Whitepaper

How Hardware Accelerated Machine Learning Transforms the Landscape For Embedded IoT Devices

Version 1.0

Table of Contents

1 - Introduction

While microcontrollers have long served as the backbone of embedded systems, many emerging applications demand more inferencing capability than traditional microcontrollers can deliver. On the other hand, employing a microprocessor-type host for these applications can be cost prohibitive and power hungry. There is a compelling case for microcontrollers with integrated hardware machine learning acceleration, bridging the gap between limited microcontroller capabilities and high-powered microprocessors.

The Arm Ethos-U55 machine learning processor, or microNPU, now delivers the hardware acceleration needed to enable a new class of enhanced embedded microcontrollers that deliver vastly greater machine learning performance than conventional MCUs based on Cortex-M cores.
The Ensemble family from Alif Semiconductor is among the first of this new generation of devices to reach the market, delivering more than 480 times the inferencing performance of the fastest conventional MCUs based on Cortex-M cores. The family contains single core and dual core Cortex-M55 MCUs, accelerated with Ethos-U55, as well as multicore fusion processors that combine Cortex-M55, Ethos-U55, and Cortex-A32.

2 - Machine Learning in the Edge: This is Only the Beginning

Machine learning inference on network edge devices is becoming increasingly ubiquitous. Using embedded machine learning inference to handle workloads such as pattern matching and anomaly detection has enabled developers to accelerate the performance and reduce the power consumption of their applications simultaneously. Ultimately, this enables superior user experiences; faster - sometimes real-time – responses to complex challenges such as activity detection, equipment condition monitoring, voice recognition, natural language processing, and healthcare monitoring.

To help embedded developers exercise the power of machine learning in their applications, established MCU vendors have introduced new tools in their conventional development ecosystems that are specifically aimed at building and optimizing neural networks for implementation within the constraints imposed by typical microcontroller-class processor architectures.

The infusion of machine learning into edge devices, and the tools to enable this trend, have merely whetted developers' (and end users'!) appetites for the enormous possibilities on offer. But there are limitations. Complex deep-learning models need to be simplified or compressed for deployment within typical MCU resource constraints. Real-time inferencing imposes strict latency requirements to ensure timely decision-making and responsiveness. Often, machine learning inference in the edge must strike a balance between accuracy and computational efficiency to ensure rapid and reliable inferencing with acceptable precision.

The scope for machine learning inference in the edge has become viewed in terms of "the three Vs" – vibration, voice, and video. While fast, low-power vibration monitoring using neural networks is comfortably within the capabilities of today's embedded-class MCUs, applications that demand voice processing capabilities beyond basic keyword detection can be taxing. Real-time response when handling video is typically not possible.

Demand for the next generation of machine learning applications is already growing, calling for more inferencing capability than traditional microcontrollers can provide. These include smart surveillance systems that need real-time object detection and tracking, which requires efficient inferencing for video analytics. In addition, autonomous drones require on-board machine learning inference for obstacle avoidance and navigation. Intelligent voice assistants, health monitoring devices, and industrial IoT applications necessitate local inferencing for enhanced functionality. Low-power MCUs require additional hardware acceleration to handle these types of machine learning algorithms efficiently.

3 - Realizing Hardware Acceleration

The time is right to realize hardware acceleration of machine learning workloads, suitable for integration within the typical microcontroller restrictions on silicon area and power consumption. And it's here now: the Arm Ethos-U55 is a new type of processor, or microNPU, designed to accelerate inference within those constraints. Working together with the Cortex-M55, currently Arm's most powerful embedded MCU core, Ethos-U55 can unleash 480 times greater inferencing performance than the Cortex-M55 can manage working alone.

The Ethos-U55 microNPU (Figure 1) is optimized to accelerate machine learning workloads based on convolutional and recurrent neural network (CNN, RNN) models. Architected for low power operation and designed for seamless interactions with the Cortex-M55 embedded core, it contains a configurable multiply-accumulate (MAC) engine that can handle 128-bit and 256-bit MACs. Also supporting 8-bit (int8) and 16-bit (int16) quantization, the microNPU enables models to be shrunk by up to 75%.

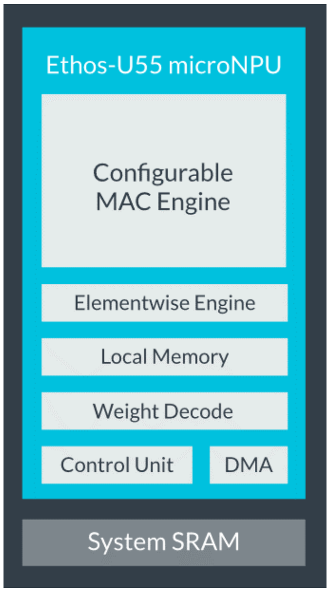

Figure 1. Ethos-U55 microNPU.

