Endpoint AI and the Advent of the microNPU

MLSys Conference

April 5th 2021

Tomas Edsö
Distinguished Engineer
Arm Machine Learning Group

Endpoint AI and tinyML

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

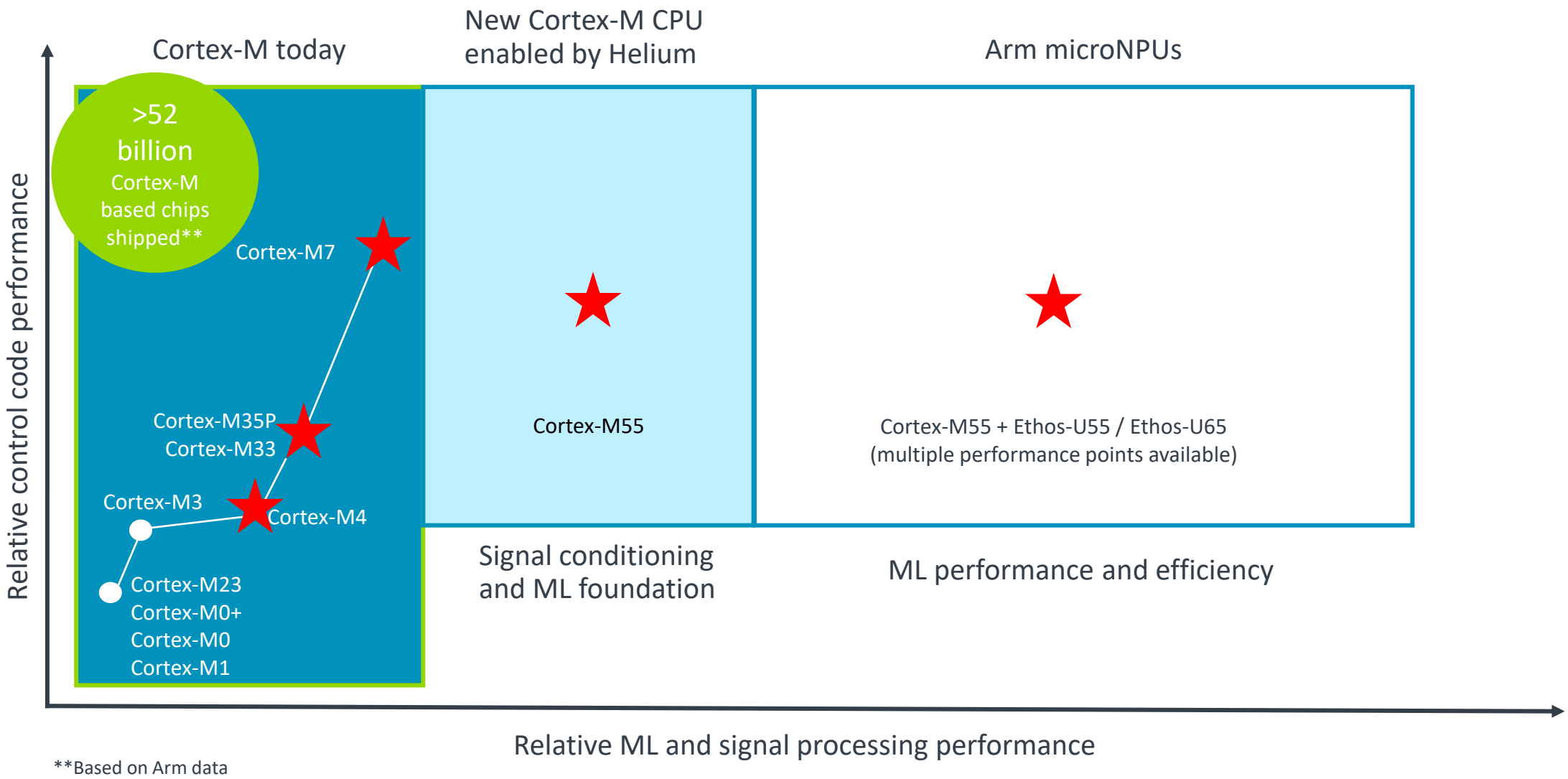
Keyword spotting, speech recognition, natural language processing, speech synthesis, sound recognition, etc.

