arm

TVM at Arm

TVM Summit – Seattle

Ramana Radhakrishnan
u99127/@ramana-arm

5th December 2019

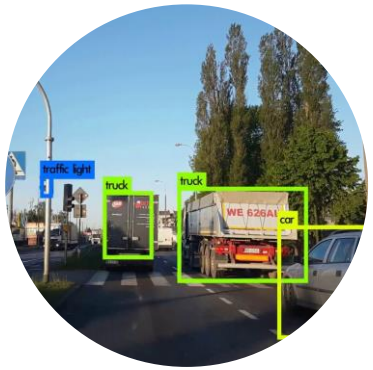
© 2019 Arm Limited

Agenda

- AI / ML in End Points
- Brief overview of Arm's ML Platform
- Using TVM in Arm
  - Current Areas of Interest
  - Future Areas of Interest
- Challenges / Observations

© 2019 Arm Limited

What is AI Being Used for in Endpoints?



Vision

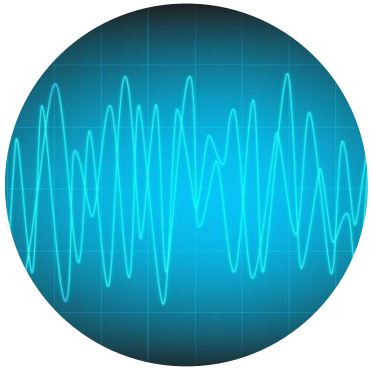
Images and video

Object detection, face unlock, defocus (bokeh), beautification, scaling, etc.

Voice

Recognition and creation

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



Vibration

Any ‘signal’

Accelerometer, pressure, lidar/radar, speed, shock, vibration, pollution, density, viscosity, etc.

AI performs well with ‘patterns’ of data

© 2019 Arm Limited

Diversity of AI Requirements in the Market

Premium

- Best user experience and responsiveness
- Highest performance in power-efficient design

Balanced

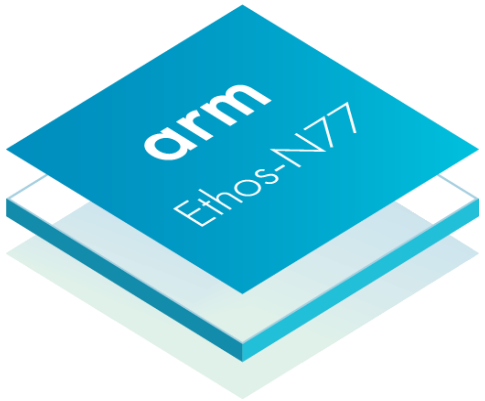
- Superior user experience in mid-range designs
- Balance performance with area and power

Cost Sensitive

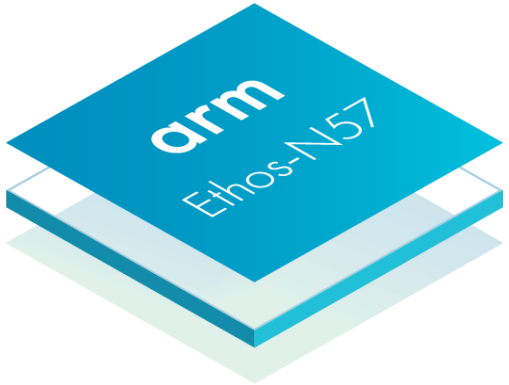
- Delivering advanced user experiences for the most cost-sensitive designs
- Optimized for performance in the smallest area

© 2019 Arm Limited
arm

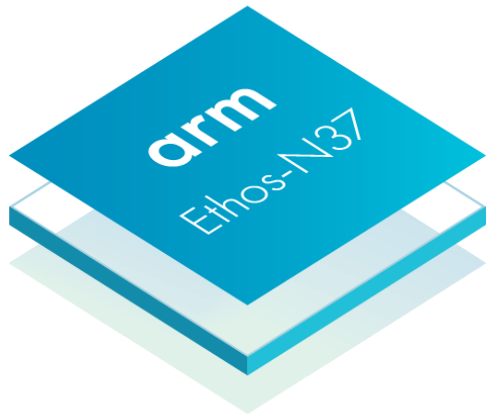
Introducing Ethos NPUs for Every Market Segment



