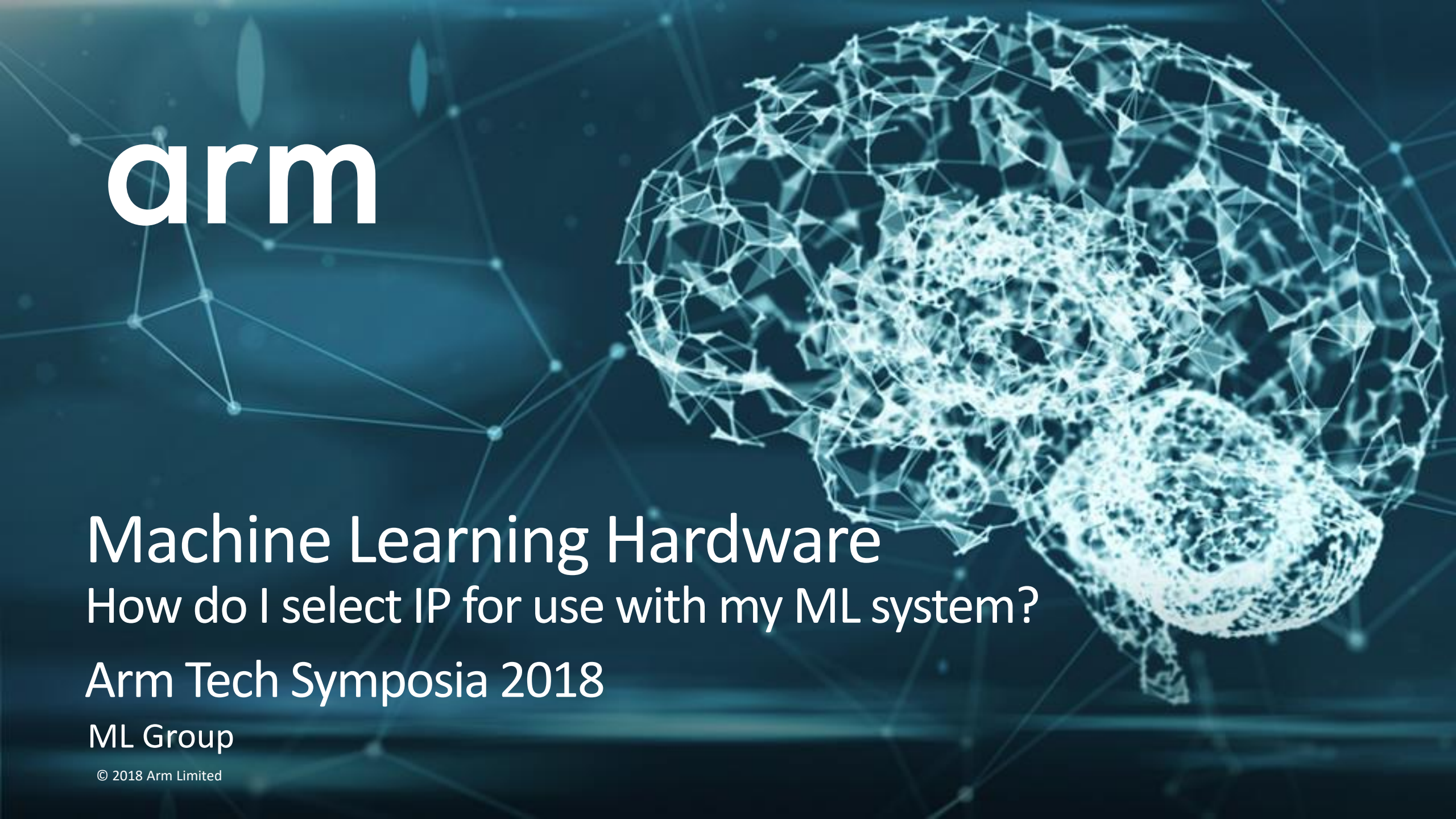


Machine Learning Hardware

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

Arm Tech Symposia 2018

ML Group

Machine Learning is Being Used Across Many Industries

IoT and Embedded

