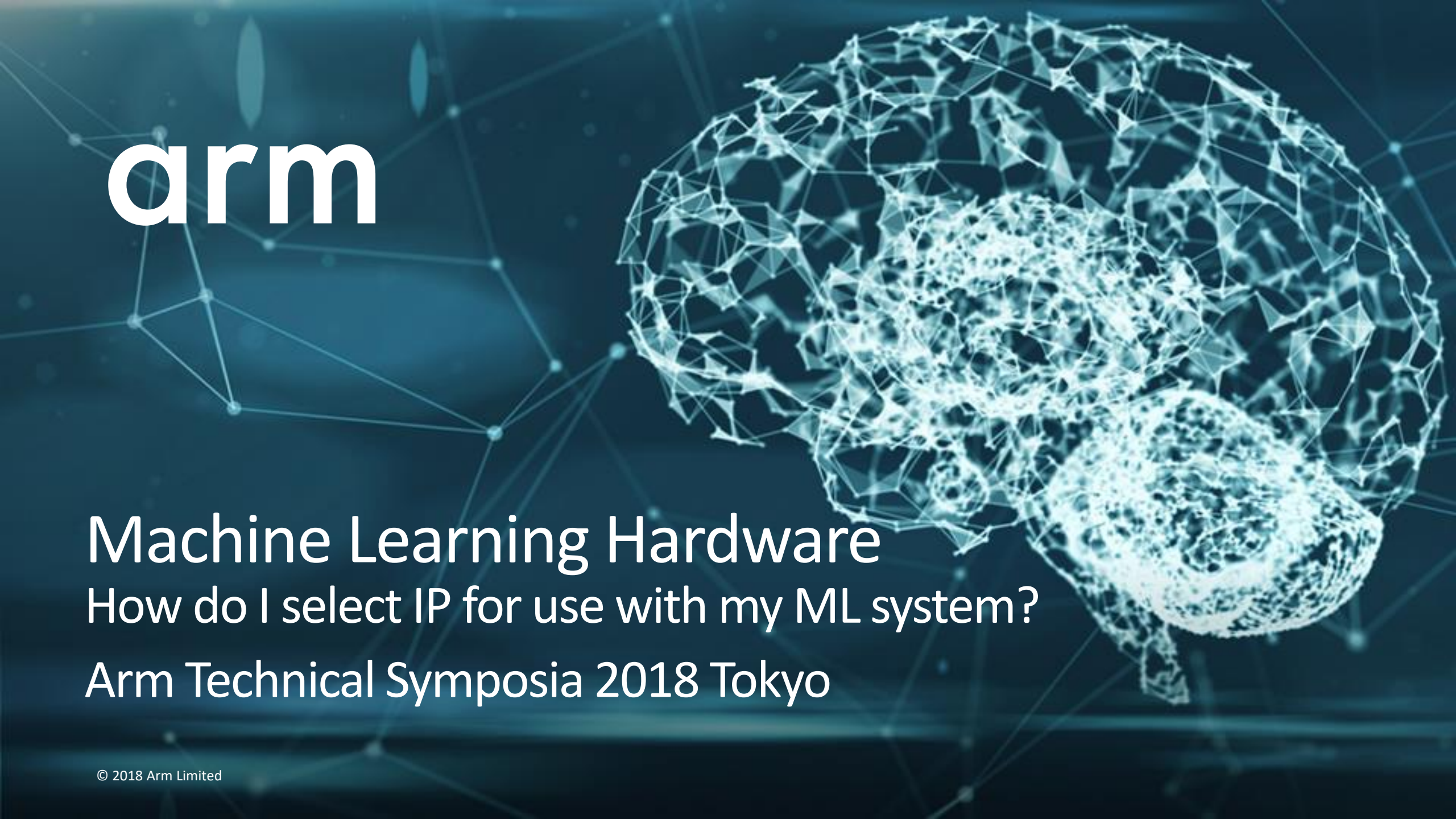


Machine Learning Hardware

How do I select IP for use with my ML system?

