A new class of machine learning (ML) processor, called a microNPU, specifically designed to accelerate ML inference in area-constrained embedded and IoT devices. The Ethos-U55 combined with the AI-capable Cortex-M55 processor provides a 480x uplift in ML performance over existing Cortex-M based systems.

**Key Features and Benefits**

- **Partner Configurable**
  
  Multiple configurations allow designers to rapidly target a wide variety of AI applications with up to 480x increase in performance.

- **Extremely Small Area**
  
  90% energy reduction in about 0.1mm² for AI applications in cost-sensitive and energy-constrained devices.

- **Single Toolchain**
  
  A unified toolchain for Ethos-U55 and Cortex-M helps ease developer use and creation of AI applications.
HIGHLIGHTS

+ **Energy Efficient**
  Provides up to 90 percent energy reduction for ML workloads, such as ASR, compared to previous Cortex-M generations.

+ **Network Support**
  Flexible design supports a variety of popular neural networks, including CNNs and RNNs, for audio processing, speech recognition, image classification, and object detection.

+ **Future-Proof Operator Coverage**
  Heavy compute operators run directly on the micro NPU, such as convolution, LSTM, RNN, pooling, activation functions, and primitive element wise functions. Other kernels run automatically on the tightly coupled Cortex-M using CMSIS-NN.

+ **Offline Optimization**
  Offline compilation and optimization of neural networks, performing operator, and layer fusion as well as layer reordering to increase performance and reduce system memory requirements by up to 90 percent. Delivers increased performance and lower power compared to non-optimized ordering.

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

+ **Element Wise Engine**
  Designed to optimize for commonly used elementwise operations such as add, mul and sub for commonly used scaling, LSTM, GRU operations. Enables future operators composed of these similar primitive operations.

---

Futureproof
Native support for the most common ML network operations, including CNN and RNN, with flexibility for future ML innovations.

Arm Ethos-U55 enables powerful embedded ML Inference, with partner configurable options that help accelerate time to market.
KEY USE CASES FOR ETHOS-U55
+ 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

+ Mixed Precision
  Supports Int-8 and Int-16: 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
  percent, increasing system inference performance and reducing power.

+ System Integration
  Works seamlessly with Cortex-M devices and Arm Corstone systems, and
  allows designers to configure and build high-performance, power-efficient
  SoCs while further differentiating with combinations of Arm processors
  and their own IP.
## Specifications

### Key Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (At 1 GHz)</td>
<td>64 to 512 GOP/s</td>
</tr>
<tr>
<td>MACs (8x8)</td>
<td>32, 64, 128, 256</td>
</tr>
<tr>
<td>Utilization On Popular Networks</td>
<td>Up to 85%</td>
</tr>
<tr>
<td>Data Types</td>
<td>Int-8 and Int-16</td>
</tr>
<tr>
<td>Network Support</td>
<td>CNN and RNN/LSTM</td>
</tr>
<tr>
<td>Winograd Support</td>
<td>No</td>
</tr>
<tr>
<td>Sparsity</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Memory System

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal SRAM</td>
<td>18 to 50 KB</td>
</tr>
<tr>
<td>External On Chip SRAM</td>
<td>KB to Multi-MB</td>
</tr>
<tr>
<td>Compression</td>
<td>Weights only</td>
</tr>
<tr>
<td>Memory Optimizations</td>
<td>Extended compression, layer/operator fusion</td>
</tr>
</tbody>
</table>

### Development Platforms

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Frameworks</td>
<td>TensorFlow Lite Micro</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>RTOS or bare-metal</td>
</tr>
<tr>
<td>Software Components</td>
<td>TensorFlow Lite Micro Runtime, CMSIS-NN, Optimizer, Driver</td>
</tr>
<tr>
<td>Debug and Profile</td>
<td>Layer-by-layer visibility with PMUs</td>
</tr>
<tr>
<td>Evaluation and Early Prototyping</td>
<td>Performance Model, Cycle Accurate Model, or FPGA Evaluations</td>
</tr>
</tbody>
</table>

To find out more about the Ethos-U55 processor, visit [developer.arm.com/ethos-u55](http://developer.arm.com/ethos-u55)