arm AI

AI Virtual Tech Talks Series

Getting started with Cortex-M software development and Arm Development Studio

Pareena Verma & Ronan Synnott, Arm

August 11, 2020

AI Virtual Tech Talks Series

Date	Title	Host
Today	Getting started with Arm Cortex-M software development and Arm Development Studio	Arm
August, 25	Efficient ML across Arm from Cortex-M to Web Assembly	Edge Impulse
Sept, 8	Running accelerated ML applications on mobile and embedded devices using Arm NN	Arm
Sept, 22	How to reduce AI bias with synthetic data for edge applications	Dori.AI

Visit: developer.arm.com/solutions/machine-learning-on-arm/ai-virtual-tech-talks

arm

Agenda

➢ **Introduction to Arm Cortex-M processor family**
- ➢ Highest performance Cortex-M7 processor
- ➢ First Helium capable Cortex-M55 processor

➢ **CMSIS-NN and TensorFlow Lite Micro**
- ➢ Build and deploy a ML application with TensorFlow Lite and CMSIS-NN kernels

➢ **Q & A 1**

➢ **Arm Development Studio and demonstration**
- ➢ Develop and debug your ML application on Arm Cortex-M7 and Cortex-M55 FVPs
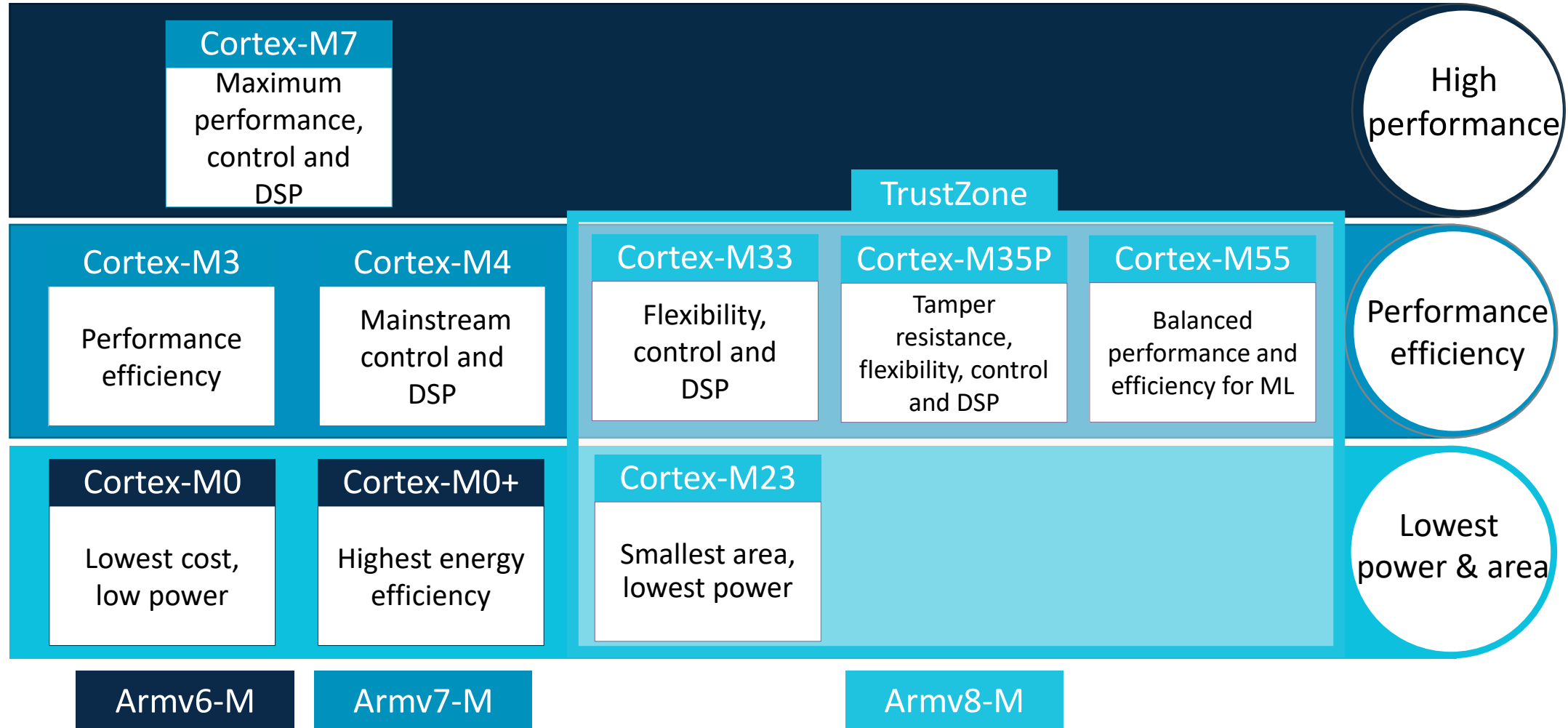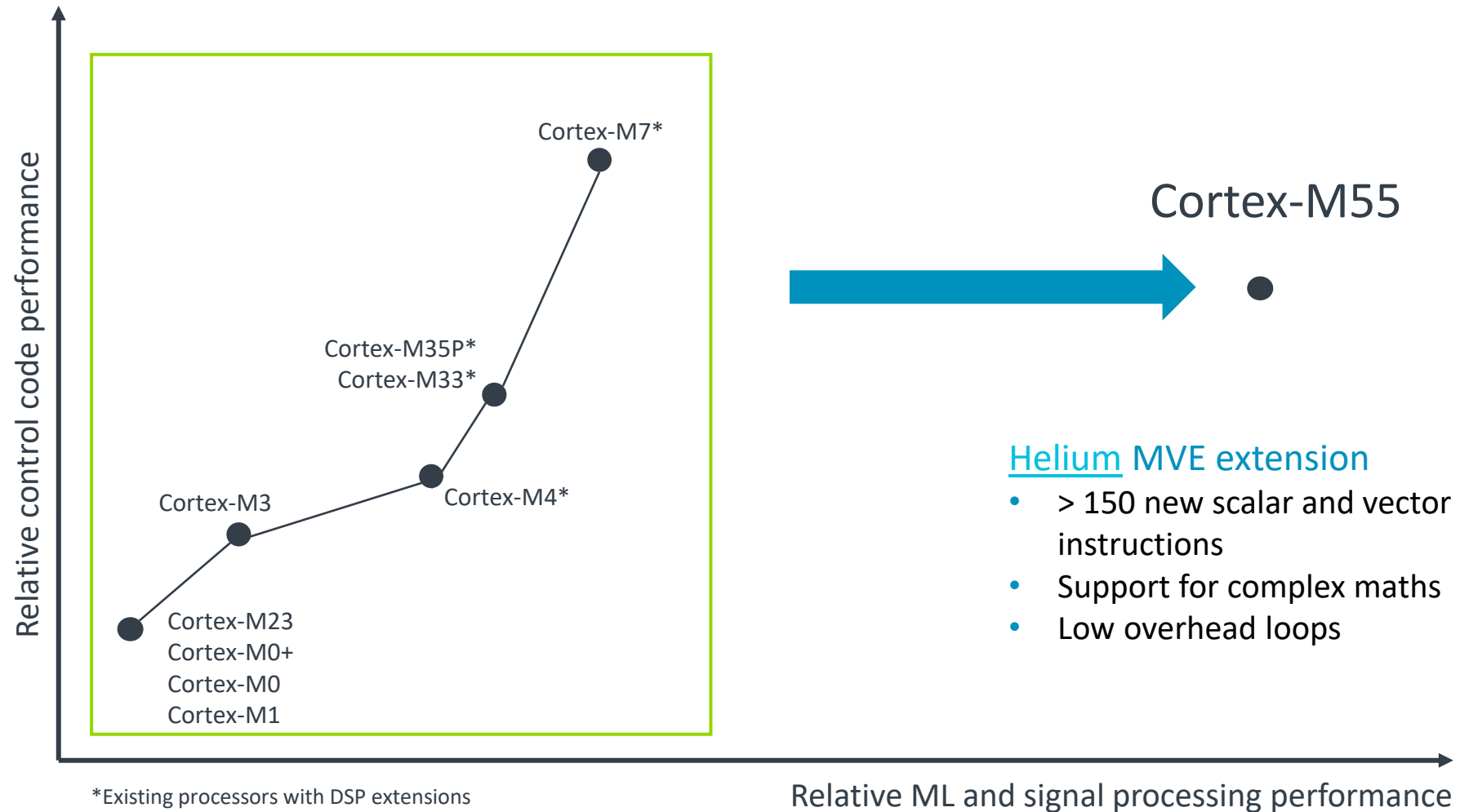
➢ **Q & A 2**

