

tinyML[®] Foundation

Enabling Ultra-low Power Machine Learning at the Edge

tinyML Summit April 22 - 24, 2024



www.tinyML.org

The background features a view of Earth from space, showing the curvature of the planet and city lights at night. A vertical line divides the image, with the right side transitioning into a vibrant, glowing digital pattern of concentric circles and lines in shades of blue and purple, suggesting a network or data flow.

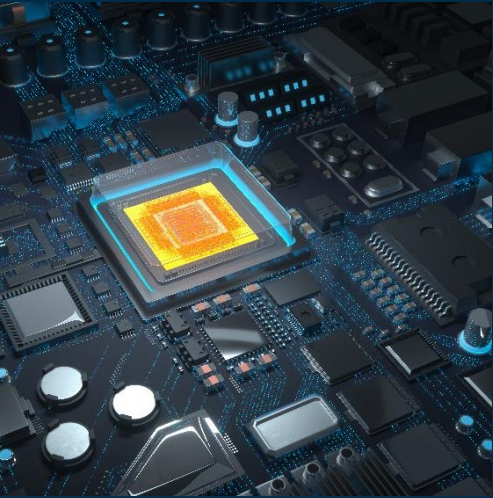
arm

Accelerating Edge AI Innovation

Parag Beeraka
April 2024

© 2024 Arm

The Evolution of Compute at the Edge



Embedded



Connectivity



AI

AI is Creating Incredible Opportunities

PAST

PRESENT

FUTURE



Home



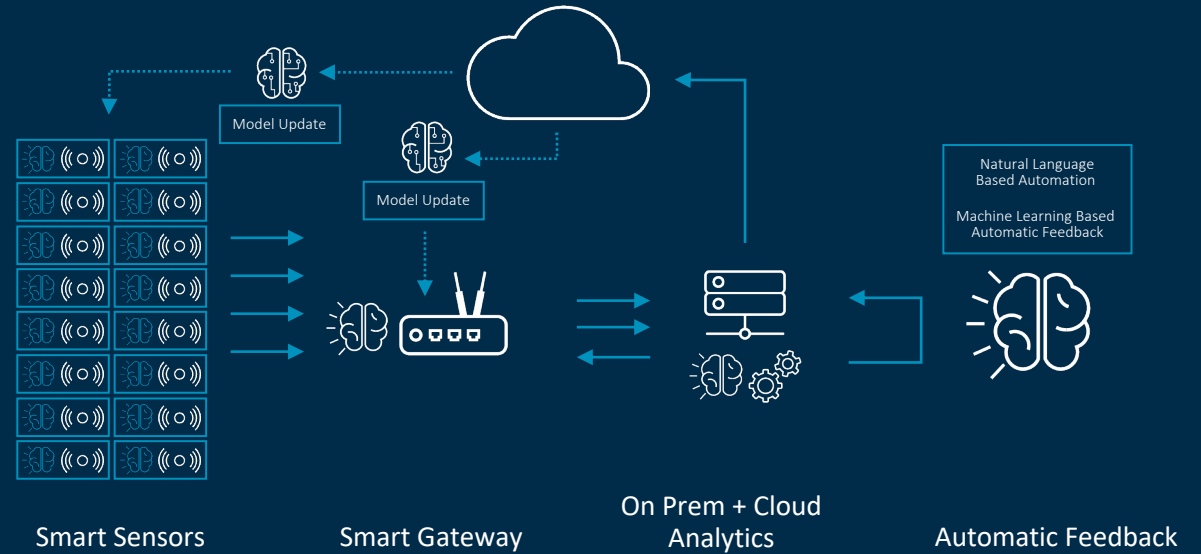
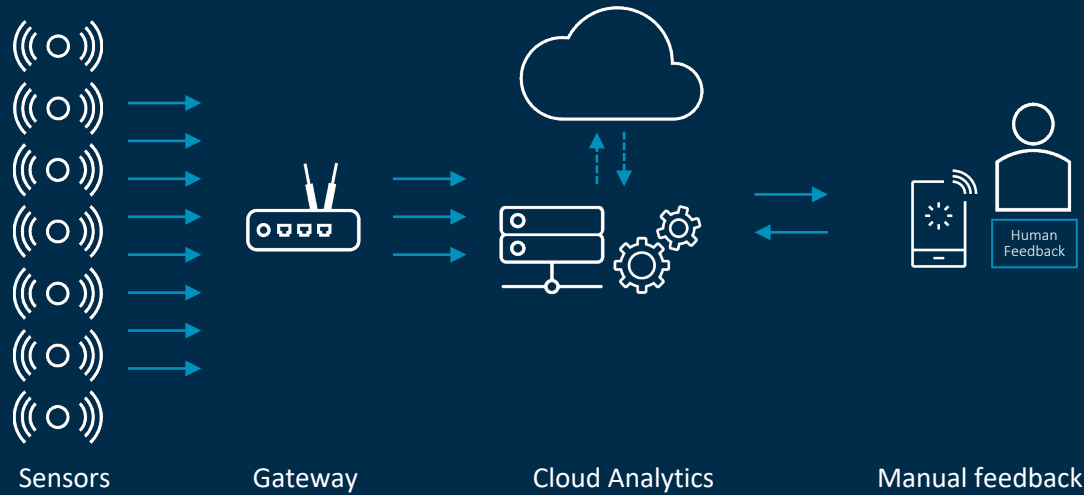
City



