SILICON LABS

works with
BY SILICON LABS

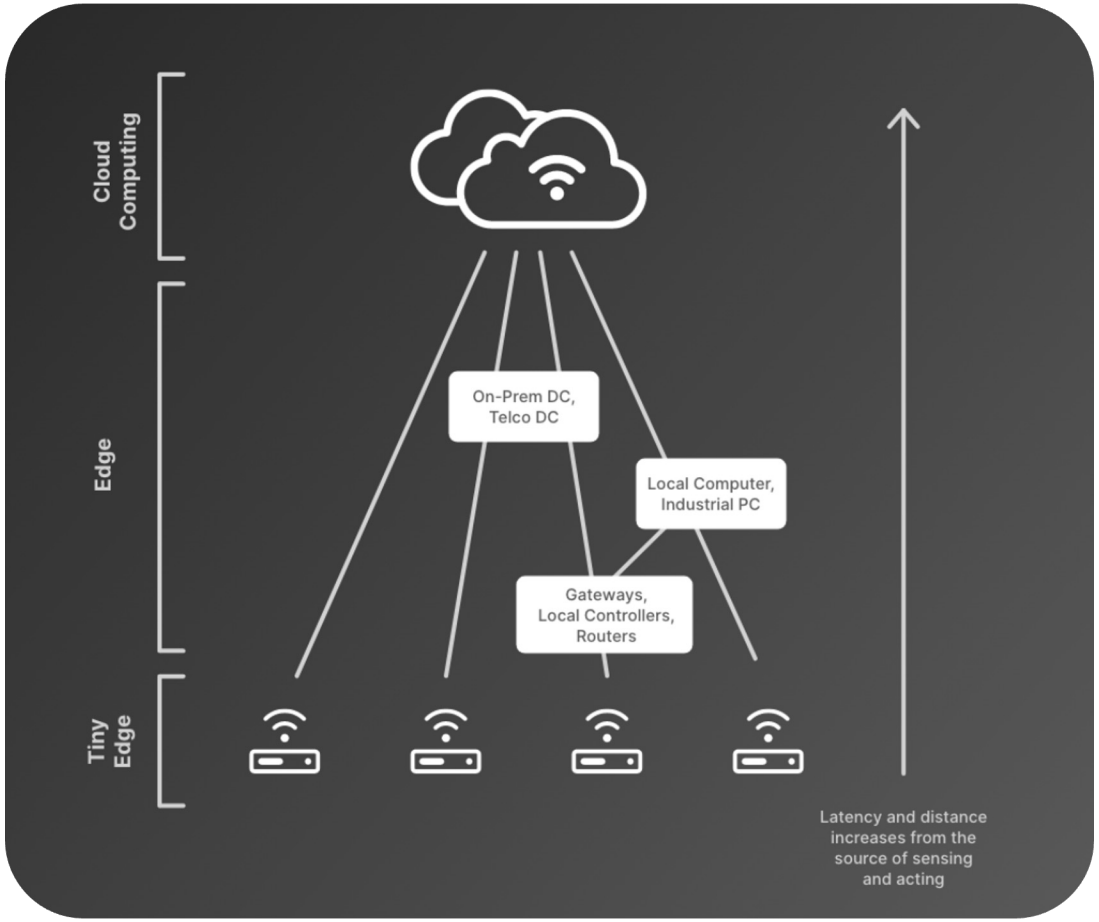
IOT - 103

IoT Trends: Embedded ML in Your Edge Devices

IoT

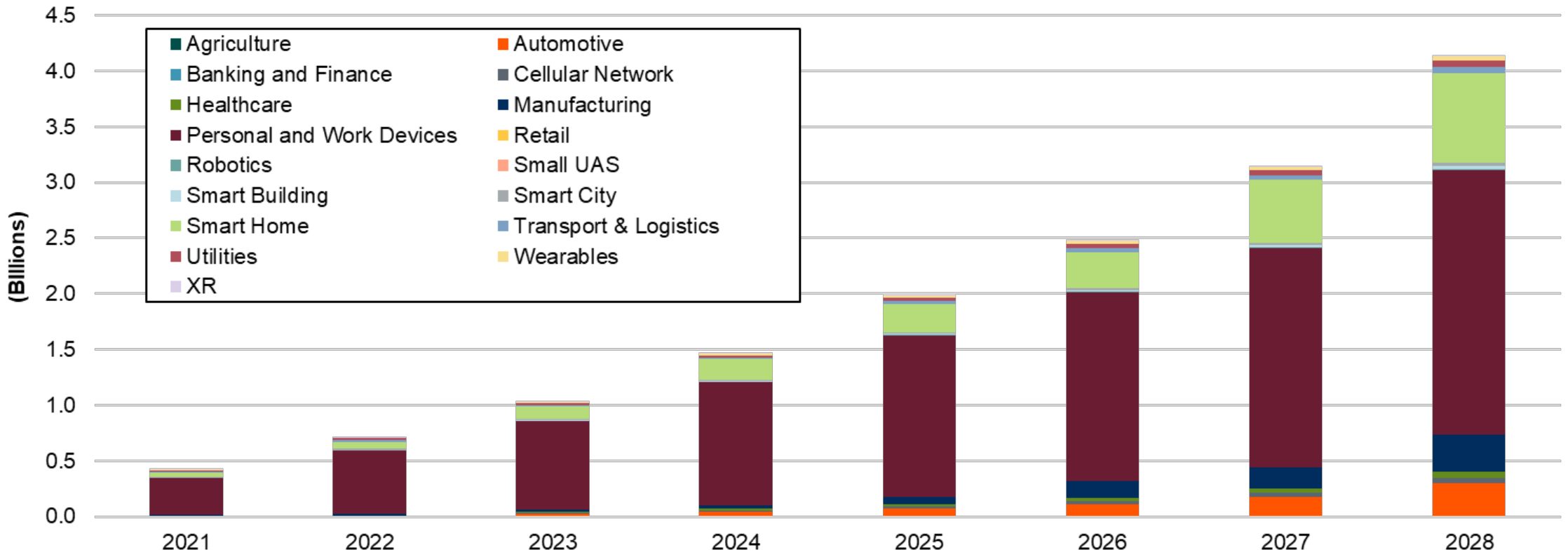
Tamas Daranyi | August 2023

Embedded ML - TinyML



- **Intelligence is moving down to the 'edge'** to keep processing as close as possible to the source of data
- **Edge AI** refers to any device with the capabilities to process AI workload
- **Tiny Edge** – sensor class devices or small application processors, typically Cortex M class (M33, M4 or equivalent/below) run TinyML applications
- **TinyML Characteristics**:
  - Ultra-low power applications and systems (<mW range)
  - Decent inference speed but more focused on efficiency
  - Algorithms, networks, and models down to < hundreds of kB
  - Efficient ML accelerators
  - Low cost, optimized, well integrated systems

An Emerging Market – Significant Growth YoY



TinyML Device Shipment to Exceed 4 Billion by 2028, 10+ Billion by 2030

Source: ABI Research, 2023

Why Machine Learning on Embedded Devices?

Low Latency Required

- Mission or safety-critical applications require real-time reactions
- Large data to process - typically at vision use cases - no time to upload to anywhere to process

Privacy and IP Protection, Security

- Data never leaves the sensing device, only inference result/metadata is transferred
- Less sensitive data to transmit, less chance to be hacked
- Protecting IP

Bandwidth Constraints

