What is tinyML and what is it used for?

On-device machine learning applications in the single mW and below

Vibration and motion

Any ‘signal’

Predictive maintenance, sensor fusion, accelerometer, pressure, lidar/radar, speed, shock, vibration, pollution, density, viscosity, etc.

Voice and sound

Recognition and creation

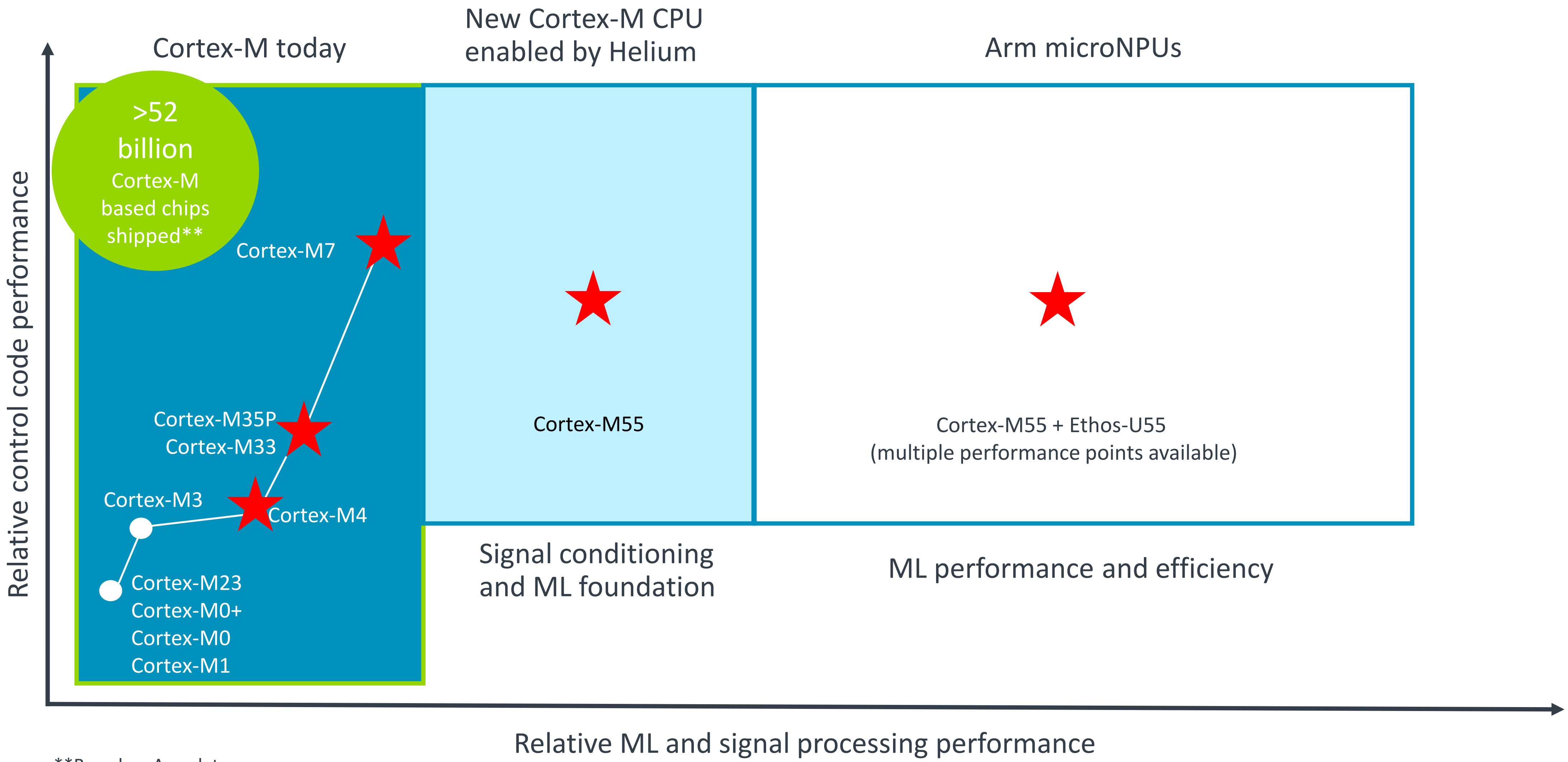
Keyword spotting, speech recognition, natural language processing, speech synthesis, sound recognition, etc.