arm

arm AI

AI Virtual Tech Talks Series

Arm Cortex-M processors

Cortex-M processor portfolio

Cortex-M7			High performance
Maximum performance, control and DSP			

TrustZone

Cortex-M3	Cortex-M4	Cortex-M33	Cortex-M35P	Cortex-M55	Performance efficiency
Performance efficiency	Mainstream control and DSP	Flexibility, control and DSP	Tamper resistance, flexibility, control and DSP	Balanced performance and efficiency for ML	

Cortex-M0	Cortex-M0+	Cortex-M23	Lowest power & area
Lowest cost, low power	Highest energy efficiency	Smallest area, lowest power	

Armv6-M	Armv7-M	Armv8-M

arm

Uplift in DSP and ML performance for Cortex-M

Relative control code performance

Cortex-M7*

Cortex-M55

Cortex-M35P*
Cortex-M33*

Cortex-M3

Cortex-M4*

Cortex-M23
Cortex-M0+
Cortex-M0
Cortex-M1

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

*Existing processors with DSP extensions

Relative ML and signal processing performance

Graph not to scale

arm

Cortex-M7 key features
https://developer.arm.com/ip-products/processors/cortex-m/cortex-m7

- High performance core with DSP capabilities
  - Powerful DSP instructions
  - SP/DP Floating Point Unit
  - Six stage dual-issue pipeline
  - 5.01 CoreMark/MHz

- Flexible memory systems
  - Up to 16MB tightly-coupled memories for real-time determinism
  - Memory Protection Unit (MPU) and up to 64kB caches
  - 64-bit AXI4 memory interface

arm

Cortex-M55 key features

https://developer.arm.com/ip-products/processors/cortex-m/cortex-m55

- High performance core with DSP and vector processing capabilities
  - Helium vector processing technology
    - Configurable as integer-only or integer + floating-point
  - Powerful DSP instructions
  - HP/SP/DP Floating Point Unit
    - Vector HP/SP
  - Four stage pipeline
  - 4.2 CoreMark/MHz

- Flexible memory systems
  - Up to 16MB tightly-coupled memories for real-time determinism
  - Memory Protection Unit (MPU) and up to 64kB caches
  - 64-bit AXI5 memory interface

arm

Cortex-M55 arithmetic configuration options

Helium Configuration

FPU Configuration	MVE=0 (No MVE)	MVE=1 (Integer MVE)	MVE=2 (Integer and Floating-point MVE)
FPU=0 (No FPU)	• Scalar Integer	• Vector Integer	• N/A
FPU=1 (With FPU)	• Scalar Integer • Scalar Float	• Vector Integer • Scalar Float	• Vector Integer • Vector Float