Arm Technical Symposia 2018 Tokyo

Machine Learning is Being Used Across Many Industries

IoT and Embedded

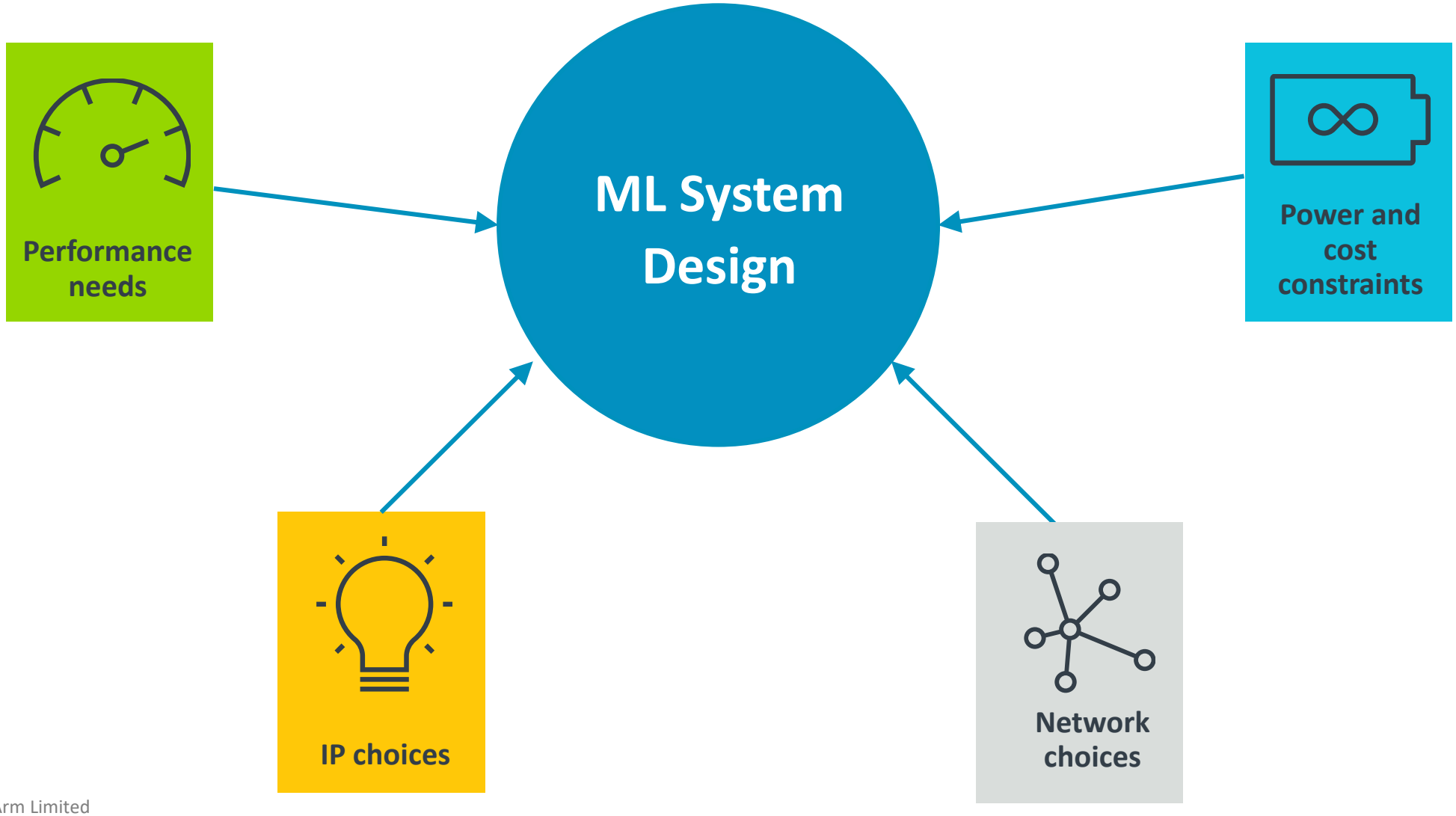
Mobile and Consumer

Automotive

Networking and Servers

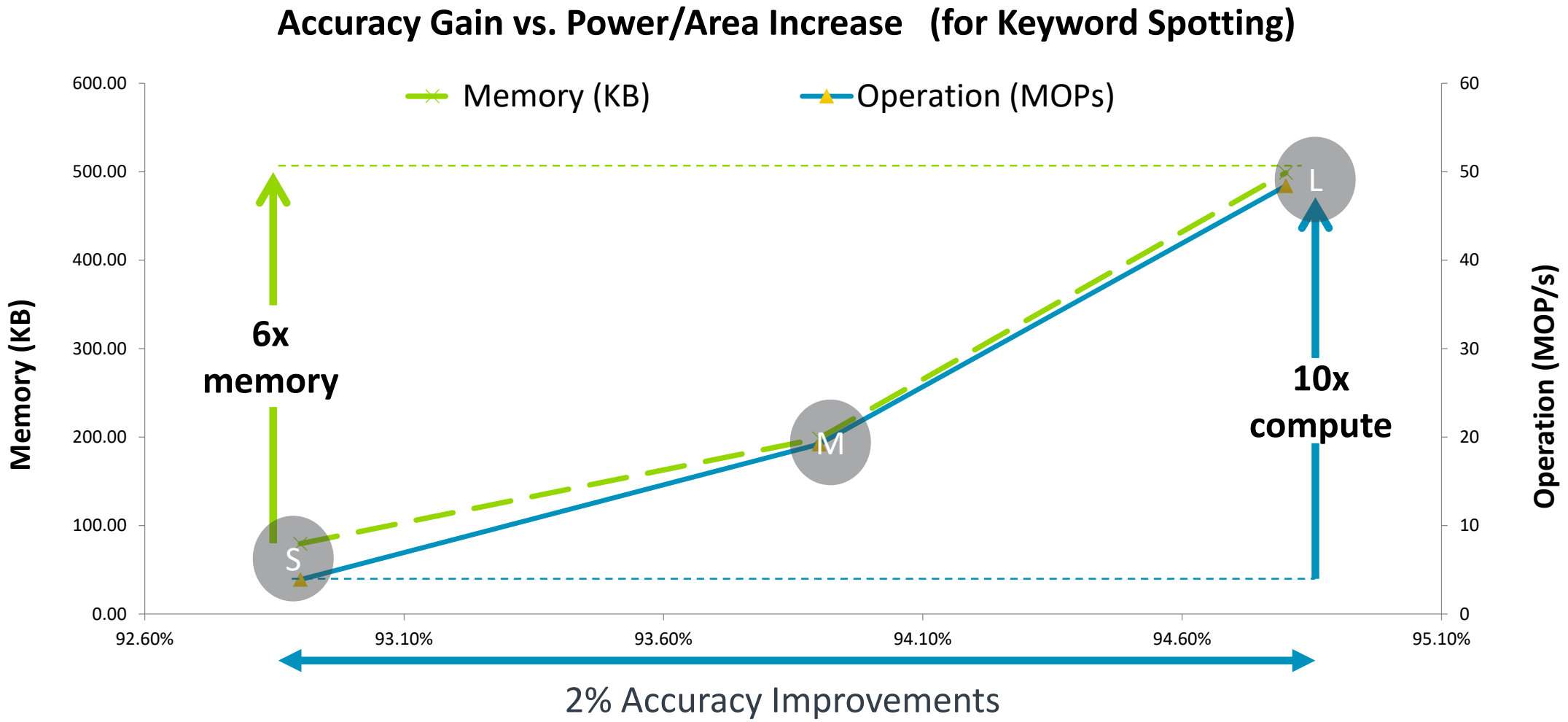