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

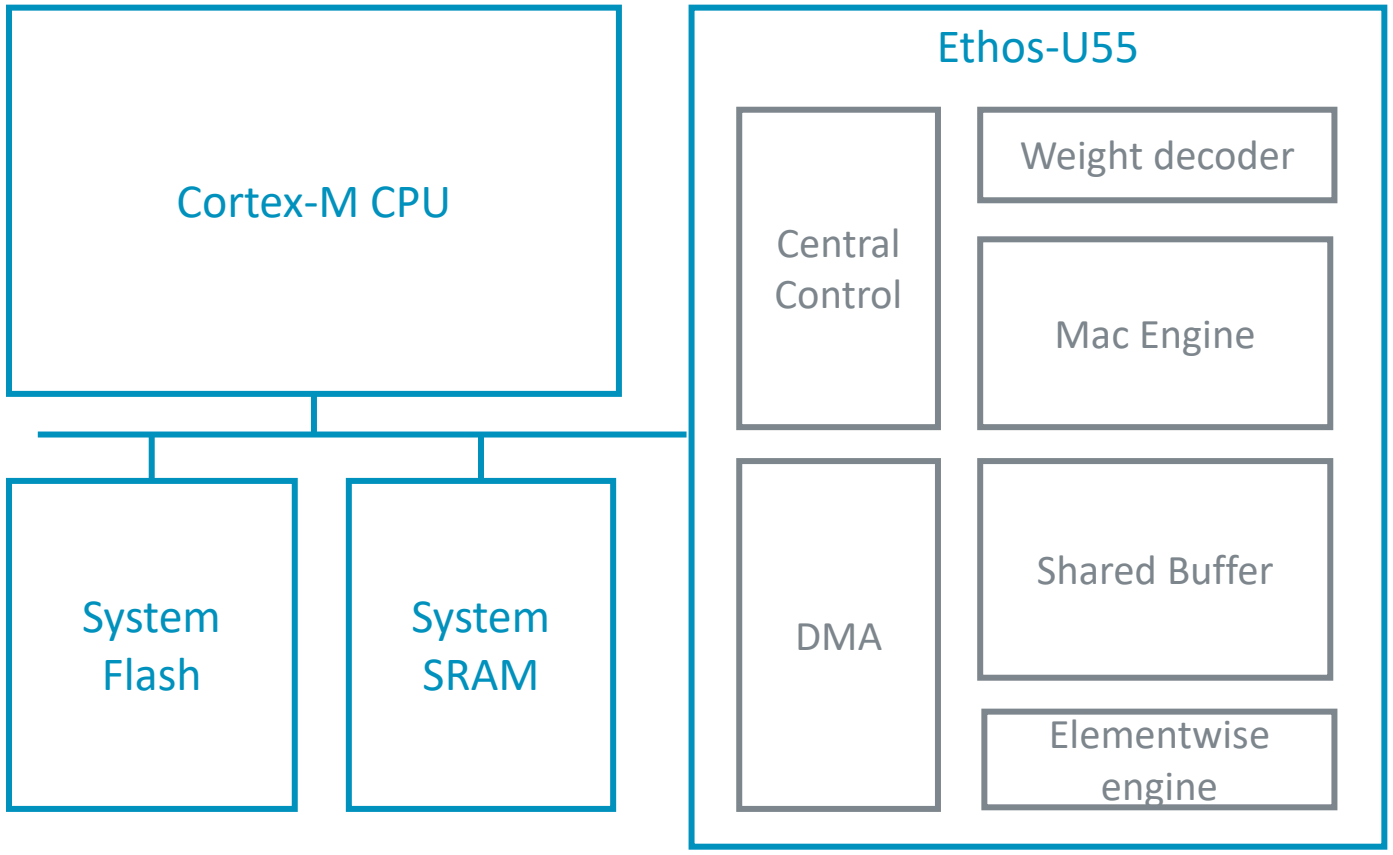
arm

Ethos-U55

The first microNPU for M-class CPUs

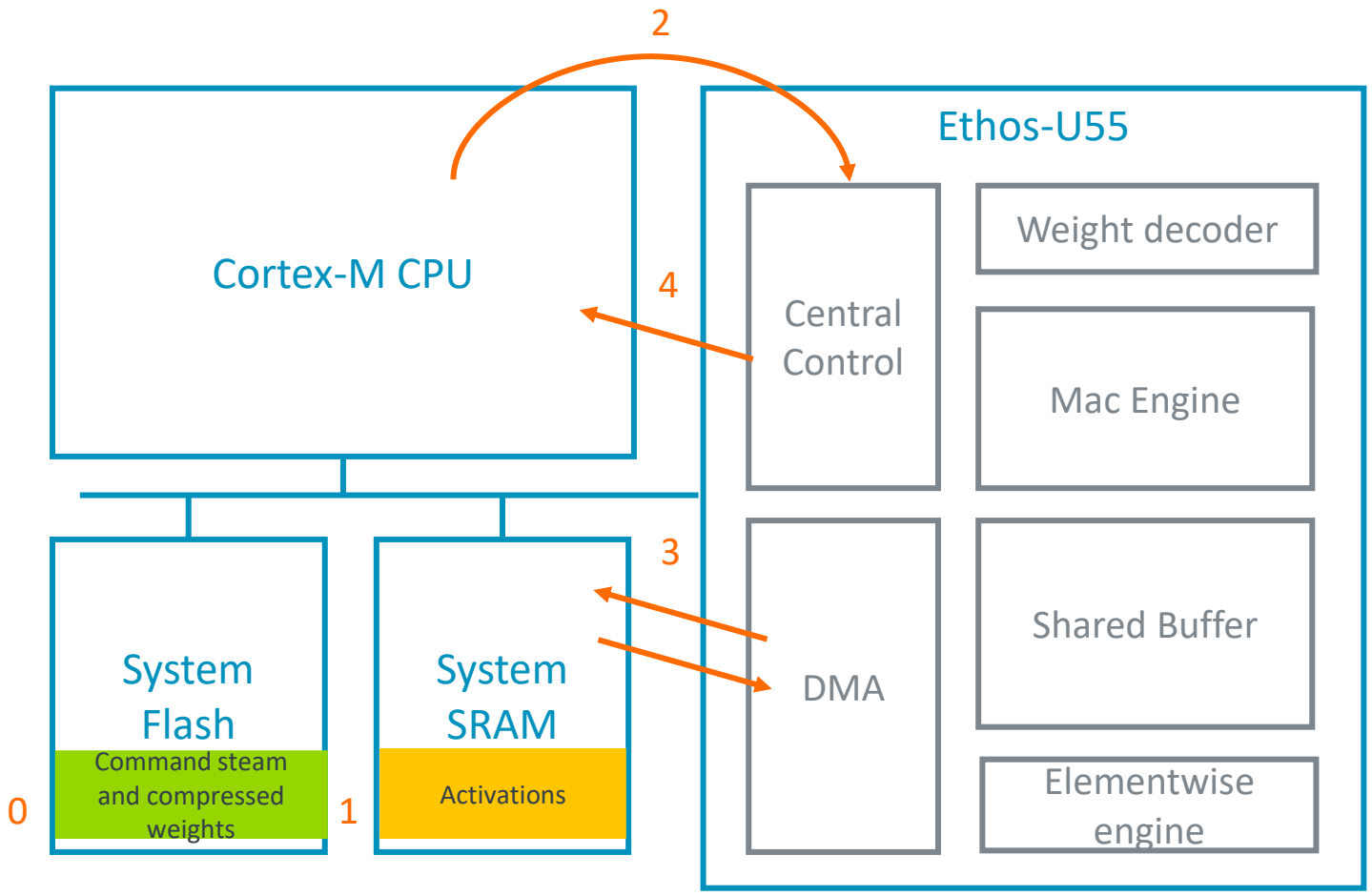
Ethos-U55 overview

- Works alongside
  Cortex-M55, Cortex-M7, Cortex-M33 and Cortex-M4 processors
- Works alongside on-chip SRAM and system flash
- Accelerates CNN and RNN operators.
- Efficient weight compression
- 8- or 16-bit activations
  Weights are always 8-bit
- 32, 64, 128 or 256 MAC/cc configurations