Alif Semiconductor has made the significant investment needed to bring the two together, enabling embedded developers to take advantage of the available performance gain and deliver the applications the market is now ready for. The Ensemble family contains the E1 and E3 single and dual Cortex-M55 MCUs with single and dual Ethos-U55 accelerators, and the higher-performing E5 and E7 fusion processors that combine a dual Cortex-M55/Ethos-U55 machine learning engine with single and dual Cortex-A32 application processors running at 800MHz.

The Ensemble E3 is the first in this new family of hardware-accelerated MCUs to enter production. Alif presented demonstrations and impressive initial data at the 2021 Arm DevSummit and has since significantly expanded the number of evaluated use cases and benchmarks. As an example, a MobileNet V2 1.0 model trained on the ImageNet dataset executes 135 times faster using Alif's microNPU accelerators than when executed on a Cortex-M55 MCU working alone, achieving an execution time of 20ms compared with almost 3 seconds when using the Cortex-M55 core. Keep in mind that the Cortex-M55 operates at considerably higher performance than previous-generation Cortex-M cores. The measured energy used per inference also drops dramatically. The accelerated operation is 108x more power efficient, consuming only 0.86mJ compared with 62.4mJ.

The data shown in Table 1 illustrates the inference acceleration and energy savings achieved with Ensemble E3 leveraging the Ethos-U55 and Cortex-M55, compared to Cortex-M55 alone, when handling popular embedded machine learning models. Models are optimized selectively for high performance and high efficiency, as indicated.

High Efficiency (HE) System: Cortex-M55 and Ethos-U55 128MAC at 160MHz

Model	Accelerated inferencing				CPU-bound inferencing (on Cortex-M55)				Improvement w. Acceleration	
	Time ms	Power mW (Δ)	Current mA (Δ)	Energy mJ (Δ)	Time ms	Power mW (Δ)	Current mA (Δ)	Energy mJ (Δ)	Time	Energy efficiency
KWS[1]: MicroNet Medium (ARM)	15.9	8.8	2.6	0.14	326	3.4	1.0	1.27	21x	19x
Object Detection[2]: YOLO-Fastest (face trained)	18.6	14.2	4.2	0.27	1373	5.4	1.6	8.3	74x	67x
Auto Speech Recognition[4]: Tiny ASR (Wav2letter)	78.6	10.0	3.0	0.69	8562	7.4	2.2	62.5	109x	104x

High Performance (HP) System: Cortex-M55 and Ethos-U55 256MAC at 400MHz

Model	Accelerated inferencing				CPU-bound inferencing (on Cortex-M55)				Improvement w. Acceleration	
	Time ms	Power mW (Δ)	Current mA (Δ)	Energy mJ (Δ)	Time ms	Power mW (Δ)	Current mA (Δ)	Energy mJ (Δ)	Time Improve	Energy Improve
KWS[1]: MicroNet Medium (ARM)	6.8	25.7	7.6	0.17	137	19.3	5.7	2.62	20x	18x
Object Detection[2]: YOLO-Fastest (face trained)	7.3	33.8	10.0	0.25	657	21.3	6.3	13.7	90x	76x
Image Classification[3]: MobileNet v2	20.1	43.3	12.8	0.86	2707	24.0	7.1	62.4	135x	108x
Auto Speech Recognition[4]: Tiny ASR (Wav2letter)	27.9	29.6	8.9	0.89	4534	17.3	5.2	77.8	163x	138x

1. KWS: From ARM MicroNets paper. Quantized int8, trained on 'Google Speech Commands' dataset. Model footprint: 154KB MRAM, 28KB SRAM
2. Object Detection: 192x192 grayscale resolution & color. Quantized int8, trained on 'WIDER FACE' dataset. Model footprint: 431KB MRAM, 433KB SRAM
3. Image Classification: 224x224 24bit resolution & color. Quantized int8, trained on 'ImageNet' dataset. Model footprint: 3,552KB MRAM, 1,47KB SRAM
4. ASR: Tiny Wav2letter Pruned slotted into ARM's ML demo app, running the ASR use case. MRAM=2346.06KB, SRAM=1197.20KB

Table 1

Moreover, as an MCU, the Ensemble E3 leverages extensive system-on-chip integration to eliminate several external ICs from the product bill of materials, including memories, power-management IC (PMIC), and secure element. Figure 2 shows the on-chip resources available to handle image input, machine learning execution, and image output for facial recognition and tracking with the Ensemble E3.

