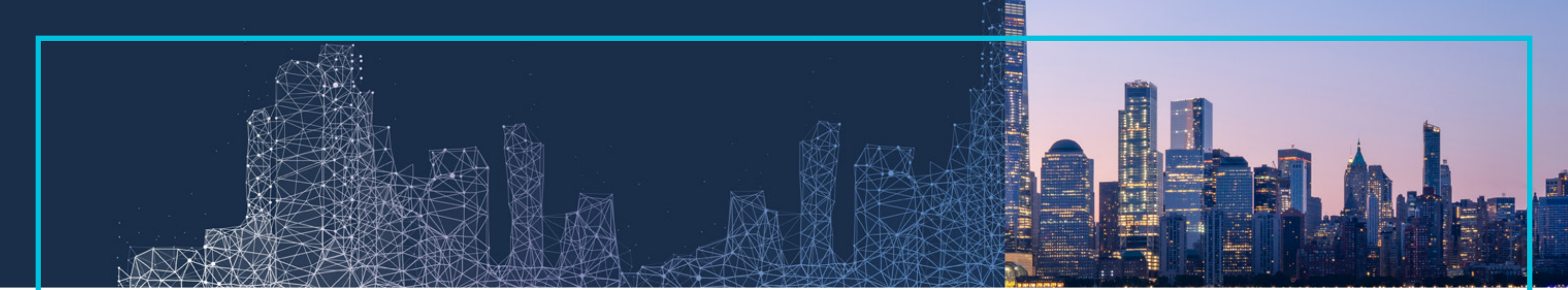
A Guide to Endpoint AI Solutions for Small, Power-constrained Devices

Cortex-M

Ethos-U

Software and Tools

Ecosystem



As the IoT intersects with artificial intelligence (AI) advancements and the rollout of 5G, more on-device intelligence means that smaller, cost-sensitive devices can be smarter and more capable. They also benefit from greater privacy, availability, and responsiveness due to less reliance on the cloud or internet. By delivering this intelligence on microcontrollers designed securely from the ground up, Arm is reducing silicon and development costs. To make it easier and more efficient to innovate devices of the future, Arm is optimizing the software development process and speeding up time to market for product manufacturers and developers who are looking to enhance digital signal processing (DSP) and machine learning (ML) capabilities on-device.

Why Arm?

Trusted Partner for Success

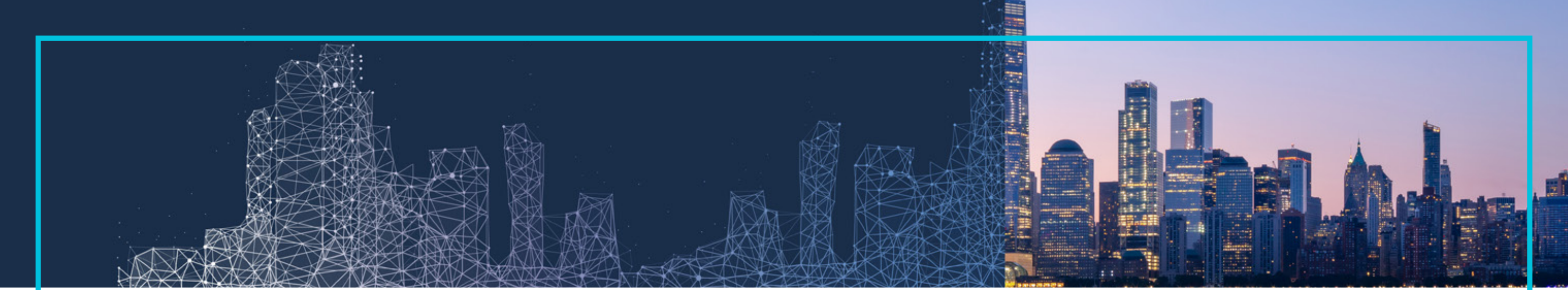
Achieve success by choosing proven Arm technology that has been integrated into billions of devices to date

