Arm Ethos-N37
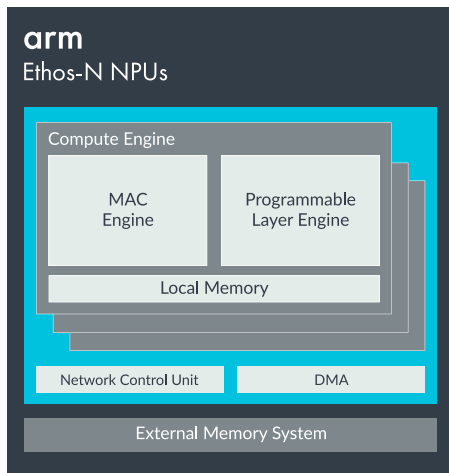## NPU

arm

AT A GLANCE

Optimized for the most cost- and power-sensitive designs, Ethos-N37 delivers premium AI experiences in entry phones, DTVs and other endpoint devices. With the highest performance, open-source software framework and the largest AI ecosystem, the Arm AI platform makes it easy to develop on Arm.

arm
Ethos-N NPUs

Compute Engine
MAC Engine
Programmable Layer Engine
Local Memory
Network Control Unit
DMA
External Memory System

Ethos-N37 entry-level ML inference processor contains four engines.

KEY FEATURES & BENEFITS

✛ **High Efficiency**
Delivers up to 1 TOP/s within 1mm ^{2} using 512 8-Bit MACs.

✛ **Lowest Area**
Smallest-area inference NPU brings battery life savings and quick integration for cost or thermally constrained designs.

✛ **Optimized Design**
Drives up to 225% convolution performance uplift using Winograd on 3x3 kernels, delivering up to 90% MAC utilization.

✛ **Futureproof**
Supports a wide range of existing ML operations and future innovations through firmware updates and compiler technology.

Small Footprint ML Inference Processor

What's New?

✛ **Network Support**
Flexible design supports a variety of popular neural networks, including CNNs and RNNs, for classification, object detection, image enhancements, speech recognition and natural language understanding.

✛ **Futureproof Operator Coverage**
The MAC engine flexibly decomposes arbitrarily sized kernels with stride and dilation modes including convolution, deconvolution, depthwise separable, and vector product. Programable Layer Engines execute layers not supported by the MAC engine, supporting various primitives, activation functions and future operators.

✛ **Mixed Precision**
Supports both Int-8 and Int-16: lower-precision Int-8 for classification and detection tasks; high-precision Int-16 for HDR image enhancements and audio tasks.

✛ **Compression and Winograd Convolution**
MAC engines provide decompression, activation, Winograd transformation and scaling. Winograd accelerates common filters by 225% compared to other NPUs, allowing actual performance to far exceed architectural performance.

✛ **Multicore**
Supports up to eight processors in a tightly coupled cluster, with the ability to process multiple networks in parallel or a single, large network split across cores. Larger configurations of up to 64 cores are supported through Arm CoreLink mesh technology.

✛ **Weight and Feature Map Compression**
Minimizes system memory bandwidth by 1.5-3x, reducing off-chip memory accesses by 90% through extended compression technologies, targeting both weight and activations.

✛ **Security**
Supports TrustZone system security with configurable secure queues for multiple users and flexible processing in the TEE or SEE, providing layered security to protect both ML models and input data.

✛ **System Integration (SMMU)**
ACE-Lite master port and optional SMMU (System Memory Management Unit) integration allows for support and protection of memory and easy handling of multiple users.
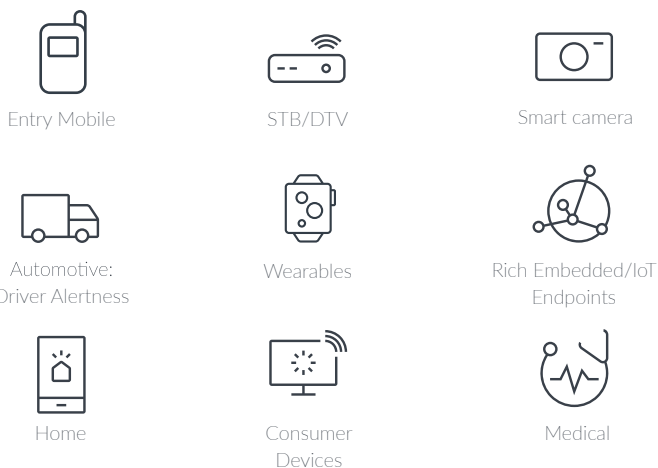
KEY USE CASES FOR ETHOS-N37

+ Object classification
+ Object detection
+ Face detection/identification
+ Human pose detection/ hand-gesture recognition
+ Image segmentation
+ Image beautification
+ Super resolution
+ Framerate adjustment (super slow-mo)
+ Speech recognition
+ Sound recognition
+ Noise cancellation
+ Speech synthesis
+ Language translation

Specifications

Key Features	Performance (at 1GHz)	Up to 1 TOP/s	
	MACs (8x8)	512	
	Data Types	Int-8 and Int-16	
	Network Support	CNN and RNN	
	Efficient Convolution	Winograd support	
	Sparsity	Yes	
	Secure Mode	TEE or SEE	
	Multicore Capability	8 NPUs in a cluster 64 NPUs in a mesh	
Memory System	Embedded SRAM	512KB	
	Bandwidth Reduction	Extended compression technology, layer/ operator fusion	
	Main Interface	1xAXI4 (128-bit), ACE-5 Lite	
Development Platform	Neural Frameworks	TensorFlow, TensorFlow Lite, Caffe2, PyTorch, MXNet, ONNX	
	Neural Operator API	Arm NN, AndroidNN	
	Software Components	Arm NN, neural compiler, driver and support library	
	Debug and Profile	Layer-by-layer visibility	
	Evaluation and Early Prototyping	Arm Juno FPGA systems and cycle models	

Market Segments

Entry Mobile

STB/DTV

Smart camera

Automotive: Driver Alertness

Wearables

Rich Embedded/IoT Endpoints

Home

Consumer Devices

Medical

To find out more about the Ethos-N37 processor, visit

developer.arm.com/ethos-n37

arm