



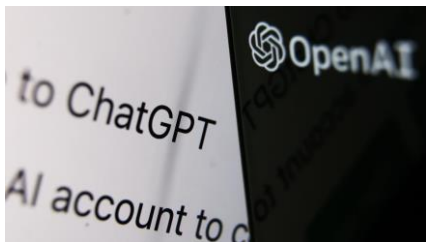
Deploy Your Embedded Vision Solution on Any Processor Using Edge Impulse

Amir Sherman

Senior Director Global Business
Development Semiconductors & Eco
Partners

Edge Impulse

From cloud AI to edge AI/ML to endpoint AI



Cameras are everywhere

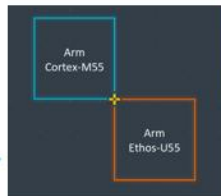
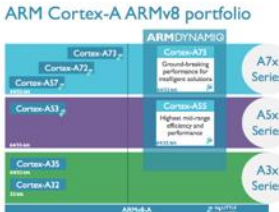


What is the "best/right" technology ?

Neural Networks



SYNOPSYS®

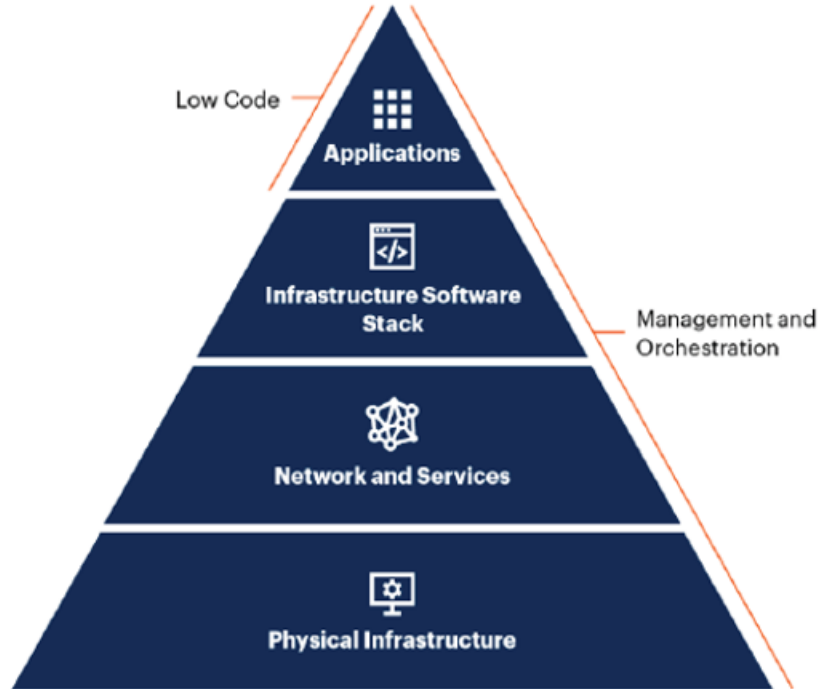


cadence®



But it is all so complex!

Reducing Edge Complexity: Edge Management and Orchestration, and Low Code Approaches



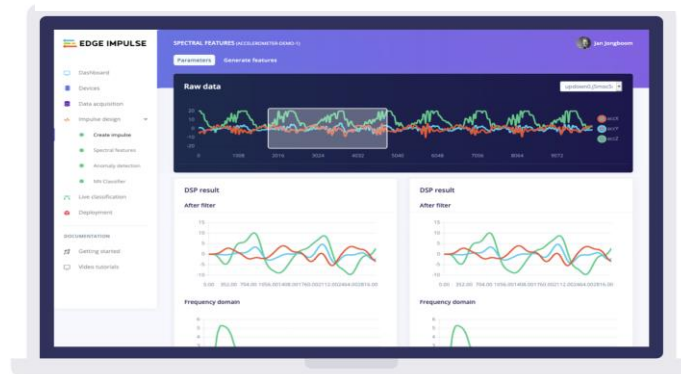
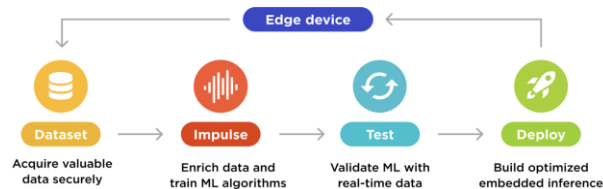
Source: Gartner
767014_C

© 2023 Edge Impulse

The leading embedded ML platform for any technology – Edge Impulse

The first fully integrated ML platform

- Royalty-free business model, therefore no impact on BOM cost
- Your IP, stays your IP
- Total explainability, no black boxes