Retail

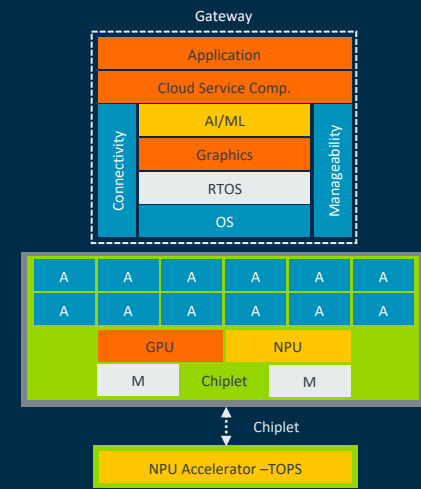
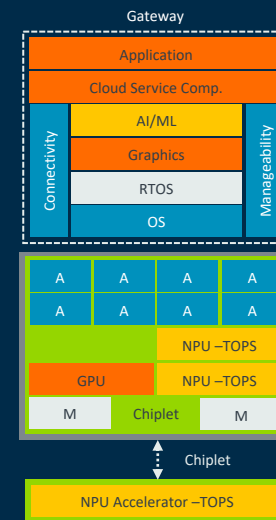
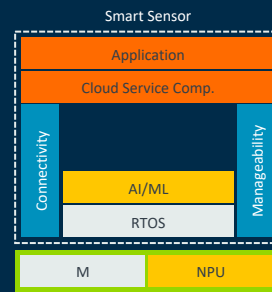
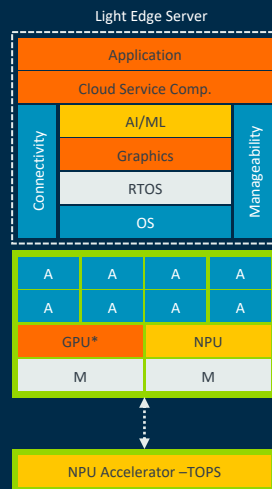
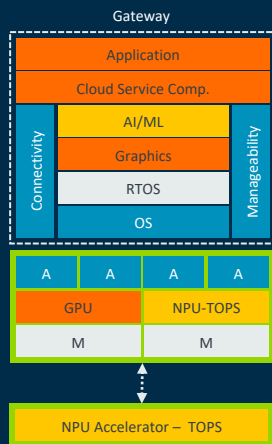
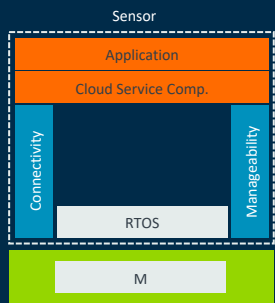


Transformation is Under Way



2024

2030



Generative AI will Further Enhance the Opportunity

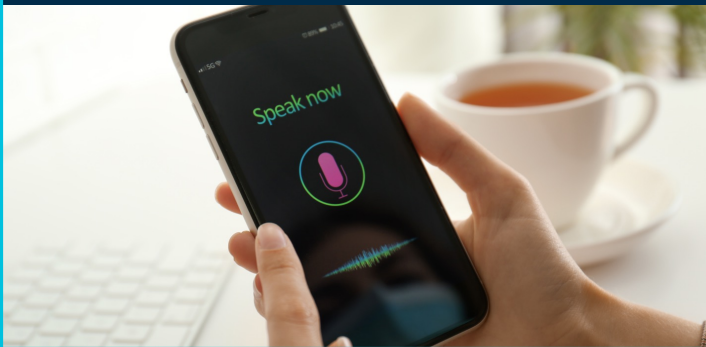
Live Language Translation



Conversational Robots



Speech/Music Generation

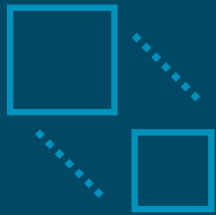


Code Generation (Industrial)



Meeting the Needs of the Market As It Evolves

Navigating growing performance demands and increased software complexity



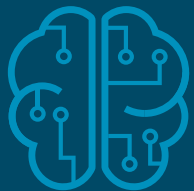
Range of
Performance



Time-To-
Market



Low-Power
Design



Machine
Learning
Integration



Software
Portability



Broad
ecosystem
support

Arm is Enabling tinyML with Three Focus Areas

Hardware



Software



Ecosystem



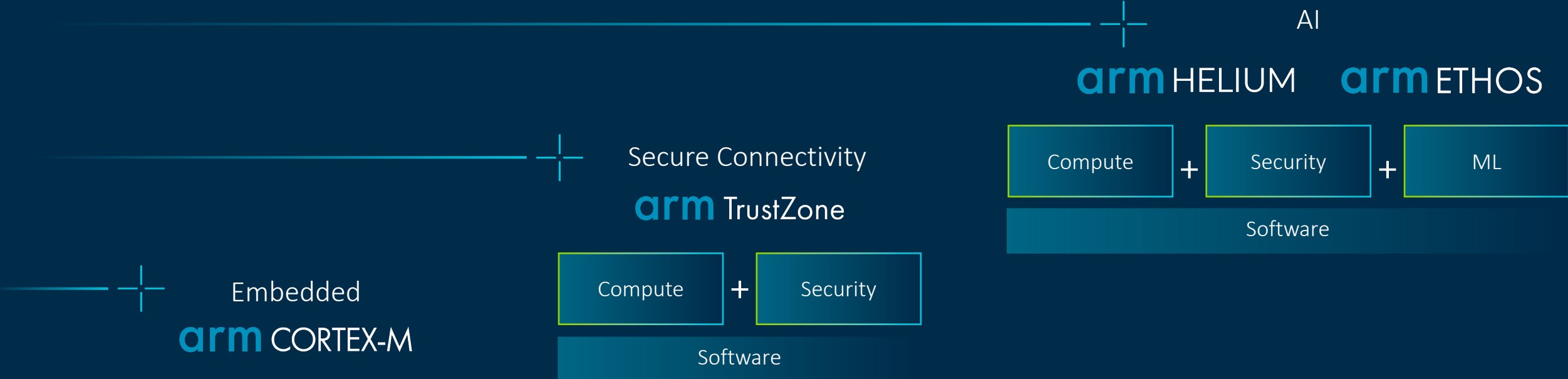
Arm is Enabling tinyML with Three Focus Areas

Hardware

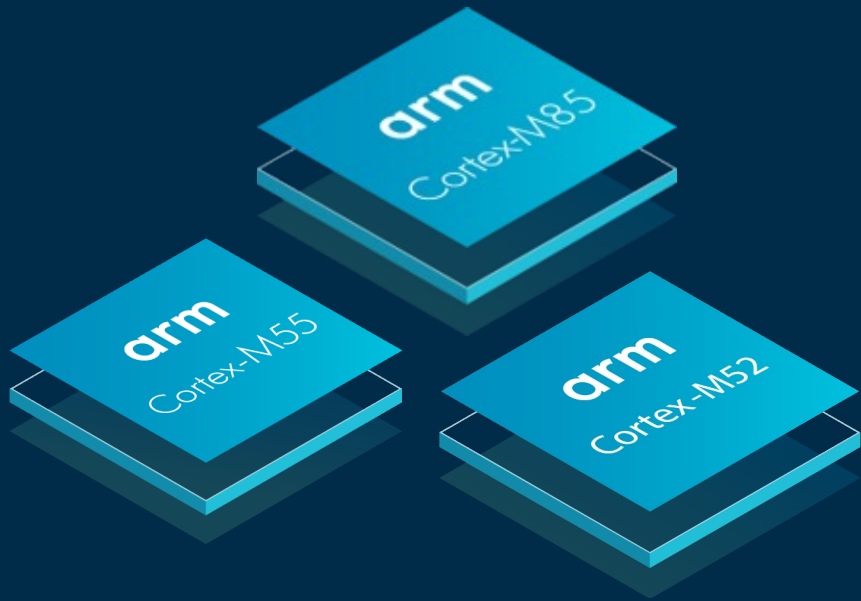


Arm has Delivered Continued Innovation for tinyML

Addressing evolving technology requirements to enable innovation and commercial success