Typical Ethos-U55 data flow

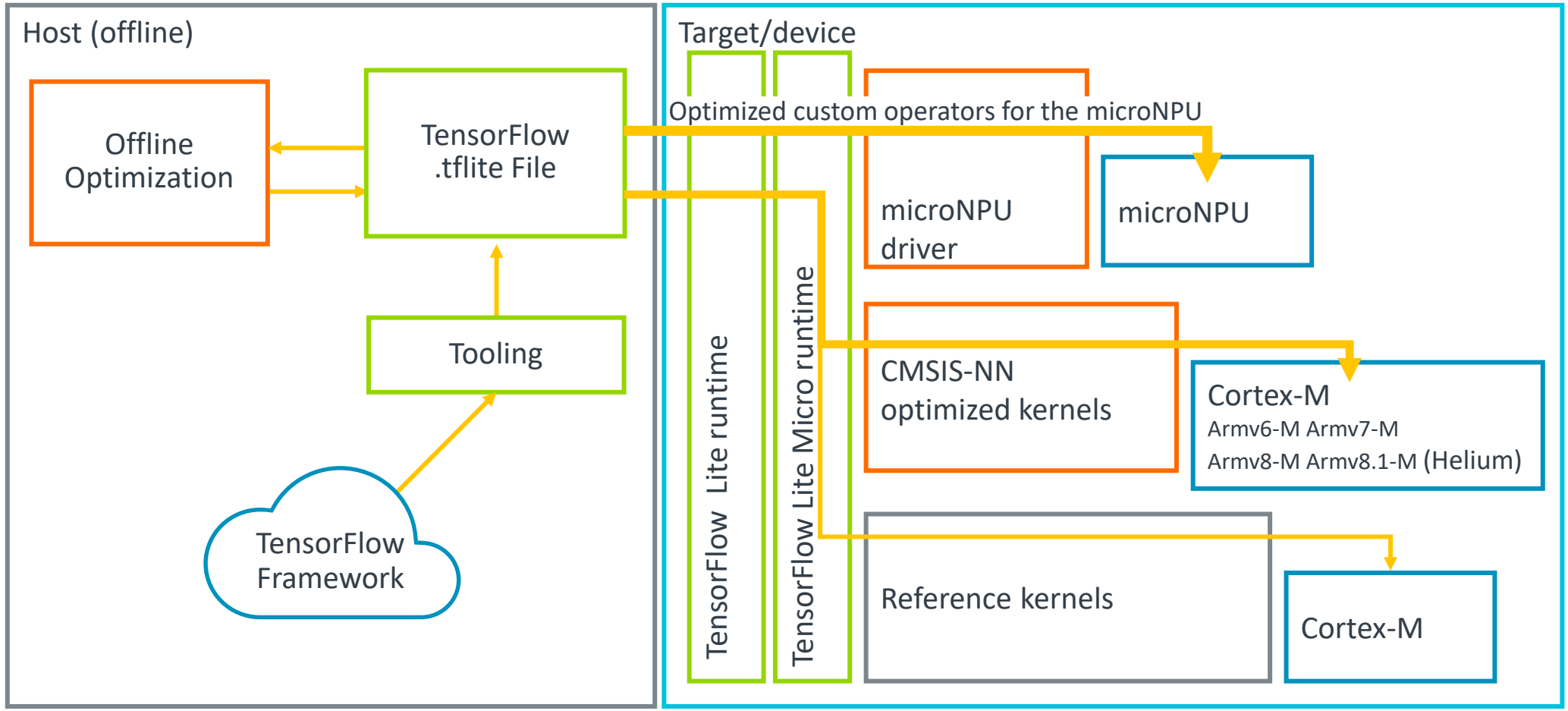
0. An offline compiled command stream with corresponding compressed weights are put into system Flash.
1. Input activations are stored in system SRAM.
2. The host starts Ethos-U55 by defining all memory regions to be used, such as the location of the command stream and input activations.
3. Ethos-U55 autonomously runs all commands, using SRAM as a scratch buffer. The final outputs are written to a defined SRAM buffer.
4. Interrupt on completion of writing the final outputs.



Operations supported by Ethos-U

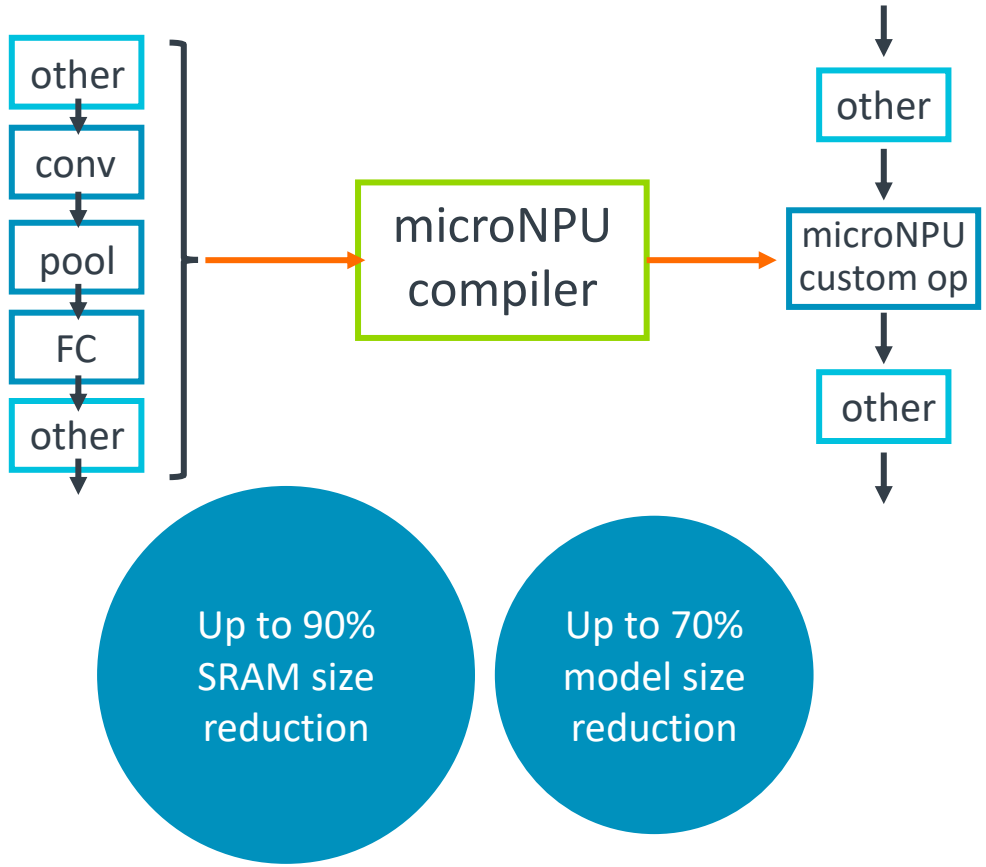
- Conv2D and DepthwiseConv2D
  - Up to 64x64 kernel
  - Up to 3x3 stride
  - Optimized kernel dilation, up to 2x2
- DeConv
  - Zero, nearest neighbor and bilinear, up to 2x2
- MaxPool and AvgPool
  - Up to 64x64 kernel with SAME padding
  - Up to 256x256 kernel with VALID padding
  - Up to 3x3 stride
- FullyConnected
  - Support for batching
- LSTM, GRU
  - Broken down into supported operations
- ADD, SUB, MUL, MIN, MAX
  - All types of broadcast allowed
- ReLU, ReLU1, ReLU6, tanh, sigmoid
  - All can be fused with other operations
- Configurable LUT
  - Can map any unary function
- PRELU, ABS
- ConCat
- Reshape
- ExpandDims
- Squeeze
- SoftMax

