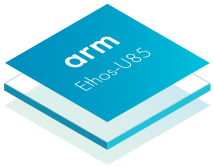
arm ETHOS

NPU

PRODUCT BRIEF

Arm Ethos-U85



KEY FEATURES AND BENEFITS

- Extending Performance and Efficiency
  Unlock future edge AI use cases with 20% more energy efficiency than Arm Ethos-U65 and scalable performance from 128 to 2048 MACs, providing up to 4 TOPs at 1GHz.
- Enabling Generative AI at the Edge
  Allowing native support for transformer networks, along with support for the Tensor Operator Set Architecture (TOSA) standard.
- Leveraged in a System-Level Solution
  Ethos-U85 is integrated into the subsystem of Arm Corstone-320, with Arm Cortex-M85, Arm DMA-350, and Arm Mali-C55.
- Unified Software and Tools
  Develop, deploy, and debug AI applications using a common toolchain across Arm Cortex and Ethos-U processors, and the Ethos-U ecosystem.

Ethos-U85 can target diverse edge AI applications.

POWERING EDGE AI INNOVATION

Ethos-U85 provides support for transformer-based models at the edge, which are the basis for newer language and vision models used to build edge AI solutions. Ethos-U85 scales from 128 to 2048 MAC units and is 20% more energy efficient than Ethos-U65.

Built upon previous Ethos-U generations, Ethos-U85 offers the same toolchain so partners can benefit from seamless migration and leverage investments in Arm-based ML.

KEY USE CASES FOR ETHOS-U85

- Speech-to-text translation
- Live translation
- Small language models
- 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
- Image de-noising

HIGHLIGHTS

- **New Use Cases**

  Enables future edge AI use cases, including generative AI on the edge, with native support for transformer networks.

- **Support Complex Models**

  Run complex models in heterogenous systems, either under a rich OS in Arm Cortex-A systems with wider AXI interfaces (128-bit) and DRAM support or an RTOS in Arm Cortex-M systems.

- **Integrated DMA**

  Weight and activations are fetched ahead of time using a DMA connected to system memory via an AXI5 master interface.

- **Energy Efficiency**

  Provides up to 20% energy efficiency improvements than Ethos-U65.

- **Future-Proof Operator Coverage**

  Heavy compute operators run directly on the NPU, such as Transpose, Gather, Matmul, Resize Bilnear, ArgMax, along with convolution, LSTM, RNN, pooling, activation functions, and primitive element-wise functions.

- **Offline Optimization**

  Increases performance and reduces system memory requirements by up to 90% with offline compilation and optimization of neural networks, performing operator, and layer fusion, as well as layer reordering. Delivers increased performance and lower power compared to non-optimized ordering.

MARKET SEGMENTS


TinyML


High-Performance Embedded


Wearables


AR/VR


ML Islands in Socs


Mobile


Smart Cameras


Smart Home


Automotive Powertrain


Sensor Fusion


Industrial Automation


Environmental Sensors


Infrastructure

Element Wise Engine

Designed to optimize for commonly used element-wise operations, such as addition, multiplication, and subtraction for commonly used scaling, LSTM, and GRU operations. Enables future operators to comprise these similar primitive operations.

Mixed Precision

Support for Int-8 weights, and Int-8 or Int-16 for activations: 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%, increasing system inference performance and reducing power.

Specifications

Key Features	Performance (At 1GHz)	256 GOPS/s to 4 TOP/s
	MACs (8x8)	128, 256, 512, 1024, 2048
	Utilization on popular networks	Up to 85%
	Data types	Int-8 weights and Int-16 activations
	Network support	CNN, RNN, and transformer networks
	Winograd support	No
	Sparsity	Yes (2/4 sparsity supported with throughput doubled)
Memory System	Internal SRAM	29 to 267 KB
	System interfaces	Up to six 128-bit AMBA 5 AXI
	External-on-chip SRAM	KB to multi-MB
	Compression	Weights only; both Standard and Fast Weight Decoder
	Memory optimizations	Extended compression, layer/operator fusion, striping capability
Development Platforms	Neural frameworks	TensorFlow Lite Micro
	Operating systems	Bare-metal, RTOS, Linux
	Software components	TensorFlow Lite Micro Runtime, CMSIS-NN, optimizer, driver
	Debug and profile	Layer-by-layer visibility with PMUs, cross trigger interface
	Evaluation and early prototyping	Performance model, FPGA evaluation platforms

To learn more about the Ethos-U85 processor, visit

developer.arm.com/ethos-u85

arm