



Energy-Efficient Tiny Machine Learning at the Edge for Next Generation of Smart Sensors

Dr. Michele Magno

Credits: Tommaso Polonelli, Jonas Erb, Moritz Scherer, Philipp Mayer, Manuel Eggimann, Luigi Piccinelli, Prof. Dr. Luca Benini,

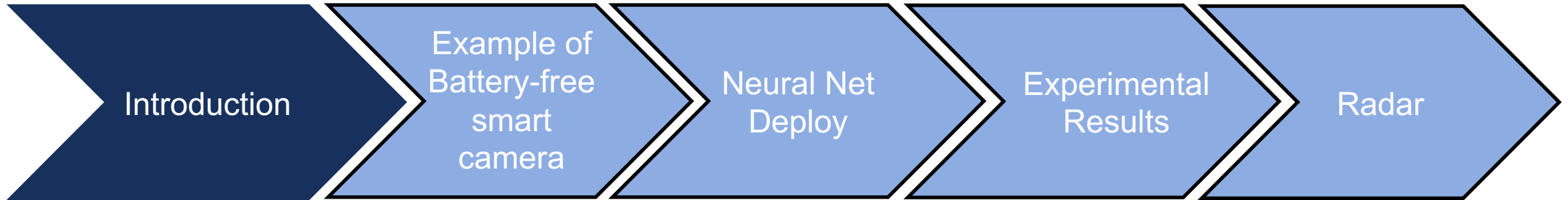
Michele Magno



Dr. Magno is a senior scientist and head of the Project-based Learning Centre at ETH Zurich.

Dr. Magno is a Senior Member of IEEE, the finalist of ETH Spark Award 2018, and a recipient of many other awards and grants. His background is in computer sciences and electrical engineering.

Overview



Introduction

- Internet of Things (IoT)
 - Variety of sensors
 - Connected to cloud (often wirelessly)
- Machine learning
 - Extract relevant information from data
 - High computational demand
- Data Processing
 - Mainly in the cloud



Internet of Things pushes AI and ML at the edge

The world is producing excessive amounts of "unstructured data" that need to be reconstructed

(IBM's CTO Rob High)

"A PC will generate 90 megabytes of data a day, an autonomous car will generate 4 terabytes a day, a connected plane will generate 50 terabytes a day."

Source: Samsung HBM

Source: Tractica

Bandwidth



1 Billion cameras WW (2020)
30B Inference/sec

Latency



Communication latency also with 5G or other networks is in the range of hundred of milliseconds

Availability

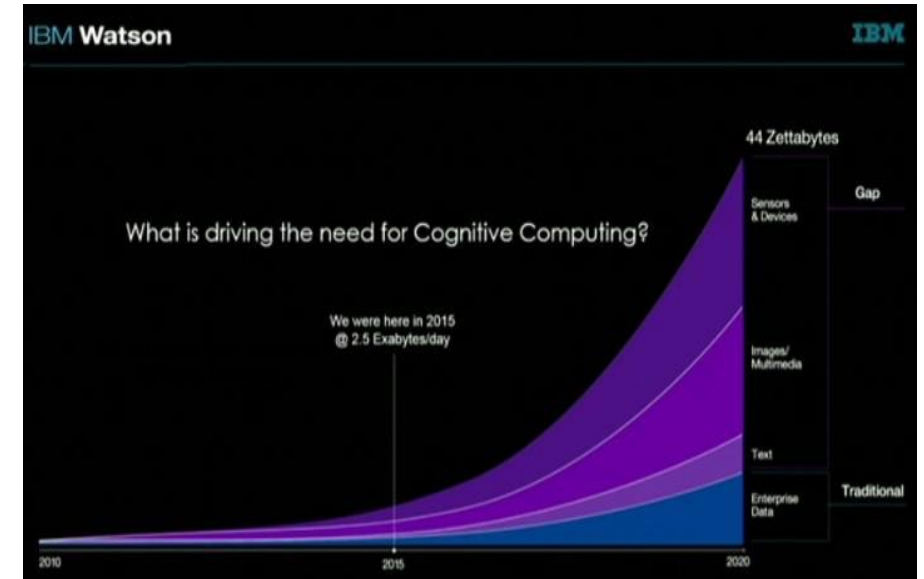


50% of world at less than 8mbps Only 73% 3G/4G availability WW

Security








Data traveling in the network are more vulnerable. Attacks to networks and communication towers



Source: IBM

Since 2015, roughly 2.5 Exabyte of data are being generated per day. Projection shows a 44 Zettabytes of data per day by 2020.

Edge Vs Cloud

- Latency/reliability 
- Data Protection 
- No Wireless Communication Needed – Lower Bandwidth requirements 
- Lower Power Consumption 
- Lower Cost 

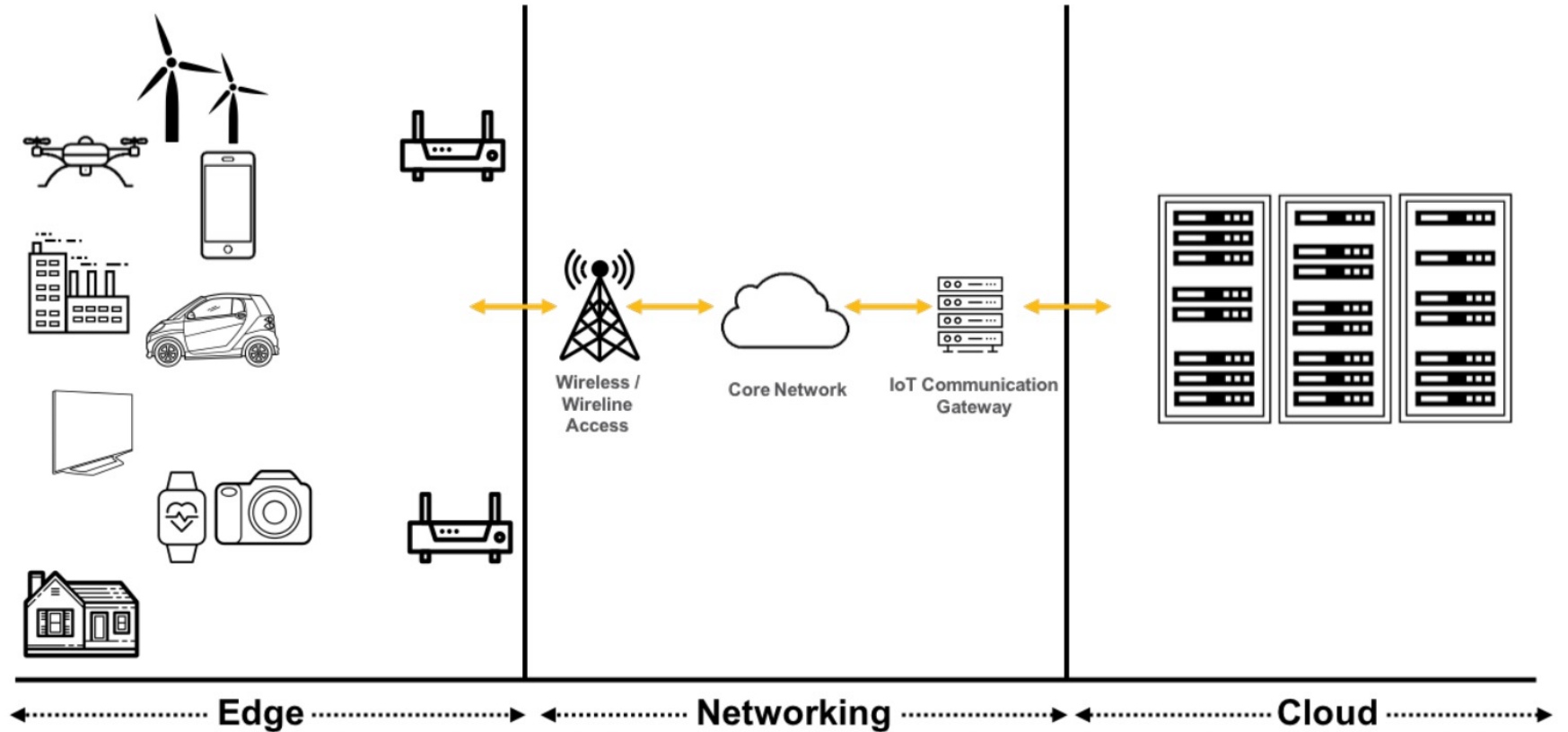


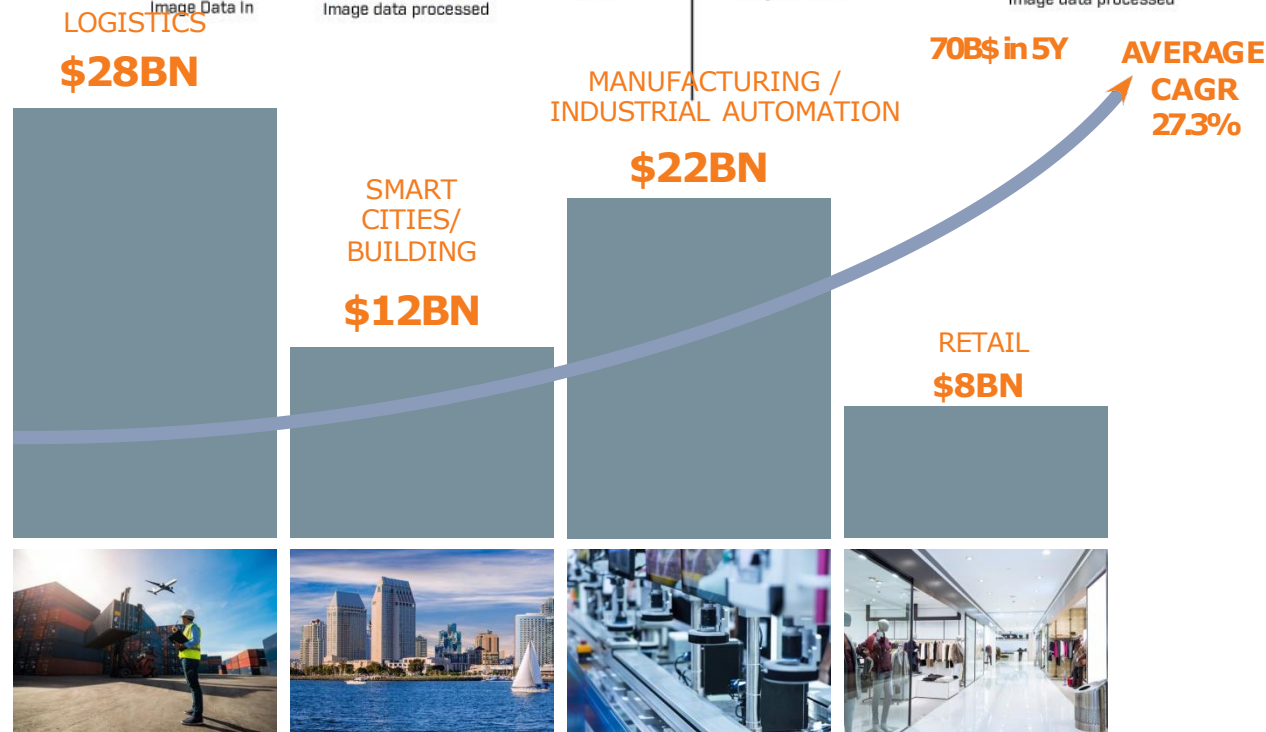
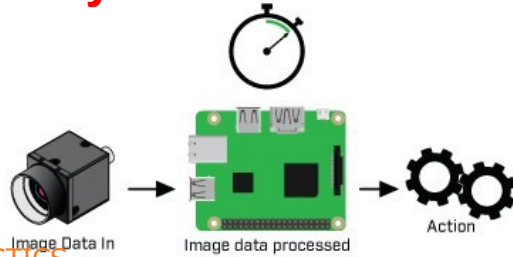
Figure reference: Accelerating Implementation of Low Power Artificial Intelligence at the Edge, A Lattice Semiconductor White Paper, November 2018