arm

Meet the TensorFlow Family

TensorFlow

Model Creation
Model Training

Inference in cloud

Computationally Large

Google TPU

TensorFlow Lite

TensorFlow models made suitable for Edge devices

Inference on device only

No reliance on network connectivity

E.g. Speech recognition and NLP in 80Mb

Smartphones

Cortex-A Class IoT

TensorFlow Lite Micro

Bare metal TensorFlow Lite runtime for Arm MCU

Inference on device only

No reliance on network connectivity

Runs on MCU in 10'sK

Ultra Low Power- always on*

Cortex-M

*May be woken often by sensors

arm
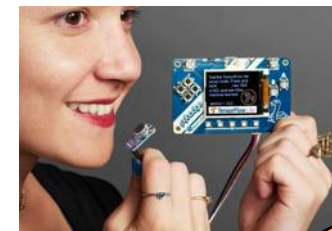
TensorFlow Lite Micro

- A version of TensorFlow Lite designed to run on Microcontrollers:
  - Less than 16KB binary footprint + Inference Model (typically < 100K)
  - No memory allocation
  - Maintains the TensorFlow Lite API's
  - Examples to get started on Arm Cortex-M
    - https://github.com/tensorflow/tensorflow/tree/master/tensorflow/lite/micro/examples/

- Several Arm Fixed Virtual Platforms (FVPs) available today for software development
  - Start porting your ML workloads to Arm Cortex-M systems without the need for a physical target
  - Programmer's view , which gives you a comprehensive model on which to build and test your software
 - https://developer.arm.com/tools-and-software/simulation-models/fixed-virtual-platforms

arm

Software Architecture - CMSIS-NN

Application

TensorFlow Lite Micro

CMSIS-NN
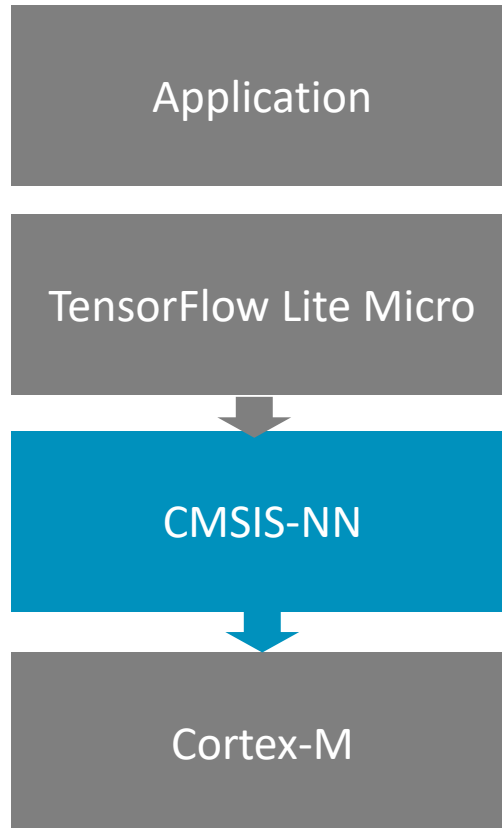
Cortex-M

- Open Source : https://github.com/ARM-software/CMSIS_5

- An optimized kernel library for Cortex-M
  - Supports TFLite operators
  - Broadly equivalent to the Arm Compute Library for Cortex-A CPUs

- Offline flow creates a binary for Cortex-M based platforms

- Targets the Cortex-M architectures
  - Armv6-M/Armv7-M/Armv8-M and Armv8.1-M with MVE support
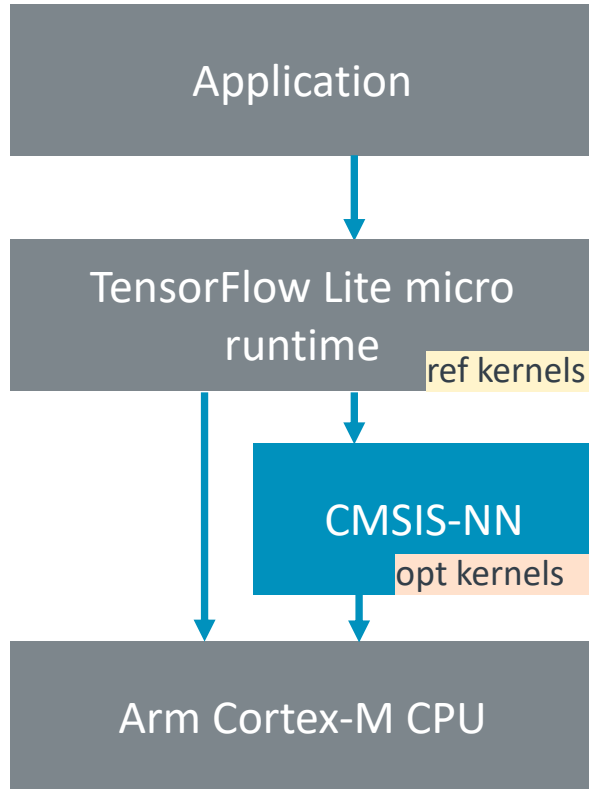 - Runs on earlier versions of the architecture