TRUSTED BY LEADING ENTERPRISES

OURA



SONY

SAIC



ECOLAB

NOWATCH



65,000+

Developers

161K+

ML projects

5,000+

Enterprises

183M+

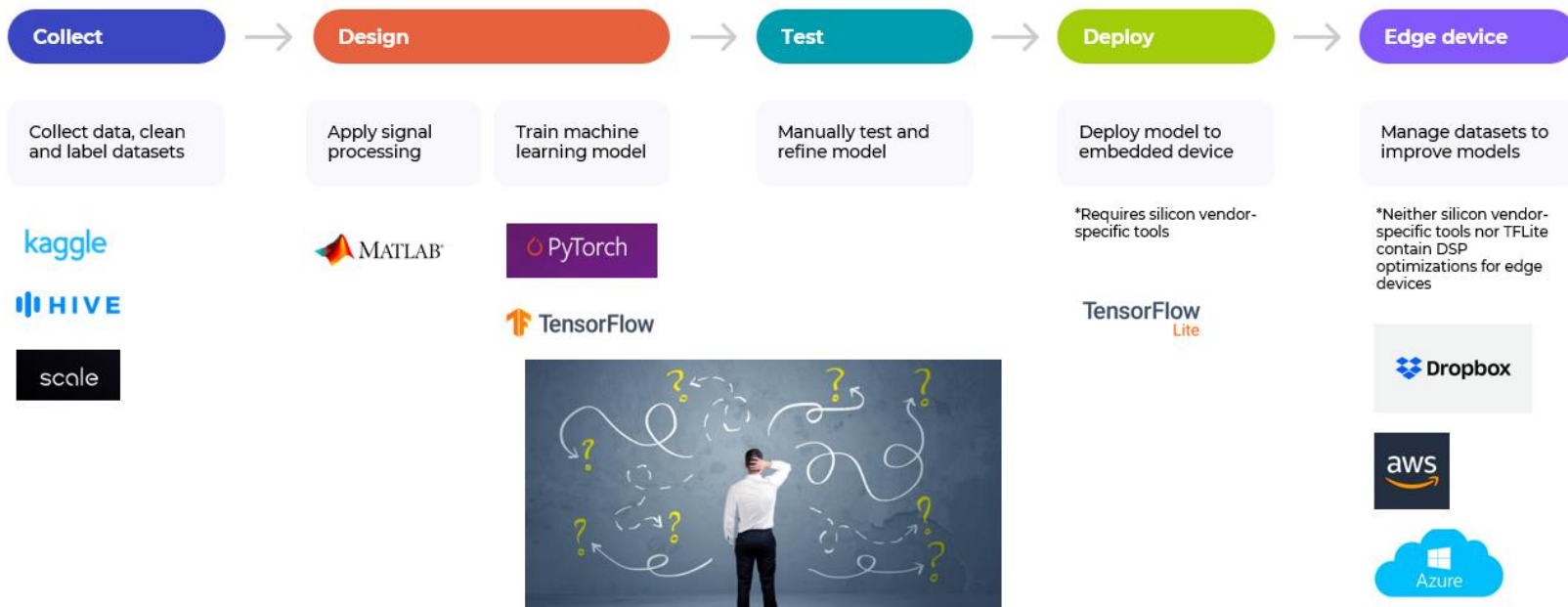
data samples

14M+

cloud jobs

Typical development of EdgeML applications

Requires 20+ man years, expertise in ML and embedded to build the infrastructure and integrate dozens of different tools.



Develop EdgeML applications with Edge Impulse

An end-to-end platform for projects using any data or device, built for developers with MLOps infrastructure built-in.

 **EDGE IMPULSE**

Collect

Design

Test

Deploy

Edge device

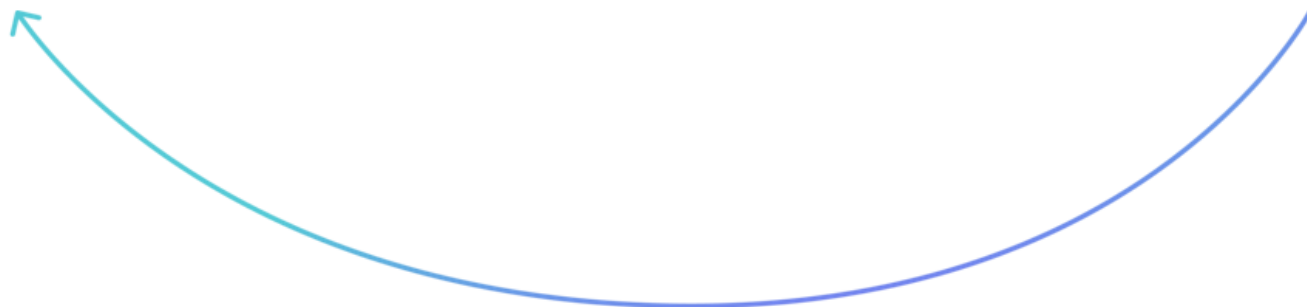
Rapidly build custom datasets

Confidently develop models & algorithms

Accurately test model performance

Easily deploy on any target

Easily improve add new device data



Some of the semiconductors & IP's we support



How it looks ?

Studio updates

Getting started

Start building your dataset or validate your model's on-device performance:

Add existing data

Collect new data

Upload your model

Step 2: Process "saved_model.zip"

Configure model settings for optimal processing.

Model input

Input shape: (28, 28)

Other

Model output

Output shape: (10)

Classification

Output labels (10)

Enter labels for your model separated by ",":

class 1, class 2, class 3, class 4, class 5, class 6, class 7, cla

Save model

On-device performance

MCUs

| DEVICE | LATENCY | EON COMPILER | | TFLITE | |
|--------------------|---------|--------------|--------|------------|---------------|
| | | RAM | ROM | RAM | ROM |
| Low-end MCU ⓘ | 124 ms. | 7.2K | 109.8K | 9.9K +2.7K | 127.7K +17.8K |
| High-end MCU ⓘ | 2 ms. | 7.2K | 109.9K | 9.9K +2.7K | 132.0K +22.1K |
| + AI accelerator ⓘ | 2 ms. | 7.2K | 109.9K | 9.9K +2.7K | 132.0K +22.1K |

Microprocessors

| DEVICE | LATENCY | MODEL SIZE |
|----------------------|---------|------------|
| MPU ⓘ | 1 ms. | 101.4K |
| GPU or accelerator ⓘ | 1 ms. | 101.4K |



BYOM-Bring Your Own Model

The diagram illustrates the BYOM (Bring Your Own Model) workflow. On the left, three logos represent input models: TensorFlow (orange), PyTorch (red), and ONNX (grey). Arrows from these logos point to a central stack of three colored layers (blue, green, red) representing the ONNX format. Below this stack is the Python SDK logo. From the Python SDK, two arrows branch out: one labeled 'profile()' pointing to a terminal window, and another labeled 'deploy()' pointing to a blue device icon. The terminal window displays the following output:

```
target: cortex-m4f-80mhz  
RAM: 39.1 kB  
flash: 37.6 kB  
latency: 145 ms
```

A man in a light blue shirt is speaking in the foreground, partially overlapping the diagram.

