# Accelerating Machine Learning Compute for the IoT and Embedded Market

Tanuj Arora, Product Manager


Machine learning (ML) technology is expanding rapidly to impact all markets, including the deeply embedded space. It is expected that AI-enabled IoT shipments will grow at a rapid rate in the coming years. The existence of a dedicated $200M VC fund from Amazon for voice startups also points to a future prevalent with 'Alexa' type personal assistants in many devices everywhere. While voice as a user interface has gained a lot of popularity, many applications use vision as well.

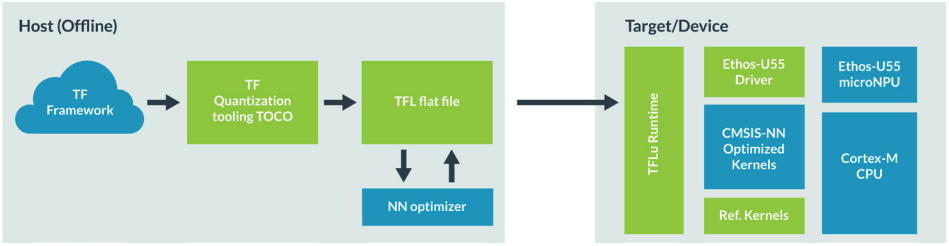
Several of these new applications, including voice assistants, medical diagnosis and treatment equipment, are possible due to the progress made in the field of neural networks. This is placing an ever-increasing demand on processing capabilities at the endpoint. The cost of processing on the cloud is too high, not just in terms of the monetary value of using a cloud service, but also in terms of power and bandwidth capacity. Additionally, processing in the cloud requires mitigating latency issues, raises security concerns, and can reduce reliability. There is a growing privacy concern as well, and generally, people prefer their data to be restricted to the device they own. All these factors drive a need to efficiently accelerate neural networks in devices that are the more resource-constrained.

Arm provides an unmatched solution to this problem by accelerating ML workloads on existing Arm Cortex-M based systems. Billions of embedded devices that use Arm Cortex-M processors already exist in the world and can easily accelerate ML workloads through the use of optimized libraries provided within CMSIS-NN. CMSIS-NN is open source and newly optimized libraries are being added at each quarterly release. To increase the ML performance of these systems, Arm has announced new technologies:

1. Cortex-M55 processor: The latest Cortex-M processor that increases the ML performance by up to 15x as compared to the previous Cortex-M generations
2. Ethos-U55 microNPU: Ethos-U55 is a first generation microNPU that can work with a Cortex-M processor. It allows acceleration of neural networks in an extremely low-area with low-power consumption. Together, Cortex-M55 and Ethos-U55 deliver up to a 480x ML performance uplift compared to previous Cortex-M generations. Plus, under the same software stack, Ethos-U55 increases the ML performance of a Cortex-M55 system by up to 32x.

## Optimized Flow for Efficient Neural Network Processing

This section describes the flow for using a network model trained in TensorFlow and running it on a combined Cortex-M processor and Ethos-U55 microNPU.




The process starts by training or acquiring a TensorFlow model that is to be accelerated. The model is then quantized to 8-bit integer format and converted to the standard TensorFlow Lite flat file format. The NN optimizer tool, provided by Arm, reads this TensorFlow Lite flat file as an input and formats it to make it ready for deployment.

TensorFlow Lite Micro is an offline runtime created to execute within the constraints of embedded devices. The TensorFlow Lite flat file created offline on the host is deployed on the target device. The flat file contains information on which layer of the neural network executes on Ethos-U55 versus the attached Cortex-M processor. The layers supported by Ethos-U55 are accelerated on it and the remaining layers execute on the attached Cortex-M. The layers that execute on the Cortex-M processor are accelerated through the CMSIS-NN software library if the corresponding kernel is available. Otherwise, the TensorFlow Lite Micro reference kernels are used.

To summarize, a neural network can be efficiently accelerated in an extremely small area and power envelope using the following:

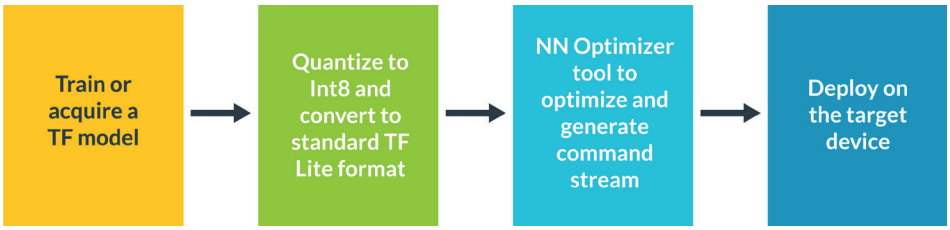
1. TensorFlow Lite micro
2. NN Optimizer Tool
3. Ethos-U55
4. Cortex-M55 and CMSIS-NN

This supports the deployment of new use cases and applications previously impossible in embedded devices.

