Deploy Your Embedded Vision Solution on Any Processor Using Edge Impulse

Amir Sherman

Senior Director Global Business Development Semiconductors & Eco Partners

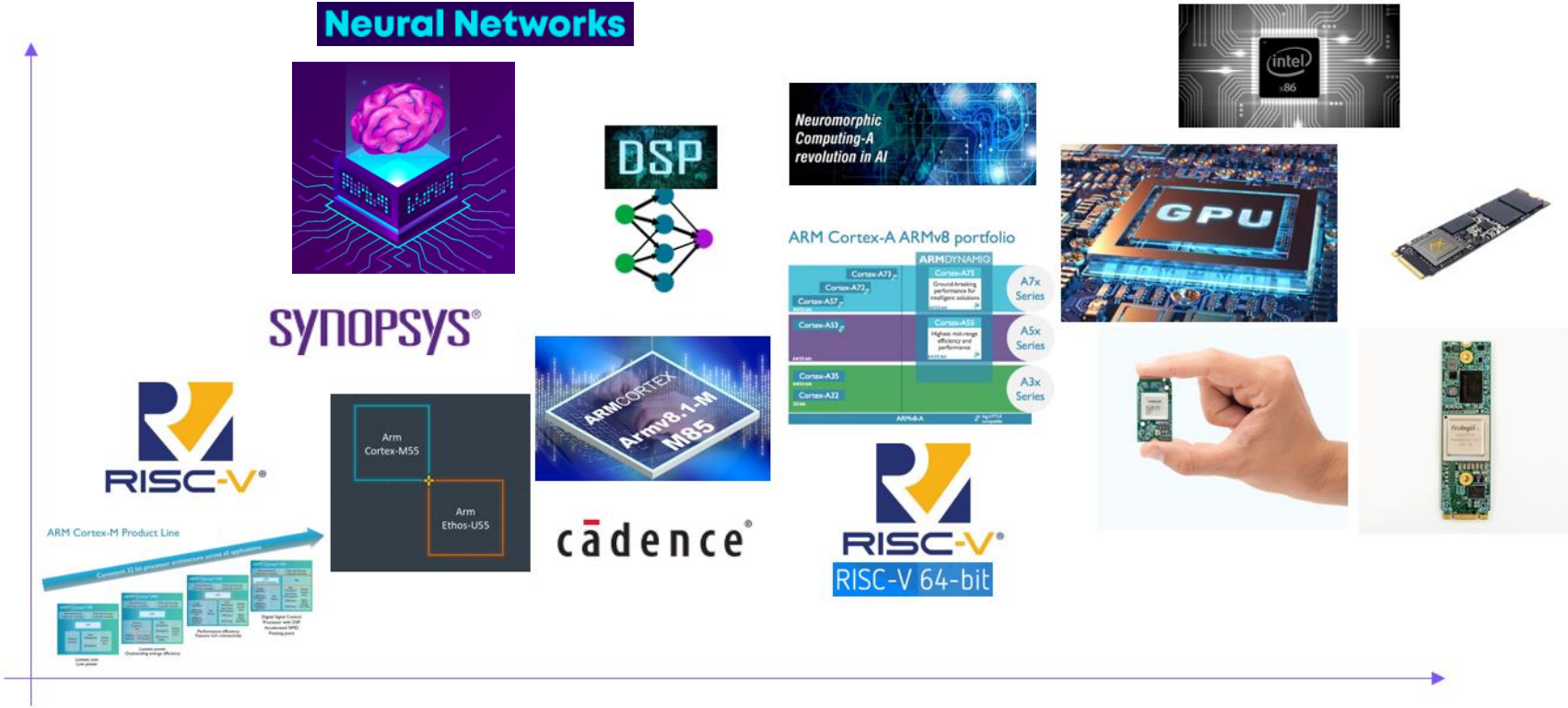
Edge Impulse

EDGE IMPULSE

From cloud AI to edge AI/ML to endpoint AI

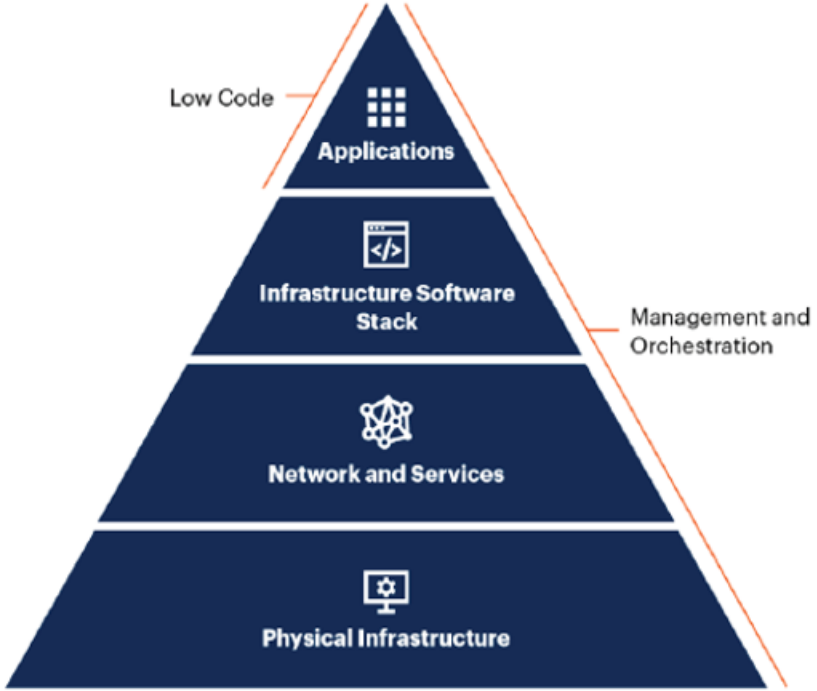
Cameras are everywhere

What is the “best/right” technology ?



But it is all so complex!

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

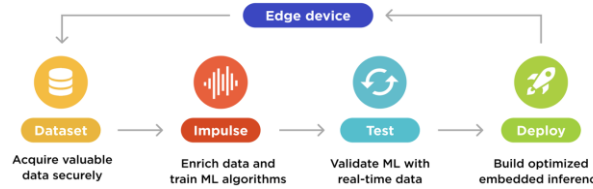


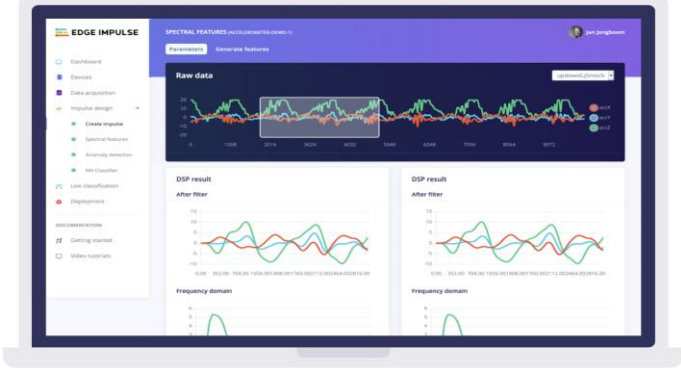
Source: Gartner
767014_C

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

ŌURA · poly · SONY · SAIC · Lexmark · NASA · ECOLAB · NOWATCH · BOSCH · KNOW LABS

65,000+ Developers

5,000+ Enterprises

161K+ ML projects

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.

