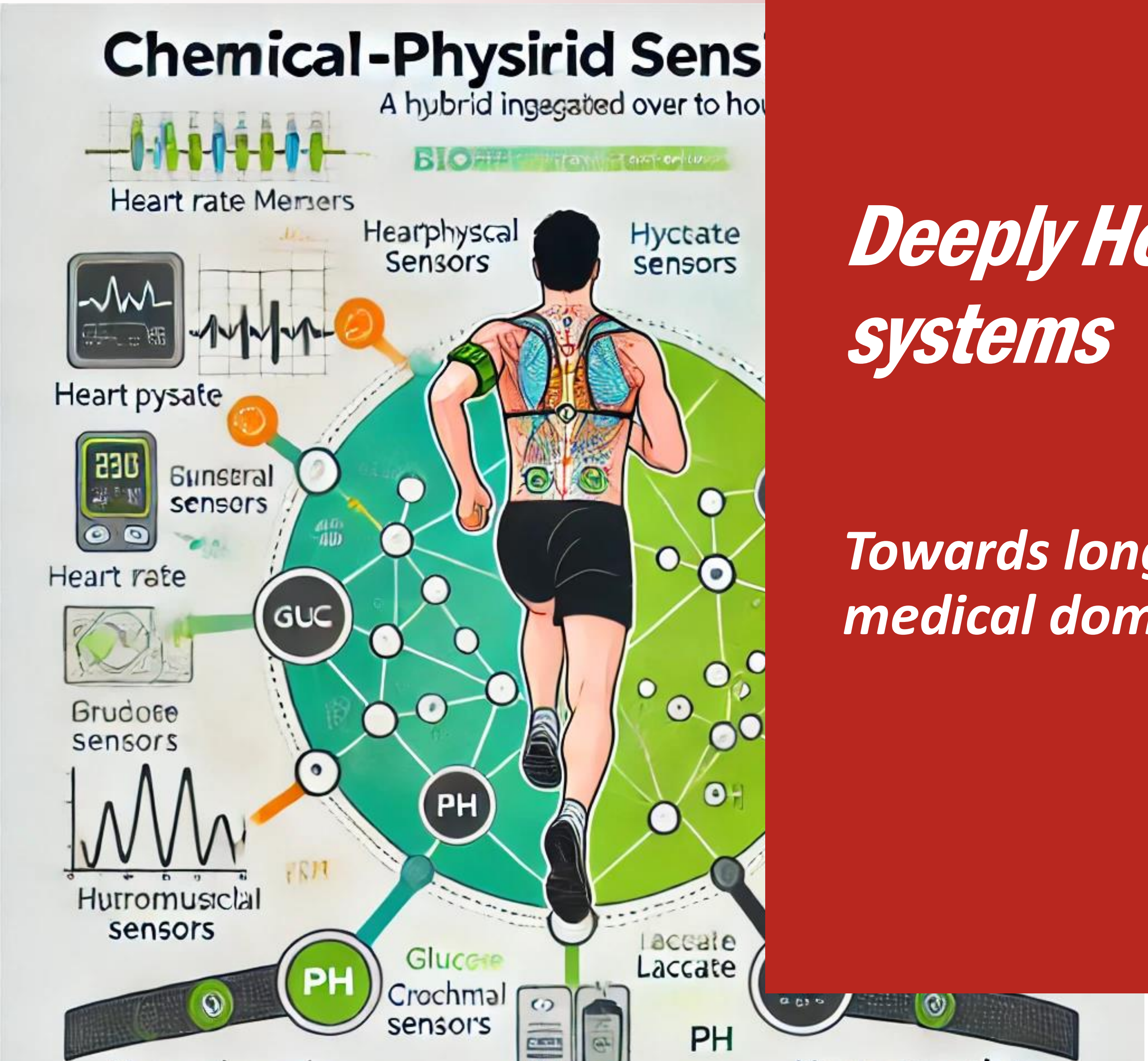


3x speed-up (60% less power) for transformers use in medical edge AI applications

Main features:

- **4x2 array of RCs with torus connection**
- **RCs synchronized per column (common PC)**
- **3 VWRs per column**
- **1 Scalar Register File (SRF) per column**
- **cialized slots LCU, MXCU)**
- **Shared scratchpad memory (SPM)**

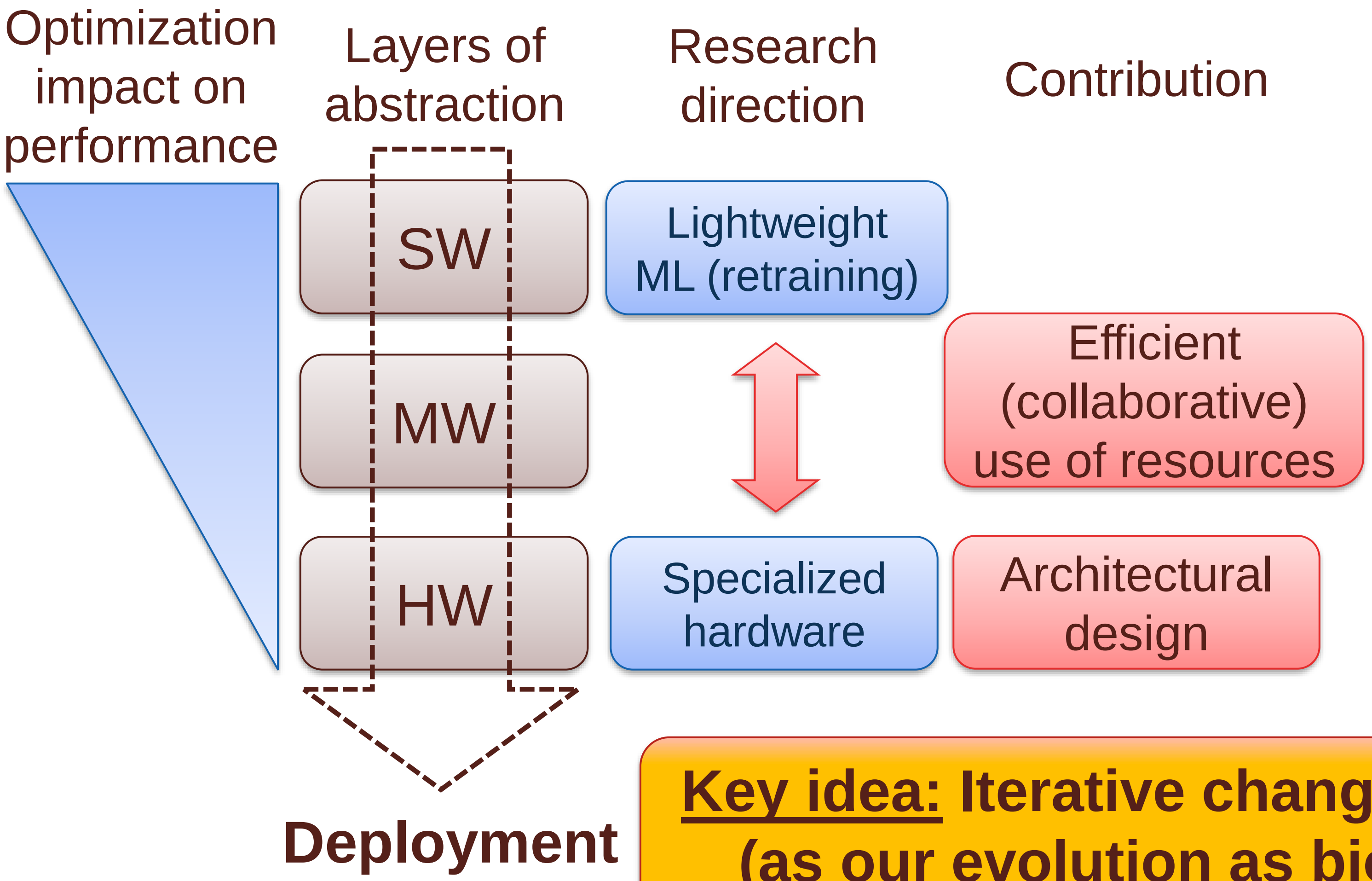
Benoıt. Denkinger et al., DAC 2022 and TC 2023