Helium Delivers Significant Uplift for ML on Cortex-M



arm CORTEX-M

Signal Processing → Machine Learning

Signal Conditioning

Feature Conditioning

Decision Algorithm 

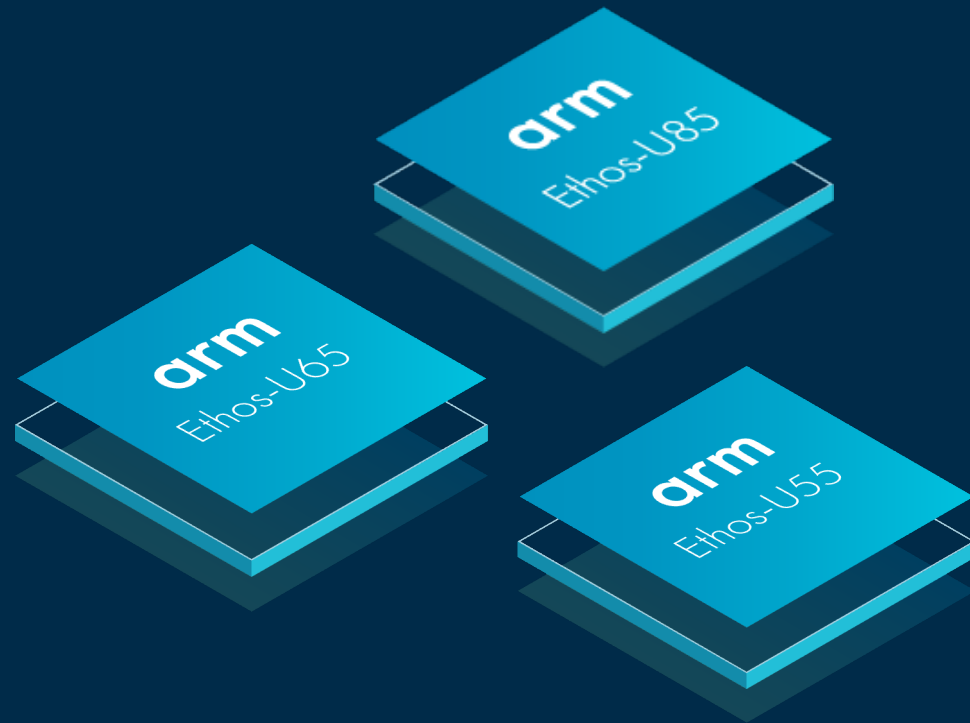
Up to 5x higher signal processing performance*
(CFFT in int32)

Up to 15x higher ML performance*
(matrix multiplication in int8)

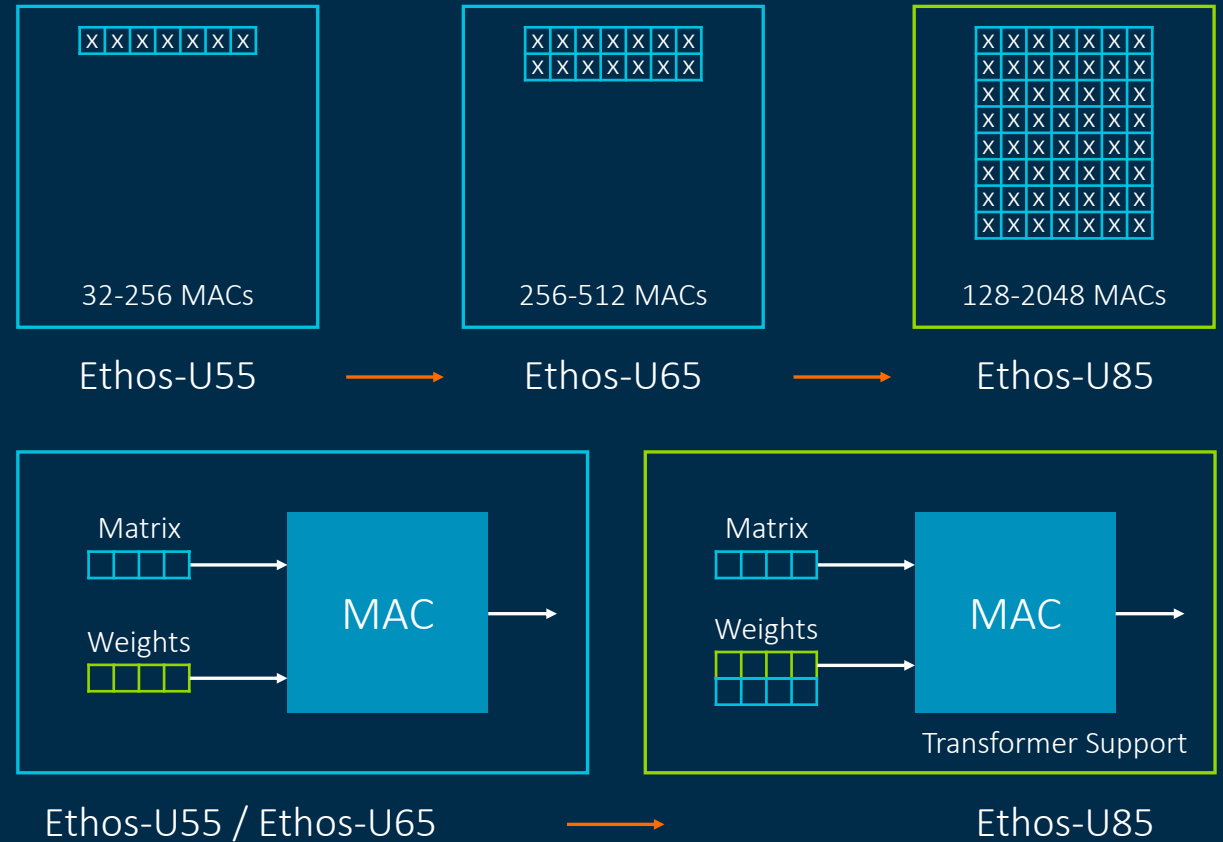
*Compared to existing Armv8-M implementations