Collect → **Design** → **Test** → **Deploy** → **Edge device**

Collect data, clean and label datasets

kaggle

HIVE

scale

Apply signal processing

MATLAB

Train machine learning model

PyTorch

TensorFlow

Manually test and refine model

Deploy model to embedded device

*Requires silicon vendor-specific tools

TensorFlow Lite

Manage datasets to improve models

*Neither silicon vendor-specific tools nor TFLite contain DSP optimizations for edge devices

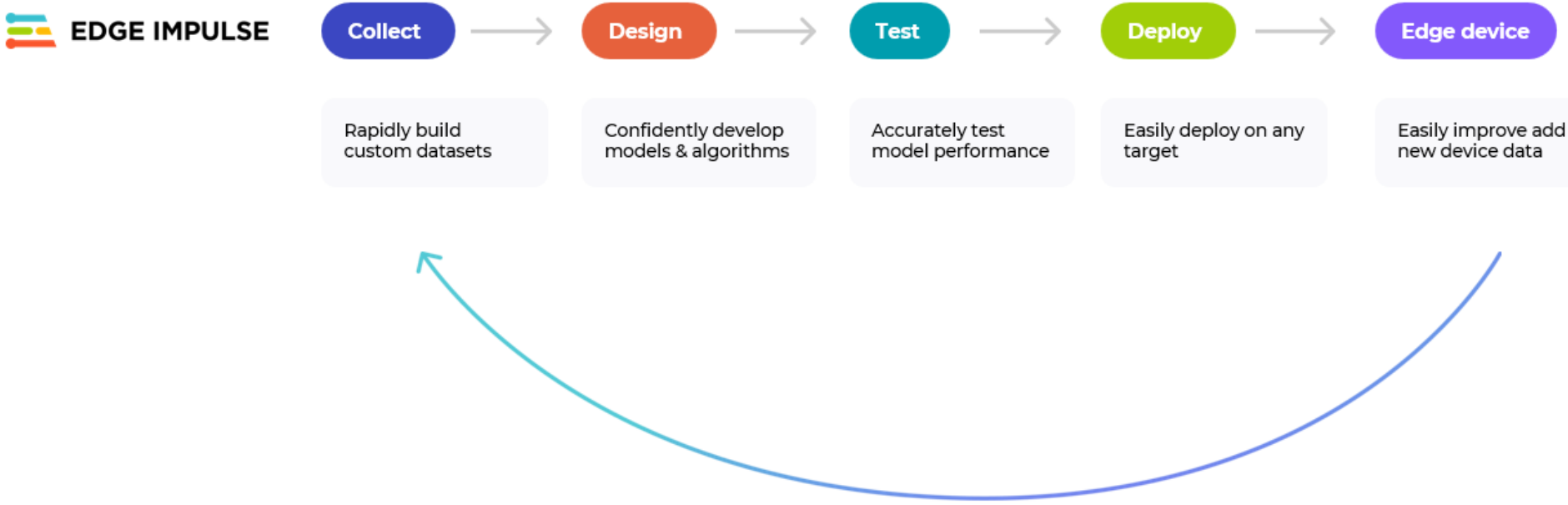
Dropbox

aws

Azure

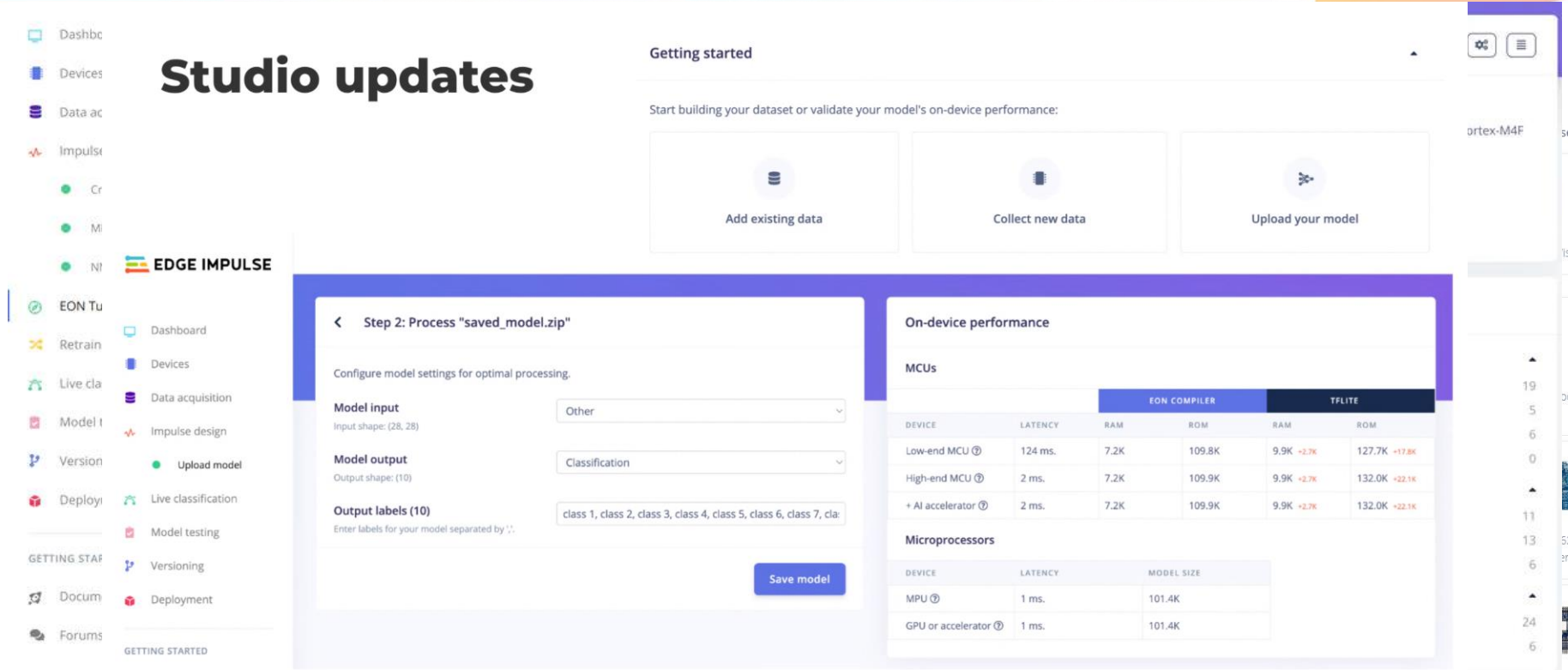
Develop EdgeML applications with Edge Impulse

