Comprehensive Arm Solutions for Innovative Machine Learning (ML) and Computer Vision (CV) Applications

Helena Zheng ML Group, Arm

Arm Technical Symposia 2017, Taipei

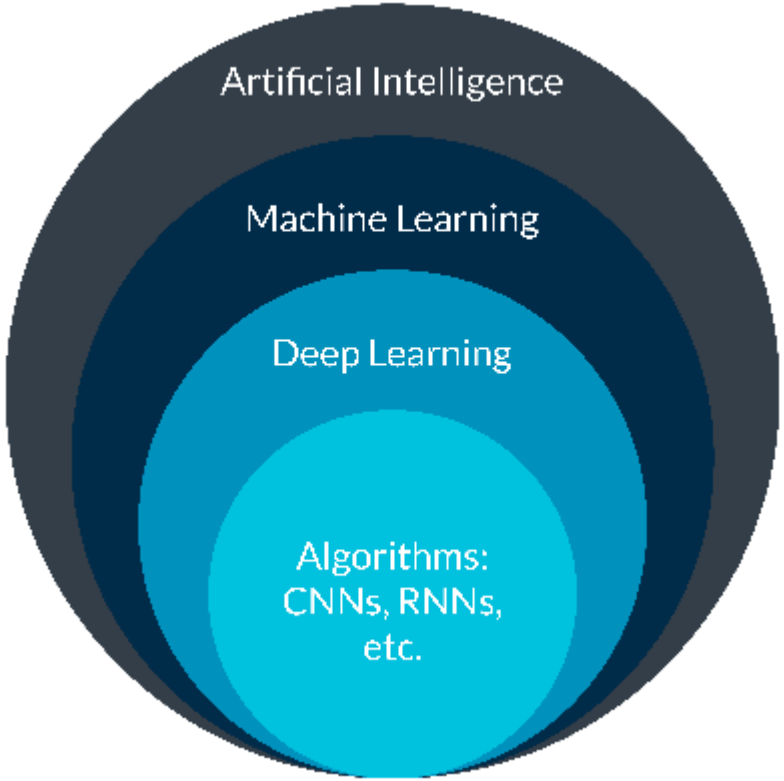
Machine Learning is a Subset of Artificial Intelligence

AI means many things to many people

Artificial Intelligence

- Machine Learning
- Perception & Vision
- Natural Language Processing
- Knowledge Representation
- Planning & Navigation
- Generalized Intelligence

ML itself has a lot of depth



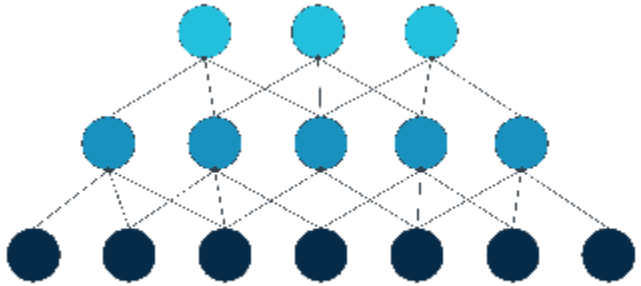
arm

Why Artificial Intelligence(AI) Exploding Now

Availability of increased data sourced at the edge with ubiquitous powerful compute!

Compute

2010



NN

2015

Data

2016 – 1 zettabyte

IP Traffic

2020 – 2.3 zettabyte

arm

AI Presents Significant Opportunity for Innovation

VR/MR

Robotics

Drones

Shipping & logistics

IoT

Home, surveillance & analytics

Automotive

Mobile

arm

Distributing Intelligence

Cloud-based training

High-performance processing

On-device learning

Security and privacy for your data

AI in your hand & cloud

Real-time inference for autonomous systems

arm

Why is On-device ML Driving AI to the Edge?