- Long range, low power, and slow networks can't transfer all Time Series data to process somewhere else
- Overloading of mesh network is an issue
- Large data to chunk e.g. hi-res images

Offline Mode Operation

- Local system keeps operating standalone in case of any network issue
- Connectivity is occasional or blocked by admin

Cost Reduction

- Network and infrastructure costs
- Data ingestion costs
- Data storage costs
- Cloud services
- Ops, maintenance
- Compact edge with ML solutions integrated to wireless SoC
- Cheaper devices

Power constraints

- Ultra-low power applications
- Always-on systems
- Healthy tradeoff in transmit to higher level compute vs. locally process

Embedded ML Trends

Check **AIML – 101:** Ensure First Time Success of Edge ML Applications

SensiML

NEW APPLICATIONS

TinyML opened new horizons in embedded application space

(MCU class devices)

→ Advanced data processing

→ Audio Voice detection

→ Vision

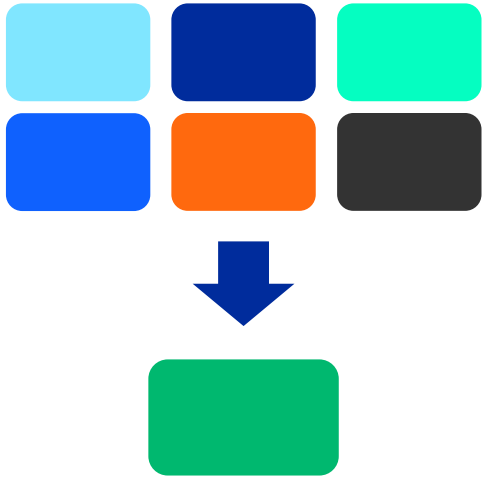
EFFICIENCY

Ultra low-power applications

Strong requirements on battery life

→ Always-on System Topology

→ HW Accelerated Machine Learning



SIMPLIFICATION

Reducing system complexity and cost

Compact size is a must (e.g. wearables)

→ Application chip and radio modem into one combined solution

→ Multi-radio, multi-protocol wireless in one single SoC

→ ML in sensor class devices

TIME TO MARKET

IOT market is growing.
Strong competition among ODMs.
Having market differentiation is key.

