2024 embedded VISION SUMMIT®

# How Arm's Machine Learning Solution Enables Vision Transformers at the Edge

Stephen Su

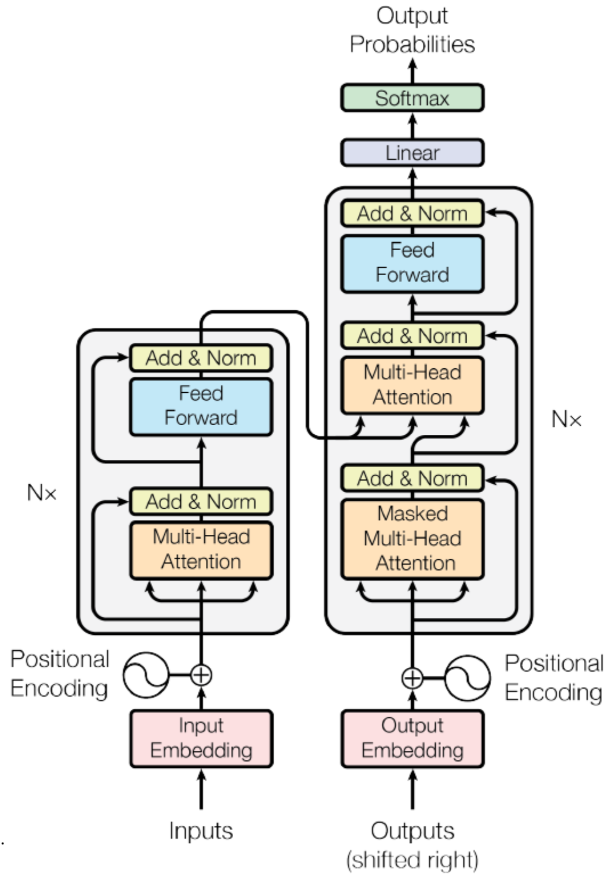
Sr. Segment Marketing Manager

Arm Inc.

arm

# Transformer Background

- What is a transformer? Ref. [1] Vaswani et al. Attention is all you need, NIPS 2017
- A highly scalable network architecture based on self-attention



# Why Transformers?

- Potentially unified architecture for text, audio, and image
- Models based on transformers perform outstandingly in natural language processing (NLP) and computer vision (CV)
- Support wide use cases, not only image classification but also applications such as super resolution, segmentation, object detection, and much more

# Transformer in Vision Applications

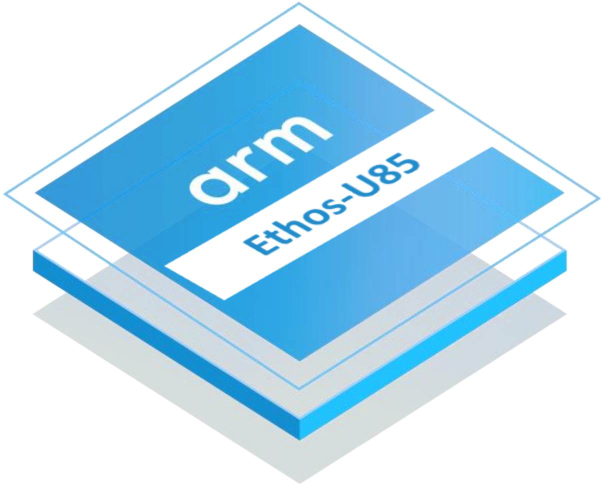
- While CNNs have inductive biases, e.g., locality and translation equivariance,
- The transformer uses self-attention to capture the dependencies within the input sequences
- Hence, models based on transformers are more extendable; i.e., work well in video understanding, image completion, multi-camera, and multi-modal domains

# Challenges in Deploying Transformer Models at the Edge

- Hardware is fragmented, ranging from CPU only, (CPU + GPU), or (CPU + accelerator), and others
  - What is the most suitable hardware solution for transformers?
- Efficiency is another challenge
  - How do you run transformer models with high power efficiency and low latency?
- Model size and memory usage
  - We need a toolset (with tutorials) to compress model size to a reasonable size so that it can be deployed at the edge.