Bandwidth

Power

Cost

Latency

Privacy

arm

Arm ML Platform Enables

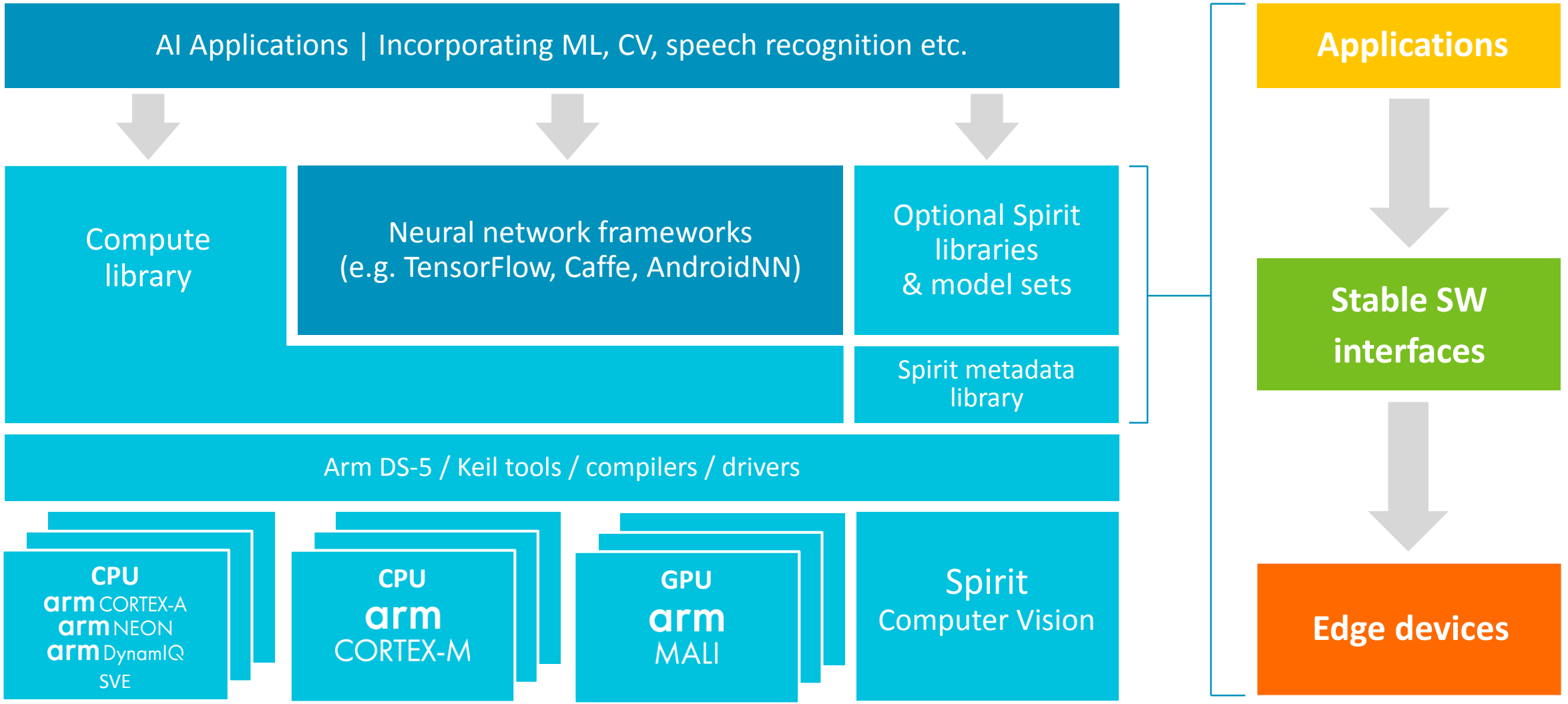
Efficiency

Flexibility

Freedom

arm

Arm's ML Computing Platform



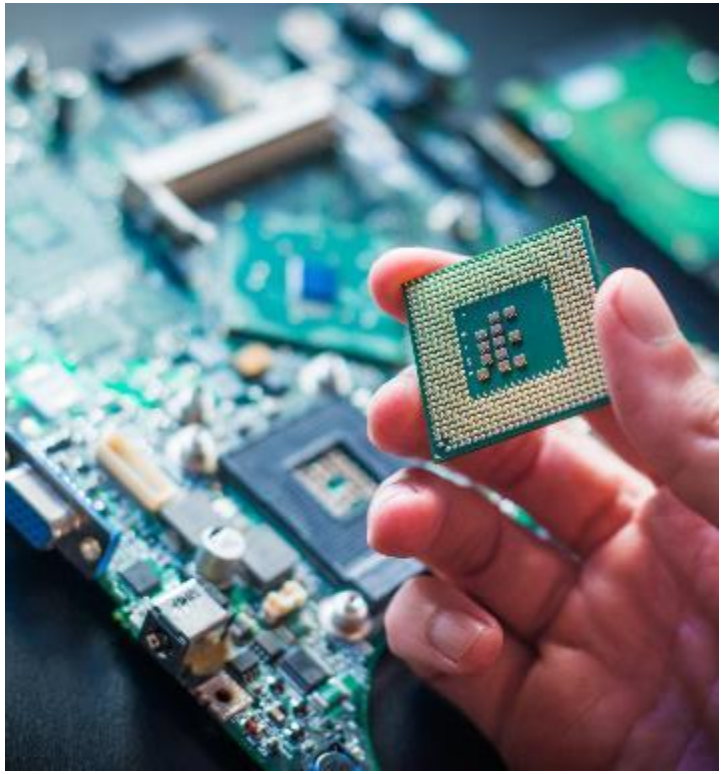
arm

Components of Arm ML Platform

Software

Specialized Acceleration

Hardware



arm

Software Development

arm

Arm Compute Library

Faster, advanced processing

What is the Arm Compute Library?