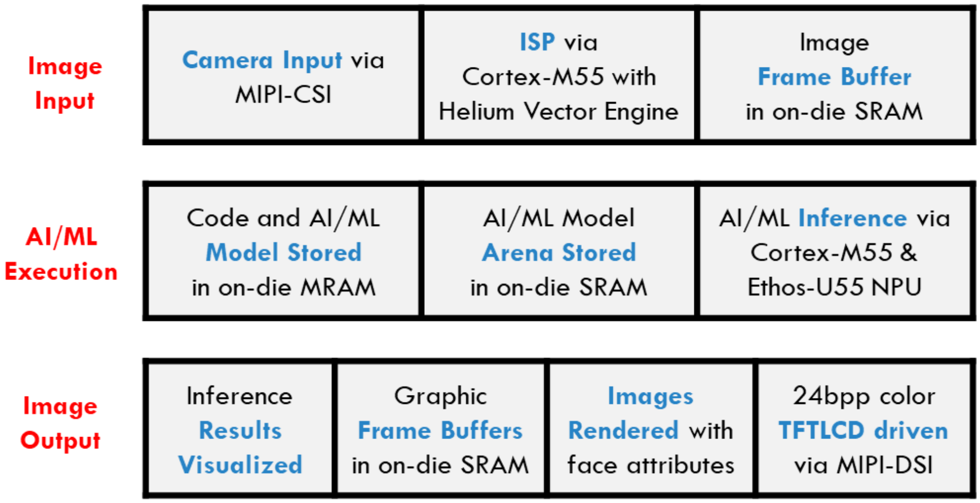

Image Input	Camera Input via MIPI-CSI	ISP via Cortex-M55 with Helium Vector Engine	Image Frame Buffer in on-die SRAM	
AI/ML Execution	Code and AI/ML Model Stored in on-die MRAM	AI/ML Model Arena Stored in on-die SRAM	AI/ML Inference via Cortex-M55 & Ethos-U55 NPU	
Image Output	Inference Results Visualized	Graphic Frame Buffers in on-die SRAM	Images Rendered with face attributes	24bpp color TFTLCD driven via MIPI-DSI

Figure 2. System-on-chip integration with Ensemble E3 MCU.

At Embedded World 2023, Alif Semiconductor demonstrated a facial recognition and tracking application, hosted on Ensemble E3, running at 148 inferences per second, which is significantly faster than is currently possible using conventional MCUs.

4 - Realizing AI Applications on Ensemble Devices

The Ensemble ecosystem contains software development tools and evaluation boards to assist application bring-up on these AI-accelerated MCUs.

Typical machine learning applications that can now be hosted on edge devices, delivering seamless natural user experiences with high accuracy and low latency include real-time object detection and recognition in equipment such as surveillance systems, smart cameras, and autonomous vehicles, enabling quick decisions based on the local environment.

Anomaly Detection in sensor data can benefit from improved performance in predictive maintenance for industrial machinery, detecting abnormalities in healthcare monitoring devices, and identifying security breaches in smart home systems.

Natural Language Processing (NLP) applications for speech recognition, language translation, and voice-assisted applications can move to the edge and reduce reliance on cloud services, enhancing privacy and reducing latency for tasks like voice-controlled home automation and voice assistants.

Real-time gesture recognition enables intuitive human-machine interaction in applications like augmented reality, virtual reality, and gaming.

Sentiment analysis of text and social media data can be performed on edge devices, allowing for real-time feedback and personalized experiences in applications like chatbots, customer service, and social media monitoring.

In addition, applications like e-commerce, content streaming, and personalized healthcare can be hosted on edge devices leveraging machine learning models to provide personalized recommendations based on user behavior and preferences and hence minimize communication with cloud servers.

5 - Conclusion

Low-power machine learning inference optimized for edge devices has quietly transformed users' experiences of equipment and applications from smart watches and healthcare wearables to home digital assistants, smart appliances, and smart factories. While enabling more satisfying interactions with the technology that powers everyday life and work, this quantum leap in performance has also alerted the world to new possibilities and further raised expectations for immediate responses to challenges that demand complex processing, particularly those that involve rapid video analytics such as facial recognition and tracking, object detection, image classification, and keyword spotting.

The accelerated inference performance needed is now possible with Alif's Ensemble MCUs. These devices enable familiar applications to deliver vastly improved user experiences at low power and low cost, while also enabling more complex machine learning workloads to migrate to the network edge and so benefit from lower latency, reduced power consumption, reduced reliance on network bandwidth, and increased data security.

6 - Document History

Version	Change Log
1.0	Initial public release