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

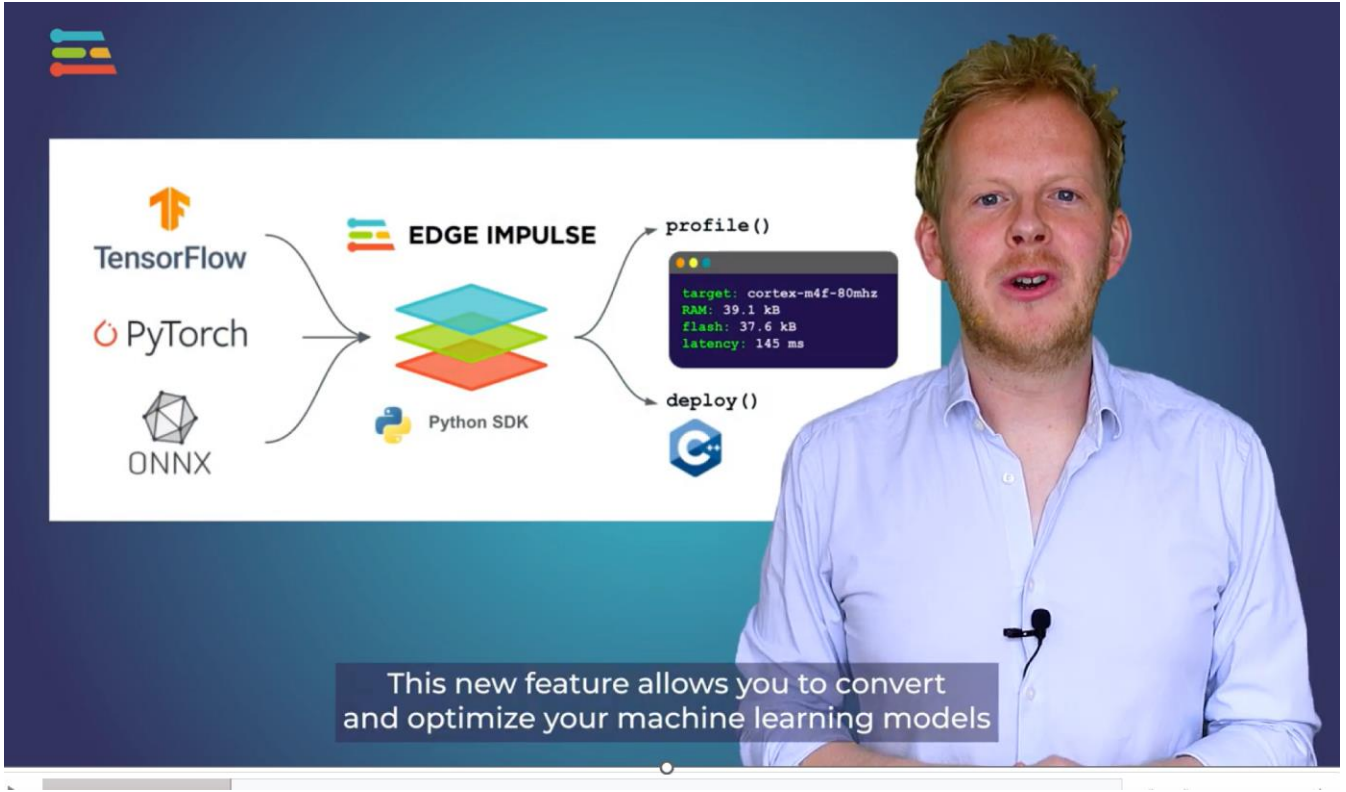


Some of the semiconductors & IP's we support

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
EDGE IMPULSE
Dashboard
Devices
Data acquisition
Impulse design
Upload model
Live classification
Model testing
Versioning
Deployment
GETTING STARTED
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
EON COMPILER
TFLITE
DEVICE
LATENCY
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



Technology Examples

EDGE IMPULSE

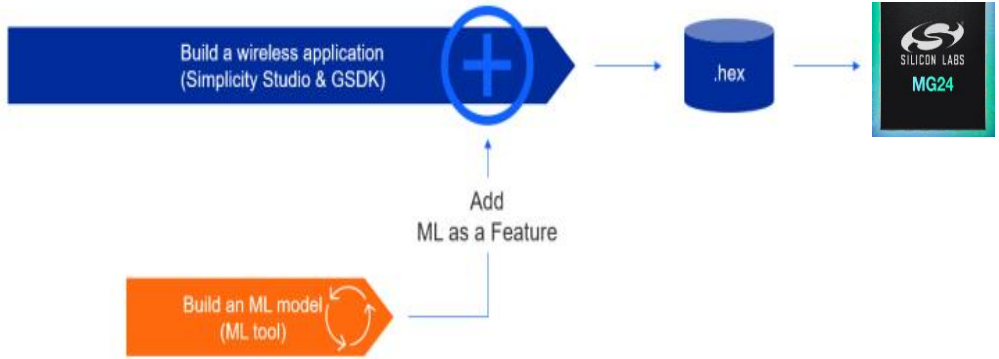
Industrial-Grade TinyML Applications with Silicon Labs

EDGE IMPULSE | 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.

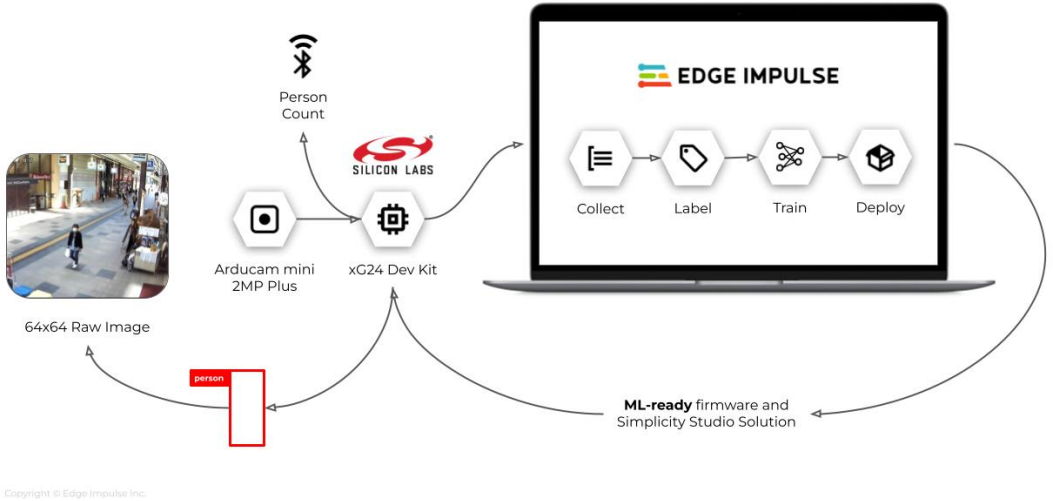


Accelerate AI/ML at the Edge with xG24 and Edge Impulse

AI/ML capabilities of the EFR32MG24 with MVP

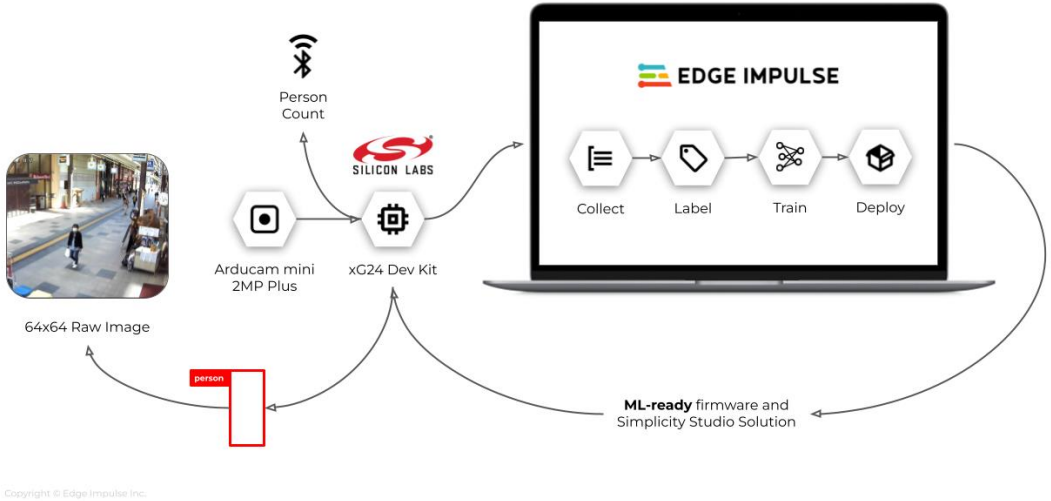
Bringing machine learning (ML) to IoT applications reduces bandwidth requirements, saves power, and increases a device's ability to make smarter decisions. Silicon Labs supports machine learning in all Series 1 and Series 2 wireless SoCs including newly released BG24 and MG24 products with built-in AI/ML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