Cloud → Edge → Extreme Edge AI a.k.a. TinyML

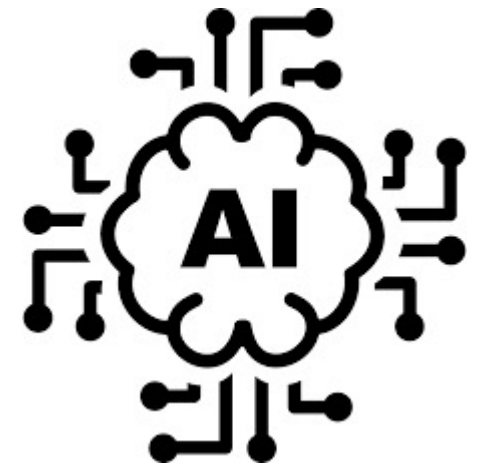
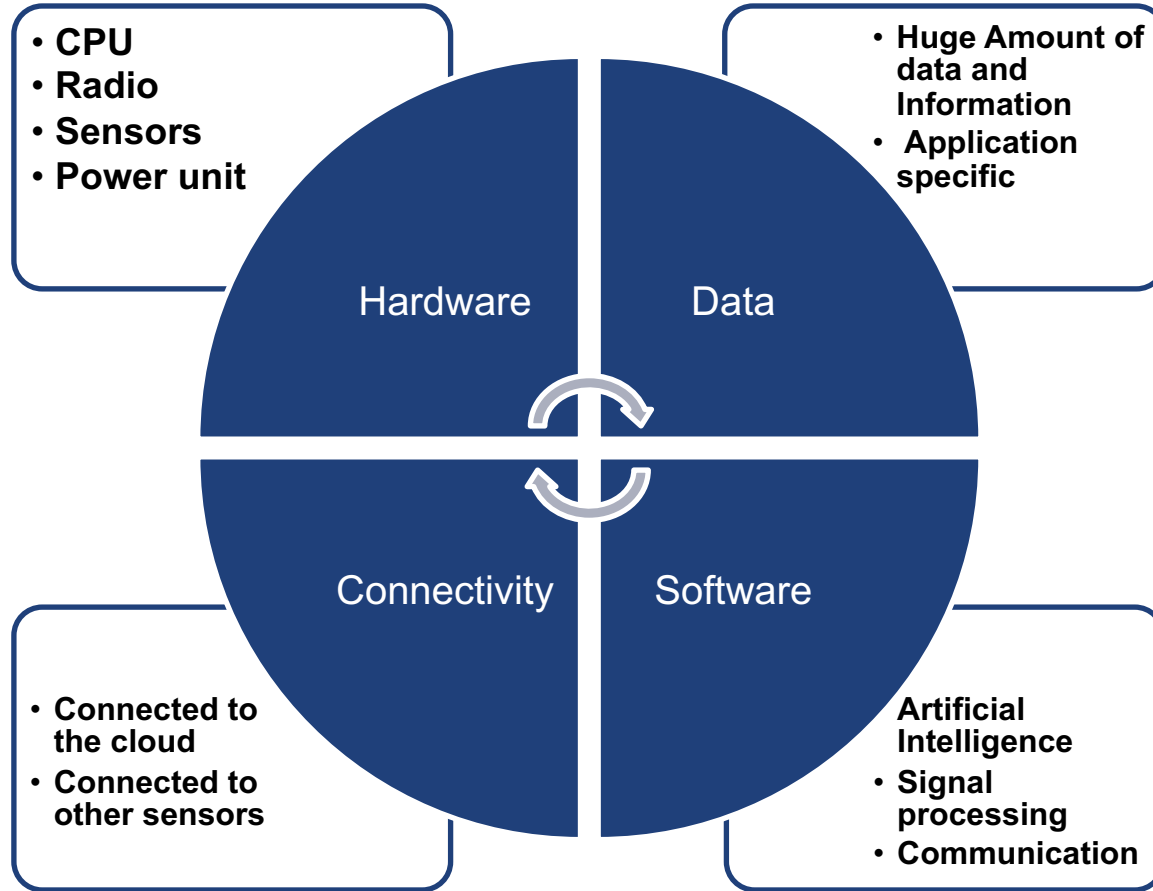
Latency, Privacy

Edge Computing

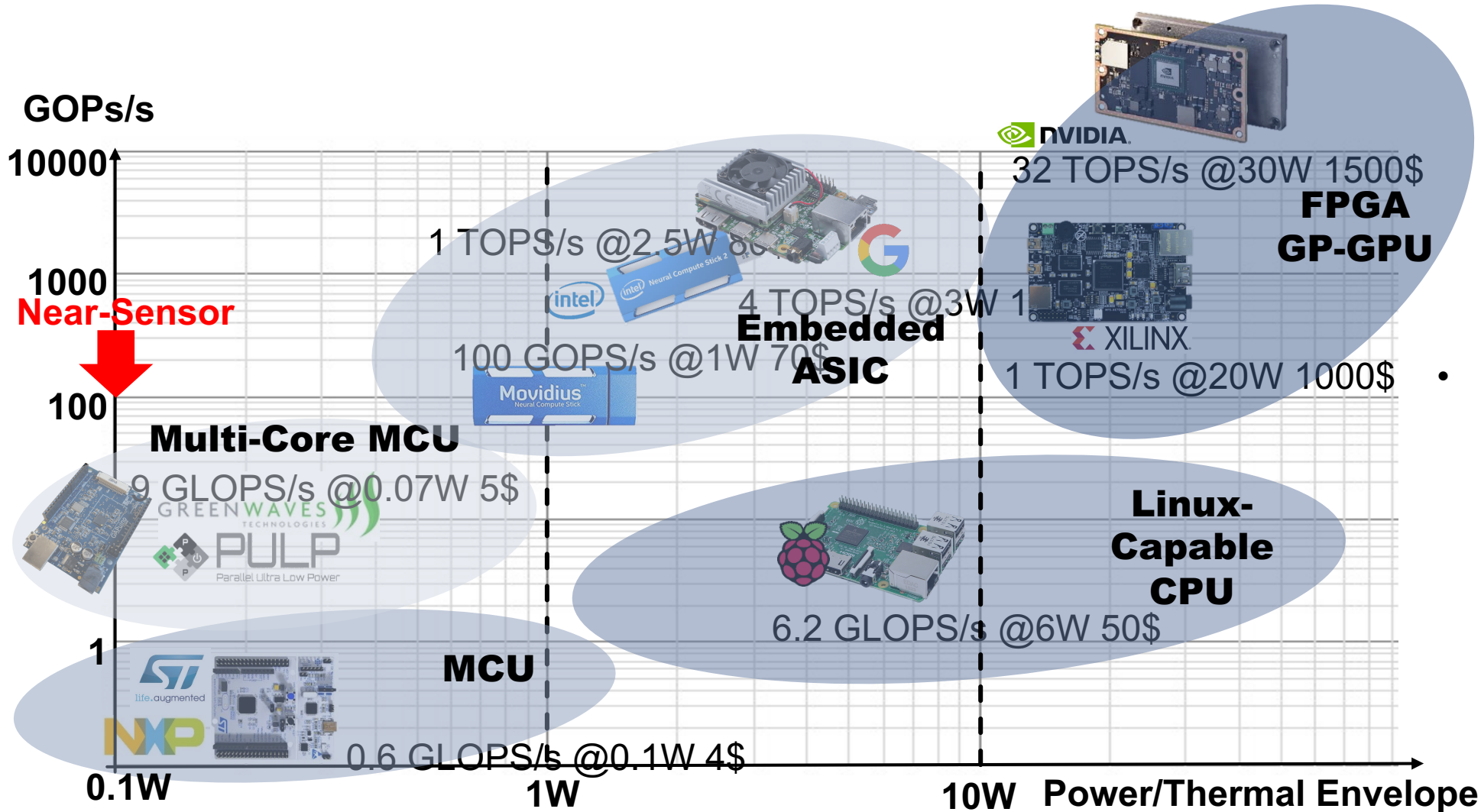
Cloud Computing



What is a IoT Device?