# Arm Machine Learning Solution Supporting Vision Transformers

# Introducing Next Generation Arm NPU—What Makes it Attractive?



## Higher power efficiency

- Targeting **20%** over current generation

## Increased performance

- Configurations from **128** MACs/cycle to **2048** MACs/cycle

## Extended operator support

- **Hardware accelerated transformer network support**

## Double MAC throughput

- For **2/4** sparse layers

# New Hardware Operators Accelerate Transformer Networks

- In addition to the operators currently supported by the original Ethos product family, the latest Arm Ethos-U85 includes native hardware support for transformer networks and DeeplabV3 semantic segmentation network, such as:

TRANSPOSE

GATHER

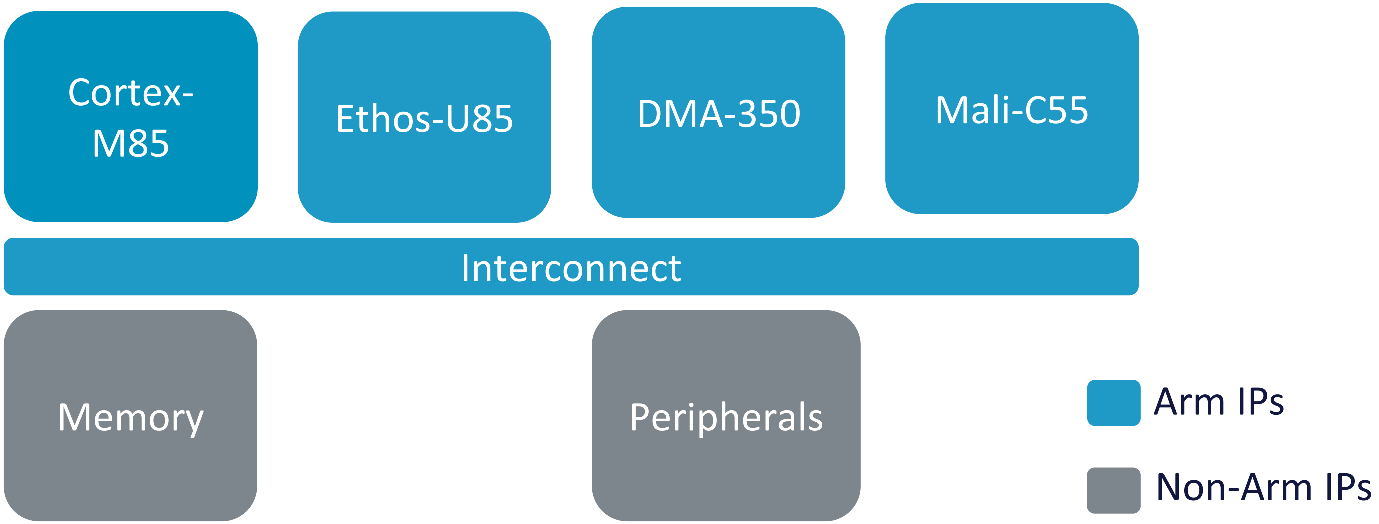
