



arm

Machine Learning Hardware

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

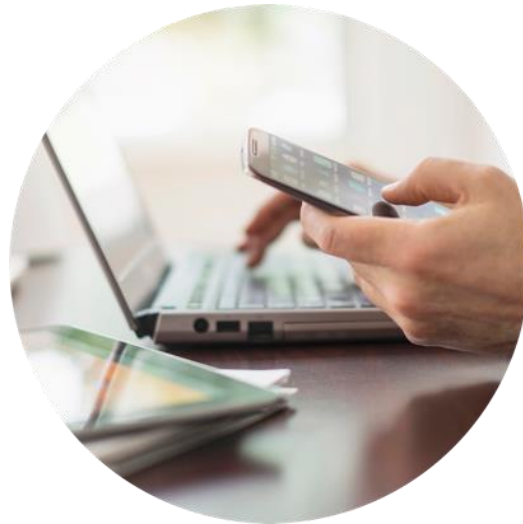
Arm Technical Symposia 2018

ML Group

Machine Learning is Being Used Across Many Industries



IoT and Embedded



Mobile and Consumer

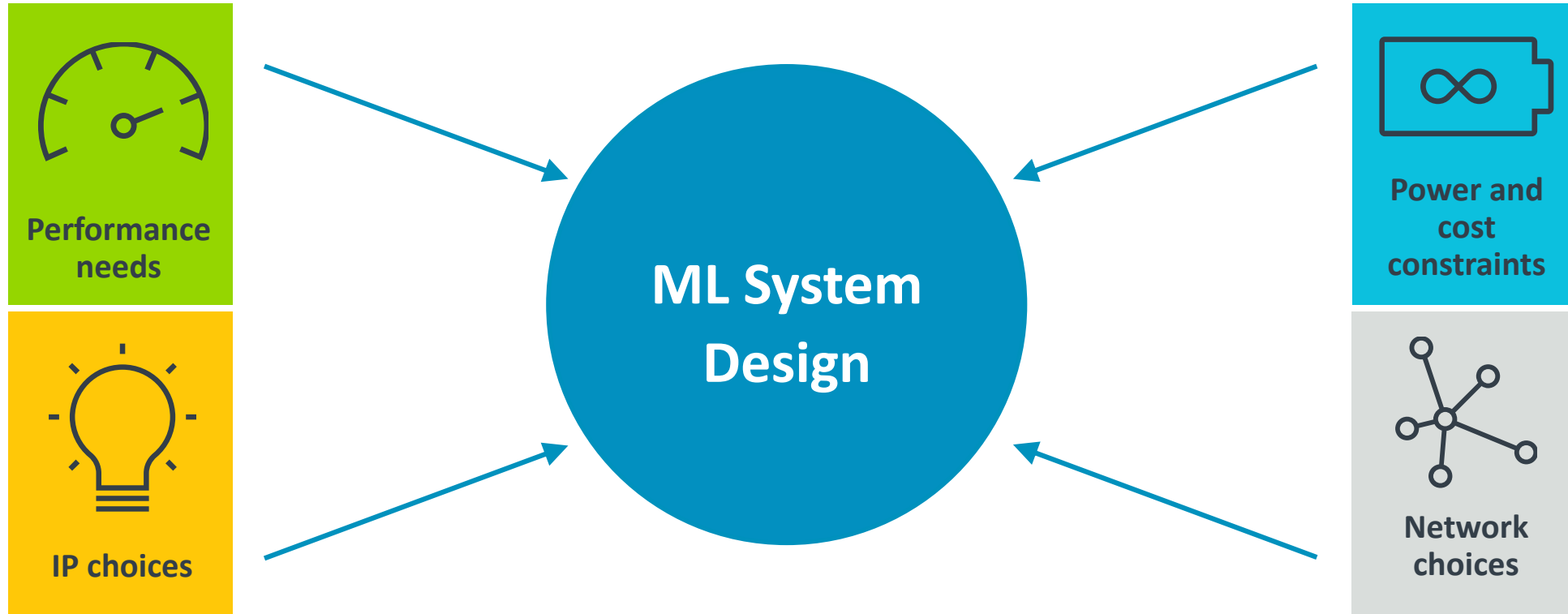


Automotive

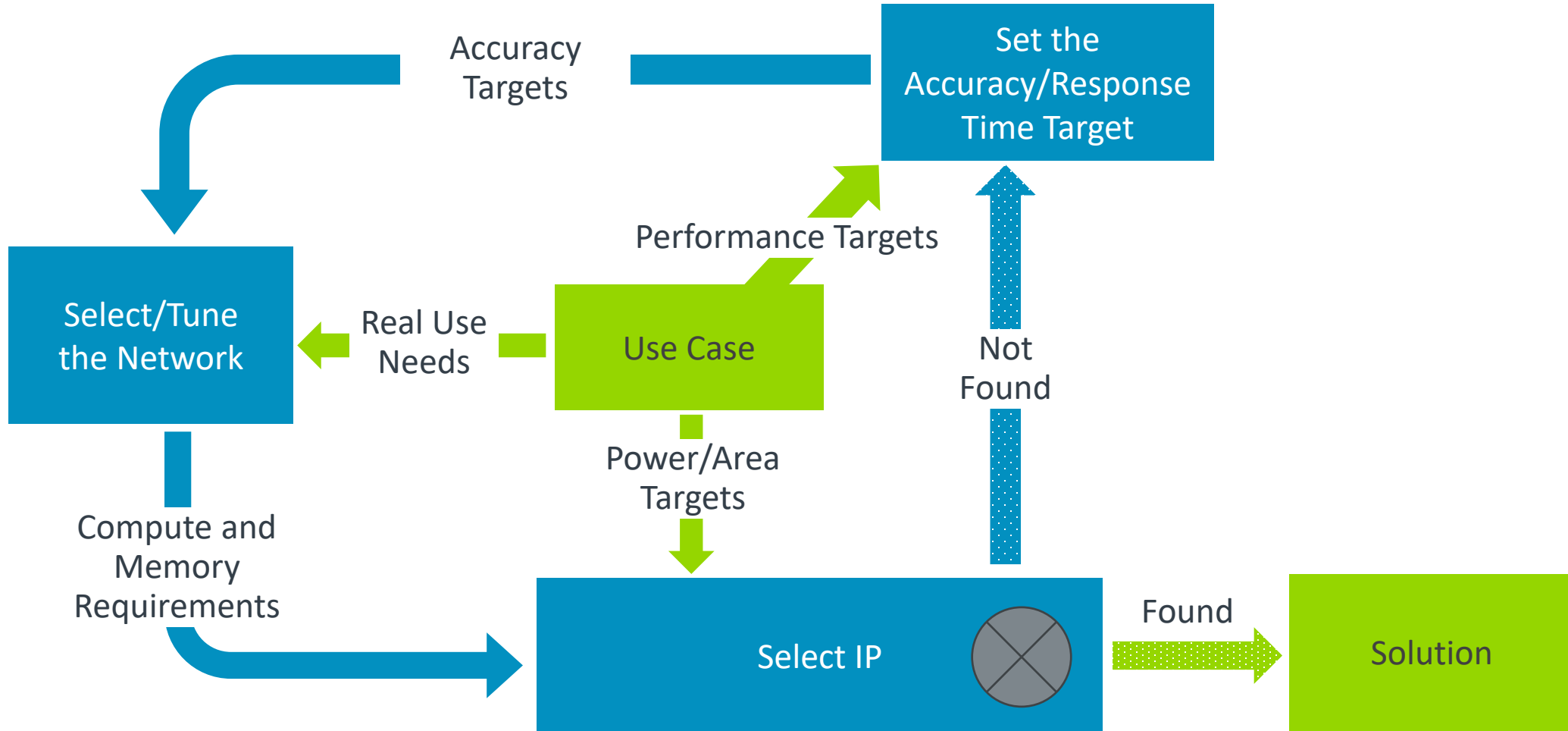


Networking and Servers

The Challenge with Balancing Product Decisions

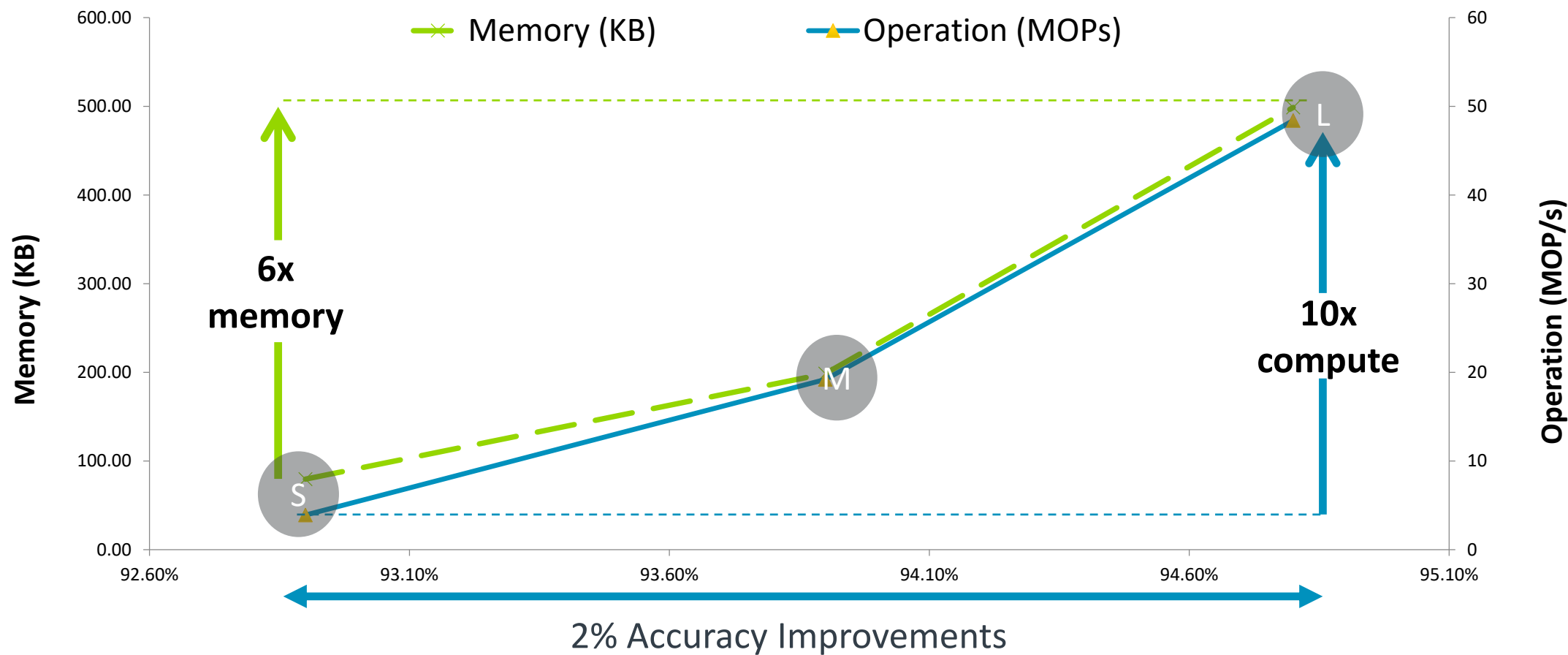


IP Selection Considerations