AI/ML capabilities of the EFR32MG24 with MVP

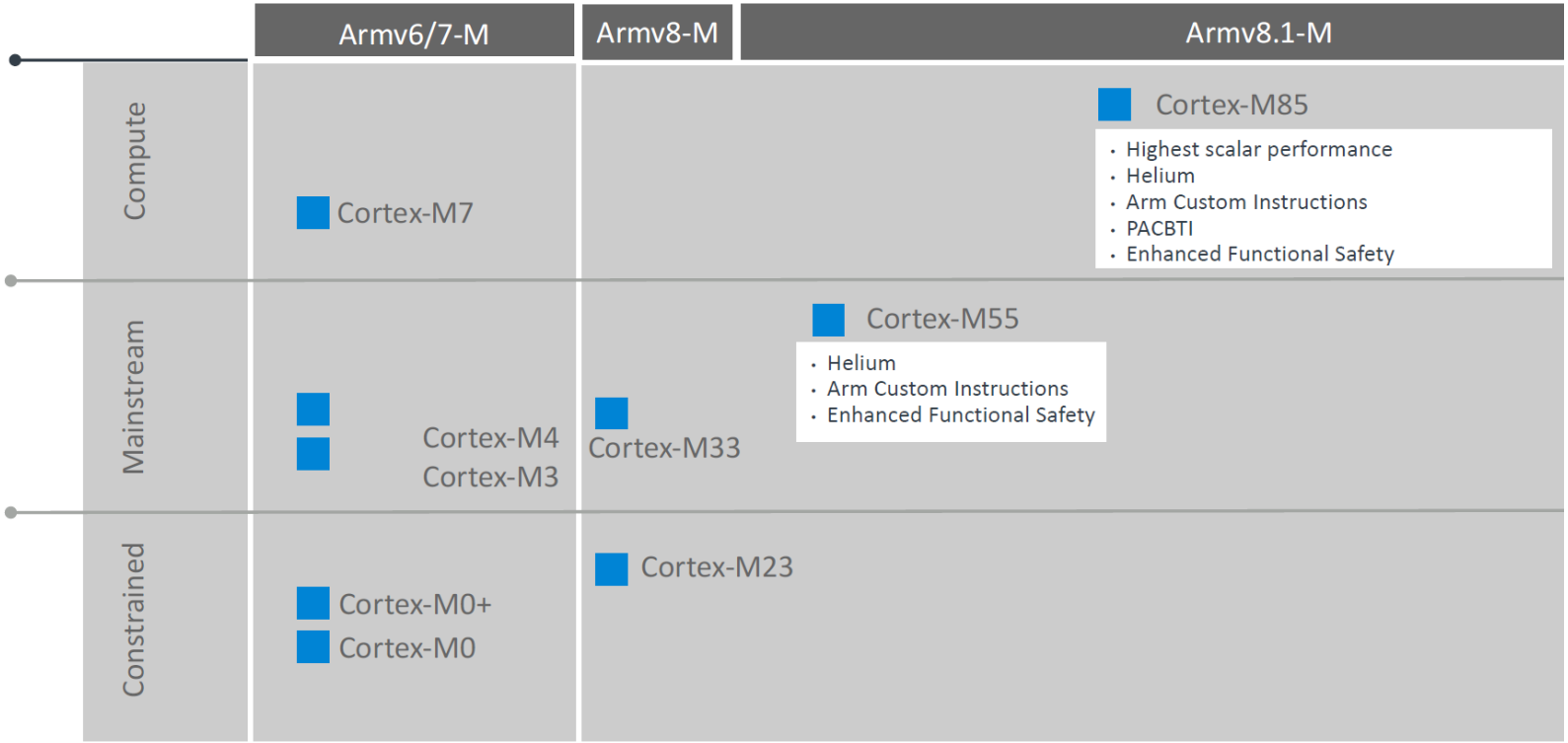
For this project, we attached an Arducam mini 2MP plus to the xG24 Dev Kit in order to capture low-res images of people flow from a real environment.



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.

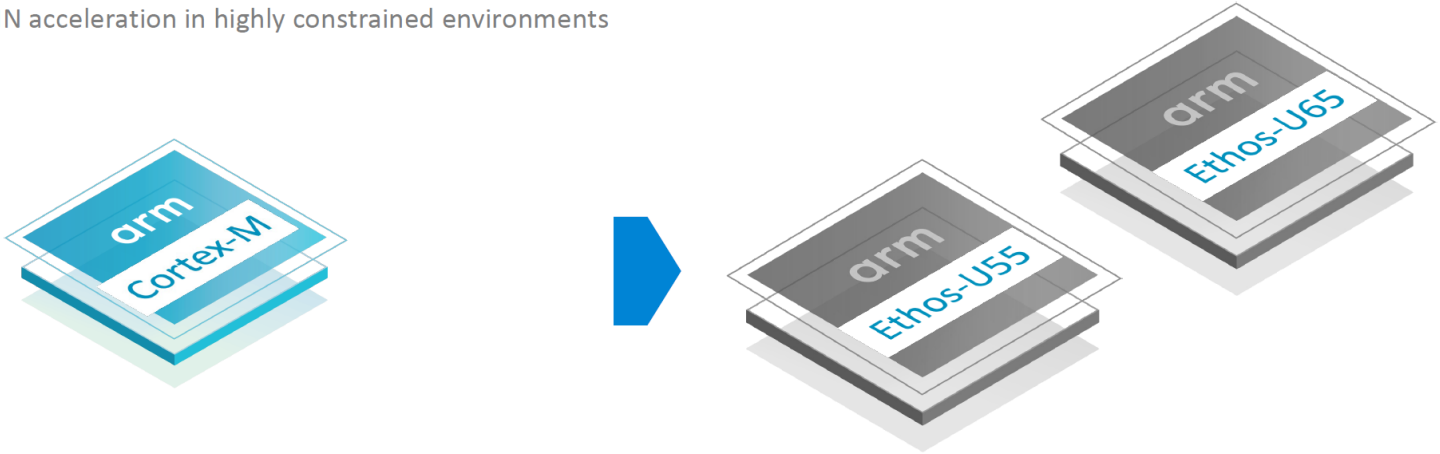
ARM's latest CortexM55 & microNPU Ethos-U55