Functions for CV and deep-learning algorithms

Optimized for Arm CPU and GPU

OS and platform agnostic

No fee, MIT license

Delivers faster processing

4.6x faster than stock OpenCV on NEON

Offers OpenCV and Open VX compatibility

Use as a plug-in backend for your own runtime implementation

Available now: https://developer.arm.com/technologies/compute-library

arm

Arm Compute Library Partners

arm

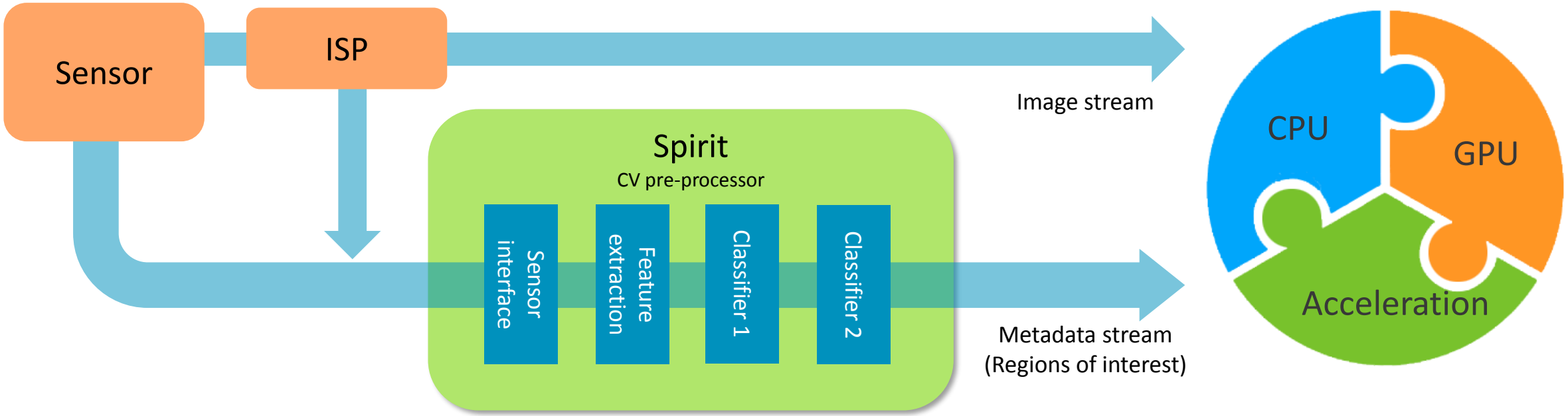
14

arm

Computer Vision (CV)

arm

Spirit for Object Detection and Localization



arm

Comparison with Neural Network Framework Solutions

arm

Spirit: Key Features

Object localization (>200k locations in full HD)

Size variation (~20x for full HD)

Scalable to 4K without performance compromise

Real time, 60 fps, no dropped frames

Invariance to optical distortions

Invariance to illumination conditions

High occlusion tolerance

Suitable for stationary and moving cameras

18

arm

Comparison with a DSP

Spirit uses a form of HOG*/ SVM* baked into an efficient hardware design

Using a DSP to achieve the same performance

E.g. Pedestrian Detection on a DSP

- Processed at VGA resolution
- Achieving 5fps
- Operating at 40MHz/50MHz
- Scaling this to Spirit performance levels of 1080p60 would require the DSP to run at 3.24GHz

