Arm Ethos-N Processor Series

NPU

KEY FEATURES & BENEFITS

- **Scalable Performance**
  Delivering up to 10 to 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 with low DRAM traffic (MB/Infr) through compression, clustering, and operator cascading.

- **Optimized Design**
  Driving up to 225% convolution performance uplift using Winograd on 3x3 kernels, delivering up to 90% MAC utilization.

- **Unified Software and Tools**
  Develop, deploy and debug with the Arm AI platform using online or offline compilation and Arm Development Studio 5 Streamline.

Powering AI Inference from Cloud to Edge to Endpoint

**Highlights**

- **Arm AI Platform**
  Arm AI platform optimally supports all popular frameworks and operators with the ability to add new networks/operators on Cortex CPU, Mali GPU, or through Ethos NPU driver updates.

- **Inference Deployment Flexibility**
  Efficiently execute your models using ahead of time compilation with TVM or online interpreted with Arm NN or through Android Neural Networks API (NNAPI).

- **Target Multiple Market Segments**
  Scalable single core from 1 to 10 TOP/s single core and up to 640 TOP/s through multi-core and mesh technologies for mobile, auto and infrastructure devices.

- **Longer Battery Life**
  Up to 40% lower DRAM traffic (MB/Infr) through improved weight and feature map compression, clustering, and cascading.

- **Advanced Operator Cascading**
  Optimized NPU Cascading to improve battery life by automatically scheduling graphs into sub-graphs of connected operators to optimally utilize on device SRAM avoiding power intensive DRAM access.

- **Early Network Performance Feedback**
  Ethos-N NPU Static Performance Analyzer (SPA) allows early performance feedback before hardware availability.

- **Security**
  Supports comprehensive security solution in conjunction with Arm SMMU & CryptoCell IP.

- **System Integration (SMMU)**
  Allows for support and protection of memory and easy handling of multiple users via tight system integration through the ACE-Lite master port and optional SMMU integration.

Arm Ethos-N NPUs address the ML inference requirements of multiple markets with a unified software platform.
## 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

### Key Features

<table>
<thead>
<tr>
<th></th>
<th>Ethos-N78</th>
<th>Ethos-N77</th>
<th>Ethos-N57</th>
<th>Ethos-N37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>10.5, 2, 1 TOP/s</td>
<td>4 TOP/s</td>
<td>2 TOP/s</td>
<td>1 TOP/s</td>
</tr>
<tr>
<td>MAC/Cycle (8x8)</td>
<td>4096, 2048, 1024, 512</td>
<td>2048</td>
<td>1024</td>
<td>512</td>
</tr>
<tr>
<td>Efficient convolution</td>
<td>Winograd support delivers 2.25x peak performance over baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configurability</td>
<td>90+ Design Options</td>
<td>Single Product Offering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network support</td>
<td>CNN and RNN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data types</td>
<td>Int-8 and Int-16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secure mode</td>
<td>TEE or SEE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multicore capability</td>
<td>8 NPUs in a cluster</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>64 NPUs in a mesh</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory System</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Embedded SRAM</td>
<td>384KB – 4MB</td>
<td>1–4 MB</td>
<td>512 KB</td>
<td>512 KB</td>
</tr>
<tr>
<td>Bandwidth reduction</td>
<td>Enhanced Compression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main interface</td>
<td>1xAXI4 (128-bit), ACE-5 Lite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural frameworks</td>
<td>TensorFlow, TensorFlow Lite, Caffe2, PyTorch, MXNet, ONNX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inference deployment</td>
<td>Ahead of time compiled with TVM Online interpreted with Arm NN Android Neural Networks API (NNAPI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software components</td>
<td>Arm NN, Arm NPU software (compiler and support library, driver)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Debug and profile</td>
<td>Heterogeneous layer-by-layer visibility in Development Studio 5 Streamline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation and early prototyping</td>
<td>Ethos-N Static Performance Analyzer (SPA), Arm Juno FPGA systems, Cycle Models</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To find out more about the Ethos processor series, visit [developer.arm.com/ethos](http://developer.arm.com/ethos)

---

All brand names or product names are the property of their respective holders. Neither the whole nor any part of the information contained in, or the product described in, this document may be adapted or reproduced in any material form except with the prior written permission of the copyright holder. The product described in this document is subject to continuous developments and improvements. All particulars of the product and its use contained in this document are given in good faith. All warranties implied or expressed, including but not limited to implied warranties of satisfactory quality or fitness for purpose are excluded. This document is intended only to provide information to the reader about the product. To the extent permitted by local laws Arm shall not be liable for any loss or damage arising from the use of any information in this document or any error or omission in such information.