tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

TINY ML

www.tinyML.org

A Battery-Free Long-Range Wireless Smart Camera for Face Detection: An accurate benchmark of novel Edge AI platforms and milliwatt microcontrollers

Dr. Michele Magno. ETH Zurich. D-ITET Center for Project-based Learning
Credits: Marco Giordano, Philipp Mayer, Xiaying Wang.

Overview

Introduction

System Architecture

Neural Net Deploy

Experimental Benchmark

Demo

Introduction

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 and few of them are **smart**

The wave of IoT is pushing the limit of **battery-less devices and Tiny ML.**

Narrative

Forbes

Time

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

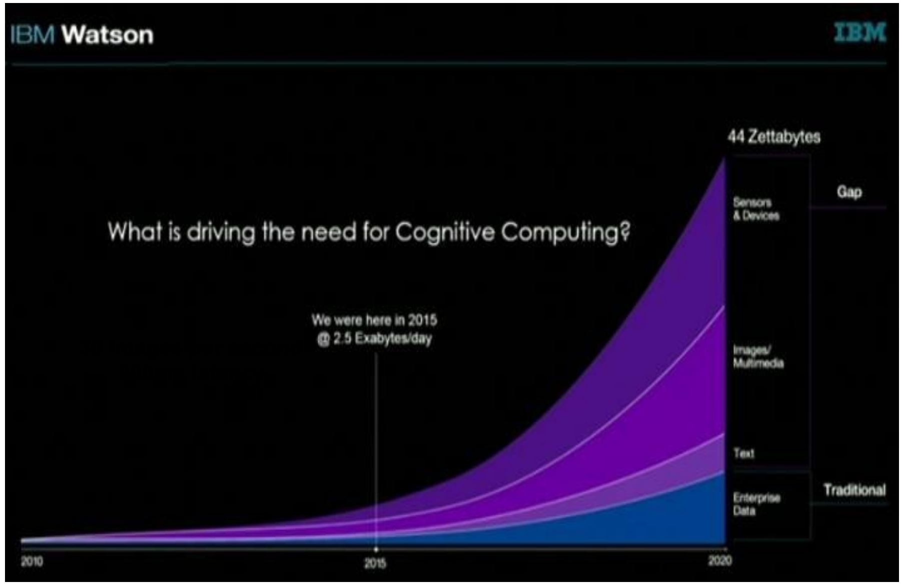
Communication latency also with 5G or other networks is in the range of hundred of milliseconds

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

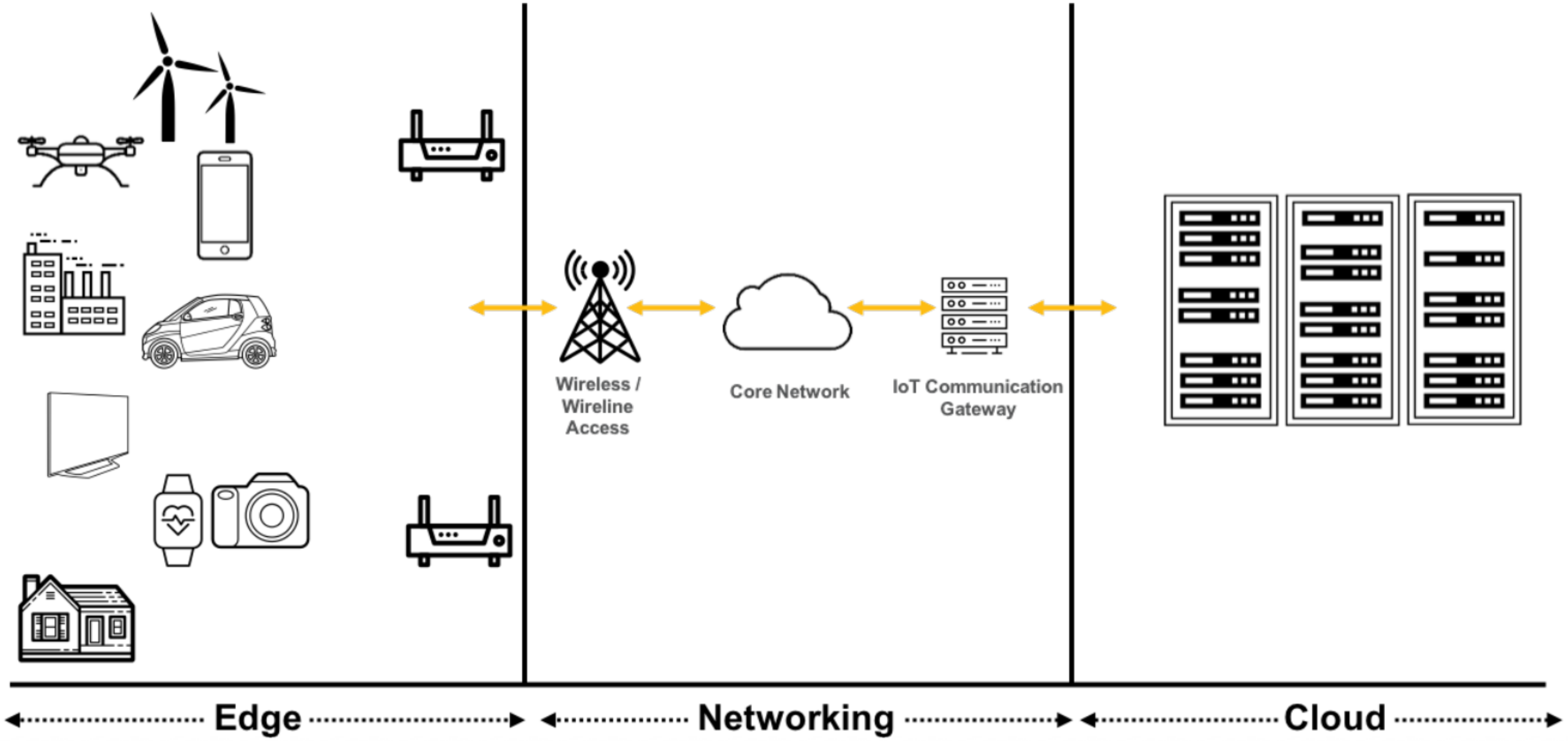


Figure reference: Accelerating Implementation of Low Power Artificial Intelligence at the Edge, A Lattice Semiconductor White Paper, November 2018