Mapping of NNs to Hardware using TensorFlow Lite



The Vela Offline Compiler

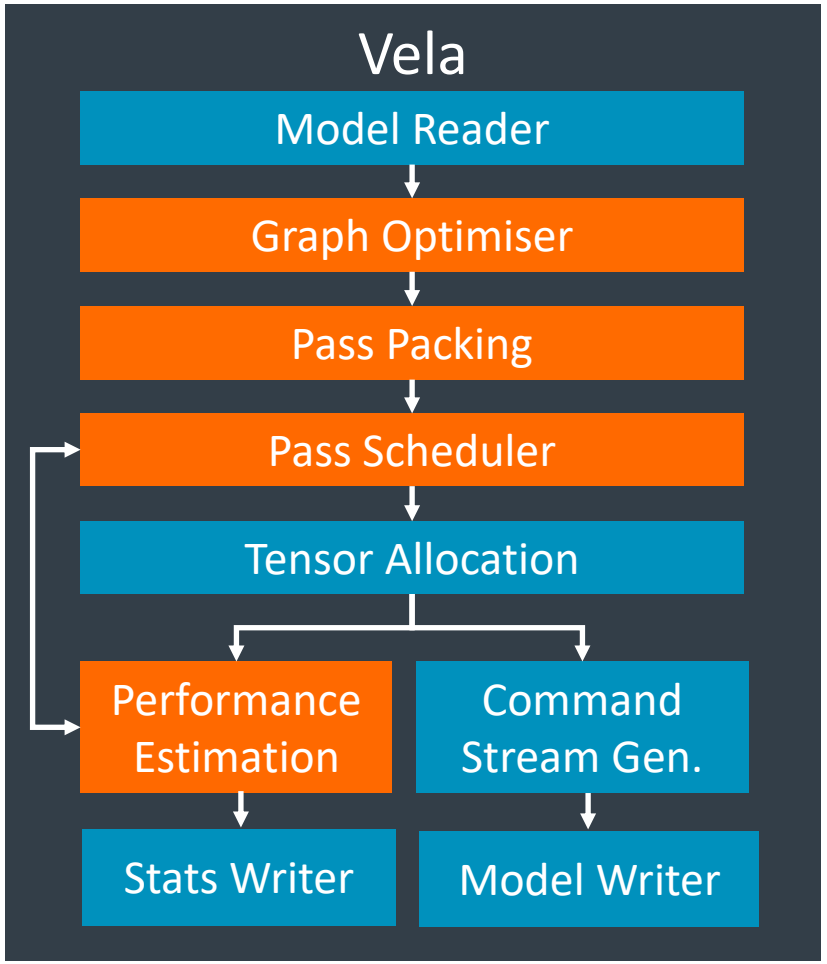
- Open source
- Reads a tflite file and identifies subgraphs
- Optimizes scheduling of subgraphs
- Loss-less compression of weights
- Execution and memory planning
- Generates commands for microNPU
- Writes out a modified tflite file



