Arm DynamIQ: Intelligent Solutions Using Cluster Based Multiprocessing

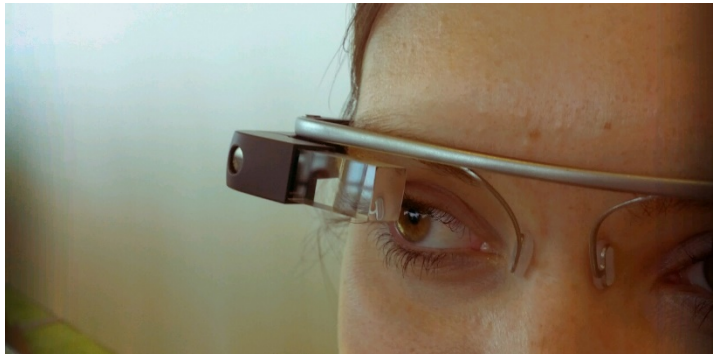
Peter Greenhalgh
VP and GM of Central Technology and Fellow

© 2017 Arm Limited

Technology innovations of 2013

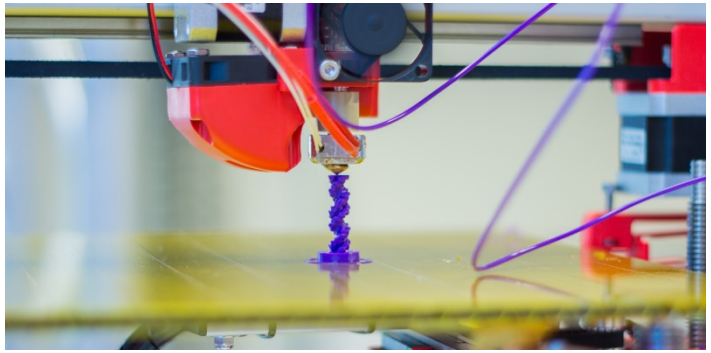
Drones

Wearable technology



Smartwatch

3D printing



Voice recognition

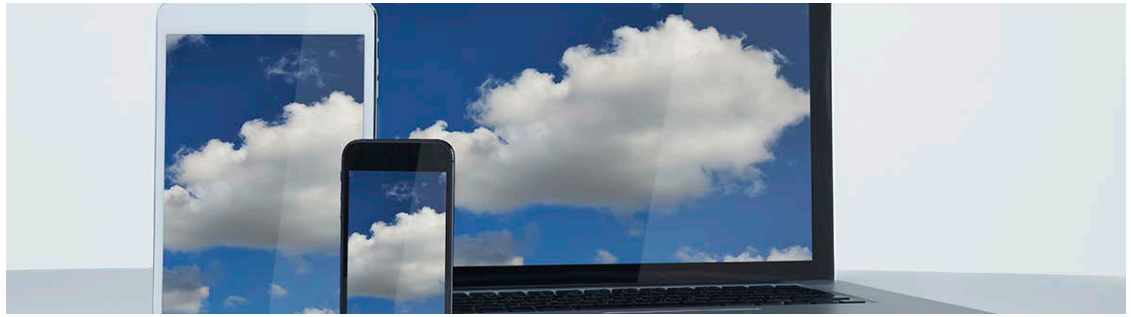
Social media

© 2017 Arm Limited
arm

Looking ahead from edge to cloud

The future requires a new approach to CPU design

Cortex beyond mobile



Mixed reality

Safe and autonomous

Hyper-efficient

Secure private compute

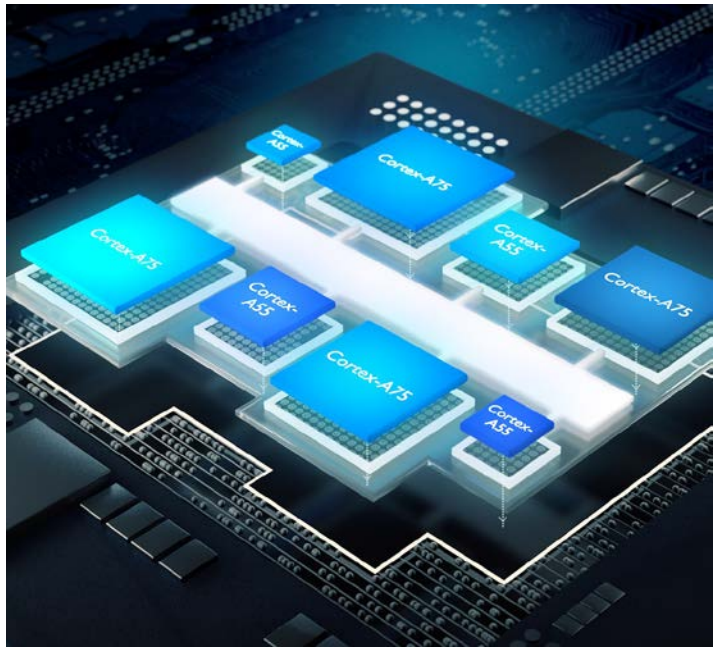
© 2017 Arm Limited

arm

Arm DynamIQ

Rearchitecting the compute experience

Multi-core redefined for broad market



Massive system performance uplift