Unlocking the Full Potential of Neural Networks

Accelerating implementation of higher performance AI enabled systems



arm ETHOS-U



Arm Ethos-U85

Offering best in class features for neural network acceleration



Increased Performance

Configurations from 128 MACs to 2048 MACs (4 TOPs)



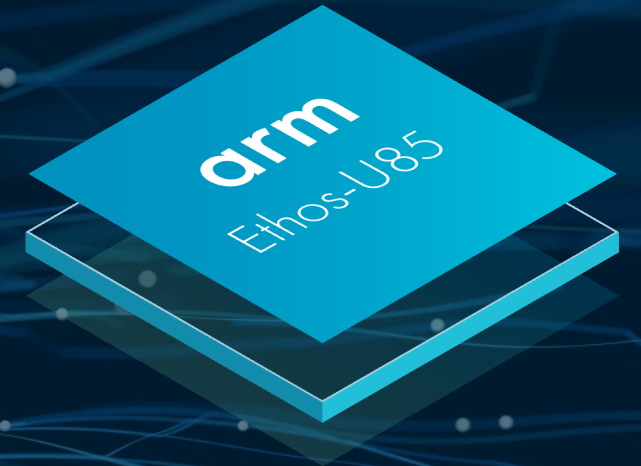
Higher Power Efficiency

20% more energy efficient than previous generation



Extended Operator Support

Transformer network support for faster customization

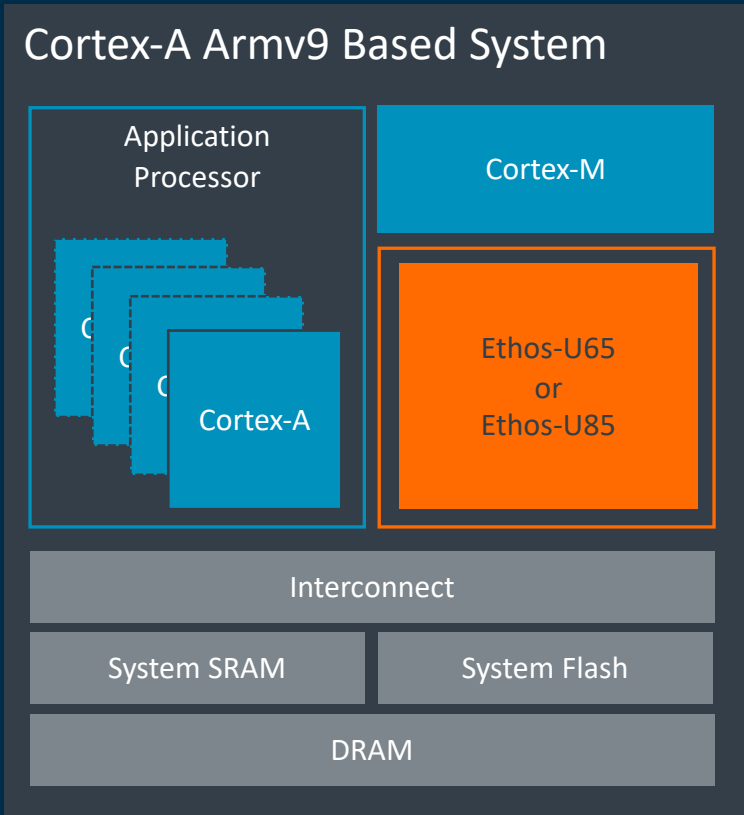
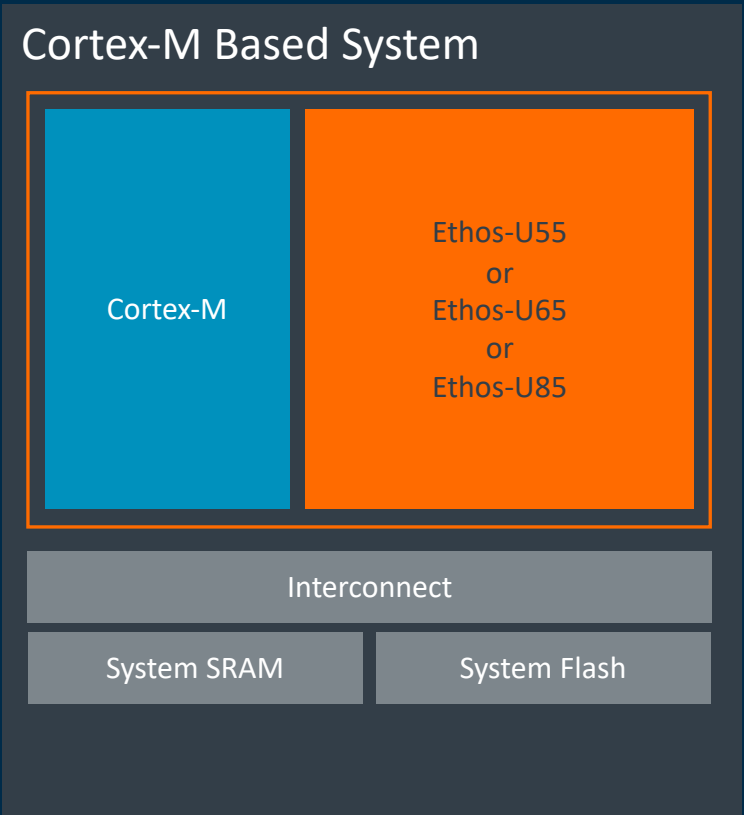