Enabling networks not before feasible in embedded systems

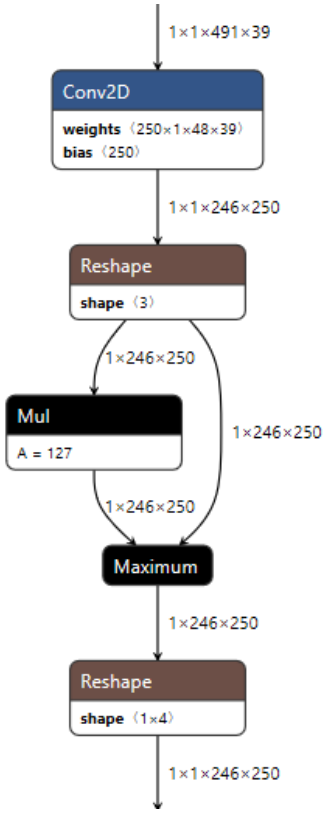
Vela Internal Components

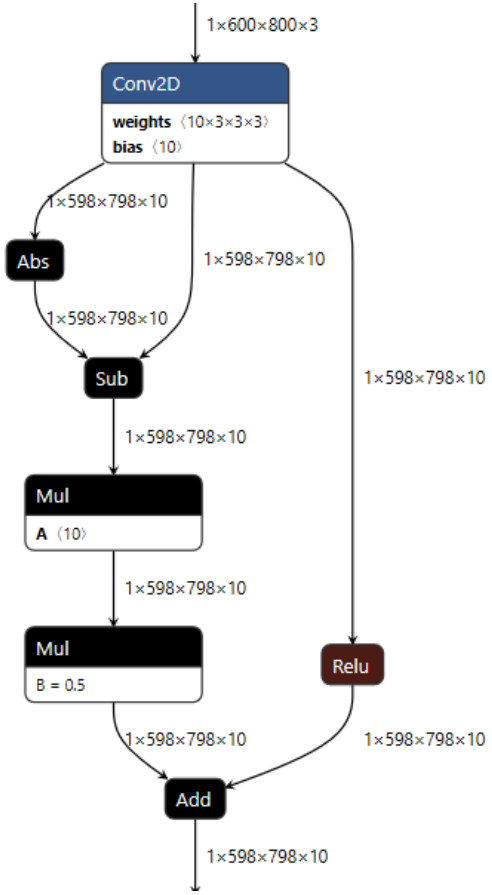
- Graph Optimiser
  - Traverses the graph performing high level checks and optimisations (e.g. LeakyReLU to Mul/Max to LUT)
- Pass Packing
  - Groups network operators that can be executed as a single hardware operation
- Pass Scheduler
  - Finds the optimal way for the hardware to process the Pass
  - Cost model scores strategies using performance estimates
  - Weight Encoding
- Performance Estimation
  - Generates Memory usage, Bandwidth, and Processing Cycles

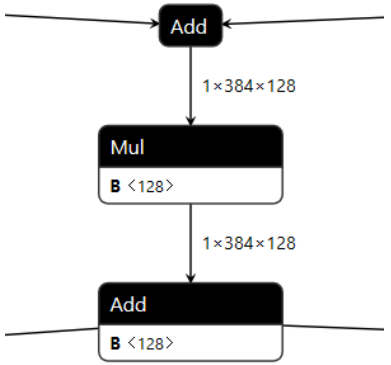


Configurable LUT

- Fusing of sequences of unary functions
  - Single LUT fused with the layer producing it

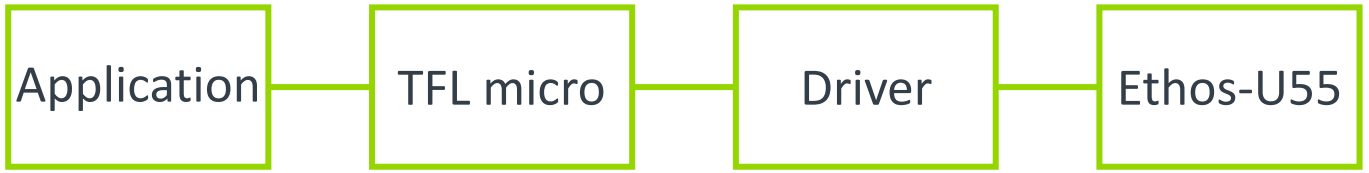




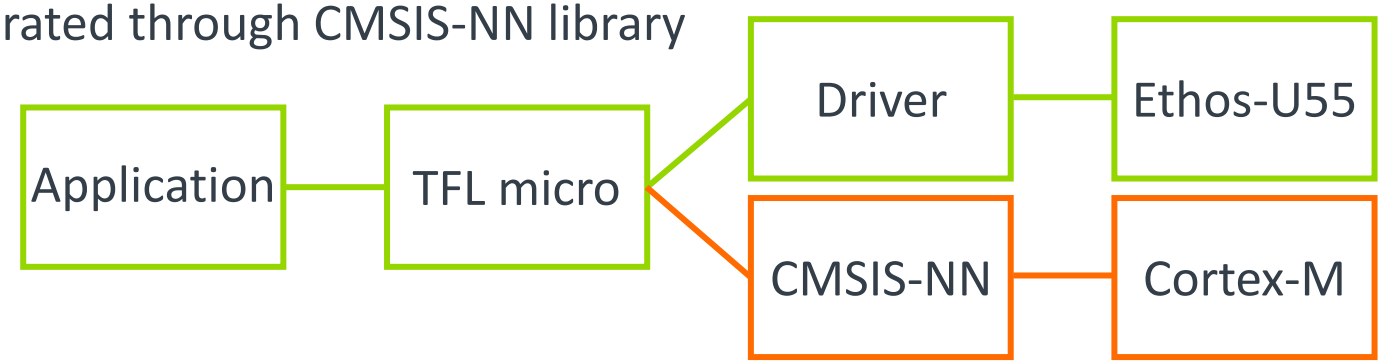


Network support in Ethos-U55

- Ethos-U55 can completely execute networks that map to the supported operator set
  - For example:
    - Deepspeech
    - RNNoise
    - Wav2letter
    - DSCNN
    - MobileNet_v1
    - MobileNet_v2
    - MobileNet_v3
    - ResNet
    - Inception_v3



- Any unsupported operation fallback to the Cortex-M processor
  - These can be accelerated through CMSIS-NN library