ARM's latest CortexM55 & microNPU Ethos-U55

Ethos-U class of NPU's for Embedded Systems

Providing NN acceleration in highly constrained environments

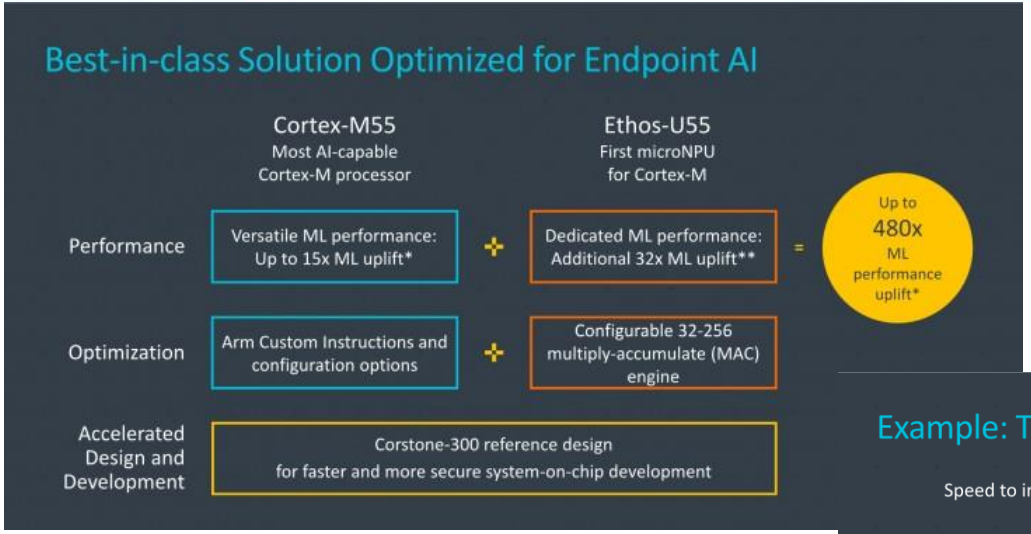


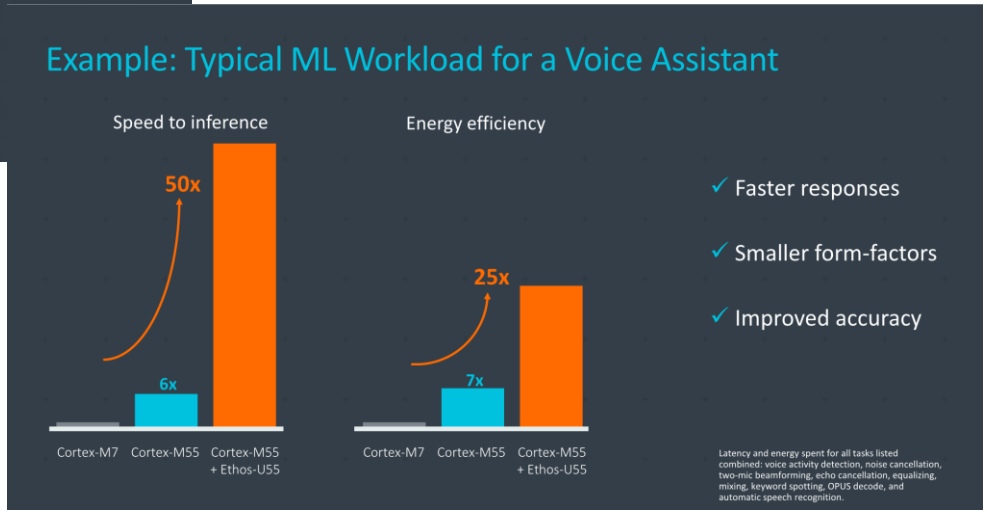
Ubiquitous presence	Orders of magnitude increase in NN perf
NN acceleration in software	Easy integration into existing design

Signal Processing → Neural network acceleration

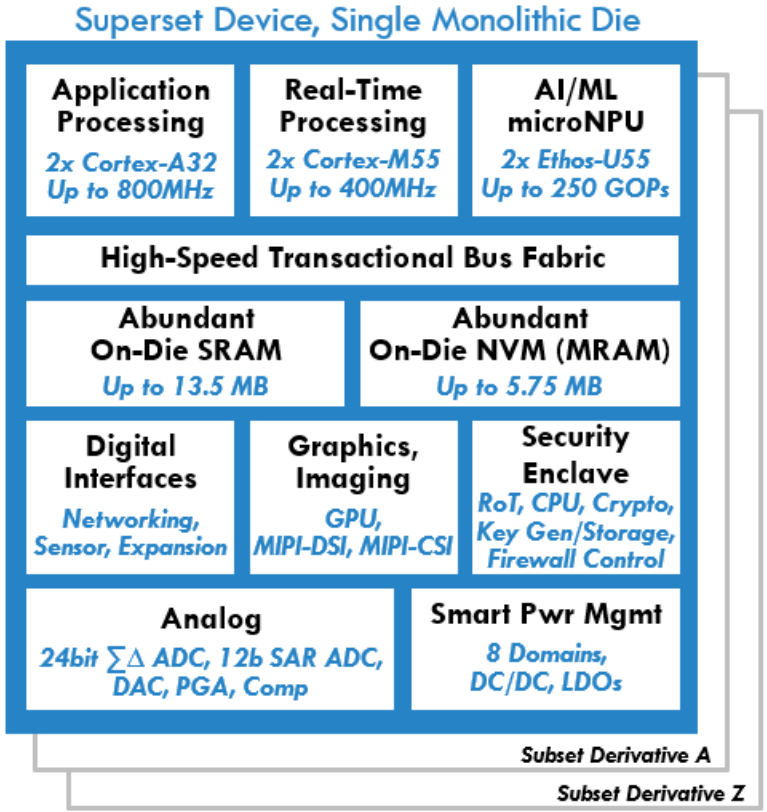
Common software development environment secures any investment made on software development

ARM's latest CortexM55 & microNPU Ethos-U55





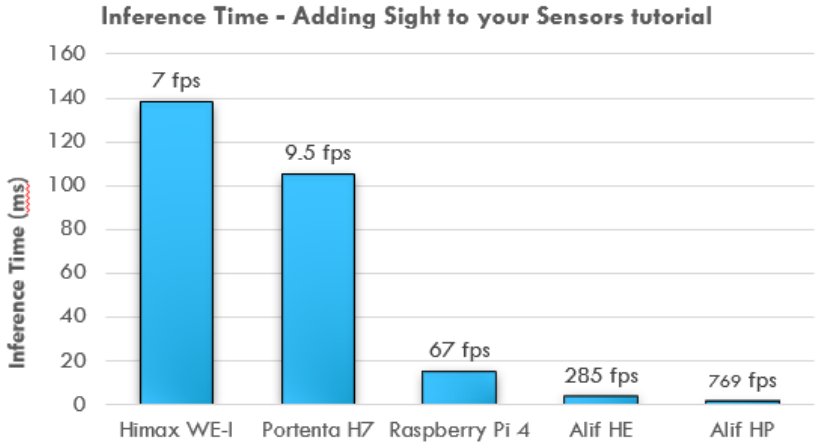
Official support for the Alif's Ensemble family



Astounding AI/ML performance benchmark

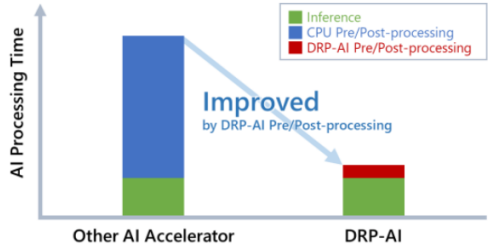
ML Model	Inference Time		
	Alif MCU	Competing MCU	Performance Improvement
Object Detection	0.786 msec Alif E3 Cortex-M55+Ethos-U55 @ 400MHz	17 msec Broadcom BCM2711B0 Quad Cortex-A72 @ 1.5GHz	22x

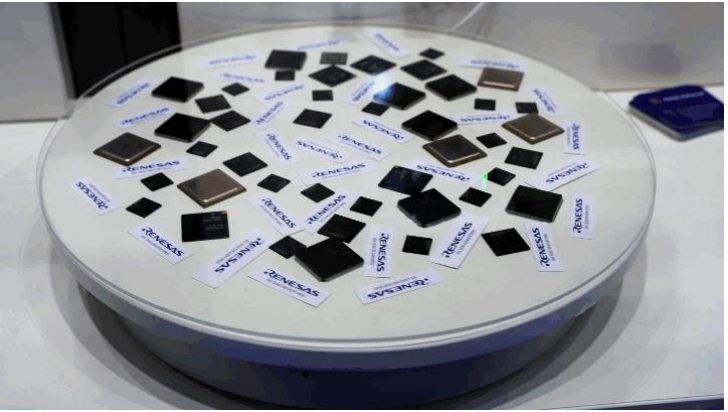
ML Model	Inference Time (msec)		
	M55 only	M55 +U55	U55 Uplift
Image Classification (Mobilenet V2)*	600	8	75x
Keyword Spotting (DS-CNN-L)**	94	3	31x
Object Detection (FOMO)*	74	0.786	94x
Face Detection (SSD Face + Yaw)*	394	4.1	96x
Face Detection (SSD Face + Yaw + Landmarks)*	418	4.7	89x
Face Detection (SSD Face + Yaw + Landmarks)**	1030	10.4	99x



Offical 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

System
- Arm Debugger (CoreSight™)
- Arm TrustZone®
- 16 x DMAC
- Interrupt Controller
- PLL/SSCG
- Standby (Sleep/Software/Module)