Vision

Images and video

Object detection, face unlock, object classification etc.

Pushing the Boundaries for Real-time On-device Processing



**Based on Arm data

★ Well suited for ML & DSP applications

Graph not to scale

Cortex-M55: The Most AI-capable Cortex-M Processor

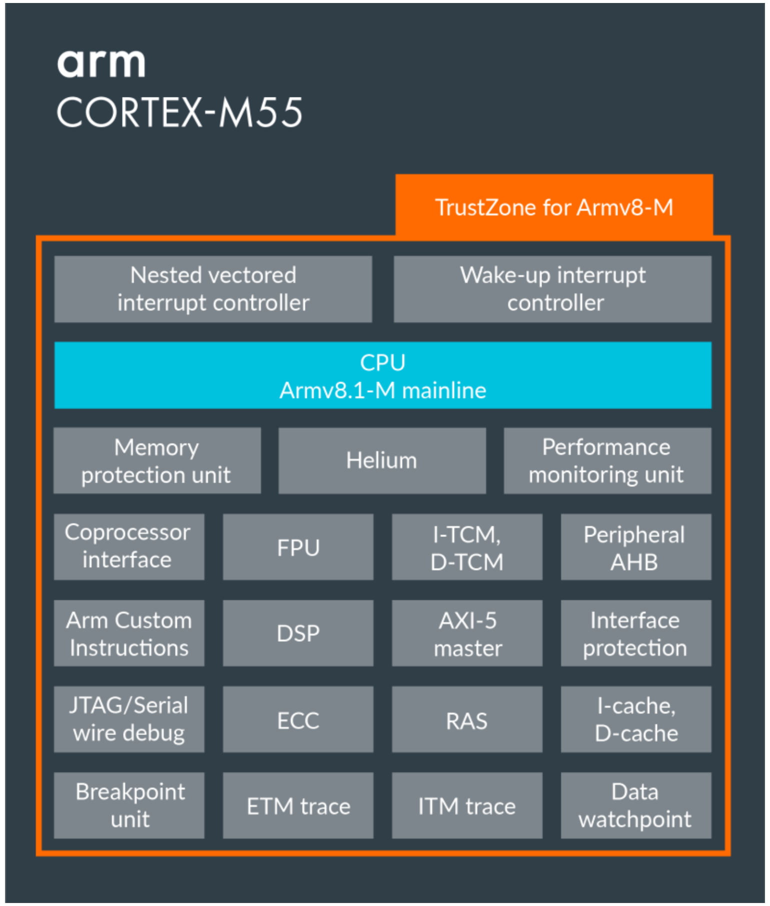
✓ First CPU based on Arm Helium technology

- Energy-efficient and configurable with vector processing capabilities
- Delivers up to 5x DSP performance and up to 15x ML performance*
- Versatile capability for both classical ML and NN inference