→ Rising of turnkey solution providers in ML space

→ More and more integrated ML toolchains and growing ecosystems

→ Tools to Vertical solutions

Industry’s Answers for the TinyML Space Growth

AI/ML on Silicon Labs' Wireless SoCs

EFR32 Series 2 and Wi-Fi SoCs

Higher Performance Platform
- ARM Cortex M33 (78 MHz)
- Improved radio performance
- Lower power (MCU active, TX/RX)

Improved Security
- Secure Vault - Mid
- Secure Vault - High (select OPNs)

Acceleration - MVP
- AI/ML acceleration
- Faster AoA/AoD calculation
- Math library (matrix and vector ops)

AI Software
- TensorFlow Lite for Microcontrollers with accelerated kernels in GSDK
- 3rd Party end-to-end tools

All Series 2 SoCs support ML

SILICON LABS MG24

78MHz CortexM33
AI/ML accelerator
1.5MB / 256kB
2.4 GHz radio
20 dBm TX Power
Secure Vault
Low power

EFR32FG28:
Launch this month!

SILICON LABS SiWx917

Wi Fi 6

180MHz CortexM4
160 MHz NWP
AI/ML accelerator
Up to 8MB / 672kB
2.4 GHz radio
21 dBm TX Power
PSA L2 Security
Low power

xG24-DK2601B Developer kit

Broad Range of Sensors
- 9-axis Inertial Sensor
- 2 Digital Microphones
- PIR sensor
- Pressure Sensor
- Relative Humidity and Temperature Sensor
- UV and Ambient Light Sensor
- Hall-effect Sensor

Ready to demonstrate ML
- Sample applications in GSDK
- Examples on GitHub
- Examples and tutorials in MLTK
- Many sample applications and demos from partners
- Plug&Play Sensor extensions with Sparkfun Qwiic connector

sparkfun ELECTRONICS

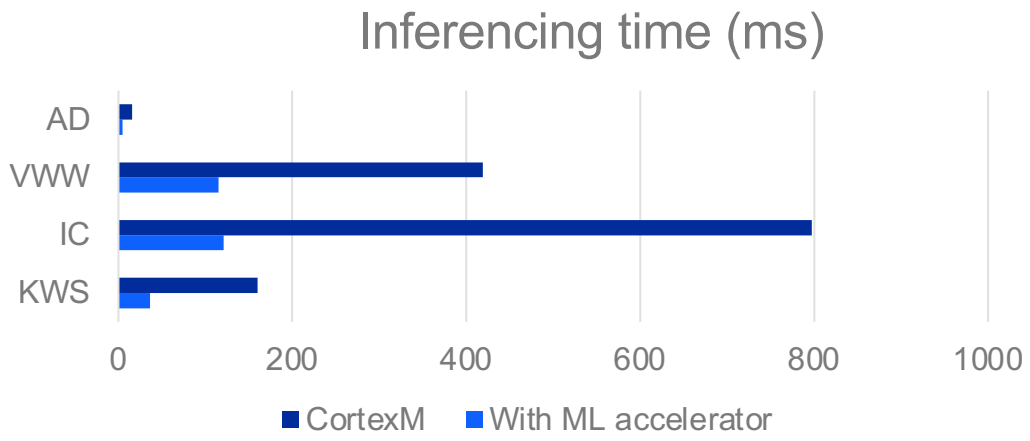
TensorFlow Lite

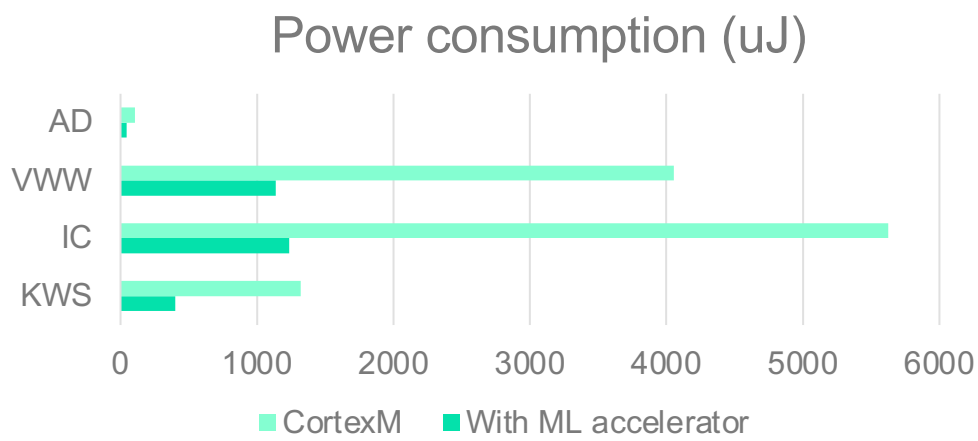
Common Machine Learning software and tools on our Wireless SoC portfolio