Timers
- 1 x 32bit MTU3
- 8 x 16bit MTU3
- 8 x 32bit PWM
- 3 x WDT

Analog
- 8 x 12bit ADC

CPU
- Arm Cortex®-A55 1.2GHz: Neon™ VFP; L1 I$: 32KB w/Parity; L1 D$: 32KB w/ECC; L2$: 0KB
- Arm Cortex®-A55 1.2GHz: Neon™ VFP; L1 I$: 32KB w/Parity; L1 D$: 32KB w/ECC; L2$: 0KB
- L3$: 256KB w/ECC
- Arm Cortex®-M33 200MHz

AI Accelerator
- DRP-AI

Memory
- RAM 128KB w/ECC

Video & Graphics
- 3D GPU Arm Mali™-G31
- Camera In (MIPI-CSI2 4lane, Parallel)
- H.264 Enc/Dec 1920 x 1080 @ 30fps
- Display Out (MIPI-DSI 4lane, Parallel)
- Image Scaling Unit

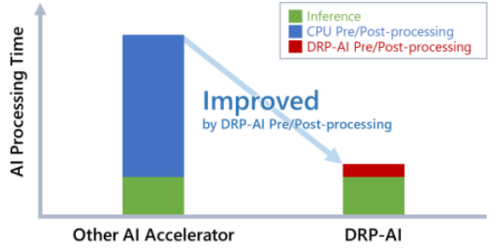
Security (Option)
- Secure Boot
- Device Unique ID
- Crypto Engine
- JTAG Disable
- TRNG
- OTP 4Kbit

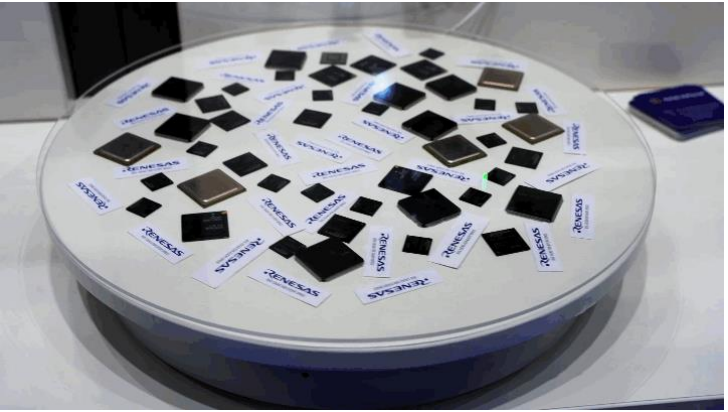
Interfaces
- 1 x DDR3L/DDR4-1600 16bit (In line ECC)
- 1 x SPI Multi I/O (8bit DDR)
- 1 x SDHI(UHS-I)/MMC
- 1 x SDHI(UHS-I)
- 1 x USB2.0 Host
- 1 x USB2.0 Host / Function
- 2 x 100/1000Mbps Ether MAC
- 4 x I2C
- 2 x SCI 8/9bit
- 5 x SCIF(UART)
- 3 x RSPI
- 2 x CAN-FD
- GPIO