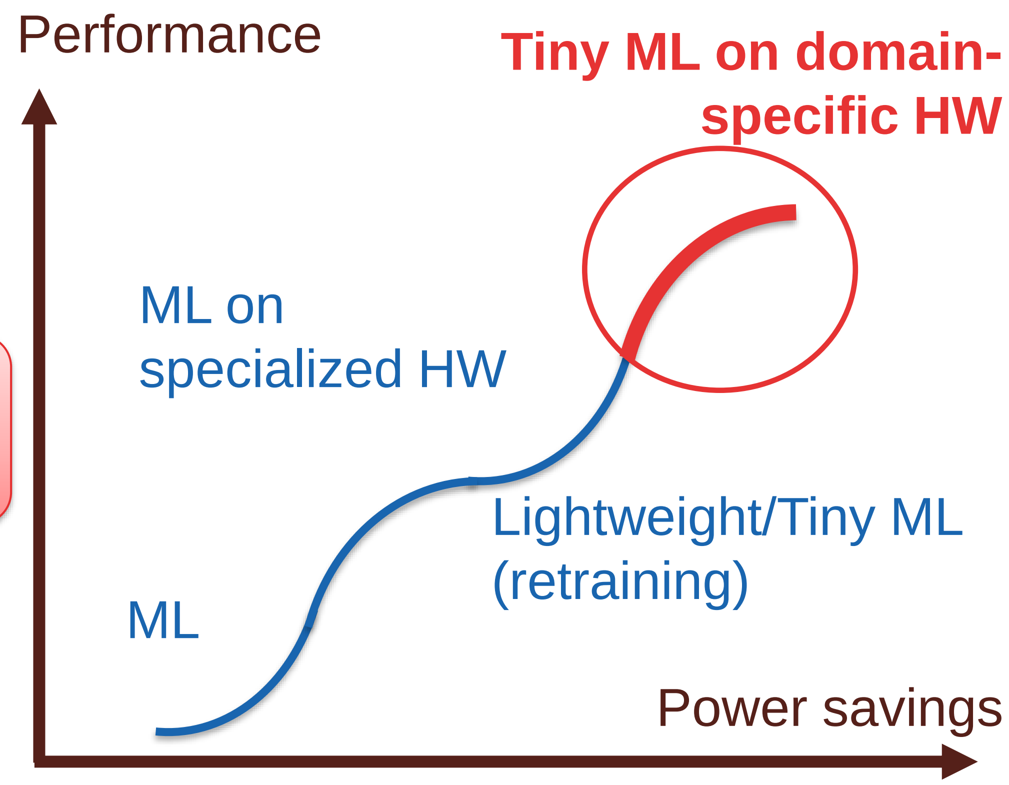
Deeply Heterogeneous sensing systems

Towards long term monitoring for the medical domain cyber-physical systems

ML Deployment

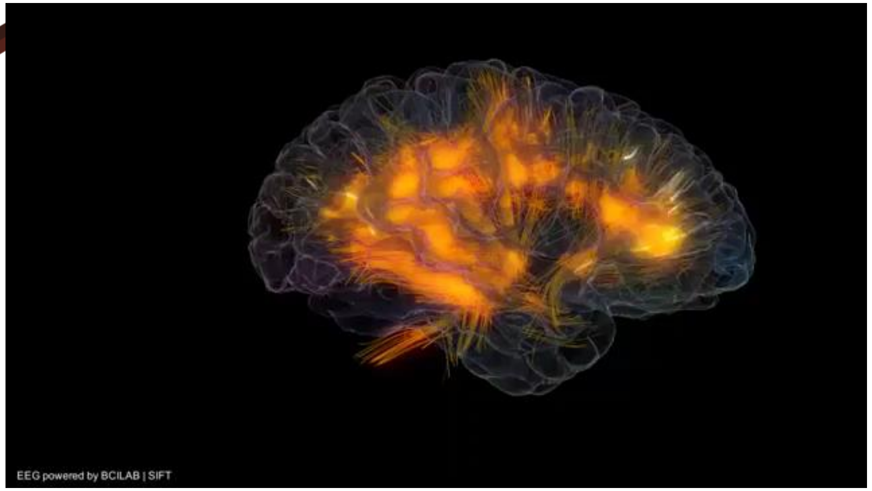


Domain-Specific Exploration

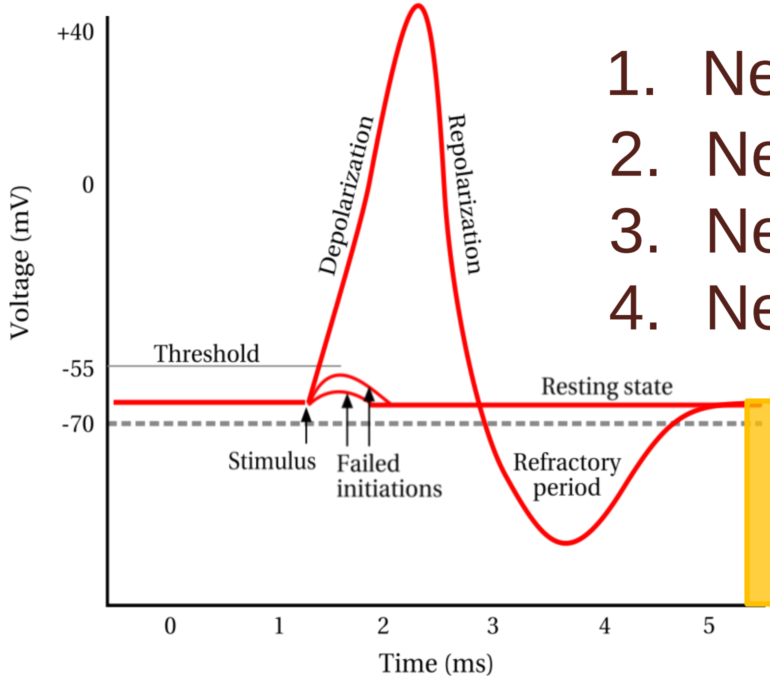


<u>Key idea:</u> Iterative changes of HW & SW edge AI system (as our evolution as biological systems took place!)

Neuro-Inspired: How Does our Brain "Work"?



- Brain "embedded" computing features:
  - Size: 4-100 μm neurons, 1.3-1.5 dm^3
  - Approx. 80B Neurons, 100 Trillion Synapses
  - **20W average** (>10,000 TFLOPS)



1. Neurons are idle most of the time (**no power** consumed)
2. Neurons react only to stimulations (**small part active**)
3. Neurons **integrate** storage with processing
4. Neurons are **configurable**

Design of medical edge AI systems based on its unique domain-specific properties and multiple accelerators

Exploiting Medical Knowledge in Edga AI Design: ECG Spectrogram

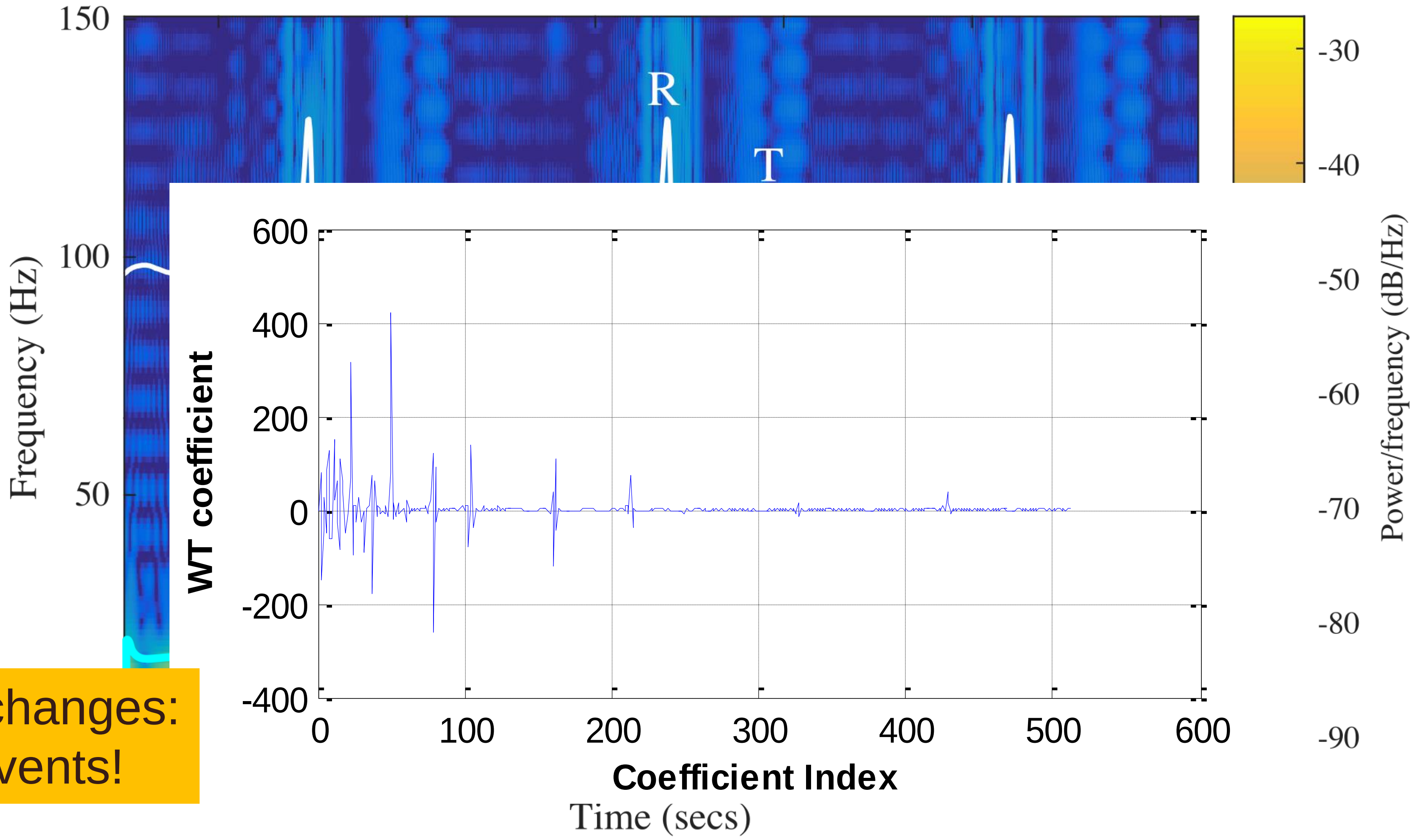
ECG temporal properties:

- High frequencies
- Low frequencies
- Changing in time

Different frequencies localized in time

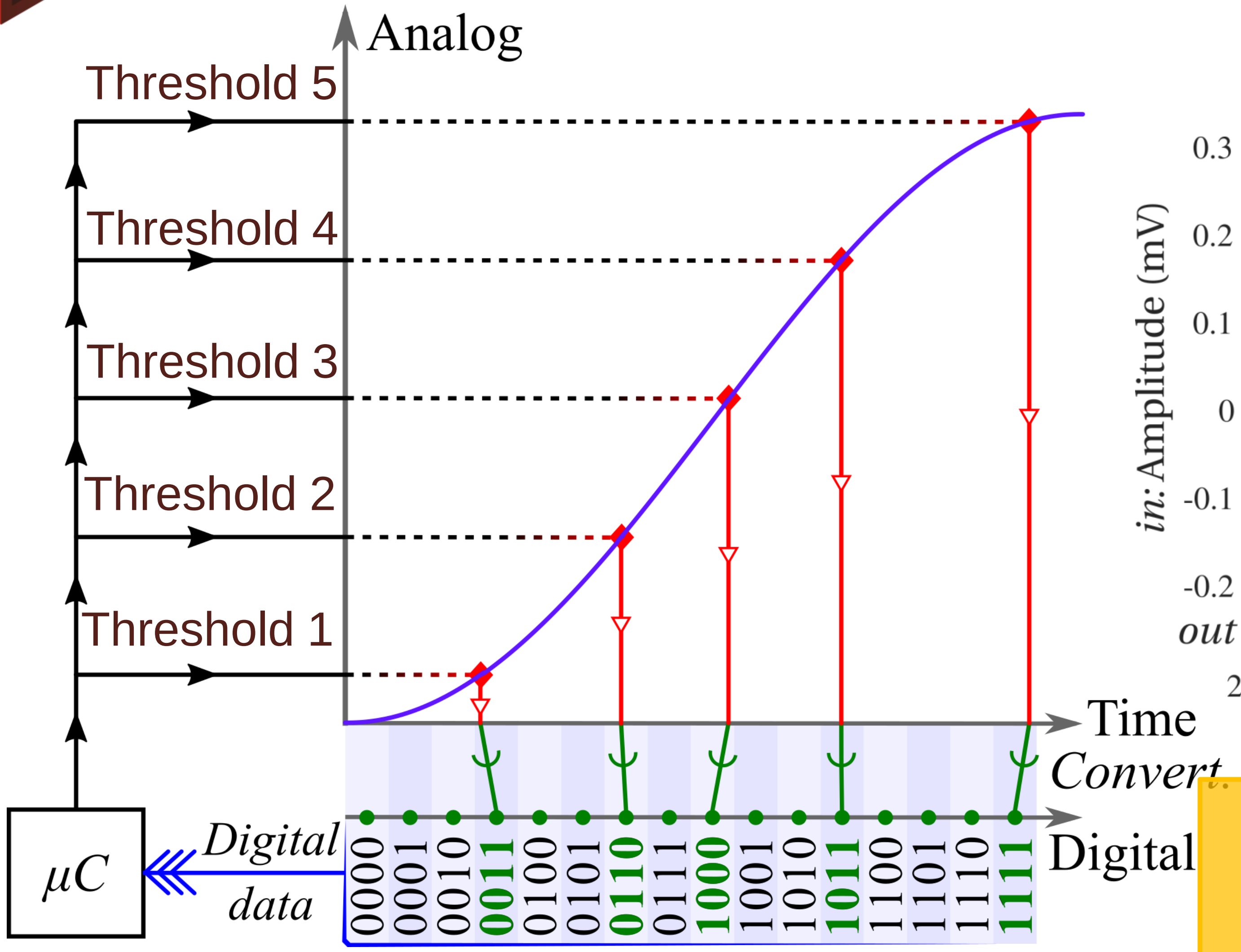
Uniform sampling is sub-optimal

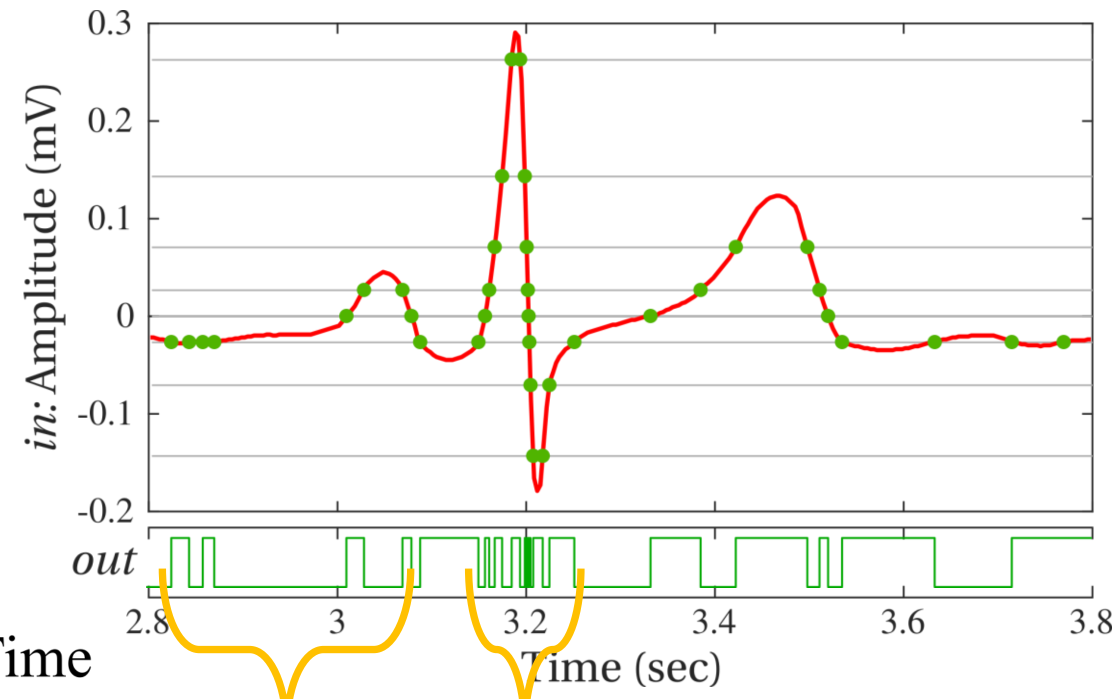
And if representation changes: sparse signal = Few events!