Choosing Realistic Accuracy Targets

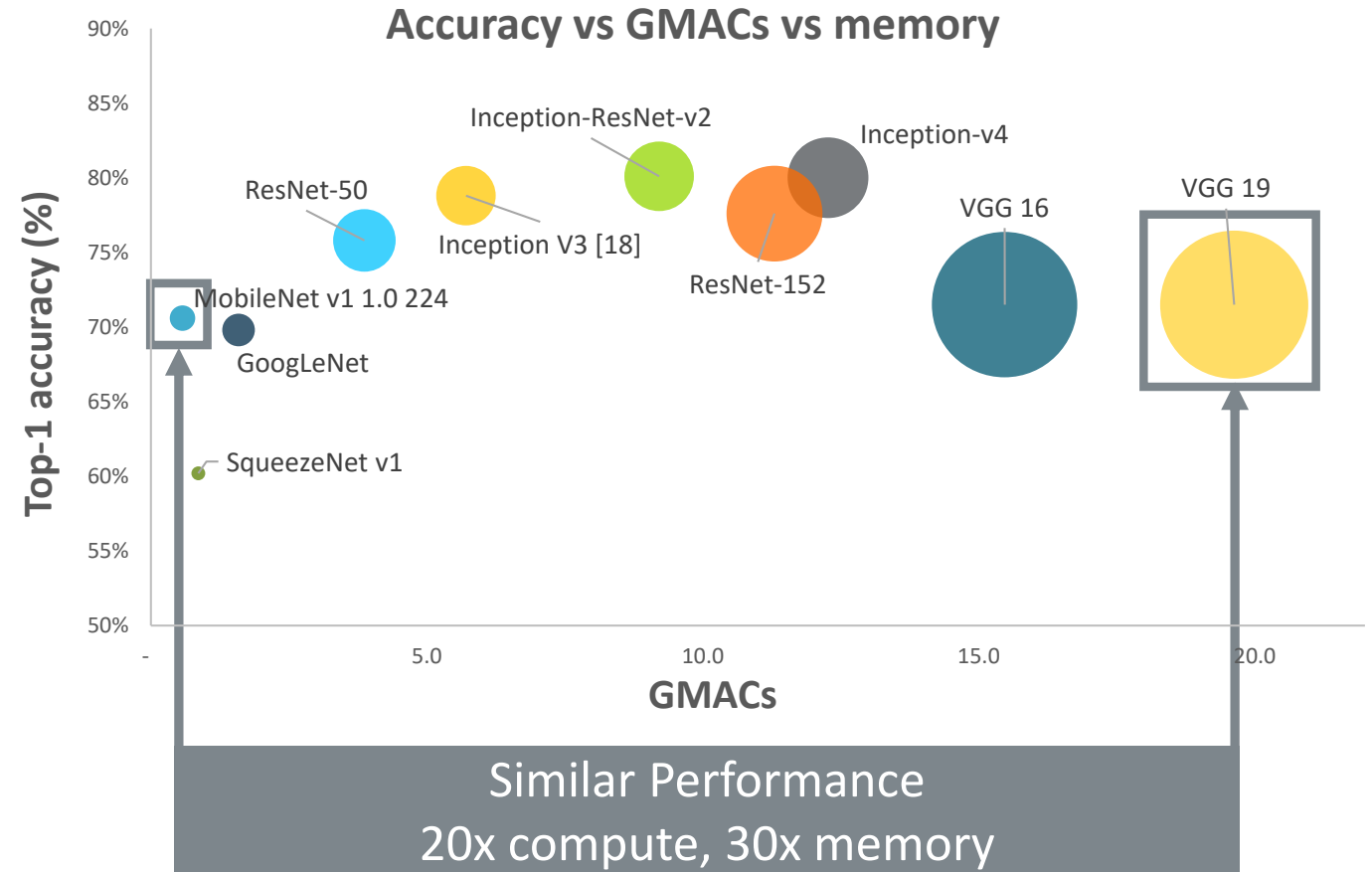
Accuracy Gain vs. Power/Area Increase (for Keyword Spotting)



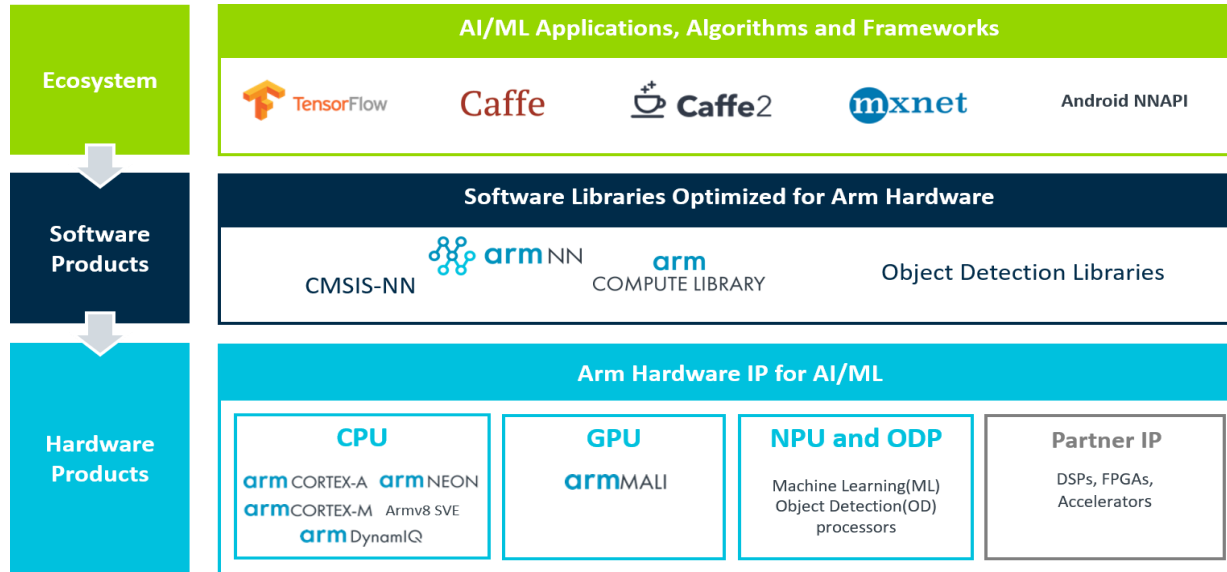
A mild accuracy improvement could result in high increase of compute and area requirements