Benefits of the MVP Hardware Accelerator

- Dedicated **ML computing subsystem** next to the CPU: Matrix Vector Processor (MVP)
- Optimized MVP to accelerate ML inferencing with a lot of processing power **offloading the CPU**
- **Up to 8x faster** inferencing over Cortex-M
- Up to **6x lower power** for inferencing

Performance data with ML hardware accelerator vs. pure SW on CortexM*



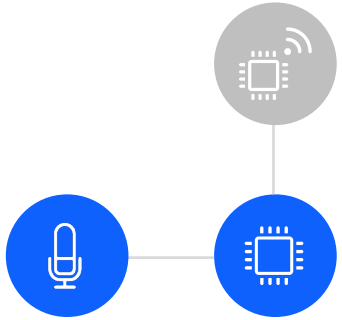


*Standardized performance benchmark validated by independent benchmarking body **MLCommons.org**. Published in MLPerf Tiny v1.0. Results are for inferencing only (not for the complete application).

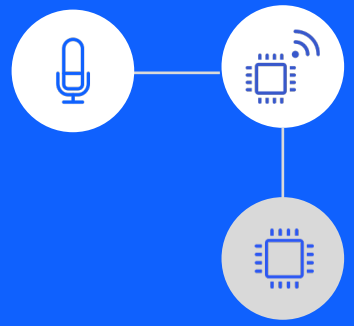
ML Commons

‘Always-on’ System Topology

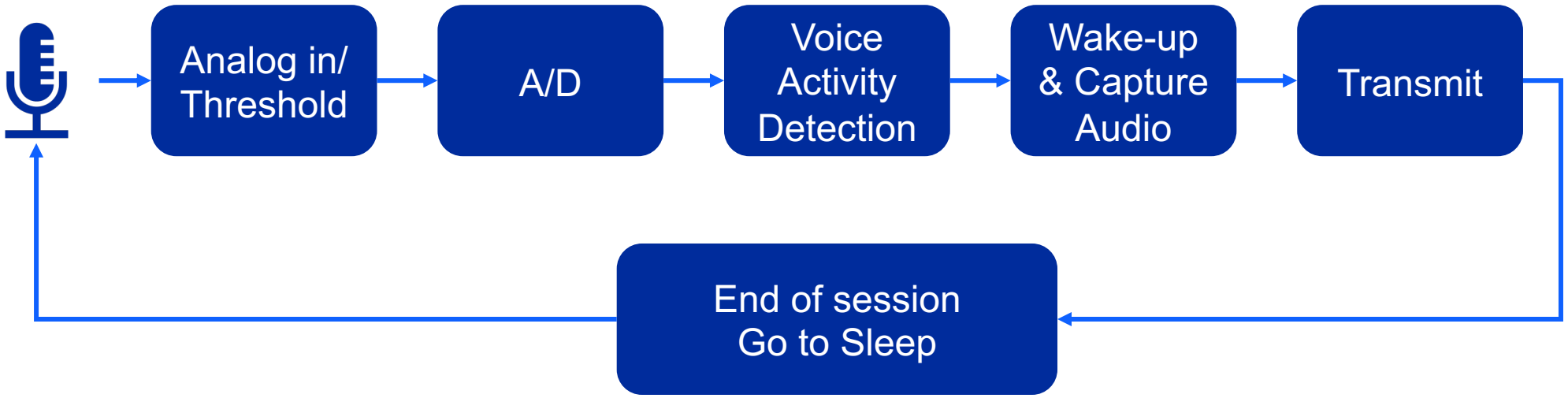
Frequent wake-up of application CPU



Application CPU can sleep more



Always-On Audio



Wake Word Detection

- Sensory's ML Wake-word detection optimized for always-on system topology
- Analog / digital mic, Audio Front-end
- EFR32MG24 Wireless SoC on xG24 DevKit
- Plug and play demo with Sensory VoiceHub available on Silicon Labs's GitHub