MATMUL

RESIZE BILINEAR

ARGMAX

# Arm Example Subsystem

- Pre-integrated and verified machine learning solution



Cortex-M85

Ethos-U85

DMA-350

Mali-C55

Interconnect

Memory

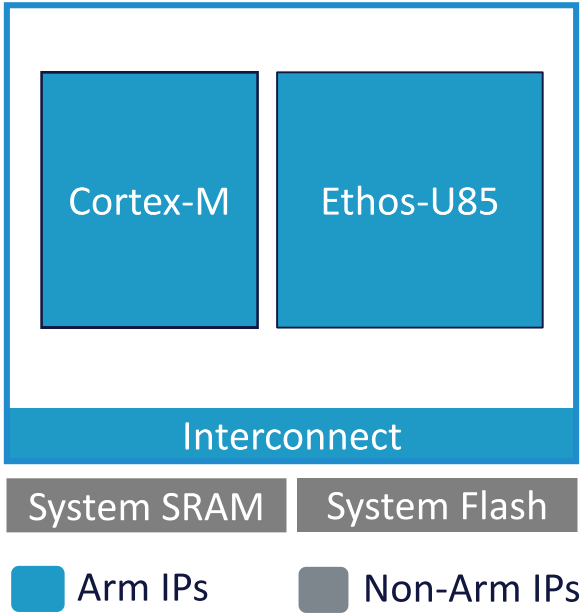
Peripherals

🟦 Arm IPs

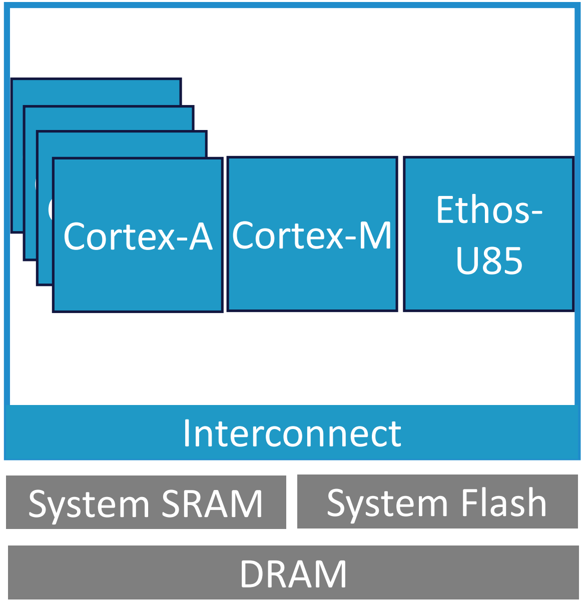
⬛ Non-Arm IPs

# How to Use Ethos-U85 in a System

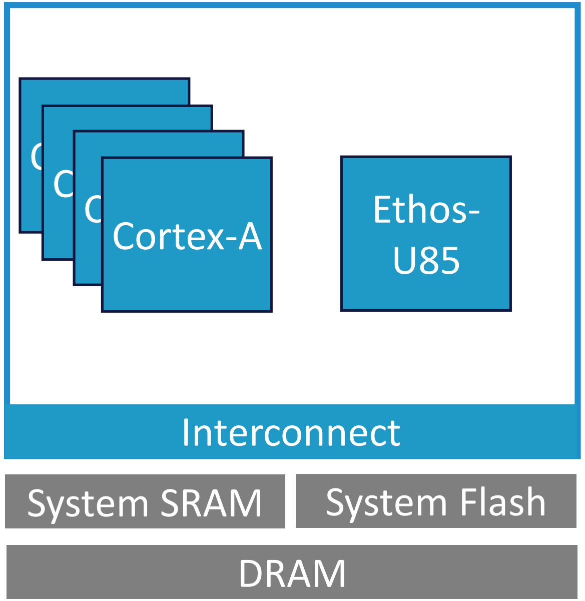