Power-efficient Design

Arm IP offers power-efficient solutions that are designed for discrete processing and microcontrollers

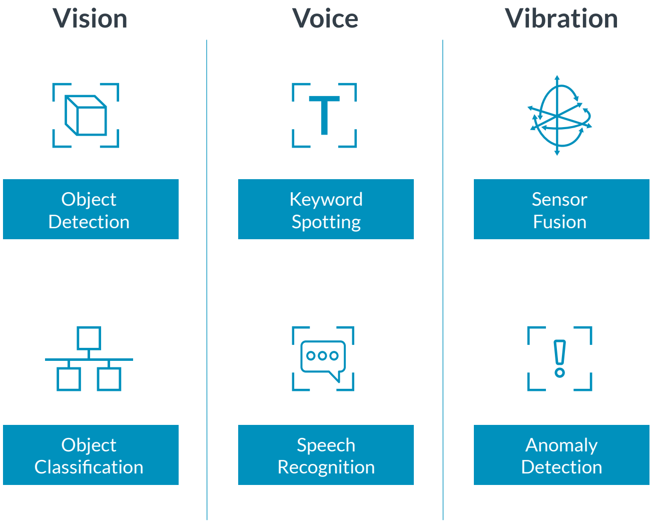
Streamlined Development Process

Arm's platform for building and deploying endpoint AI devices enables software developers to unlock the full potential of Cortex-M based devices by supporting any workload, device, software stack and cloud



Endpoint AI is Driven by Vision, Voice, Vibration

Endpoint AI is the convergence of IoT and ML. McKinsey has identified more than 100 use cases across 11 sectors which could create a value of $250 billion in hardware value for edge and endpoint processing by 2025. By analyzing these use cases specifically for endpoint AI, we can identify that the main use cases align around three areas: Vision, Voice, and Vibration. These categories are applicable to all IoT segments, from smart home to healthcare to industrial, and include a wide range of functions.



Vision

Object Detection

Object Classification

Voice

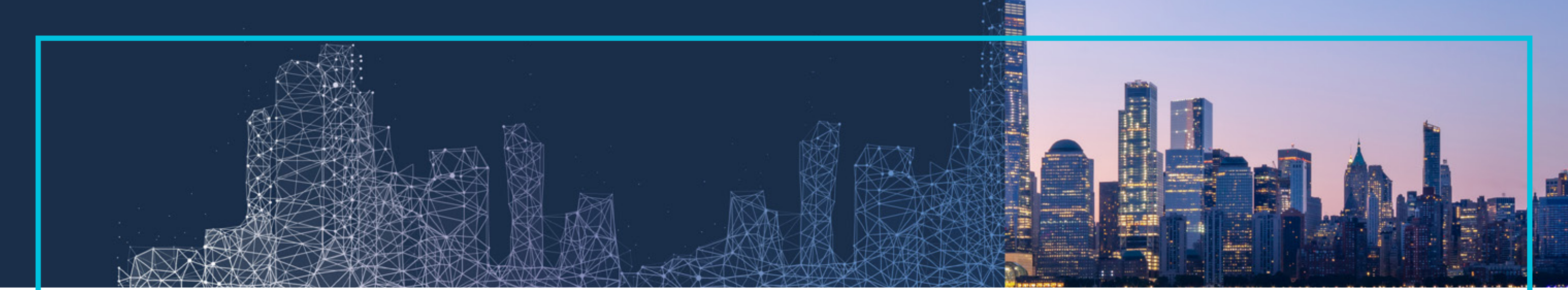
Keyword Spotting

Speech Recognition

Vibration

Sensor Fusion

Anomaly Detection



Vision

Image and Low-power Object Detection

Vision functions include:

- Visual inspection, counting, and sorting
- Object detection and classification
- Visual identification and authentication

Typical applications range from processing of images in smart doorbells to industrial applications.

Vision applications require low-cost imaging sensors working with microcontrollers, which are economical to deploy.