*Histogram of Oriented Gradients
*Support Vector Machines

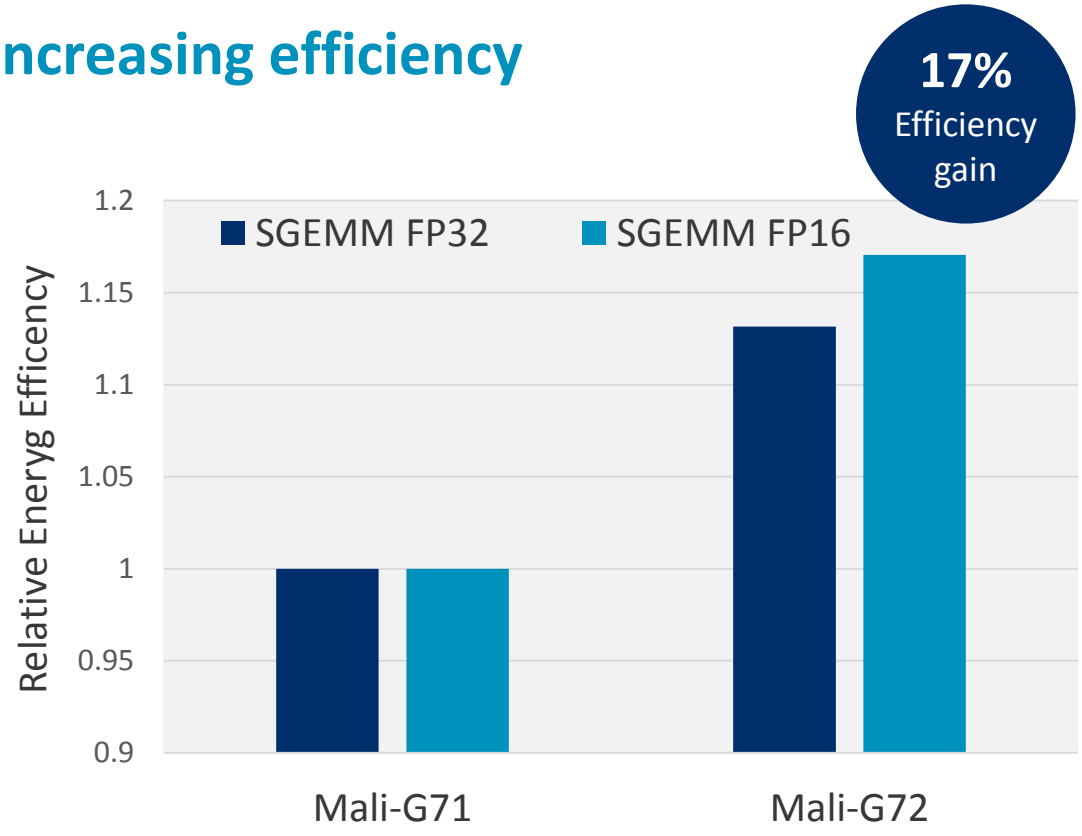
19
arm

ML on Mali GPUs

arm

Mali GPUs: Increasing ML Throughput and Efficiency

Increasing efficiency



- GEMM depicts core functionality of ML algorithms
- Mali-G72 has several optimizations to improve ML inference
  - Less power-hungry FMA unit
  - Bigger L1 cache in the execution engine
- Mali-G72 is the most efficient Mali GPU for machine learning

arm

Hardware

22

arm

AI Applications at the Edge on Arm

Detect plant diseases

Sort cucumbers

Detect Caltrain delays

arm

Instruction Sets for AI

Cortex-A

- Additional dot product instructions (Cortex-A55 and Cortex-A75)
- New Scalable Vector Extensions (SVE) instructions

Cortex-M

- Optimized CMSIS-DSP libraries for matrix multiplication

Closely-coupled acceleration

- Improved performance and efficiency (for broader use cases)
- Flexibility in multi-core computing with Arm DynamIQ technology

25
arm

DynamIQ: New Cluster Design for New Cores

Arm DynamIQ big.LITTLE systems:

- Greater product differentiation and scalability
- Improved energy efficiency and performance
- SW compatibility with Energy Aware Scheduling (EAS)

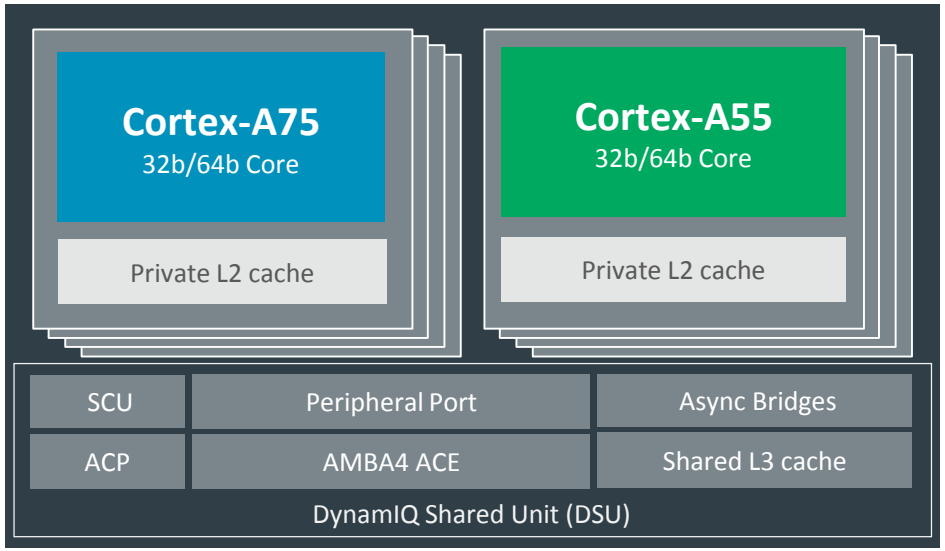
Private L2 and shared L3 caches

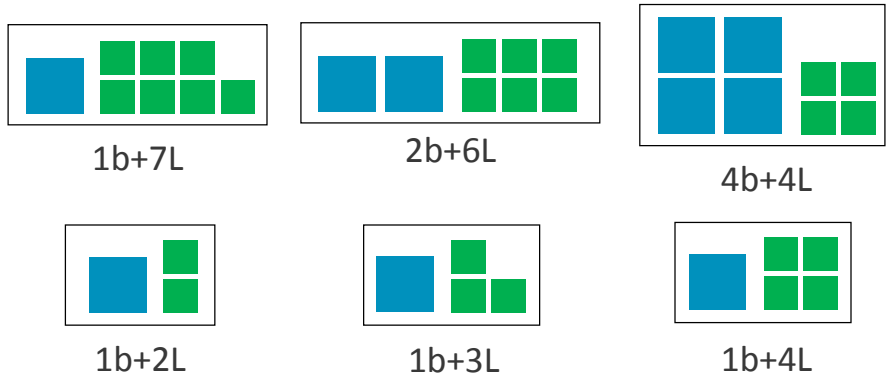
- Local cache close to processors
- L3 cache shared between all cores

DynamIQ Shared Unit (DSU)