arm

CMSIS-NN and TensorFlow Lite for Microcontrollers

Access to optimized kernels through TensorFlow Lite micro
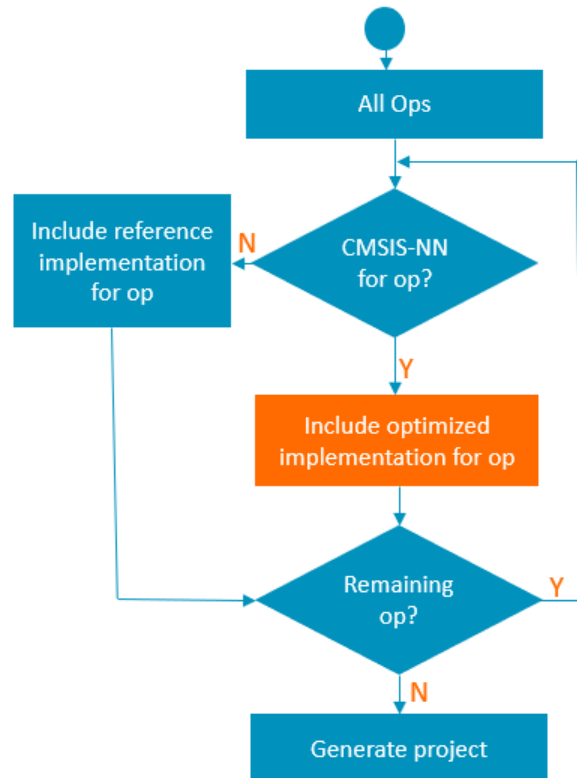
- Support for optimized bit exact int8 kernels

- Fallback on reference kernels when optimization is not available

arm

# Build TensorFlow Lite micro examples for Cortex-M7 and Cortex-M55 FVPs

```
┌─────────────────────────────┐
│         Application         │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   TensorFlow Lite micro     │
│         runtime             │
│                  ┌──────────┤
│                  │ref kernels│
└──────────────────┴──────────┘
    │              │
    │              ▼
    │    ┌────────────────────┐
    │    │     CMSIS-NN       │
    │    │         ┌──────────┤
    │    │         │opt kernels│
    │    └─────────┴──────────┘
    ▼              ▼
┌─────────────────────────────┐
│      Arm Cortex-M CPU       │
└─────────────────────────────┘
```

- Start with the examples in the TensorFlow repo https://github.com/tensorflow/tensorflow/tree/master/tensorflow/lite/micro/examples/

- Add support for your target platform (Cortex-M7 and Cortex-M55 FVP)

- Generate project/binary with 'TAGS=cmsis-nn armclang'

- Build the binary for a simple application (micro_speech_test) that runs an inference on Cortex-M FVP

Docker project in Arm Tool-Solutions repo on Github
https://github.com/ARM-software/Tool-Solutions/tree/master/docker/tensorflow-lite-micro-fvp

arm

TensorFlow Lite Micro examples – micro_speech

Audio detection to detect the words "yes" and "no"