This new feature allows you to convert and optimize your machine learning models

Technology Examples



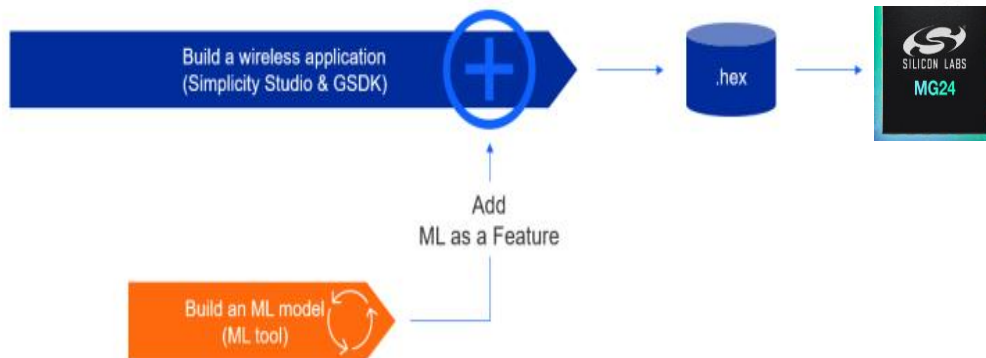
Industrial-Grade TinyML Applications with Silicon Labs



AI/ML capabilities of the EFR32MG24 with MVP

Bringing machine learning (ML) to IoT applications reduces bandwidth requirements, saves power, and increases a device's ability to make smarter decisions. Silicon Labs supports machine learning in all Series 1 and Series 2 wireless SoCs including newly released BG24 and MG24 products with built-in AI/ML hardware accelerator.

The MVP accelerator is a co-processor designed to perform matrix and vector operations. Using hardware accelerated kernel implementations will reduce neural network inference time, as well as off-load the main processor to allow it to perform other tasks or go to sleep.

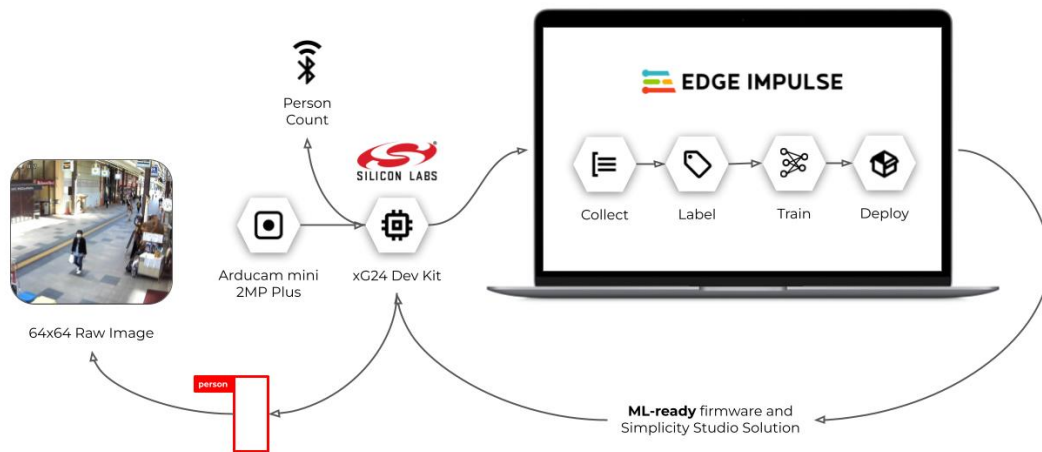


Accelerate AI/ML at the Edge with xG24 and Edge Impulse

AI/ML capabilities of the EFR32MG24 with MVP

Bringing machine learning (ML) to IoT applications reduces bandwidth requirements, saves power, and increases a device's ability to make smarter decisions. Silicon Labs supports machine learning in all Series 1 and Series 2 wireless SoCs including newly released BG24 and MG24 products with built-in AI/ML hardware accelerator.

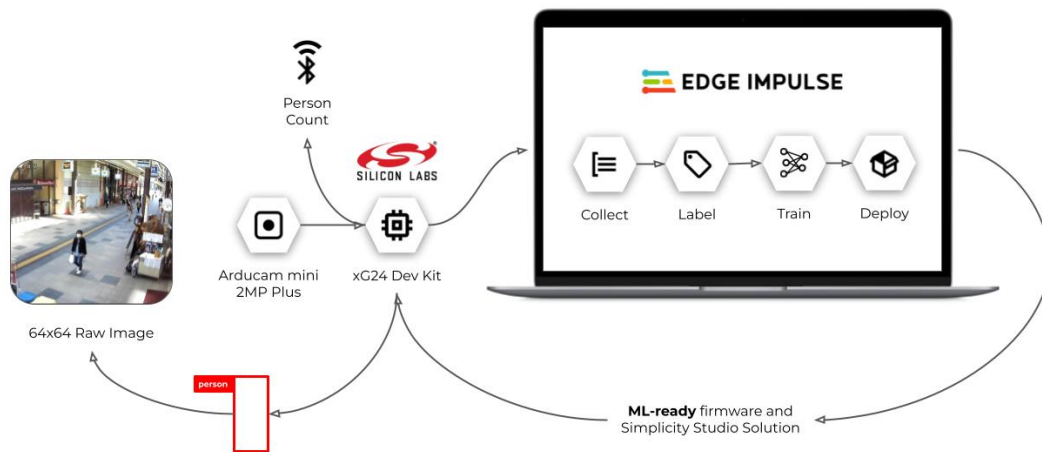
The MVP accelerator is a co-processor designed to perform matrix and vector operations. Using hardware accelerated kernel implementations will reduce neural network inference time, as well as off-load the main processor to allow it to perform other tasks or go to sleep.



Copyright © Edge Impulse Inc.

AI/ML capabilities of the EFR32MG24 with MVP

For this project, we attached an Arducam mini 2MP plus to the xG24 Dev Kit in order to capture low-res images of people flow from a real environment.



Copyright © Edge Impulse Inc.

AI/ML capabilities of the EFR32MG24 with MVP

For this project, we attached an Arducam mini 2MP plus to the xG24 Dev Kit in order to capture low-res images of people flow from a real environment.