- Smart Lock with voice id authentication and activity detection on xG24_DK2601B; Lock opens only for the dedicated person, ignores everyone else
- Analog / Digital mic with AFE
- TFLite micro ML application, Leveraging internal accelerator
- Check SensiML & Silicon Labs voice ID demo video post from EmbeddedWorld23

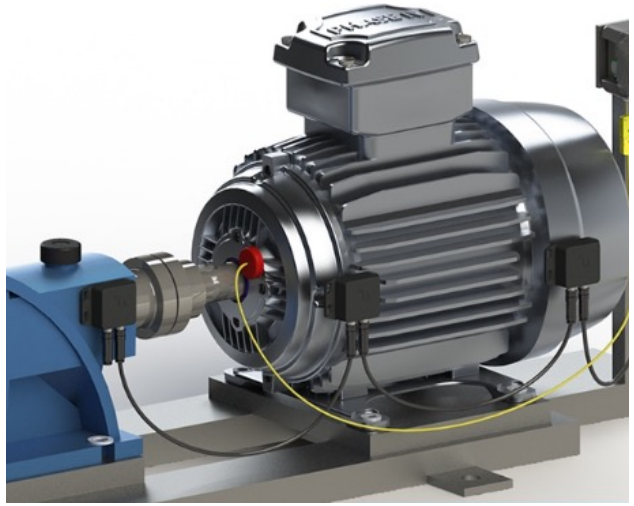
Various ML Applications from partner AIZIP

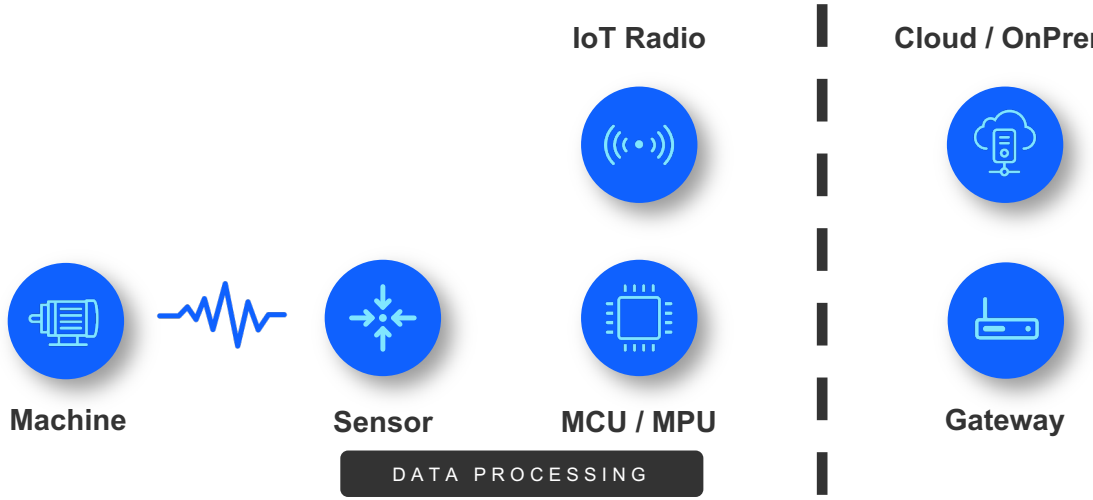
- **AIZIP is a ML solution provider offering wide range of deep neural networks (DNN) with superior performance, running on EFR32 parts**
- **Turnkey Glass Break solution on MG24 available**

Customer case study: Smart Circuit Breaker with anomaly detection

- Wireless smart circuit breaker
- Continuous monitoring of voltage and current signal and predict safety critical events to act earlier than with traditional method (e.g. arc fault detection)
- Speed and accuracy is important – below 4 msec response time – acceleration is a must have
- All in one wireless SoC

