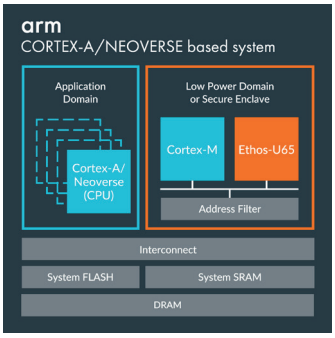
Arm Ethos-U65

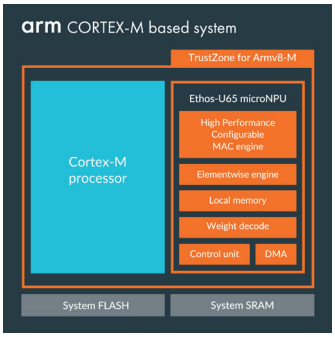
microNPU

Product Brief

KEY FEATURES & BENEFITS

- **Extending Performance and Efficiency**
  Unlock new vision and voice use cases in minimum area with 2x performance uplift (over Ethos-U55) and reach 1 TOP/s in 0.6mm^2 (in 16nm).
- **Flexible Integration**
  Build low-cost, highly efficient systems with rich OS and DRAM support in Cortex-A and Neoverse systems, and on BareMetal or RTOS SRAM/FLASH systems on Cortex-M with the highly successful Ethos-U architecture.
- **Unified Software and Tools**
  Develop, deploy, and debug AI applications with the Arm Endpoint AI solution using a common toolchain across Arm Cortex, Neoverse, and Ethos-U processors.
- **Enhanced Design**
  Supports popular networks with extended operator support, provides wider AXI interfaces, and improves reliability with ECC added into internal RAMs.





Ethos-U65 can target numerous different applications with use in high-performance Cortex-A and Neoverse systems or low-power embedded devices based on Cortex-M.

Powering Innovation in a New World of AI Devices

Build low-cost, highly efficient AI solutions in a wide range of embedded devices with Arm's latest addition to the Ethos-U microNPU family. The Ethos-U65 maintains the power efficiency of the Arm Ethos-U55, while extending its applicability to Arm Cortex-A, Cortex-R and Arm Neoverse-based systems, and at the same time delivers twice the on-device machine learning (ML) performance.

Highlights

- **New Use Cases**
  Enables demanding AI use cases, such as object detection and segmentation, with 150% higher performance (Inf/s) supporting read/write from DRAM.
- **Support Complex Models**
  Process complex workloads under a rich OS in Cortex-A systems with wider AXI interfaces (128-bit) and DRAM support with an average 150% improvement in inf/s for popular networks.
- **Integrated DMA**
  Weights and activations are fetched ahead of time using a DMA connected to system memory via an AXI5 master interface.
- **Energy Efficient**
  Provides up to 90% energy reduction for ML workloads such as ASR, compared to previous Cortex-M generations.
- **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%. Delivers increased performance and lower power compared to non-optimized ordering.
- **Element-wise Engine**
  Designed to optimize for commonly used element-wise operations, such as addition, multiplication, and subtraction for commonly used scaling, LSTM, GRU operations. Enables future operators composed of these similar primitive operations.

KEY USE CASES FOR ETHOS-U65

- 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


tinyML Endpoints


High Performance Embedded


Wearables


AR/VR


ML Islands in SoCs


Mobile


Smart Cameras


Automotive Powertrain


Sensor Fusion


Industrial Automation


Environmental Sensors


Infrastructure

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%, increasing system inference performance and reducing power.

Specifications

Key Features	Performance (At 1GHz)	512 GOPS/s to 1 TOP/s
	MACs (8x8)	256, 512
	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	55 to 104 KB
	System Interfaces	Two 128-bit AXI
	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	Bare-metal, RTOS, Linux
	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-U65 processor, visit developer.arm.com/ethos-u65