- End Point AI: Cortex-M based system



- ML Island: Cortex-A based system

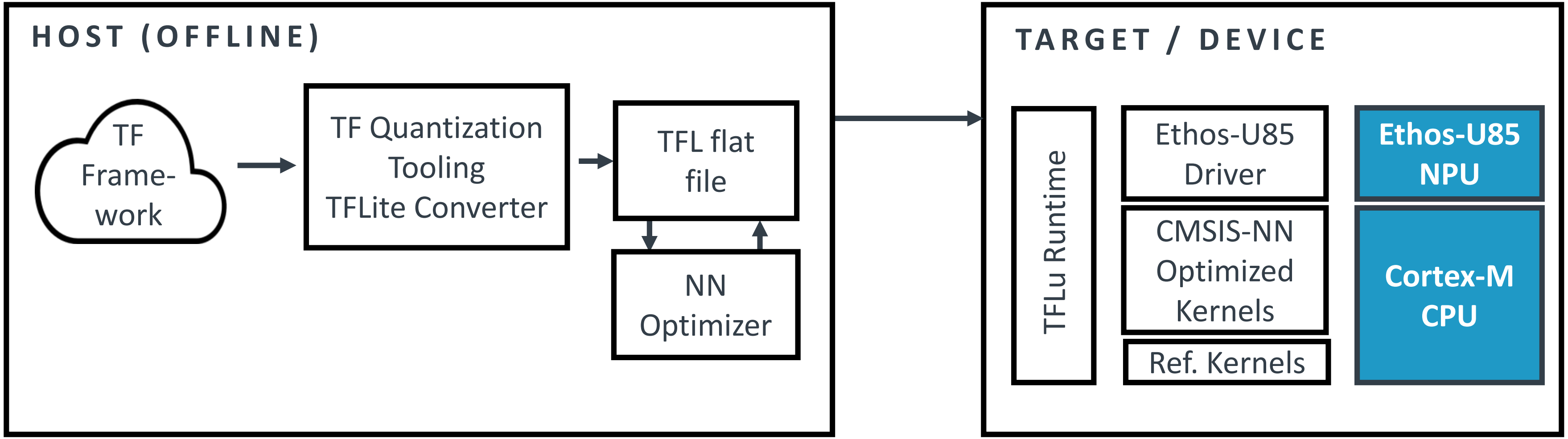


- Discrete NPU: Cortex-A only



# Software Flow on Arm Machine Learning Solution

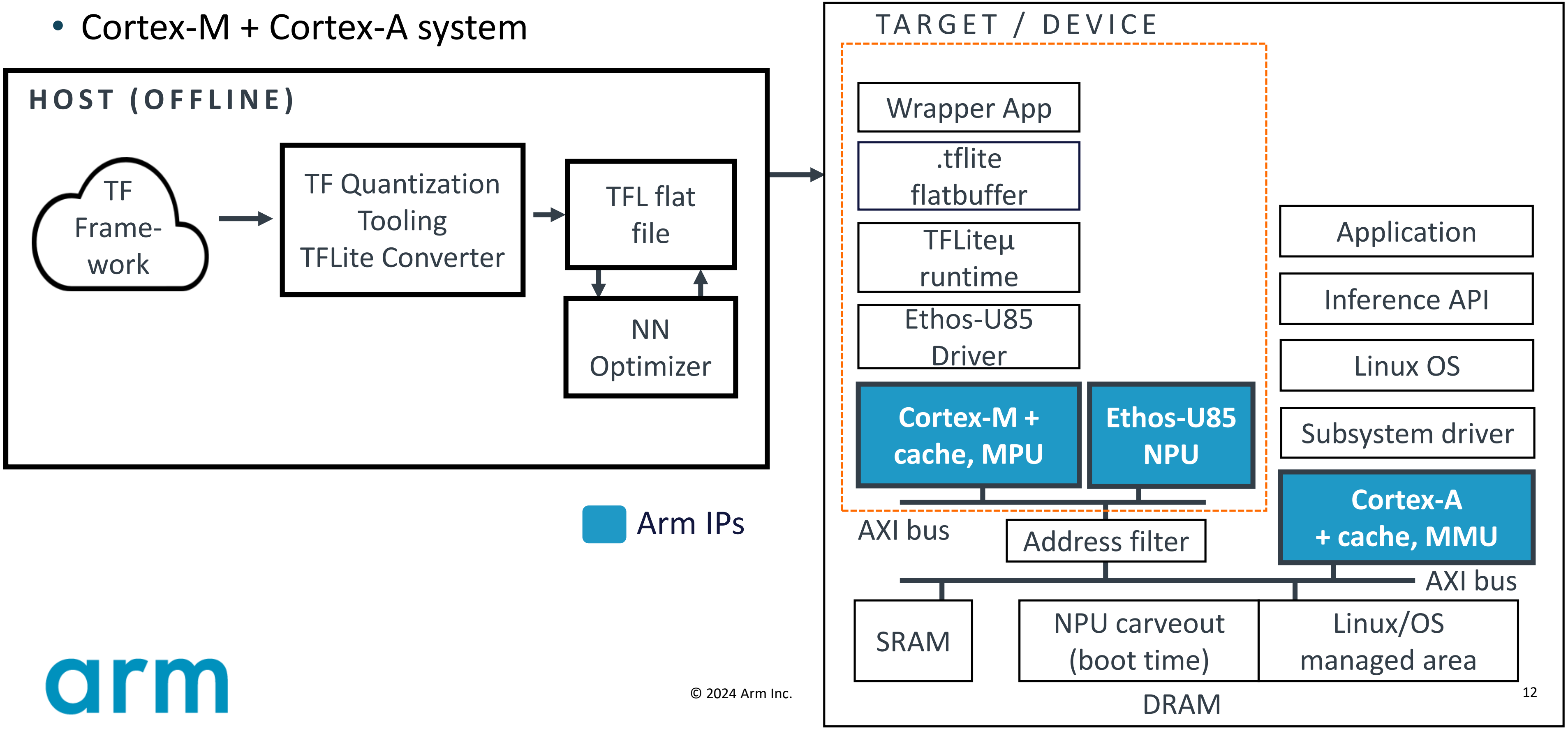
- Cortex-M CPU with Ethos-U85



🟦 Arm IPs

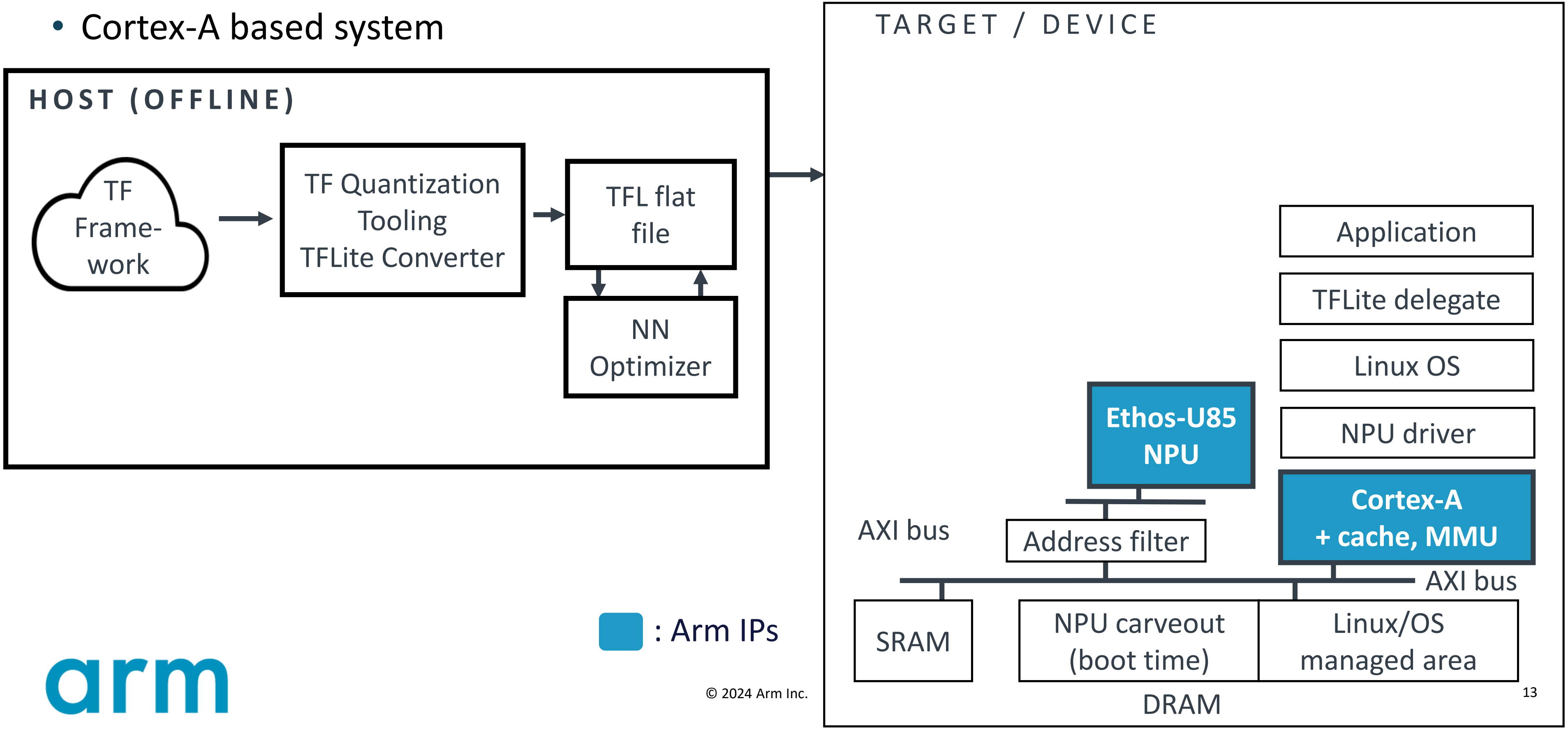
# Software Flow on Arm Machine Learning Solution