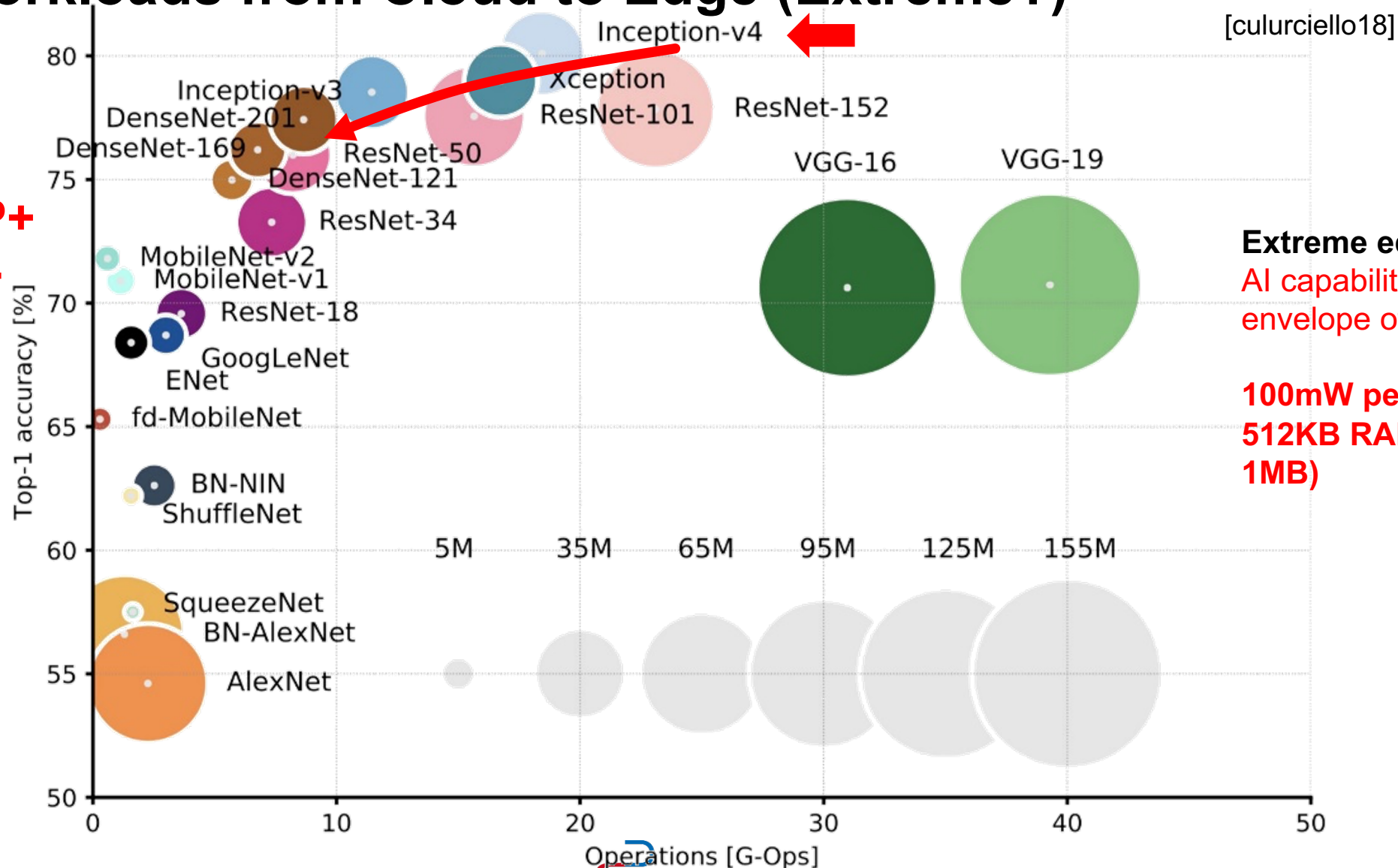
- New platform for edge computing every year. Edge-AI Platforms



- **Trends:**
 - Optimized accelerator for NN
 - Parallel Solutions
 - Custom Architectures

AI Workloads from Cloud to Edge (Extreme?)

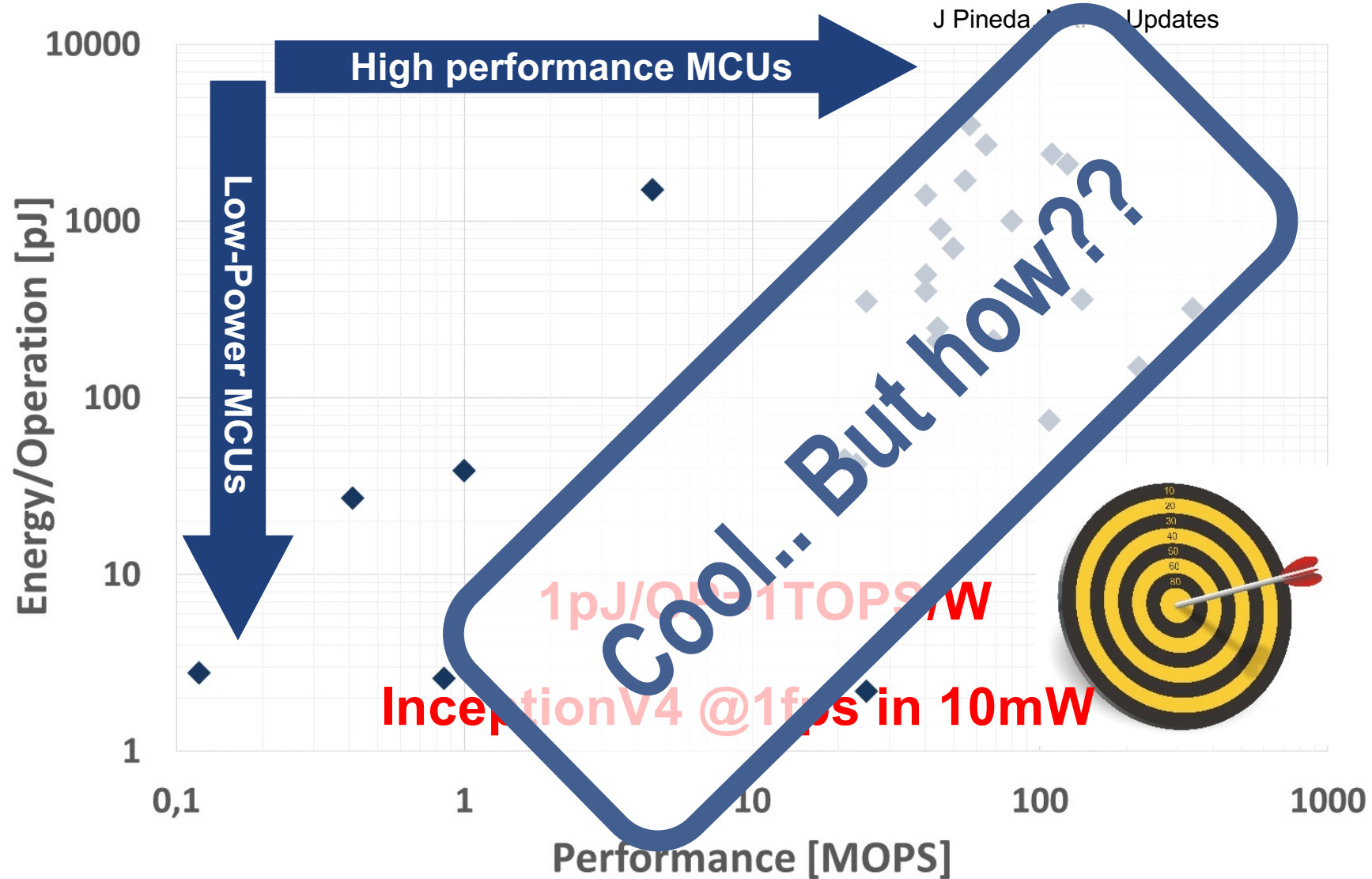
GOP+
MB+



Extreme edge AI challenge
AI capabilities in the power envelope of an MCU:

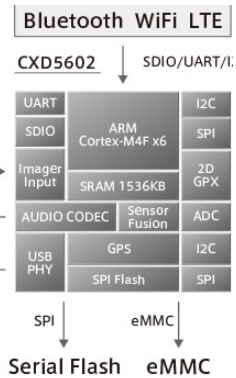
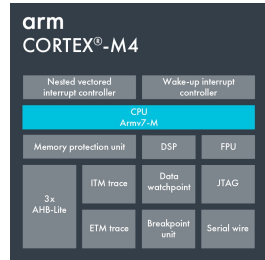
100mW peak (1mW avg)
512KB RAM (best case 1MB)

Energy efficiency @ GOPS is THE Challenge

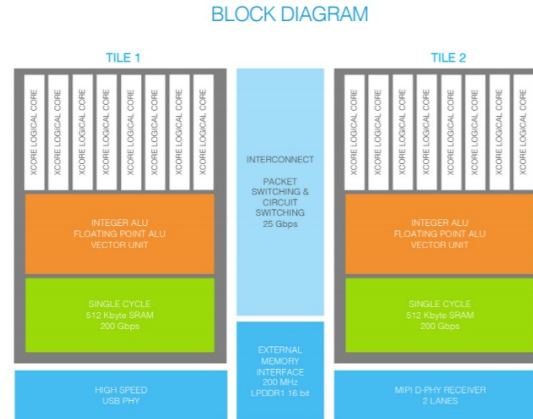


New Trend is improving operations for Cycle : parallel + accelerators + computational efficient architectures

Evaluate the performance in terms of Latency, Computation power, Energy Efficiency, of below 200mW power platforms.

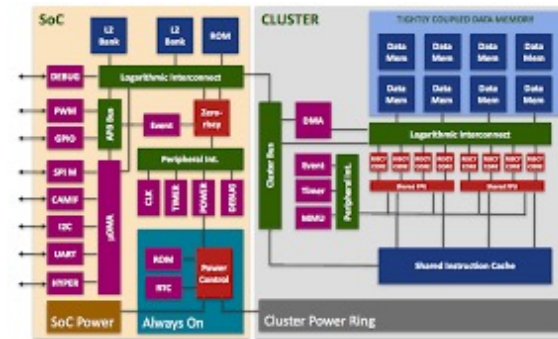


Sony
Spresense
CXD5602
6x ARM-
Cortex-M4



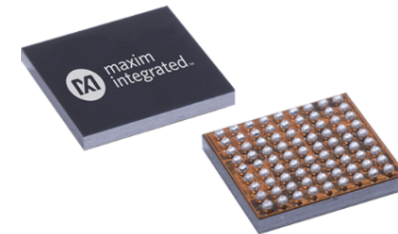
XMOS – XCORE.AI
Parallel – HW C

PULP
(ETH Prof. Benini)
8X-RISC-V Parallel



GAP8
GREENWAVES TECHNOLOGIES

MAX78000



Dual Core
Arm Cortex-M4
RISC-V
HW Accelerator

Next generation of IoT devices for surveillance: **Always-on Smart Sensors.**

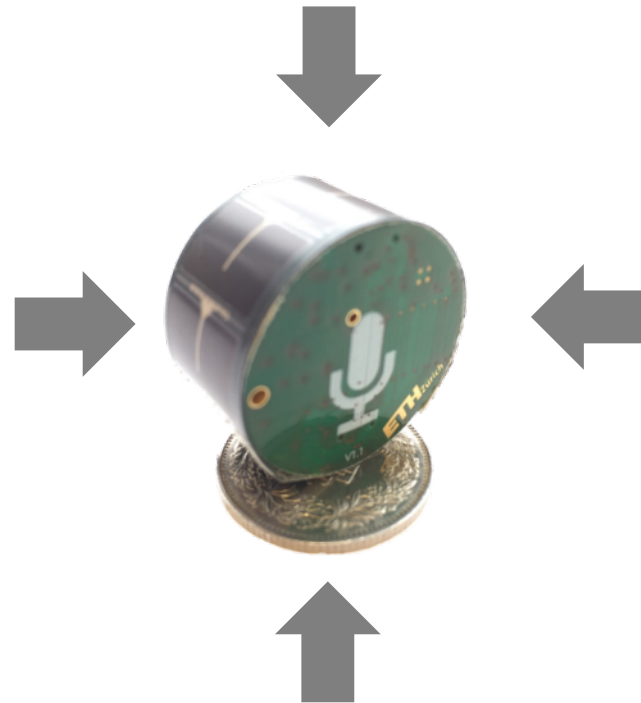
1.) Edge Signal Processing and AI

Smart devices
for perpetual operation

3.) Low power system design

2.) Energy harvesting

4.) Low Power and long range communication



Next generation of IoT device for surveillance: **Always-on Smart Sensors.**

1.) Edge Signal Processing and AI

Smart devices
for perpetual operation

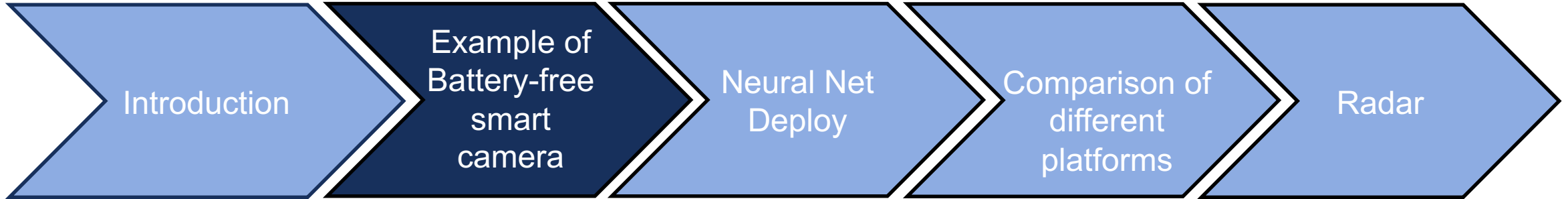
3.) Low power system design

2.) Energy harvesting

4.) Low Power and long range communication



Overview



Goal of this previous work

Miniaturized camera devices are today a commercial reality, widely used by:

- Surveillance
- Monitoring
- Controlling access

They rely on batteries with few hours of operation time

The wave of IoT is pushing the limit of battery-less devices:

- In this work we propose a Battery-Free Long-Range Wireless Smart Camera for Face Detection



Goal of the previous project

The need:

- A device to recognize a person and monitor access to a door
- Line cabling can be very expensive
- If battery powered extremely long lifetime is needed

The solution:

- Energy harvesting to achieve potential infinite lifetime
- Long range wireless communication to report access in real time, low datarate is not a limit

Remarkable results:

- Less than one second per inference
- >90% accuracy over 5 faces
- Proposed neural network model fit in only 115kByte of ROM

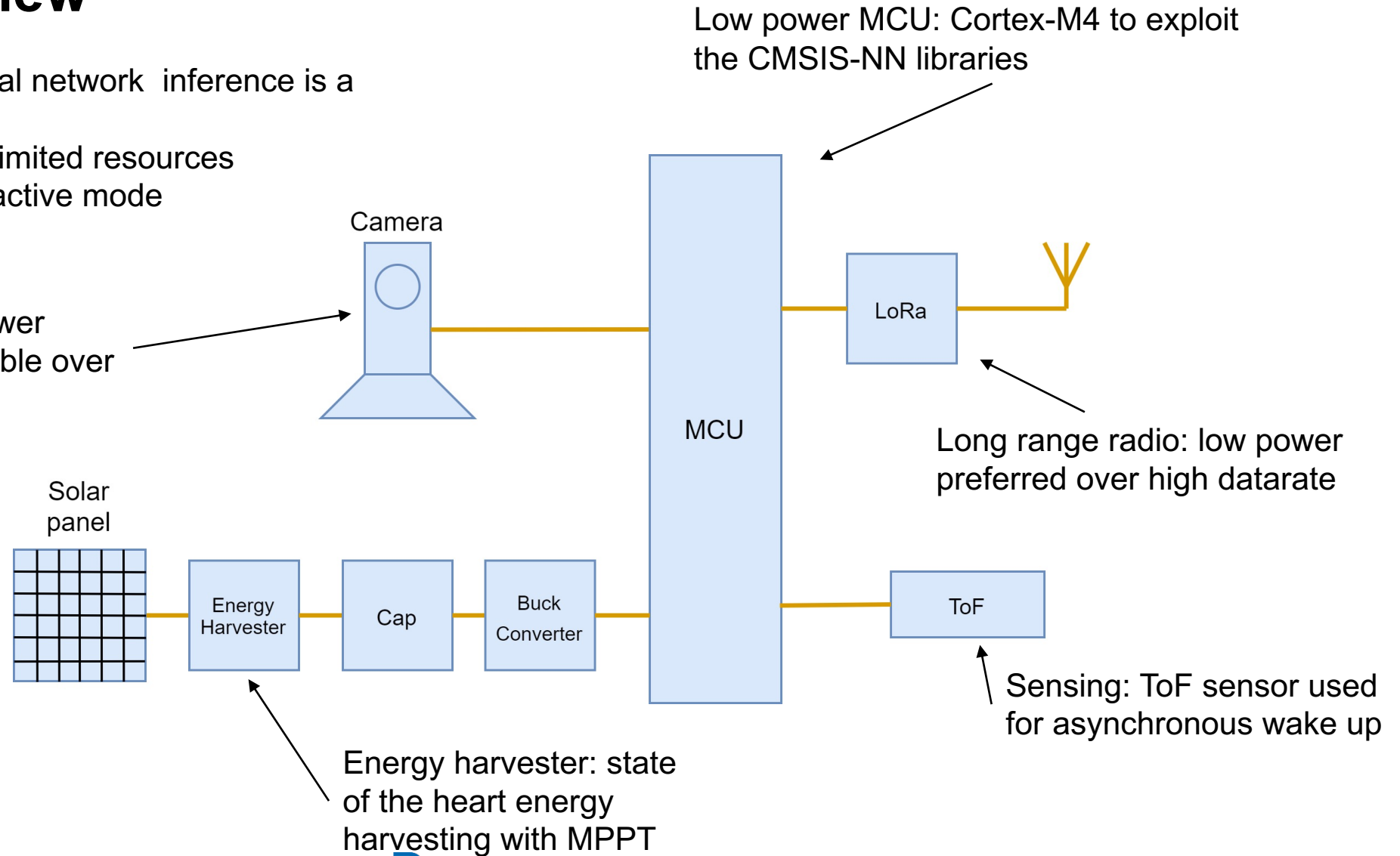


System overview

Self-sustainability for neural network inference is a challenging task:

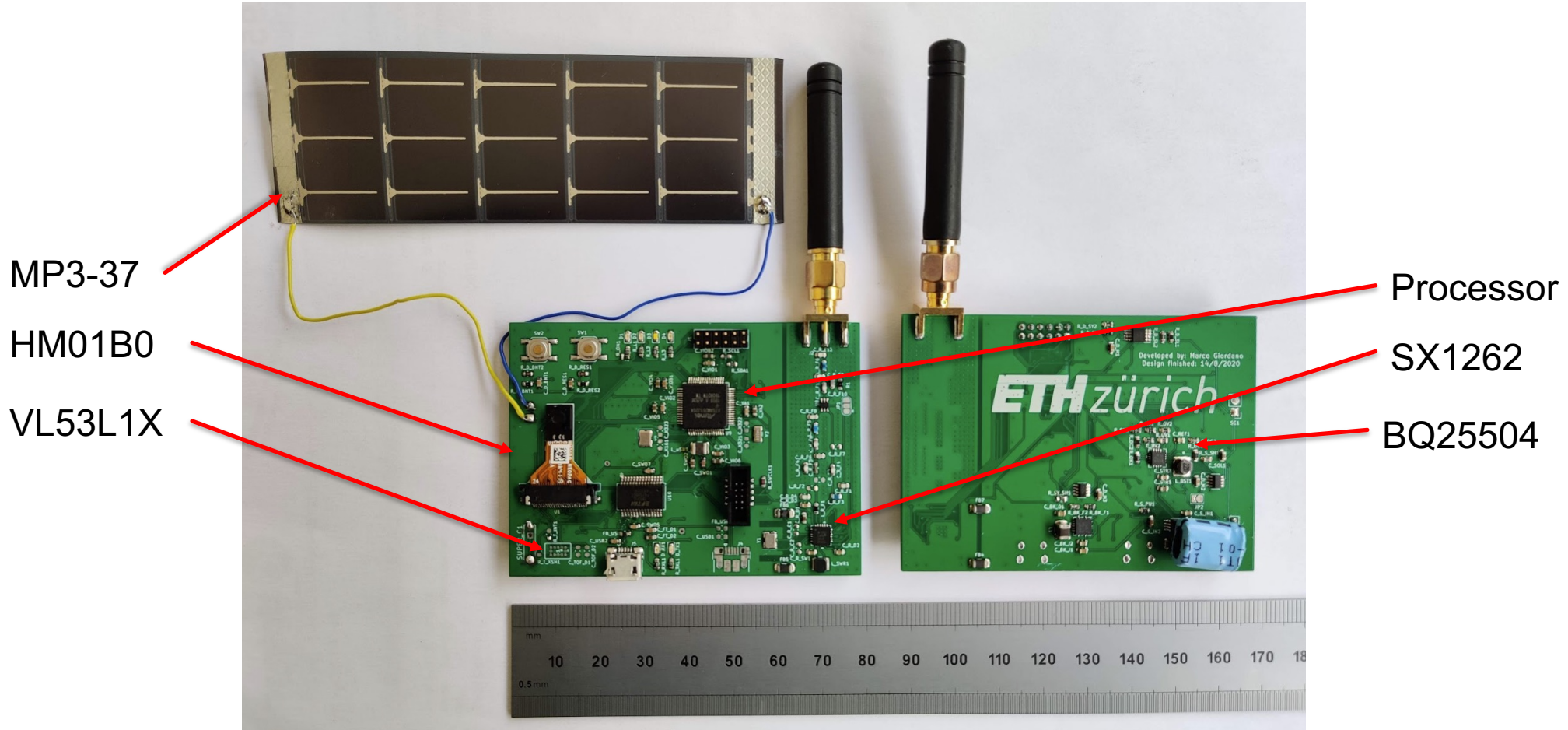
- A microcontroller has limited resources
- long-running times in active mode

Sensing: camera's low power characteristics are preferable over high resolution

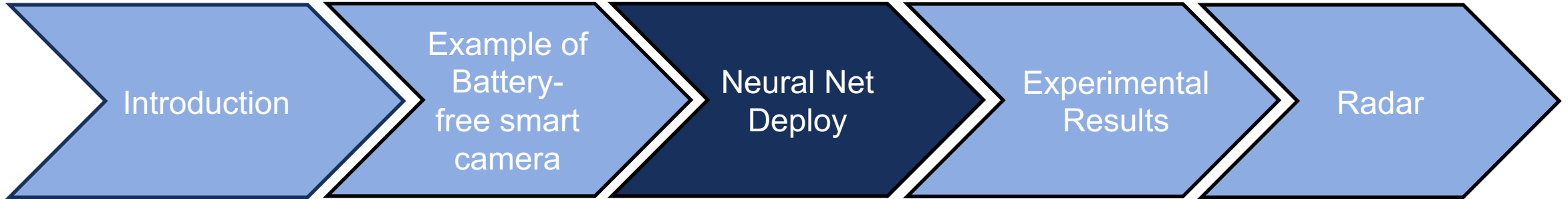


System overview – working prototype

We realized a working prototype of a small battery-less system



Overview



The proposed neural network

Goal:

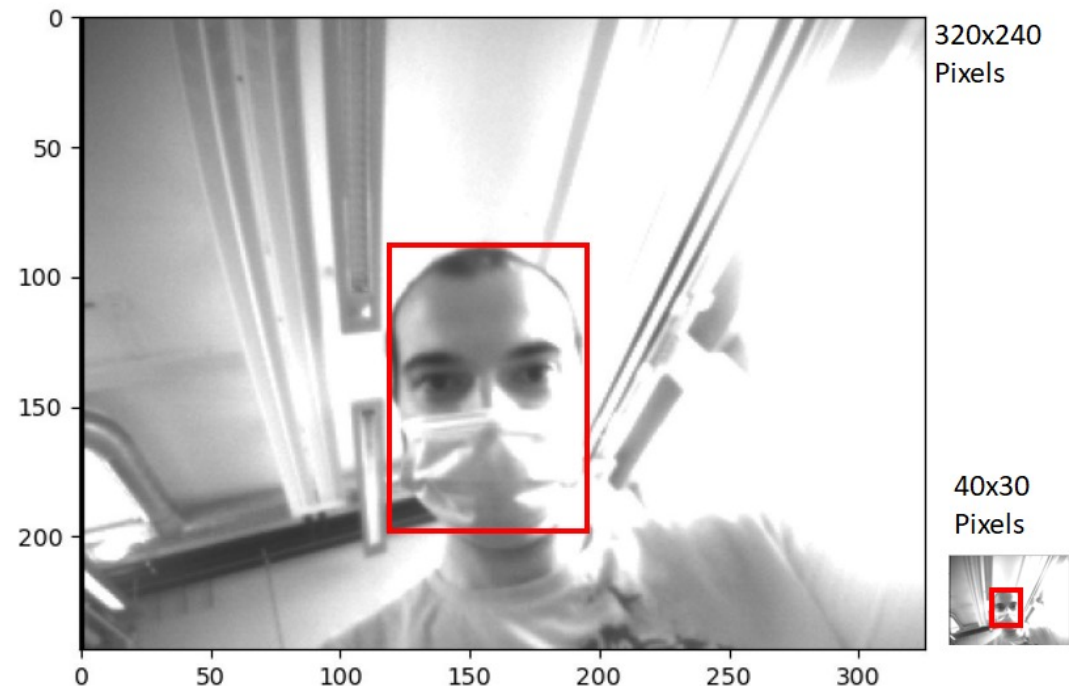
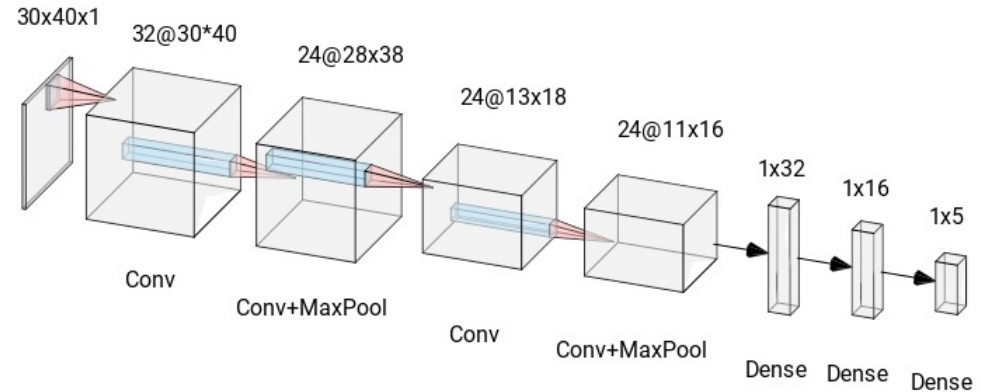
- design a neural network to support inference with limited resources

Challenges:

- Needed to reduce the input size
- Camera shoots greyscale images
- Quantized weights were used to optimize network size and speed up inference

Results:

- More than 90% accuracy
- Only 50kB memory footprint
- Less than 30mJ per inference
- less than 650ms per inference

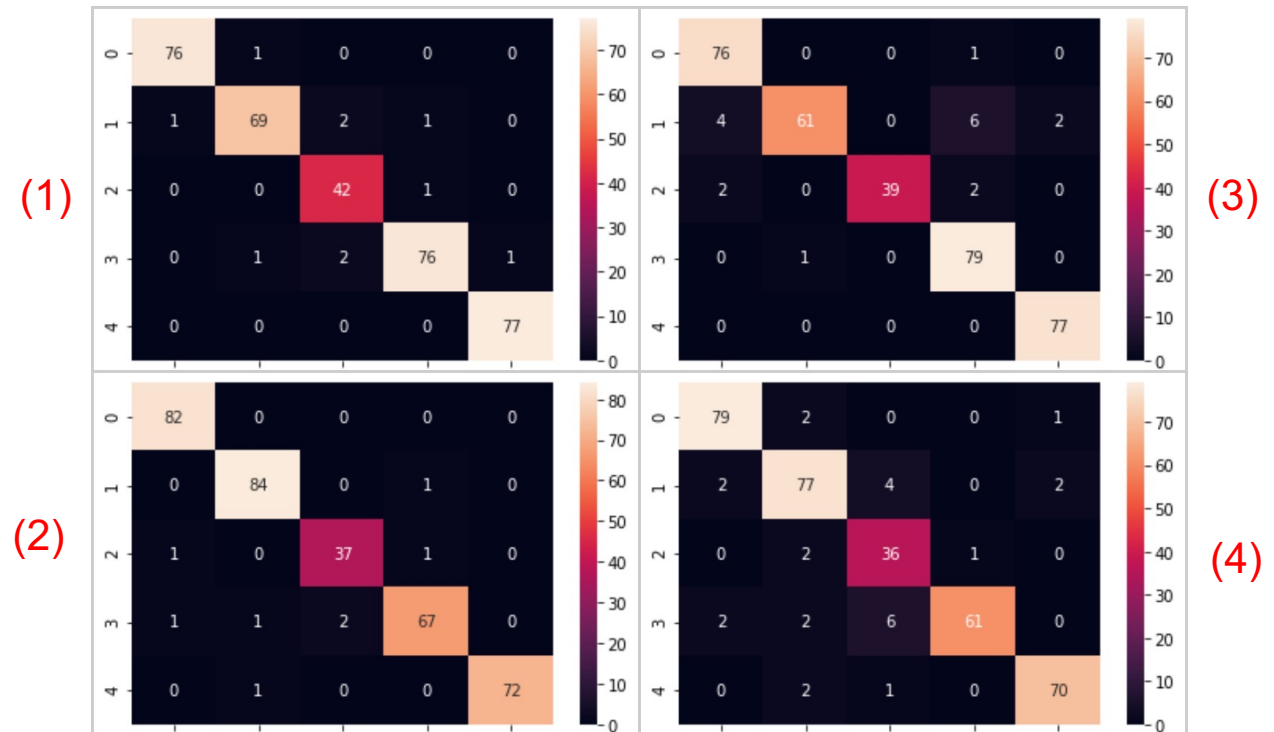


Confusion matrix – Float vs int8

The resolution of 30*40 has been chosen as the best tradeoff between the model's memory occupancy and computing time over a very moderate loss of accuracy.

The numbers in the confusion matrix represent the 5 different faces

Input size/ data type	Accuracy	Precision	Recall	F1-Score
240x320, float (1)	0.97	0.97	0.97	0.97
240x320, int8 (2)	0.95	0.96	0.94	0.95
30x40, float (3)	0.97	0.97	0.97	0.97
30x40, int8 (4)	0.93	0.91	0.92	0.92



Confusion matrix – Float vs int8

We exploited the maximum from Cortex M4:

- Used CMSIS-NN and SIMD (Single Instruction Multiple Data)
- Manually implemented the CMSIS-NN over Tensorflow Lite framework

In order to use CMSIS-NN:

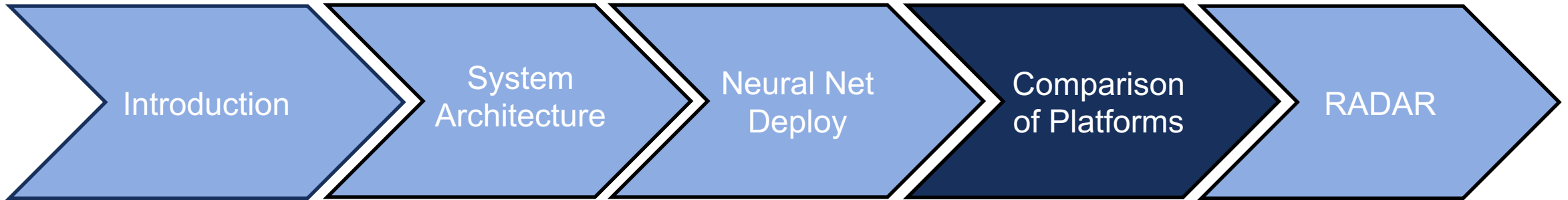
- The model must be quantized to 8 or 16 bits
- Quantized model outperforms floating point in inference time with a little loss in accuracy

Model	ROM	Cycles	Accuracy	Δ Accur	Inference time	Δ Time
Float FPU	- 182192 bytes	110091614	0.97	0	2.29 s	0
Int8 CMSIS	- 114048 bytes	30949719	0.93	-4%	0.64 s	-357%
Int8 Google Kernel	- 114048 bytes	523084654	0.93	-4%	10.89 s	+475%