- Contains L3, Snoop Control Unit (SCU) and all cluster interfaces

Additional instructions for ML





Example: DynamIQ big.LITTLE configurations

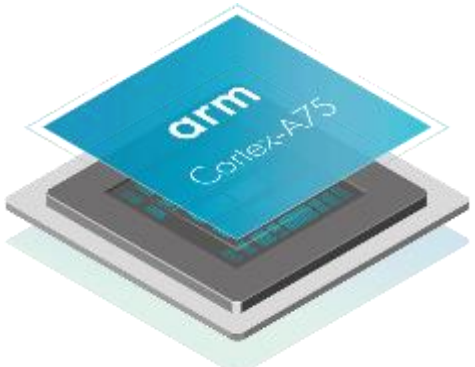
arm

New DynamIQ-based CPUs for New Possibilities

Cortex-A75 processor

>50%

more performance
compared to current devices

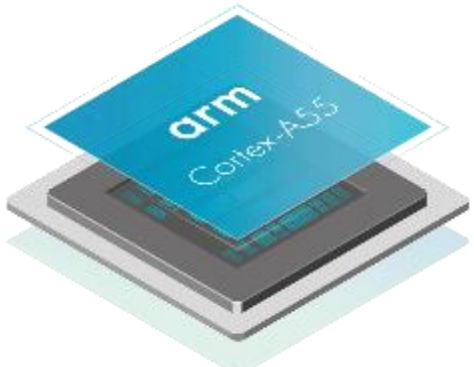


Estimated device performance using SPECINT2006, final device results may vary
Comparison using Cortex-A73 at 2.4GHz vs Cortex-A75 at 3GHz