Next generation of IoT devices: 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

Introduction

System Architecture

Neural Net Deploy

Experimental Benchmark

Demo

System overview Always-on Smart Camera

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

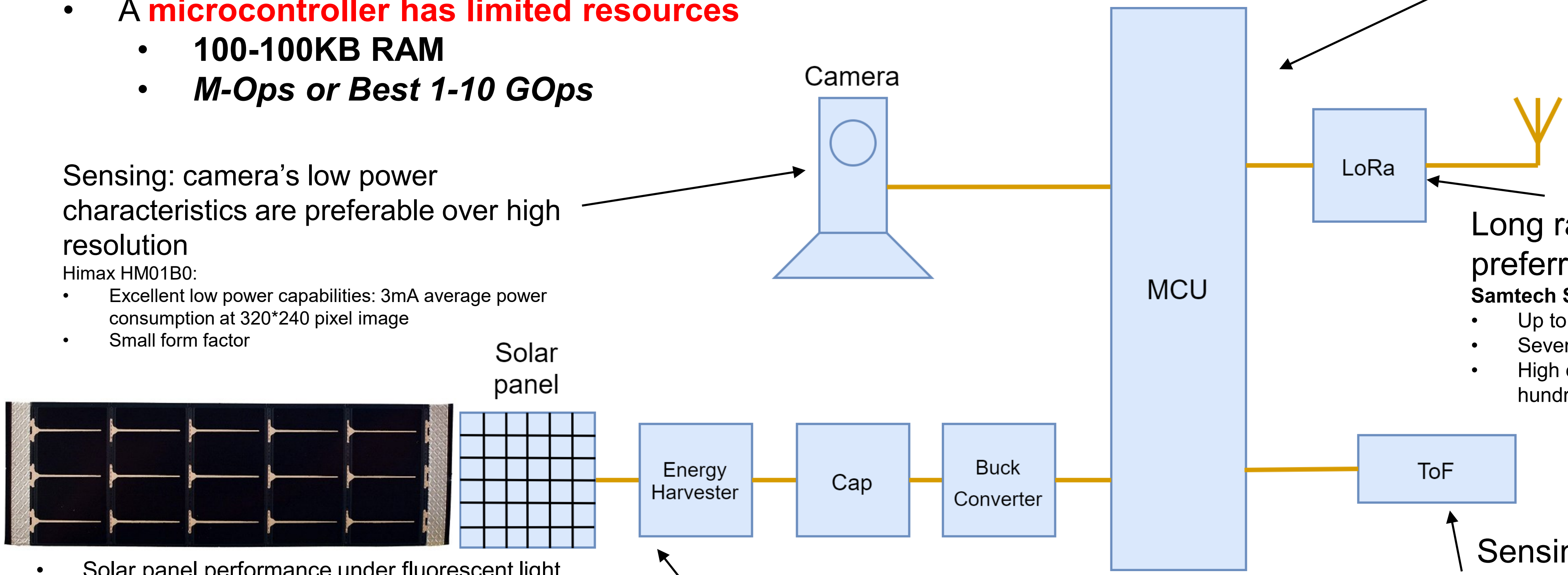
- A **microcontroller has limited resources**
  - **100-100KB RAM**
  - ***M-Ops or Best 1-10 GOps***

mW MCU: (Benchmark)

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

Himax HM01B0:

- Excellent low power capabilities: 3mA average power consumption at 320*240 pixel image
- Small form factor



- Solar panel performance under fluorescent light performs much worse than solar light:
  0.15W -> 0.3-3mW

Energy harvester:with Maximum power Point Trackig

Long range radio: low power preferred over high datarate.

Samtech SX1262:

- Up to 22dBm of transmitting power
- Several km range
- High current scalability: 30mA in TX mode, hundreds of nA in Sleep mode.

