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
- Frame rate adjustment (super slow-mo)
- Speech recognition
- Sound recognition
- Noise cancellation
- Speech synthesis
- Language translation

### Specifications

<table>
<thead>
<tr>
<th>Key Features</th>
<th>Performance (at 1GHz)</th>
<th>Data Types</th>
<th>Network Support</th>
<th>Efficient Convolution</th>
<th>Sparsity</th>
<th>Secure Mode</th>
<th>Multicore Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 TOP/s</td>
<td>Int-8 and Int-16</td>
<td>CNN and RNN</td>
<td>Winograd support</td>
<td>Yes</td>
<td>TEE or SEE</td>
<td>8 NPUs in a cluster</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64 NPUs in a mesh</td>
</tr>
<tr>
<td>Memory System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Embedded SRAM</td>
<td>1MB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bandwidth Reduction</td>
<td>Extended compression technology, layer/operator fusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Main Interface</td>
<td>1xAXI4 (128-bit), ACE-5 Lite</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development Platform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neural Frameworks</td>
<td>TensorFlow, TensorFlow Lite, Caffe2, PyTorch, MXNet, ONNX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neural Operator API</td>
<td>Arm NN, AndroidNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Software Components</td>
<td>Arm NN, neural compiler, driver and support library</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Debug and Profile</td>
<td>Layer-by-layer visibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evaluation and Early Prototyping</td>
<td>Arm Juno FPGA systems and cycle models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Market Segments

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

To find out more about the Arm Machine Learning processor, visit [www.developer.arm.com/ml-processor](http://www.developer.arm.com/ml-processor)

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.