More intelligent systems

Confidential © Arm 2017

Innovating for the scalable future

Key Arm technologies

- Arm NEON
- Arm AMBA
- Arm big.LITTLE
- Arm CoreLink
- Arm TrustZone

Expanding Arm technology processor architecture for broad market

Up to 8 CPUs
'Octacore' smartphones

Dual cluster

Heterogeneous processing

arm DynamIQ

Nearly **"Unlimited"** design spectrum

Covers all existing use cases

DynamIQ cluster

Dynamic flexibility

2013 → 2017

© 2017 Arm Limited
arm

DSU – Broadening the reach of technology

© 2017 Arm Limited

arm

DynamIQ: New cluster design for new cores

DynamIQ big.LITTLE systems:

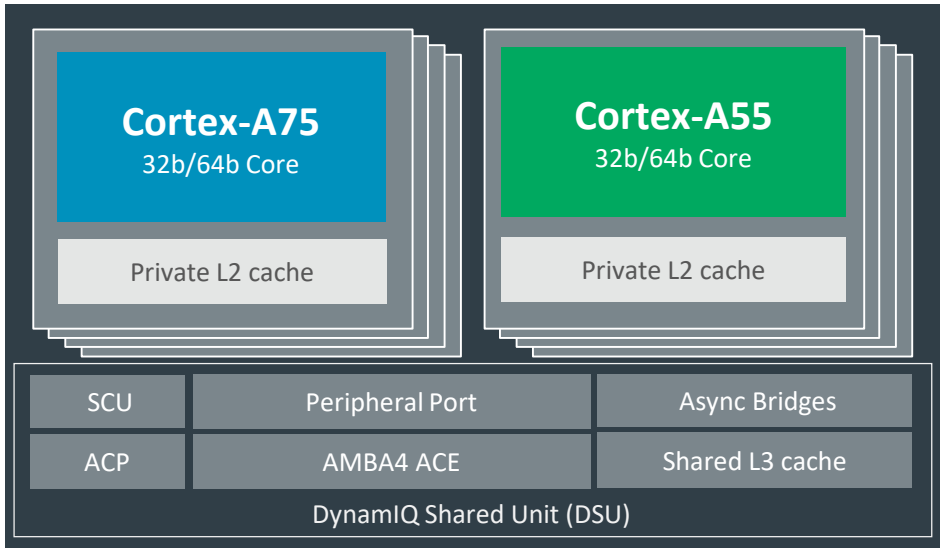
- Greater product differentiation and scalability
- Improved energy efficiency and performance
- SW compatibility with Energy Aware Scheduling (EAS)