## 1. TensorFlow Lite for Microcontrollers

TensorFlow Lite for Microcontrollers is a version of TensorFlow Lite, specifically designed to execute ML models on microcontrollers and embedded devices with only a few kilobytes of memory.

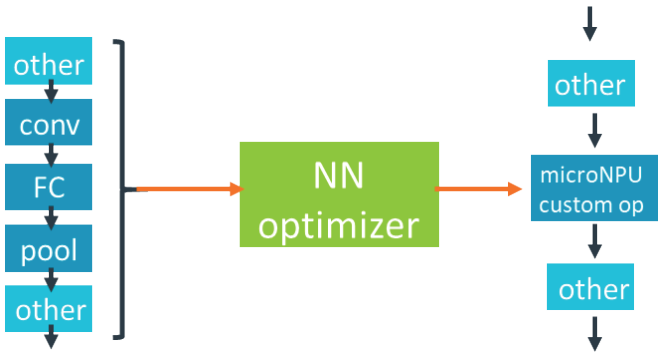
The following figure describes the easy flow of using TensorFlow Lite micro to deploy a model on an embedded platform with the Ethos-U55 microNPU.




Arm is working on developing TensorFlow Lite Micro in an open source collaboration with Google to help accelerate neural networks in an extremely small memory footprint. TensorFlow Lite Micro doesn't require operating system support, standard C or C++ libraries, or dynamic memory allocation. The core runtime fits in 16KB on a Cortex-M3 processor, and with enough operators to run a speech keyword detection model, takes up a total of 22KB. [1]

## 2. The NN Optimizer Tool

The NN Optimizer is a tool provided by Arm to format a neural network for the Ethos-U55 microNPU. It takes a trained neural network in a TensorFlow Lite flat file as an input and formats that network to output a modified flat file ready to be deployed on the target device, as seen in the diagram below:




After reading the flat file, the tool identifies the subgraphs that can execute on the Ethos-U55 microNPU and optimizes the scheduling of these subgraphs. Several factors are taken into consideration for this optimization:

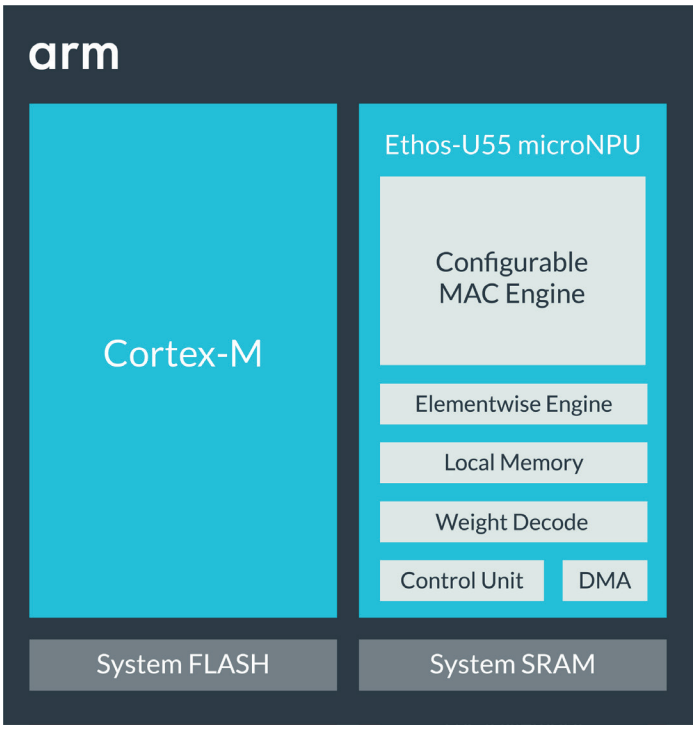
- Fusing of multiple operators
- Balancing available bandwidth vs compute
- Prefetching certain parameters from Flash
- Scheduling across several layers to reduce SRAM footprint

Next, the tool compresses the weights to reduce the SRAM and Flash footprint thereby improving performance of memory bandwidth bound layers. The compression is extremely important as it allows the execution of bigger networks that wouldn't fit in the available memory before. Finally, a command stream is generated for the microNPU and the modified flat file is written.