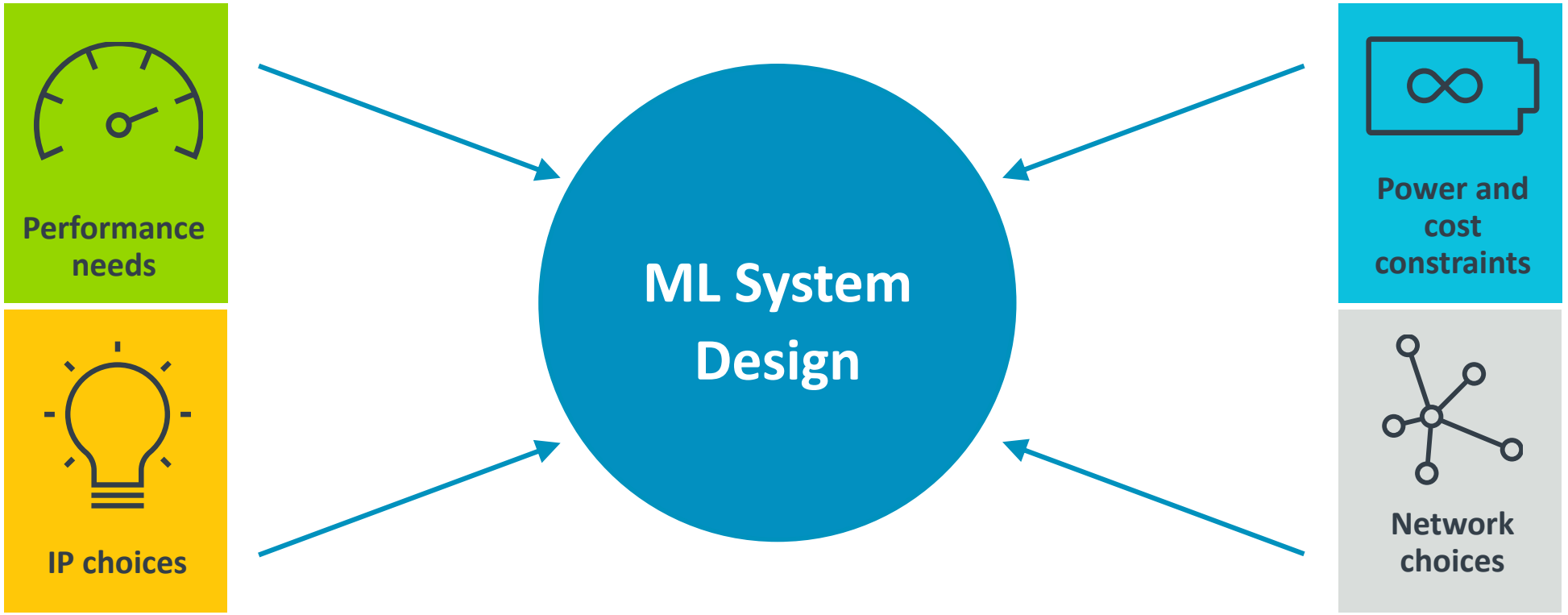
Mobile and Consumer

Automotive

Networking and Servers

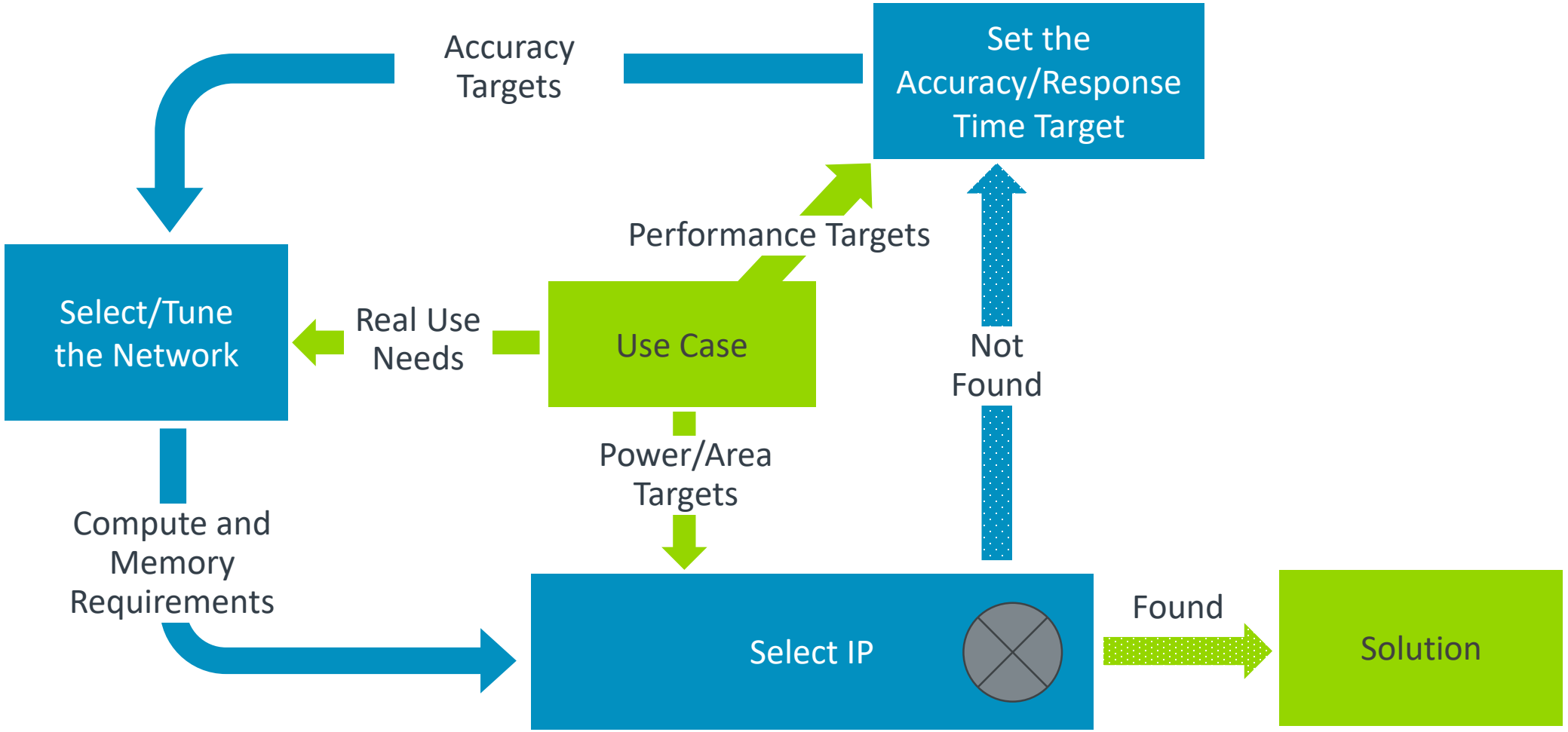
arm

The Challenge with Balancing Product Decisions



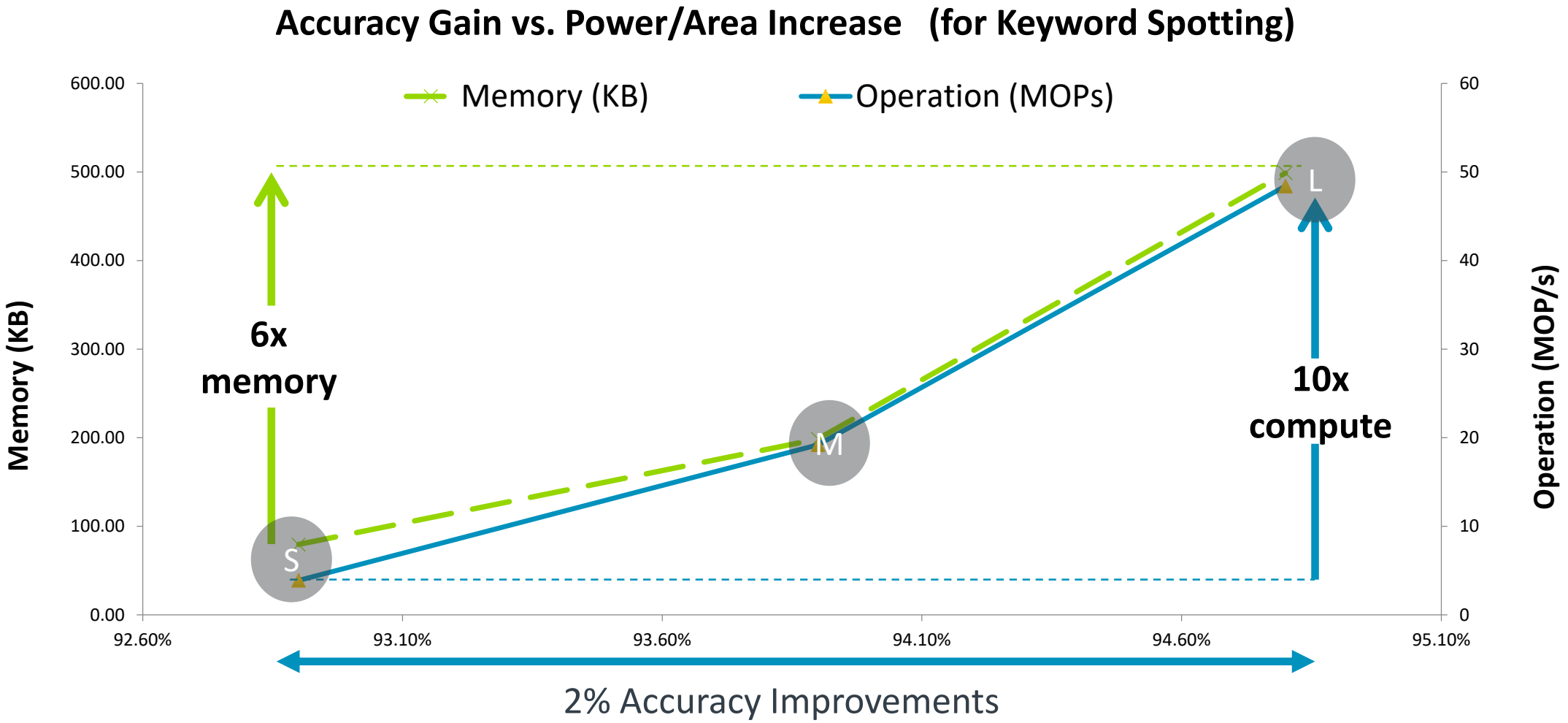