- Cortex-M + Cortex-A system



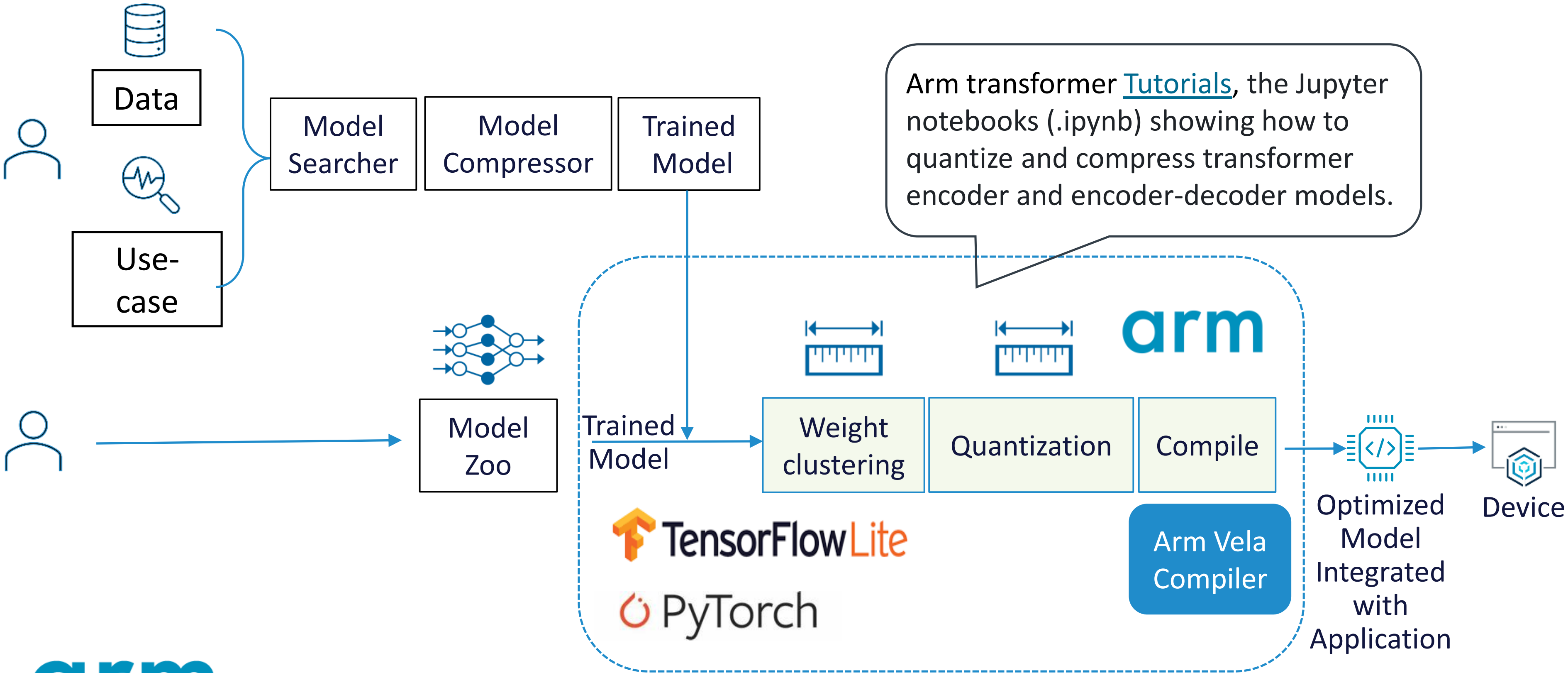
# Software Flow on Arm Machine Learning Solution