G. Surrel, et al., "Online Obstructive Sleep Apnea Detection on Wearable Sensors", IEEE Transactions on Biomedical Circuits and Systems (TBioCAS), May 2018.

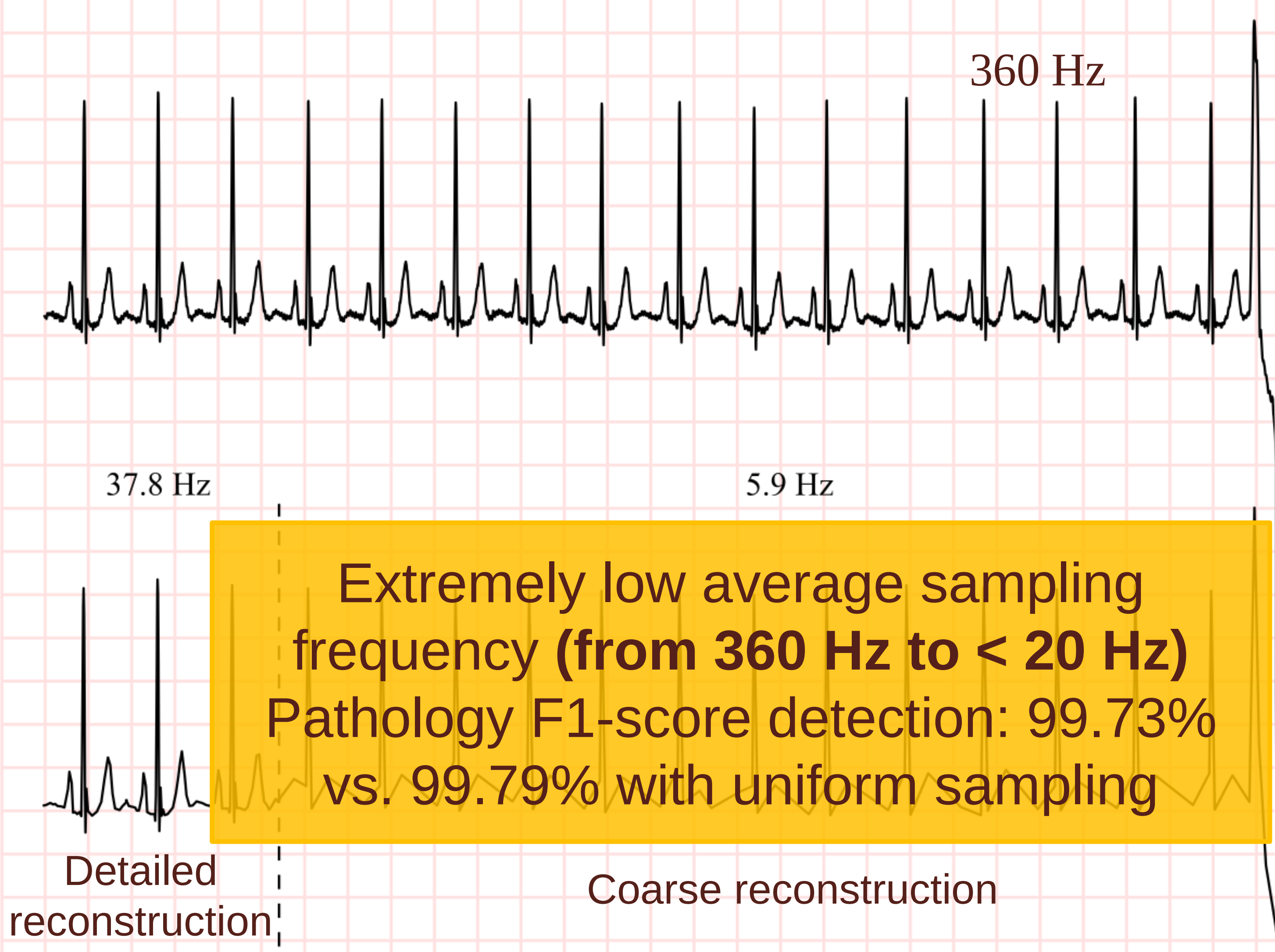
Event-Based Sampling in Medical Edge AI Systems





Useful sampling driven by **biosignal's properties + pathology: dynamically tuned**

G. Surrel *et al.*, "Event-Triggered Sensing for High-Quality and Low-Power Cardiovascular Monitoring Systems", IEEE D&T 2019



Extremely low average sampling frequency **(from 360 Hz to < 20 Hz)**
Pathology F1-score detection: 99.73% vs. 99.79% with uniform sampling

S. Zanoli, Flavio Ponzina, Tomas Teijeiro, Alexandre Levisse, David Atienza, "An Error-Based Approximation Sensing Circuit for Event-Triggered Low-Power Wearable Sensors", IEEE JETCAS, June 2023.

Extra challenges than our brain in edge AI: Epilepsy monitoring

1. Sparse events (few / month): Accurate monitoring but **long-term**
2. Real-time and personalized: Not only inference, but **training too!**
3. **User experience**

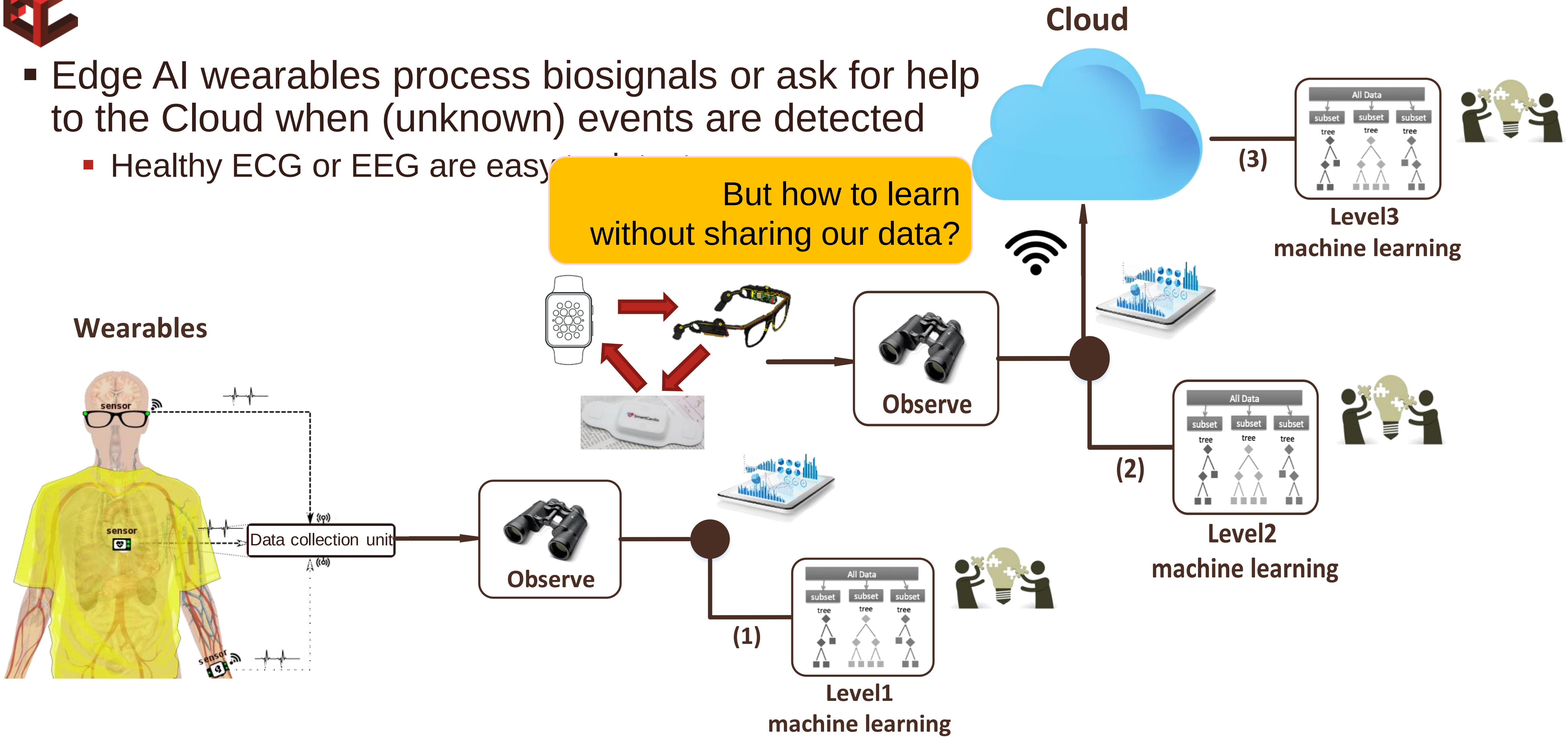
Social stigma: Patients refuse to wear EEG caps

Need for high sensing accuracy with **suboptimal positions of edge AI systems**!

Multi-Level AI for Healthcare: Edge AI + Cloud Big Data

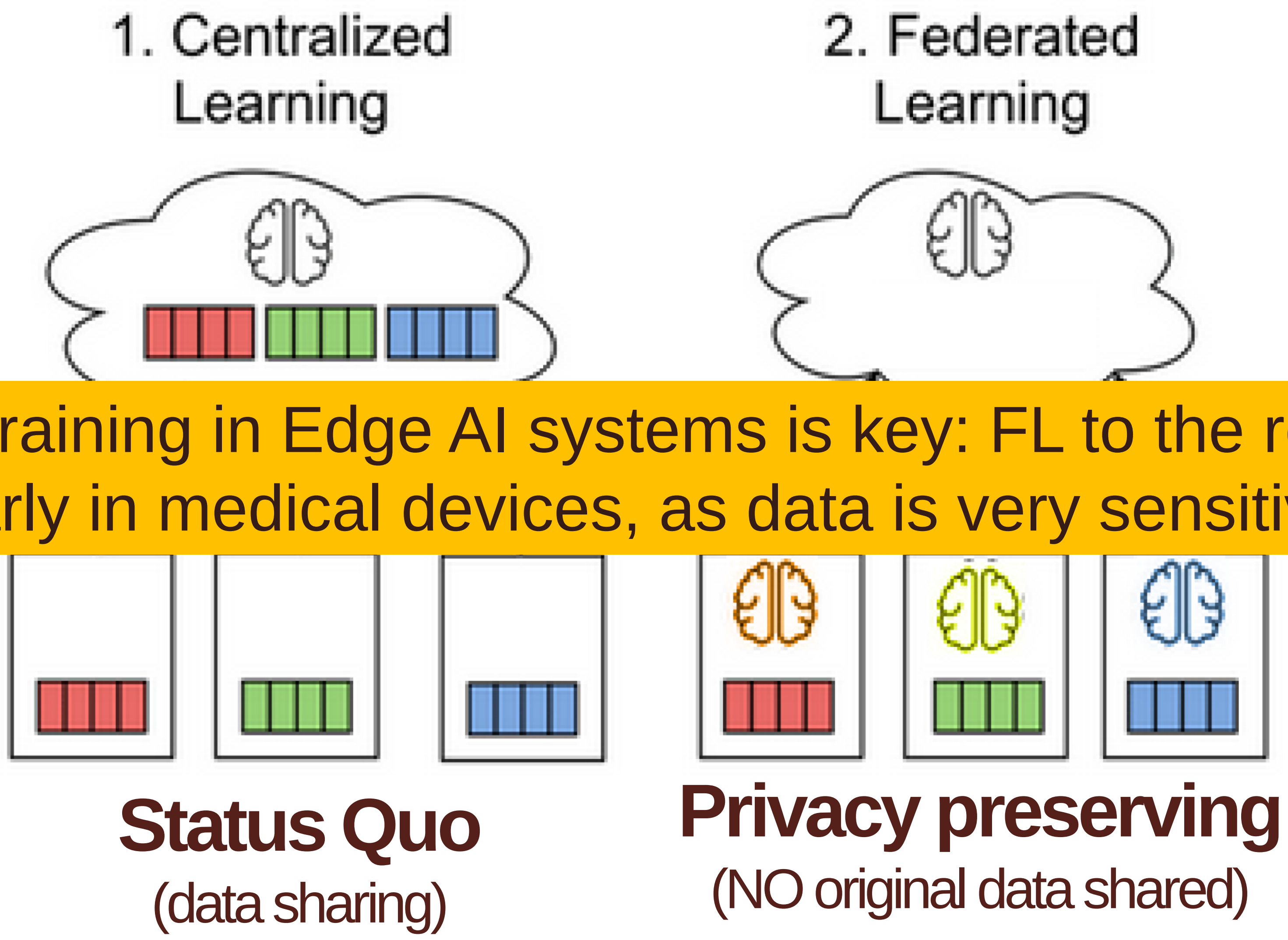
- Edge AI wearables process biosignals or ask for help to the Cloud when (unknown) events are detected
  - Healthy ECG or EEG are easy

But how to learn without sharing our data?