arm

IP Selection Considerations



arm

Choosing Realistic Accuracy Targets

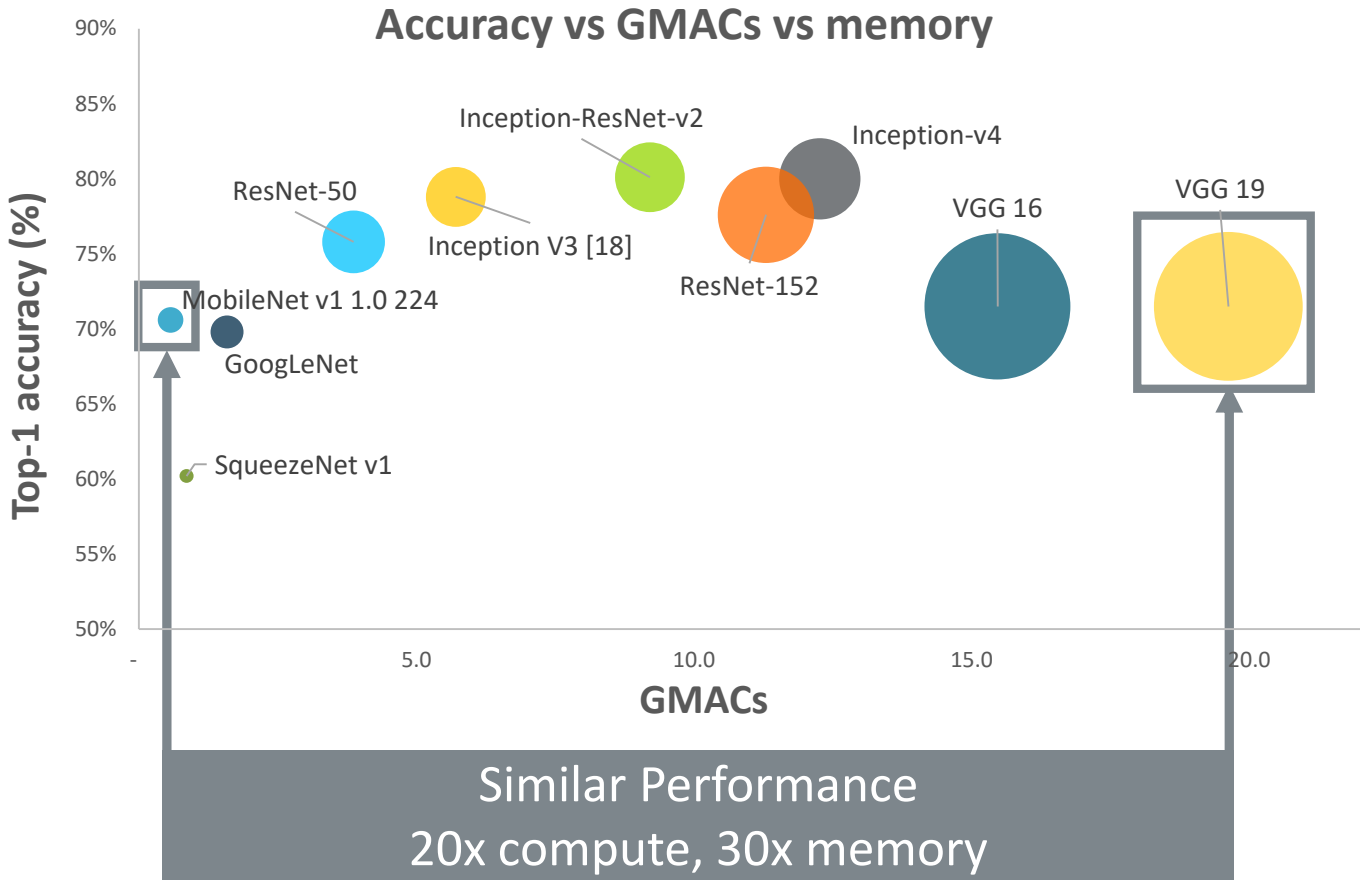


A mild accuracy improvement could result in high increase of compute and area requirements

