Arm Ethos-U55

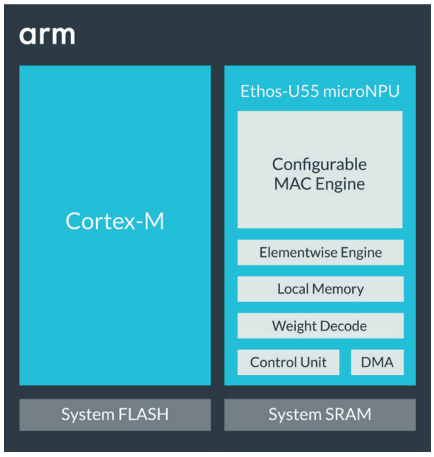
microNPU

arm

Product Brief

KEY FEATURES & BENEFITS

- **Partner Configurable**
  Multiple configurations allow designers to rapidly target a wide variety of AI applications with up to 480x increase in performance
- **Extremely Small Area**
  90% energy reduction in about 0.1mm² for AI applications in cost-sensitive and energy-constrained devices.
- **Single Toolchain**
  A unified toolchain for Ethos-U55 and Cortex-M helps ease developer use and creation of AI applications.
- **Futureproof**
  Native support for the most common ML network operations, including CNN and RNN, with flexibility for future ML innovations.



Arm Ethos-U55 enables powerful embedded ML Inference, with partner configurable options that help accelerate time to market

Embedded ML Inference for Cortex-M systems

A new class of machine learning (ML) processor, called a microNPU, specifically designed to accelerate ML inference in area-constrained embedded and IoT devices. The Ethos-U55 combined with the AI-capable Cortex-M55 processor provides a 480x uplift in ML performance over existing Cortex-M based systems.

Highlights

- **Energy Efficient**
  Provides up to 90 percent energy reduction for ML workloads, such as ASR, compared to previous Cortex-M generations.
- **Network Support**
  Flexible design supports a variety of popular neural networks, including CNNs and RNNs, for audio processing, speech recognition, image classification, and object detection.
- **Future-Proof Operator Coverage**
  Heavy compute operators run directly on the micro NPU, such as convolution, LSTM, RNN, pooling, activation functions, and primitive element wise functions. Other kernels run automatically on the tightly coupled Cortex-M using CMSIS-NN.
- **Offline Optimization**
  Offline compilation and optimization of neural networks, performing operator, and layer fusion as well as layer reordering to increase performance and reduce system memory requirements by up to 90 percent. Delivers increased performance and lower power compared to non-optimized ordering.
- **Integrated DMA**
  Weights and activations are fetched ahead of time using a DMA connected to system memory via AXI5 master interface.
- **Element Wise Engine**
  Designed to optimize for commonly used elementwise operations such as add, mul and sub for commonly used scaling, LSTM, GRU operations. Enables future operators composed of these similar primitive operations.
- **Mixed Precision**
  Supports Int-8 and Int-16: lower precision for classification and detection tasks; high-precision Int-16 for audio and limited HDR image enhancements.
- **Lossless Compression**
  Advanced, lossless model compression reduces model size by up to 75 percent, increasing system inference performance and reducing power.

KEY USE CASES FOR ETHOS-U55

- Object classification
- Object detection
- Face detection/identification
- Human pose detection/ hand-gesture recognition
- Image segmentation
- Image beautification
- Super resolution
- Speech recognition
- Sound recognition
- Noise cancellation

MARKET SEGMENTS


Constrained Embedded


Tiny ML Endpoints


Wearables


AR/VR


ML Islands in SoCs


Premium Mobile


Smart Cameras


Automotive Powertrains


Sensor Fusion


Industrial


Environmental Sensors


Infrastructure

System Integration

Works seamlessly with Cortex-M devices and Arm Corstone systems, and allows designers to configure and build high-performance, power-efficient SoCs while further differentiating with combinations of Arm processors and their own IP.

Specifications

Key Features	Performance (At 1GHz)	64 to 512 GOP/s
	MACs (8x8)	32, 64, 128, 256
	Utilization on popular networks	Up to 85%
	Data Types	Int-8 and Int-16
	Network Support	CNN and RNN/LSTM
	Winograd Support	No
	Sparsity	Yes
Memory System	Internal SRAM	18 to 50 KB
	External On Chip SRAM	KB to Multi-MB
	Compression	Weights only
	Memory Optimizations	Extended compression, layer/operator fusion
Development Platform	Neural Frameworks	TensorFlow Lite Micro
	Operating Systems	RTOS or bare-metal
	Software Components	TensorFlow Lite Micro Runtime, CMSIS-NN, Optimizer, Driver
	Debug and Profile	Layer-by-layer visibility with PMUs
	Evaluation and Early Prototyping	Performance Model, Cycle Accurate Model, or FPGA Evaluations

To find out more about the Ethos-U55 processor, visit developer.arm.com/ethos-u55