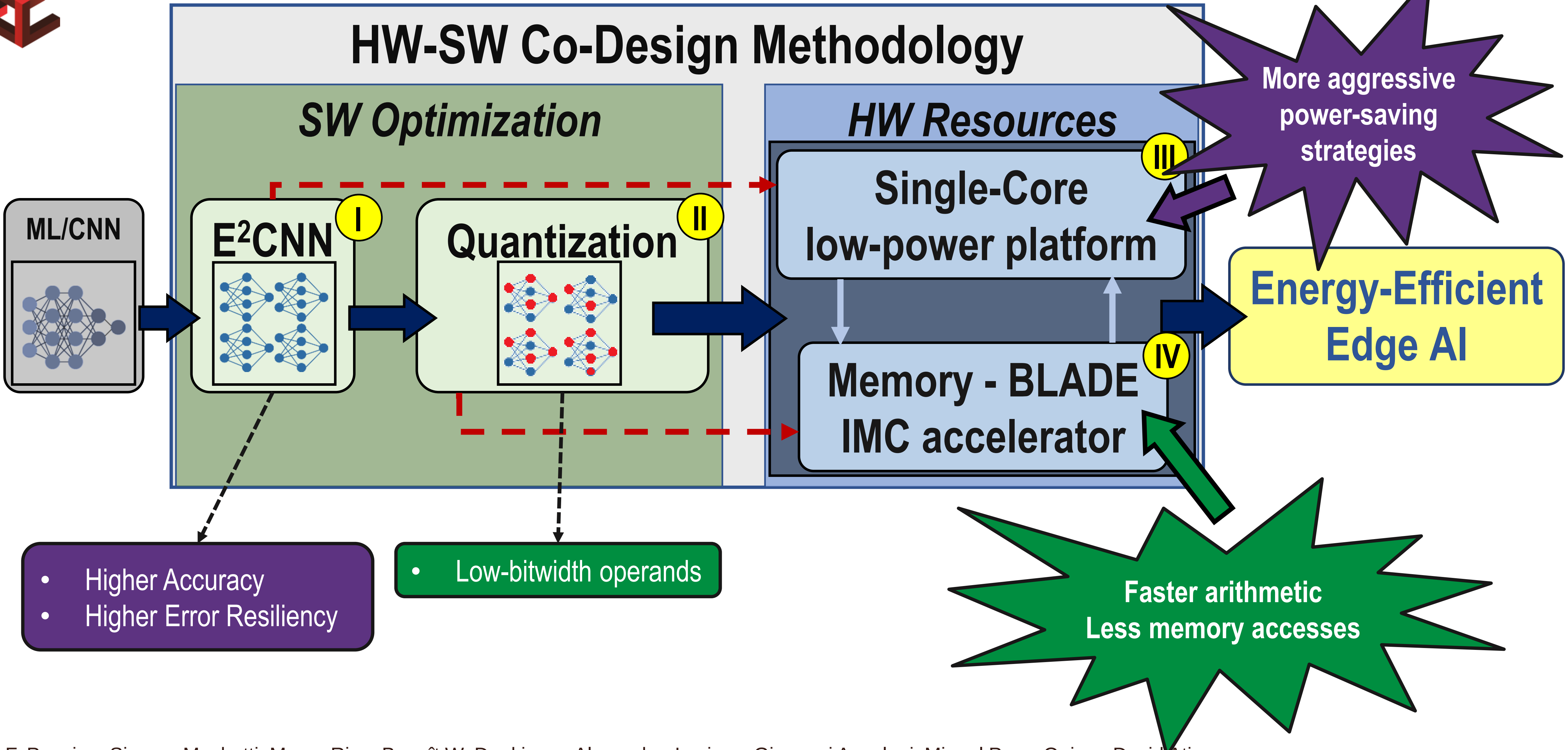
F. Forooghifar, Amir Aminifar, David Atienza, "Resource-Aware Distributed Epilepsy Monitoring Using Self-Awareness from Edge to Cloud", IEEE Transactions on Biomedical Circuits and Systems (TBioCAS), December 2019

- Components
  - ML model:
  - Data:



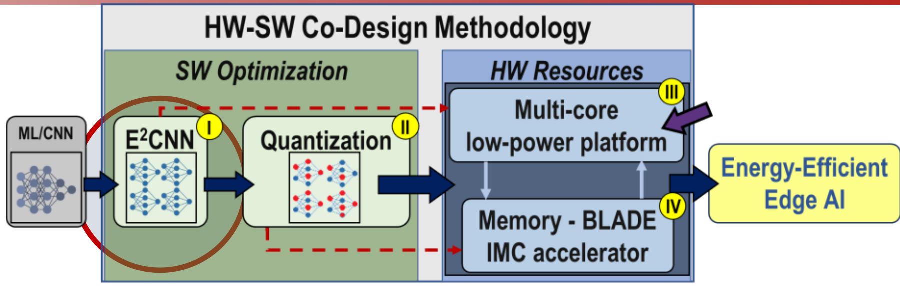
Adaptation/training in Edge AI systems is key: FL to the rescue, particularly in medical devices, as data is very sensitive!

System Co-Design Methodology for Medical Edge AI

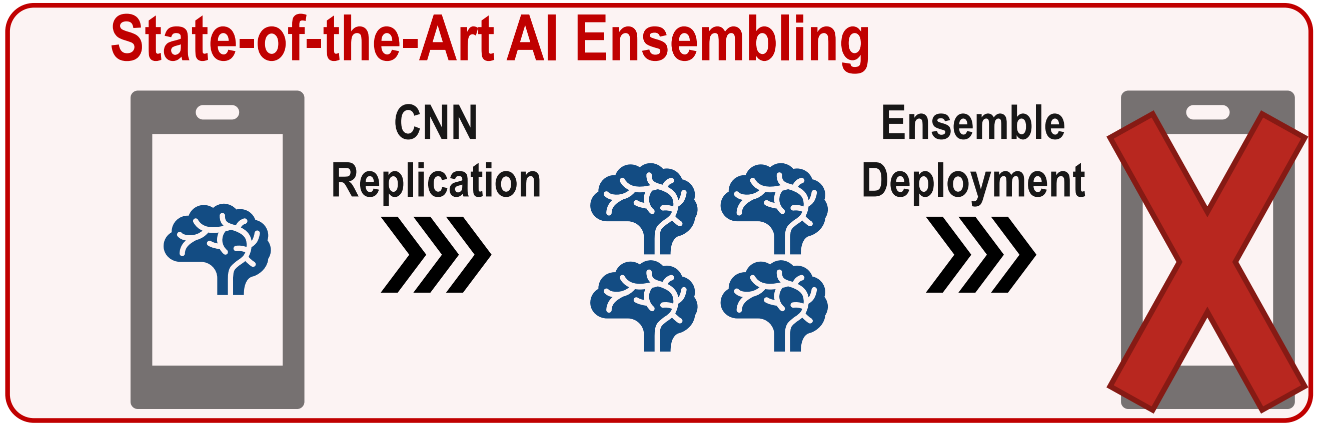


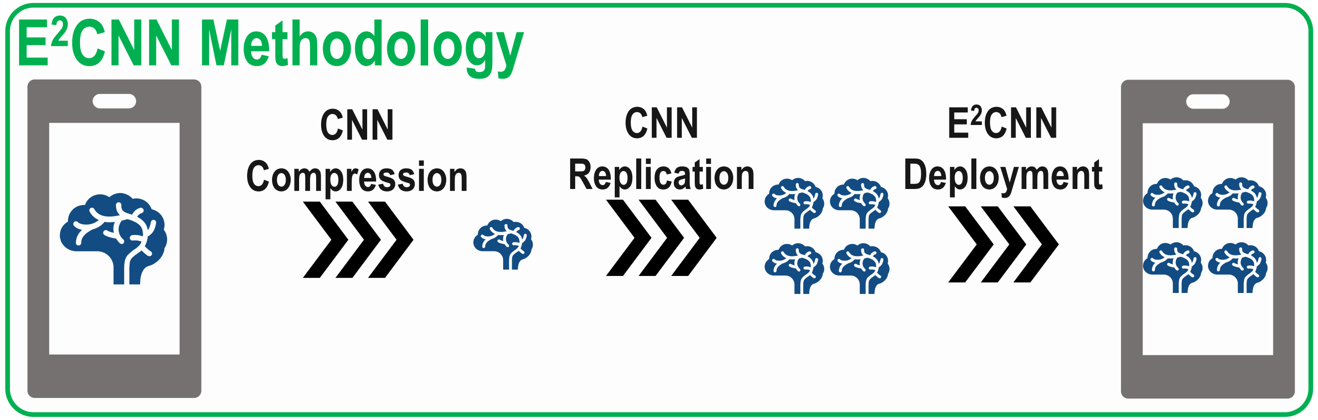
F. Ponzina, Simone Machetti, Marco Rios, Benoît W. Denkinger, Alexandre. Levisse, Giovanni Ansaloni, Miguel Peon-Quiros, David Atienza, "A hardware/software co-design vision for deep learning at the edge", IEEE Micro Magazine, November 2022

E^{2}CNN: Embedded Ensembles of ML/CNNs



- Designed to be deployed in EdgeAI systems
  - Ensembles of AI can significantly decrease workload and memory requirements
- How to build E^{2}CNN
  - Be *N* the desired number of instances forming the ensemble (*N*=4 in the example)
  - Before training, compress the initial CNN via filter pruning by a factor *N*
  - Replicate the obtained structure *N* times
  - Train each CNN independently





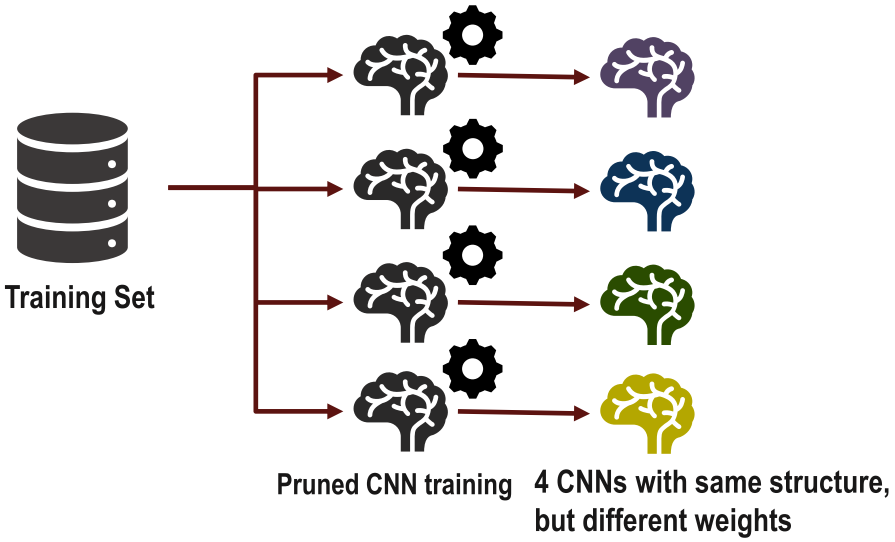
1. Reduced computation for edge AI
2. Low memory use for final AI/ML, benefit from multiple combined models!