arm

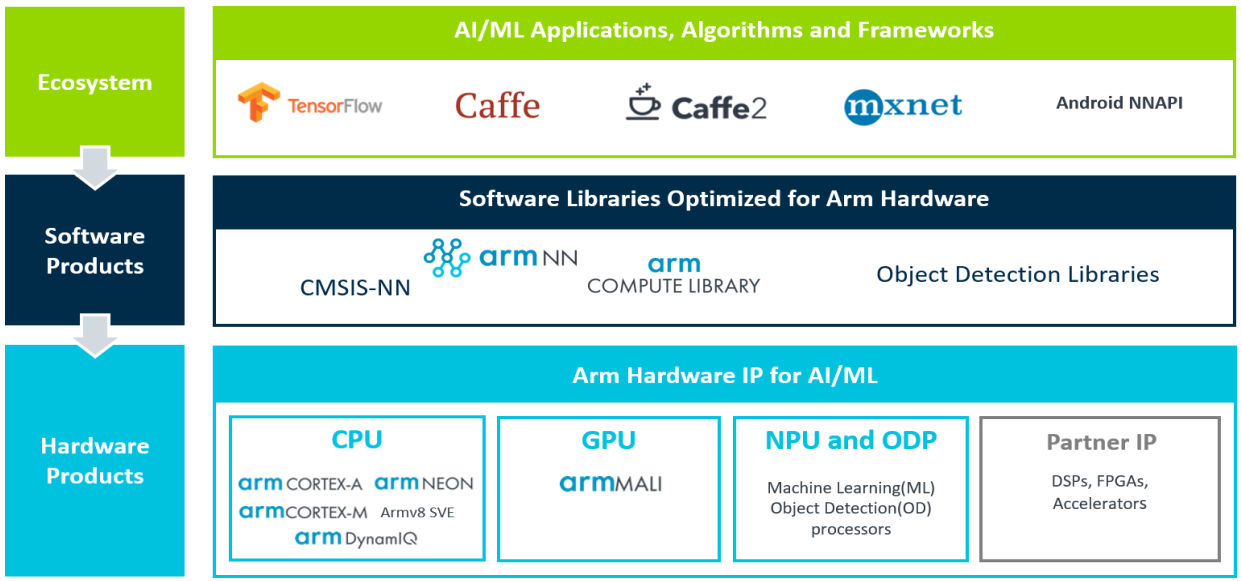
Choosing the Optimal Neural Network

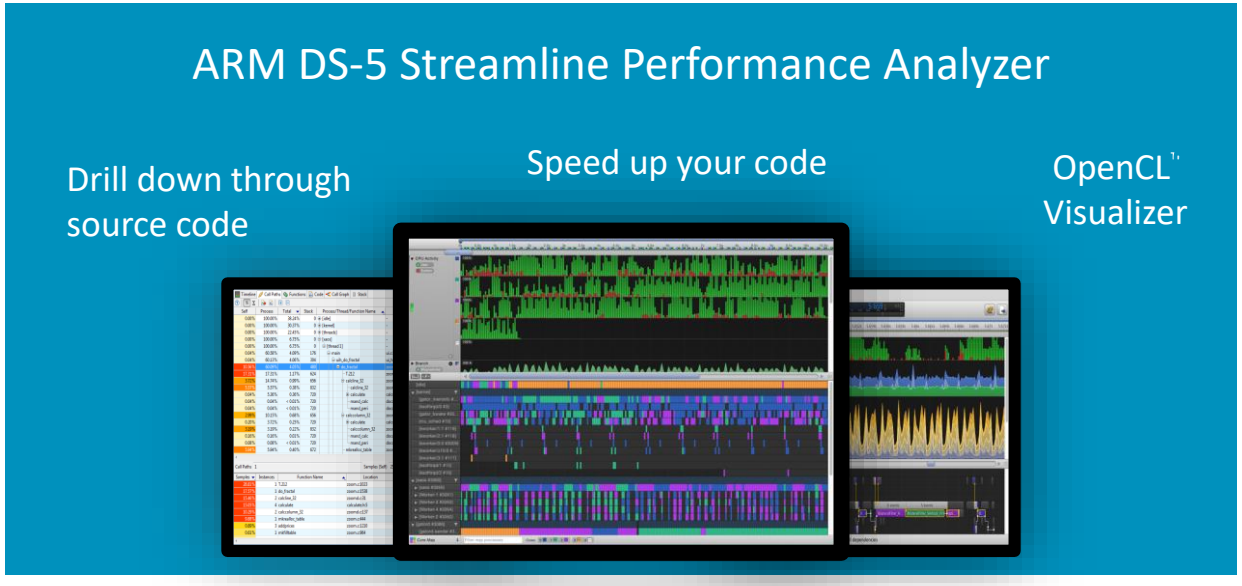
- To achieve similar levels of accuracy
  - 10x or more difference in compute
  - 10x or more difference in parameters
- Network preferences on hardware
  - Some algorithms are more effective in CPU/GPU
  - Some algorithms are better accelerated by specialized ML processors



arm

Switching from IP to IP with Ease



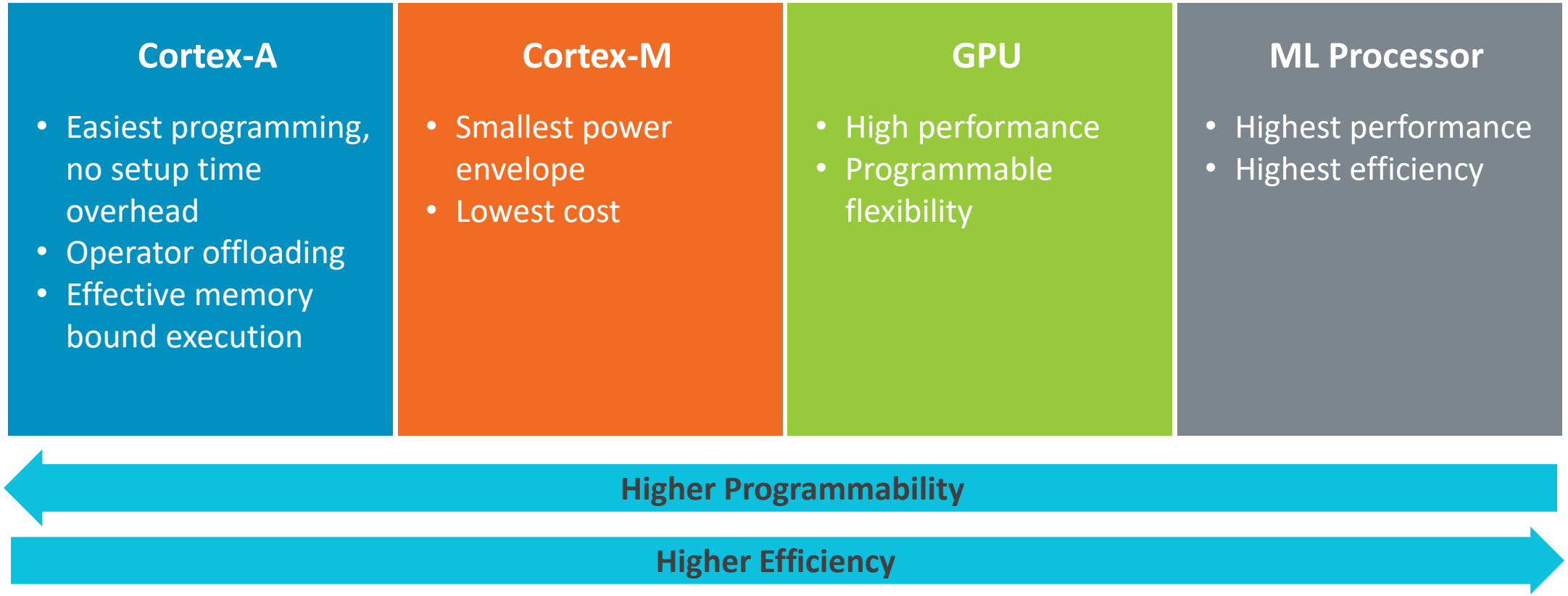


- Arm NN: Hides hardware complexity from the application
- Compute Library, CMSIS-NN: Targeted performance optimization for each processor
- ARM DS-5: Visualized heterogenous view of CPU, GPU, and ML Processors
  - Full compatibility with Arm NN and Compute Library, enables network layer visibility
  - In development, coming to market soon