Sensing: ToF sensor used for asynchronous wake up

ST VL53L1X

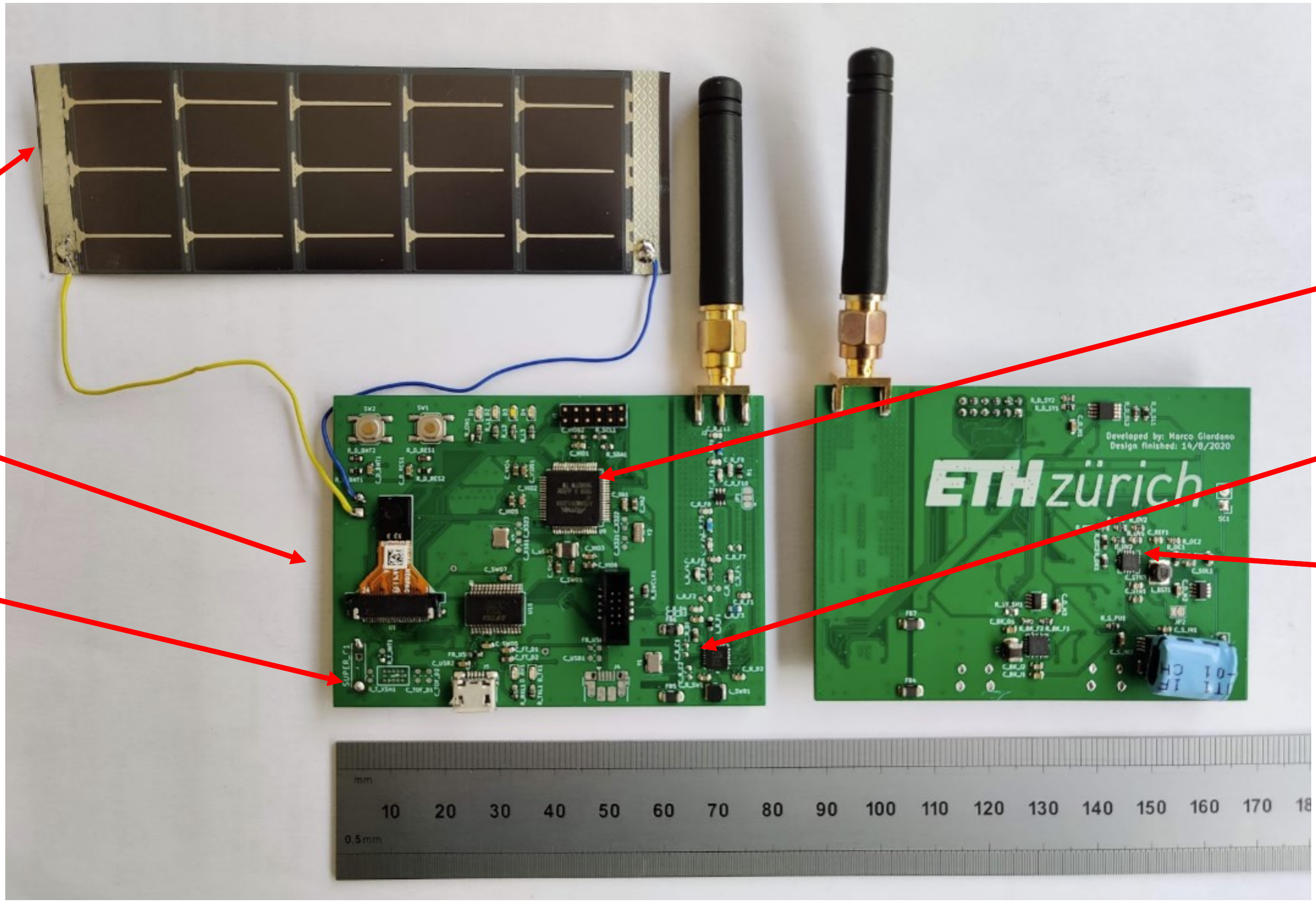
System overview – working prototype

We realized a working prototype of a small always-on system with Long Range communication.

Flexible Solar panel

Camera Himax

Time of Flight Sensor



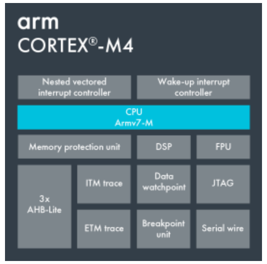
Processor

LoRa Radio SX1262

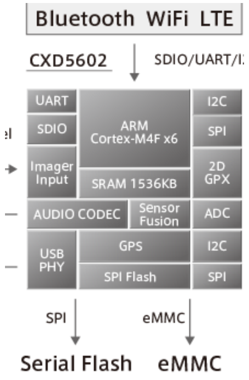
Energy Harvesting power management

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

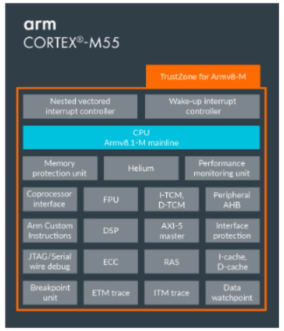
Evaluate the performance in terms of Latency, Computation power, Energy Efficiency, of below 200mW power platforms.



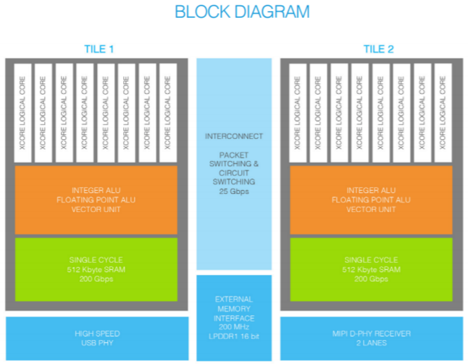
Apollo Ambiq 3.
ARM Cortex-M4F



Sony
Spresense
CXD5602
6x ARM-Cortex-M4

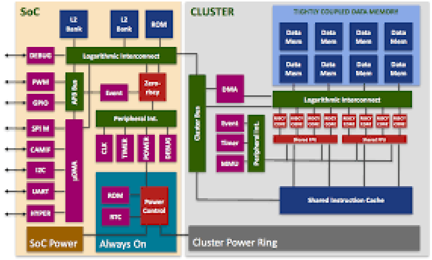


ARM Cortex M55



XMOS.AI.