Predictive Maintenance & Condition Monitoring Applications





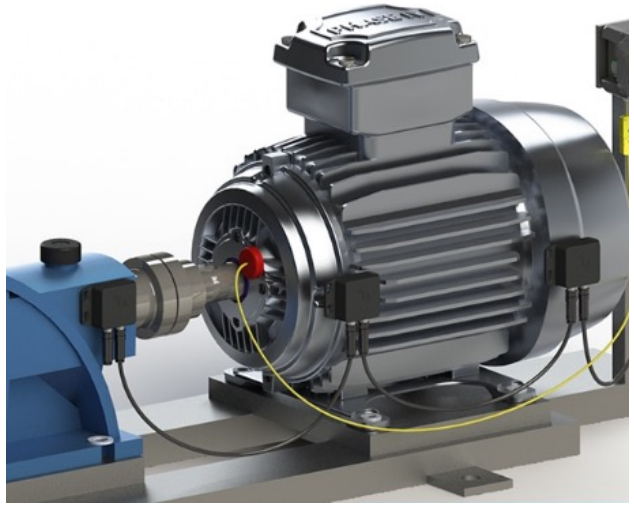
Industries

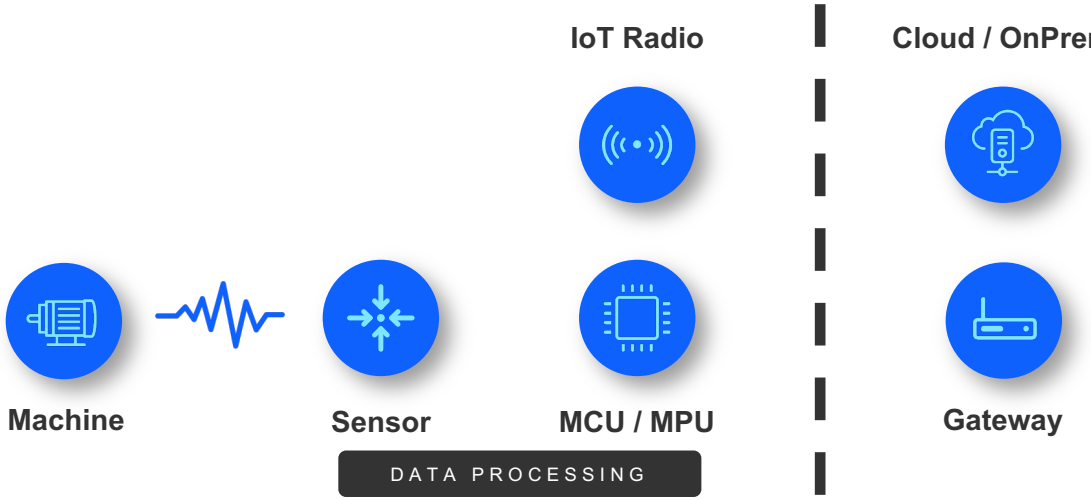
- Chemical and Petroleum refinement
- Water filtration
- Pulp & Paper processing
- Power Plants
- Manufacturing & Warehousing
- Rail, Shipping & Logistics
- Construction & Farming vehicles
- HVAC & Refrigeration

Applications

- Motor & electrical drives
- Factory machinery/tool vibration
- Valves and pressure sensors and pumps
- Noise detection from bearings
- Heat measurement of lubricant/fluids

Predictive Maintenance & Condition Monitoring Applications





Sign up for other WorksWith sessions to learn more about the topic!

AIML-102: *Machine Health and Condition Monitoring using Edge Impulse*

EDGE IMPULSE

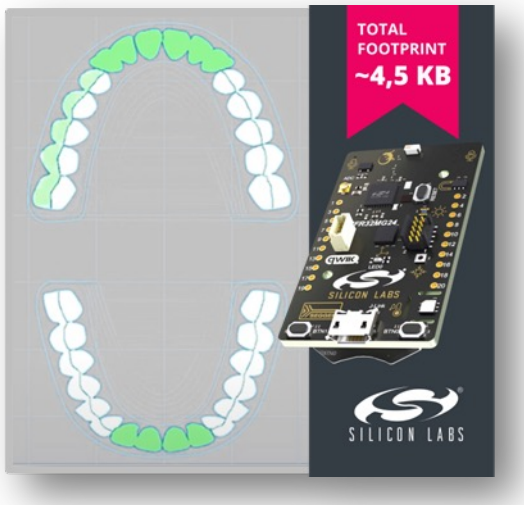
AIML-103: *Machine Learning Techniques for Predictive Maintenance*

SILICON LABS

Ultra low footprint IMU use cases by partner NeutonAI

Intelligent Agent Neuton

Teeth-brushing Tracking Solution



- The solution can identify which specific area of your oral cavity was being cleaned
- Classes: 11
- Accuracy: >97%
- Inference Time: **<2 ms**
- Total footprint: **~4,5 KB**