✓ Advanced memory interfaces for fast access to ML data and weights

✓ TrustZone support

✓ Extensive configurability



arm
CORTEX-M55

TrustZone for Armv8-M

Nested vectored interrupt controller
Wake-up interrupt controller

CPU
Armv8.1-M mainline

Memory protection unit
Helium
Performance monitoring unit

Coprocessor interface
FPU
I-TCM, D-TCM
Peripheral AHB

Arm Custom Instructions
DSP
AXI-5 master
Interface protection

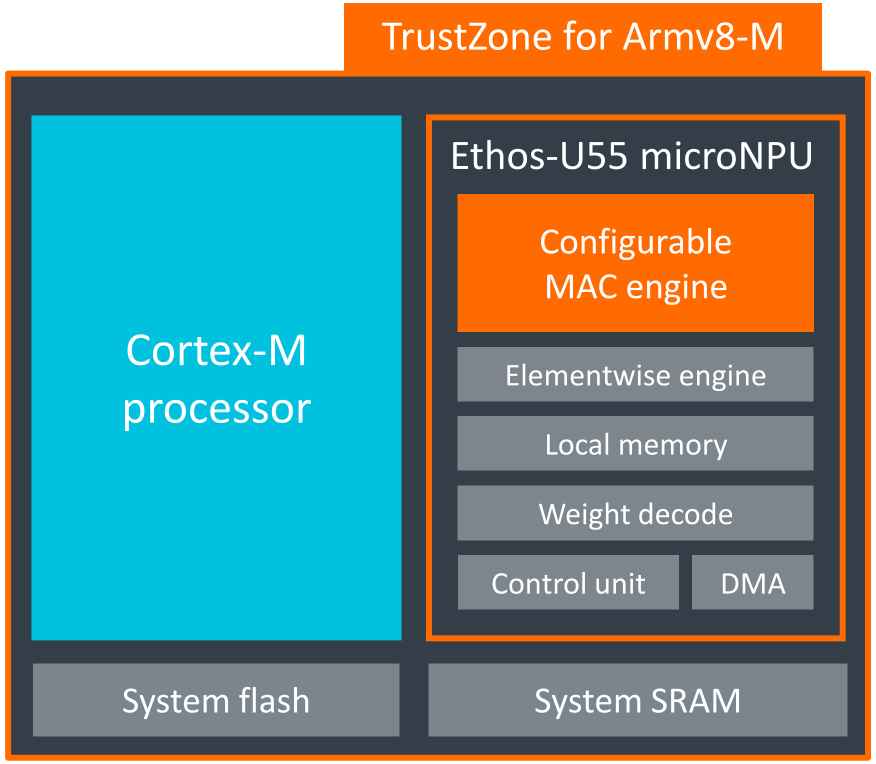
JTAG/Serial wire debug
ECC
RAS
I-cache, D-cache

Breakpoint unit
ETM trace
ITM trace
Data watchpoint

*Compared to previous Cortex-M generations

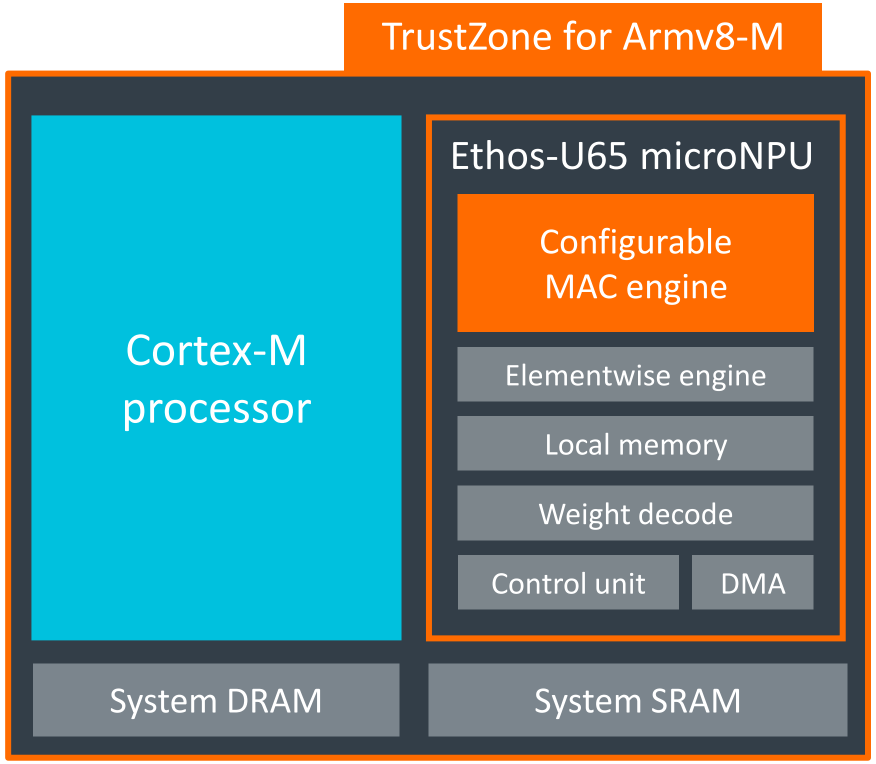
Ethos-U55: The first microNPU for Cortex-M