### 3. Ethos-U55 microNPU

Ethos-U55 is a new class of ML processor called a microNPU and specifically designed to accelerate ML inference in area-constrained embedded and IoT devices. Paired with a Cortex-M processor, the microNPU's configurability allows designers to target a wide variety of AI applications with:

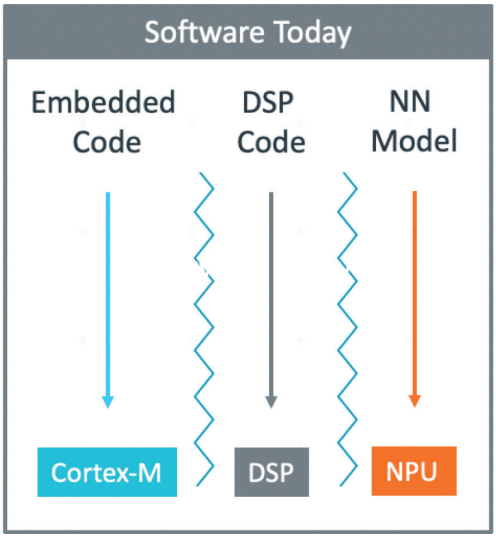
- Four different configurations: 32/64/128/256 MACs/cycle
- Maximum performance up to a 0.5 TOP/s in 16nm process
- Four possible host processors: Cortex-M55, Cortex-M7, Cortex-M33 and Cortex-M4




Ethos-U55 accelerates a fixed set of high compute operators in hardware. Other operators run on the attached microcontroller. It is the easiest way to add more neural network acceleration capability to Cortex-M based systems in a very small silicon area cost.
The smallest version of the IP can fit in approximately 0.1mm^2 area in 16nm process.

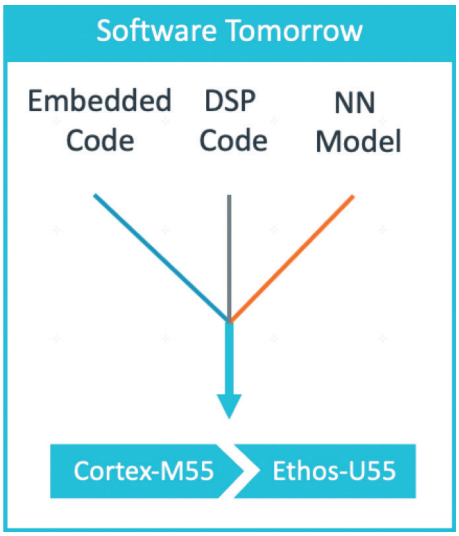
Ethos-U55 leverages a single Cortex-M toolchain to simplify and reduce AI application development time. Most of the applications running ML today run embedded code, neural network models and DSP code. Using multi-vendor toolchain results in:

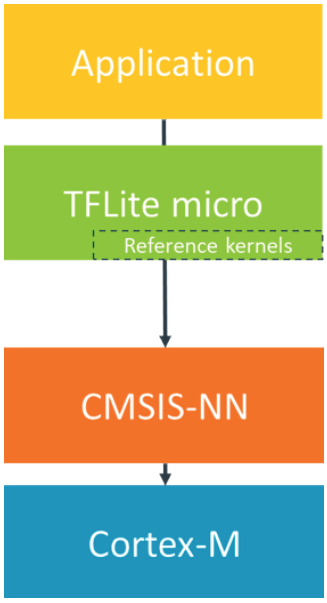
- Multiple software development flows
- Multiple toolchains, complex debug
- Increased programming, maintenance, and support issues




Ethos-U55 and Cortex-M55 provide an opportunity to run this code on a single toolchain. The following features help reduce complexity and eases software development:

- Unified software development flow
- Optimized open source libraries
- Full Integration with Arm Development Studio






### 4. Cortex-M55 Processor and CMSIS-NN

Cortex-M55 is the first processor with Helium vector extensions for enhanced performance and efficiency. Cortex-M55 provides up to 15x increase in ML performance and 5x increase in signal processing performance as compared to the previous generation Cortex-M processors. CMSIS-NN is a library of optimized low-level kernels for Cortex-M CPUs. It can be called from TensorFlow Lite Micro or bare metal implementations.

It is an open source library available with an Apache-2 license. Key operators are accelerated through CMSIS-NN with fallback to reference kernels in TensorFlow Lite Micro.

## Summary

The greatest potential for the next computing revolution lies in scaling AI to the billions of smaller, power-constrained endpoint devices. Innovative signal processing and ML techniques will open up new opportunities for SoC architects to deliver these new levels of efficient AI performance for microcontrollers.

The Cortex-M55 processor, Ethos-U55 microNPU and Arm's industry-leading embedded ecosystem of software libraries and tools support will bring AI to the billions, removing barriers to ML adoption and deployment. These processors will securely and efficiently increase ML and signal processing performance for the next generation of world changing IoT and embedded devices. They extend the performance of Arm's AI platform for microcontroller-based endpoint devices, offering silicon providers a more diverse range of hardware choices and empowering developers to deliver this next revolution in computing

For further information, explore the below links:

- Ethos-U55 microNPU web page
- Cortex-M55 web page including processor datasheet
- CMSIS-NN open source software library

## References

[1] https://www.tensorflow.org/lite/microcontrollers