4X

Performance improvement
over previous generation

Arm Ethos-U85 Scales to Support tinyML and other Workloads

Supports small MCU implementations to high performance Armv9 MPU systems



A Holistic Approach for tinyML

Enabling rapid time to market for edge AI device deployment

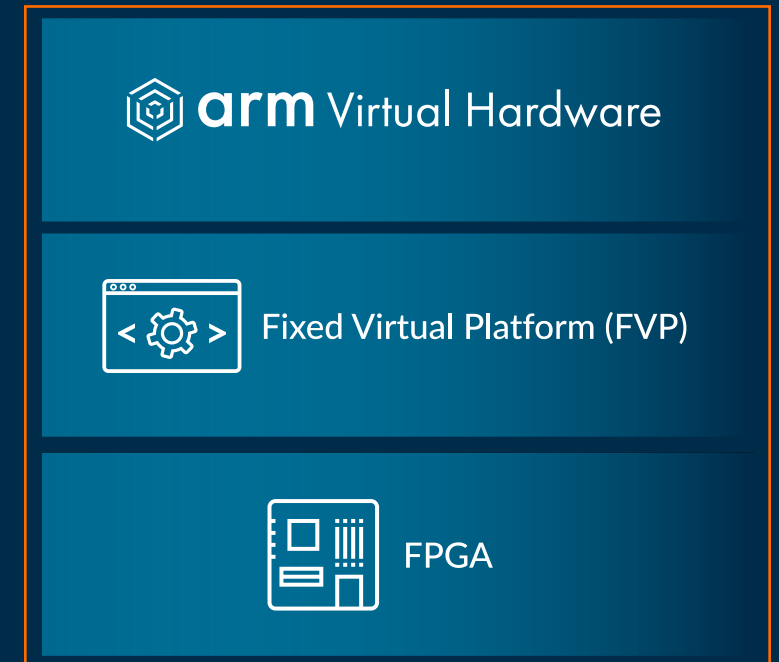
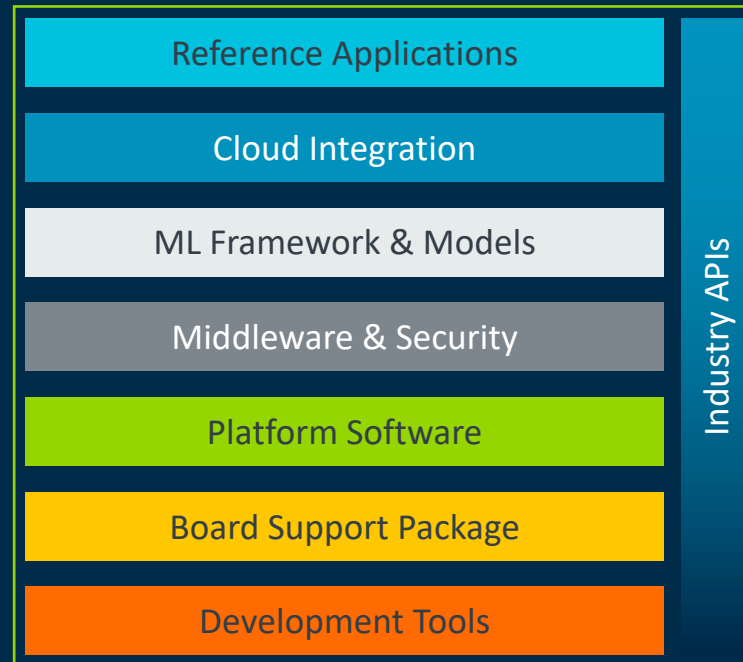
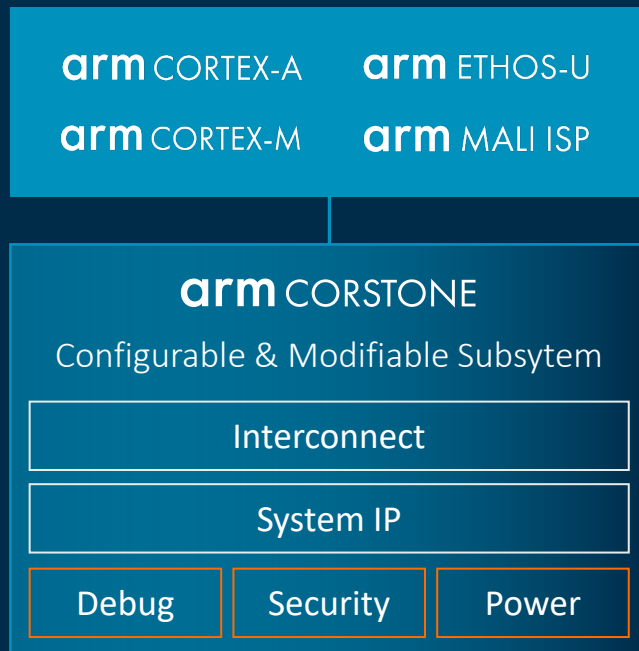
Reference Design