arm

Heterogeneous Compute

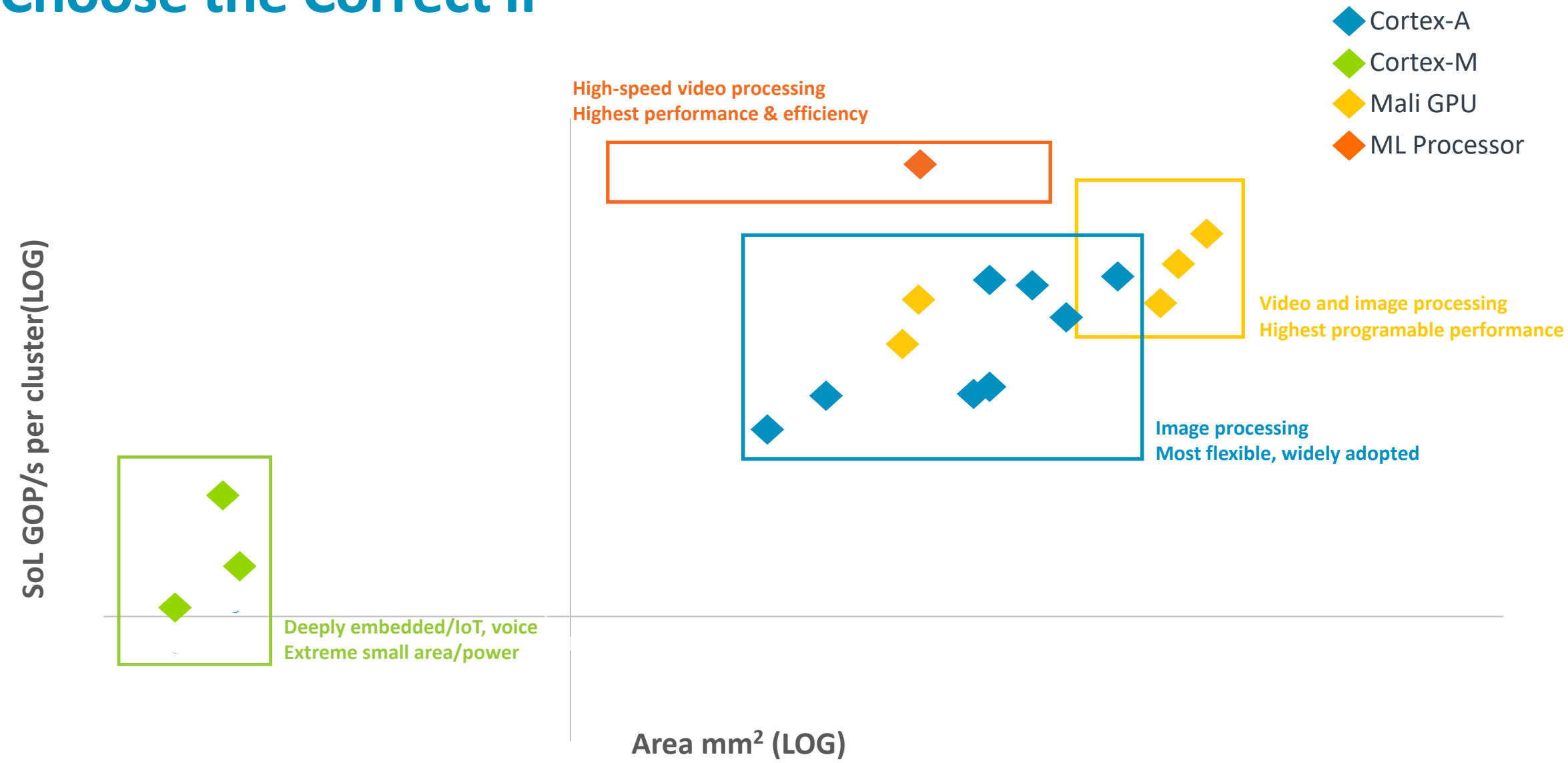
Maximize the benefits from all IP families