EDGE IMPULSE



ALIF
SEMICONDUCTOR

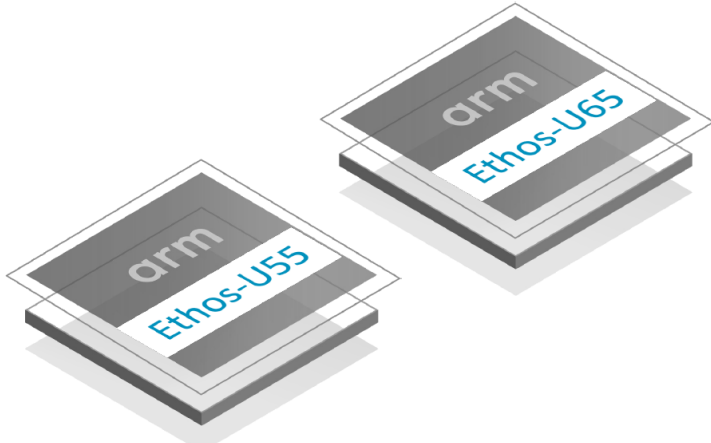
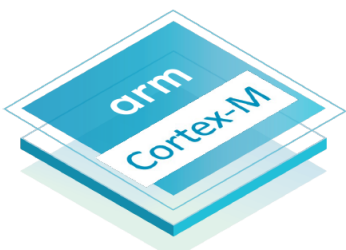
ARM's latest CortexM55 & microNPU Ethos-U55

| | Armv6/7-M | Armv8-M | Armv8.1-M |
|-------------|---|--|---|
| Compute | <ul style="list-style-type: none"> Cortex-M7 | | <ul style="list-style-type: none"> Cortex-M85 <ul style="list-style-type: none"> • Highest scalar performance • Helium • Arm Custom Instructions • PACBTI • Enhanced Functional Safety |
| Mainstream | <ul style="list-style-type: none"> Cortex-M4 Cortex-M3 | <ul style="list-style-type: none"> Cortex-M33 | <ul style="list-style-type: none"> Cortex-M55 <ul style="list-style-type: none"> • Helium • Arm Custom Instructions • Enhanced Functional Safety |
| Constrained | <ul style="list-style-type: none"> Cortex-M0+ Cortex-M0 | <ul style="list-style-type: none"> Cortex-M23 | |

ARM's latest CortexM55 & microNPU Ethos-U55

Ethos-U class of NPU's for Embedded Systems

Providing NN acceleration in highly constrained environments



Ubiquitous presence

NN acceleration in software

Orders of magnitude increase in NN perf

Easy integration into existing design

Signal Processing

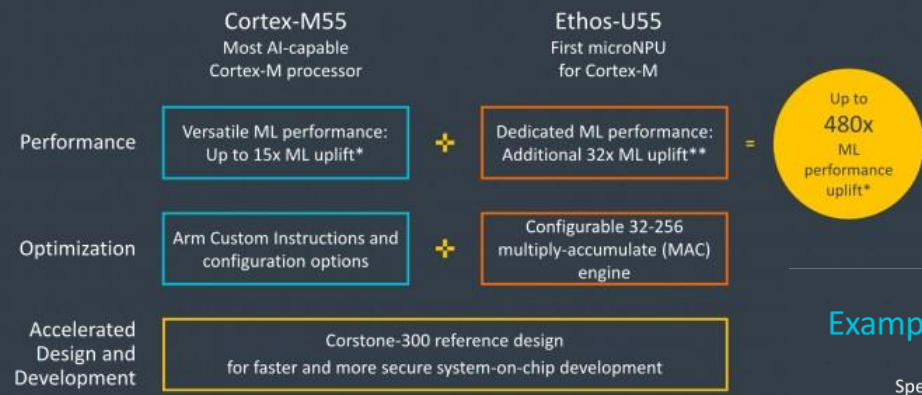


Neural network acceleration

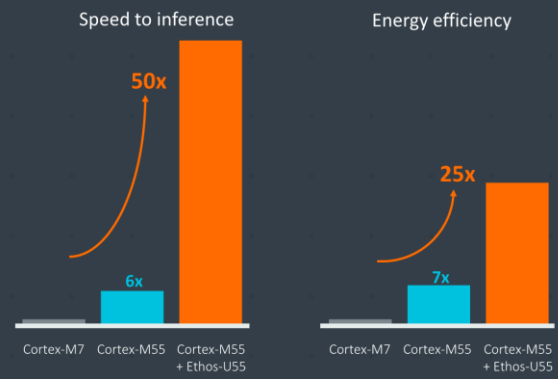
Common software development environment secures any investment made on software development

ARM's latest CortexM55 & microNPU Ethos-U55

Best-in-class Solution Optimized for Endpoint AI



Example: Typical ML Workload for a Voice Assistant

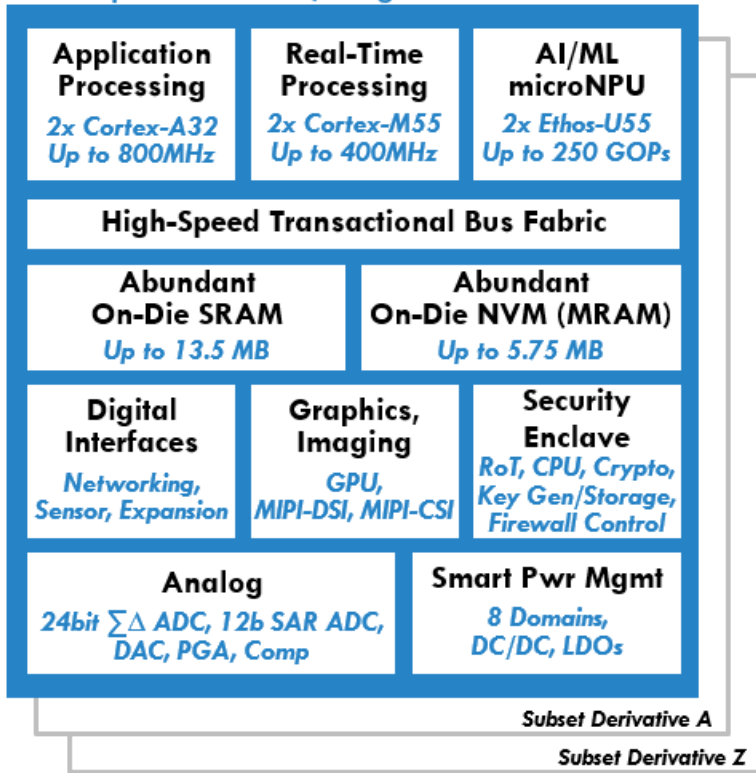


- ✓ Faster responses
- ✓ Smaller form-factors
- ✓ Improved accuracy

Latency and energy spent for all tasks listed combined: voice activity detection, noise cancellation, two-mic beamforming, echo cancellation, equalizing, mixing, keyword spotting, OPUS decode, and automatic speech recognition.