- ✓ Highest efficiency and small memory footprint
- ✓ 32, 64, 128, or 256 unit multiply-accumulate (MAC) engine
- ✓ Weight decoder and DMA for on-the-fly weight decompression
- ✓ Tooling available for offline optimization
- ✓ Works with a range of Cortex-M processors:
  - Cortex-M55
  - Cortex-M33
  - Cortex-M7
  - Cortex-M4

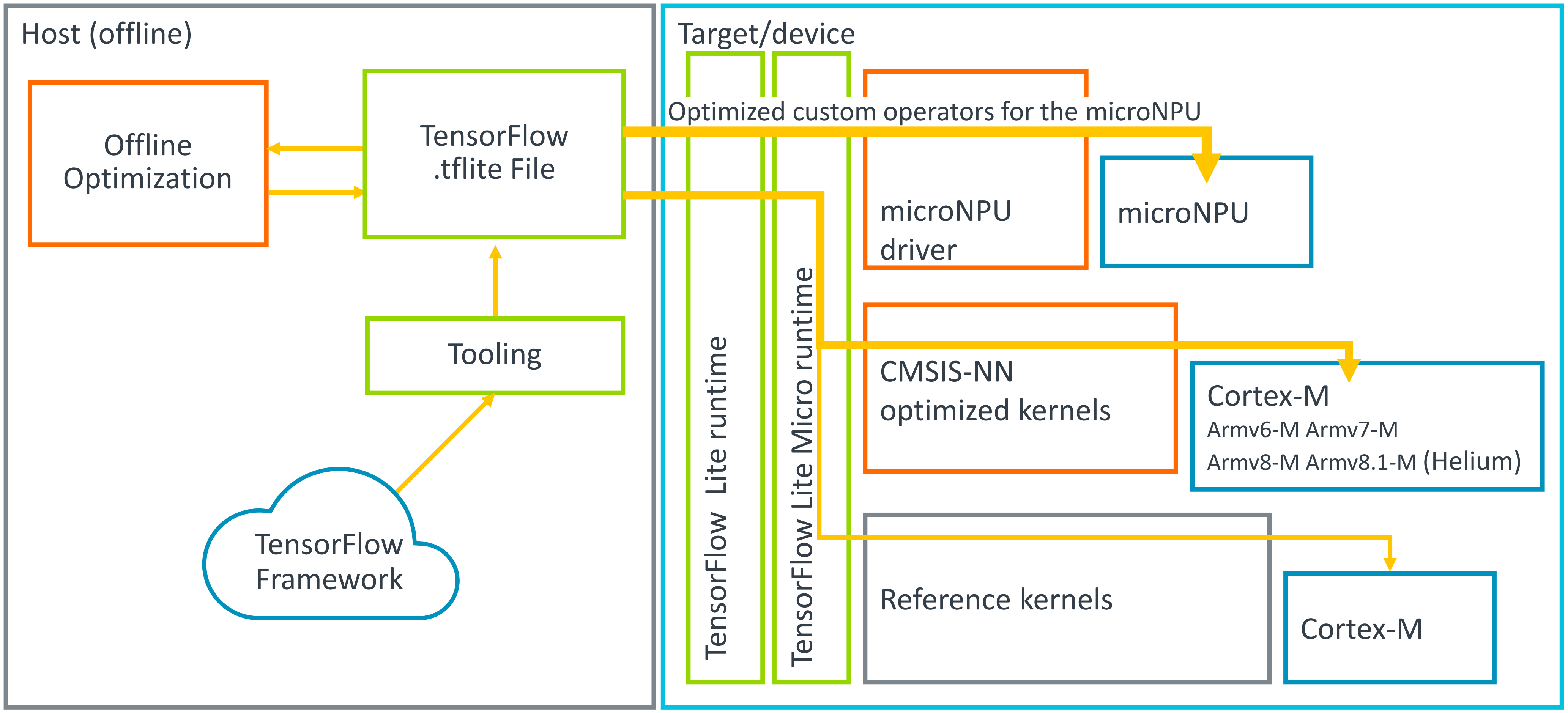


Ethos-U65: The second generation microNPU

- ✓ 256 or 512 unit multiply-accumulate (MAC) engine
- ✓ DMA update for DRAM as well as flash support
- ✓ Can be an M-class subsystem inside an A-class system