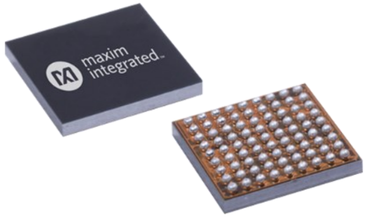
PULP
RISC-V Parallel



Mr- Wolf
GAP8

GREENWAVES TECHNOLOGIES

MAX78000



Dual Core
Arm Cortex-M4
RISC-V
NN Accelerator

Overview

Introduction

System Architecture

Neural Net Deploy

Experimental Benchmark

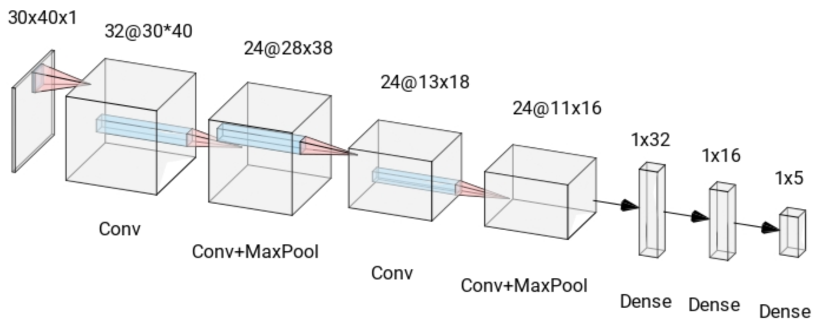
Demo

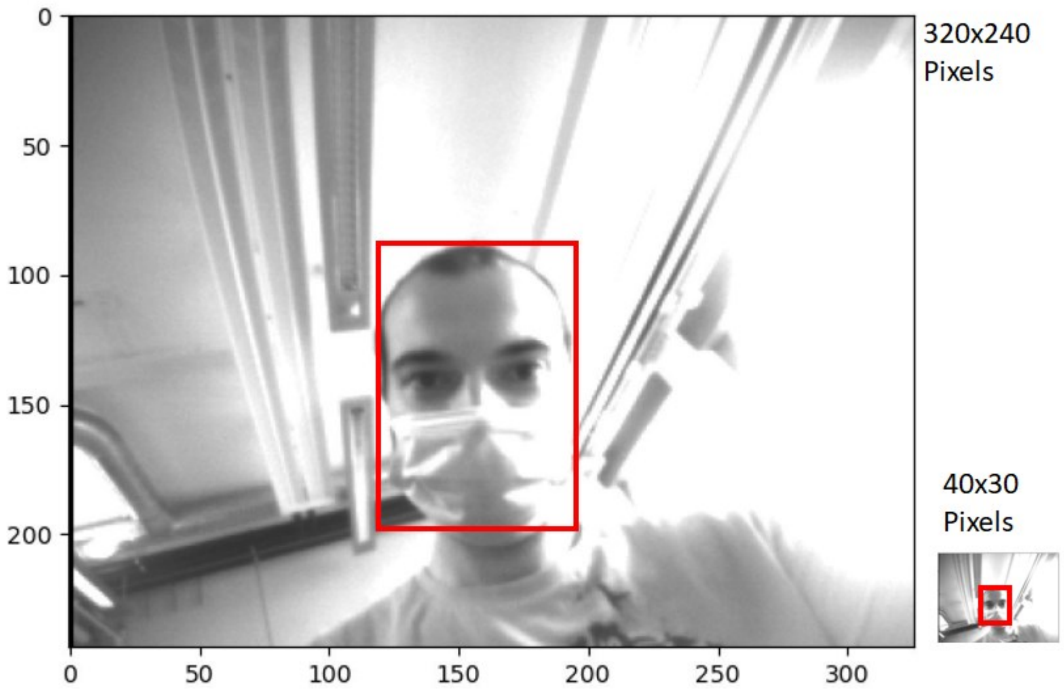
The proposed Tiny Neural Network

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

Expected Results:
- More than 95% accuracy with 5 classes
- Around 50kB memory footprint
- Less than 30mJ per inference (For self-sustainability)
- less than few hundreds ms per inference





Data augmentation

Data augmentation represented an important step towards a successful training of the neural network:

- Used the open CelebA[1] dataset as a reference
- Only 20 to 30 images per actor are provided

To overcome this shortcoming:

- Images rotation (-20, -10, 0, +10, +20 degrees)
- Exposition alteration (gamma transform coefficient: 0.1, 0.4, 1, 2.5, 5)

Benefit:

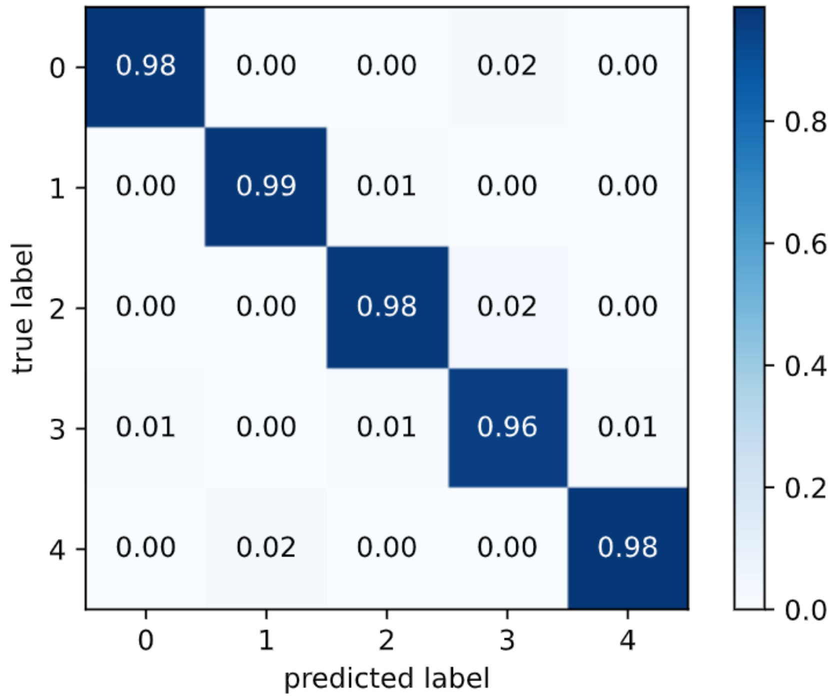
- Better simulation of possible working conditions
- Improved generalization towards subject inclined, different light conditions