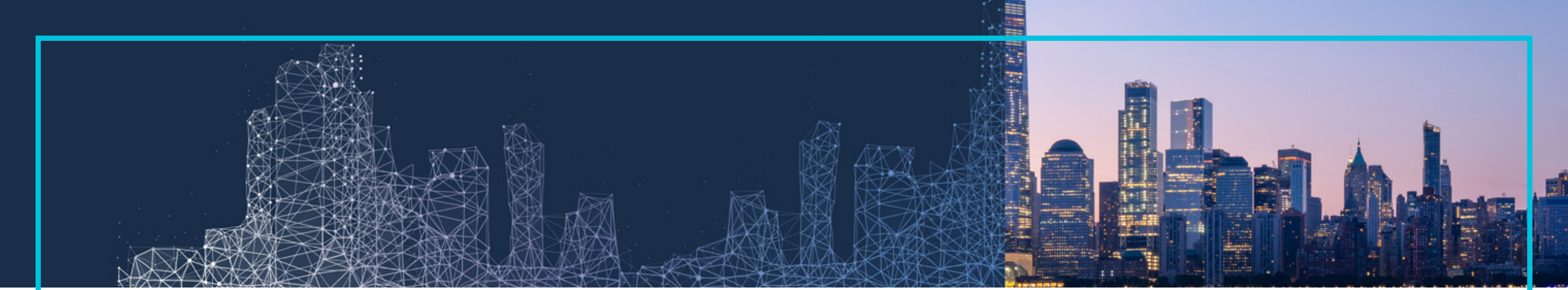
Example

Facial recognition can be used to make sure that the person using the cooker in the kitchen is an adult and not a child. If the sensor detects a child, then the cooker is locked and cannot be used.

Machine Vision with OpenMV

The OpenMV Cam is a low-cost, extensible, Python-powered machine vision module based on a Cortex-M processor. See how the device can be used to acquire, digitize, and process images to take an appropriate action.

Get started with the OpenMV Cam for image classification



Voice

Voice Identification, Sound Recognition, and Audio Processing

Voice functions include:

- Audio pre-processing and post-processing
- Keyword spotting or limited commands
- On-device natural language processing (NLP) without the cloud
- Voice biometrics and user identification
- Sound recognition and classification

Typical applications include voice activation of devices, which incorporates audio pre-processing such as beam forming or noise cancellation, post-processing such as MP3 decode, and sound recognition of important events such as smoke alarms or a baby crying.

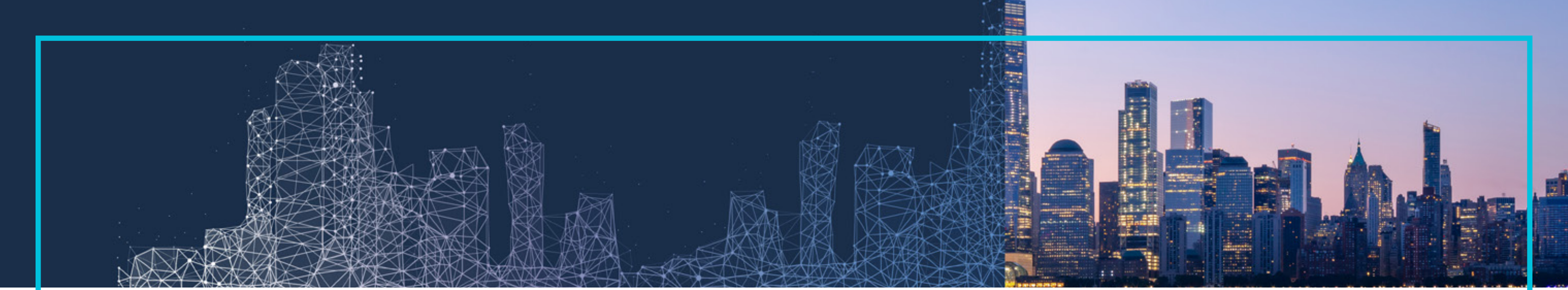
Example

A smart wearable could catch any sound in your environment and alert you when something unexpected happens: when the water is still running or the cooker gas is still on.