arm

The Challenge with Balancing Product Decisions



arm

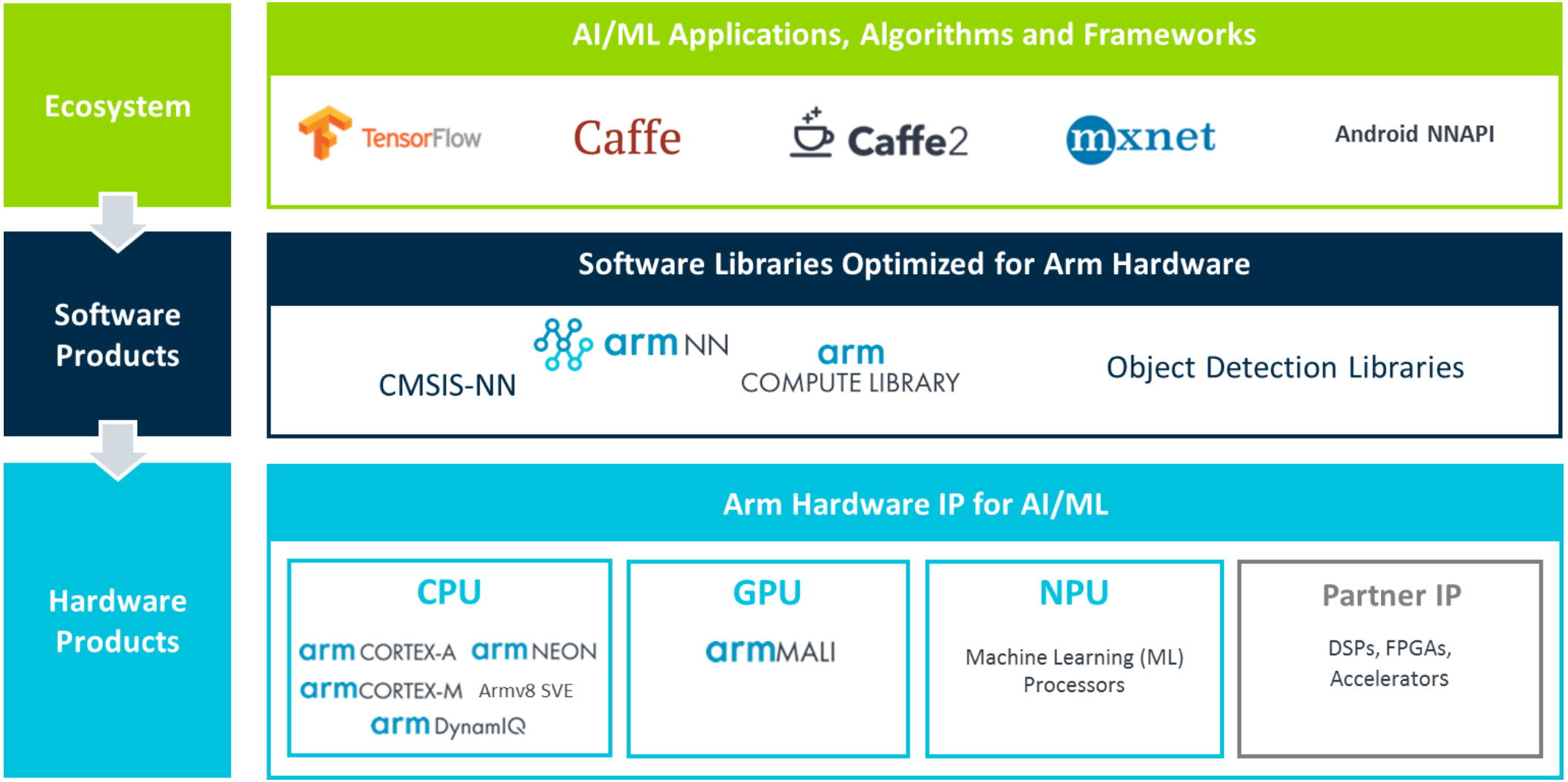
Choosing Realistic Accuracy Targets



A mild accuracy improvement could result in high increase of compute and area requirements

