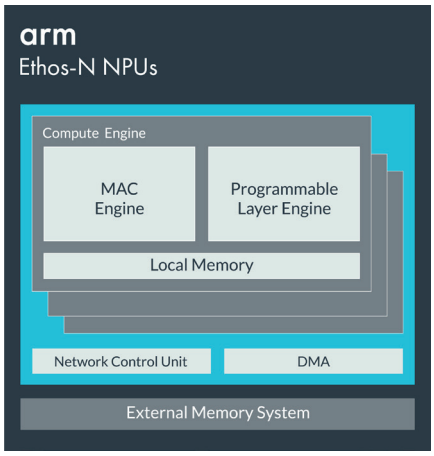
Arm Ethos-N78 Processor

NPU

Product Brief

KEY FEATURES & BENEFITS

- **Increased Performance:** Improves user experience with 2.5x increased single core performance scalable from 1 to 10 TOP/s and beyond through many-core technologies
- **Higher Efficiency:** Up to 40% lower DRAM bandwidth (MB/Infr) and up to 25% increase in efficiency (inf/s/mm2) enables demanding neural networks to be run in diverse solutions
- **Extended Configurability:** Target multiple markets with flexibility to optimize the ML capability with 90+ configurations and the Ethos-N Static Performance Analyzer
- **Unified Software and Tools:** Develop, deploy and debug with the Arm AI platform using online or offline compilation and Arm Development Studio 5 Streamline



Scalable single core from 1 to 10 TOP/s for multiple solutions.

Highly scalable and efficient second-generation Machine Learning processor

Arm's second-generation, highly scalable and efficient processor, the Ethos-N78 enables premium AI solutions with low cost in multiple markets segments. Build new immersive applications with 2.5x increased single core performance now scalable from 1 to 10 TOP/s and beyond through many-core technologies. It provides flexibility to optimize the ML capability with over 90 configurations.

Highlights

- **Longer battery life**
  Up to 40% lower DRAM traffic (MB/Infr) through improved compression, clustering, and cascading
- **Target multiple market segments**
  Scalable single core from 1 to 10 TOP/s and beyond through multi-core and mesh technologies for mobile, auto and infrastructure devices
- **Optimize the ML capability**
  Configurable with more than 90 unique configurations with fine grained optimizations for performance (TOP/s), area (infr/s/mm2), throughput (inf/s) and average DRAM bandwidth (GB/s)
- **Low cost solutions**
  Up to 25% greater performance efficiency (inf/s/mm2) with improved utilization and optimized for use of high latency commodity DRAM
- **Secure platform**
  Supports comprehensive security solution in conjunction with Arm SMMU and CryptoCell IP
- **Inference deployment flexibility**
  Online Android NN or Offline TVM compiled flow based on open source software allows heterogeneous support for ML workloads across the Arm AI platform
- **Ethos-N NPU Static Performance Analyzer (SPA)**
  Enables pre-silicon network performance tuning with interactive speed, bandwidth, and utilization reports

KEY USE CASES FOR THE ETHOS-N78

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

MARKET SEGMENTS


Premium Mobile


Smart Cameras


Smart Home


STB/DTV


Consumer


AR/VR


Medical


Robotics


Drones


Rich IoT


Automotive IVI/ADAS


Infrastructure

Specifications

Key Features	Performance	10, 5, 2, 1 TOP/s
	MAC/Cycle (8x8)	4096, 2048, 1024, 512
	Efficient convolution	Winograd support delivers 2.25x peak performance over baseline
	Configurability	90+ Design Options
	Network support	CNN and RNN
	Data types	Int-8 and Int-16
	Sparsity	Yes
	Secure mode	TEE or SEE
	Multicore capability	8 NPUs in a cluster 64 NPUs in a mesh
Memory System	Embedded SRAM	384KB–4MB
	Bandwidth reduction	Enhanced Compression
	Main interface	1xAXI4 (128-bit), ACE-5 Lite
Development Platform	Neural frameworks	TensorFlow, TensorFlow Lite, Caffe, PyTorch, MXNet, ONNX
	Inference deployment	Ahead of time compiled with TVM Online interpreted with Arm NN Android Neural Networks API (NNAPI)
	Software components	Arm NN, Arm NPU software (compiler and support library, driver)
	Debug and profile	Heterogeneous layer-by-layer visibility in Development Studio 5 Streamline
	Evaluation and early prototyping	Ethos-N Static Performance Analyzer (SPA), Arm Juno FPGA systems, Cycle Models

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