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



Switching from IP to IP with Ease



ARM DS-5 Streamline Performance Analyzer

Drill down through source code

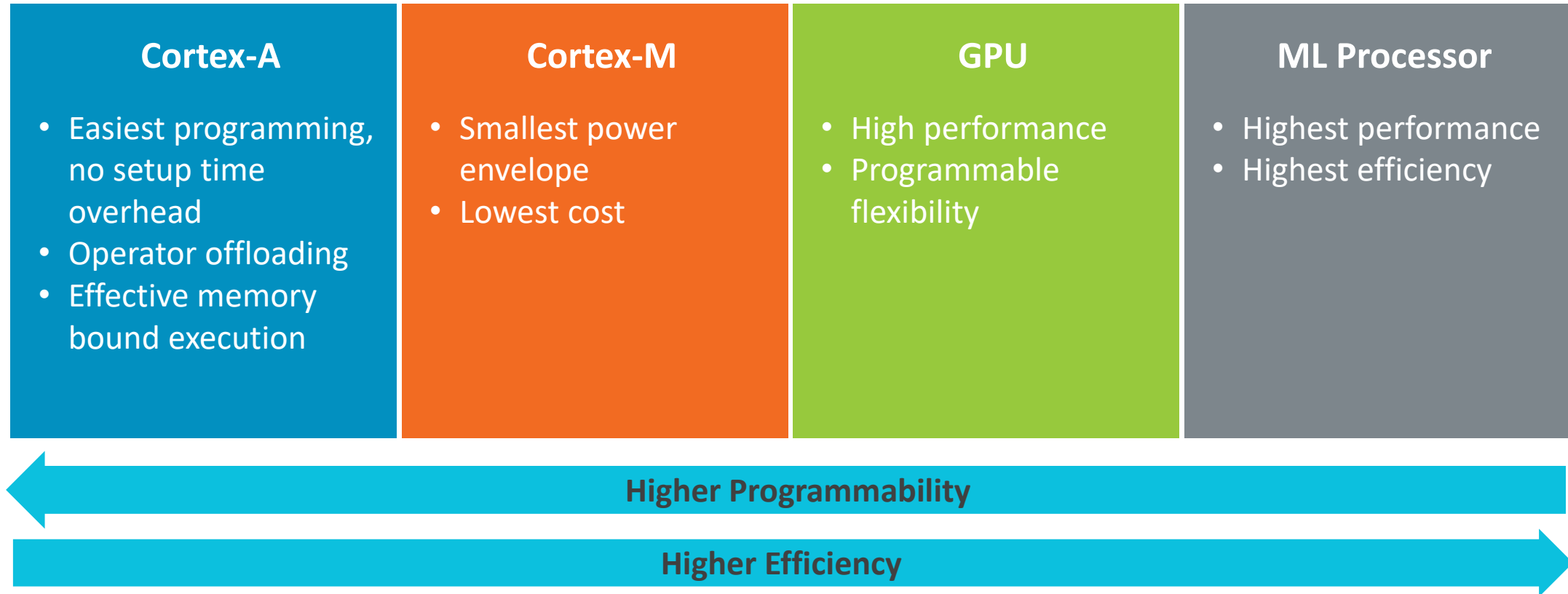
Speed up your code

OpenCL[™] Visualizer

- 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