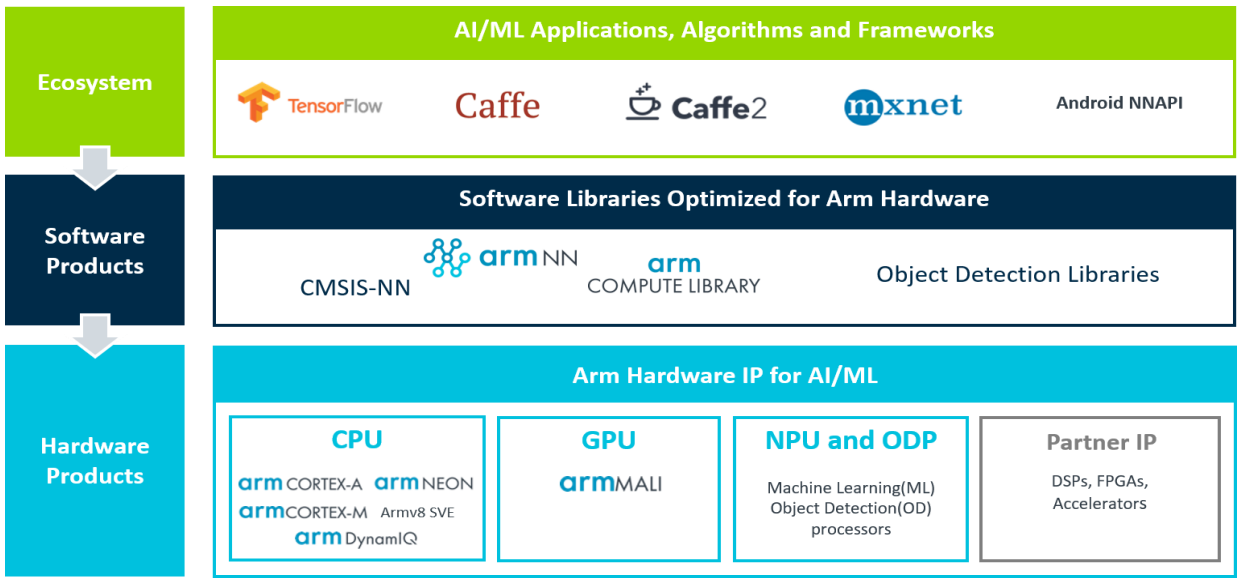
arm

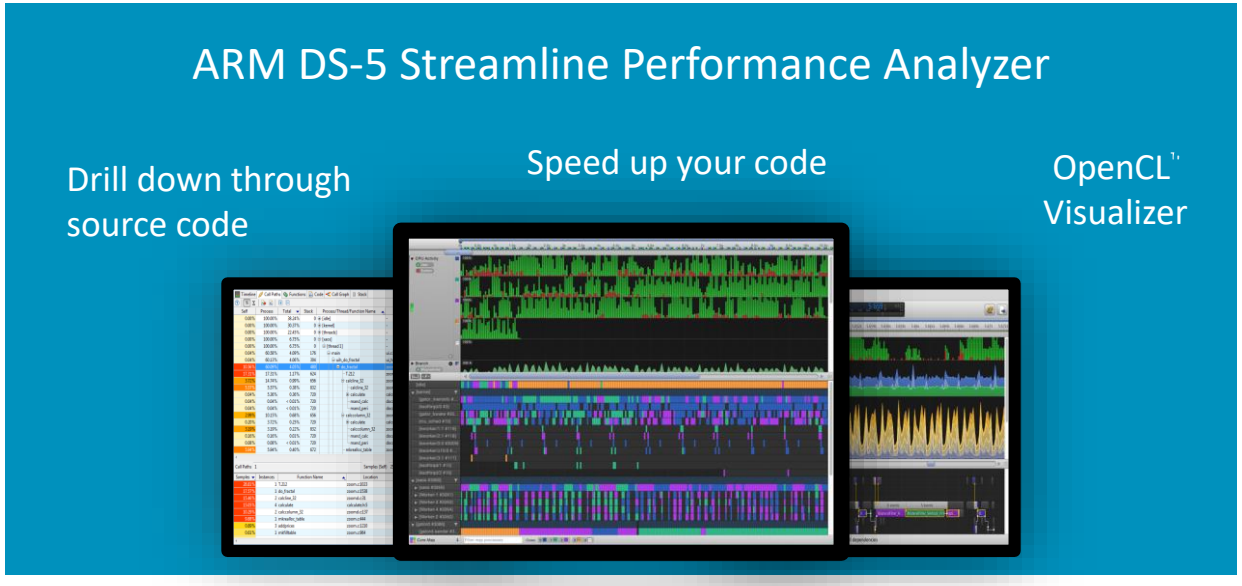
Project Trillium: Arm's ML Computing Platform