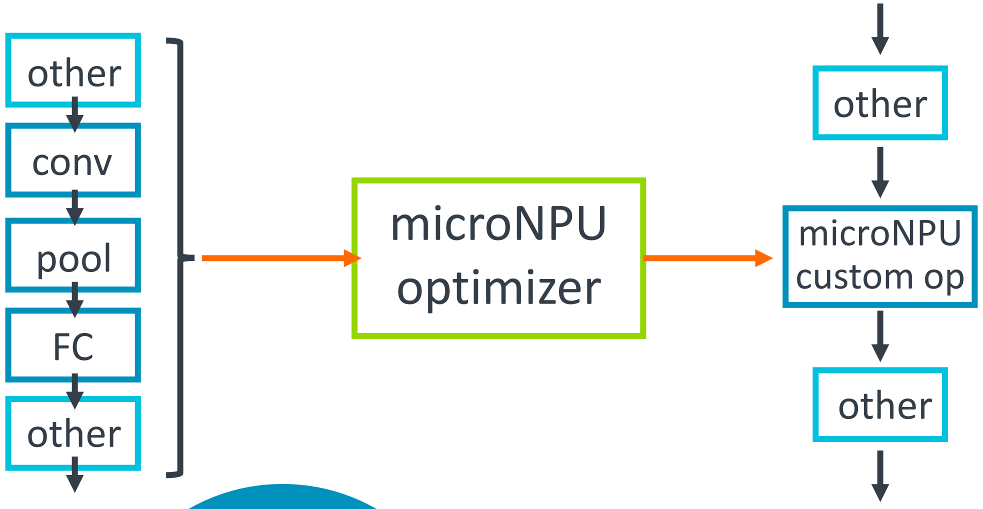


Mapping of NNs to Ethos-U using TensorFlow Lite



The Vela Optimizer

- Open source compiler
- Reads a tflite file and identifies subgraphs
- Optimizes scheduling of subgraphs
- Loss-less compression of weights
- Generates commands for microNPU
- Writes out a modified tflite file