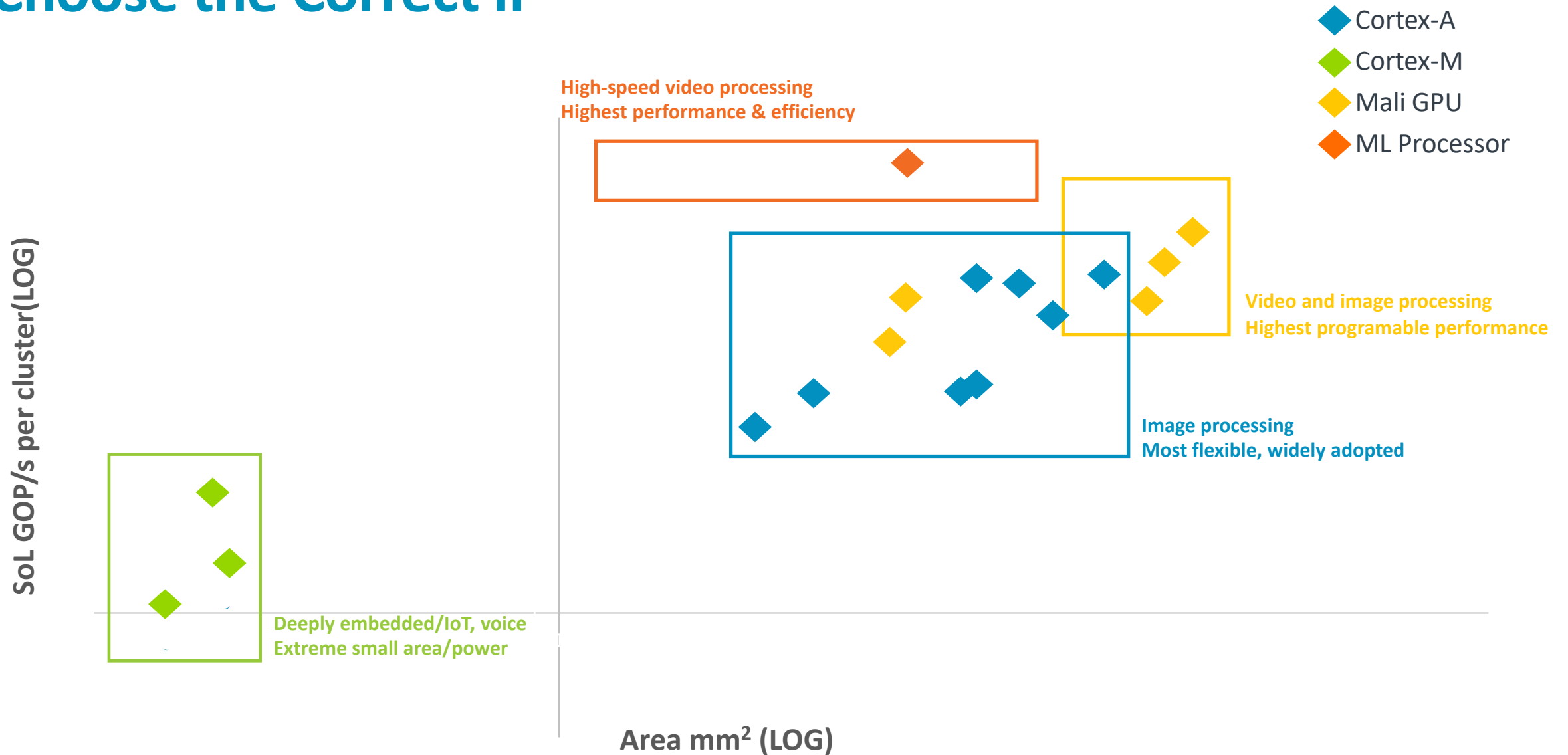
Heterogeneous Compute

Maximize the benefits from all IP families



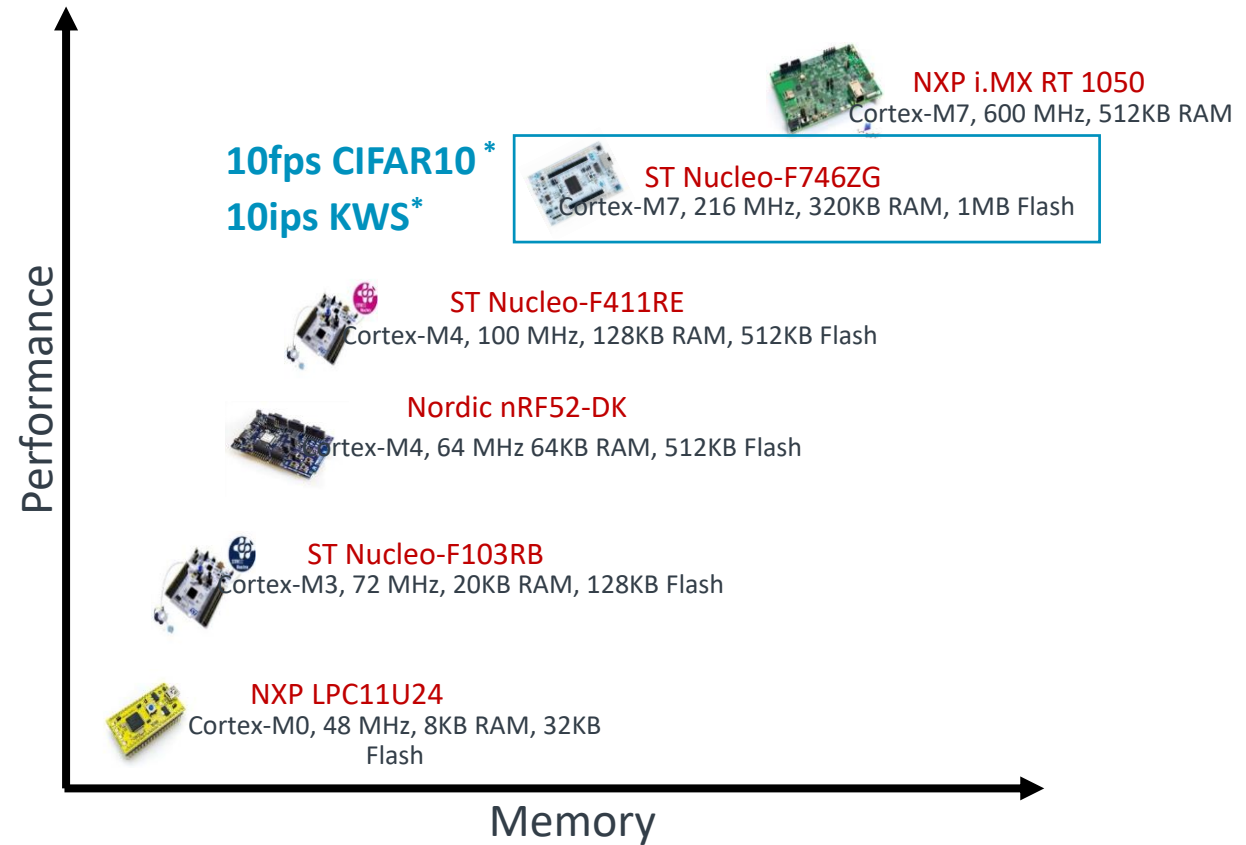
A platform built on heterogeneous compute provides the flexibility needed to match PPA across a wide range of use cases, workloads and market segments