Performance-critical AI applications delivering premium experiences



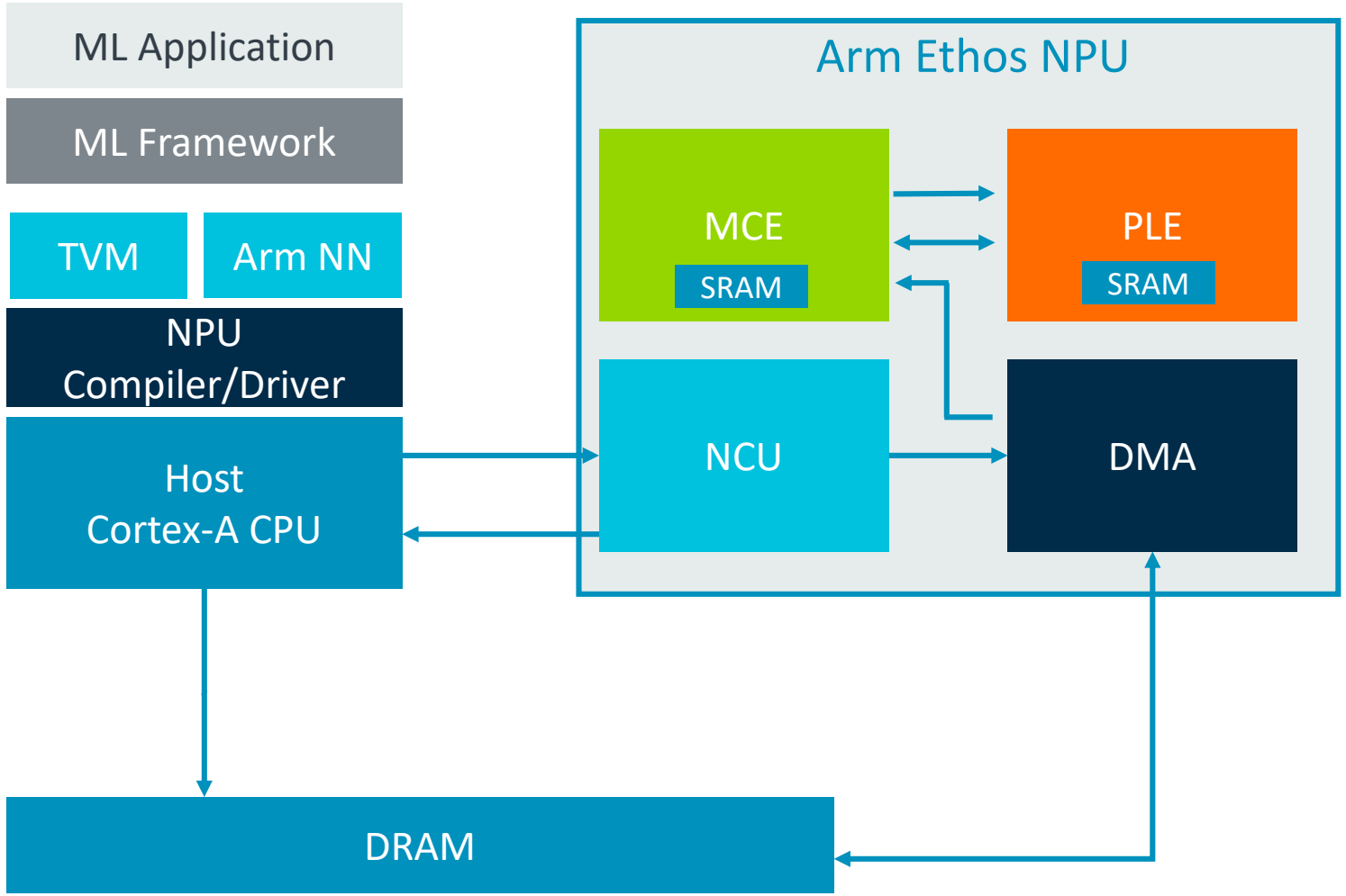
Enabling AI applications in mid-range devices balancing performance with cost and battery life constraints