Cortex-A55 processor

2.5x

higher power efficiency
compared to current devices

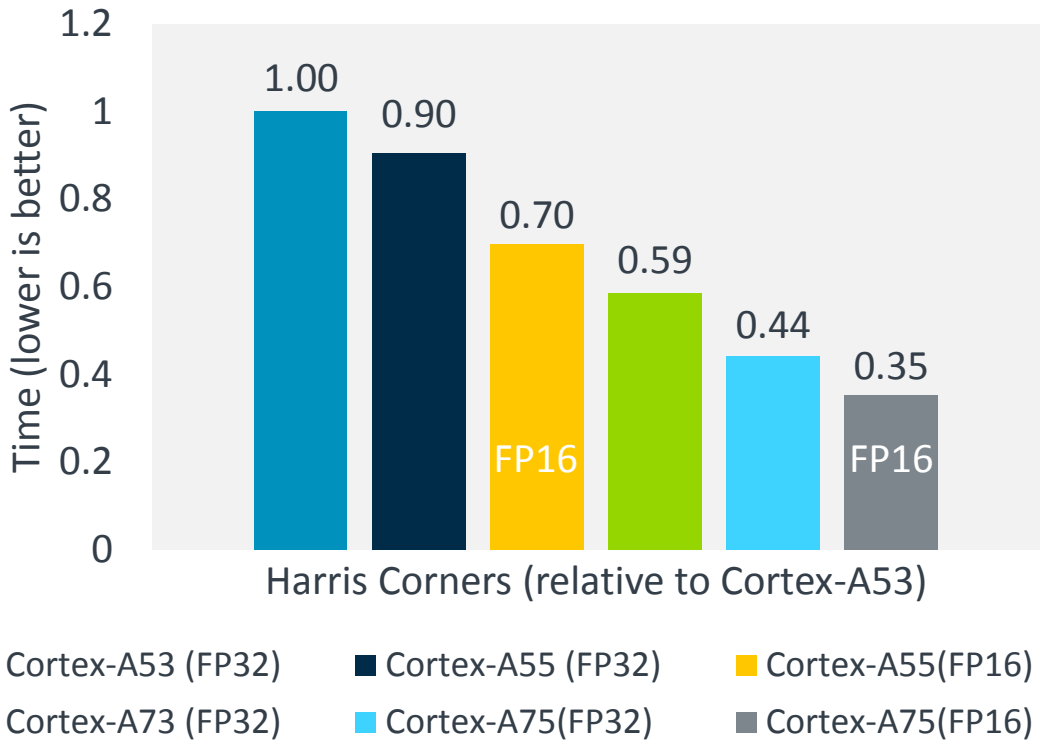


Comparison using Cortex-A53 in 28nm devices vs Cortex-A55 in 16nm devices

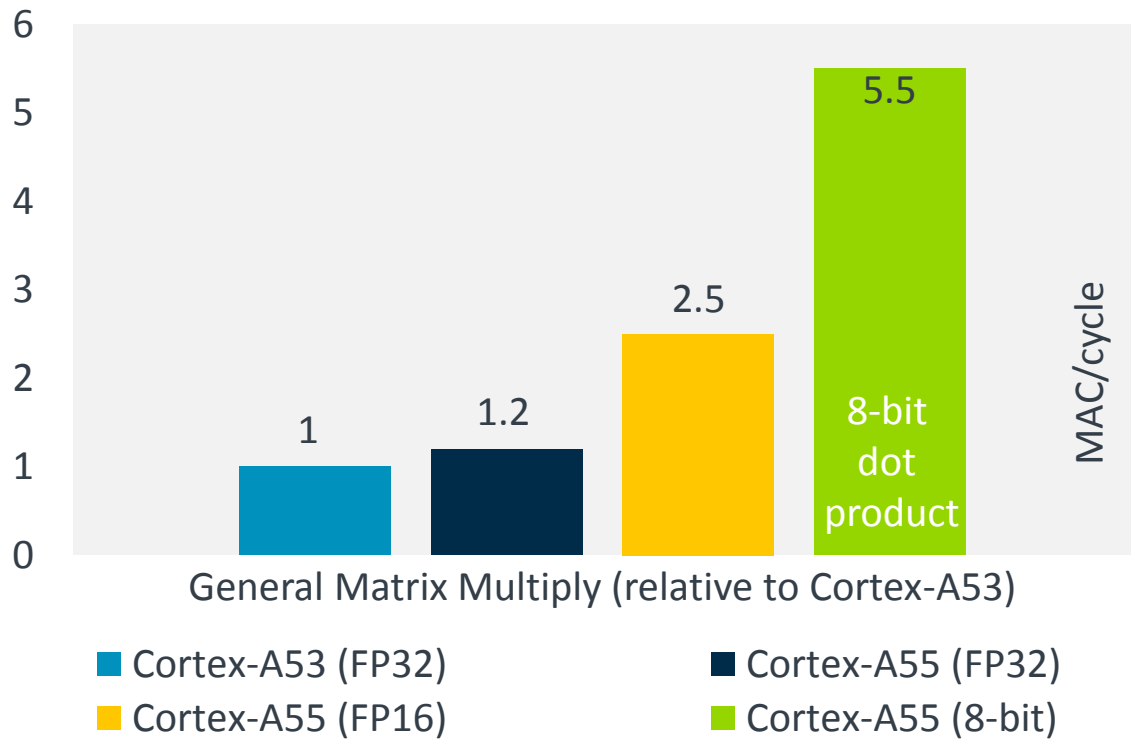
arm

Enhanced Architecture for Emerging Use Cases

Computer Vision

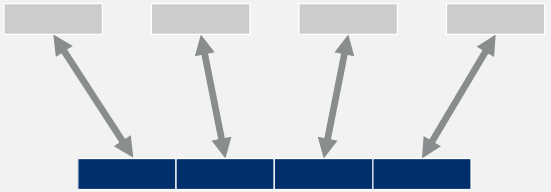


Machine Learning



arm

Introducing the Scalable Vector Extension (SVE)



Gather-load
and scatter-store

	1	2	3	4
+	5	5	5	5
pred	1	0	1	0
=	6	2	8	4

Per-lane predication

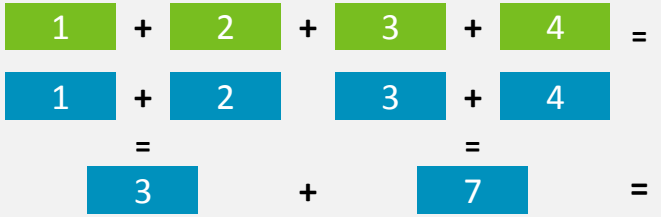
for (i = 0; i < n; ++i)

INDEX i	n-2	n-1	n	n+1
CMPLT n	1	1	0	0

Predicate-driven loop
control and management

	1	2		
+	1	2	0	0
pred	1	1	0	0

Vector partitioning and
software-managed speculation



Extended floating-point
horizontal reductions

arm

Instruction Sets for AI

Cortex-A

- Additional dot product instructions (Cortex-A55 and Cortex-A75)
- New SVE instructions

Cortex-M

- Optimized CMSIS-DSP libraries for matrix multiplication

Closely-coupled acceleration

- Improved performance and efficiency (for broader use cases)
- Flexibility in multi-core computing with DynamIQ technology

arm

Software Optimizations: Cortex-M Example

Convolution

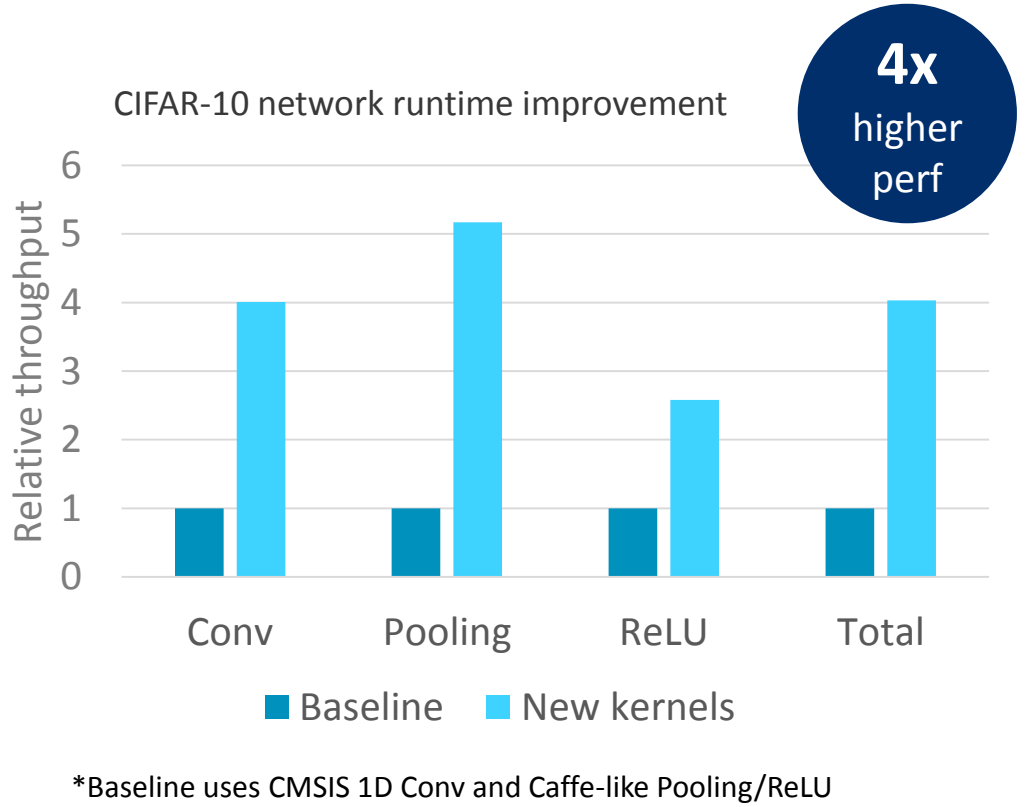
- Use of partial im2col to reduce the memory footprint
- Optimized data dimension layout (Height-Width-Channel) to save im2col overhead

