Next-generation endpoint AI for IoT with the new Arm Cortex-M55 and Ethos-U55 processors

Mark Quartermain, Cortex-M55 Product Manager
Arm Enables AI Everywhere, On Any Device

Arm’s AI platform delivers comprehensive hardware IP, software frameworks, and ecosystem

Cloud AI

Edge AI

Endpoint AI

AI-enabled IoT device shipments forecast to increase by almost 20% per year through 2024*

*Source: Arm forecast based on industry data
### Best-in-class Solution Optimized for Endpoint AI

<table>
<thead>
<tr>
<th><strong>Cortex-M55</strong></th>
<th><strong>Ethos-U55</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Most AI-capable Cortex-M processor</td>
<td>First microNPU for Cortex-M</td>
</tr>
</tbody>
</table>

#### Performance

- **Cortex-M55**
  - Versatile ML performance: Up to 15x ML uplift*

- **Ethos-U55**
  - Dedicated ML performance: Additional 32x ML uplift**

#### Optimization

- **Cortex-M55**
  - Arm Custom Instructions*** and configuration options

- **Ethos-U55**
  - Configurable 32-256 MACs

#### Accelerated Design and Development

- **Corstone-300 reference design** for faster and more secure system-on-chip development

---

*Compared to previous Cortex-M generations
**Compared to the Cortex-M55
***available in 2021
Unified Software Development: Fastest Path to Endpoint AI

Multiple software development flows
- Harder to program and debug
- More complex, longer time to market

Unified software development flow
- Works with common ML frameworks and existing tools
- More productivity, faster time to market
Arm Cortex-M55 Processor

Arm’s most AI-capable Cortex-M processor and the first to feature Arm Helium vector processing technology.
Cortex-M55: The Most AI-capable Cortex-M Processor
Cortex-M processor with enhanced DSP/ML compute capabilities

Helium vector processing
- > 150 new scalar and vector instructions
- Support for complex maths
- Low overhead loops

Vector processing support
- 2 x 32-bit MAC/cycle
- 4 x 16-bit MAC/cycle
- 8 x 8-bit MAC/cycle

Extended datatype support
- Half-precision float
- 8-bit integer
- Floating-point (half, full and double-precision)

Cortex-M ease of use
- Unified instruction set
- Single toolchain, simplified debug
- No need for separate DSP engine
- Cortex-M ecosystem

High performance system
- Memory system designed for DSP and ML applications
- Optional I/D caches
- Optional Tightly Coupled Memory

Security
- TrustZone for system wide security
- Protect software investments
Cortex-M55: Accelerating Embedded DSP/ML Performance

**Signal processing**

Average performance per datatype for selected CMSIS-DSP kernels vs Cortex-M4

- **Q7**
- **Q15**
- **Q31**
- **float16**
- **float32**

- **Cortex-M4**
- **Cortex-M33**
- **Cortex-M55**
- **Cortex-M7**

**Machine learning**

- **CIFAR10**
  - ≈7x higher perf. vs Cortex-M4

- **Keyword spotting**
  - >8x higher perf. vs Cortex-M4

Highest performance and efficiency for ML and signal processing across Cortex-M portfolio

Cortex-M55 performance results are based on selected CMSIS-DSP kernels such as CFFT, FIR, RFFT, matrix mul, vector dot product. Cortex-M55 data based on LAC RTL and ACC v6.14 compiler in development. Data subject to change.
Cortex-M55 Energy Efficiency by Datatype

Compared to the Cortex-M4 processor

- Measured average energy consumption based on selected DSP kernels from CMSIS-DSP
- Energy efficiency measured as: \( \frac{\text{Cortex-M4 cycles to complete kernel}}{\text{Cortex-M55 cycles to complete kernel}} \times \frac{\text{Cortex-M4 Power}}{\text{Cortex-M55 Power}} \)
- > 1 indicates Cortex-M55 energy efficiency is greater than Cortex-M4

Cortex-M55 data based on LAC RTL. Data subject to change.
Arm Ethos-U55 microNPU

The first Arm microNPU for Cortex-M based systems
Ethos-U55: The First microNPU for Cortex-M

- Configurations 32/64/128/256 MACs
- High compute operators accelerated in hardware. Other operators run on the microcontroller
- Works alongside Cortex-M55, Cortex-M7, Cortex-M33 and Cortex-M4 processors
- Connected to the memory bus
- Uses existing system SRAM and flash storage
- Comes with a high-performance, high-efficiency Mac engine
- Weight decoder and DMA for on-the-fly weight decompression
- 8-bit input x 8-bit weights
- 16-bit input x 8-bit weights
Ethos-U55 Performance Results

Using **128 MACs/Cycle** configuration of Ethos-U55

**Wav2letter**

- **230x**

  - Cortex-M4
  - Ethos-U55 and Cortex-M55

**MobileNet V2**

- **150x**

  - Cortex-M4
  - Ethos-U55 and Cortex-M55

*Based on early estimates*
CMSIS-NN and TensorFlow Lite Micro for Cortex-M

Optimized low-level kernels for the embedded market

- The optimized kernel library for Cortex-M
  - Called from TF Lite Micro or bare metal implementations
  - Offline flow creates a binary for Cortex-M based platforms
- Targets all Cortex-M architectures
  - Armv6-M/Armv7-M/Armv8-M/Armv8.1-M with Helium support
  - Runs on earlier versions of the architecture
- Key operators accelerated by CMSIS-NN
  - Fallback to TF Lite Micro reference kernels
- Open-source, via Apache 2.0 license
  [https://github.com/ARM-software/CMSIS_5](https://github.com/ARM-software/CMSIS_5)
Accelerating Embedded Machine Learning

Add Ethos-U55 under the same stack

- Boosts ML performance beyond Cortex-M alone
  - Small memory footprint
  - Low-power for 'always-on' applications
- Operators are accelerated by the microNPU
  - Fallback to CMSIS-NN, then reference kernels
- Tooling for offline optimization
  - Quantize and tune data structures for the microNPU
Summary: Bringing the Benefits of AI to Billions More - Devices

- Significant uplift in DSP/ML performance
- Meets efficiency needs of IoT endpoint
- Built-in system-wide security
- Unified toolchain enabling ease-of-use
- Simplified SoC and software development
- Industry-leading ecosystem

Arm Cortex-M55

Arm Ethos-U55

Arm Corstone-300
Industry-wide Effort: The Most Extensive AI Ecosystem

Significant silicon partner collaboration

Algorithm, software, tools and RTOS partners

...and others
Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
תודה

Find Out More:
Cortex-M55: developer.arm.com/cortex-m55
Ethos-U55: developer.arm.com/ethos-U55
Corstone-300: developer.arm.com/corstone-300

Alternatively, get in touch with an expert to learn more:
pages.arm.com/cortex-M55-consultation