Arm and Ecosystem Software and Tools



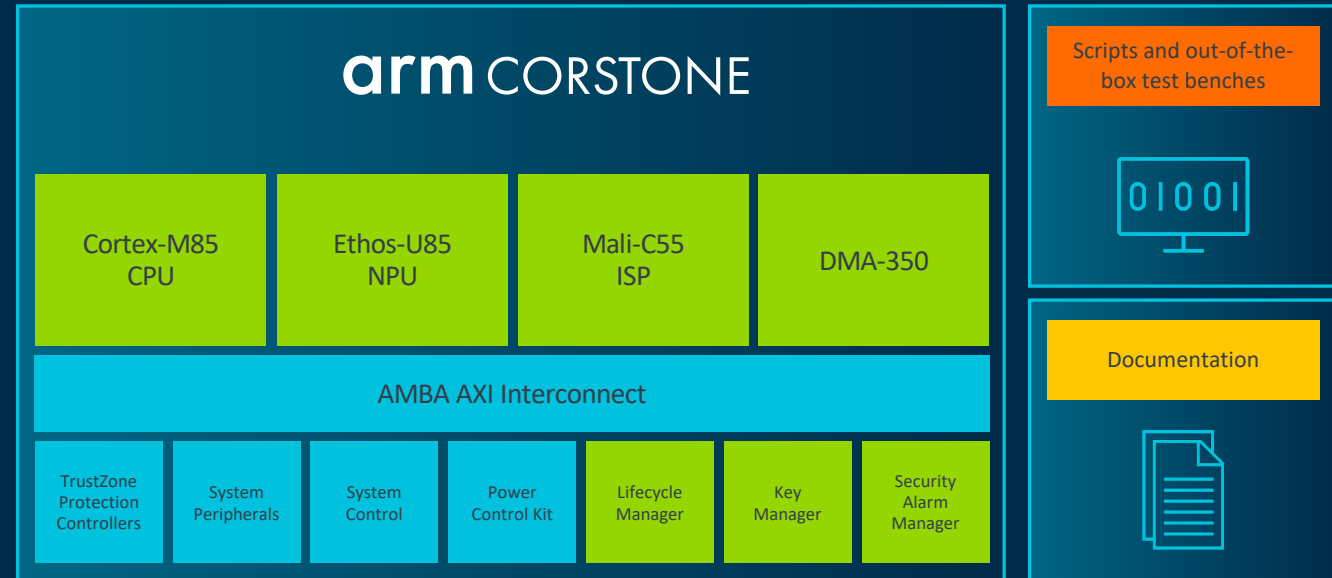
Arm Prototyping Platforms



Corstone-320

Arm's highest performing IoT reference design

- + Bringing together leading-edge embedded IP
 - Cortex-M85, the highest performance MCU processor with Helium technology
 - New Arm Ethos-U85 for up to 4 TOPs AI/ML workload acceleration
 - Mali-C55 image signal processor
- + Suitable for use cases including vision, voice, audio and general-purpose Edge AI
- + Designed to meet IoT security standards



Arm is Enabling tinyML with Three Focus Areas

Software



Enabling Comprehensive Software and Tools

MACHINE LEARNING

FRAMEWORKS

COMPILERS
RUNTIMES

 TensorFlow Lite

 PyTorch

Vela 

 TFLM

 ptvm

SOFTWARE

LIBRARIES

EVAL KITS

CMSIS-NN

ML Eval Kit

 mlia

STANDARDS

IMPLEMENTATION

SPECIFICATIONS

 TOSA

TOOLS

ECOSYSTEM TOOLS

ARM DEVELOPMENT TOOLS

 arm IP Explorer

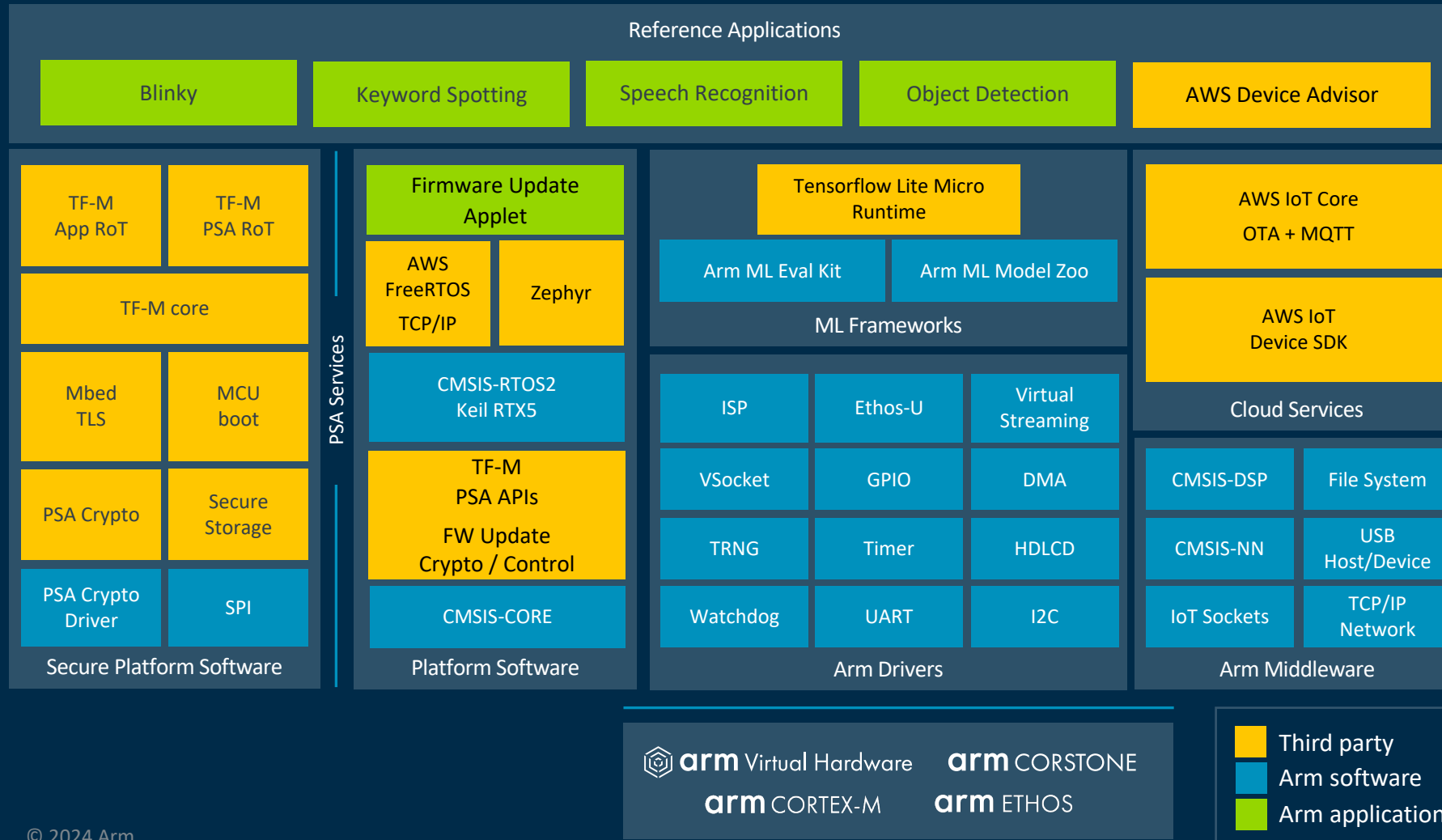
 arm Virtual Hardware

 KEIL
Tools by ARM

 arm DS

Enabling Open Source to Accelerate tinyML

<https://github.com/FreeRTOS/iot-reference-arm-corstone3xx/releases/tag/v202403.00>