Demo

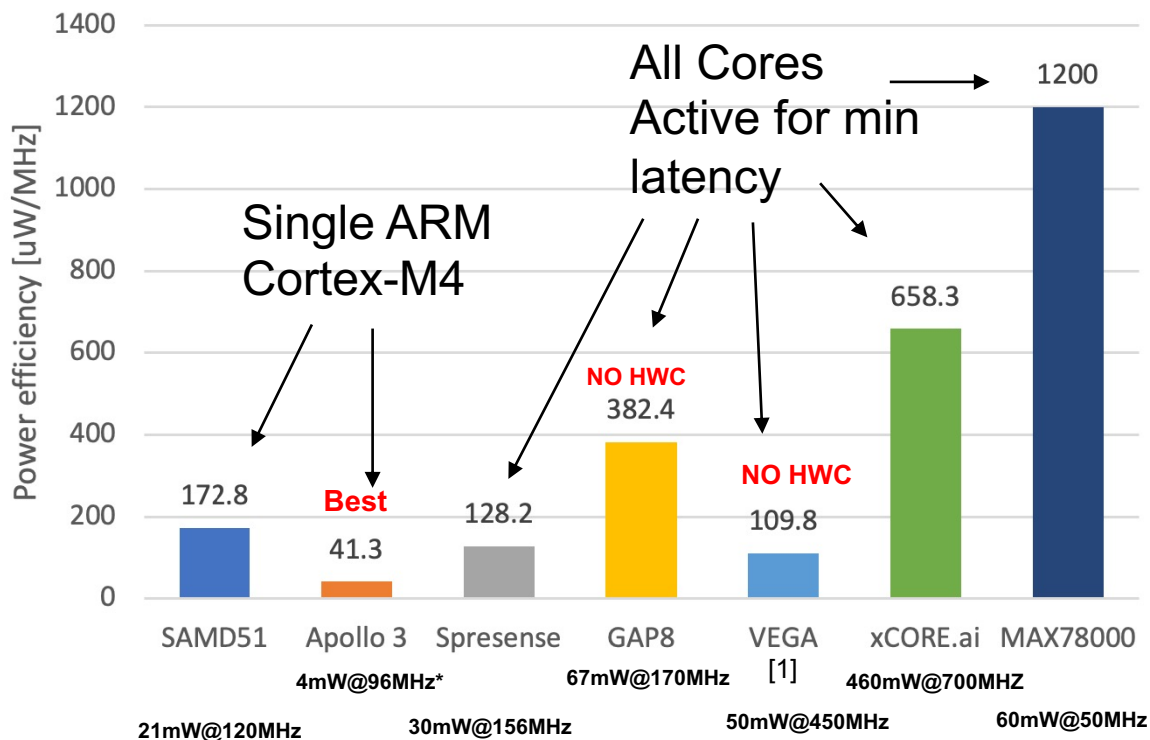


Overview

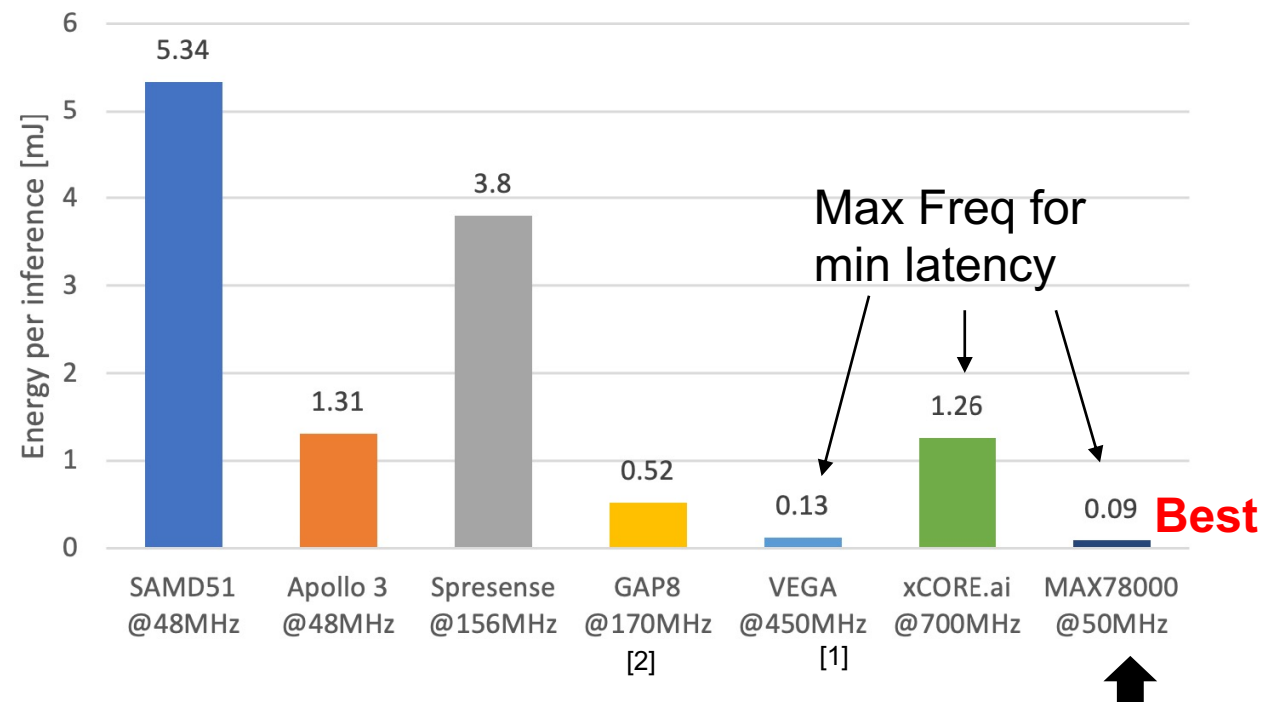


General Comparison: Power vs energy efficiency

Platforms power efficiency



Platform energy consumption



[1] "A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Cores SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up from MRAM-Based State Retentive Sleep Mode", ISSCC2021

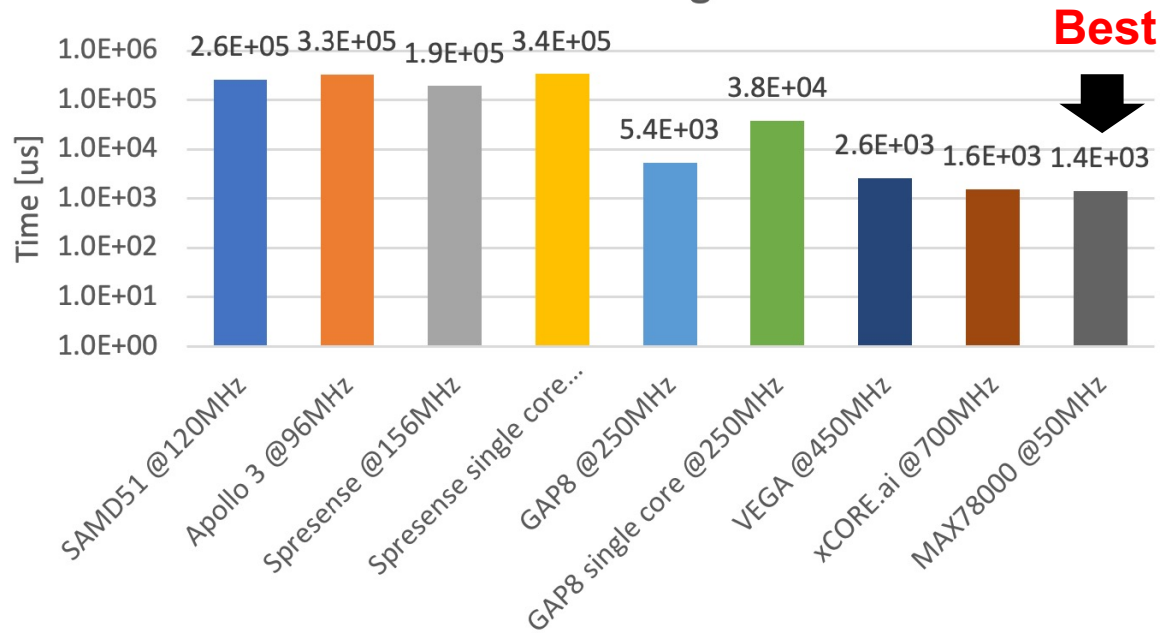
[2] "PULP-NN: Accelerating Quantized Neural Networks on Parallel Ultra-Low-Power RISC-V Processors", Arxiv

*Measured data during the inference, Amiq datasheet claims half current consumption

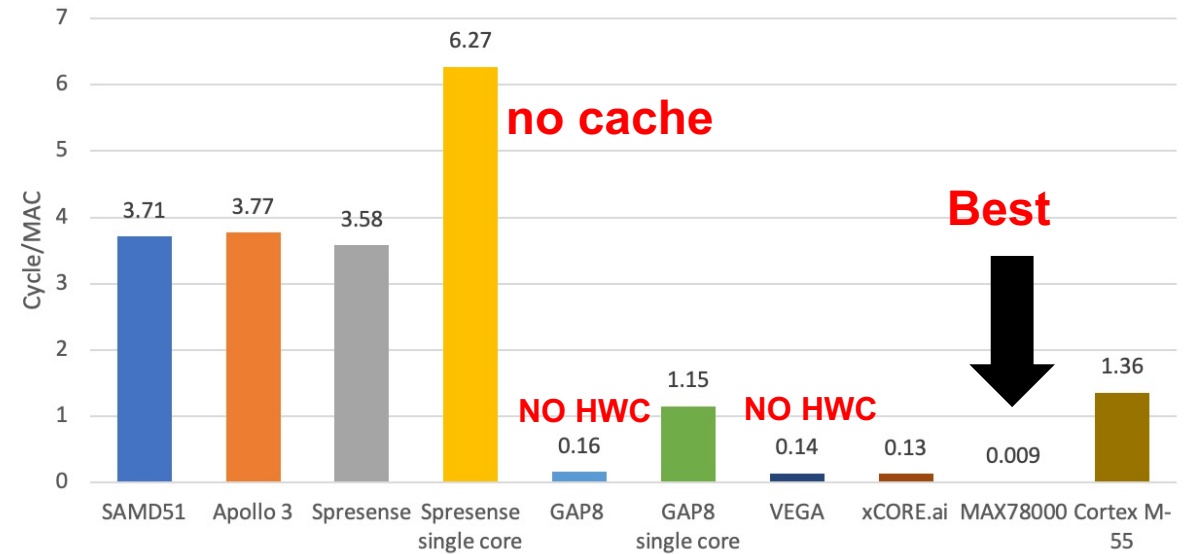
General comparison: Computational efficiency vs min Latency

Max Freq for
min latency

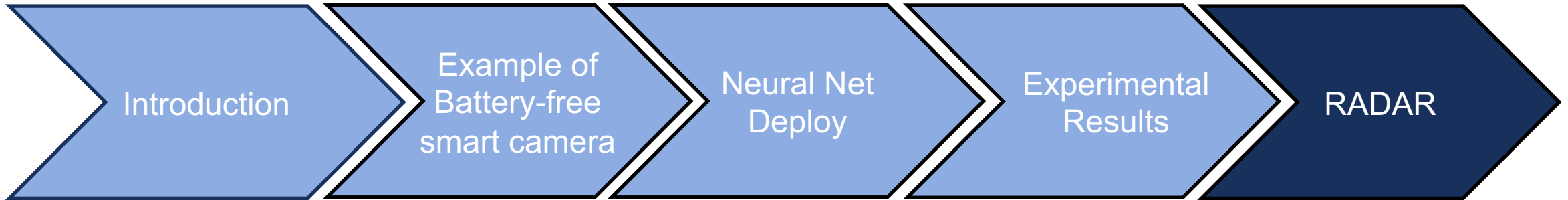
Inference time - log scale



Inference performance

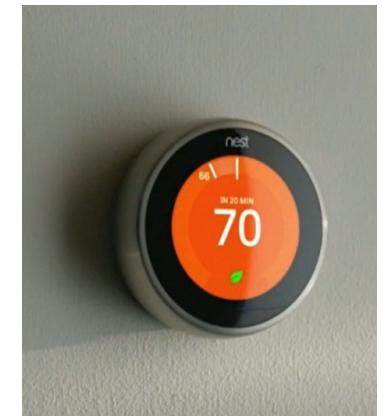


Overview



Introduction: Hand Gesture Recognition

- Gesture Recognition is the topic of computer science that interpret human gesture via Mathematical algorithms.
- Human machine Interfaces
- Camera, inertial sensors and other sensors