Choose the Correct IP



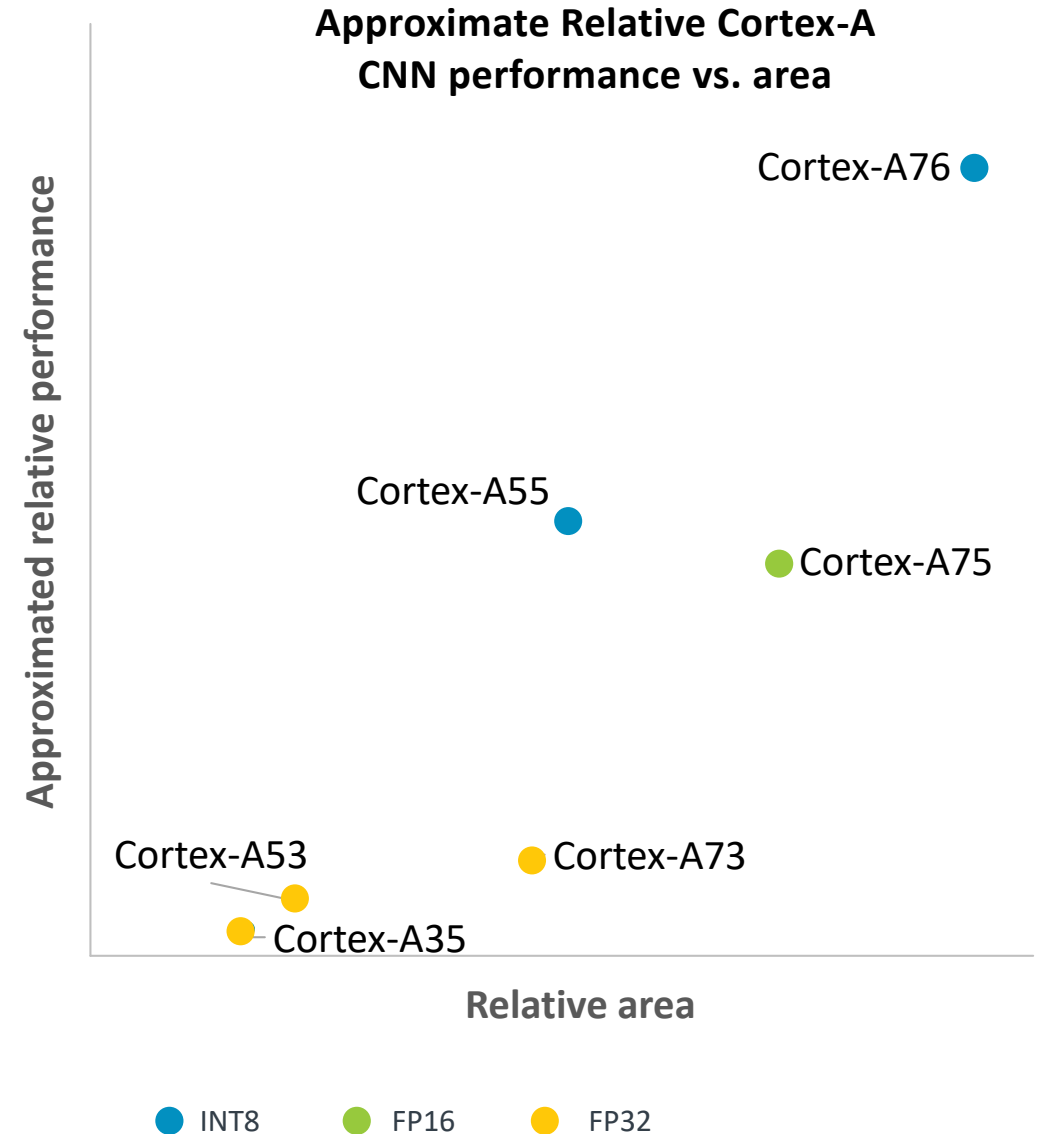
Cortex-M Microcontrollers

- Widely available in embedded hardware
 - Fully programmable
 - Extreme low power and small area
 - $\sim 0.1\text{mm}^2$, $\sim 10\text{mW}$ in 16FF
 - ML speech and image recognition
- Software support
 - CMSIS-NN, CMSIS-DSP
 - Tuned ML functions
 - General purpose DSP functions



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



Mali GPUs

Available in a range of devices

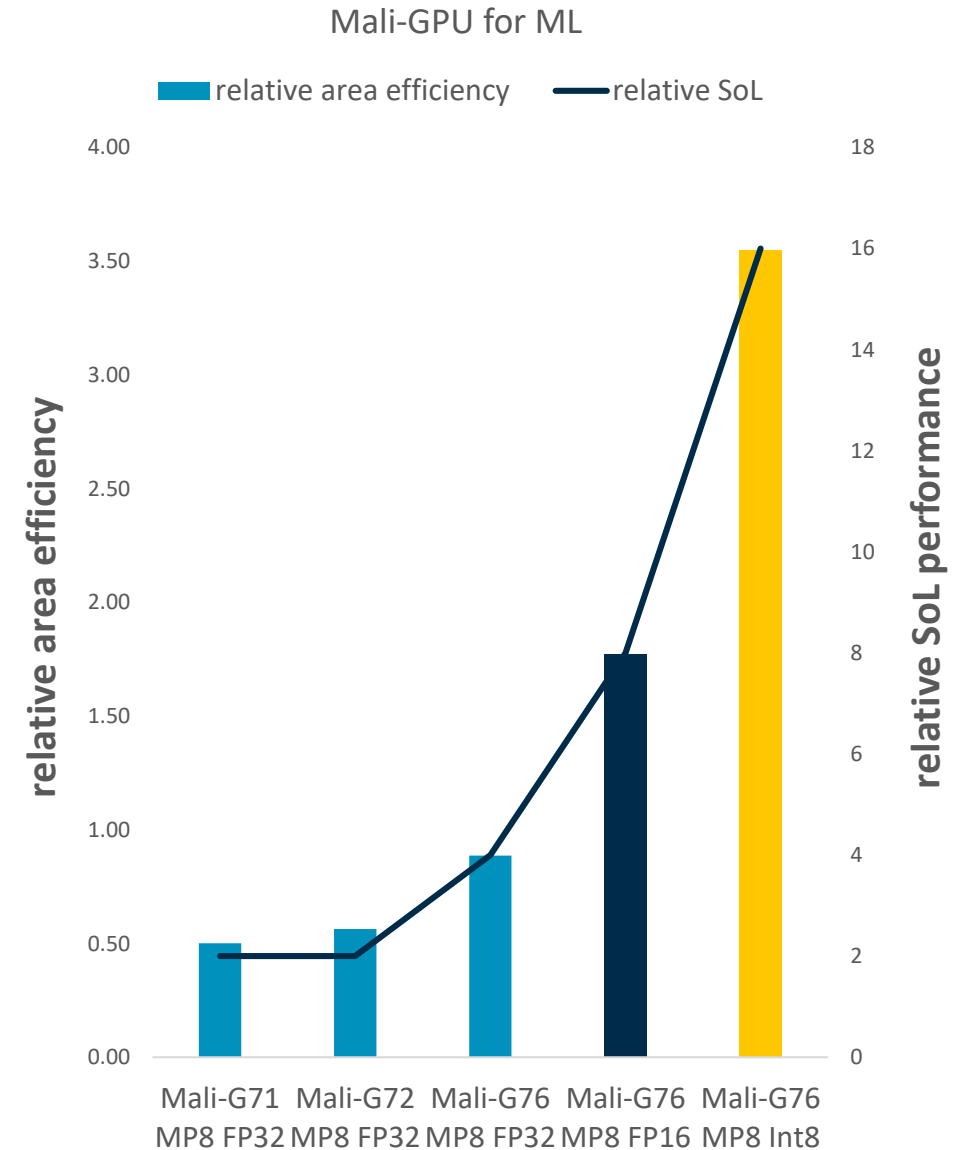
- Mobile phones, DTV, surveillance cameras, automotive IVI etc