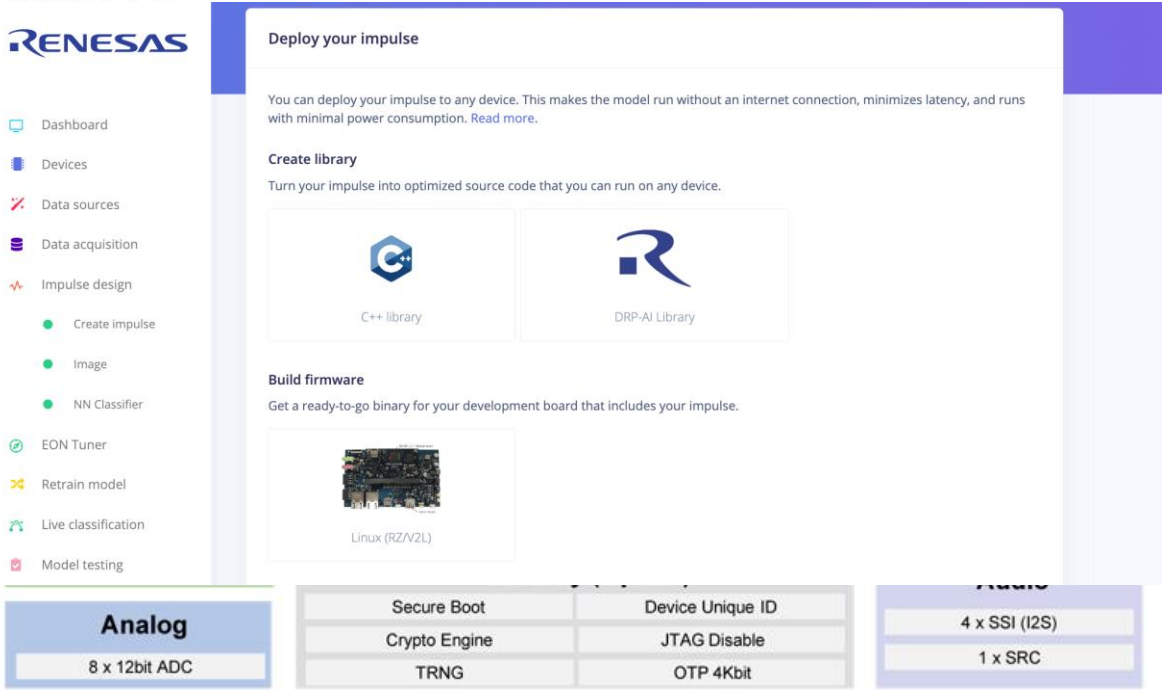
Audio
- 4 x SSI (I2S)
- 1 x SRC

Offical 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



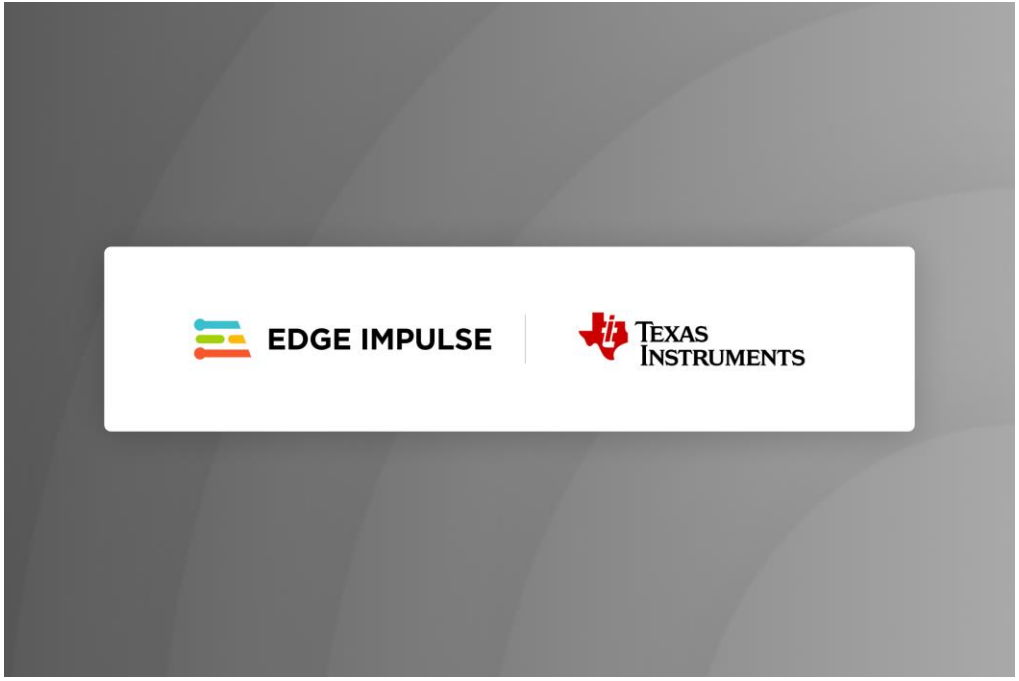




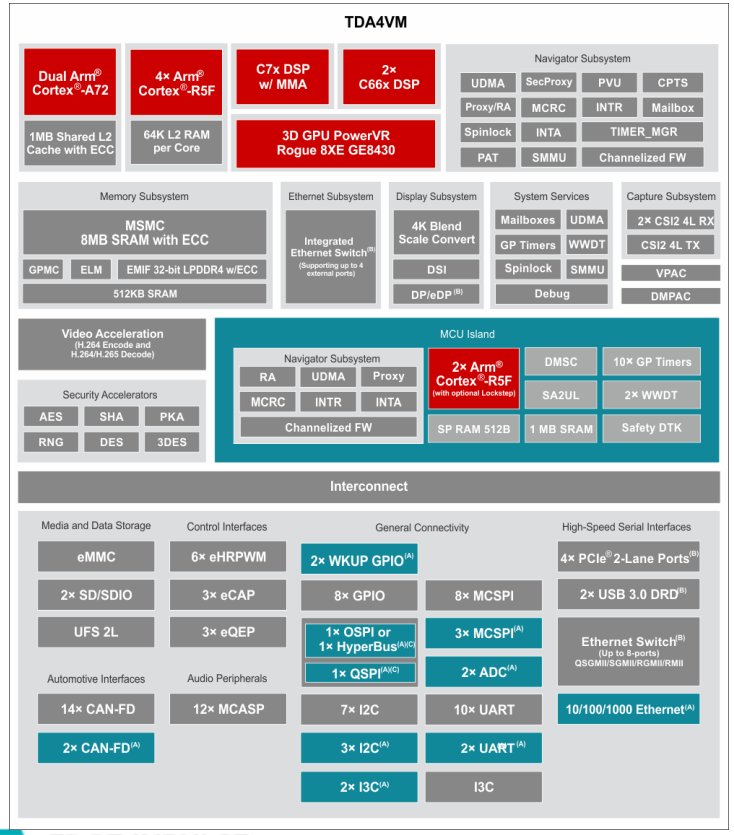
Analog
8 x 12bit ADC

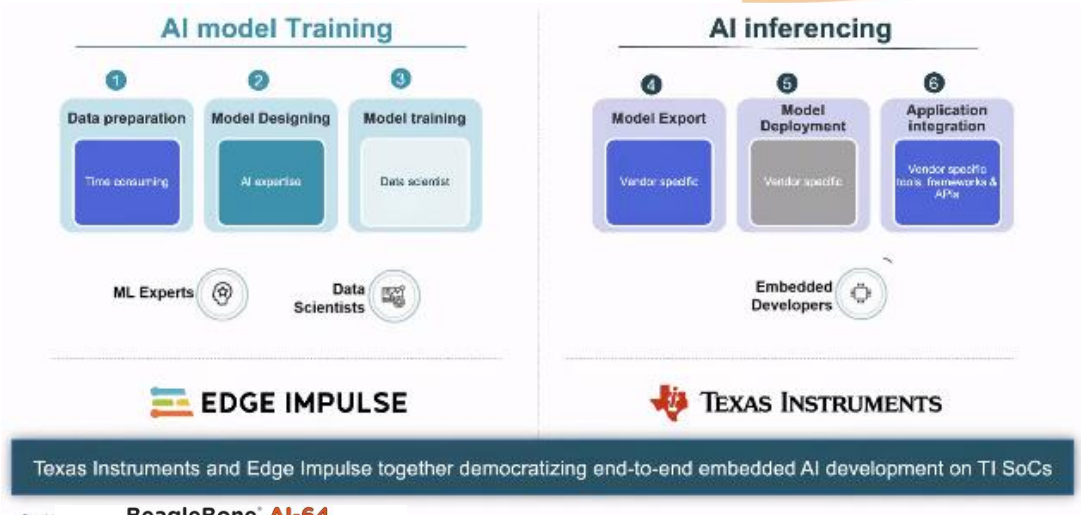
Secure Boot	Device Unique ID
Crypto Engine	JTAG Disable
TRNG	OTP 4Kbit

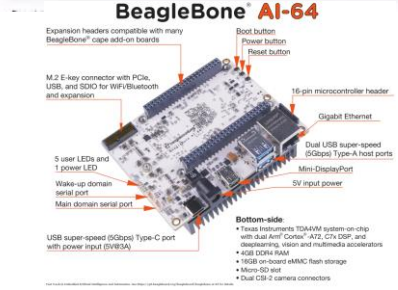
Audio
4 x SSI (I2S)
1 x SRC



TDA4VM multi-core embedded vision processor







Industrial edge compute based on NVIDIA

Lexmark

optra edge

NVIDIA

And so many other cores are supported

Tensilica Datasheet

Xtensa LX7 Processor

cadence

High-performance, configurable, and extensible controllers and DSPs

Xtensa LX7 Processors for Today's SoC Challenges

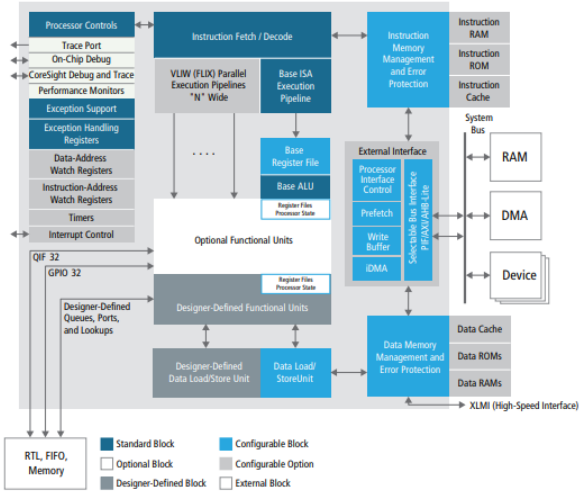
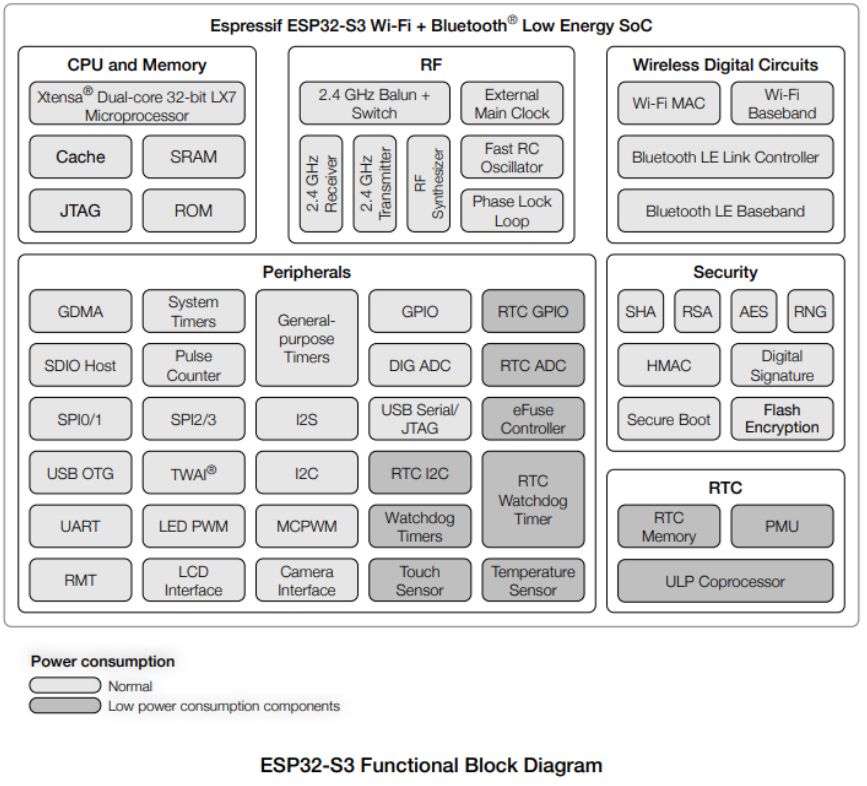