Flavio Ponzina, Miguel Peon, Andreas Burg and David Atienza, "E2CNNs: Ensembles of Convolutional Neural Networks to Improve Robustness Against Memory Errors in Edge-Computing Devices", IEEE Transactions on Computers, August 2021

E^2CNN – Use in Medical Edge AI systems

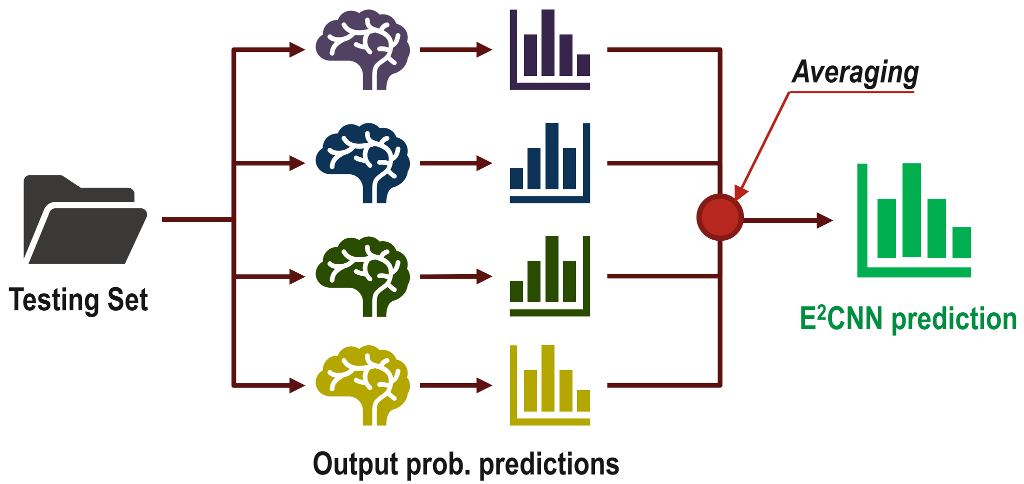
How to train the E^2CNN design

- Train each CNN structure independently on target dataset
- Use a (different) random weights initialization for each CNN structure



How to run inference with E^2CNN

- Feed each instance/wearable with the target input data to be classified
- Average the individual output predictions to get E^2CNN output



Quantization (Epilepsy Monitoring as Case Study)

a) **Uniform Quantization**

- Multipliers: 8 bits
- Multiplicands: 16 bits

CNN STRUCTURE		
Layer	**Weights**	**Activations**
Conv	8	16
Conv	8	16
FC	16	8

b) **Multipliers Optimization**

- Layer-based quantization
- Target: multipliers
- 2 ≤ N ≤ 8 quant. bits

CNN STRUCTURE		
Layer	**Weights**	**Activations**
Conv	4	16
Conv	5	16
FC	16	3

c) **Filter-level Optimization**

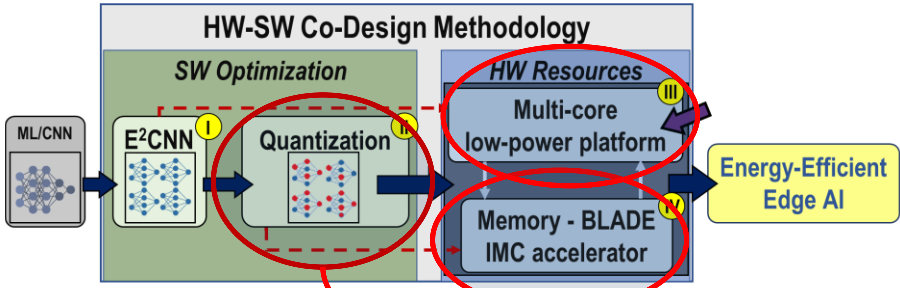
- Filter-based Optimization
- Redundant bits removal

CNN STRUCTURE		
Layer	**Weights**	**Activations**
Conv	Mixed	16
Conv	Mixed	16
FC	16	3

d) **Multiplicands Optimization**

- Layer-based quantization
- Target: multiplicands
- N=8 or N=16 quant. bits

CNN STRUCTURE		
Layer	**Weights**	**Activations**
Conv	Mixed	8
Conv	Mixed	16
FC	8	3

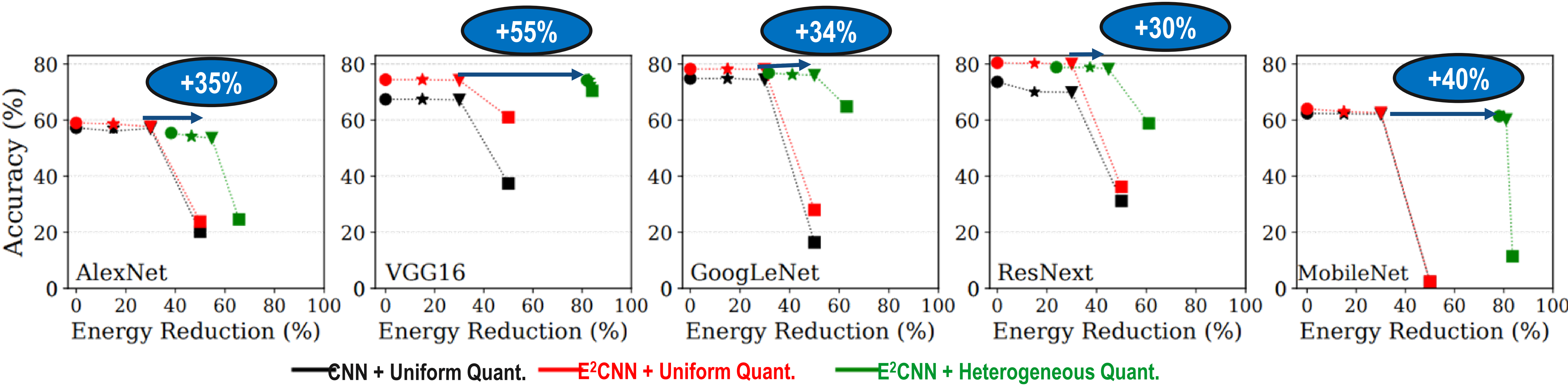


- **Heterogeneous** per-layer quantization
  - Applied on top of a uniformly quantized baseline (a)
- Accuracy-driven
  - Quantization in steps (b) and (d) constrained by a user-defined accuracy level (e.g., medical application)
- Custom bitwidths
- Filter-level optimization
  - Remove convolutional filters if containing only 0s weights
  - No impact on accuracy, but significant MACs reduction

Medical system co-optimization calls for exploration of different types of accelerators: **heterogeneous edge AI architectures!**

Experimental Results: Heterogeneous Quantization with Edge AI Co-Design

- Co-design enables competitive detection of epileptic seizures at minimal energy
  - **80% accuracy** on average with best E2CNN models of AI/ML instances
  - Energy savings up to ***55%*** at system level, ***without any relevant accuracy drop***



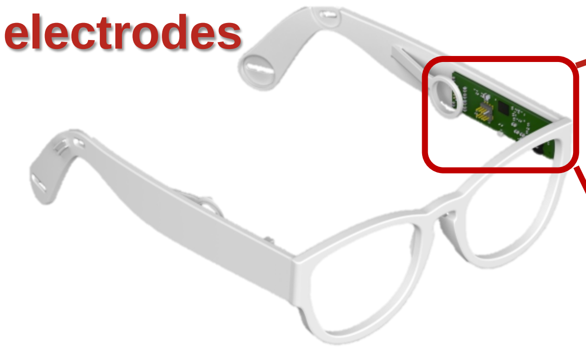
Markers correspond to different approx. layers used to improve energy efficiency

Some of the ESL gadgets

building an smarter edge much faster

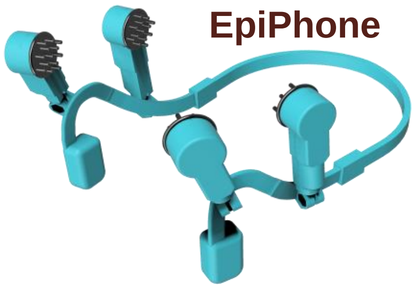
e-Glass[1]: A system for real-time seizure monitoring

Comparison: 24 vs 4 electrodes



e-Glass first prototype.

Glasses embodiment minimizes social stigma
(new: EpiPhone - 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 – Medical Edge AI (3rd Gen.):

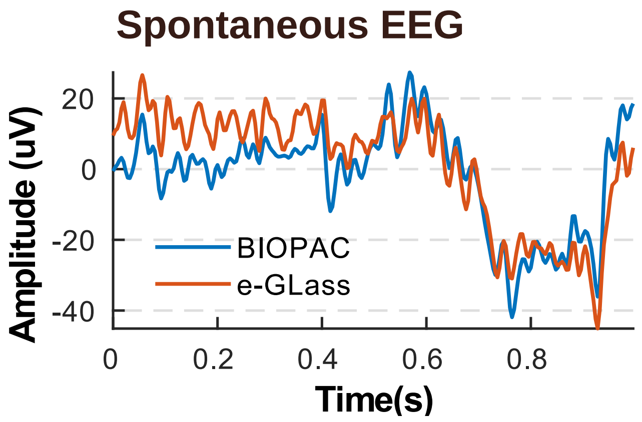
- HEEPocrates – Ultra-low power edge AI
- Onboard memory: 64 MB (up to 7 days of recording of EEG signals)

Battery powered: up to 96h monitoring

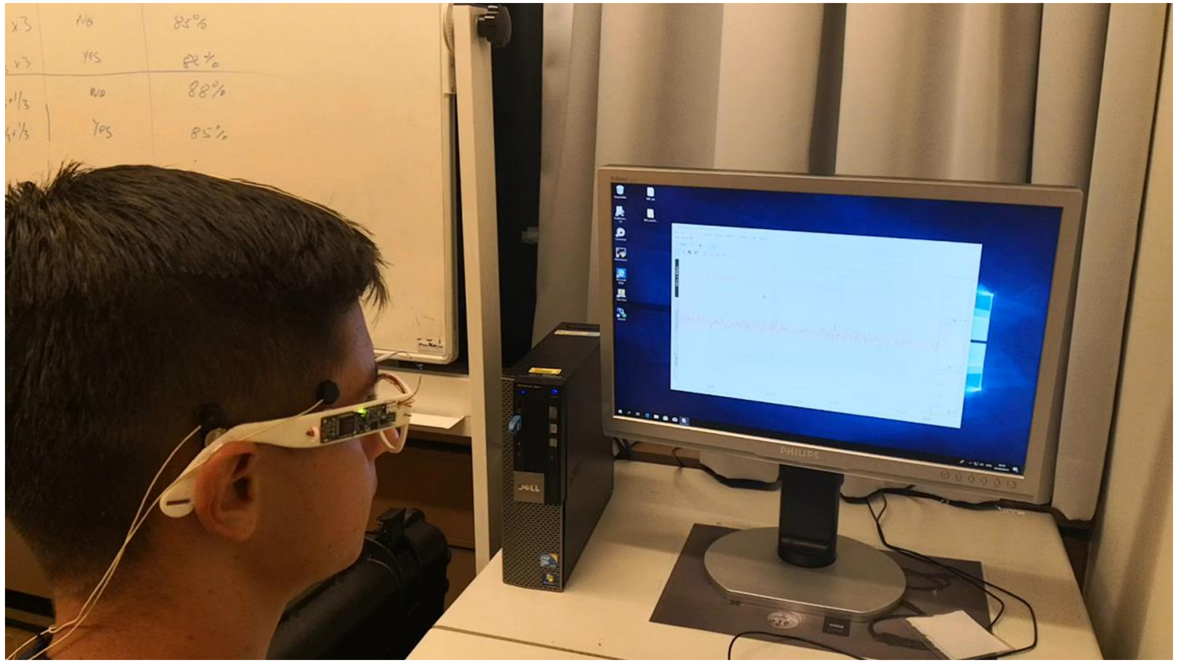
D. Sopic, A. Aminifar, and D. Atienza, “e-Glass: A Wearable System for Real-Time Detection of Epileptic Seizures,” in 2018 IEEE International Symposium on Circuits and Systems (ISCAS). Florence, Italy: IEEE, may 2018, pp. 1–5.

