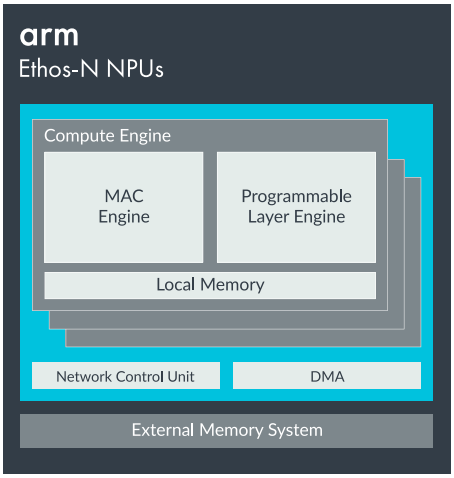
Arm Ethos-N Processor Series

NPU

Product Brief

AT A GLANCE

The Arm Ethos-N processor series delivers the highest throughput and efficiency in the lowest area for machine learning (ML) inference from cloud to edge to endpoint.



Arm Ethos-N NPUs address the ML inference requirements of multiple markets, providing 16, 8 or 4 engines with identical software.

KEY FEATURES & BENEFITS

- **Scalable Performance**
  Delivering up to 4, 2 or 1 TOP/s of single core performance with multicore scalability, supporting up to eight NPUs in a cluster, and up to 64 NPUs in mesh systems.
- **Highly Efficient**
  Achieving up to 5 TOPs/W (including on-chip memory accesses) through internally distributed SRAM, storing data close to the compute elements to save power and reduce DRAM access.
- **Optimized Design**
  Driving up to 225% convolution performance uplift using Winograd on 3x3 kernels, delivering up to 90% MAC utilization.
- **Futureproof**
  Supporting a wide range of existing ML operations, as well as future innovations through firmware updates and compiler technology.

Powering AI Inference from Cloud to Edge to Endpoint

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..

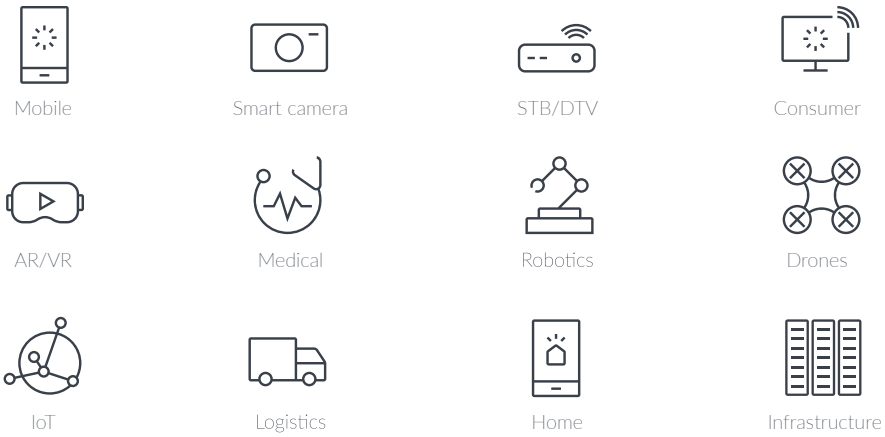
KEY USE CASES FOR THE ETHOS PROCESSOR SERIES

- 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

		Ethos-N77	Ethos-N57	Ethos-N37
Key Features	Performance (at 1GHz)	4 TOP/s	2 TOP/s	1 TOP/s
	Mac/Cycle (8x8)	2048	1024	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	1-4 MB	512 KB	512 KB
	Bandwidth Reduction	Extended compression technology, layer/operator fusion, clustering, and workload tilling		
	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



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