Pooling

- Split into x-pooling and y-pooling instead of window-based
- 5.1X improvements compared to Caffe-like implementation

Activation

- ReLU: use SIMD within a register, 2.6X improvement compared to Caffe-like implementation
- Sigmoid and Tanh: use table look-up



*Baseline uses CMSIS 1D Conv and Caffe-like Pooling/ReLU

The new kernels will be integrated into future versions of CMSIS

arm

Instruction sets for AI

Cortex-A

- Additional dot product instructions (Cortex-A55 and Cortex-A75)
- New SVE instructions

Cortex-M

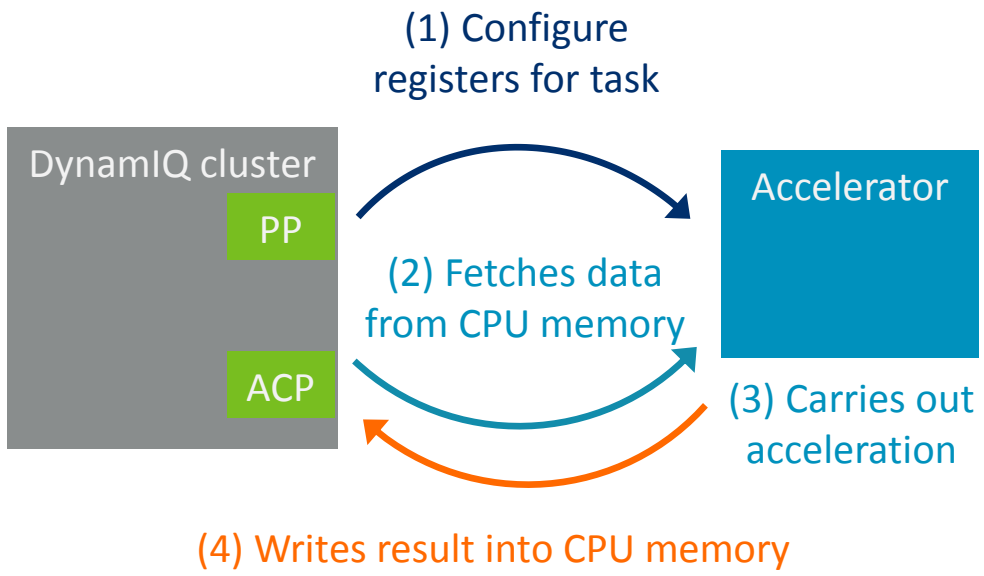
- Optimized CMSIS-DSP libraries for matrix multiplication