arm

Switching from IP to IP with Ease



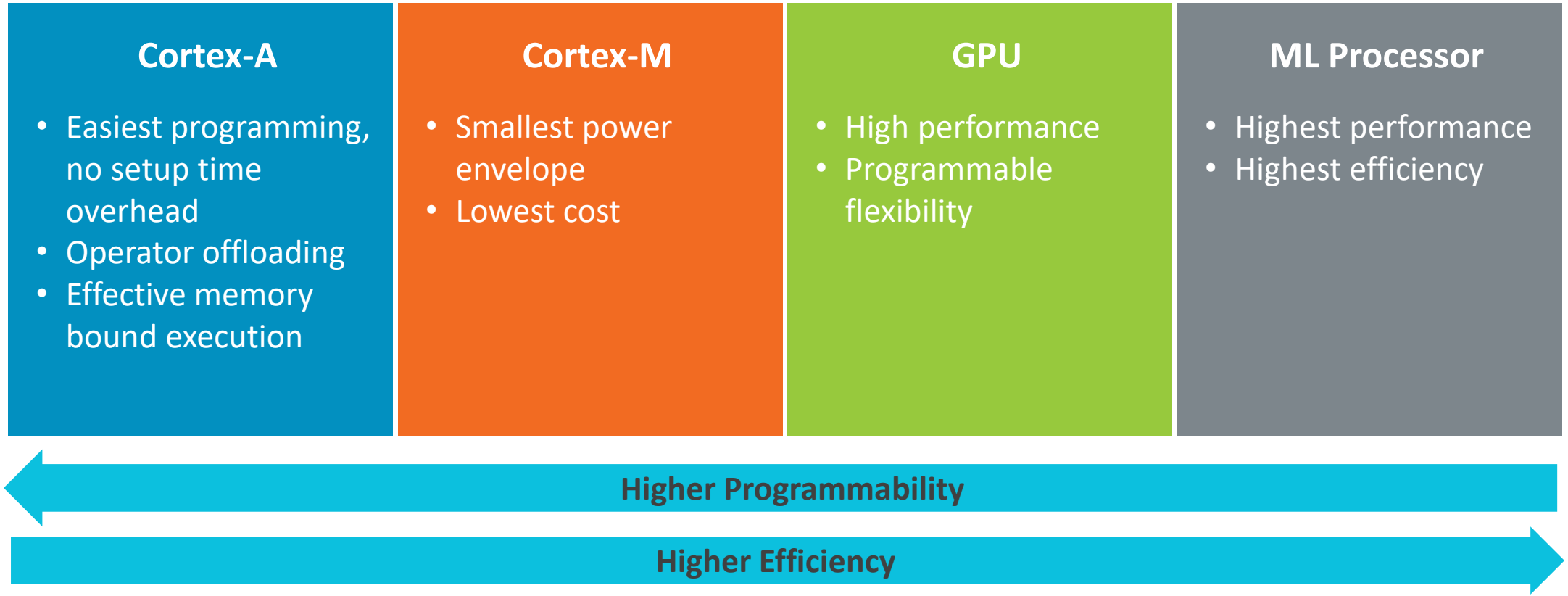


- Arm NN: Hides hardware complexity from the application
- Compute Library, CMSIS-NN: Targeted performance optimization for each processor
- ARM DS-5: Visualized heterogenous view of CPU, GPU, and ML Processors

arm

Heterogeneous Compute

Maximize the benefits from all IP families

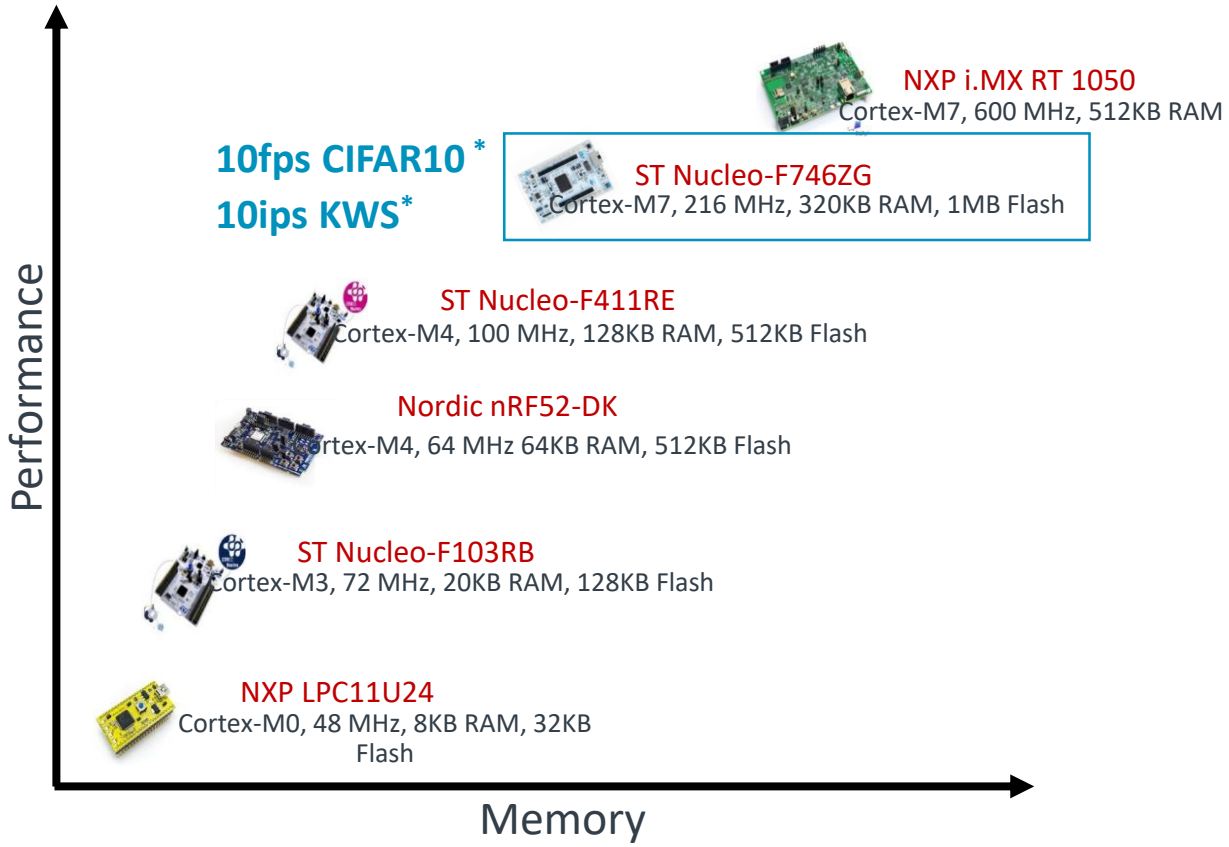


A platform built on heterogeneous compute provides the flexibility needed to match PPA across a wide range of use cases, workloads and market segments