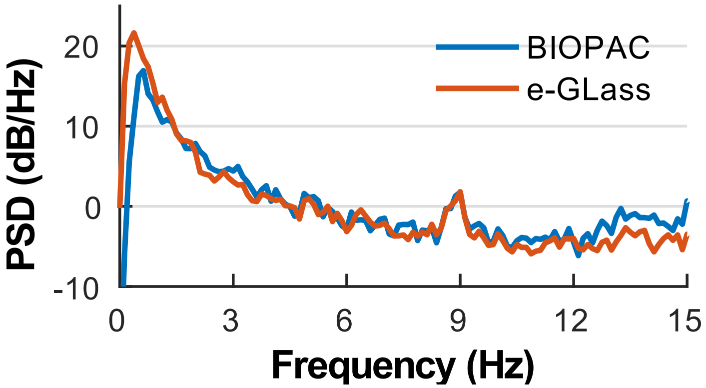
e-Glass vs BIOPAC (commercial EEG recording equipment)

- Deep learning filter per person



Correl. r = 0.85



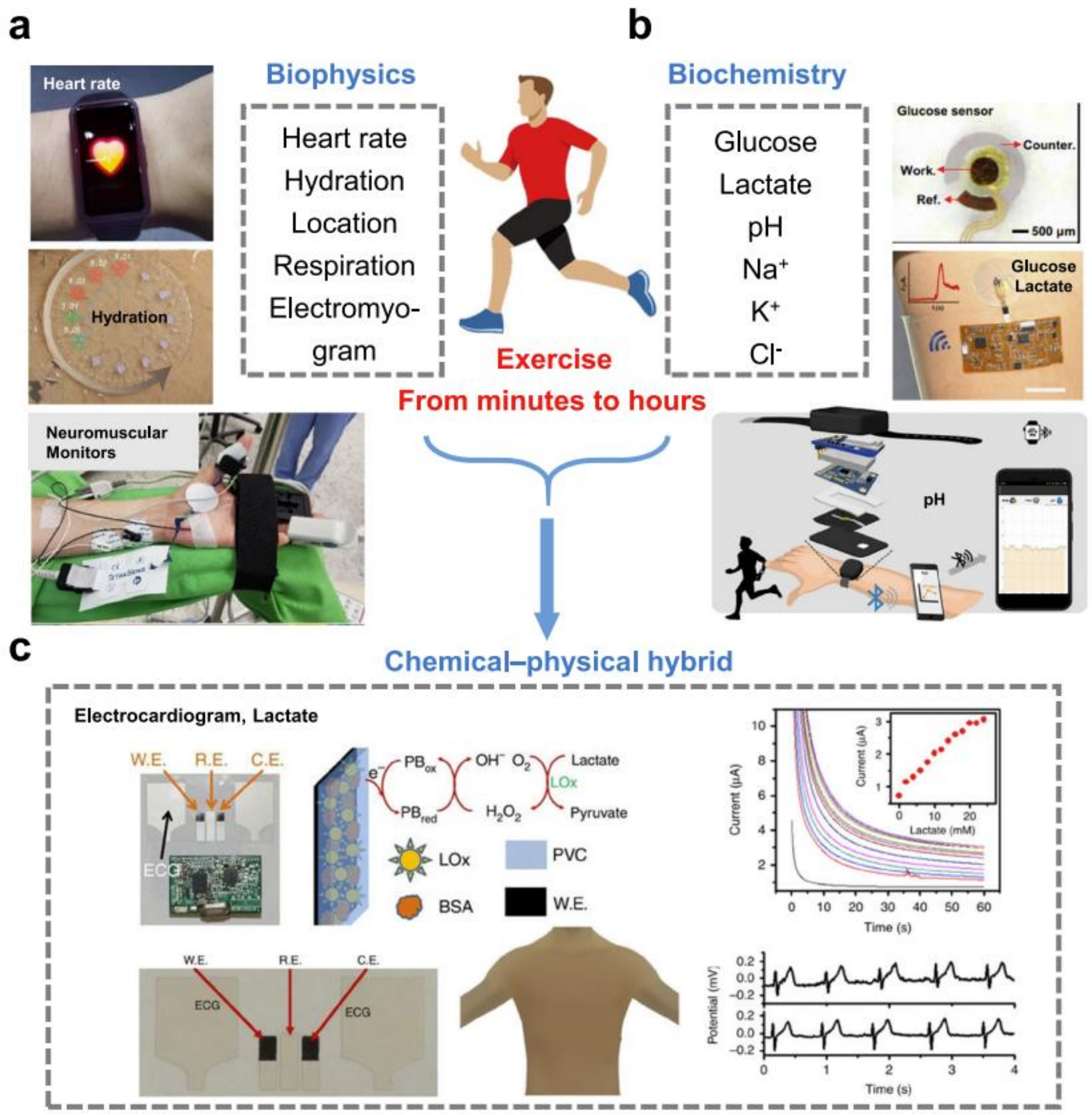


e-Glass and (expensive) BIOPAC show high correlation

Deeply Heterogeneous medical Sensing Systems

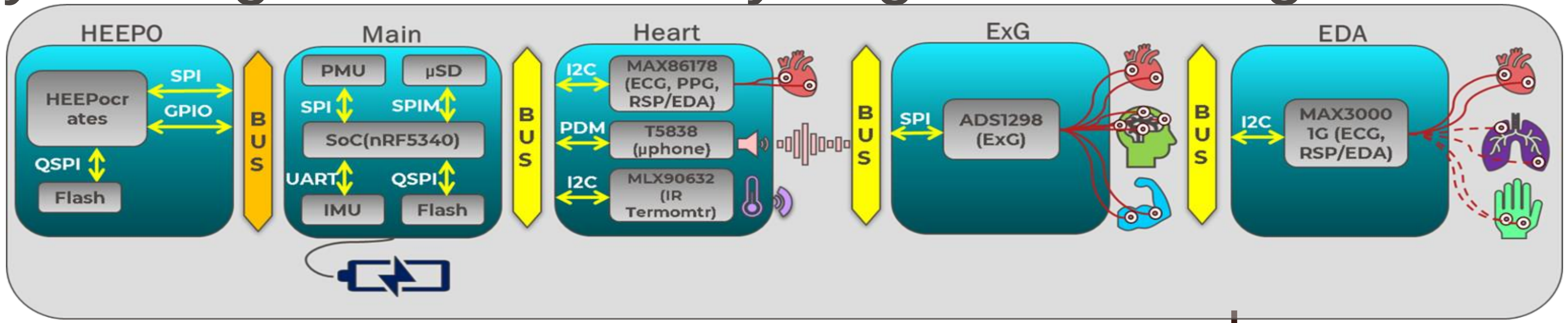
The sensing task is becoming <u>inter</u> and <u>multi</u> modality

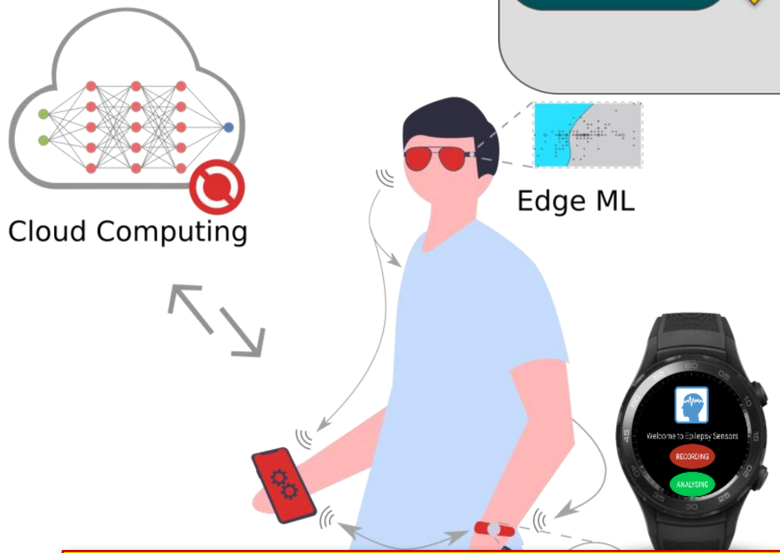
Are our sensing medical technology ready to take such hetereogeneous challenge?



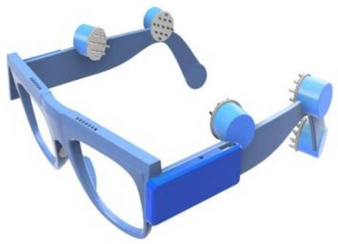
VersaSens: Multi-Parametric Edge AI Network of Wearables