Box Shipment Tracking Solution

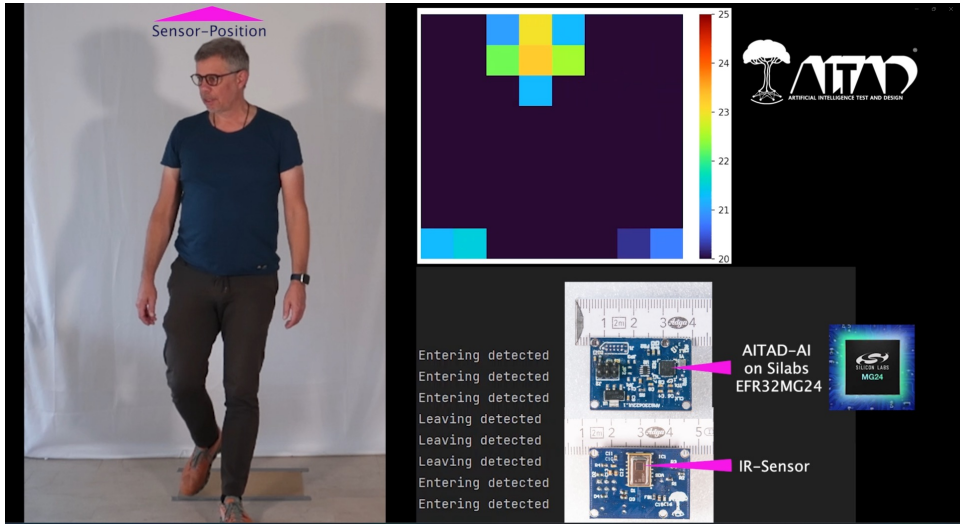
- Recognizes the following box states: moving, stationary, impacted, thrown, picked, placed in the wrong position and unknown class.
- Accuracy: >96,3 %
- Inference Time: **<1 ms**
- Total footprint: **~3 KB**

Gesture Recognition Control for Smart Watch

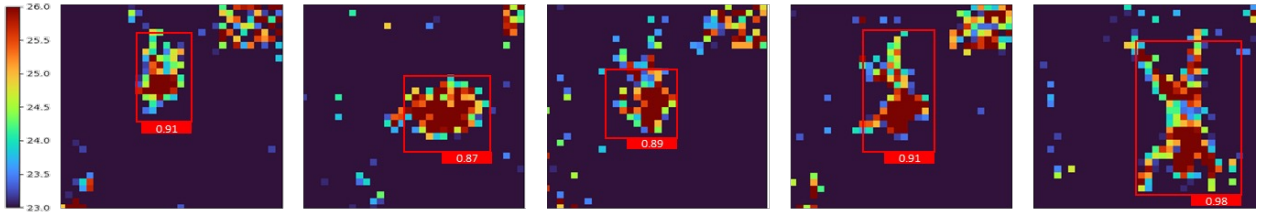
- Allows to control a watch by recognizing the following gestures: pinch, double pinch, clench, double clench and unknown class.
- Accuracy: >97,6 %
- Inference Time: **<1 ms**
- Total footprint: **~3,24 KB**

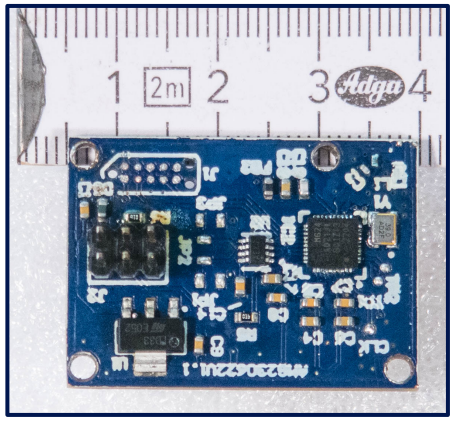
ML Vision: IR camera product by partner AITAD

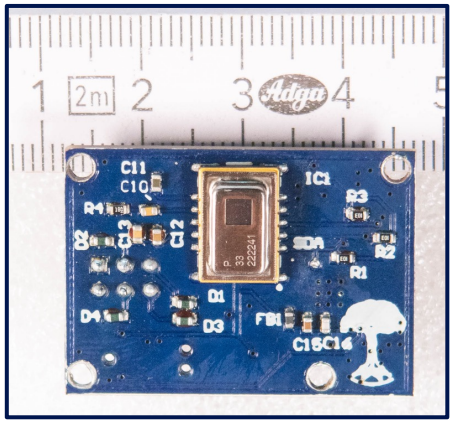
- Certified solution partner
- Use case: AI based wireless IR-person detection for showers on EFR32MG24
  - Reasonable price
  - Wireless, self-sufficient and powerful
  - Unique AITAD embedded AI process
  - Low power and low-resolution IR sensor
    Scalability (from 8x8 to higher pixels / angels)



