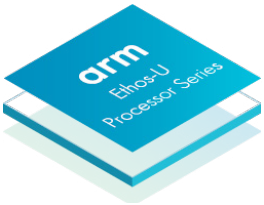
arm ETHOS

NPU

PRODUCT BRIEF

Arm Ethos-U Processor Series



KEY FEATURES AND BENEFITS

Performance Efficiency

Unlock new edge AI use cases with minimal impact to existing design across a wide range of ML performance requirements (GOP/s to TOP/s).

Flexible Integrations

Build low cost, highly efficient systems in high-performance Cortex-A, and low-power Cortex-M embedded systems.

Optimized Design

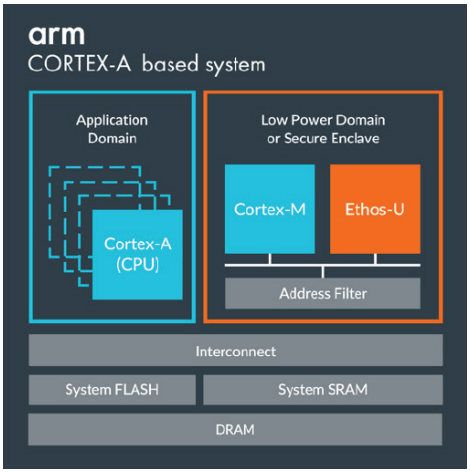
Highly optimized with integrated DMA, MAC array, and element wise engines.

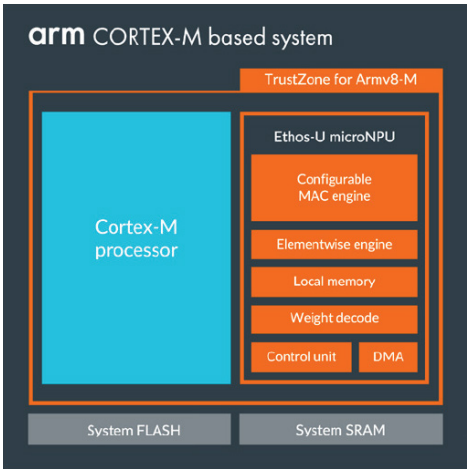
UNLOCK A NEW WORLD OF EDGE AI APPLICATIONS

Build low-cost, highly efficient AI solutions in a wide range of embedded devices with Arm's Ethos-U NPUs which enable systems based on Arm Cortex and Arm Neoverse. Ethos-U provides a scalable range of performance and memory interfaces and integrates low-power Cortex-M SoCs, as well as SoCs based on high-performance Arm Cortex-A, Cortex-R, and Arm Neoverse. Develop, deploy, and debug AI applications using a common toolchain across Arm Cortex and Ethos-U processors.

Unified Software and Tools

Develop, deploy, and debug AI applications using a common toolchain across Arm Cortex and Ethos-U processors.





Ethos-U85 can target numerous different applications with use in high-performance Cortex-A or low power Cortex-M based embedded devices.

HIGHLIGHTS

Energy Efficient

Provides up to 90% 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 NPU, such as convolution, transformer, LSTM, RNN, pooling, activation functions, and primitive element-wise functions. Other kernels can run automatically on the tightly coupled Cortex-M using CMSIS-NN or Arm Compute Library on Cortex-A.

Reduce Memory Footprint

Up to 70% reduction in model size with compression allows for the execution of larger networks and speeds up the execution of networks.

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%. Delivers increased performance and lower power compared to non-optimized ordering.

Develop Multiple Products

Products Target multiple markets with a single processor IP architecture that provides the system flexibility to configure the performance and desired memory system and OS.

Quickly Create Applications

Accelerate time to market by leveraging the Arm AI ecosystem with partners developing optimized algorithms ahead of hardware availability.

KEY USE CASES FOR THE ETHOS-U PROCESSOR SERIES

- Object classification
- Object detection
- Face detection/identification
- Human pose estimation
- Image segmentation
- Image beautification
- Super resolution
- Speech recognition
- Sound recognition
- Noise cancellation
- Speech synthesis
- Language translation
- Natural language processing

MARKET SEGMENTS


Mobile


Smart Camera


Smart Home


STB/DTV


Consumer


AR/VR


Medical


Robotics


Drones


Rich IoT

Specifications

Key Features		Ethos-U55	Ethos-U65	Ethos-U85
	Performance (At 1 GHz)	64 to 512 GOP/s	512 GOP/s to 1 TOP/s	256 GOPS/s to 4 TOP/s
	MACs (8x8)	32, 64, 128, 256	256, 512	128, 256, 512, 1024, 2048
Memory System	Internal SRAM	18 to 50 KB	55 to 104 KB	29 to 267 KB
	System Interfaces	Two 64-bit AXI	Two 128-bit AXI	Up to six 128-bit AMBA 5 AXI
	External Memory	SRAM and Flash	SRAM, DRAM, and/or FLASH	SRAM, DRAM and/or FLASH
Development Platforms	Recommended Host CPU	Cortex-M55, Cortex-M7, Cortex-M4, Cortex-M33	Cortex-M55, Cortex-M7	Cortex-M85, Cortex-M55, Cortex-M7 Cortex-A520, Cortex-A510, Cortex-A57, Cortex-A55, Cortex-A53, Cortex-A35
	Operating Systems	Bare-metal or RTOS	Bare-metal, RTOS, or Linux	RTOS, Bare-metal, or Linux

To find out more about the Ethos-U processor series, visit developer.arm.com/ethos-u

arm