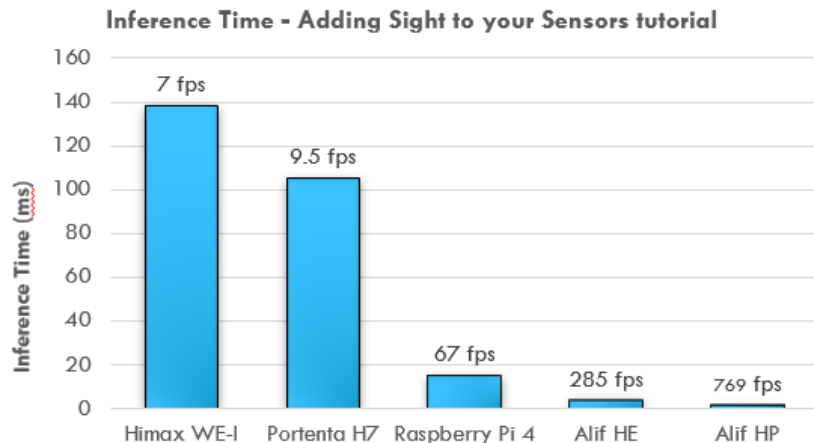
Official support for the Alif's Ensemble family

Superset Device, Single Monolithic Die



Astounding AI/ML performance benchmark

| ML Model | Inference Time | | | | |
|---|--|--|-------------------------|----------|------------|
| | Alif MCU | Competing MCU | Performance Improvement | | |
| Object Detection | 0.786 msec Alif E3 Cortex-M55+Ethos-U55 @ 400MHz | 17 msec Broadcom BCM2711B0 Quad Cortex-A72 @ 1.5GHz | 22x | | |
| | | ML Model | Inference Time (msec) | | |
| | | | M55 only | M55 +U55 | U55 Uplift |
| | | Image Classification (Mobilenet V2)* | 600 | 8 | 75x |
| | | Keyword Spotting (DS-CNN-L)** | 94 | 3 | 31x |
| | | Object Detection (FOMO)* | 74 | 0.786 | 94x |
| | | Face Detection (SSD Face + Yaw)* | 394 | 4.1 | 96x |
| | | Face Detection (SSD Face + Yaw + Landmarks)* | 418 | 4.7 | 89x |
| Face Detection (SSD Face + Yaw + Landmarks)** | 1030 | 10.4 | 99x | | |



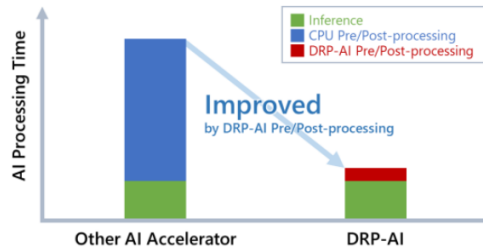


EDGE IMPULSE

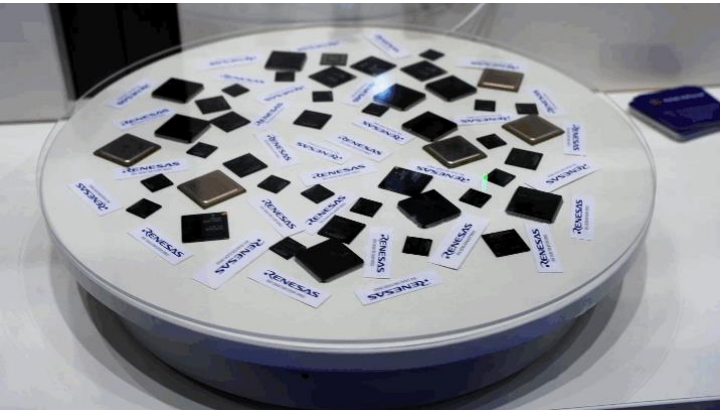
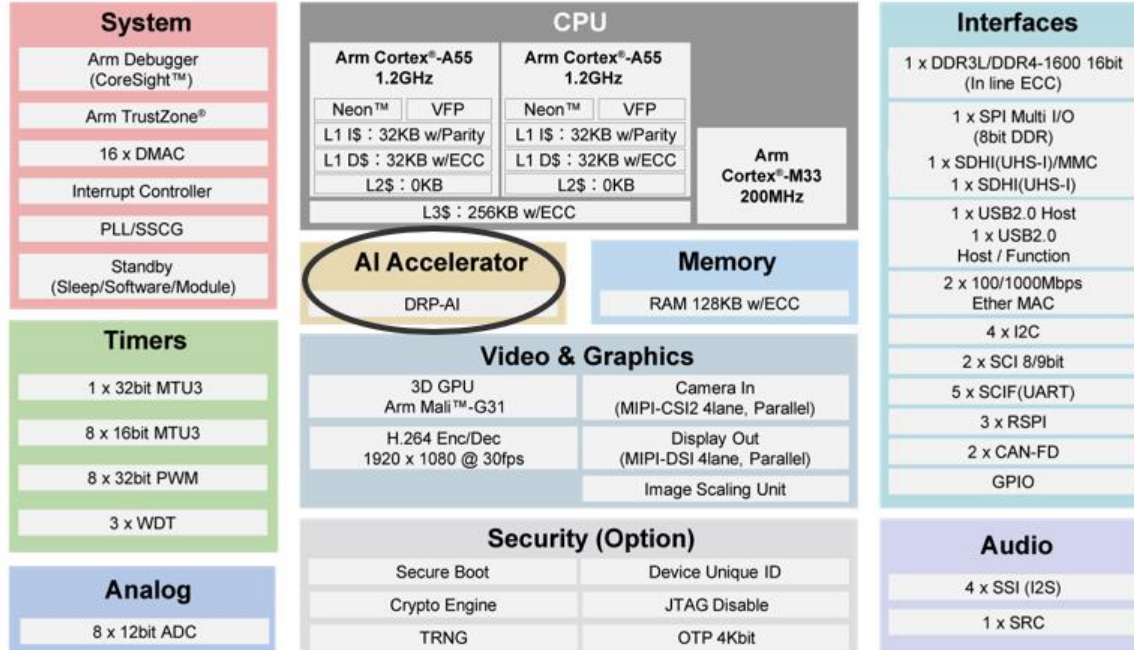
RENESAS
BIG IDEAS FOR EVERY SPACE