Voice Recognition with Alexa on Cortex-M

See how to build low-power voice-enabled products, such as light switches, thermostats, and small appliances, with Alexa Voice Service (AVS) on Cortex-M-based microcontrollers.

Get started with Alexa for voice recognition



Vibration

Using any Signal From the Physical World, Relying on Sensors

Vibration functions include:

- Predictive analytics
- Anomaly detection
- Sensor inputs for context
- Security and threat analysis

Predictive maintenance is a significant growth driver in many mass-market applications as machine downtime costs are significant.

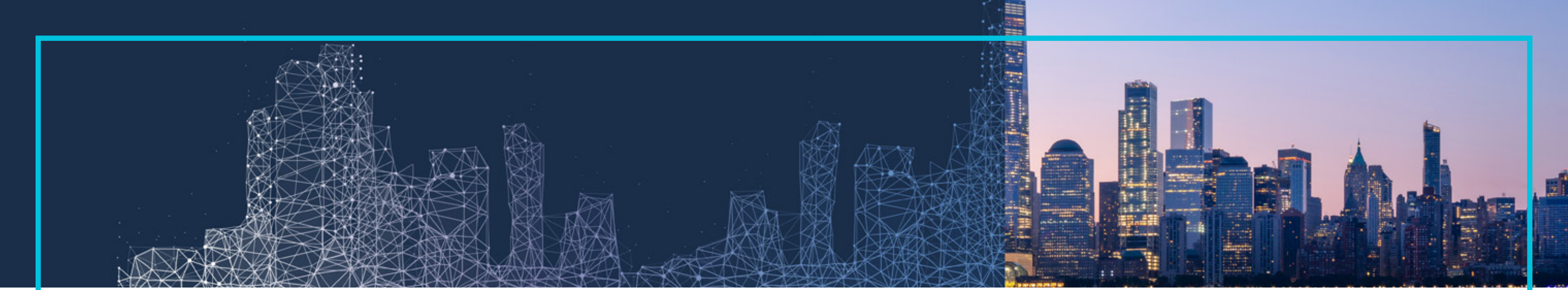
Example

Sensors in shoes could monitor your walking or running style. The results could be used to detect diseases or as feedback on your running technique.

Motor Control and Predictive Maintenance with Renesas

RA6T1 microcontrollers from Renesas are based on Cortex-M4 and optimized for enhanced motor control and support predictive maintenance with Google's TensorFlow Lite for microcontrollers (TFLµ)

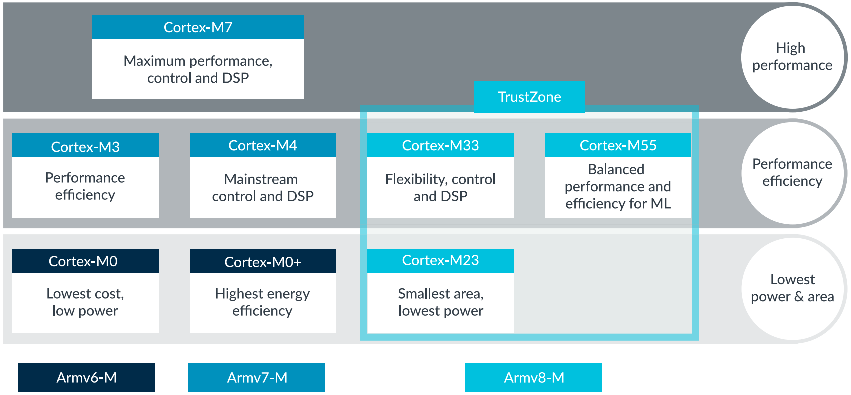
Get started with Renesas and TFLµ for predictive maintenance



What are the Different Arm Processors for Endpoint AI Devices?

Arm Cortex-M processors are optimized to enable cost and energy-efficient microcontrollers. These processors are based on the M-Profile architectures that provide low-latency and a highly deterministic operation for deeply embedded systems.

