Arm Machine Learning Processor

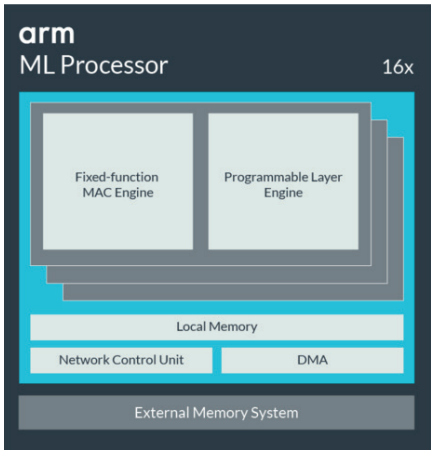
NPU

arm

Datasheet

AT A GLANCE

Based on a new, class-leading architecture, the Arm Machine Learning (ML) 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 massive uplift in efficiency compared to CPUs, GPUs and DSPs through efficient convolution, sparsity and compression.



KEY FEATURES & BENEFITS

- **Outstanding Performance**
  Delivers 4 TOP/s of performance, scaling to hundreds of TOPS in multicore deployments.
- **Highly Efficient**
  Achieving 5 TOPs/W through internally distributed SRAM storing data close to the compute elements to save power and reduce DRAM access.
- **Optimized Design**
  An innovative architecture drives high MAC utilization, improving convolutional efficiency.
- **Futureproof**
  Supports future innovation in network architecture and algorithms through programmable engines.

Industry-leading Inference Performance and Efficiency at the Edge

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**
  MAC engine processes convolution, deconvolution, depthwise separable, vector product and stride modes, plus efficient decomposition of arbitrarily sized kernels. 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 – able to process multiple networks in parallel – or a single, large network split across cores. Larger configurations are supported through Arm CoreLink mesh technology.
- **Weight and Feature Map Compression**
  Minimizes system memory bandwidth by 1.5-3x through a variety of compression technologies, targeting both weight and activation feature maps.
- **Security**
  Supports TrustZone system security with configurable secure queues for multiple users, flexible processing in the TEE or SEE for secure cases like biometric payment, protecting content for high-value media streams.
- **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.

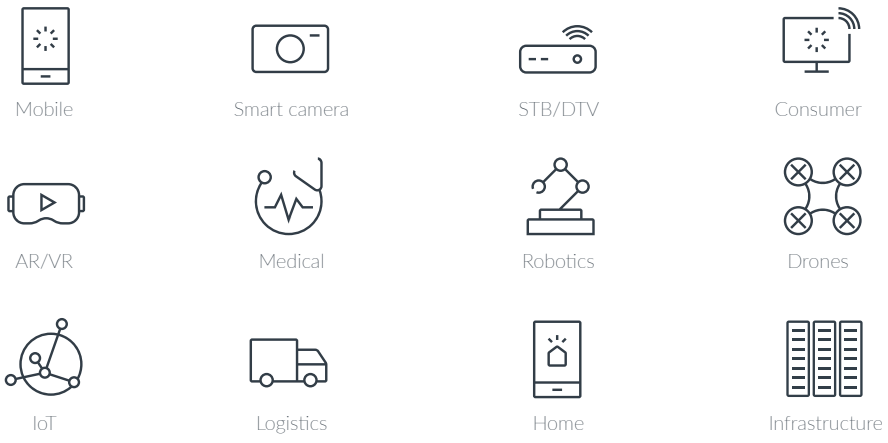
KEY USE-CASES FOR THE ML PROCESSOR

- 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
	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	1MB
	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 Arm Machine Learning processor, visit **www.developer.arm.com/ml-processor**