- Cortex-A based system



# Arm Toolset Enables the Efficient Implementation of Transformers on Ethos

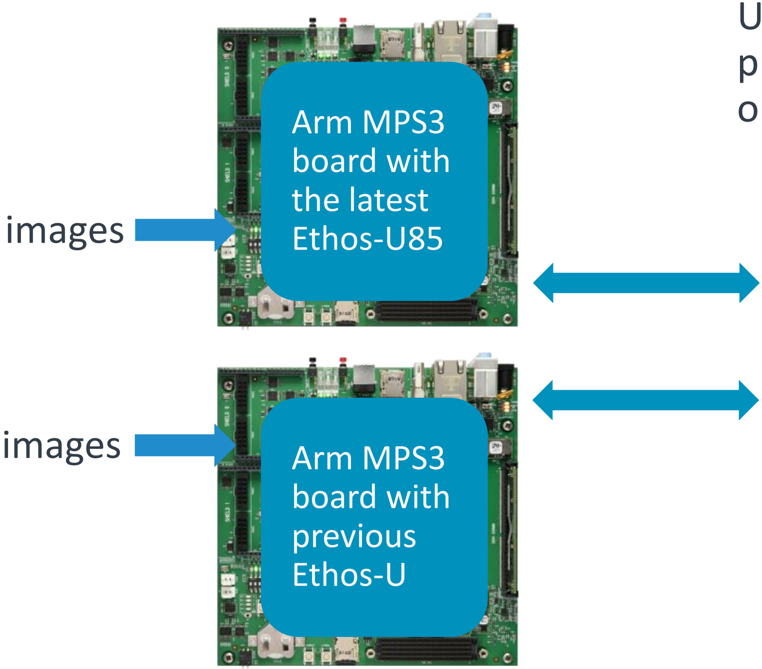
Arm transformer Tutorials, the Jupyter notebooks (.ipynb) showing how to quantize and compress transformer encoder and encoder-decoder models.



# Vision Transformer Example Implementation

# DEiT Tiny Runs on Ethos-U85

Demo is to compare how much faster the latest Ethos-U85 runs a transformer network compared to the previous Ethos, since there is no fall back for those operators with Ethos-U85