Up to 90% SRAM size reduction

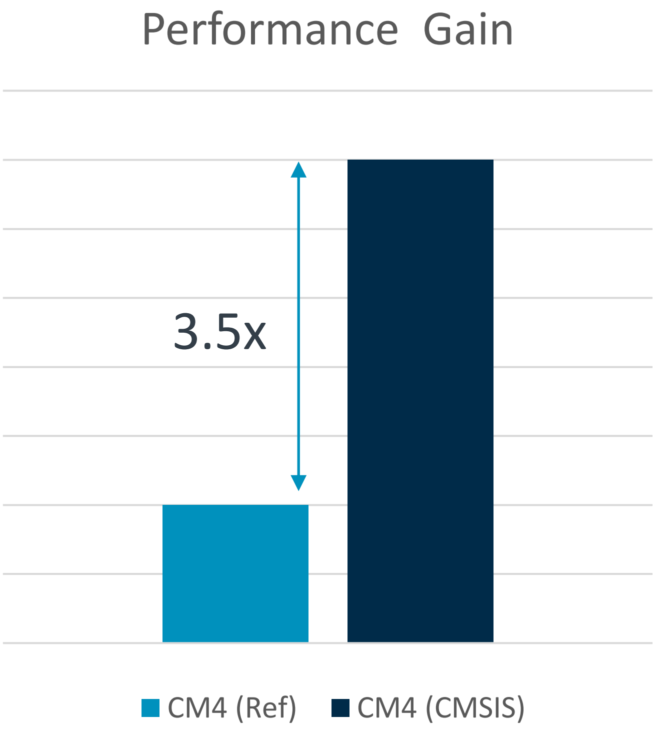
Up to 70% model size reduction

Enabling networks not before feasible in embedded systems

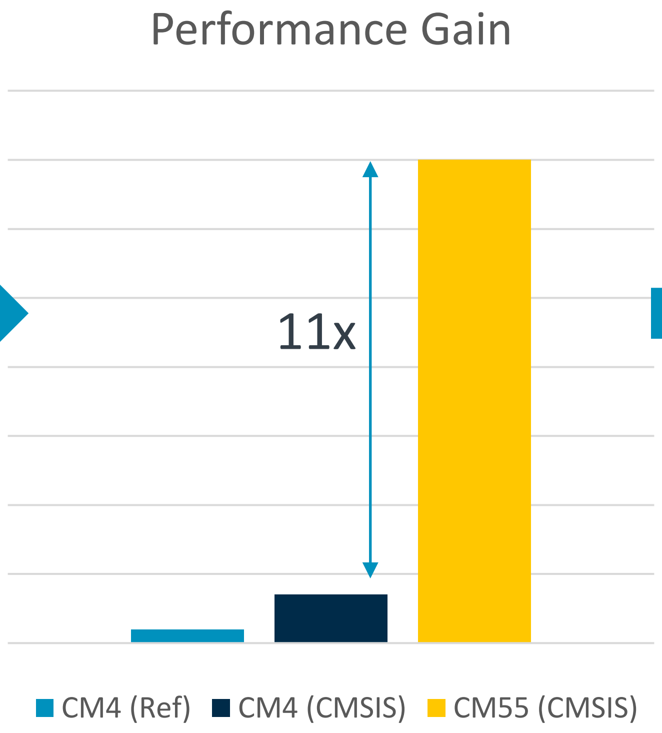
Neural Network performance Across ARM IPs

Wav2Letter

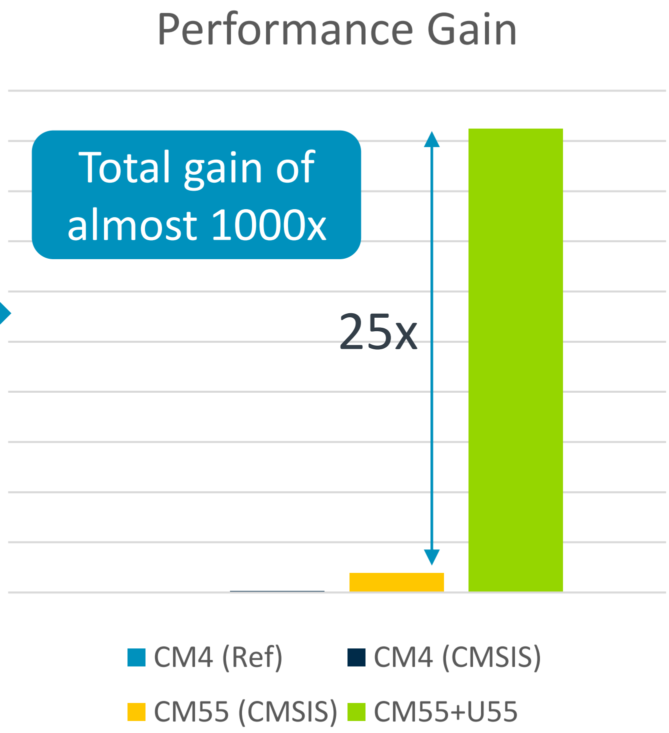
Efficient Software (CMSIS-NN)