Code Development Infrastructure



Standards & Certifications

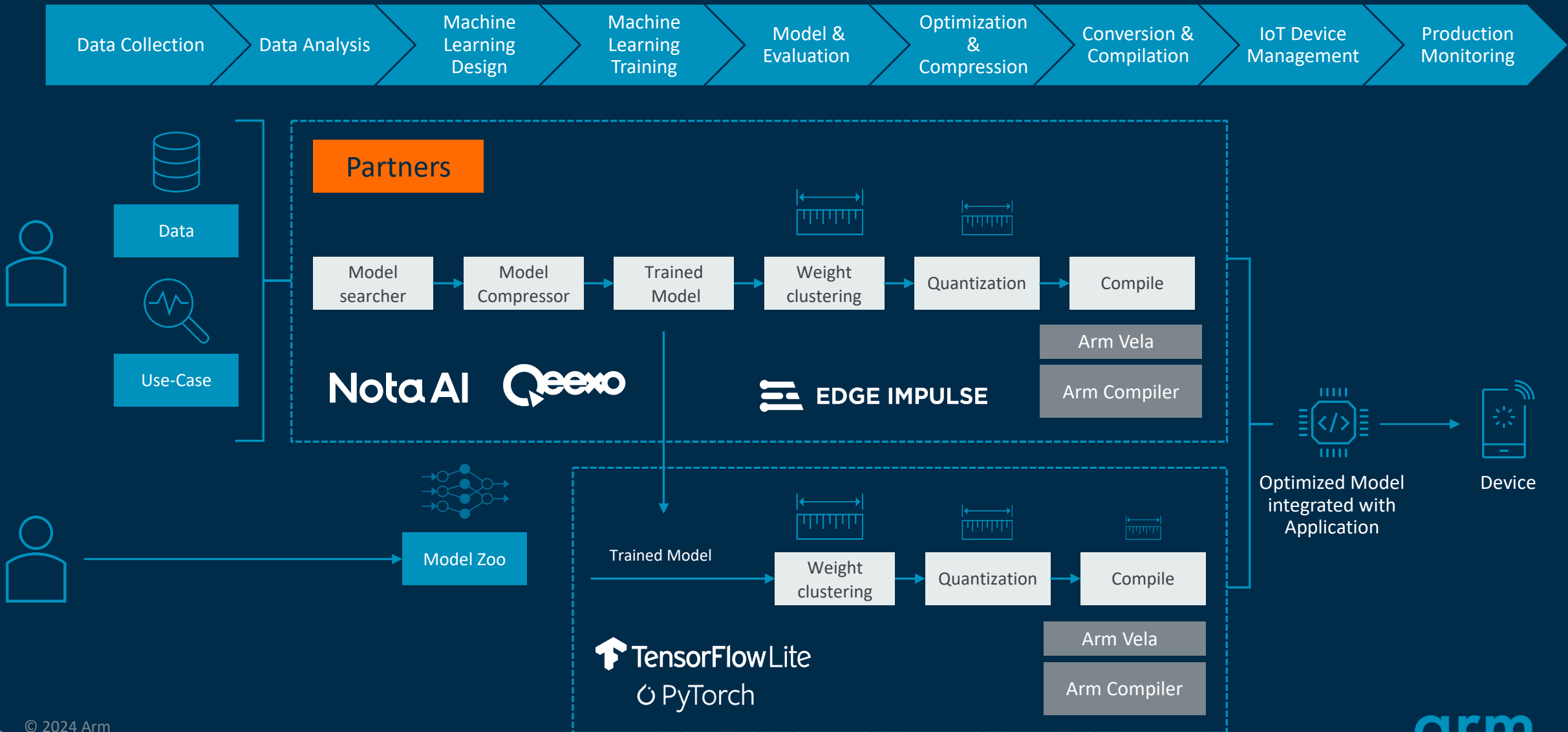
Build Tools

Arm is Enabling tinyML with Three Focus Areas

Ecosystem



Enabling Partners to Have Workflows with Arm Tools



Few specific examples on Ecosystem Enablement for tinyML

 Plumerai

seed studio



embed 

 NVIDIA





RENESAS

 synaptics

NXP

*People
Identification*

*Pose
Estimation*

*Defect
Detection*

*Image
Segmentation*

*People
Detection*

+ many more

Arm is Enabling a Vibrant Ecosystem for tinyML and Edge AI

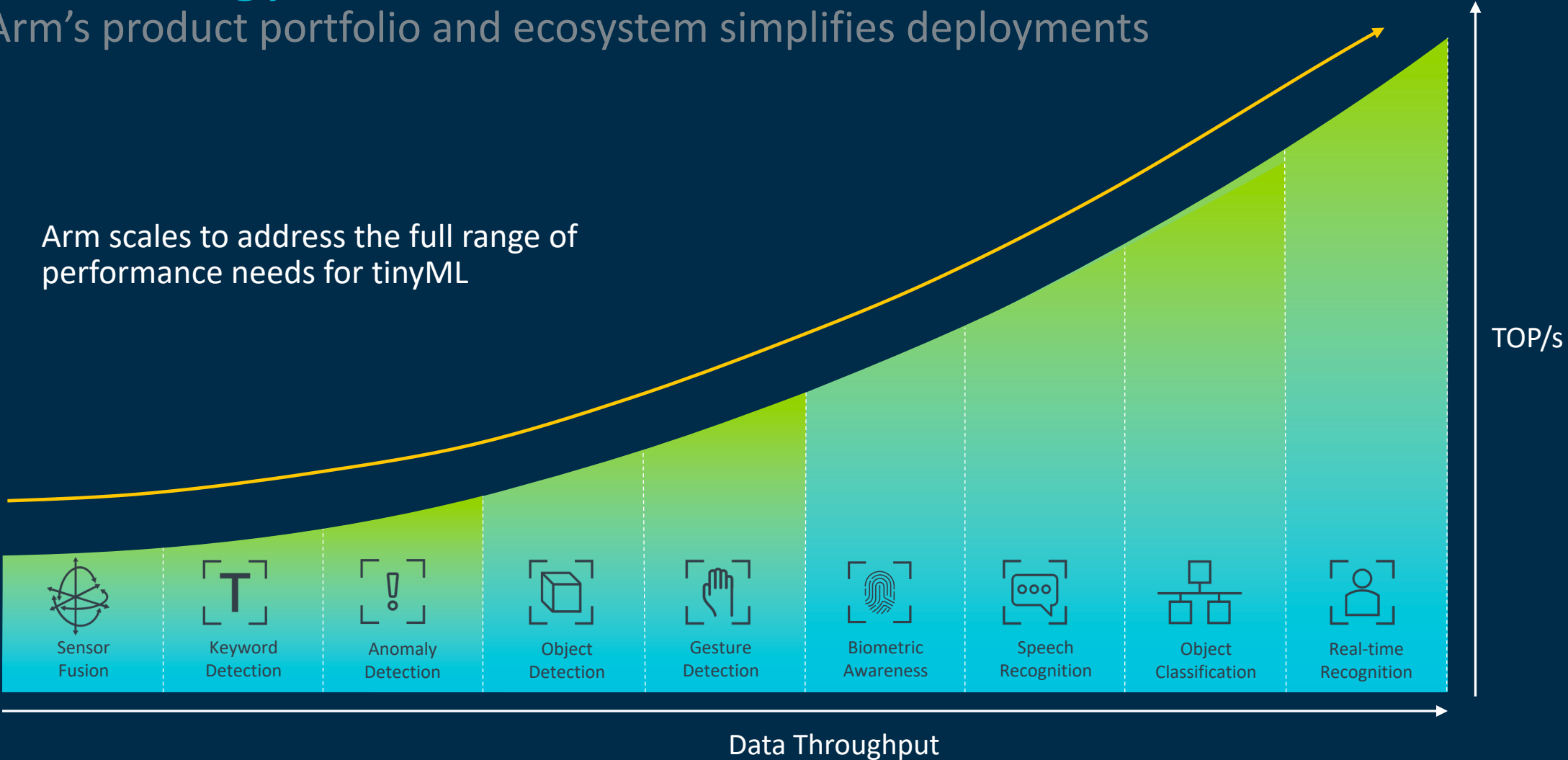
Industry leaders are collaborating to realize the potential of tinyML