arm

Cortex-M Microcontrollers

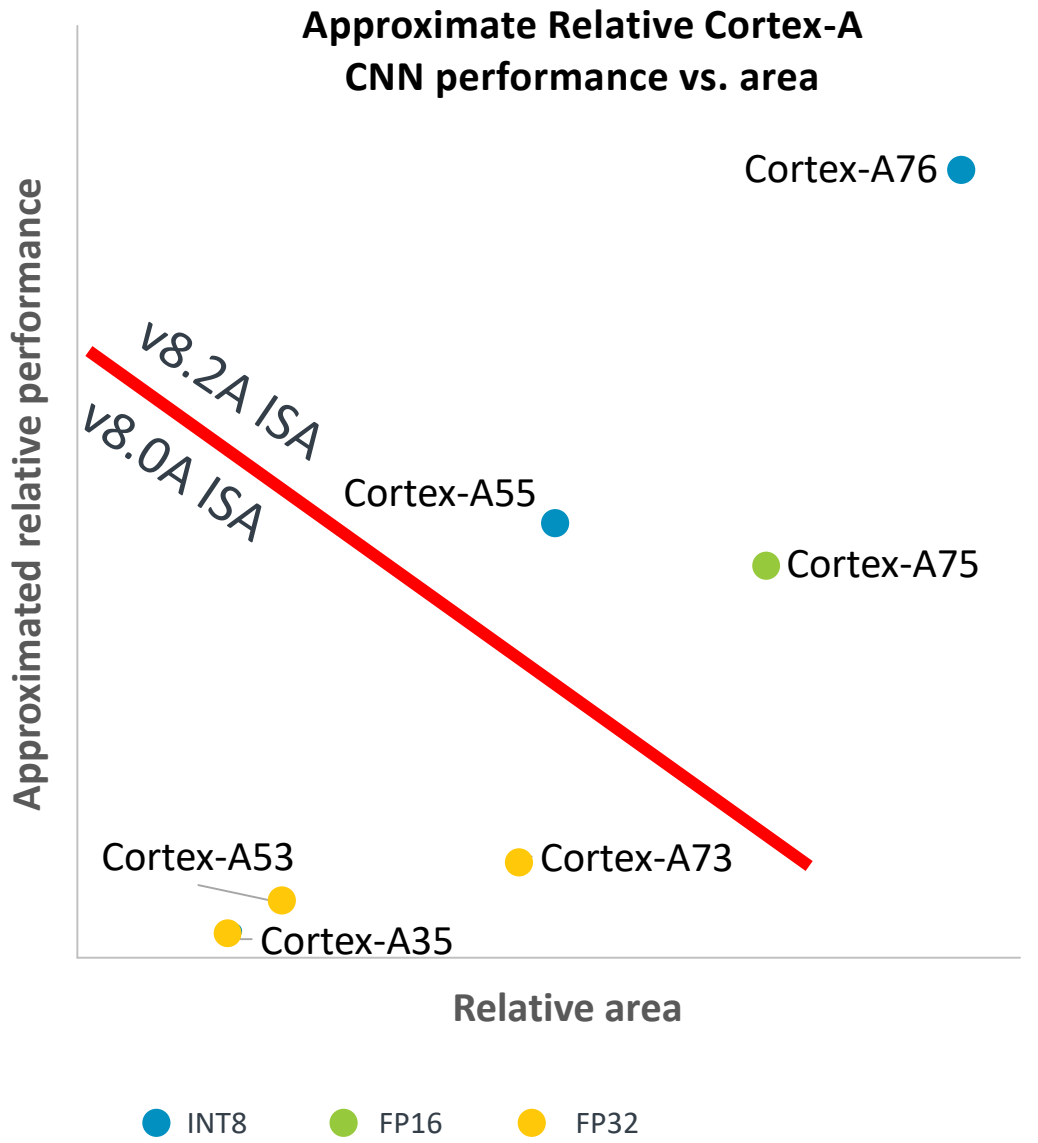
- Widely available in embedded hardware
  - Fully programable
  - Extreme low power and small area
    - ~0.1mm2, ~10mW in 16FF
  - ML speech and image recognition
- Software support
  - CMSIS-NN, CMSIS-DSP
  - Tuned ML functions
  - General purpose DSP functions



arm

Cortex-A CPUs

- In all chips needing general programmability
  - Embedded, mobile, automotive, infrastructure
  - SIMD, SVE and better memory system
  - Fallback for future operators
- Software platform support
  - Portable across platforms
  - Arm NN, Compute Library
  - Hand-tuned code for individual CPUs
  - Quarterly release with new features and better performance