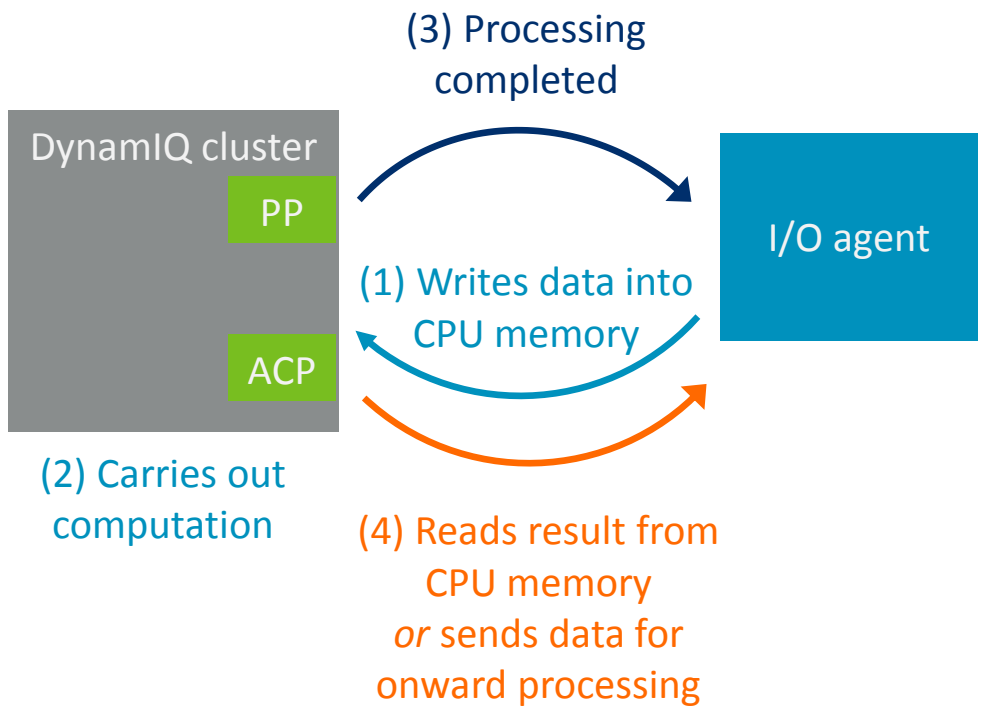
Closely-coupled acceleration

- Improved performance and efficiency (for broader use cases)
- Flexibility in multi-core computing with DynamIQ technology

arm

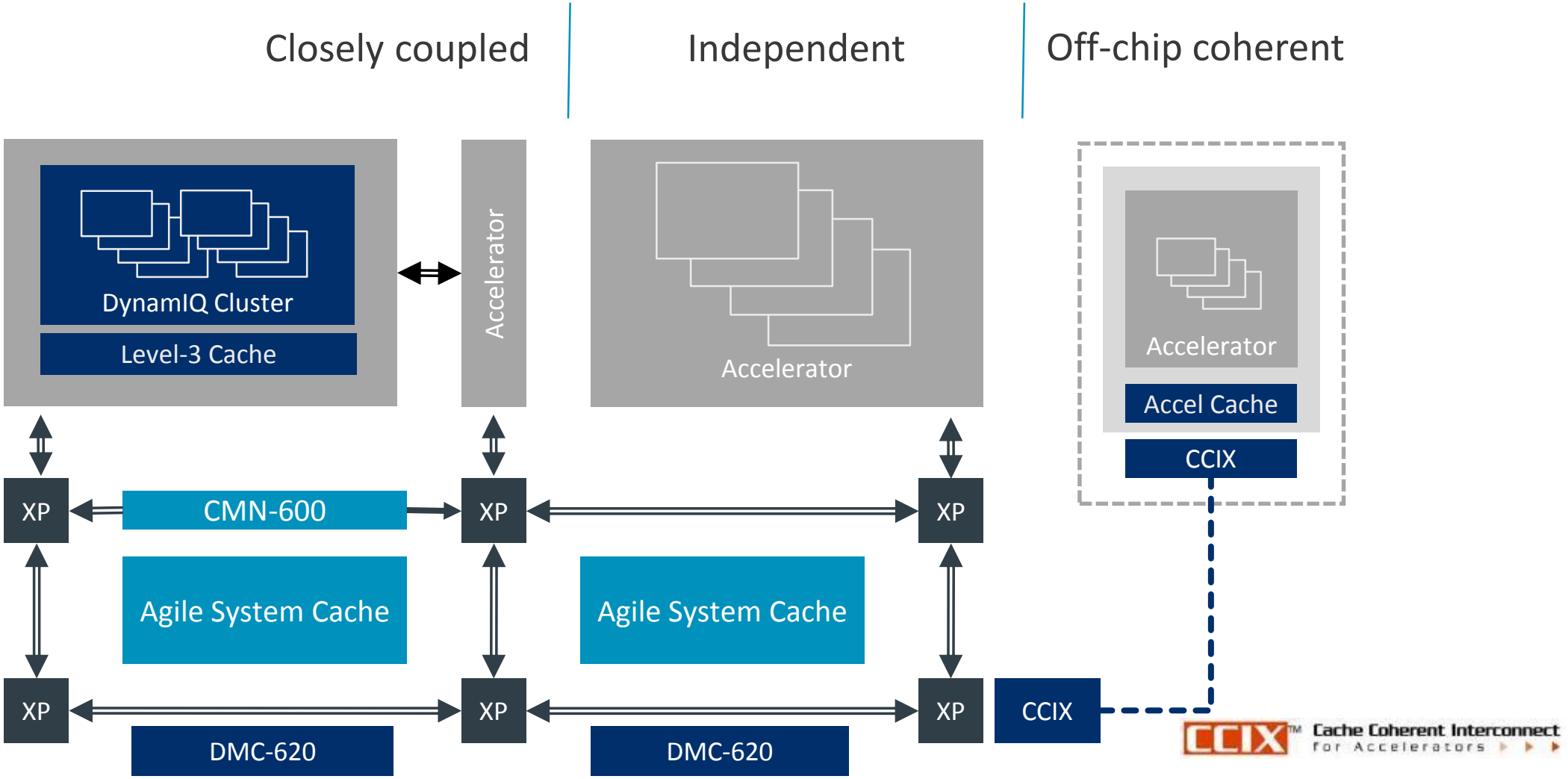
Interface to Acceleration Logic





arm

Flexible Acceleration Platform



arm

Arm ML Platform Enables

Efficiency

Flexibility

Freedom

arm

Thank You!
Danke!
Merci!
謝謝!
ありがとう!
Gracias!
Kiitos!