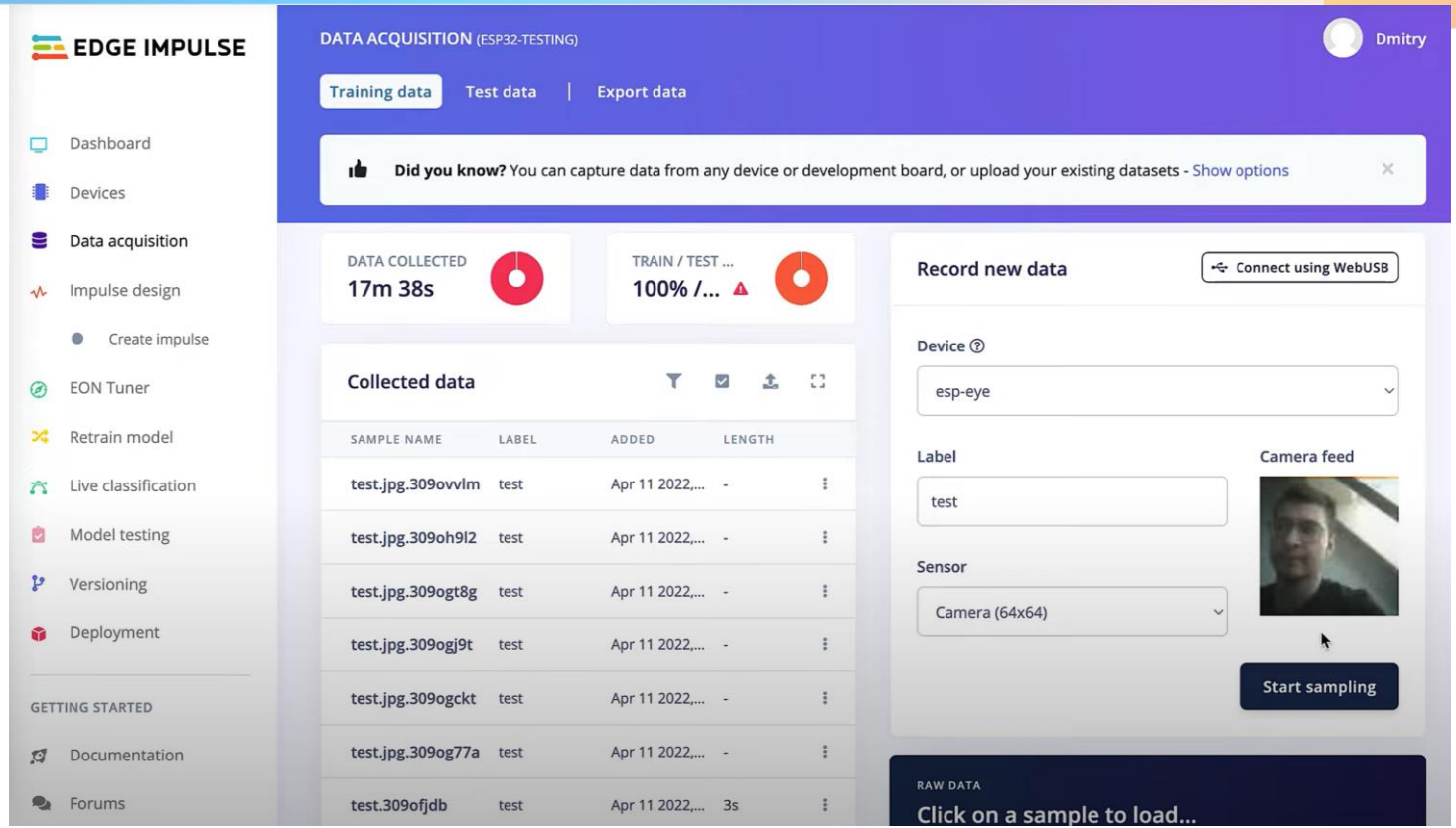
Figure 1: Block diagram of Xtensa LX7 processor architecture

ESP32-S3



ESP32-S3 Functional Block Diagram

And so many other cores are supported

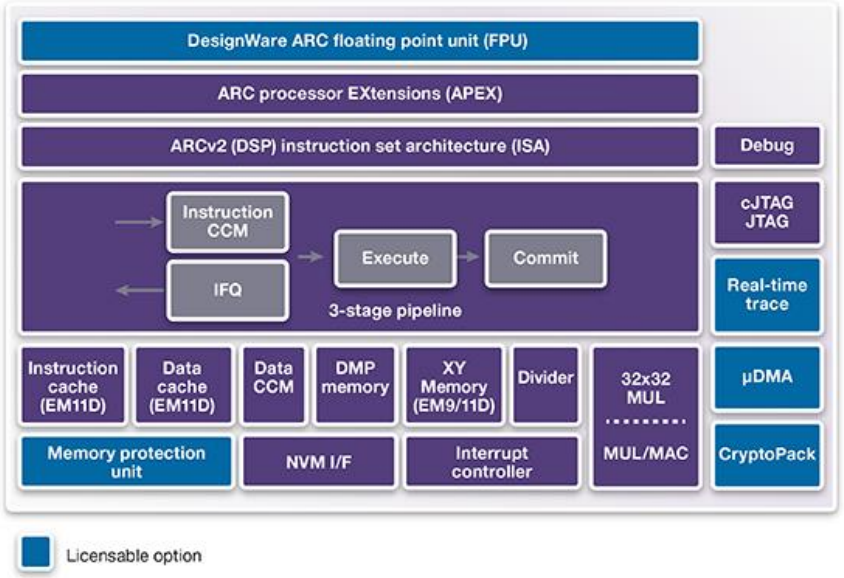


And so many other cores are supported

HX6537-A

WE-I Plus ASIC

SYNOPSYS® ARC EM9D / EM11D Processors

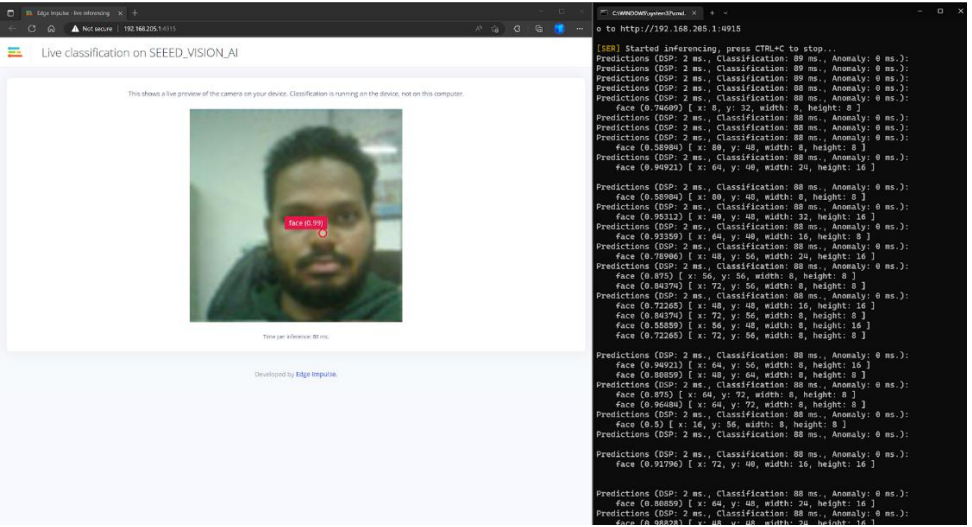


And so many other cores are supported

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

Sensor Audio Vision

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

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

EDGE IMPULSE