- Plug&Play your edge AI devices as you go to work together

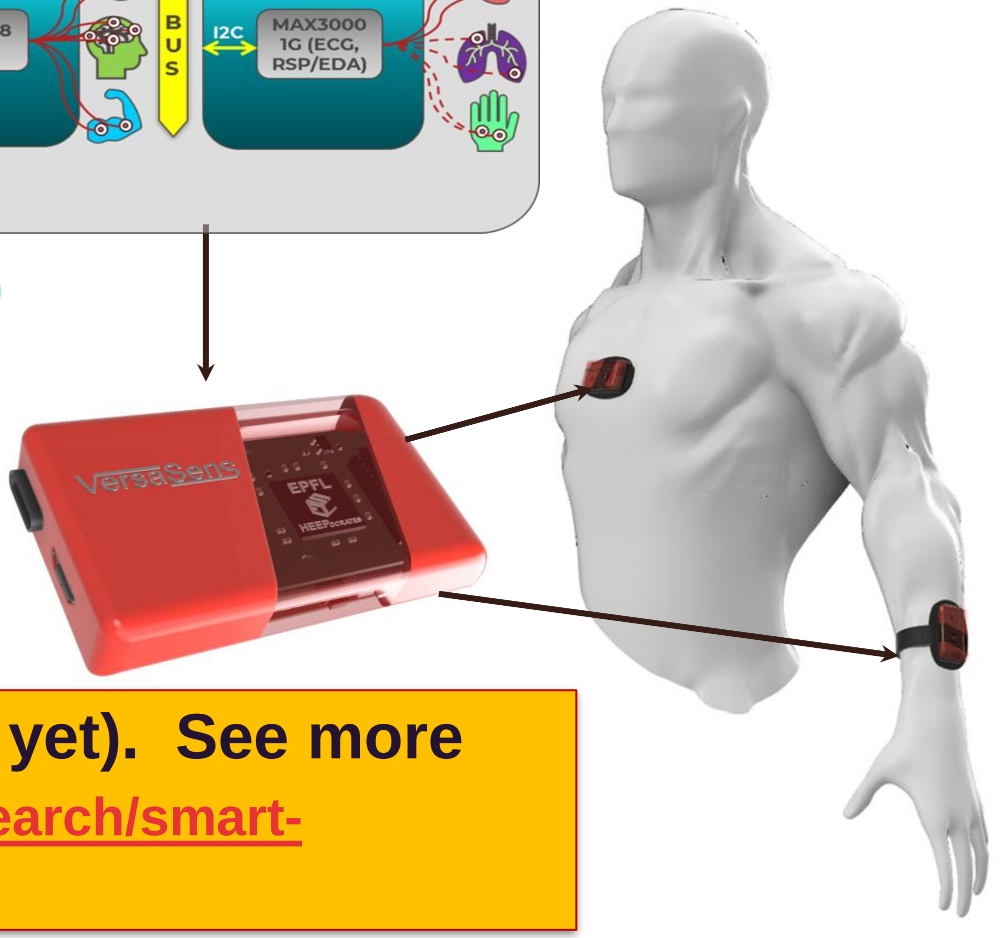




eGlass

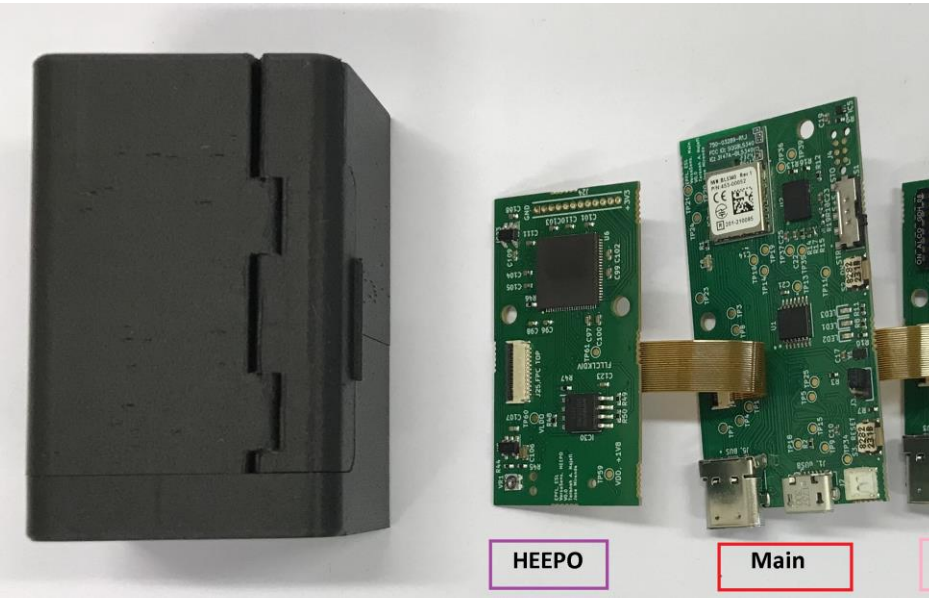


EpiPhone

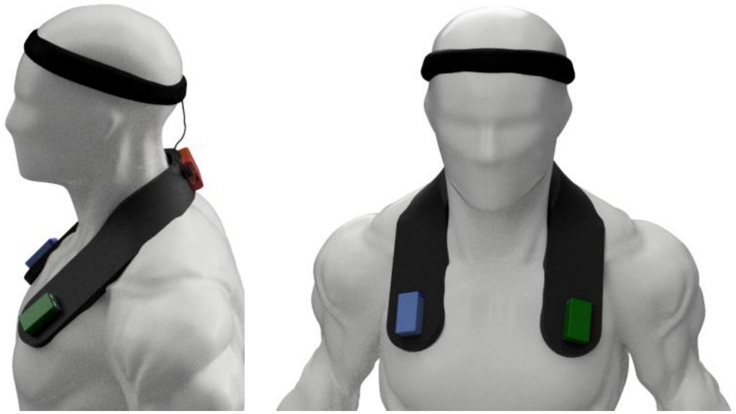


Medge AI devices are possible (but not there yet). See more details here: https://www.epfl.ch/labs/esl/research/smart-wearables/versasens/

- Open-source, easily extensible



Wris



EEG + ECG + EDA: Stress monit. config.

- Multi-location sensing and processing

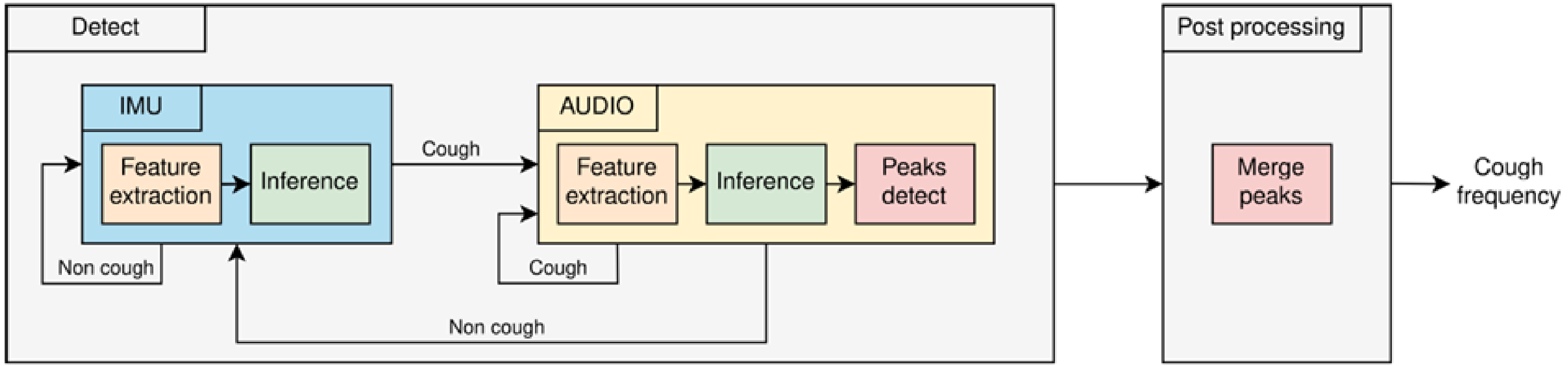
Sensemodi

Edge AI for knee monitoring, see www.sensemodi.com for more details

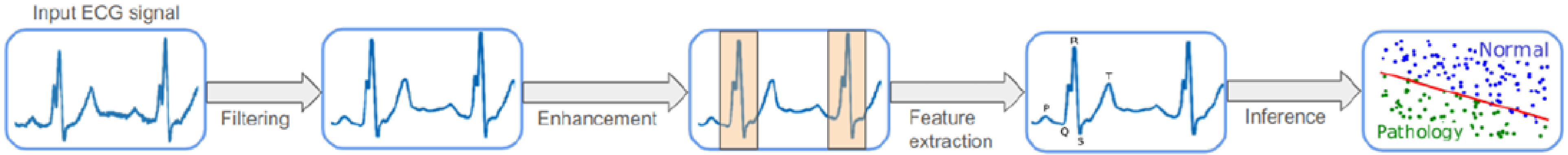
config.

VersaSens: A Modular Multimodal Platform (Different Medical Uses)

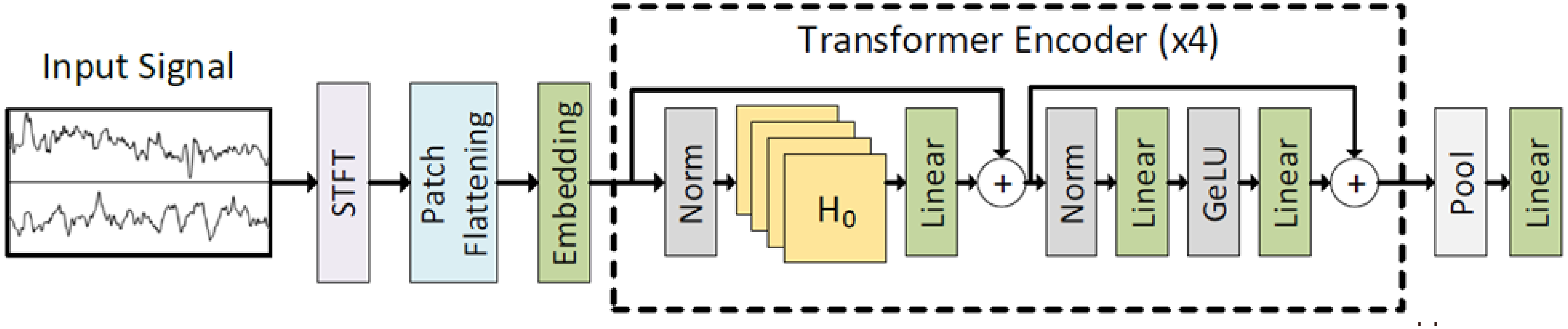
Cough frequency monitoring



Heart Beat Classifier



Seizure detection (transformer)



VersaSens: A Modular Multimodal Platform (Different Medical Uses)

Cough frequency monitoring

Parameter		Processing	Deep sleep
Duration	IMU	11 ms	489 ms
	Audio	114 ms	686 ms
Power consumption	IMU	27.7 mW	9.05 μW
	Audio	30.3 mW	
Voltage		3.3 V	3.3 V
Frequency		128 MHz	32 KHz
Energy consumption	IMU	0.30 mJ	4.42 μJ
	Audio	3.45 mJ	6.2 μJ
Total energy		IMU	**0.31 mJ**
		Audio	**3.51 mJ**

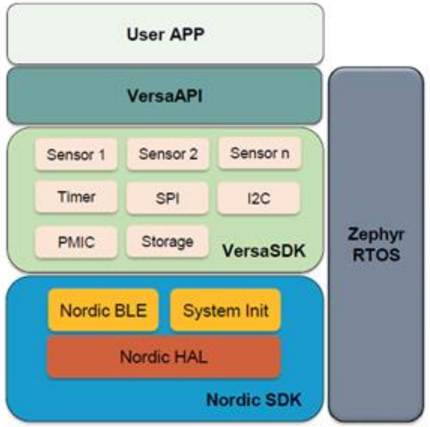
Heart Beat Classifier

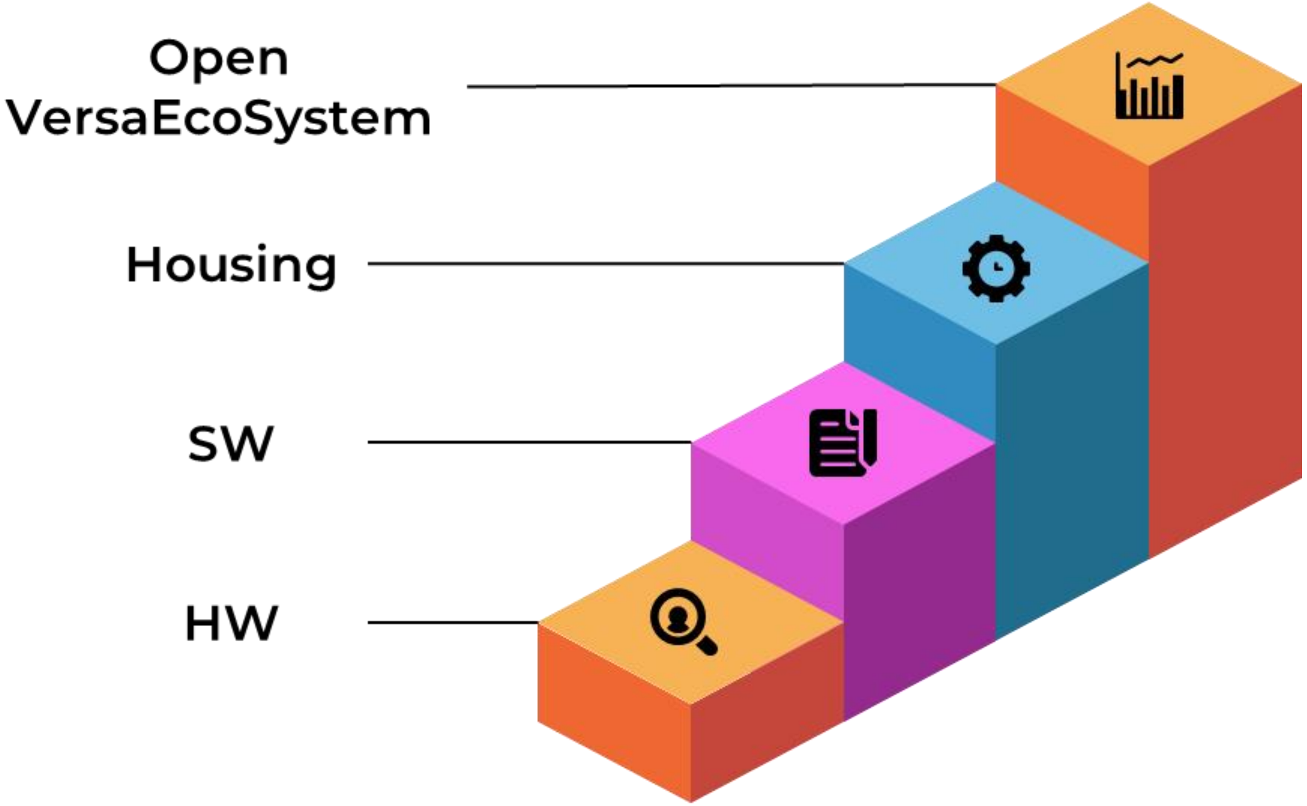
Parameter	Processing	Deep Sleep
Duration	22 ms	11978 ms
Power Consumption	8.68 mW	0.29 mW
Voltage	830 mV	830 mV
Frequency	170 MHz	32 KHz
Energy Consumption	0.19 mJ	3.47 mJ
Total Energy		**3.66 mJ**