All are MP4 configurations, except for Cortex-A55 which is MP8

arm

Mali GPUs

Available in a range of devices

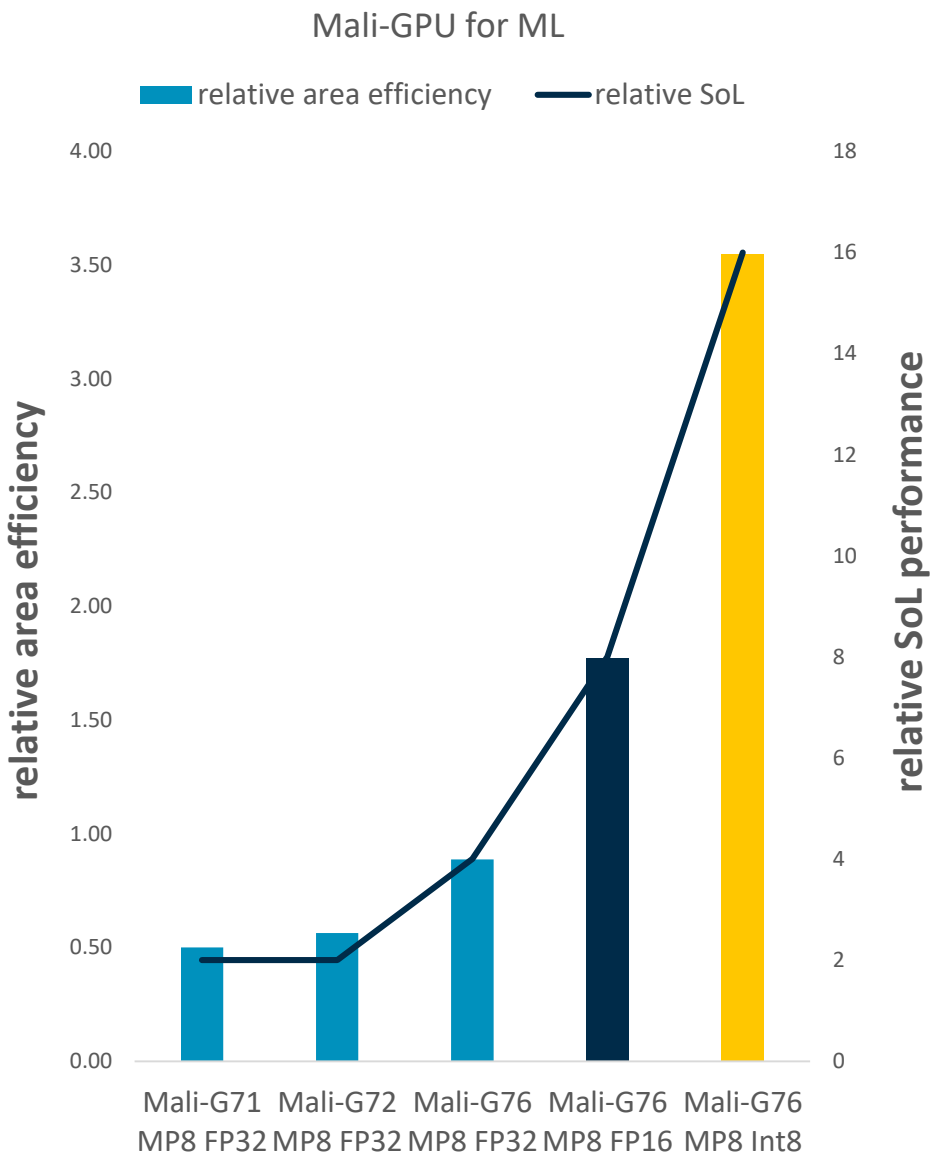
- Mobile phones, DTV, surveillance cameras, automotive IVI etc

Highly aggregated performance

- Family of GPUs for efficiency and performance
- Redesigning execution and compression units
- 4x SoL MAC performance in Mali-G76
- Reaching TOP/s performance in large configurations

Software support

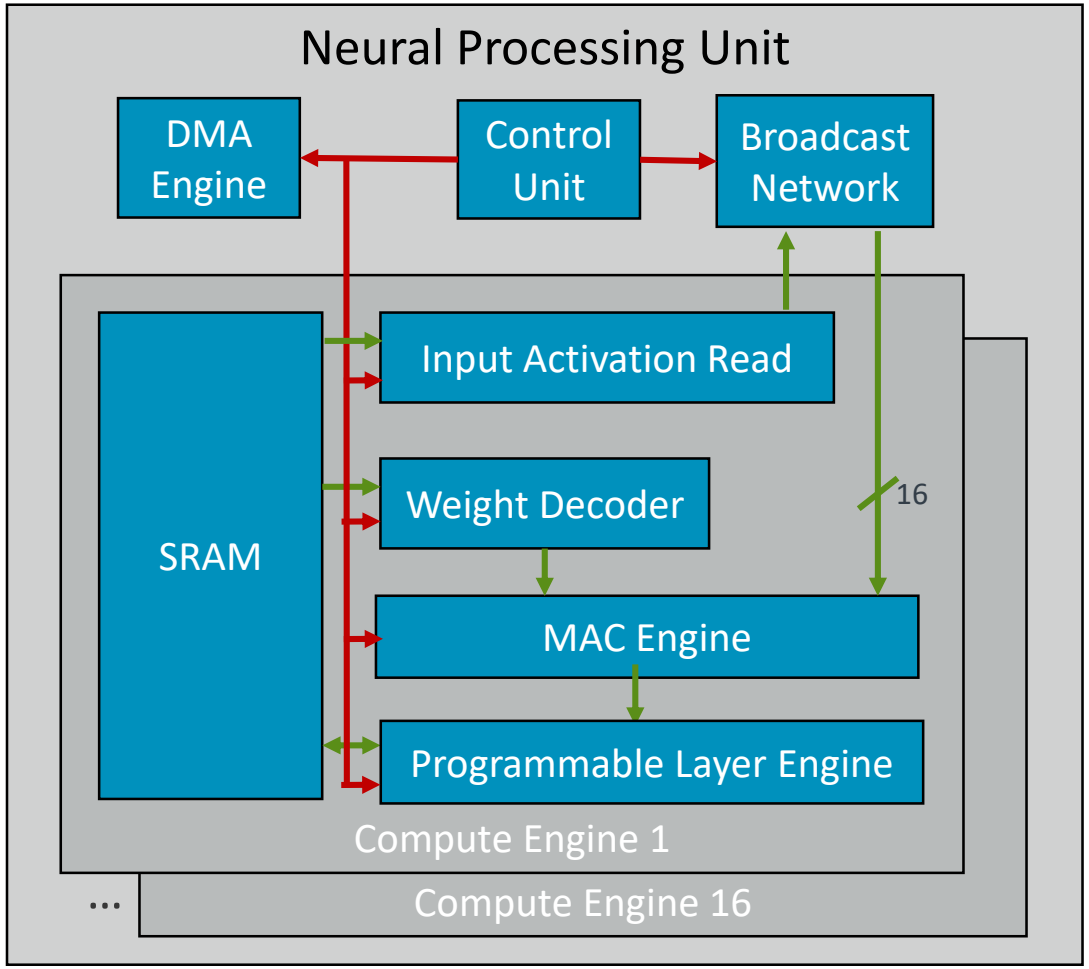
- Fully programmable
- Arm NN, Compute Library