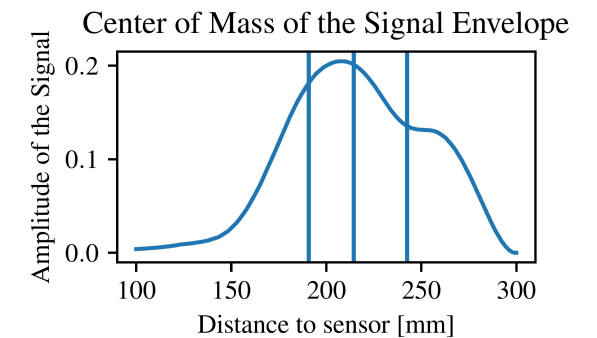
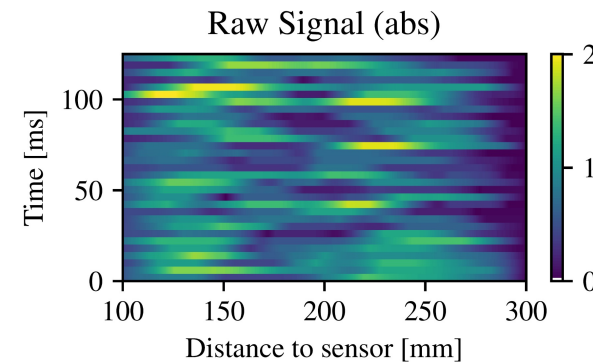
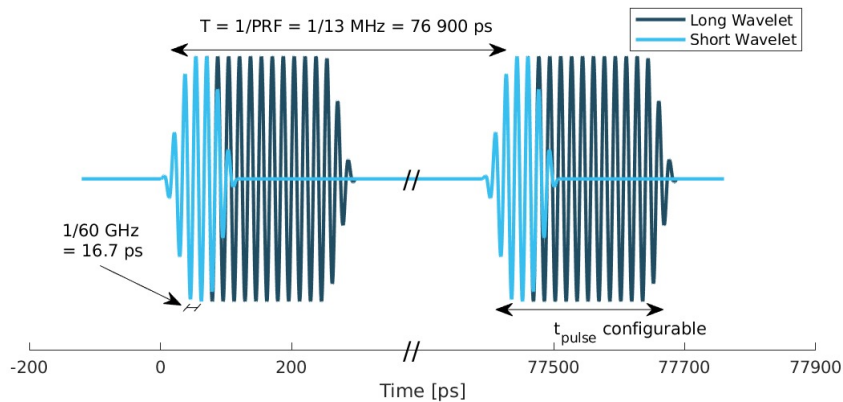
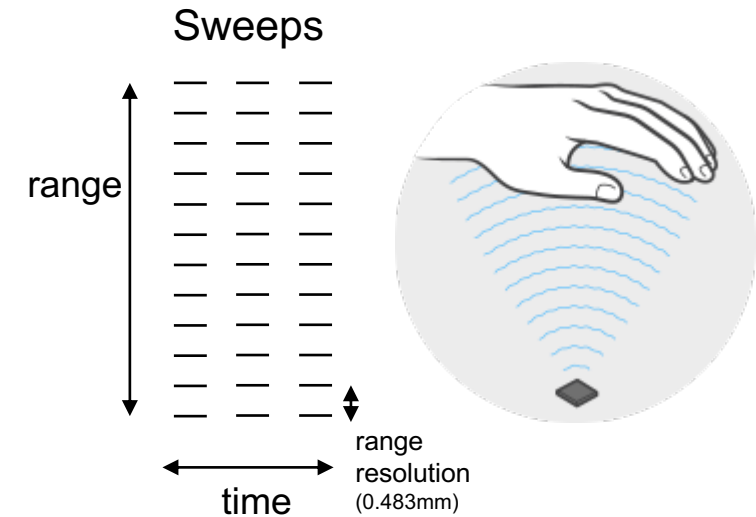


Soli: ubiquitous gesture sensing with millimeter wave radar (SIGGRAPH 2016)

Promising Technology: Short Range Radar

■ Radar

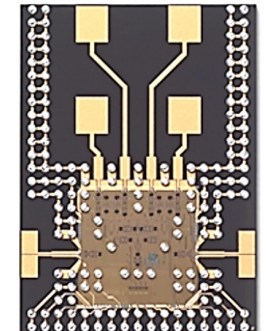
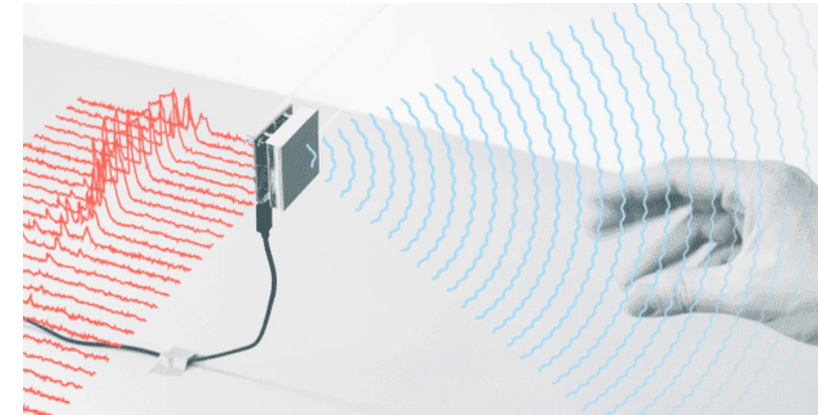
- Classical airplane detection radar: 1-12 GHz (wavelength = 2.5-30cm)
- Short-range Radar: 60GHz (wavelength = 5cm)
- Data from sensor:
 - Sweep @ 160Hz or higher → each over range
 - Time and range discrete signal $S[t, r]$



mm wavelength pulsed coherent radar, which means that it transmits radio signals in short pulses where the starting phase is well known.

Gesture Recognition Based on Radar: Google Soli Sensor

- Increasing research on radar for gesture recognition^{1,2,3,4}
- Google developed micro-radar for gesture recognition
- Good results on difficult hand-gestures: **87%** accuracy on 11 gestures and 10 people
- Why Google Soli is not tinyML?**
 - Sensor consumes too much power (**300mW**)
 - Too much data to process (**10'000 sweeps/s**)
 - Prediction model too big for embedded systems (**689MB**)
 - CNN+RNN (LSTM)⁴
 - Pixel 4 Algorithm could have different model (Realised 2020)



¹Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar, 2016

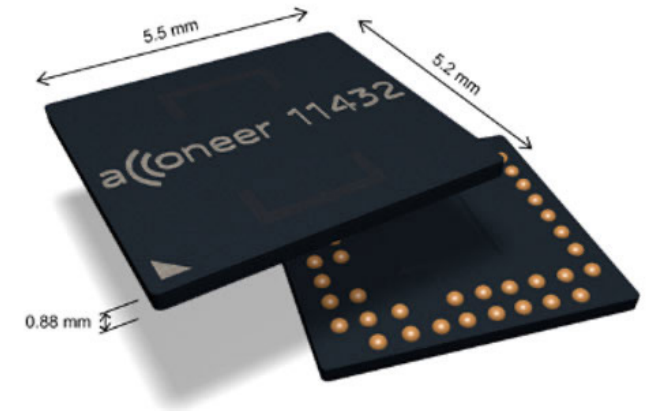
⁴Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. 2016

³Sparsity-based Dynamic Hand Gesture Recognition Using Micro-Doppler Signatures, 2017

^{4A}A Hand Gesture Recognition Sensor Using Reflected Impulses. 2017

Main contribution of this work

- **Implement radar based hand-gesture recognition in embedded system Based on TCN+CNN, below 512KiB**
- Create dataset with fine-grained hand-gestures
 - at least 1000 samples per class and 20 users
- Algorithm development suitable for embedded systems
 - less than 1MB, at least 700x smaller than I. w. Soli
- Achieving similar accuracy as I. w. Soli
 - 85% (single-user), 87% (10 people) on 11 Gestures¹
- Algorithm implementation in a mW Parallel RISC-V bases processor and experimental evaluation on efficiency (power, run-time)



¹Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. 2016

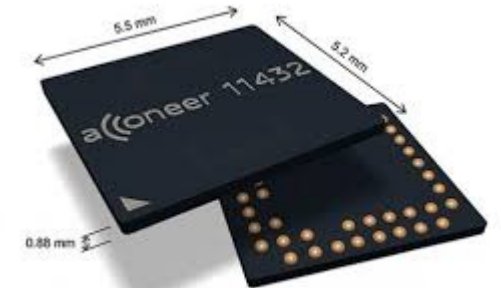
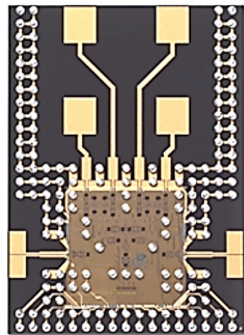
Toward TinyML and Low Power Embedded System.

We want to bring radar based gesture recognition into low embedded system

But how?