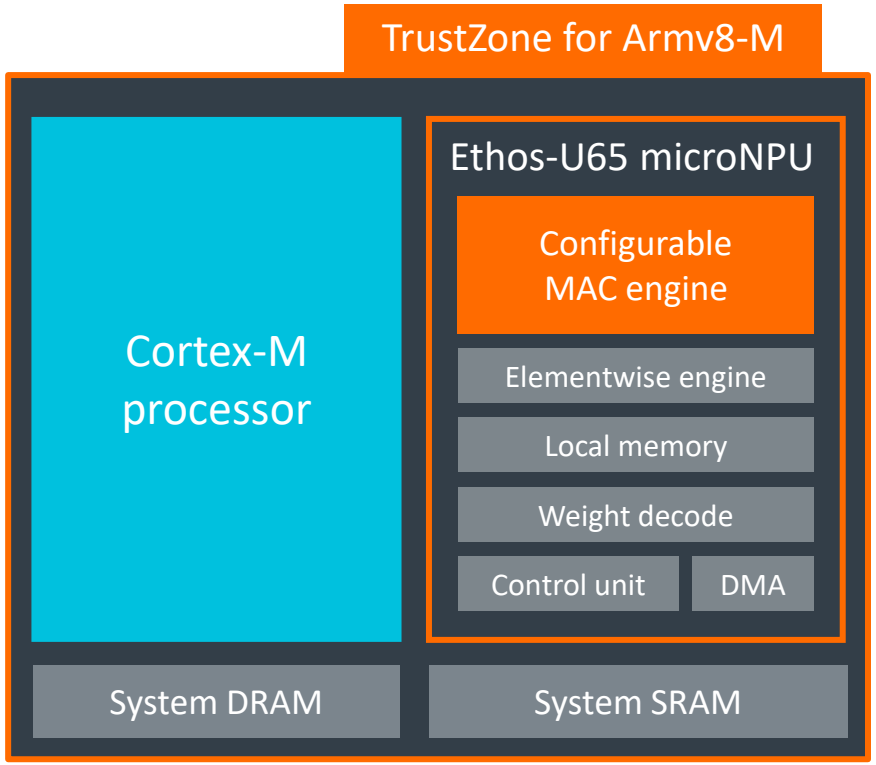
arm

Ethos-U65

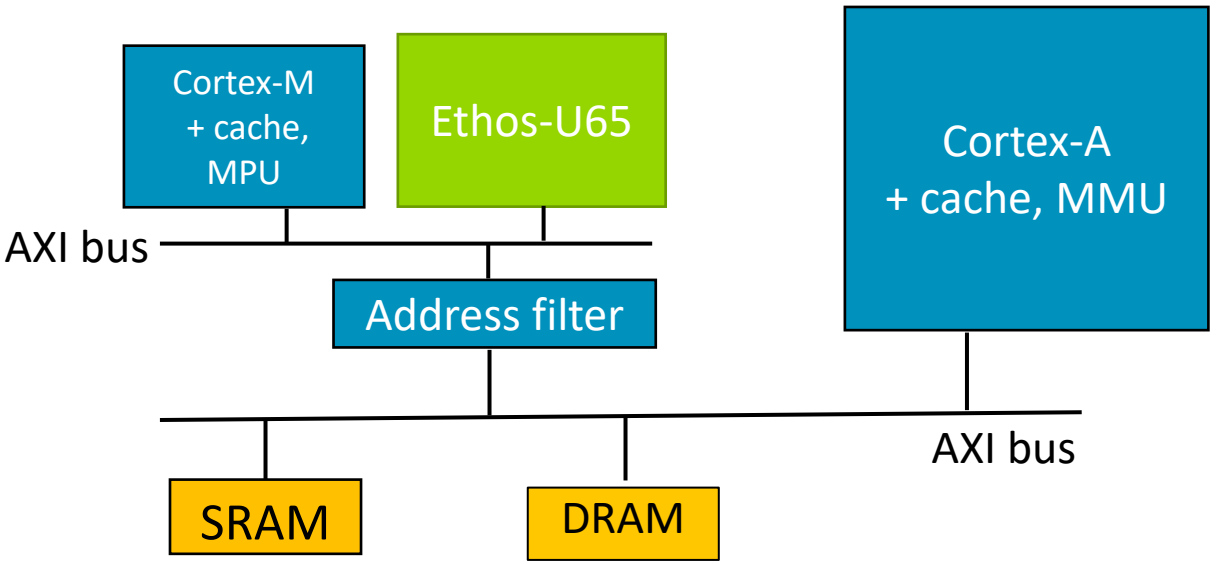
A microNPU for a new class of AI devices

Ethos-U65 overview

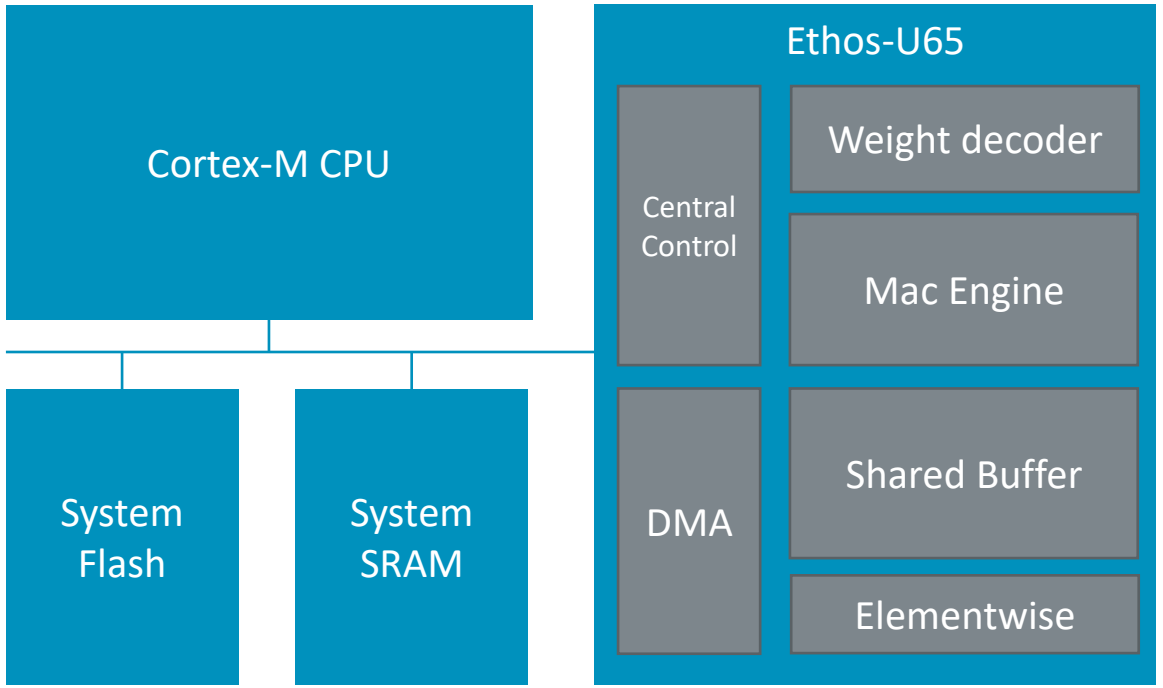
- 256 or 512 unit multiply-accumulate (MAC) engine
- DMA update for DRAM as well as flash support
- Can be an M-class subsystem inside an A-class system