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

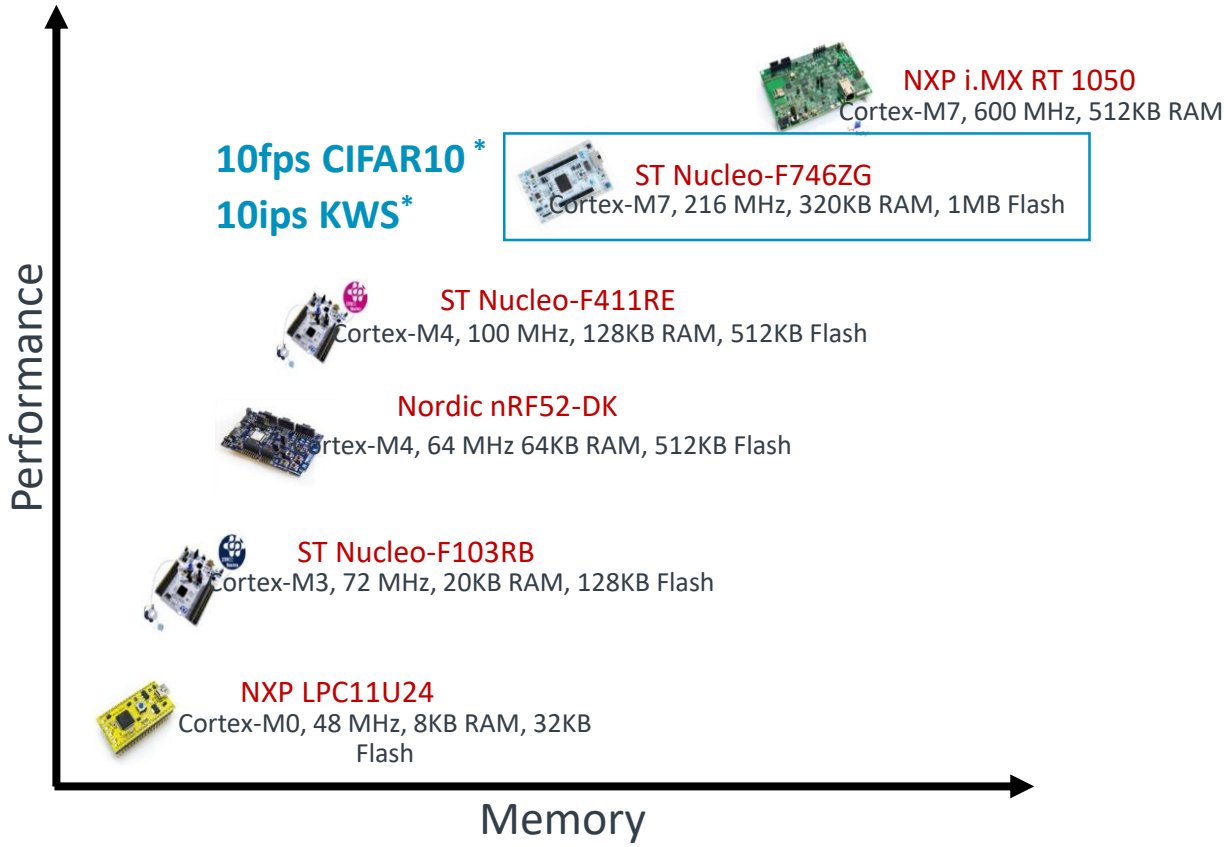
Choose the Correct IP



arm

Cortex-M Microcontrollers

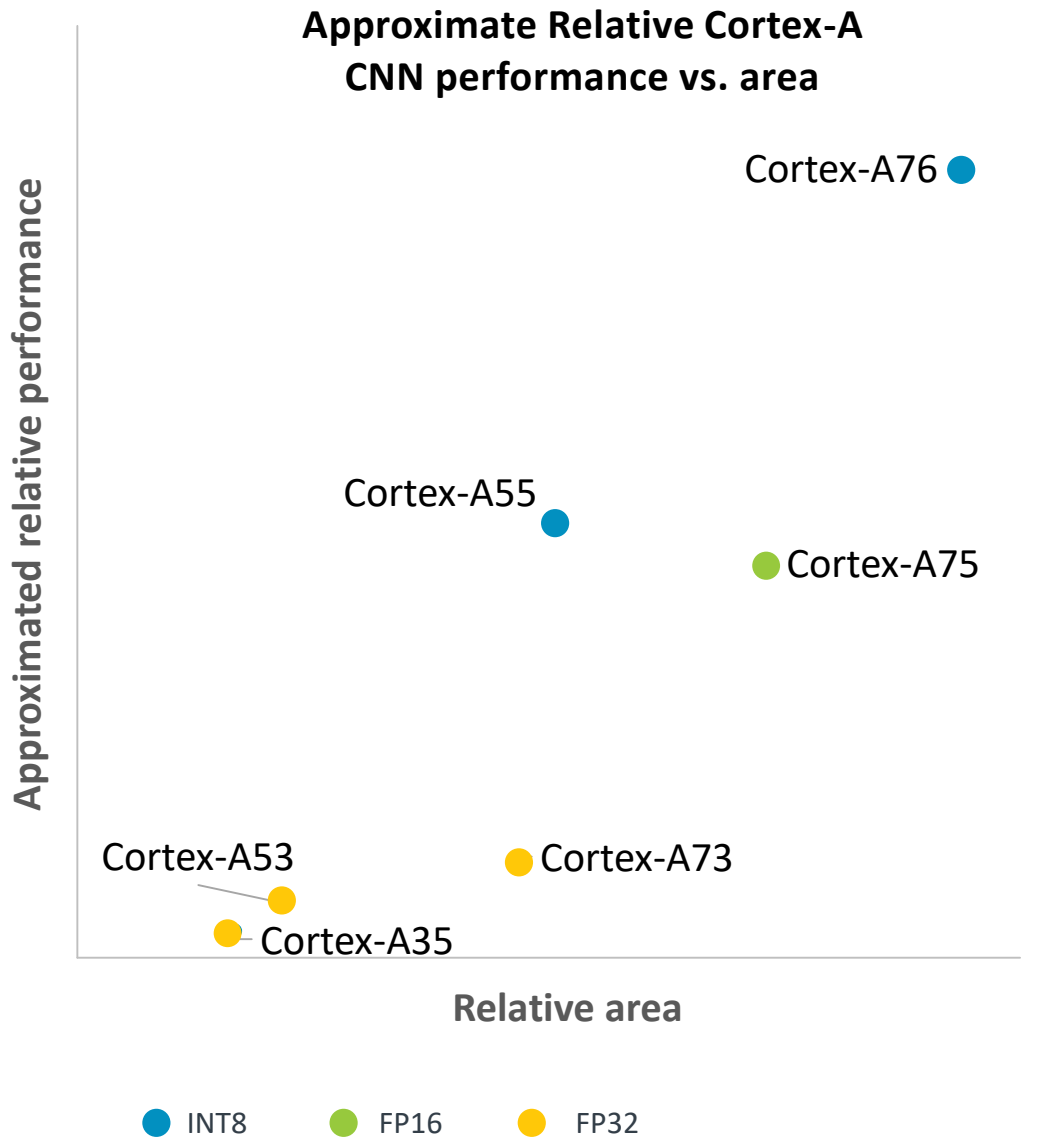
- Widely available in embedded hardware
  - Fully programable
  - Extreme low power and small area
    - ~0.1mm2, ~10mW in 16FF
  - ML speech and image recognition
- Software support
  - CMSIS-NN, CMSIS-DSP
  - Tuned ML functions
  - General purpose DSP functions



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

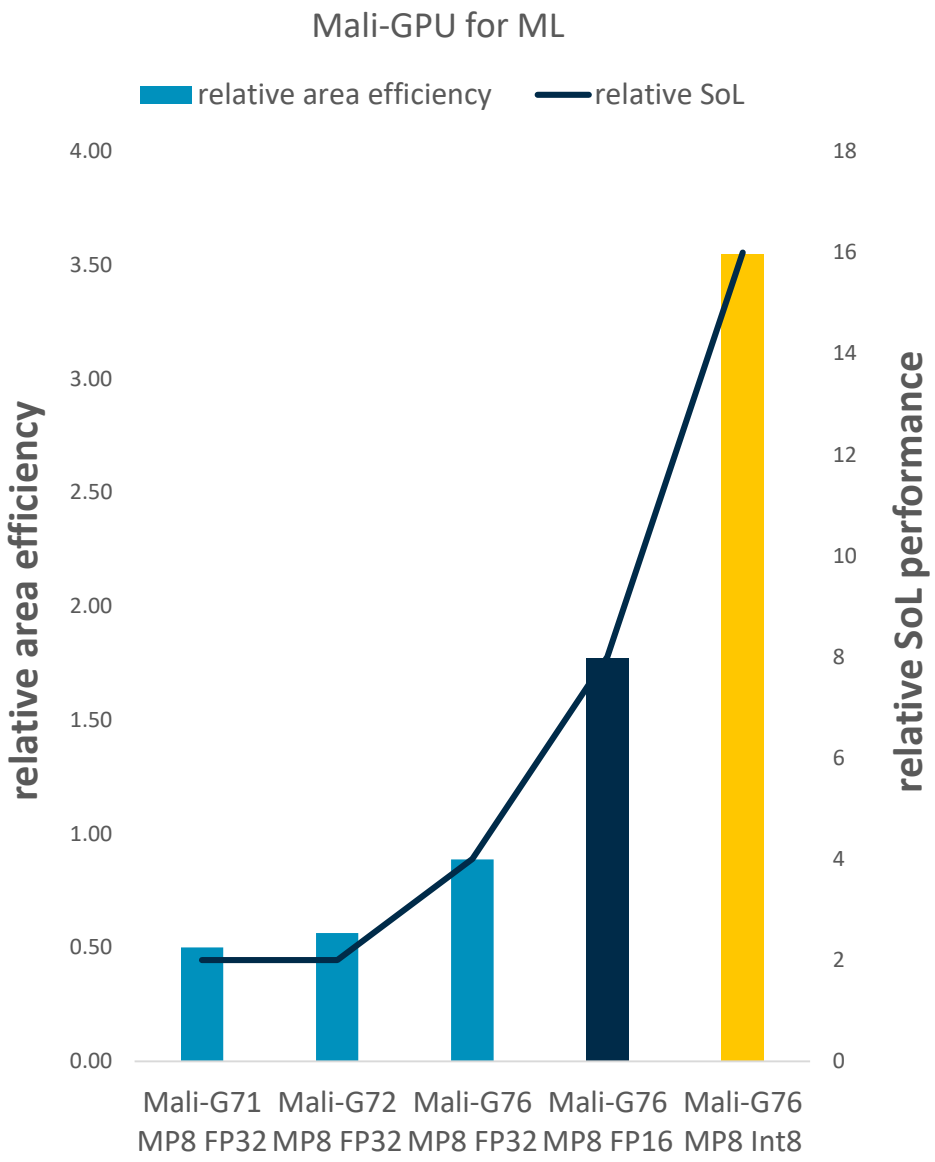
- Mobile phones, DTV, surveillance cameras, automotive IVI etc