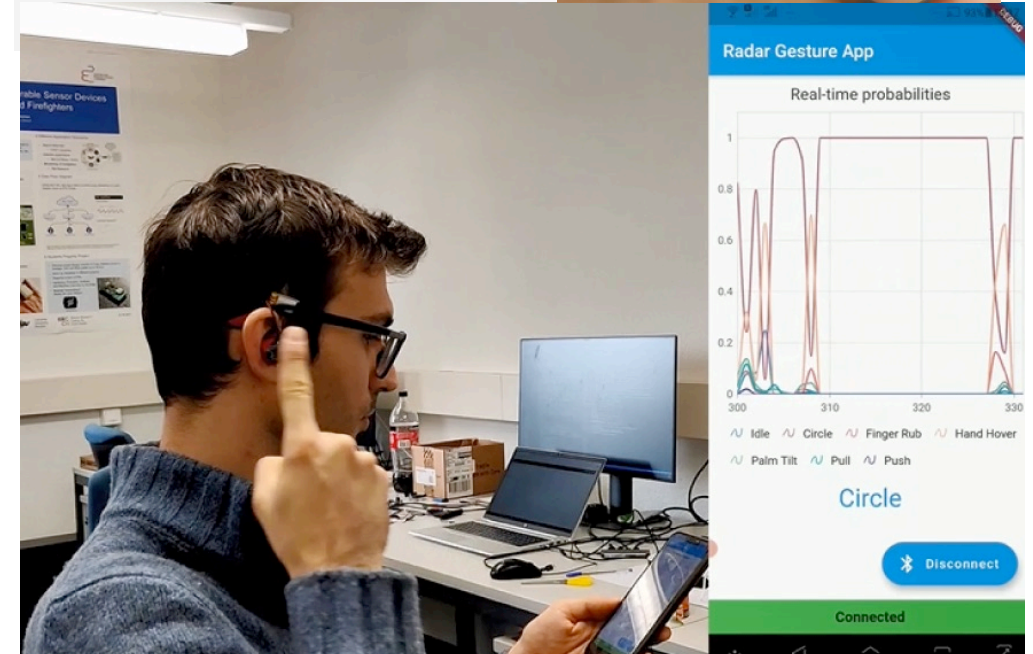
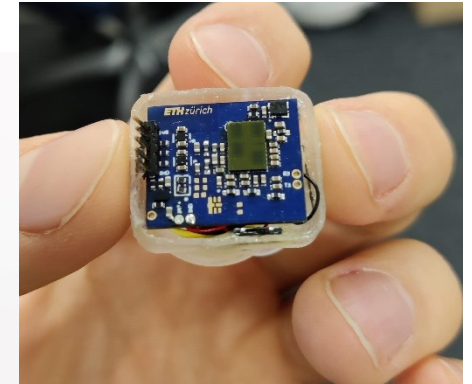
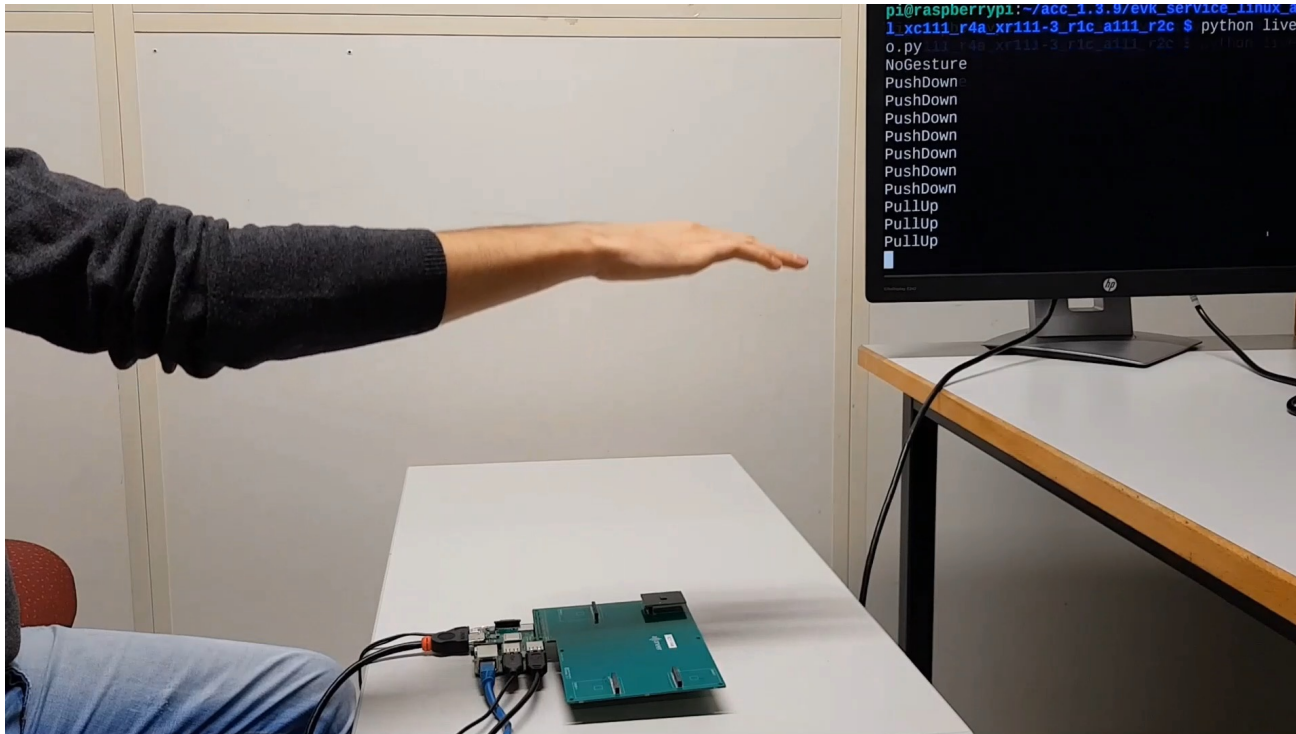
- Different sensor: **Acconeer**
 - Lower power
 - Less data:
 - lower sweep rate, fewer Tx, Rx

Sensors:	Soli	Acconeer
Carrier Frequency	60GHz	60GHz
Sweep Rate (up to)	10'000Hz	1'500Hz
Power	300mW	20mw @100Hz
Transmitters/Receivers	Tx:2, Rx:4	Tx:1, Rx:1



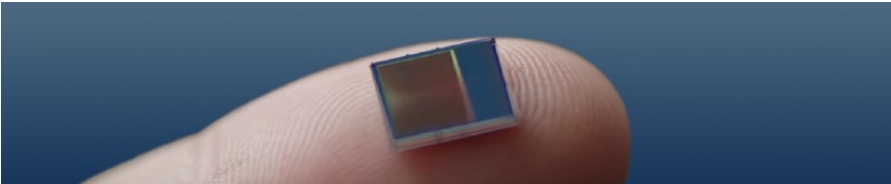
XR112

Demo fix deployment and earbuds

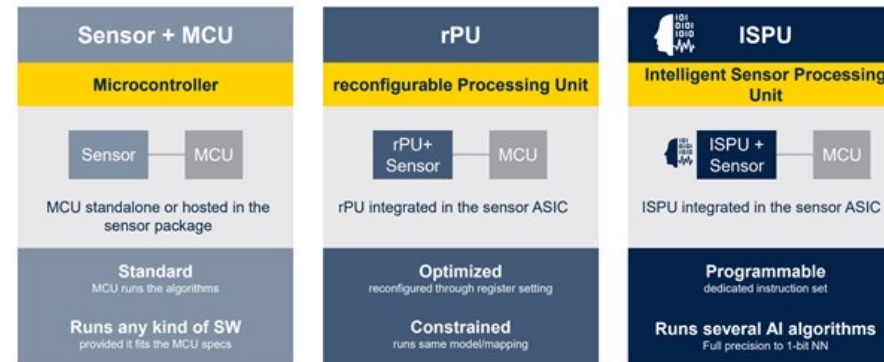
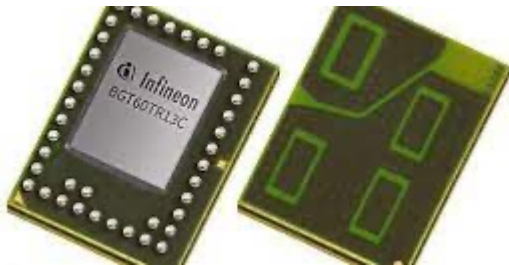


Access to novel sensors and sensors that start to include ML

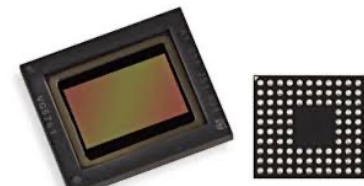
- Sony IMX-500
- ST Sensors with ISPU
 - Predictive maintenance
 - Condition monitoring etc.



- Infineon Radars

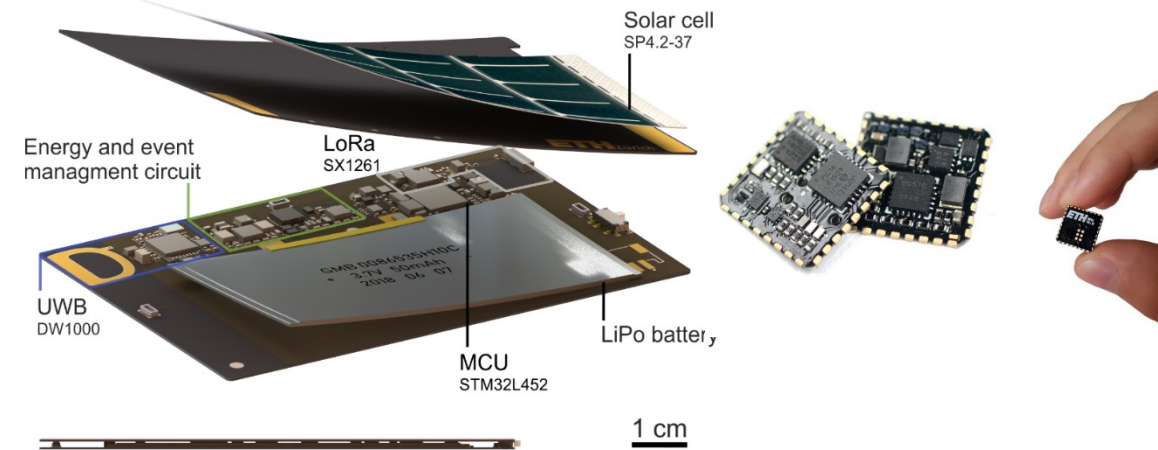
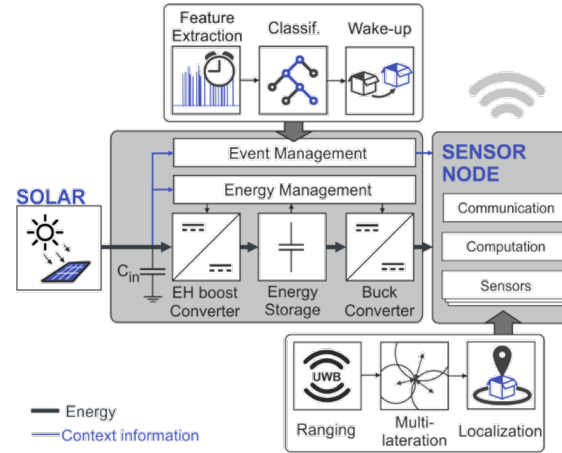
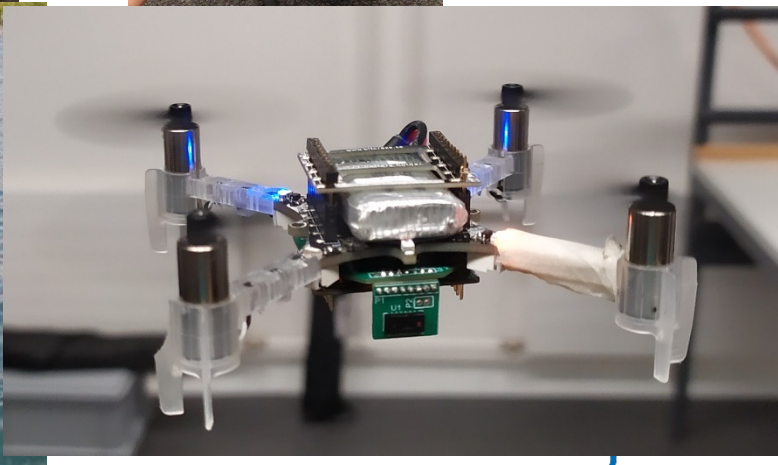


- ST ToF, Low Latency Camera and more



Many work on Tiny ML with different applications

- TinyML and Embedded Control on Robots
- Self Sustaining Smart-Sensors for Indoor and outdoor Applications and Industrial Applications



Thank you for your attention

- Edge Ai and Tiny ML is just the begin for next generation of smart IoT sensors and devices.

- Michele.magno@pbl.ee.ethz.ch