Official support for Renesas RZ/V2L 2 x CortexA55 and DRP-AI ML accelerator

| | Other AI Accelerator | DRP-AI |
|-----------------|----------------------|--------|
| Pre-processing | CPU | DRP-AI |
| Inference | AI Accelerator | DRP-AI |
| Post-processing | CPU | DRP-AI |

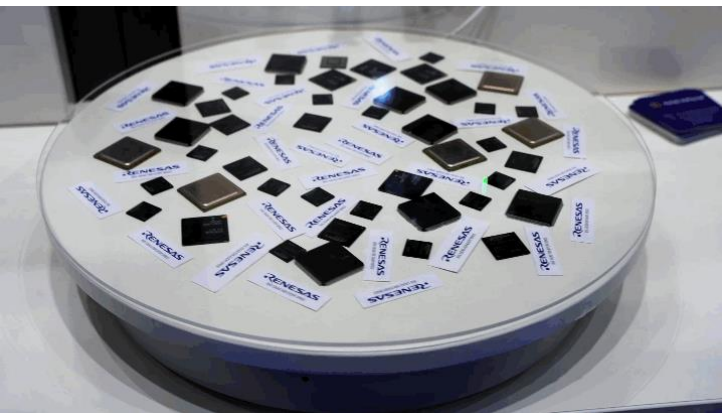
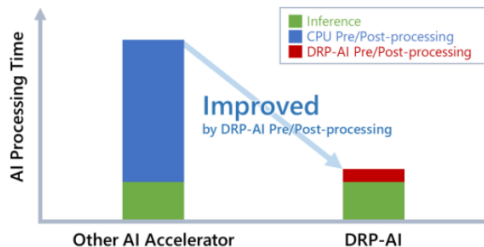


RZ/V2L Block Diagram



Official support for Renesas RZ/V2L 2 x CortexA55 and DRP-AI ML accelerator

| | Other AI Accelerator | DRP-AI |
|-----------------|----------------------|--------|
| Pre-processing | CPU | DRP-AI |
| Inference | AI Accelerator | DRP-AI |
| Post-processing | CPU | DRP-AI |



RENESAS

- Dashboard
- Devices
- Data sources
- Data acquisition
- Impulse design
 - Create impulse
 - Image
 - NN Classifier
- EON Tuner
- Retrain model
- Live classification
- Model testing

Deploy your impulse

You can deploy your impulse to any device. This makes the model run without an internet connection, minimizes latency, and runs with minimal power consumption. [Read more.](#)

Create library

Turn your impulse into optimized source code that you can run on any device.



C++ library



DRP-AI Library

Build firmware

Get a ready-to-go binary for your development board that includes your impulse.



Linux (RZ/V2L)

Analog

8 x 12bit ADC

Secure Boot

Crypto Engine

TRNG

Device Unique ID

JTAG Disable

OTP 4Kbit

4 x SSI (I2S)

1 x SRC

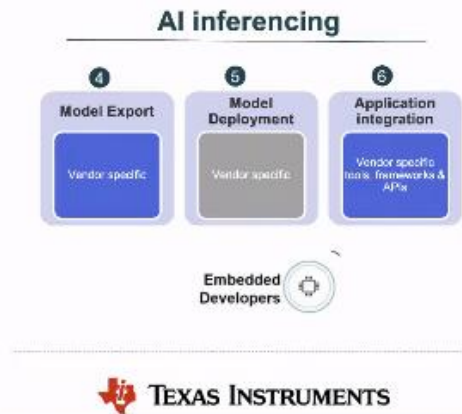
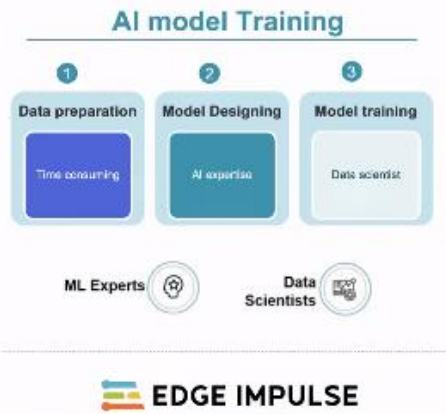
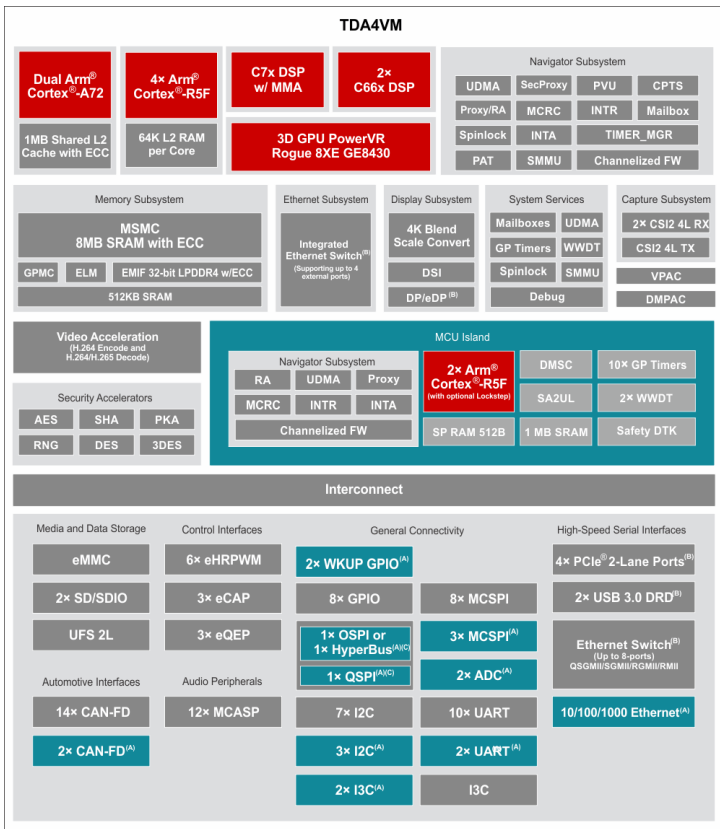