The Cortex-M55 processor is Arm's most AI capable Cortex-M processor and the first to feature Arm Helium vector processing technology for enhanced, energy-efficient DSP and ML performance.



Cortex-M7
Maximum performance, control and DSP

High performance

TrustZone

Cortex-M3
Performance efficiency

Cortex-M4
Mainstream control and DSP

Cortex-M33
Flexibility, control and DSP

Cortex-M55
Balanced performance and efficiency for ML

Performance efficiency

Cortex-M0
Lowest cost, low power

Cortex-M0+
Highest energy efficiency

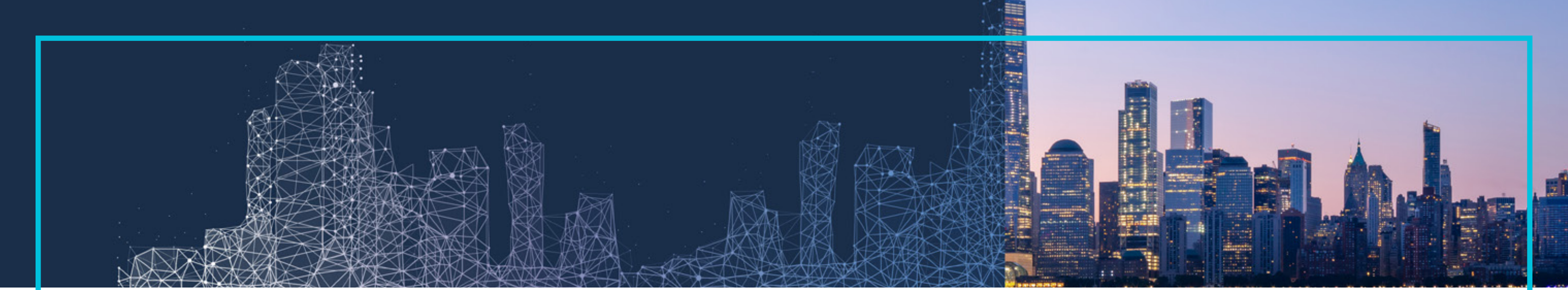
Cortex-M23
Smallest area, lowest power

Lowest power & area

Armv6-M

Armv7-M

Armv8-M

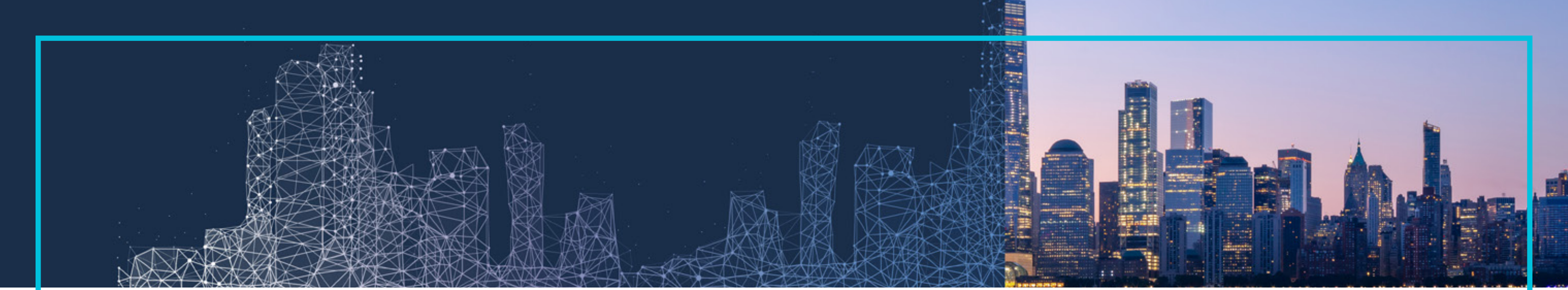


How to Accelerate ML Inference Even Further?