Scaling tinyML and Edge AI with the Broadest Range of Technology

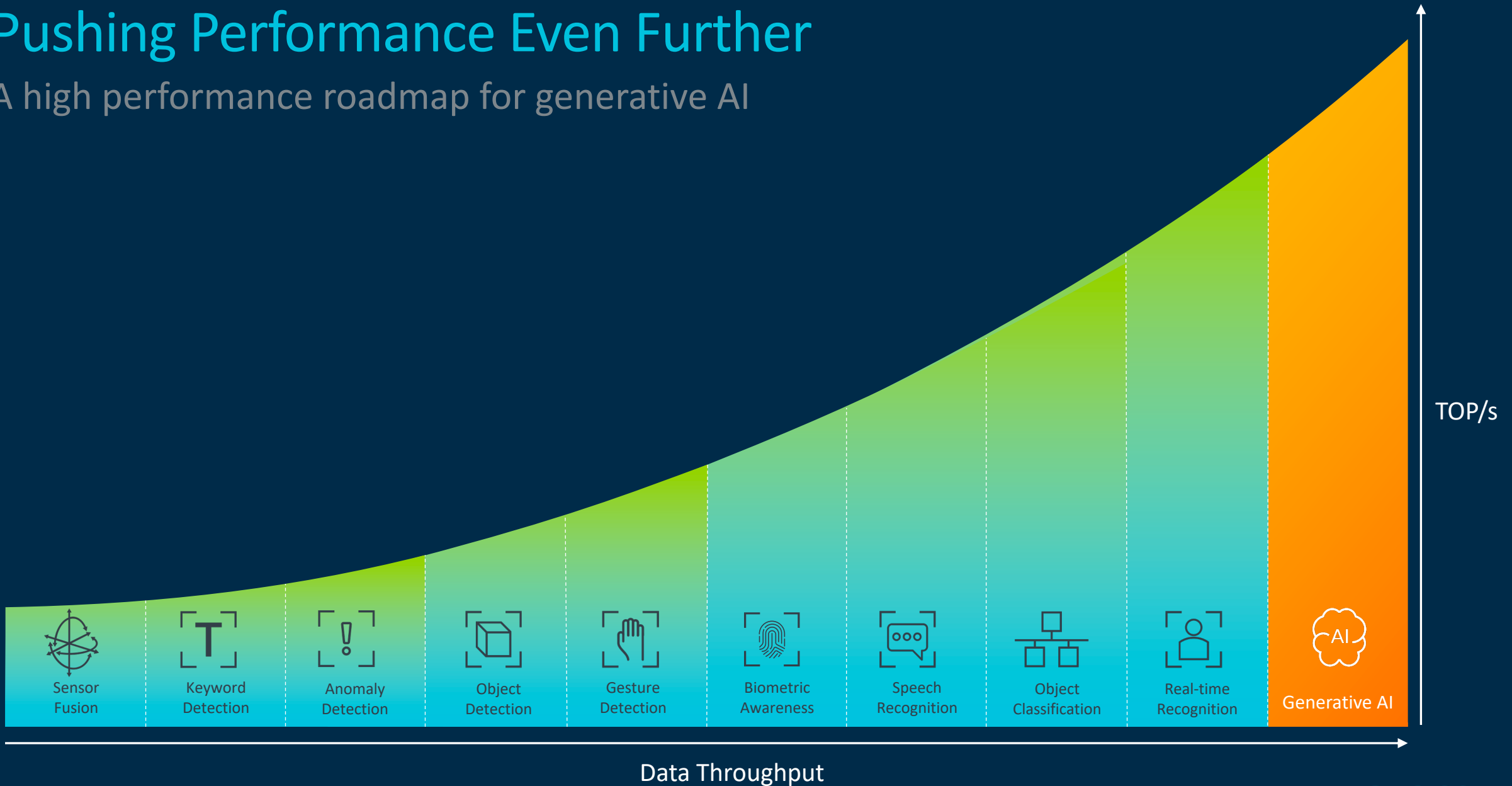
Arm's product portfolio and ecosystem simplifies deployments

Arm scales to address the full range of performance needs for tinyML



Pushing Performance Even Further

A high performance roadmap for generative AI



The Future of tinyML and Edge AI is Being Built on Arm

Enormous opportunity ahead of us to enable tinyML and EdgeAI

Arm is enabling tinyML across multiple vectors - Hardware, Software and Ecosystem

It is important for us to work together to enable a vibrant ecosystem for making tinyML successful

arm

Thank You

Danke

Gracias

Grazie

谢谢

ありがとう

Asante

Merci

감사합니다

धन्यवाद

Kiitos

شكرًا

ধন্যবাদ

תודה

ధన్యవాదములు

The ARM logo is displayed in a white, lowercase, sans-serif font against a dark blue background. The letters are bold and closely spaced.

The Arm trademarks featured in this presentation are registered trademarks or trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All rights reserved. All other marks featured may be trademarks of their respective owners.

www.arm.com/company/policies/trademarks

Copyright Notice

This presentation in this publication was presented at the tinyML[®] Summit 2024. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org