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

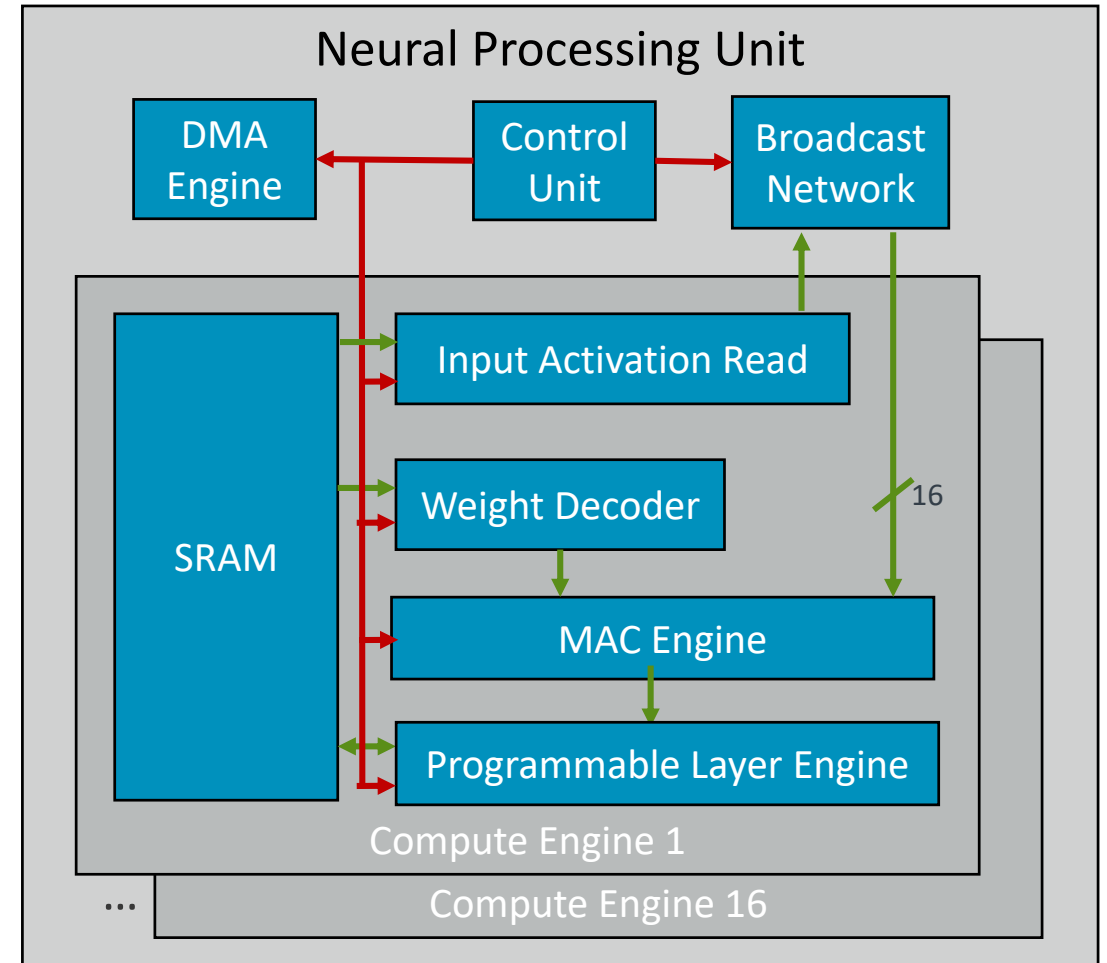
Software support

- Fully programmable
- Arm NN, Compute Library



Neural Processing Units (NPUs)

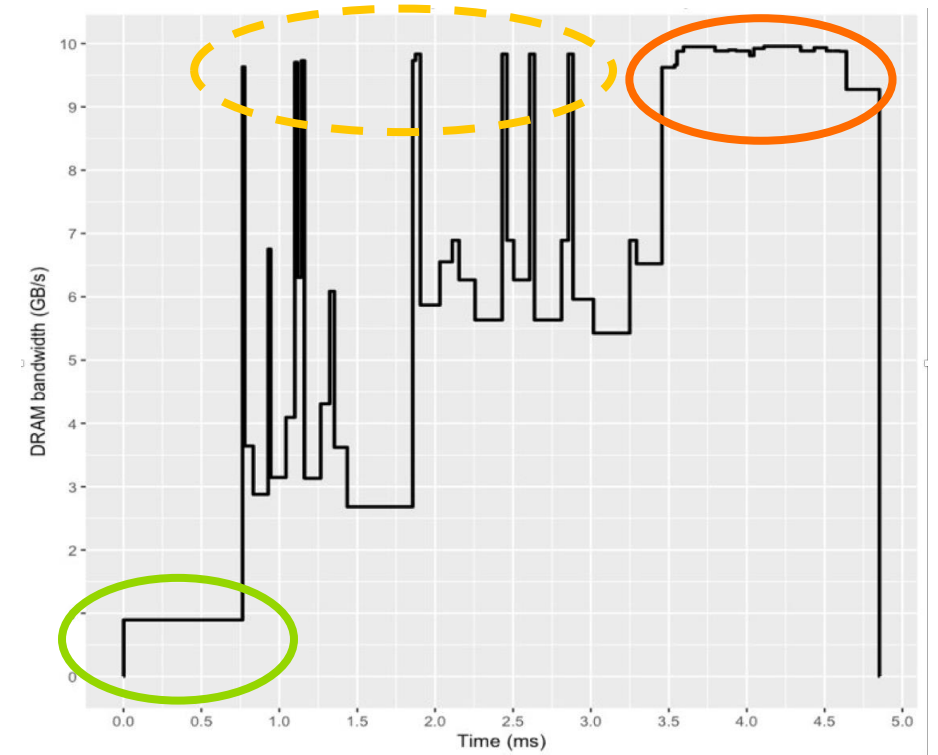
- Highest performance and efficiency
 - Scalable with family of ML processors
- Programmability for futureproofing
 - Based on Arm microcontroller technology with tool support
 - Operators can be added after tape out
 - Encompassing a range of data types in the product line
- Supported by Arm NN and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