EDGE IMPULSE



**TEXAS
INSTRUMENTS**

TDA4VM multi-core embedded vision processor



Texas Instruments and Edge Impulse together democratizing end-to-end embedded AI development on TI SoCs



Industrial edge compute based on NVIDIA



And so many other cores are supported

Tensilica Datasheet

Xtensa LX7 Processor **cādence**
High-performance, configurable, and extensible controllers and DSPs

Xtensa LX7 Processors for Today's SoC Challenges



Espressif ESP-32S3

ESP32-S3

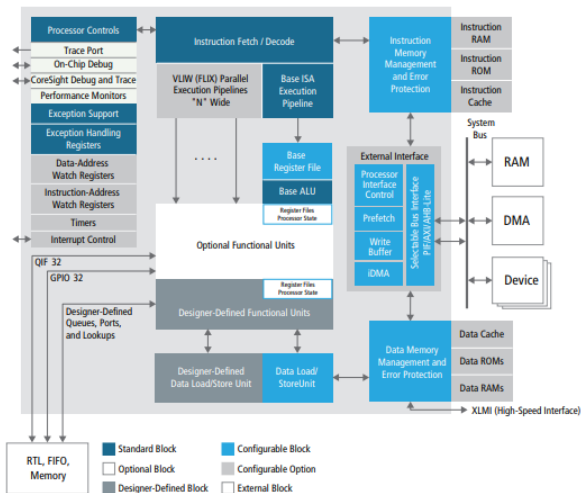
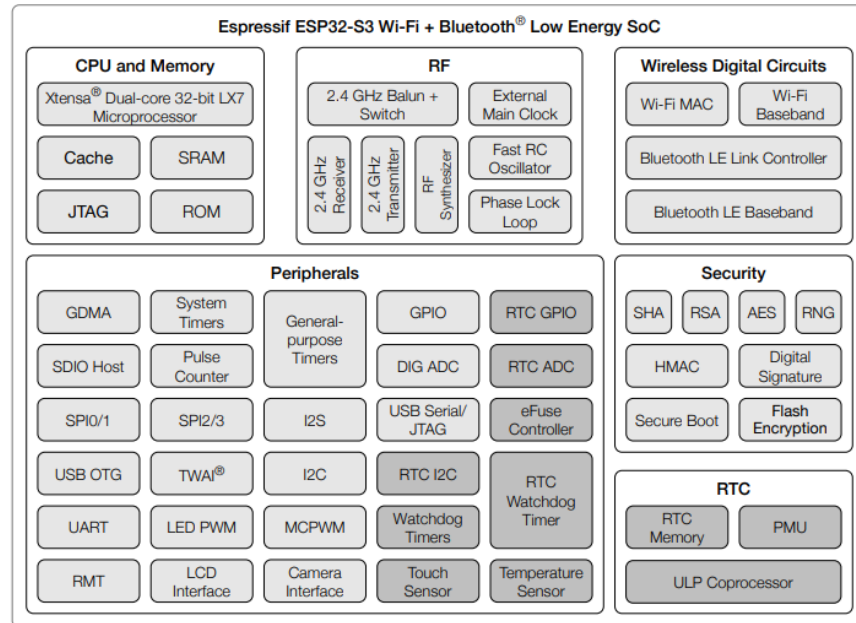


Figure 1: Block diagram of Xtensa LX7 processor architecture



Power consumption



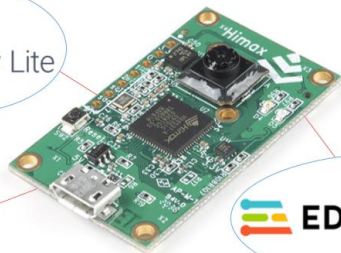
ESP32-S3 Functional Block Diagram

And so many other cores are supported

The screenshot displays the Edge Impulse web interface. On the left is a navigation sidebar with the 'EDGE IMPULSE' logo and menu items: Dashboard, Devices, Data acquisition, Impulse design, Create impulse, EON Tuner, Retrain model, Live classification, Model testing, Versioning, Deployment, GETTING STARTED, Documentation, and Forums. The main content area is titled 'DATA ACQUISITION (ESP32-TESTING)' and includes tabs for 'Training data', 'Test data', and 'Export data'. A notification banner reads: 'Did you know? You can capture data from any device or development board, or upload your existing datasets - Show options'. Below this, there are two progress indicators: 'DATA COLLECTED 17m 38s' and 'TRAIN / TEST ... 100% / ...'. A 'Record new data' section features a 'Connect using WebUSB' button, a 'Device' dropdown set to 'esp-eye', a 'Label' input field with 'test', a 'Sensor' dropdown set to 'Camera (64x64)', and a 'Camera feed' video window showing a person's face. A 'Start sampling' button is positioned below the camera feed. At the bottom, a 'RAW DATA' section prompts the user to 'Click on a sample to load...'. A central table titled 'Collected data' lists the following entries:

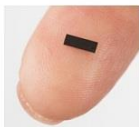
| SAMPLE NAME | LABEL | ADDED | LENGTH |
|-------------------|-------|-----------------|--------|
| test.jpg.309ovlm | test | Apr 11 2022,... | - |
| test.jpg.309oh9l2 | test | Apr 11 2022,... | - |
| test.jpg.309ogt8g | test | Apr 11 2022,... | - |
| test.jpg.309ogj9t | test | Apr 11 2022,... | - |
| test.jpg.309ogckt | test | Apr 11 2022,... | - |
| test.jpg.309og77a | test | Apr 11 2022,... | - |
| test.309ofjdb | test | Apr 11 2022,... | 3s |

And so many other cores are supported

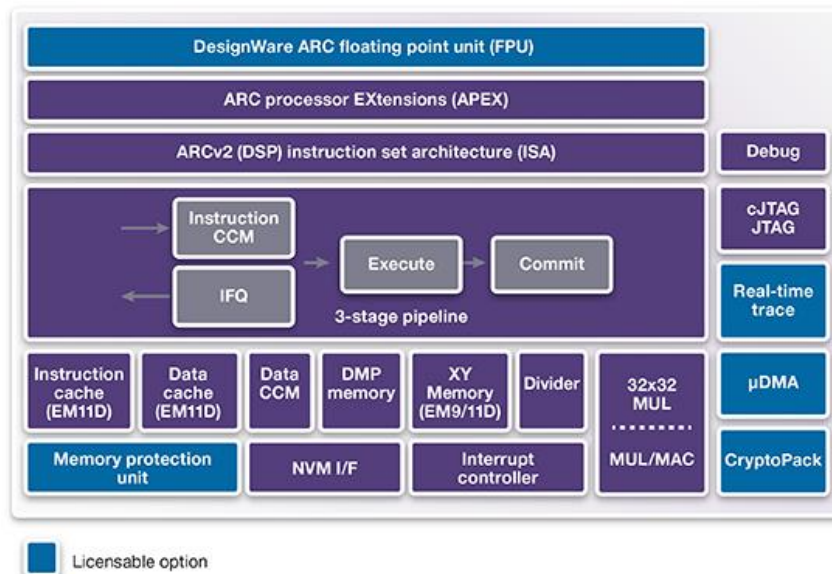


EDGE IMPULSE

➤➤ **HX6537-A**
WE-I Plus ASIC



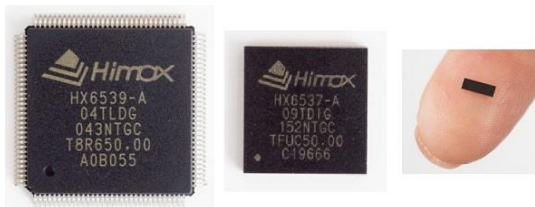
SYNOPSIS® ARC EM9D / EM11D Processors



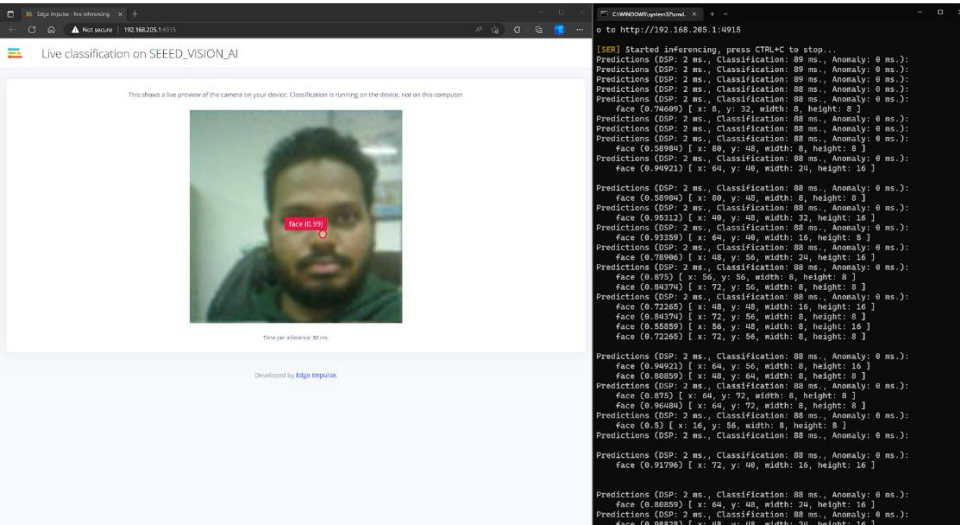
And so many other cores are supported



➤ **HX6537-A** WE-I Plus ASIC



And so many other cores are supported



Our business model

Developer

For developers looking to deploy ML on any edge device

Free

- ✓ **Unlimited** projects with 20 min per job and 4GB/4hr of data per project
- ✓ 1 seat included & basic collaboration
- ✓ Automatic platform updates
- ✓ Shared data storage
- ✓ Community-based forum
- ✓ Limited EON Tuner & Application Testing

Enterprise

For enterprise companies with large-scale projects

\$ per project/mo (royalty-free)

**Includes everything in Developer plus...*

- ✓ **Additional** projects with unlimited compute time and data storage (10,000 min included in plan)
- ✓ 5 seats included & full team collaboration
- ✓ Managed platform updates
- ✓ Hybrid data storage
- ✓ Private pre-processing and deployment blocks
- ✓ Enterprise-grade dedicated support
- ✓ Full EON Tuner & Application Testing

Add-ons: Private cloud, white label, additional users, additional compute time

A platform that goes from data to algorithms

Raw data
Unlabeled
Unstructured
Noisy



Source Code
Algorithms
exported as
libraries,
optimized for
hardware

- Data management
- Data visualization and exploration
- Built in optimizations for DSP + ML models
- Hardware virtualization
- Other tools for model validation and AutoML

*Thank
you*



COME VISIT OUR BOOTH

Amir Sherman
Senior Global Director, Semiconductor @ Ecosystem Business Development
+972-52-2240811
+49-173-3232288
amir@edgeimpulse.com
www.edgeimpulse.com