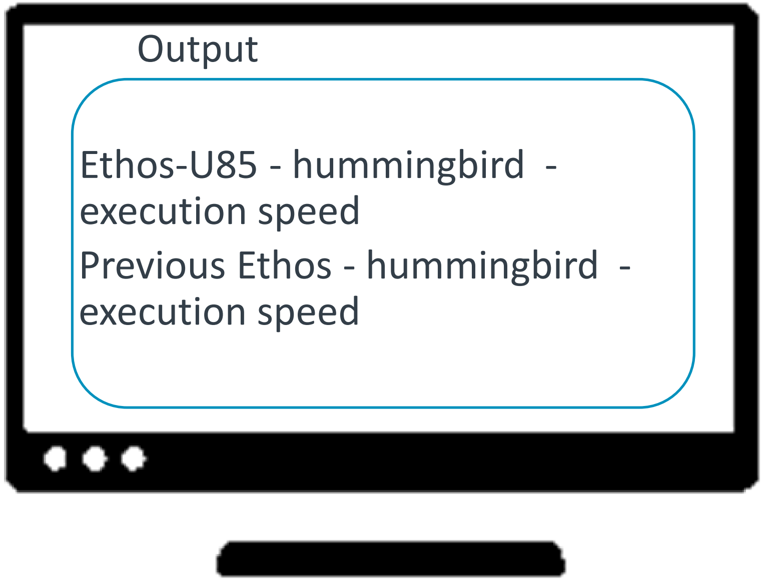


images

Arm MPS3 board with the latest Ethos-U85

images

Arm MPS3 board with previous Ethos-U



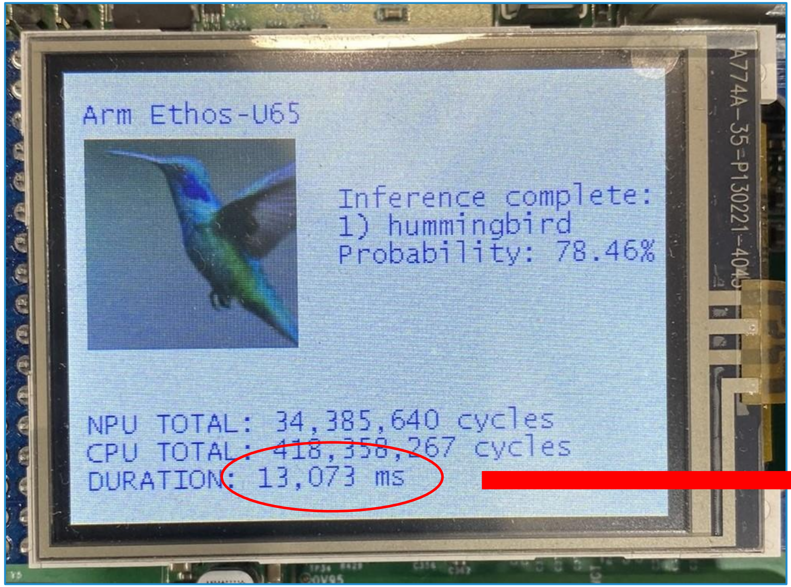
Output

Ethos-U85 - hummingbird - execution speed

Previous Ethos - hummingbird - execution speed

# Up to 8X Acceleration in Inference time

- Previous Ethos

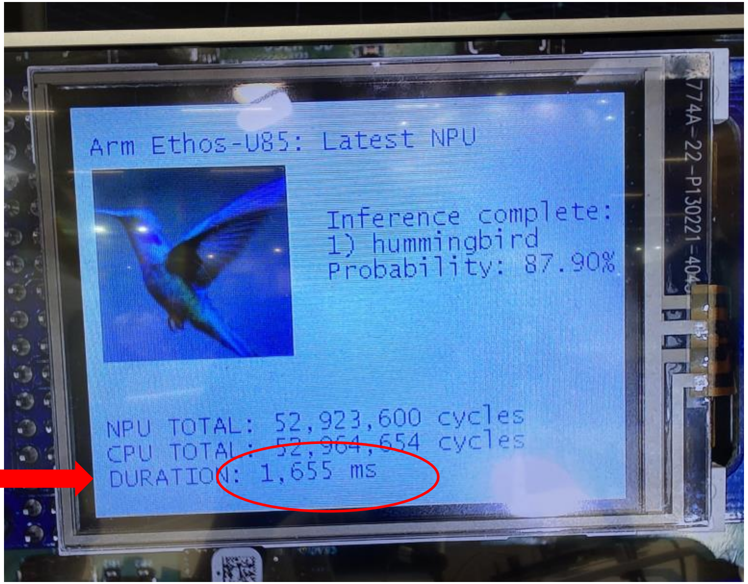


```
Arm Ethos-U65

Inference complete:
1) hummingbird
Probability: 78.46%

NPU TOTAL: 34,385,640 cycles
CPU TOTAL: 418,358,267 cycles
DURATION: 13,073 ms
```

- The Latest Ethos-U85



```
Arm Ethos-U85: Latest NPU

Inference complete:
1) hummingbird
Probability: 87.90%

NPU TOTAL: 52,923,600 cycles
CPU TOTAL: 52,964,654 cycles
DURATION: 1,655 ms
```

For more details, please visit Arm booth at **#409**.

# Summary

- Machine learning (ML) is everywhere, and its landscape is evolving from CNNs to transformer-based models
- Arm just launched the latest NPU in the Arm Ethos product family to extend the support of accelerating transformers at the edge
- Finally, “Edge AI runs on Arm.”

# Resources

Arm Ethos-U product page
https://www.arm.com/products/silicon-ip-cpu?families=ethos%20npus

Arm transformer tutorials
https://github.com/ARM-software/ML-zoo/tree/master/tutorials/transformer_tutorials

Arm keyword-transformer
https://github.com/ARM-software/keyword-transformer

**Please visit Arm booth #409 at the 2024 Embedded Vision Summit for more demos:**

"The Newly Launched Arm Ethos-U85 NPU"

"Renesas RZ/V2H- Qual-core Cortex-A55 Vision AI MPU"

"Arm-Himax, the High-efficiency Embedded Computer Vision"

# Reference

- Reference [1]: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010

# Thank You