Reference: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

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

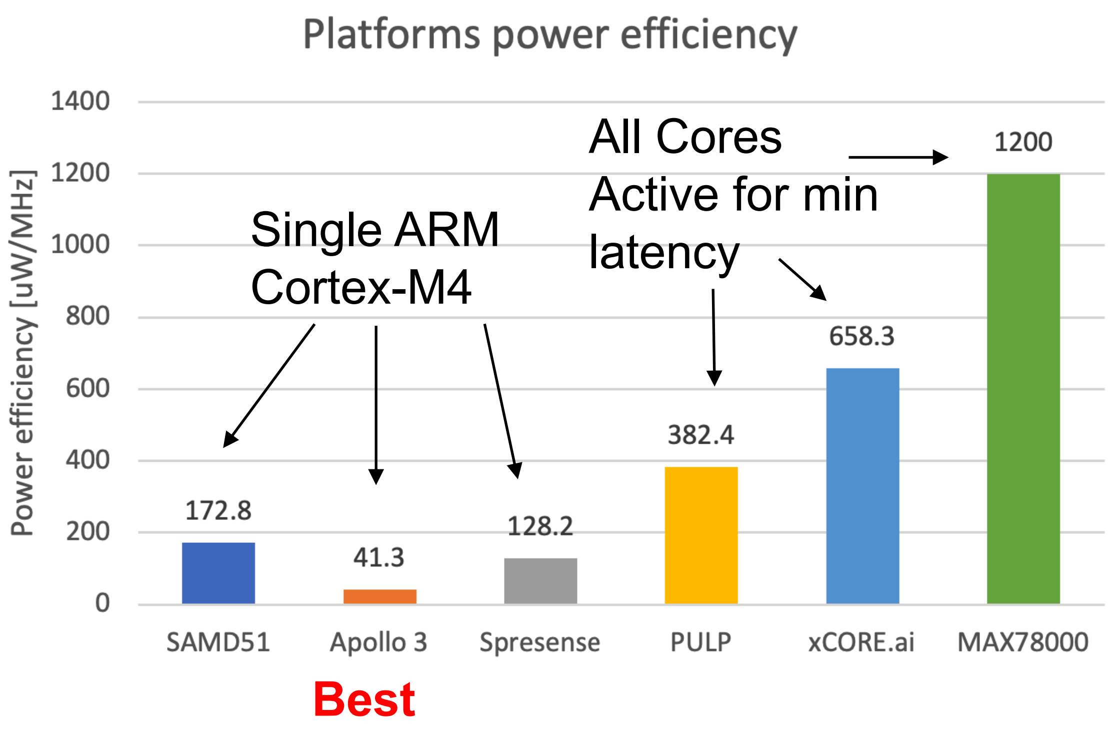
Input size/ data type	Accuracy	Precision	Recall	F1-Score
240x320, float (1)	0.97	0.97	0.97	0.97
240x320, int8 (2)	0.95	0.96	0.94	0.95
30x40, float (3)	0.97	0.97	0.97	0.97
30x40, int8 (4)	0.93	0.91	0.92	0.92