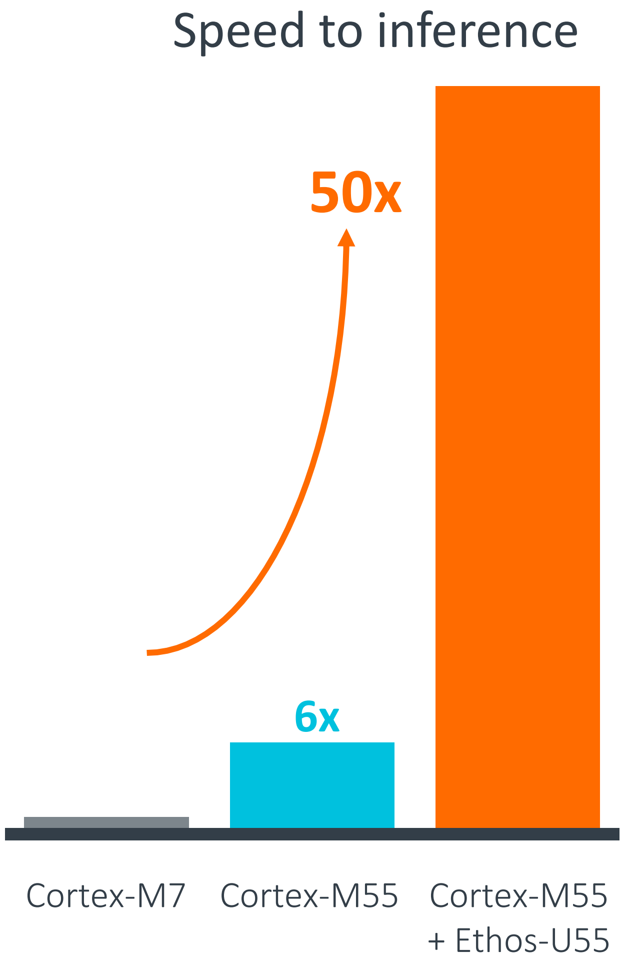
AI Capable Cortex-M55

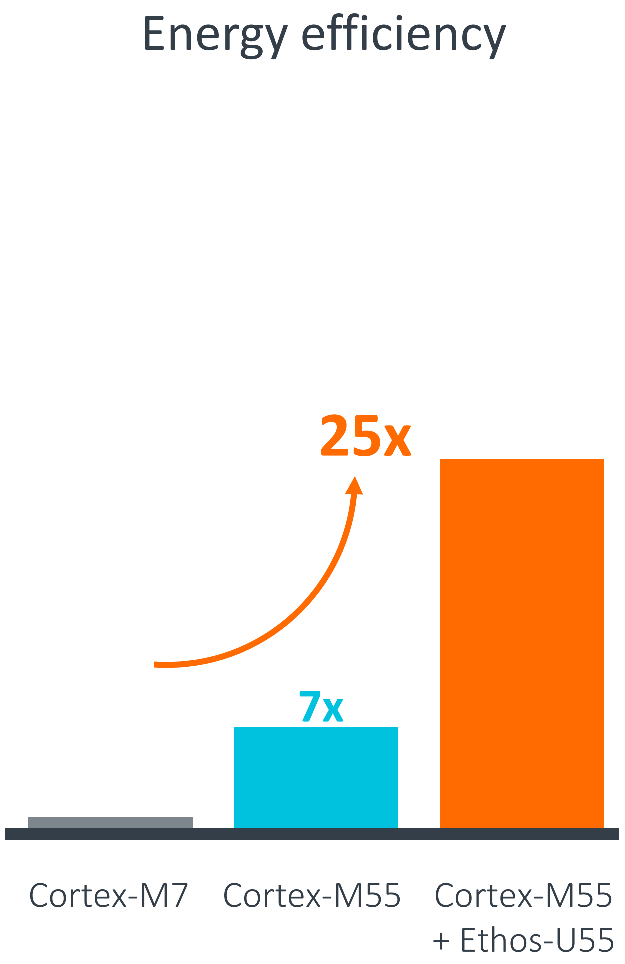


AI Dedicated U55 256 MAC/cycle



Full example: Typical ML Workload for a Voice Assistant





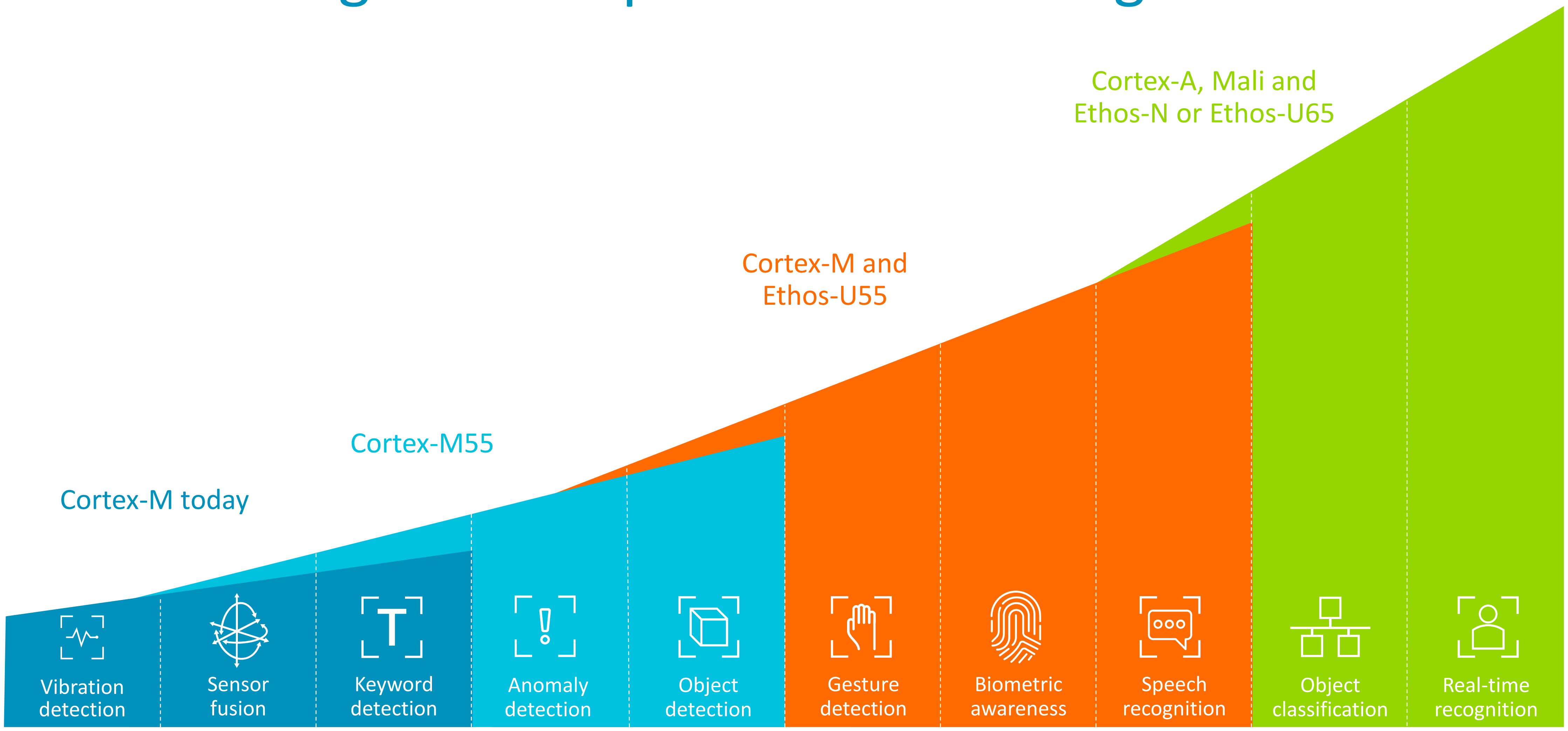
✓ Faster responses

✓ Smaller form-factors

✓ Improved accuracy

Latency and energy spent for all tasks listed combined: voice activity detection, noise cancellation, two-mic beamforming, echo cancellation, equalizing, mixing, keyword spotting, OPUS decode, and automatic speech recognition.

Broadest Range of ML-optimized Processing Solutions


Cortex-A, Mali and Ethos-N or Ethos-U65
Cortex-M and Ethos-U55
Cortex-M55
Cortex-M today
Vibration detection
Sensor fusion
Keyword detection
Anomaly detection
Object detection
Gesture detection
Biometric awareness
Speech recognition
Object classification
Real-time recognition