**Arm Ethos-U85**

**KEY FEATURES AND BENEFITS**

- **Extending Performance and Efficiency**
  Unlock future edge AI use cases with 20% more energy efficiency than Arm Ethos-U65 and scalable performance from 128 to 2048 MACs, providing up to 4 TOPs at 1GHz.

- **Enabling Generative AI at the Edge**
  Allowing native support for transformer networks, along with support for the Tensor Operator Set Architecture (TOSA) standard.

- **Leveraged in a System-Level Solution**
  Ethos-U85 is integrated into the subsystem of Arm Corstone-320, with Arm Cortex-M85, Arm DMA-350, and Arm Mali-C55.

- **Unified Software and Tools**
  Develop, deploy, and debug AI applications using a common toolchain across Arm Cortex and Ethos-U processors, and the Ethos-U ecosystem.

**POWERING EDGE AI INNOVATION**

Ethos-U85 provides support for transformer-based models at the edge, which are the basis for newer language and vision models used to build edge AI solutions. Ethos-U85 scales from 128 to 2048 MAC units and is 20% more energy efficient than Ethos-U65.

Built upon previous Ethos-U generations, Ethos-U85 offers the same toolchain so partners can benefit from seamless migration and leverage investments in Arm-based ML.

Ethos-U85 can target diverse edge AI applications.
**KEY USE CASES FOR ETHOS-U85**

- Speech-to-text translation
- Live translation
- Small language models
- Object classification
- Object detection
- Face detection/identification
- Human pose detection/hand-gesture recognition
- Image segmentation
- Image beautification
- Super resolution
- Speech recognition
- Sound recognition
- Noise cancellation
- Image de-noising

**HIGHLIGHTS**

- **New Use Cases**
  Enables future edge AI use cases, including generative AI on the edge, with native support for transformer networks.

- **Support Complex Models**
  Run complex models in heterogenous systems, either under a rich OS in Arm Cortex-A systems with wider AXI interfaces (128-bit) and DRAM support or an RTOS in Arm Cortex-M systems.

- **Integrated DMA**
  Weight and activations are fetched ahead of time using a DMA connected to system memory via an AXI5 master interface.

- **Energy Efficiency**
  Provides up to 20% energy efficiency improvements than Ethos-U65.

- **Future-Proof Operator Coverage**
  Heavy compute operators run directly on the NPU, such as Transpose, Gather, Matmul, Resize Bilinear, ArgMax, along with convolution, LSTM, RNN, pooling, activation functions, and primitive element-wise functions.

- **Offline Optimization**
  Increases performance and reduces system memory requirements by up to 90% with offline compilation and optimization of neural networks, performing operator, and layer fusion, as well as layer reordering. Delivers increased performance and lower power compared to non-optimized ordering.
**Element Wise Engine**

Designed to optimize for commonly used element-wise operations, such as addition, multiplication, and subtraction for commonly used scaling, LSTM, and GRU operations. Enables future operators to comprise these similar primitive operations.

**Mixed Precision**

Support for Int-8 weights, and Int-8 or Int-16 for activations: lower precision for classification and detection tasks; high-precision Int-16 for audio and limited HDR image enhancements.

**Lossless Compression**

Advanced, lossless model compression reduces model size by up to 75%, increasing system inference performance and reducing power.
### Specifications

#### Key Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Performance (At 1GHz)</th>
<th>MACs (8x8)</th>
<th>Utilization on popular networks</th>
<th>Data types</th>
<th>Network support</th>
<th>Winograd support</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>256 GOPS/s to 4 TOP/s</td>
<td>128, 256, 512, 1024, 2048</td>
<td>Up to 85%</td>
<td>Int-8 weights and Int-16 activations</td>
<td>CNN, RNN, and transformer networks</td>
<td>No</td>
<td>Yes (2/4 sparsity supported with throughput doubled)</td>
</tr>
</tbody>
</table>

#### Memory System

<table>
<thead>
<tr>
<th>Memory System</th>
<th>29 to 267 KB</th>
<th>Up to six 128-bit AMBA 5 AXI</th>
<th>KB to multi-MB</th>
<th>Weights only; both Standard and Fast Weight Decoder</th>
<th>Extended compression, layer/operator fusion, striping capability</th>
</tr>
</thead>
</table>

#### Development Platforms

<table>
<thead>
<tr>
<th>Development Platforms</th>
<th>TensorFlow Lite Micro</th>
<th>Bare-metal, RTOS, Linux</th>
<th>TensorFlow Lite Micro Runtime, CMSIS-NN, optimizer, driver</th>
<th>Layer-by-layer visibility with PMUs, cross trigger interface</th>
<th>Performance model, FPGA evaluation platforms</th>
</tr>
</thead>
</table>

To learn more about the Ethos-U85 processor, visit [developer.arm.com/ethos-u85](http://developer.arm.com/ethos-u85)