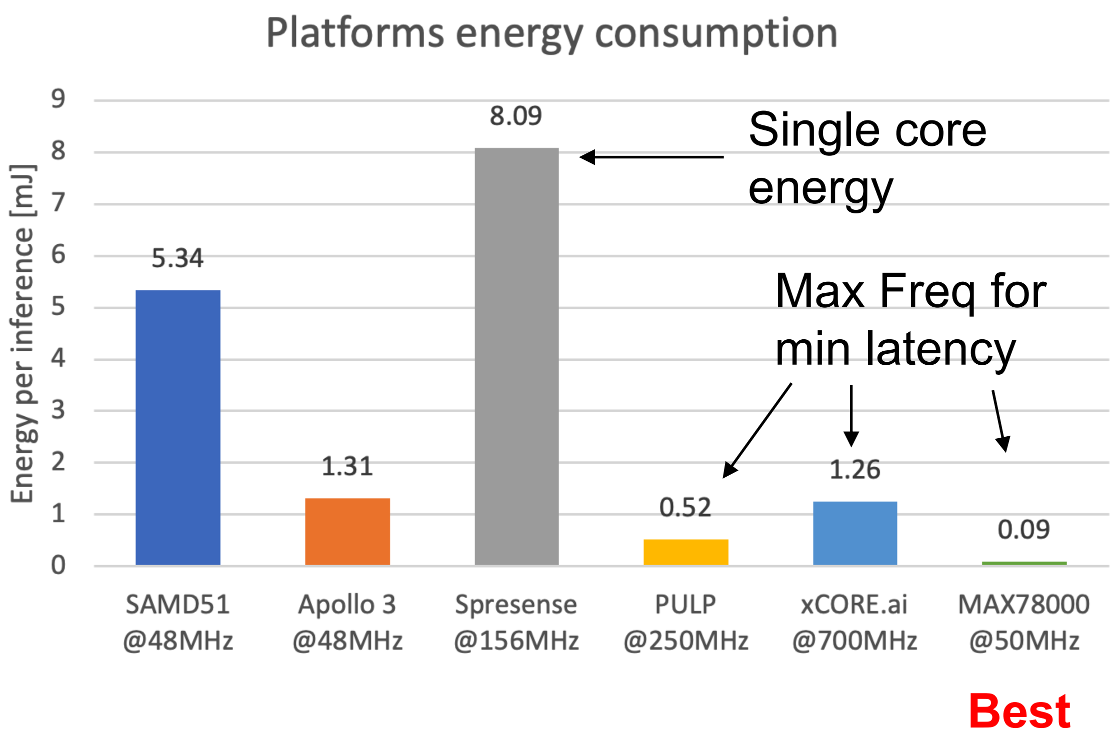


Overview

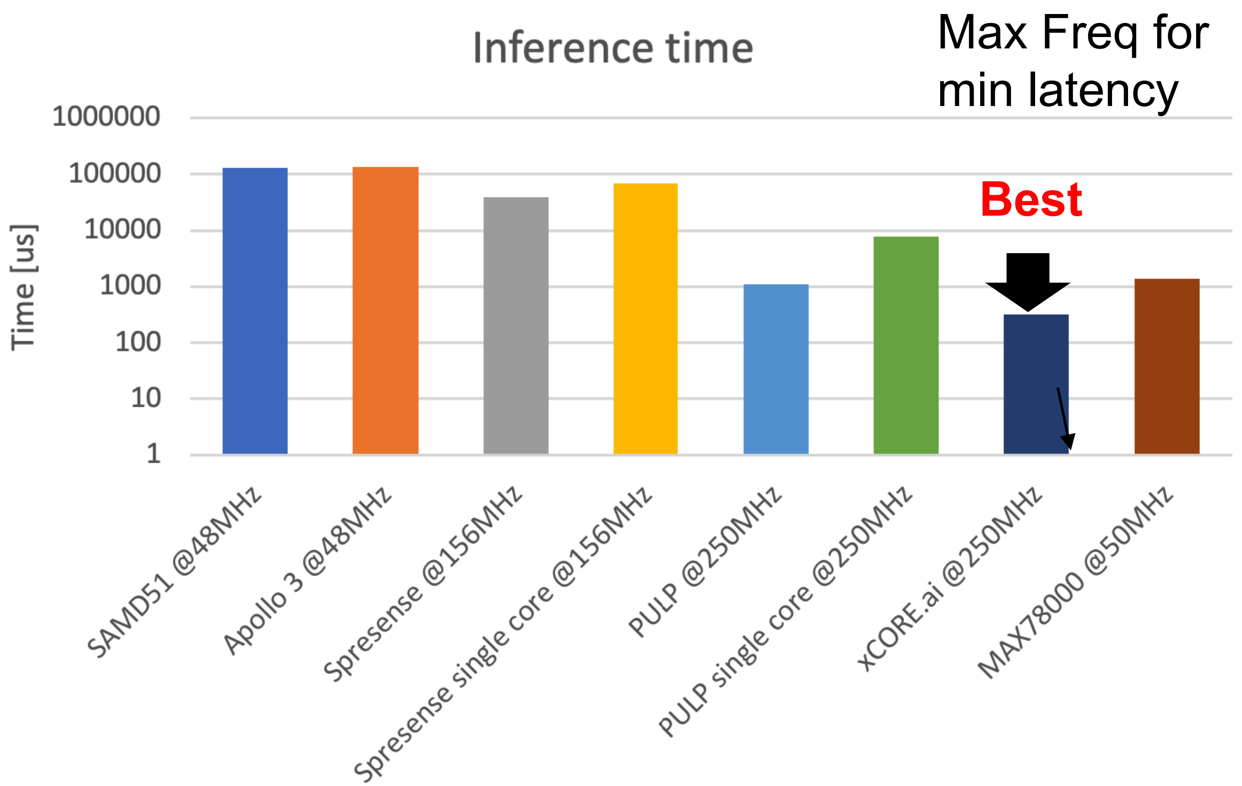
- Introduction
- System Architecture
- Neural Net Deploy
- **Experimental Benchmark**
- Demo

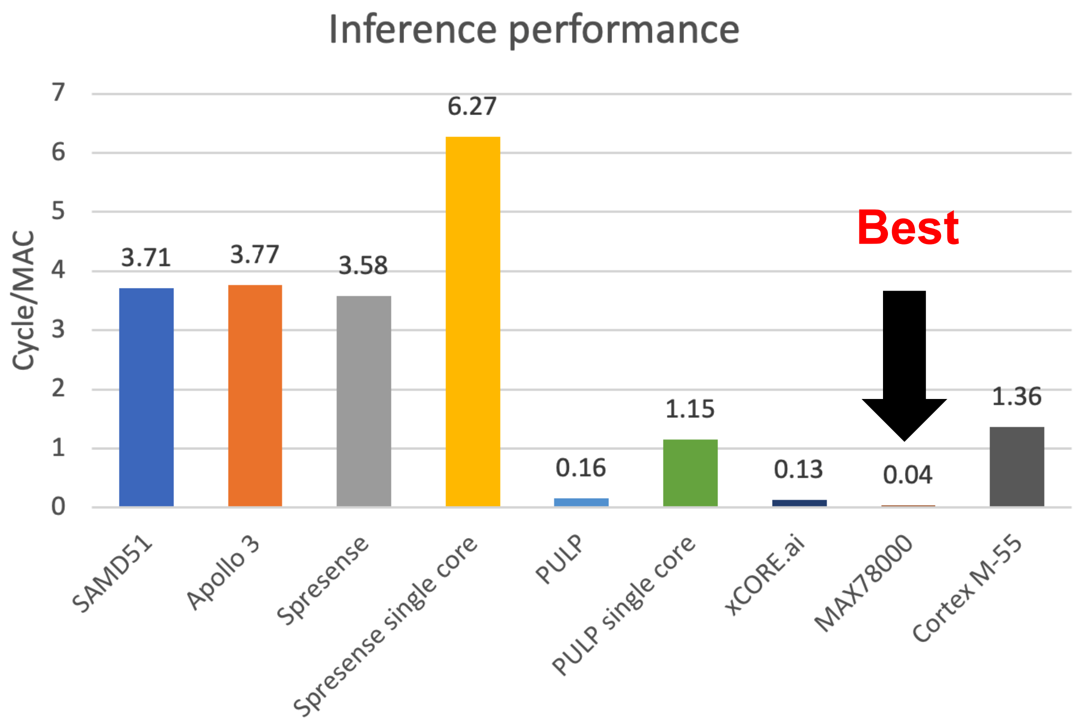
General Comparison: Power vs energy efficiency





General comparison: Computational efficiency vs min Latency





Proof-of-Concept standalone smart camera

Assumptions:

- Trigger time: once per minute.
- Battery capacity: 8.64J.
- Energy per camera image captured: 0.5 mJ.