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

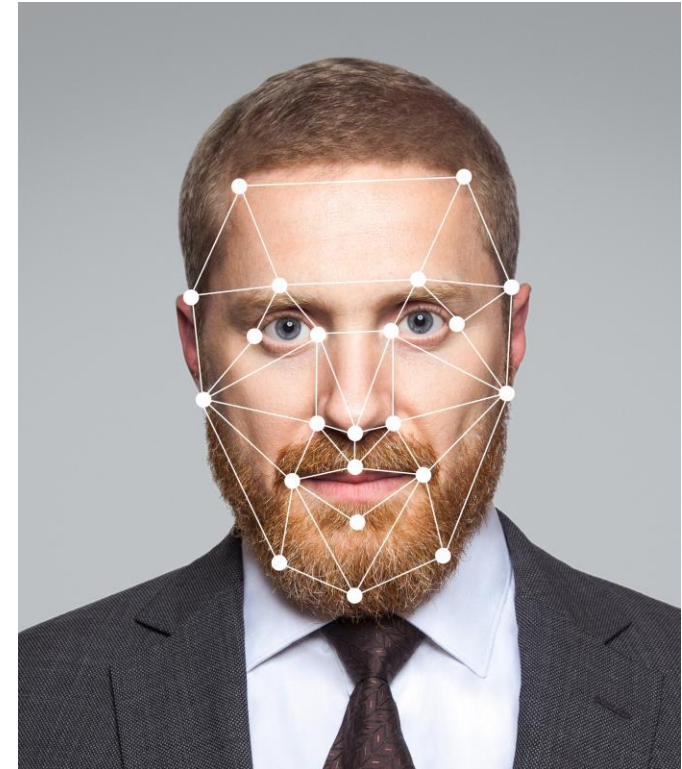
Memory requirements during inference run



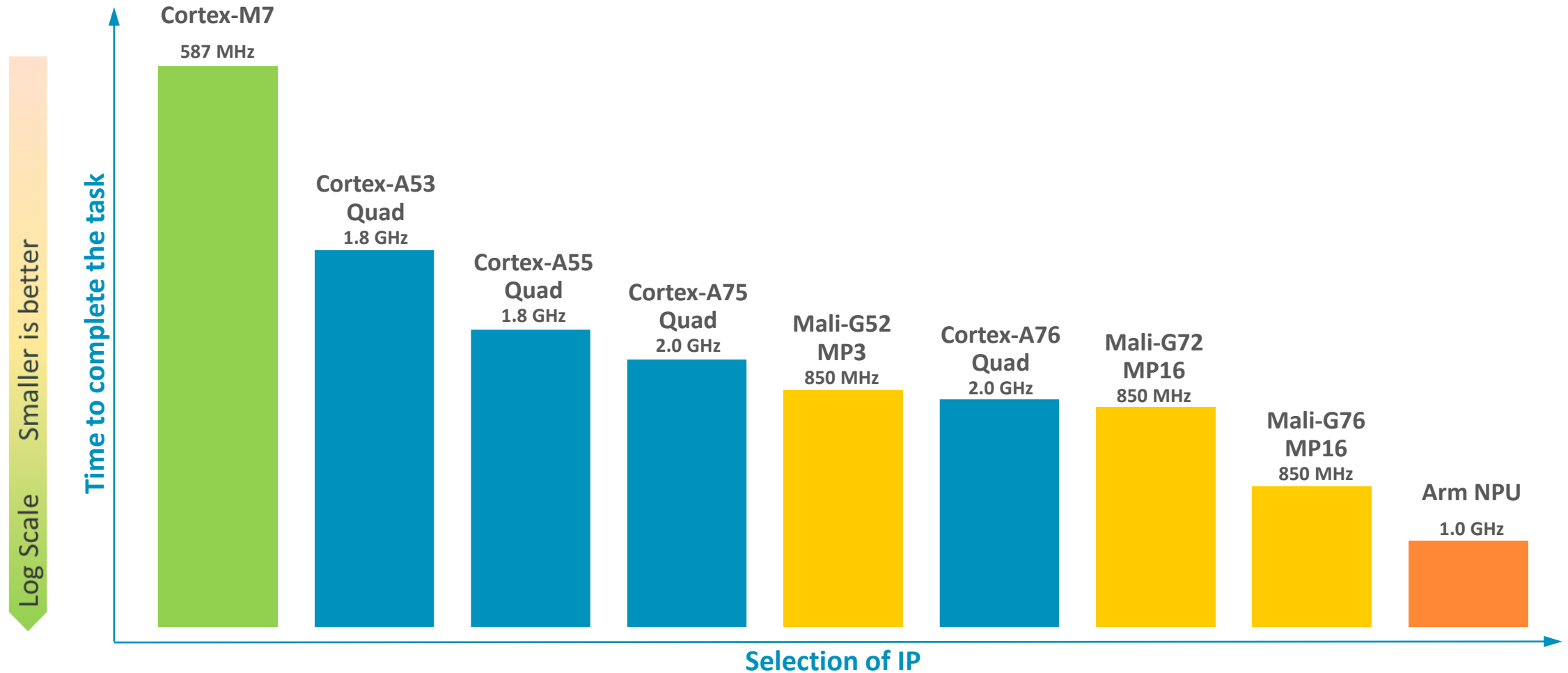
- Compute bound
- Mixed
- Memory bound

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



Example: IP Selection for Face Unlock (16FF)



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

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 0.1mm² to 100s of mm²
 - 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

arm

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