arm

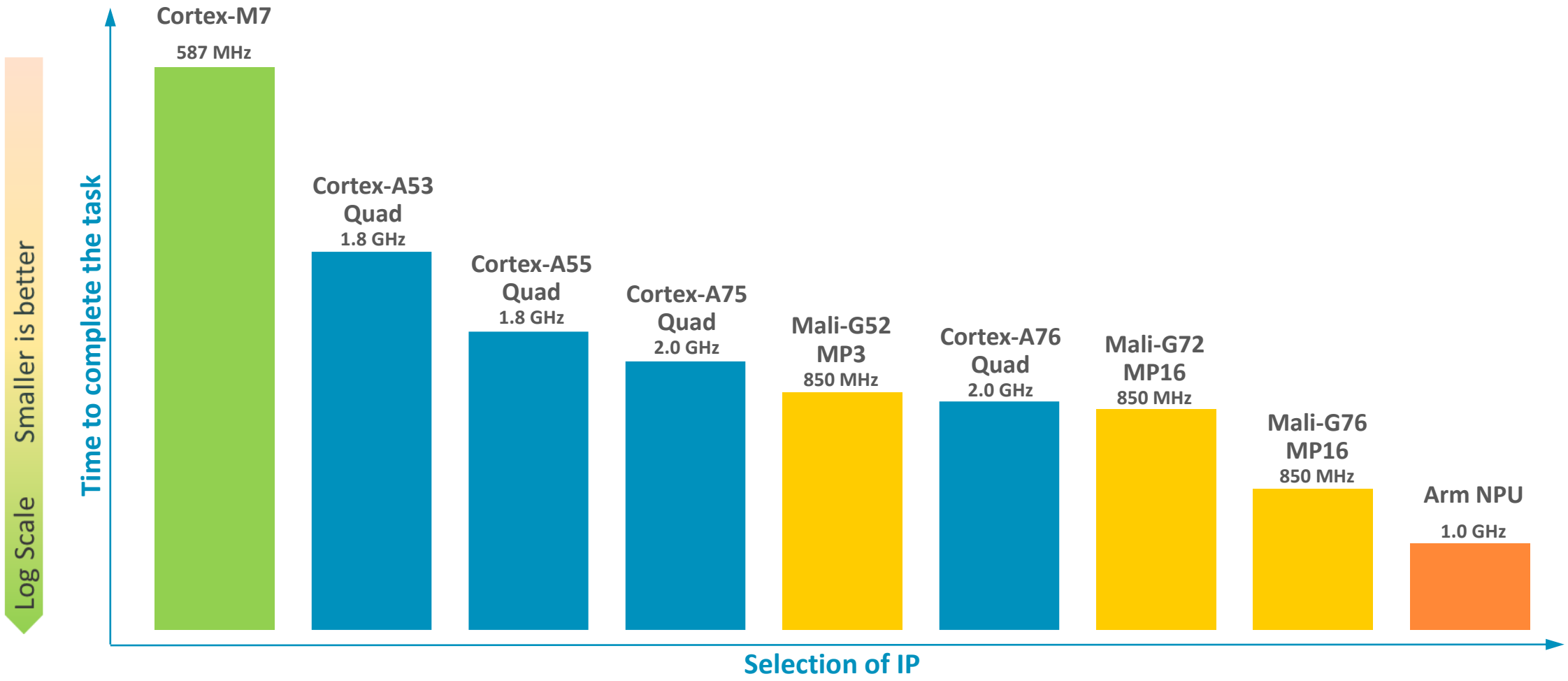
Neural Processing Units (NPUs)

- Highest performance and efficiency
  - Scalable with family of ML processors
- Programmability for futureproofing
  - Based on Arm microcontroller technology with tool support
  - Operators can be added after tape out
  - Encompassing a range of data types in the product line
- Supported by Arm NN and Arm Compute Library



arm

Example: IP Selection for Face Unlock (16FF)



Pick the right solution based on performance, availability, area and energy requirements

arm

Summary

- Arm offers a choice of ML solutions across many markets and use cases
- Arm helps our Partners to make informed choices, Cortex-M/A/Mali-GPU and Arm NPU
- It is a continuing process for both HW and SW

arm

Thank You!
Danke!
Merci!
谢谢!
ありがとう!
Gracias!
Kiitos!
감사합니다
धन्यवाद