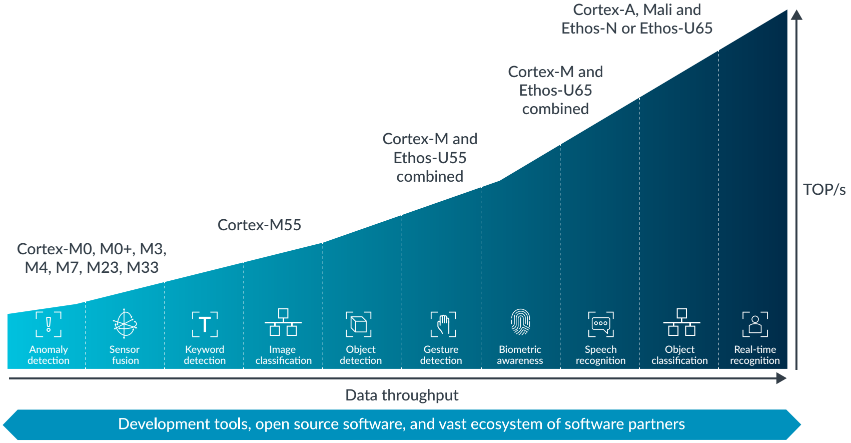
Combining Arm Cortex-M processors with Arm Ethos-U microNPUs (Neural Processing Units) accelerates ML inference and brings innovative AI applications with high performance and efficiency to embedded systems.

Ethos-U55 and Ethos-U65 address different performance points. The workloads that must be run, cost targets, and energy efficiency requirements typically drive the decision making process.

Ethos-U55	Ethos-U65
Designed for SRAM + flash	Designed for SRAM + DRAM and/or flash
Energy efficiency	High performance
Ideal in Cortex-M based microcontroller systems for area-constrained, battery-powered devices	Ideal in higher performance Cortex-M based microntroller systems or Cortex-A based systems that require running complex workloads under a rich OS



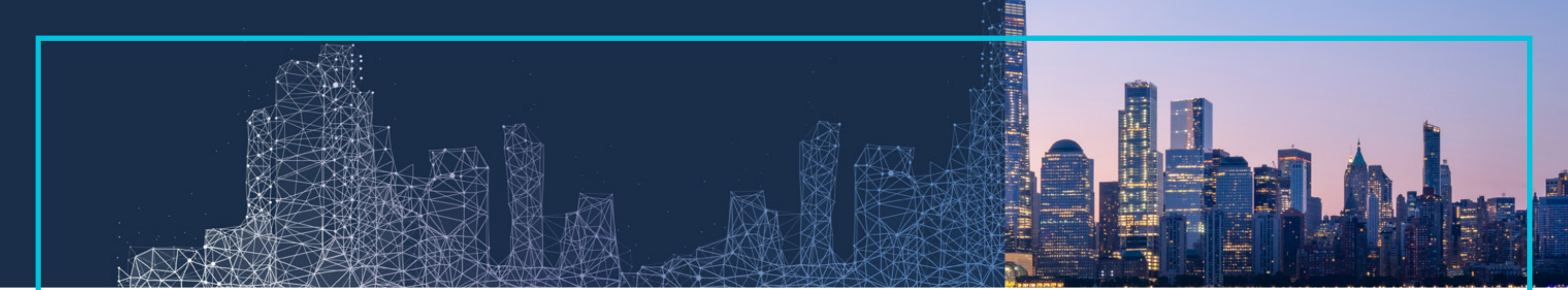
Arm Technology Provides the Right Feature Set for Vision, Voice, and Vibration Endpoint AI Applications



Get Started with Arm Technology

Arm Flexible Access is a fast and low-risk path to market for Endpoint AI applications.

Arm Flexible Access allows SoC designers to download, experiment, design and manufacture with a wide range of Arm IP, including Cortex M55 and Ethos-U55. Tools, training, and engineering support are provided. As license fees are not payable until the finished design is ready to tape out, Flexible Access members can start, stop or pivot their projects at zero commercial risk.