Platform	Energy per inference (mJ)	Battery Lifetime
Apollo3	1.31	80h00’
Spresense	8.09	16h45’
PULP	0.52	140h15’
xCORE.ai	1.26	81h50’
MAX78000	0.09	244h00’

Overview

Introduction → System Architecture → Neural Net Deploy → Experimental Results → Demo

Video Demo (in preparation –just an example)

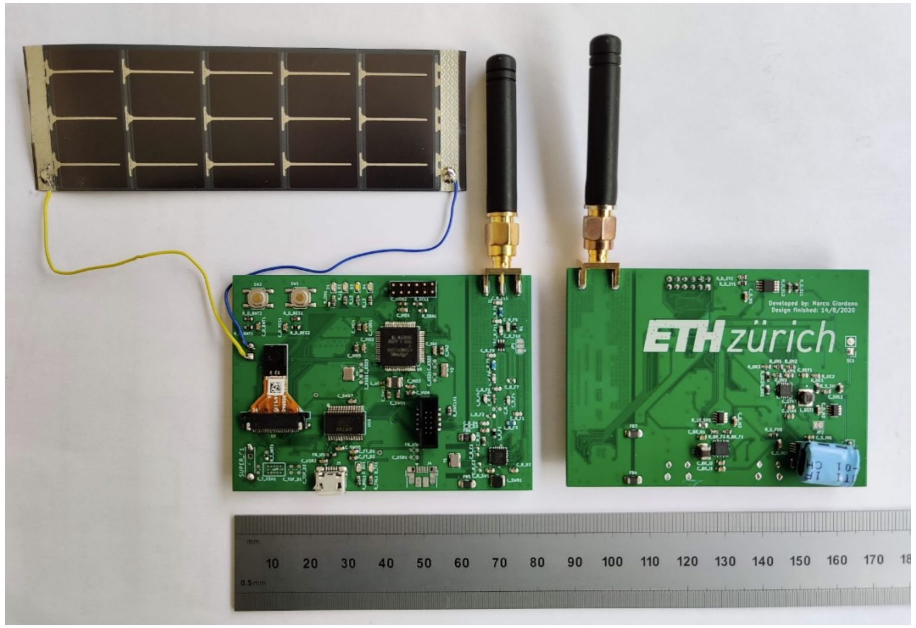
Conclusions - Battery-less long-range wireless smart camera

- Battery-less design with energy harvesting
- Tiny machine learning on the edge
- Efficient neural network for face ID
- Long range LoRa communication
- Two different implementation has been evaluated

- >95% accuracy over 5 faces
- Proposed neural network model fit in only 115kByte

- Benchmark of novel and promising processors below 100mW

- This work has been accepted in ENSSys workshop: **A Battery-Free Long-Range Wireless Smart Camera for Face Detection**



Giordano, Marco, Philipp Mayer, and Michele Magno. "A Battery-Free Long-Range Wireless Smart Camera for Face Detection." *Proceedings of the 8th International Workshop on Energy Harvesting and Energy-Neutral Sensing Systems*. 2020.

Thank you for your attention!

Premier Sponsor

Intelligent Agent
Neuton

Automated TinyML

Zero-code SaaS solution

Create tiny models, ready for embedding, in just a few clicks!

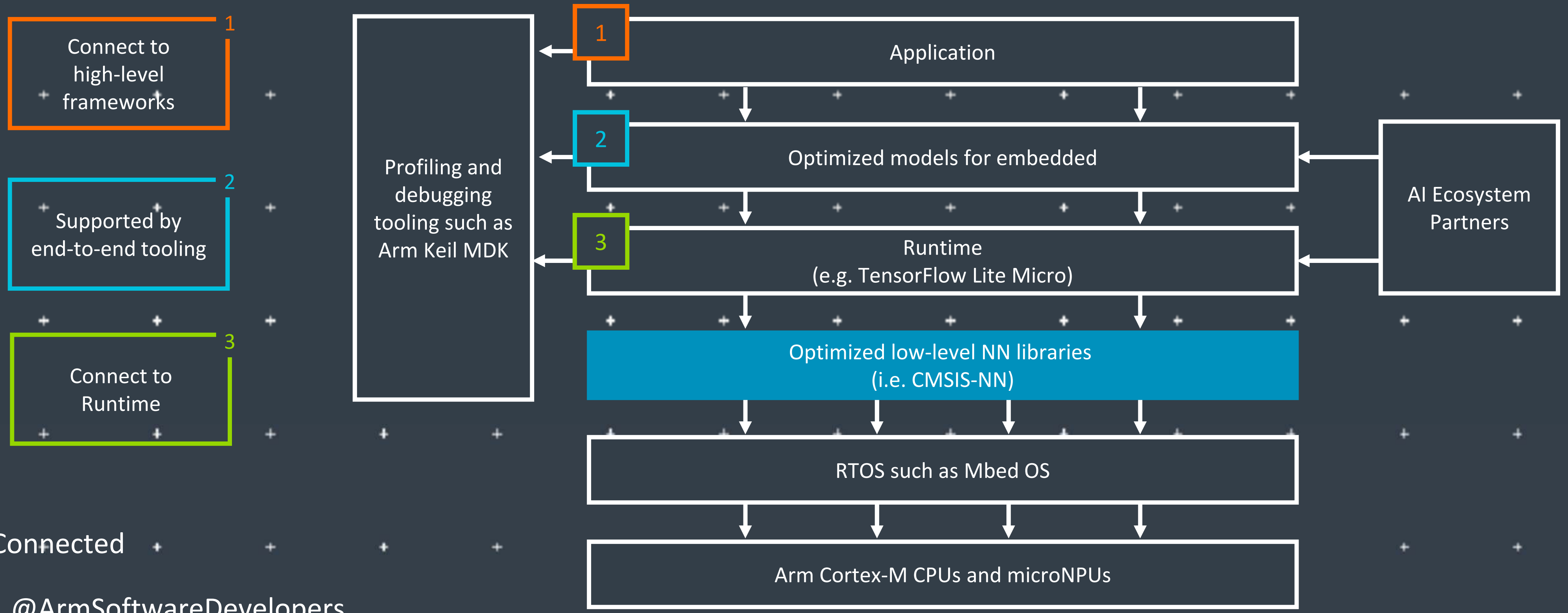
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.