Safety Fall detection use case, Mounted on the ceiling





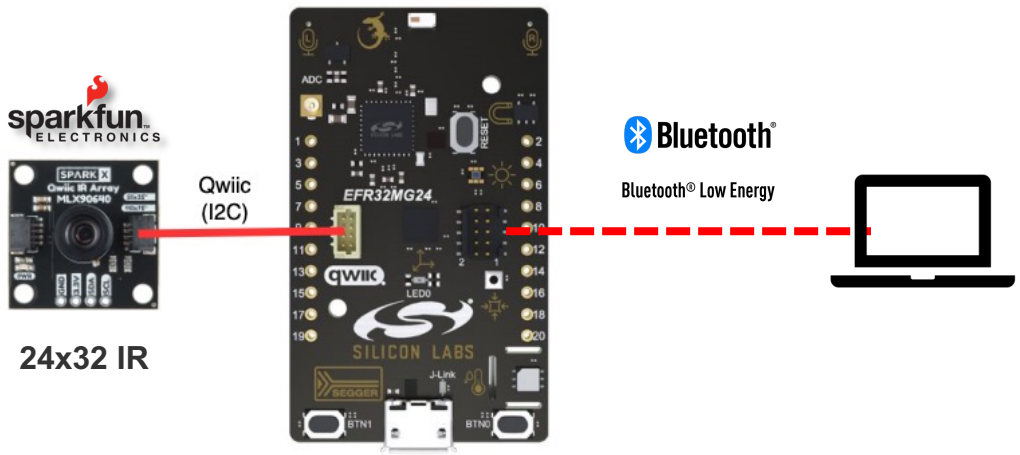


embeddedaward2023
WINNER Artificial Intelligence

& Won National Innovation Award in Germany

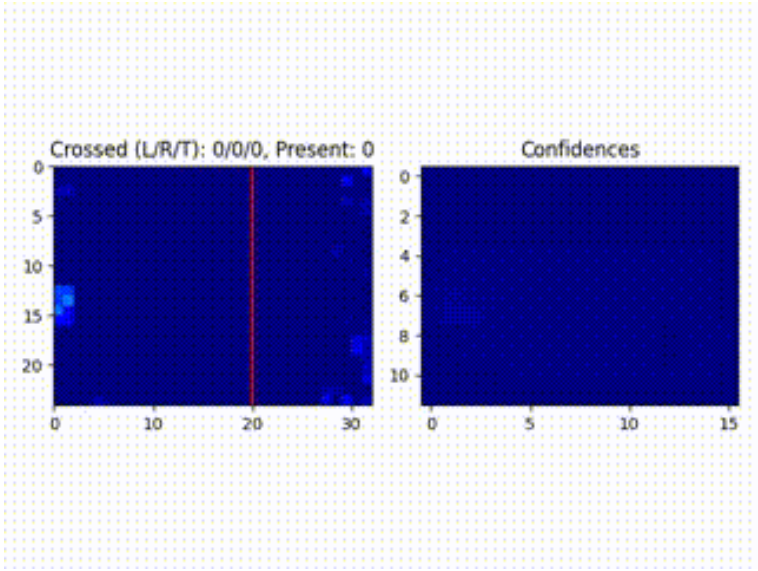
New Sample Application: People-flow counting

- **Machine Learning based People counting app available on Silicon Labs GitHub**
- **Bi-directional people flow counting**
- **Without any privacy issues**
- **Real time operation with BLE interface to the PC**
- **HW accelerated TF Lite Micro ML algorithm**



24x32 IR

xG24_DK2601B




```
Name: people_flow_counter          Clock Rate: 78.0MHz
Accelerator: MVP                   Energy: 416.6 uJ
Input Shape: 1x24x32x1             J/Op: 118.5p
Input Data Type: float32           J/MAC: 245.0p
Output Shape: 1x12x16x6            InferenceTime: 17.1msec
Output Data Type: float32          Ops/s: 205.7M
Flash, Model File Size: 26.5kB     MACs/s: 99.5M
RAM, Runtime Memory Size: 66.6kB   Inference/s: 58.5
Operation Count: 3.5M
Multiply-Accumulate Count: 1.7M
Layer Count: 23
Unsupported Layer Count: 0
Accelerator Cycle Count: 1.0M
CPU Cycle Count: 433.3k
CPU Utilization: 32.5%
```


Learn more about TinyML

- WorksWith ‘23
  - AIML-101: Ensure First Time Success of Edge ML Applications
  - AIML-102: Machine Health and Condition Monitoring using Edge Impulse
  - AIML-103: Machine Learning Techniques for Predictive Maintenance
  - AIML-201: ML Development Tools & Building and Voice Application
- www.silabs.com/AI-ML
- Silicon Labs’s tech talks, webinars
- TinyML Foundation
  - Live Events
  - Online forums
  - YouTube Webinars

SILICON LABS

works with
BY SILICON LABS

IOT - 103

Thank you!

IoT