Supporting AI applications in the most cost-sensitive endpoint devices

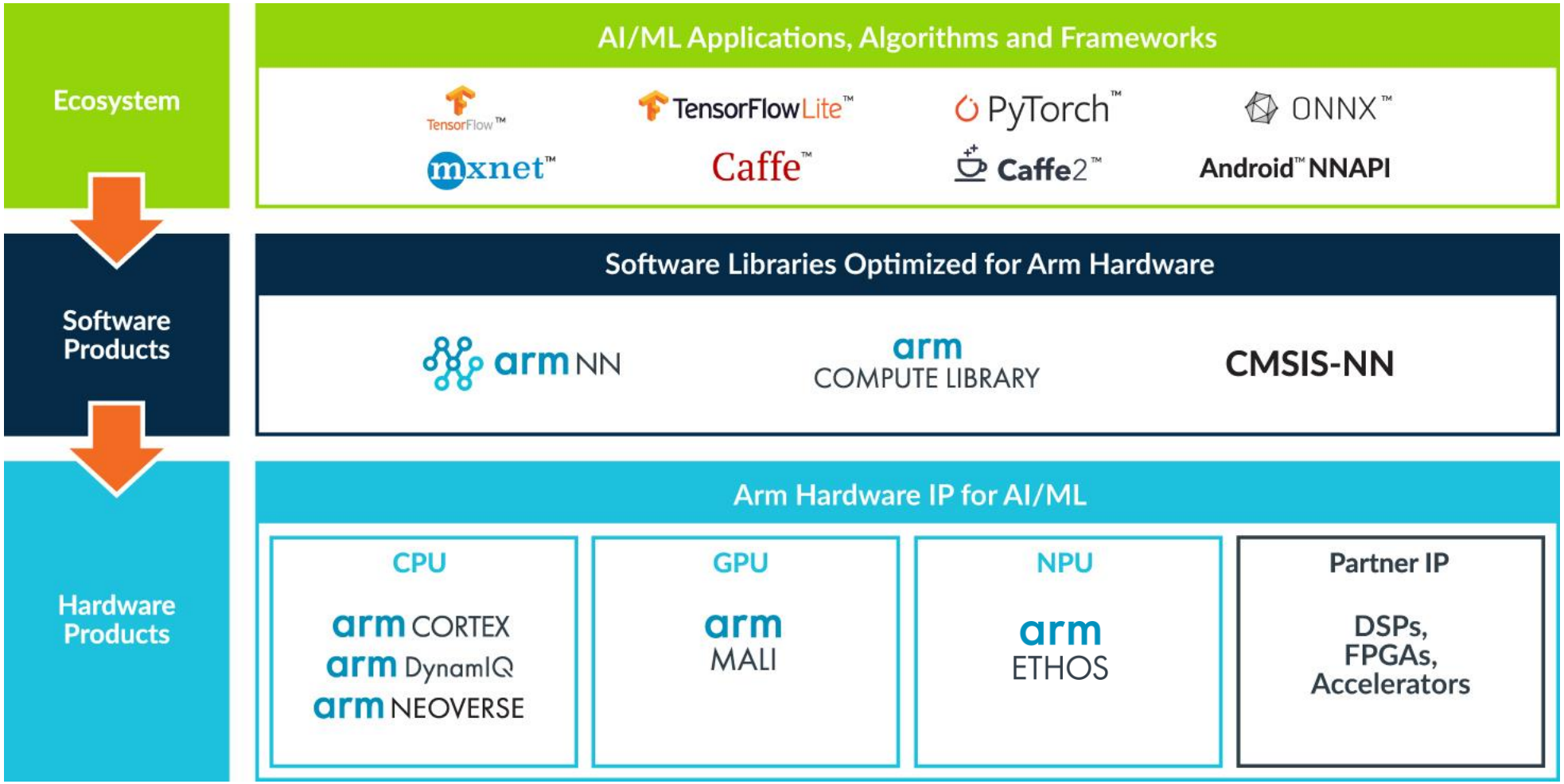
© 2019 Arm Limited
arm

Ethos NPU Software Stack



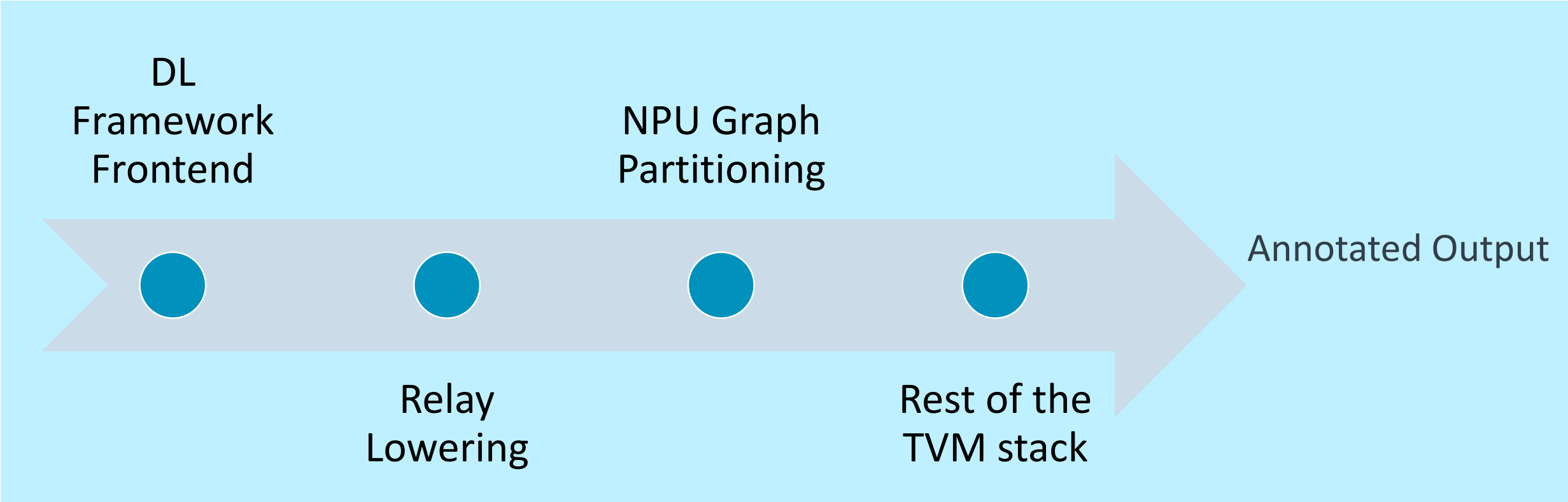
© 2019 Arm Limited
arm

Comprehensive ML Platform Makes Developing AI Easy



© 2019 Arm Limited
arm

Ethos Integration into TVM – Compile Time



© 2019 Arm Limited
arm

Current Areas of Work with TVM

Arm CPU and GPU

- Support for Mali Bifrost schedules.
  - Improvements by about 20-70%
  - Interested in Arm CPU architecture support
  - Investigating Arm Compute Library integration

General Areas

- Pre-quantized TensorFlow-Lite
  - Some operator support
- Framework versioning.
- Reviewing various bits of Arm architecture support.
- LLDB Pretty Printers
- Investigating µTVM

Ethos NPU

- Graph Partitioning for NPU
- Integrating support for Ethos-N77, Ethos-N57 and Ethos-N37

© 2019 Arm Limited
arm

Future Areas of Interest

General Framework

- Graph partitioning and Relay optimizations.
- Improvements to code generation and generic optimizations.
- Auto-tuning
- Common command line utilities for using TVM.
- Improvements to Continuous Integration / Testing
- µTVM

Arm Architecture Support

Armv8-A architecture

- Scalable Vector Extensions (SVE/SVE2)
- Matrix Multiplication Support
- BFloat16
- Improved Advanced SIMD Support

GPU Support

- Mali Bifrost
- Mali Valhall

Cortex-M Architecture

- DSP Instructions support.
- Support for Helium / MVE.

© 2019 Arm Limited
arm

Challenges / Opportunities

Deployment

Getting ready for packaging

- conda
- pypi packaging
- Integration with native packaging

Release process

Continuous Integration

- Execution tests and performance monitoring.
- Managing version updates in frameworks

Scalability

Developmental Practice

- Features vs Bug fixes.
- Isolation of changes.

Developer efficiency

- Better explanation with changes
- Debug helpers.
- Understanding the test infrastructure.

Getting Started

- Make it easier !

© 2019 Arm Limited
arm

And finally, we are hiring!

© 2019 Arm Limited
arm

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

© 2019 Arm Limited

tvm-driver

- tvm-driver –help
  - compile
    - --debug-relay-all
    - --debug-tvm-all
    - --debug-all
    - --print-llvm
    - --print-assembler
  - execute
    - --native
    - --remote
  - auto-tune

Motivation

- Ease of use of TVM stack
- Common way of getting hold of outputs.

© 2019 Arm Limited
arm