Arm Ethos-N77

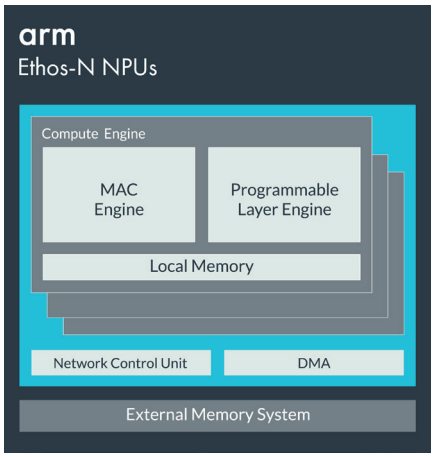
NPU

arm

Datasheet

AT A GLANCE

Based on a new, class-leading architecture, the Arm Ethos-N77 processor's optimized design enables new features, enhances user experience and delivers innovative applications for a wide array of market segments including mobile, IoT, embedded, automotive, and infrastructure. It provides a 64x uplift in efficiency compared to CPUs, GPUs and DSPs through efficient convolution, sparsity and compression.



Ethos-N77 highest efficiency ML inference processor contains 16 compute engines

KEY FEATURES & BENEFITS

- **Highest Performance**
  Delivers up to 4 TOPS of performance (2048 8-Bit MACs), scaling to 100s of TOPs in multicore deployments.
- **Optimized Design**
  Up to 225% convolution performance uplift using Winograd on 3x3 kernels, delivering up to 90% MAC utilization.
- **Highest Efficiency**
  Achieving 5 TOPs/W through internally distributed SRAM, storing data close to the compute elements to save power and reduce DRAM access.
- **Futureproof**
  Supports a wide range of existing ML operations and future innovations through firmware updates and compiler technology.

Highest inference performance and efficiency

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.

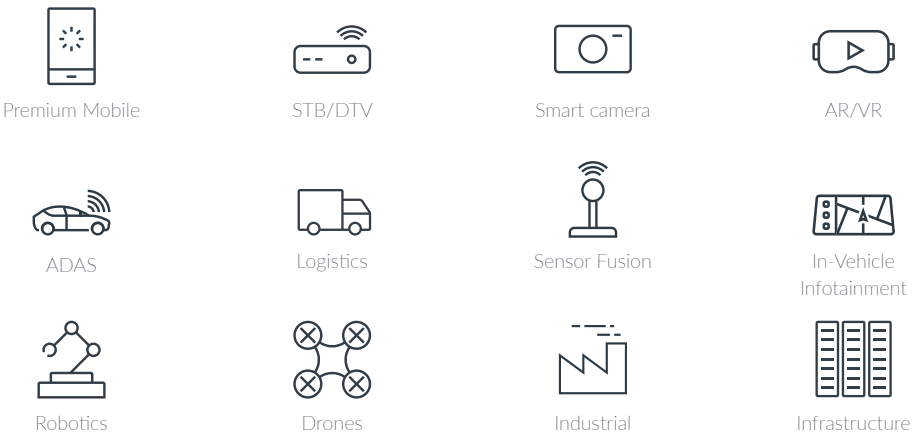
KEY USE CASES FOR ETHOS-N77

- 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)	4 TOP/s
	MACs (8x8)	2048
	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	1–4MB
	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



To find out more about the Ethos-N77 processor, visit developer.arm.com/ethos-n77