TINY ML EMEA
Virtual
JUNE 7-10, 2021

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

TinyML for all developers

C++ library

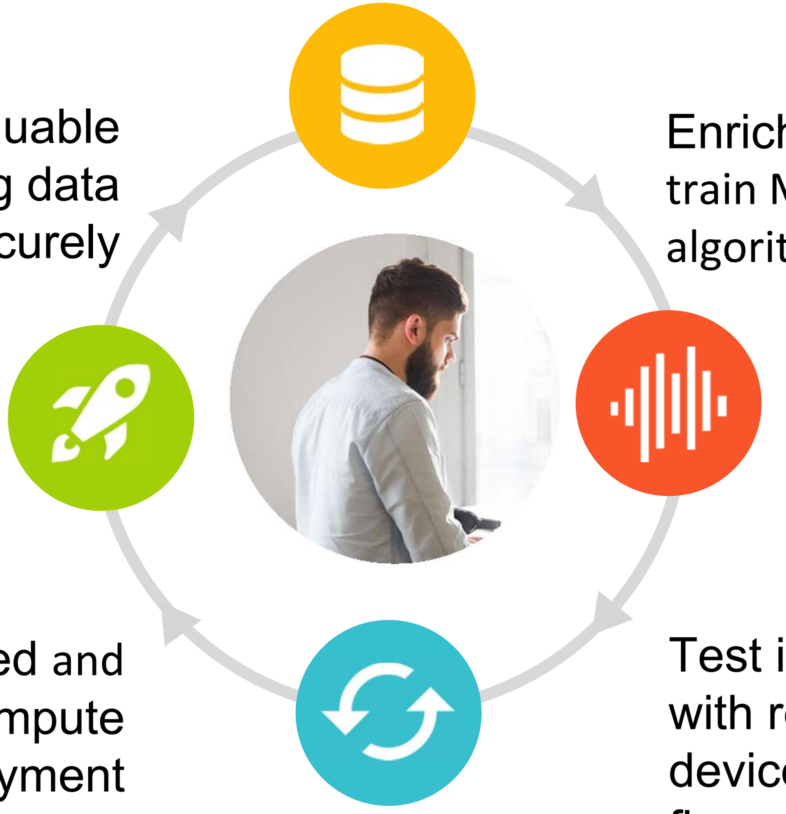
Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms



Edge Device

Real sensors in real time

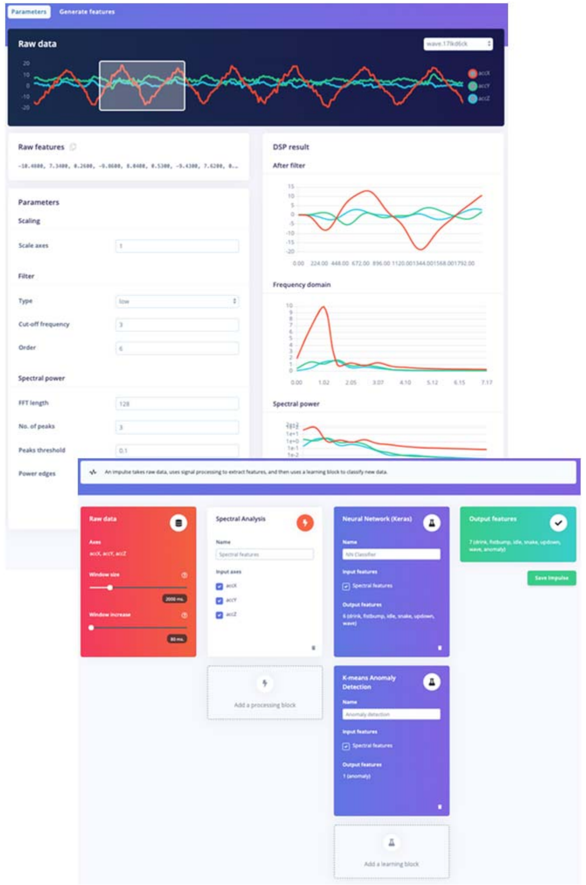
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



www.edgeimpulse.com

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

SYNTIANT

Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp

TINY ML EMEA
Virtual
JUNE 7-10, 2021

Platinum Sponsors

Part of your life. Part of tomorrow.

www.infineon.com

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai   info@reality.ai   @SensorAI   Reality AI

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions

- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars "see with sound"

Reality AI Tools® software

- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

TINY ML EMEA
Virtual
JUNE 7-10, 2021

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

SensiML™

Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

sensiml.com

TINY ML EMEA
Virtual
JUNE 7-10,2021

Silver Sponsors

emza
visual sense

GREENWAVES
TECHNOLOGIES

HOTG

imagimob

Qeexo

seeed studio
The IoT Hardware Enabler

ST
life.augmented

Copyright Notice

www.tinyML.org