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

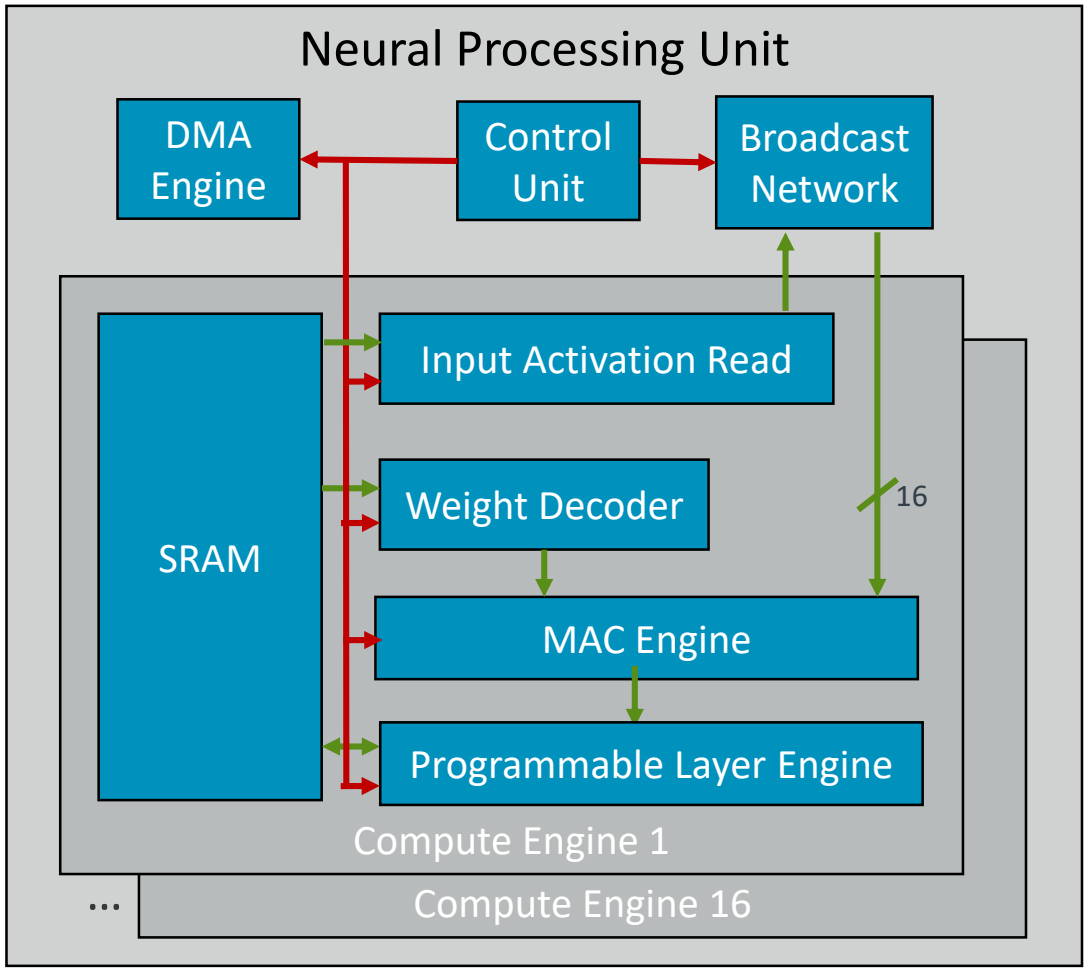
- Fully programmable
- Arm NN, Compute Library



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 the Compute Library

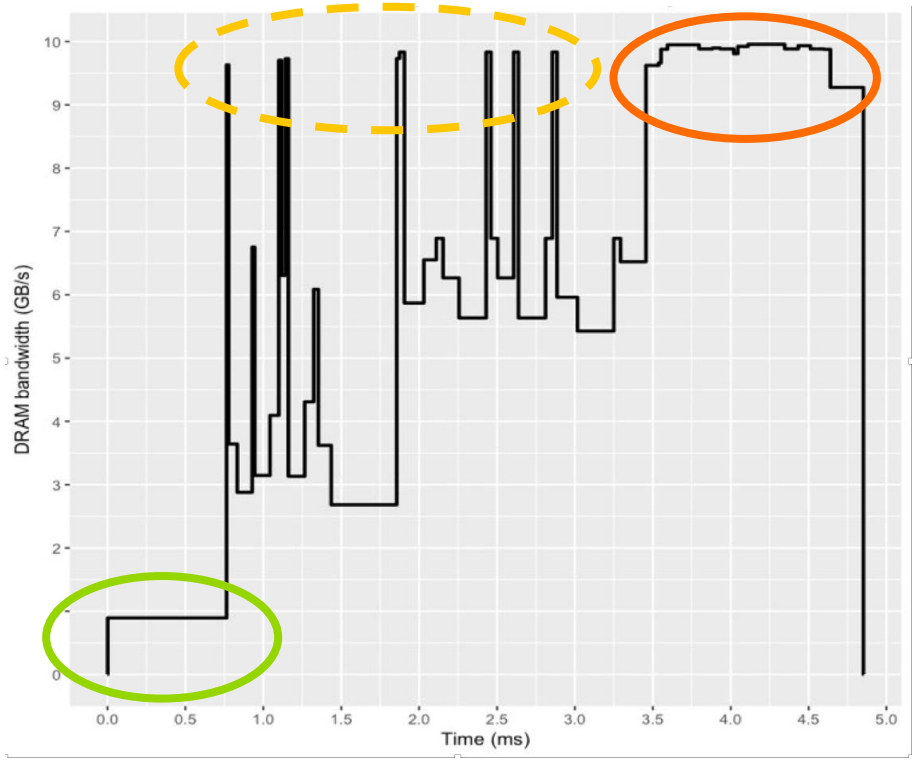


arm

System Considerations

- Memory bandwidth impacts performance
  - The memory requirement is not uniform
  - Convolution layer is compute bound
  - Pooling layer generates spike requirements
  - Fully connected layer uses the full memory bandwidth
- System impact
  - Memory impacted by other components on the SoC
- Arm is introducing NPU designs that balance the needs of compute and memory bandwidth

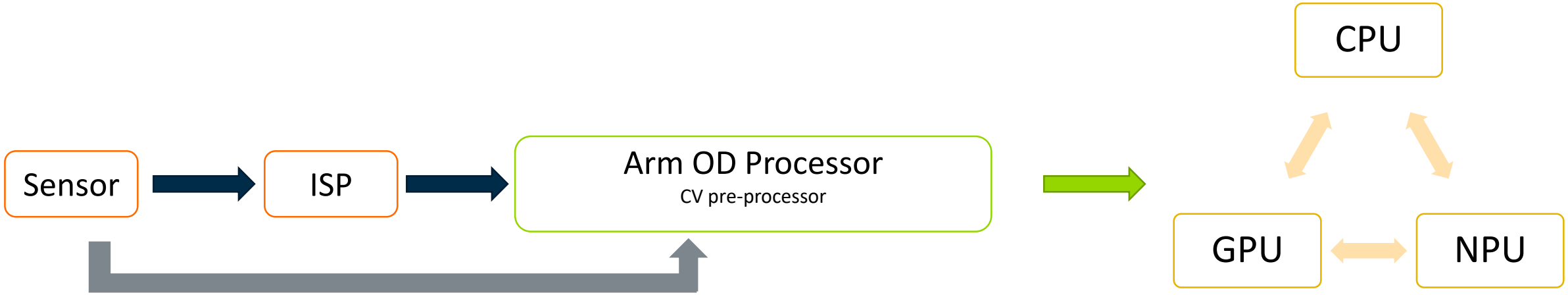
Memory requirements during inference run