Seizure detection (transformer)

Parameter	With CGRA	Without CGRA	Deep Sleep
Processing time	53 ms	79 ms	11947 ms
Power Consumption	8.86 mW	8.83 mW	0.29 mW
Voltage	830 mV	830 mV	830 mV
Frequency	160 MHz	160 MHz	32 KHz
Energy Consumption	0.47 mJ	0.70 mJ	3.46 mJ
Total Energy	**3.93 mJ**	**4.16 mJ**	

VersaSens: A Modular Multimodal Platform (Different Medical Uses)





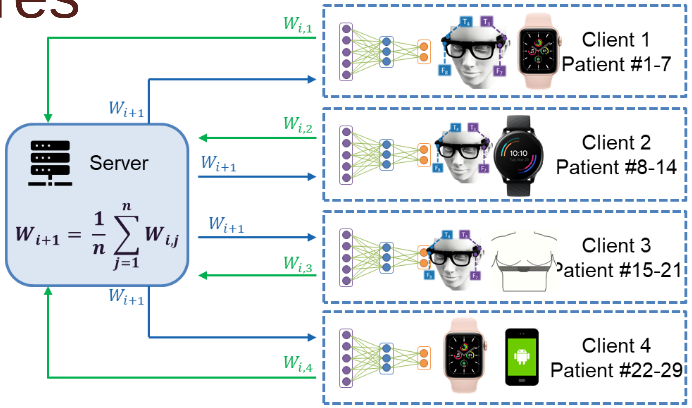
Conclusion

- New domain-specific edge AI systems: follow the brain!
  - Not just Von Neumann, evolution is needed!

- Democratizing edge AI accel.-based systems co-design

X-HEEP

  - Use of application characteristics: analog and digital features
  - Accelerator set (and architecture) keeps evolving
  - Use of FL for efficient edge AI training



- Next-frontier:
  - Mapping more efficiently AI-based medical applications (not C, but Pytorch…)
  - More efficient **on-device learning for large AI models at the edge**

Thanks for Your Attention!

Questions?

jose.mirandacalero@epfl.ch

https://www.epfl.ch/labs/esl/research/

Acknowledgements:
Flavio Ponzina,
Simone Machetti,
Saleh Baghersalimi,
Benoît Denkinger,
Dr. Giovanni Ansaloni,
Dr. Christoph Müller,
Dr. Davide Schiavone,
Dr. Miguel Peón-Quirós,

And thanks to:

Key References and Bibliography (1/3)

Single- vs. multi-core smart wearables platform design

- Michele Caon, et al., "*Scalable and RISC-V Programmable Near-Memory Computing Architectures for Edge Nodes*", ArXiv, 2024.
- Cristian Tirelli, et al., "*SAT-based Exact Modulo Scheduling Mapping for Resource-Constrained CGRAs*", ACM JETC, June 2024.
- Flavio Ponzina, et al., "*A hardware/software co-design vision for deep learning at the edge*", IEEE Micro, 2022.
- B. Denkinger, et al., "*A Very-Wide-Register Reconfigurable-Array Architecture for Low Power Embedded Devices*", Proc. DAC, 2022.
- E. De Giovanni, et al., "*Modular Design and Optimization of Biomedical Applications for Ultra-Low Power Heterogeneous Platforms*", IEEE T-CAD (ES-WEEK Spec. Issue), November 2020.
- L. Duch, et al., "*i-DPs CGRA: An Interleaved-Datapath Reconfigurable Accelerator for Embedded Bio-signal Processing*", Embedded Systems Letters, June 2019.
- L. Duch, et al., "*HEAL-WEAR: An Ultra-Low Power Heterogeneous System for Bio-Signal Analysis*", IEEE TCAS-I, September 2017.
- R. Braojos, et al. "*Nano-Engineered Architectures for Ultra-Low Power Wireless Body Sensor Nodes*", Proc. of CODES-ISSS, 2016.

ULP wearables computation optimization and ECG/EEG application mapping

- R. Zanetti, et al., "*Real-Time EEG-Based Cognitive Workload Monitoring on Wearable Devices*", IEEE TBME, 2022.
- D. Pascual, et al., "*A Self-Learning Methodology for Epileptic Seizure Detection with Minimally Supervised Edge Labeling*", DATE 2019.
- D. Sopic, S. Murali, F. Rincon, D. Atienza, "*Touch-Based System for Beat-to-Beat Impedance Cardiogram Acquisition and Hemodynamic Parameters Estimation*", Proc. DATE, 2016.
- D. Bortolotti, et al., "*Approximate Compressed Sensing: Ultra-Low Power Biosignal Processing via Aggressive Voltage Scaling on a Hybrid Memory Multi-core Processor*", Proc. of ISLPED, 2014.
- R. Braojos, H. Mamaghanian, A. Junior, G. Ansaloni, F. Rincón, S. Murali, D. Atienza, "*Ultra-Low Power Design of Wearable Cardiac Monitoring Systems*", Proc. DAC, 2014.
- F. Rincon, J. Recas, N. Khaled, D. Atienza, "*Development and Evaluation of Multi-Lead Wavelet-Based ECG Delineation Algorithms for Embedded Wireless Sensor Nodes*", IEEE TITB, Nov. 2011

EPFL

Key References and Bibliography (2/3)

- **Epileptic seizure and ECG monitoring and detection with edge AI wearable devices**
  - S. Baghersalimi, et al., "*M2SKD: Multi-to-Single Knowledge Distillation of Real-Time Epileptic Seizure Detection for Low-Power Wearable Systems*", ACM TIST, June 2024.
  - S. Baghersalimi, et al., "*Real-Time Federated Learning for Epileptic Seizure Detection*", IEEE Journal of Biomedical and Health Informatics (J-BHI), August 2021.
  - F. Forooghifar, et al., "*A Self-Aware Epilepsy Monitoring System for Detection of Seizures in Real Time*", Elsevier MONET, August 2019. Link: https://rdcu.be/bN1Bo
  - U. Pale, et al., "*Combining general and personal models for epilepsy detection with hyperdimensional computing*", Nature Artificial Intelligence In Medicine (ARTMED), January 2024.
  - L. Orlandic, et al., "*Design, Optimization, and Evaluation of a Novel Energy-Efficient, Low-Memory, Real-Time R-Peak Detection Algorithm for Wearable Devices*", Proc. of EMBC, 2019.
  - D. Sopic, et al., "*Real-Time Event-Driven Classification Technique for Early Detection and Prevention of Myocardial Infarction on Wearable Systems*", IEEE TBioCAS, October 2018.
  - D. Sopic, et al., "*e-Glass: A Wearable System for Real-Time Detection of Epileptic Seizures in Children*", Proc. of ISCAS, 2018.
  - H. Mamaghanian, et al., "*Design and Exploration of Low-Power Analog to Information Conversion Based on Compressed Sensing*", IEEE JETCAS, 2012.
- **ULP multi-biosignal analysis and classification flows**
  - L. Orlandic, et al., "*The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms*", Scientific Data (SDATA), Nature Research, September 2021.
  - S. Zanoli, et al., "*An Event-Based System for Low-Power ECG QRS Detection*", Proc. of DATE 2020.
  - G. Surrel, et al., "*Event-Triggered Sensing for High-Quality and Low-Power Cardiovascular Monitoring Systems*," IEEE Design & Test (D&T), IEEE Press, October 2019.

Key References and Bibliography (3/3)

E2CNN, Self-Awareness and Federated/Distributed Learning on edge AI wearables

- S. Baghersalimi, T. Teijeiro, D. Atienza, A. Aminifar, “*Real-Time Federated Learning for Epileptic Seizure Detection*”, IEEE Journal of Biomedical and Health Informatics (J-BHI), August 2021.
- Ponzina, Flavio, et al. "*E2CNNs: Ensembles of Convolutional Neural Networks to Improve Robustness Against Memory Errors in Edge-Computing Devices*“, IEEE Transactions on Computers, August 2021.
- Ponzina, Flavio, et al. "*A Flexible In-Memory Computing Architecture for Heterogeneously Quantized CNNs*“, Prof. of ISVLSI. 2021.
- F. Forooghifar, A. Aminifar, D. Atienza, “*Resource-Aware Distributed Epilepsy Monitoring Using Self-Awareness from Edge to Cloud*”, IEEE Transactions on Biomedical Circuits and Systems (TBioCAS), December 2019.
- F. Forooghifar, A. Aminifar, L. Cammoun, I. Wisniewski, C. Ciumas, P. Ryvlin, D. Atienza, “*A Self-Aware Epilepsy Monitoring System for Detection of Seizures in Real Time*”, Elsevier Mobile Networks and Applications (MONET), August 2019. Link: https://rdcu.be/bN1Bo
- L. Duch, P. Garcia, S. Ganapathy, A. Burg, D. Atienza, “*Energy vs. Reliability Trade-offs Exploration in Biomedical Ultra-Low Power Devices*”, Proc. DATE, 2016.
- M. Sabry, D. Atienza, F. Catthoor, “*OCEAN: An Optimized HW/SW Reliability Mitigation Approach for Scratchpad Memories in Real-Time SoCs*”, ACM TECS, 2014.
- G. Karakonstantis, M. Sabry, D. Atienza, A. Burg, “*A Quality-Scalable Spectral Analysis System for Energy Efficient Health Monitoring*”, Proc. of DATE, 2014.

BONUS

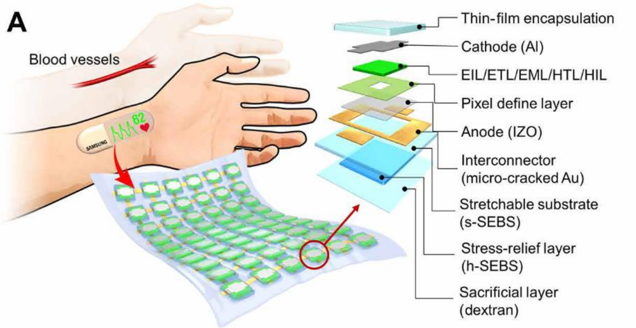
Dr. Jose Miranda (ESL-EPFL)

The Edge: How real is the sixth-sense?

Enabled activity and fitness tracking

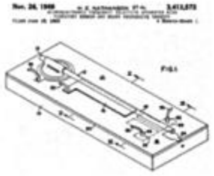
Tracks oxygen levels and easy way to track pulse

Adds more measurements to track fitness



First Pedometer (1760) → First Electrocardiogram (1895) → First Pulse Oximeter (1935) → First MEMS Device (1965) → FitBit Releases First Product (20… → Apple Releases Apple Watch with ECG → Future

Tracks electrical activity of heart



Enables miniaturization of electrical systems

Can this thing "sense"?...

The Edge: How real is the sixth-sense?

- From sensing (VR) to sense (Mixed Reality):
  - Head tracking: 4 visible light camera
  - Eye tracking: 2 IT camera
  - Depth sensors
  - IMUs
  - uPhones, speakers: VAD, KWS
  - Real-time environment mesh
  - Ability to identify and differentiate between objects

https://www.youtube.com/watch?v=FZhbJZEgKQ4
(55:07 – 57:00)

What about battery?