Different systems applicable to Ethos-U65



Cortex A + Cortex M system



Cortex M only system

arm

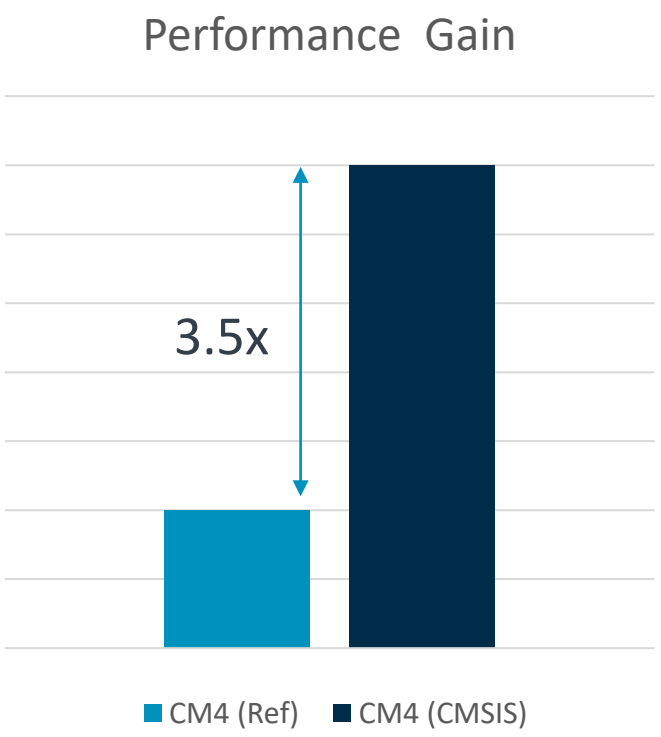
Ethos-U performance

Enabling AI at the endpoint

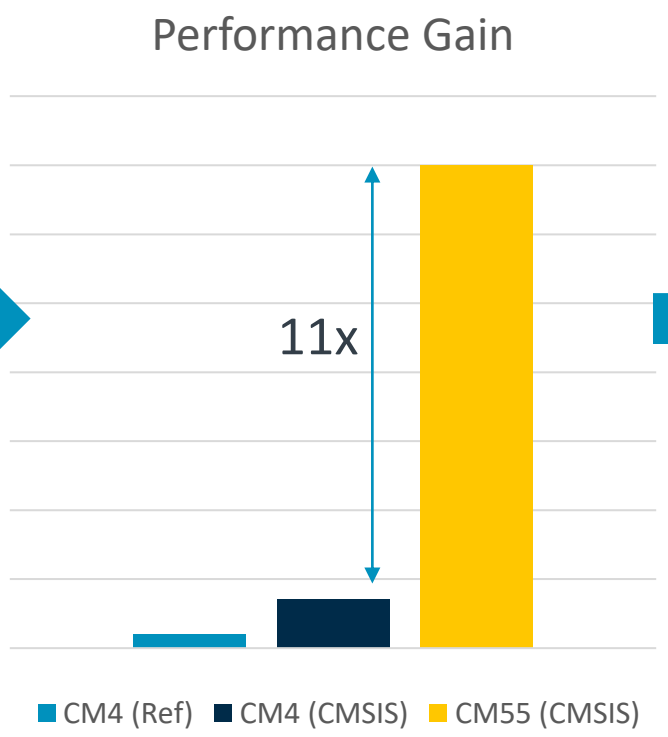
Neural Network performance Across ARM IPs

Wav2Letter

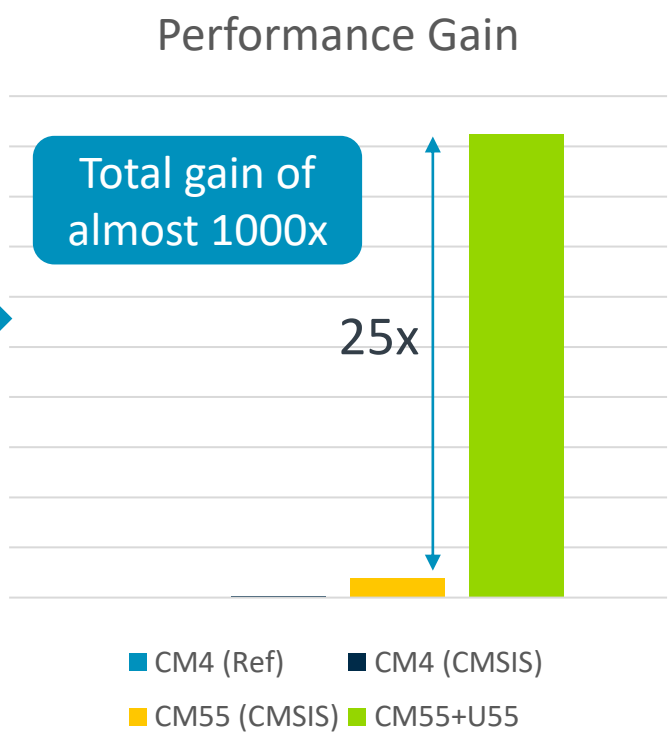
Efficient Software (CMSIS-NN)