Compute bound
Mixed
Memory bound

arm

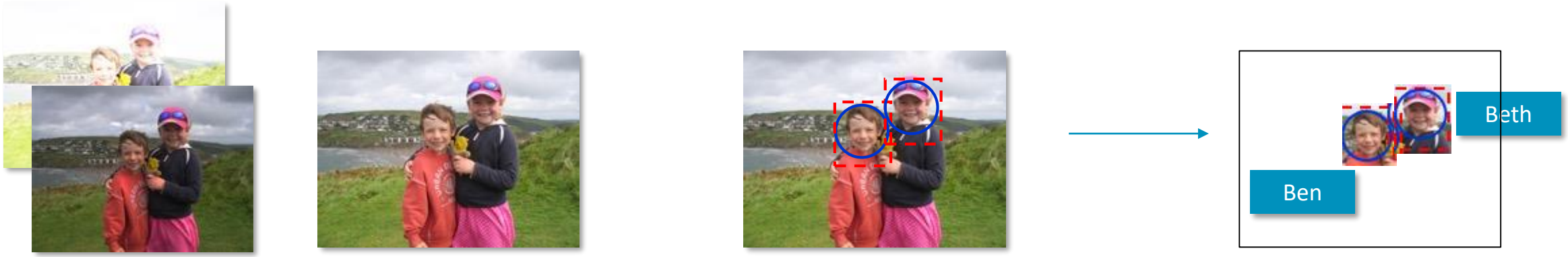
Arm OD Processor as a pre-processor



Input from Sensor
- Directly into Object Detection processor
- To an ISP and then to OD processor

People Detection
- Regions of interest calculation
- Metadata passed to CPU/GPU/NPU

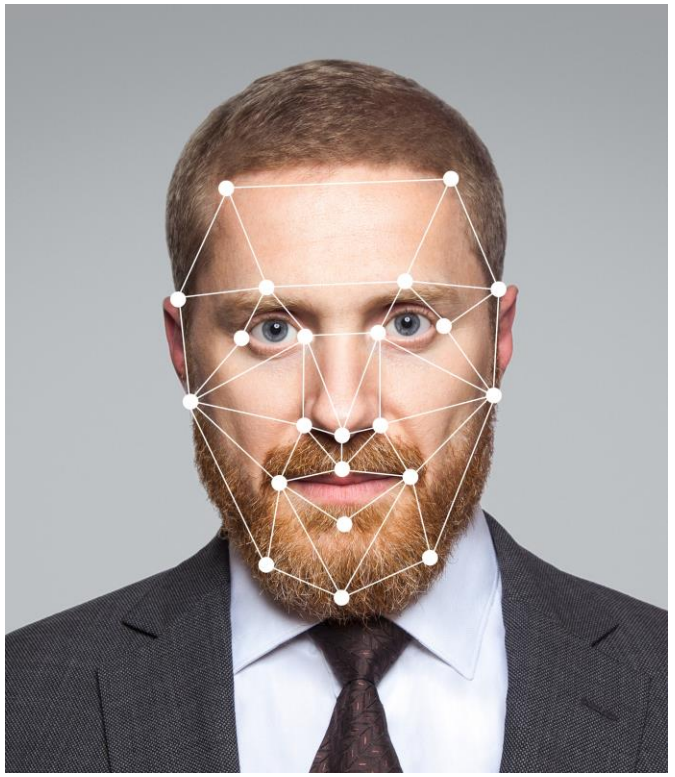
People Identification
- Tracking and other processing
- Combining with Arm NPU provides 30x more efficient solution



arm

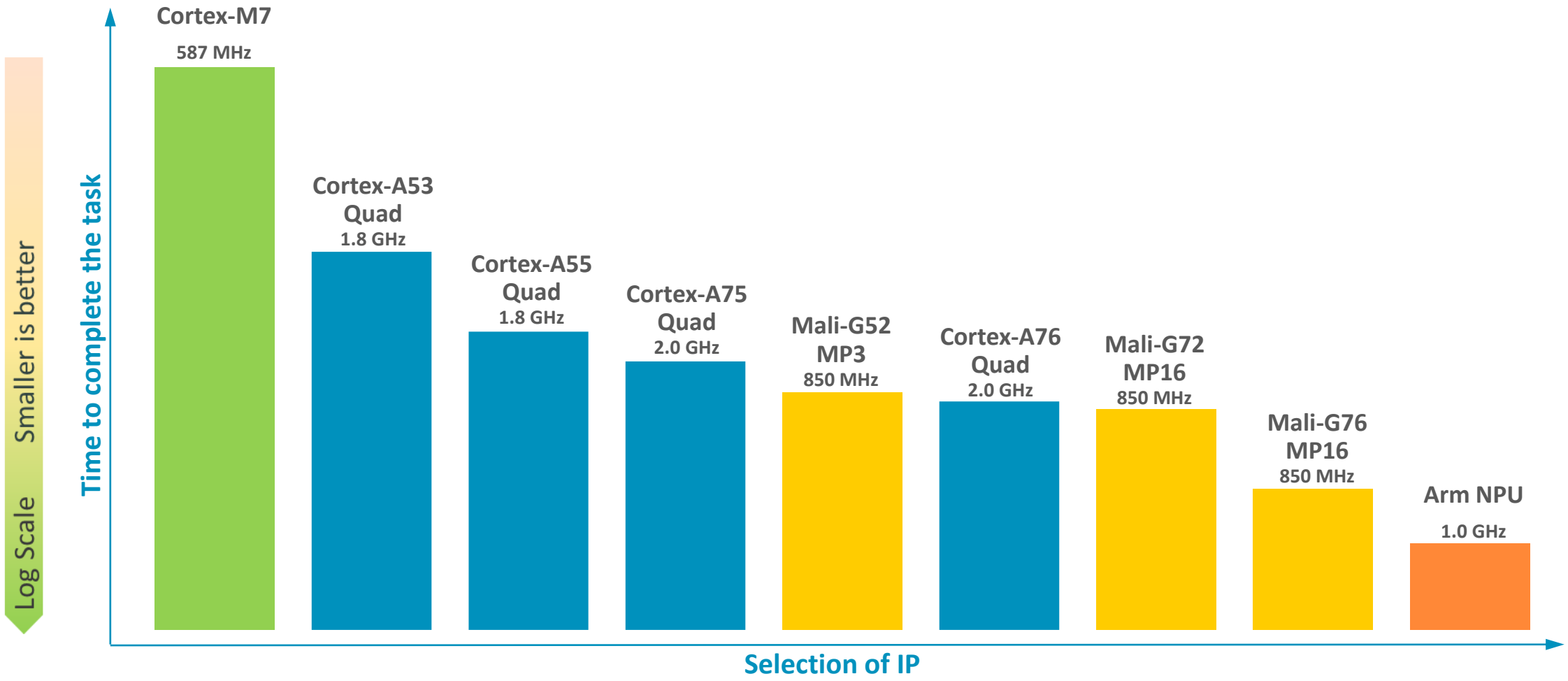
Example: IP Selection for Face Unlock

- Face unlock use case
  - Detect real faces, extract features, and filter false positive
  - Within reasonable wait time
- Feature extraction takes majority of the compute
  - 90% compute in feature extraction
  - Different networks were analyzed for compute and accuracy



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
  - Performance range from a few GOP/s to greater than 100 TOP/s
  - Power envelope from mW to 100s of W
  - Area from 0.1mm^2 to 100s of mm^2
  - Supported with Arm libraries, tool chains and continued improvements
- At Arm we help our Partners to make informed choices
  - Guidance in choosing the IP and solutions through to performance benchmarking
  - Design references with use cases such as face unlock, keyword spotting and others
- It is a continuing process
  - We are always exploring new use cases and network types to enable even more effective guidance

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