Micro_speech
(int8 model)
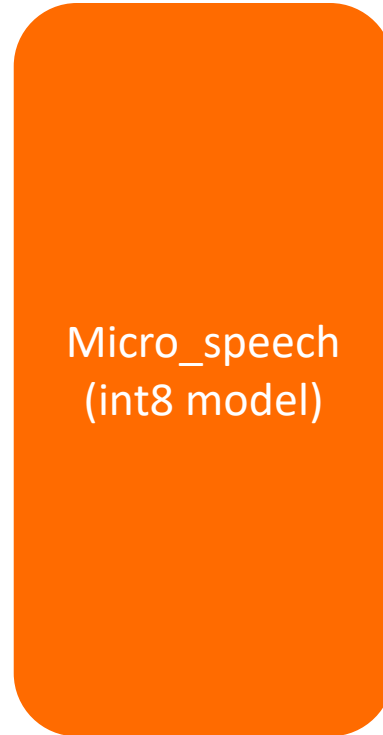
```
Silence: 0%
Unknown: 1.00%
Yes     : 86.00%
No      : 12.00%
```

arm

arm AI

AI Virtual Tech Talks Series

Q & A

arm AI

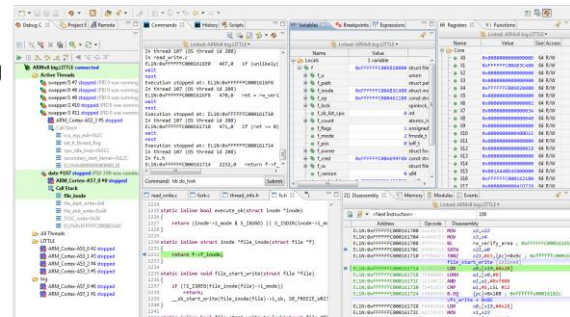## AI Virtual Tech Talks Series

Arm Development Studio

Arm Development Studio

The most comprehensive embedded C/C++ dedicated software development solution

Compiler

Best-in-class, safety-certified code compilation designed alongside Arm IP

Virtual prototypes

Architectural exploration and early software development

Multi-Core Debugger

Family of debug tools for silicon bring-up and software development for complex SoCs
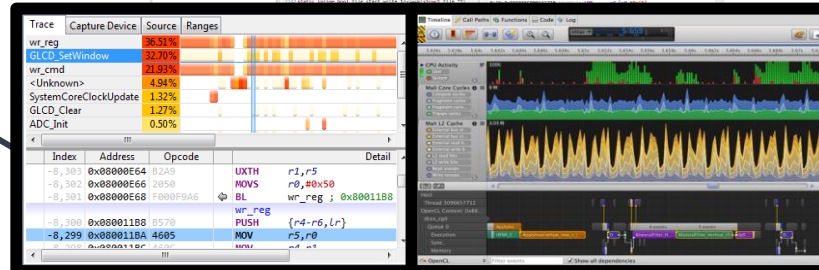
Streamline

Visibility of software performance running in Linux or baremetal

arm

Arm Fast Models and Fixed Virtual Platforms (FVPs)

Earliest architecture support

Accelerate time to market

Supported by experts

- Software development platforms for leading edge Arm IP
  - Available as pre-configured FVPs and model portfolio to build custom platforms
  - Earliest access to simulation models for Arm CPU and SystemIP
  - Ideal for software validation and continuous integration environments

- Performance, fidelity and flexibility
  - High-performance models for software developers
  - Accuracy proven against IP validation suites, used in IP validation flows

arm

arm AI
AI Virtual Tech Talks Series

Demonstration

arm AI
AI Virtual Tech Talks Series

Try it for yourself

Get started

- Get a free 30-day evaluation of Arm Development Studio
  - https://developer.arm.com/tools-and-software/embedded/arm-development-studio/evaluate

- Download the examples used in this webinar
  - git clone https://github.com/ARM-software/Tool-Solutions/

- For more information on Helium technology and get access to programmer's guides
  - https://developer.arm.com/architectures/instruction-sets/simd-isas/helium

- Contact us
  - Arm-Tool-Solutions@arm.com

arm

arm AI

AI Virtual Tech Talks Series

Q & A

arm AI
AI Virtual Tech Talks Series

Join our next virtual tech talk:
Efficient ML across Arm
from Cortex-M to Web
Assembly
by Edge Impulse

Tuesday 25 August

Register here:
developer.arm.com/solutions/machine-learning-on-arm/ai-virtual-tech-talks

arm