Unleash Developer Creativity for Endpoint AI

The IoT relies on technologies that can orchestrate the different kinds of software applications and keep devices performing optimally at scale. Arm empowers millions of software developers around the world with IoT and ML technologies, making it easier and more efficient to innovate devices of the future. Arm's platform for building and deploying endpoint AI devices enables software developers to unlock the full potential of Cortex-M based devices by supporting any workload, device, software stack and cloud.

Any Workload
ML, DSP, traditional embedded, real-time

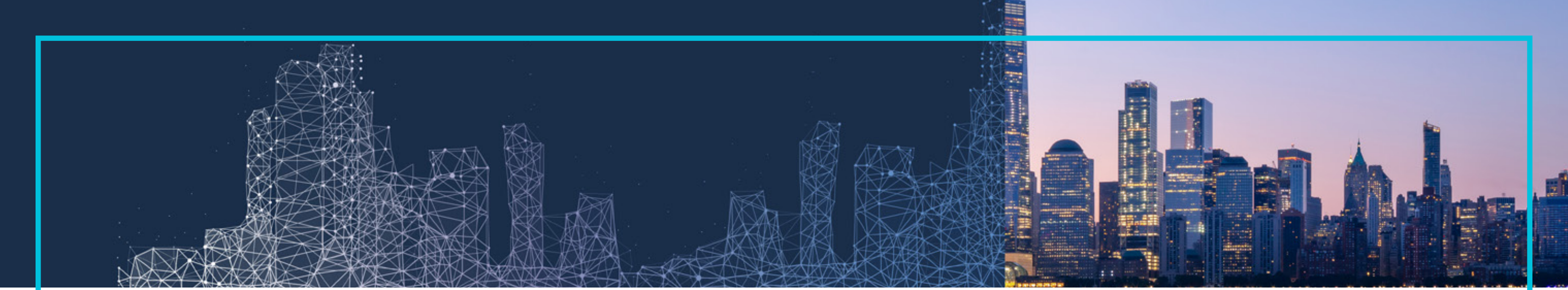
Any Device
Support a wide range of MCUs and development boards

Any OS
Support for popular RTOS, IoT stacks and embedded middleware

Any Cloud
Software and workflows to connect devices to any cloud

Get Started with Tools and Software:

- IDE targeted towards Cortex-M processors: Arm Keil MDK
- Software development tool suite: Arm Development Studio
- Free IoT operating system: Mbed OS
- Open-source software libraries: CMSIS-DSP and CMSIS-NN
- Built-in support for common ML frameworks, such as TensorFlow Lite Micro
- Functionally accurate programmer's view model: Fast Models
- Functional model of a system based on the Cortex-M55 processor enabling execution of a software stack ahead of silicon: Corstone-300 Ecosystem FVP

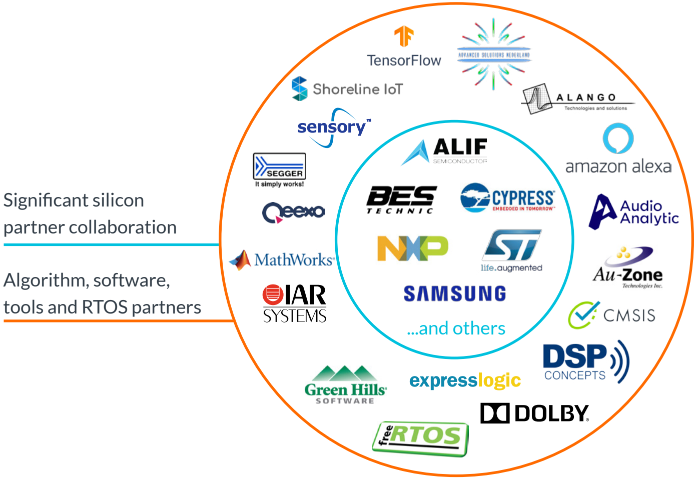


Accelerate AI Innovation with Arm's Partner Ecosystem

The Arm ecosystem comprises of leading global AI software and hardware providers who are ready to accelerate your product journey wherever you want to make compute happen.

