Arm Ethos-N Processor Series

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.

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

<table>
<thead>
<tr>
<th>Key Features</th>
<th>Ethos-N77</th>
<th>Ethos-N57</th>
<th>Ethos-N37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (at 1GHz)</td>
<td>4 TOP/s</td>
<td>2 TOP/s</td>
<td>1 TOP/s</td>
</tr>
<tr>
<td>Mac/Cycle (8x8)</td>
<td>2048</td>
<td>1024</td>
<td>512</td>
</tr>
<tr>
<td>Data Types</td>
<td>Int-8 and Int-16</td>
<td>CNN and RNN</td>
<td></td>
</tr>
<tr>
<td>Network Support</td>
<td>CNN and RNN</td>
<td>CNN and RNN</td>
<td></td>
</tr>
<tr>
<td>Efficient Convolution</td>
<td>Winograd support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sparsity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Secure Mode</td>
<td>TEE or SEE</td>
<td>TEE or SEE</td>
<td></td>
</tr>
<tr>
<td>Multicore Capability</td>
<td>8 NPUs in a cluster</td>
<td>64 NPUs in a mesh</td>
<td></td>
</tr>
<tr>
<td>Embedded SRAM</td>
<td>1-4 MB</td>
<td>512 KB</td>
<td>512 KB</td>
</tr>
</tbody>
</table>

### Memory System

- Bandwidth Reduction: Extended compression technology, layer/operator fusion, clustering, and workload filling
- Main Interface: 1xAXI4 (128-bit), ACE-5 Lite
- 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

### Development Platform

- Market Segments
  - Mobile
  - Smart camera
  - STB/DTV
  - Consumer
  - AR/VR
  - Medical
  - Robotics
  - Drones
  - IoT
  - Logistics
  - Home
  - Infrastructure

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