AI Capable Cortex-M55

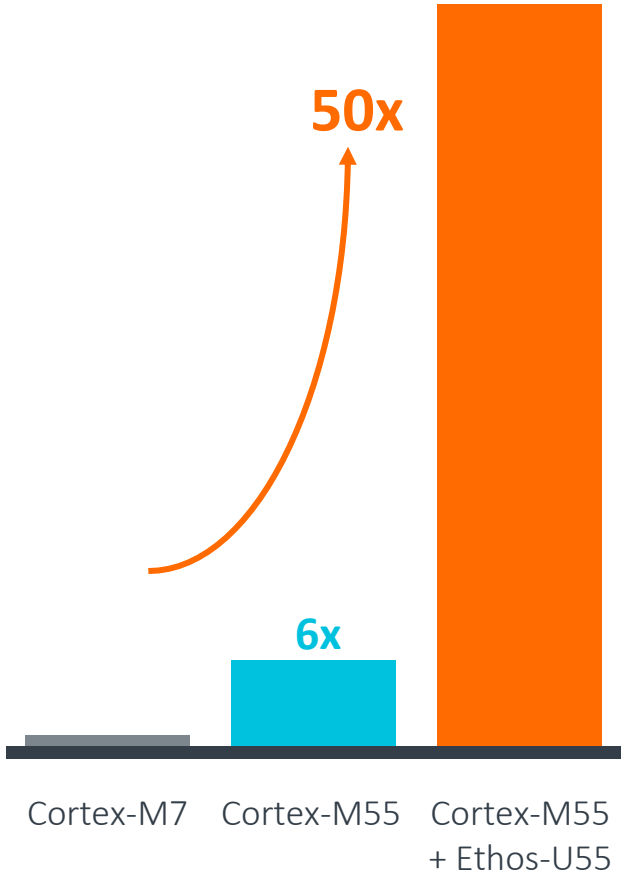


AI Dedicated U55 256 MAC/cycle

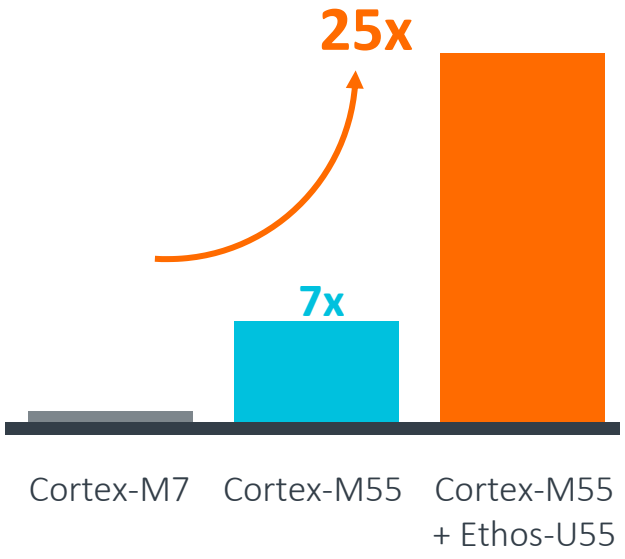


Full example: Typical ML Workload for a Voice Assistant

Speed to inference



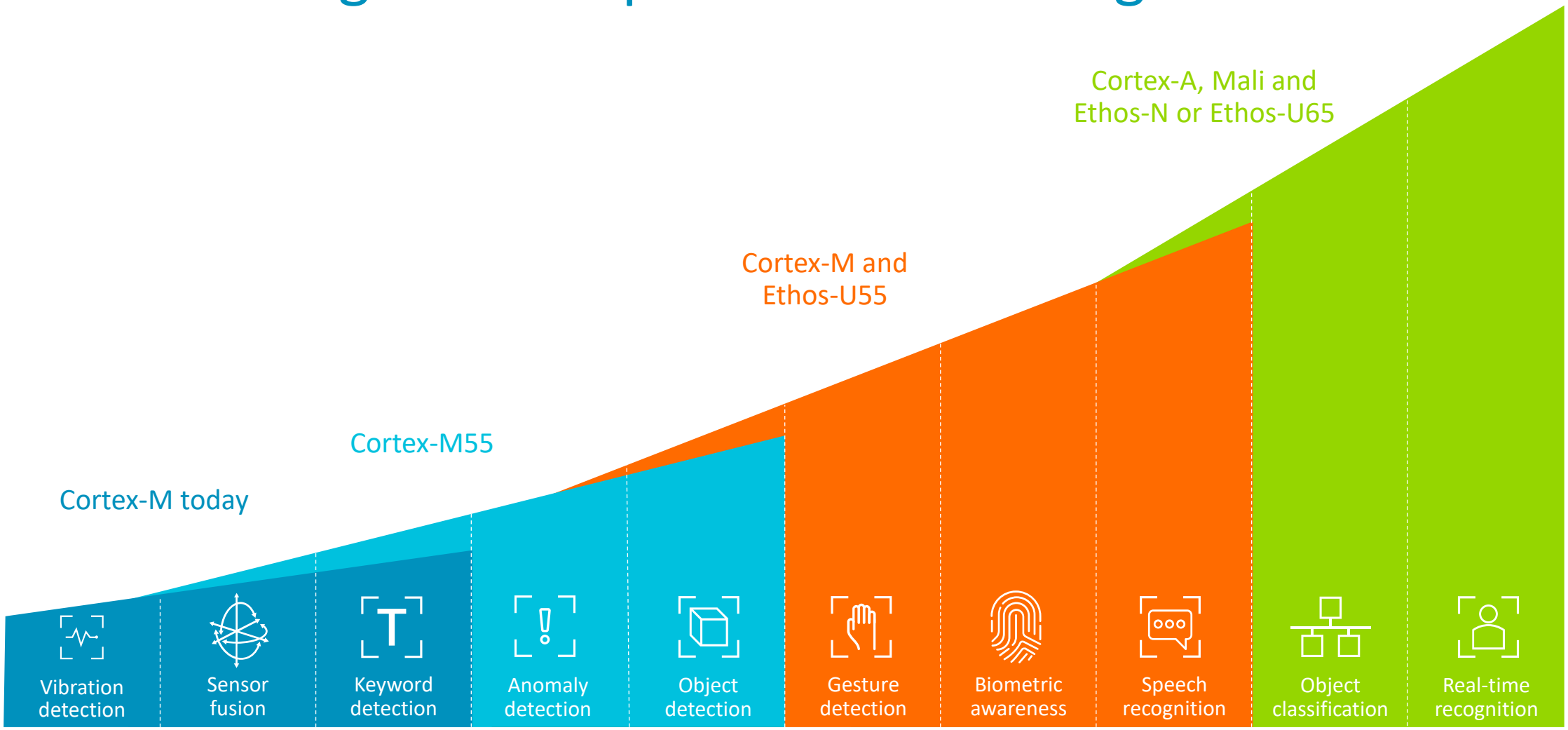
Energy efficiency



- ✓ 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



arm

Thank You
Danke
Gracias
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
Kiitos
شكرًا
ধন্যবাদ
תודה