Our extensive AI ecosystem simplifies deployment on endpoint AI devices by providing best-in-class hardware, tools, algorithms, and applications for developers to easily prototype and produce AI devices.

Learn more about the [Arm AI Partner Program]

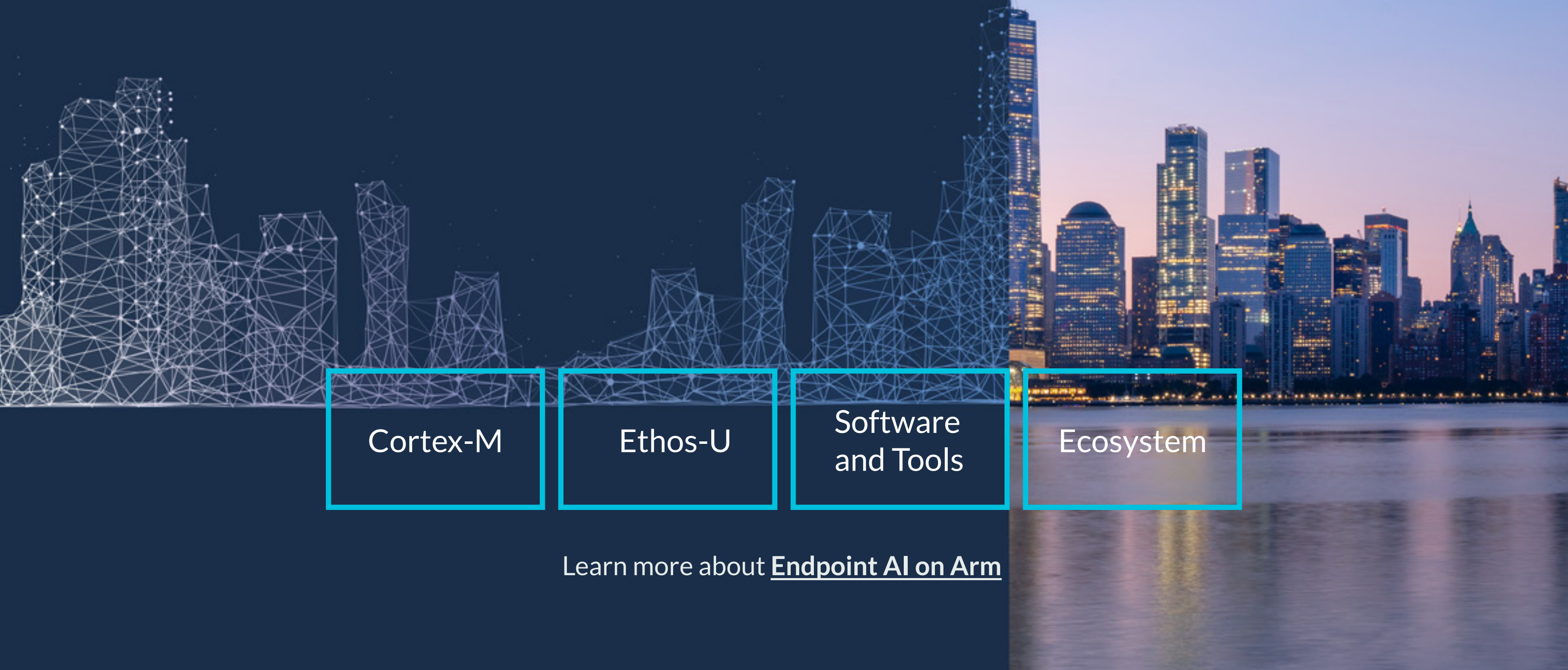


Significant silicon partner collaboration

Algorithm, software, tools and RTOS partners

...and others

A Guide to Endpoint AI Solutions for Small, Power-constrained Devices



Cortex-M

Ethos-U

Software and Tools

Ecosystem

Learn more about **Endpoint AI on Arm**