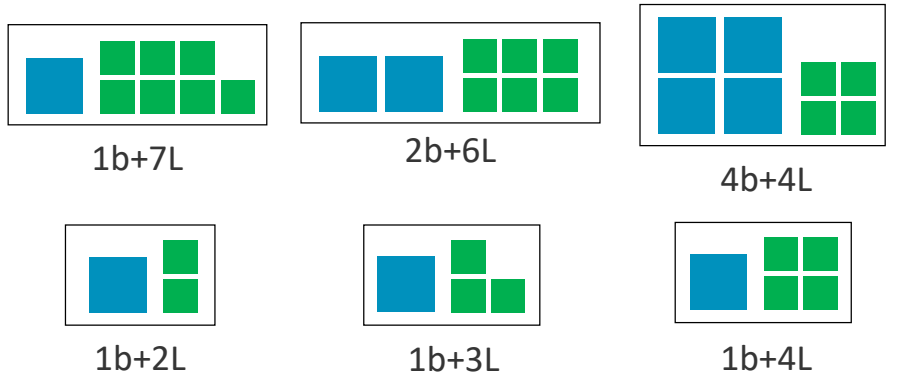
Private L2 and shared L3 caches

- 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

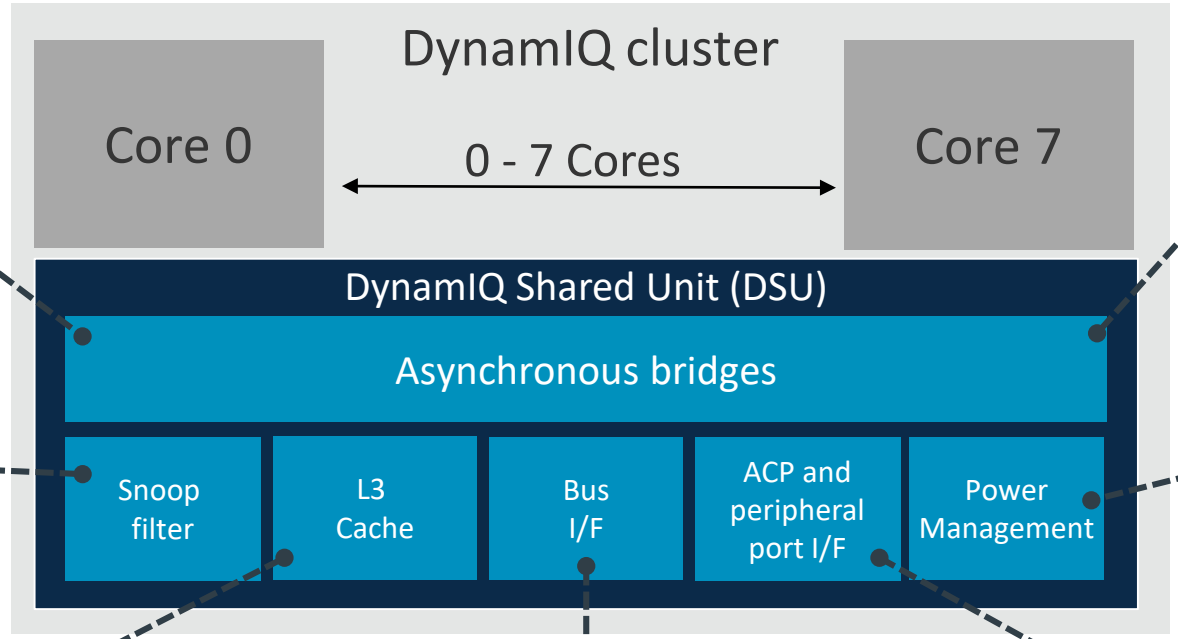




Example DynamIQ big.LITTLE configurations

© 2017 Arm Limited
arm

DynamIQ Shared Unit (DSU)



Support for multiple performance domains

Latency and bandwidth optimizations

Streamlines traffic across bridges

Advanced power management features

Supports large amounts of local memory

Scalable interfaces for edge to cloud applications

Low latency interfaces for closely coupled accelerators

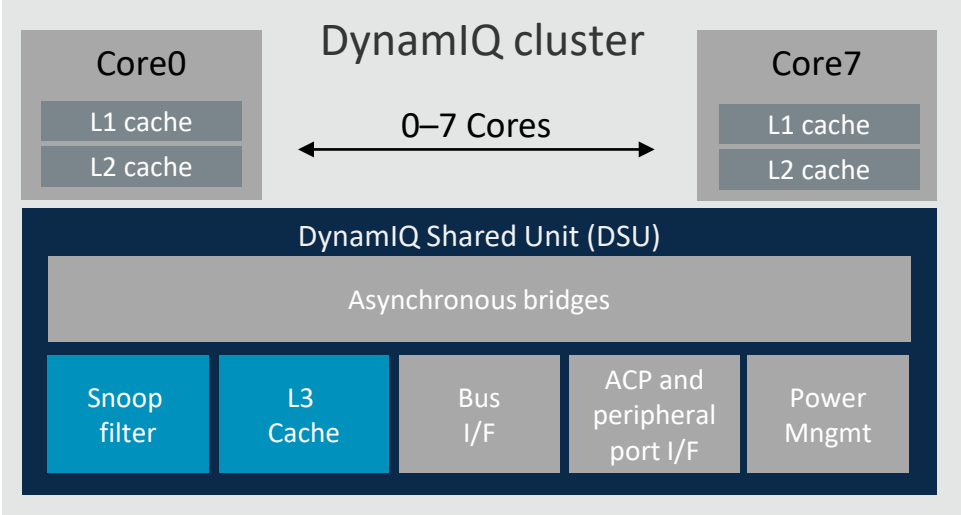
© 2017 Arm Limited
arm

Level 3 cache memory system

New memory system for Cortex-A clusters

Integrated snoop filter to improve efficiency

Enabling lower cache latencies



Load to Use Cycles*	Cortex-A53	Cortex-A55	Cortex-A73	Cortex-A75
L1 hit	3	2	3	3
L2 hit	13	6	19	8
L3 hit	-	21	-	25
Interconnect boundary	20	21	26	25

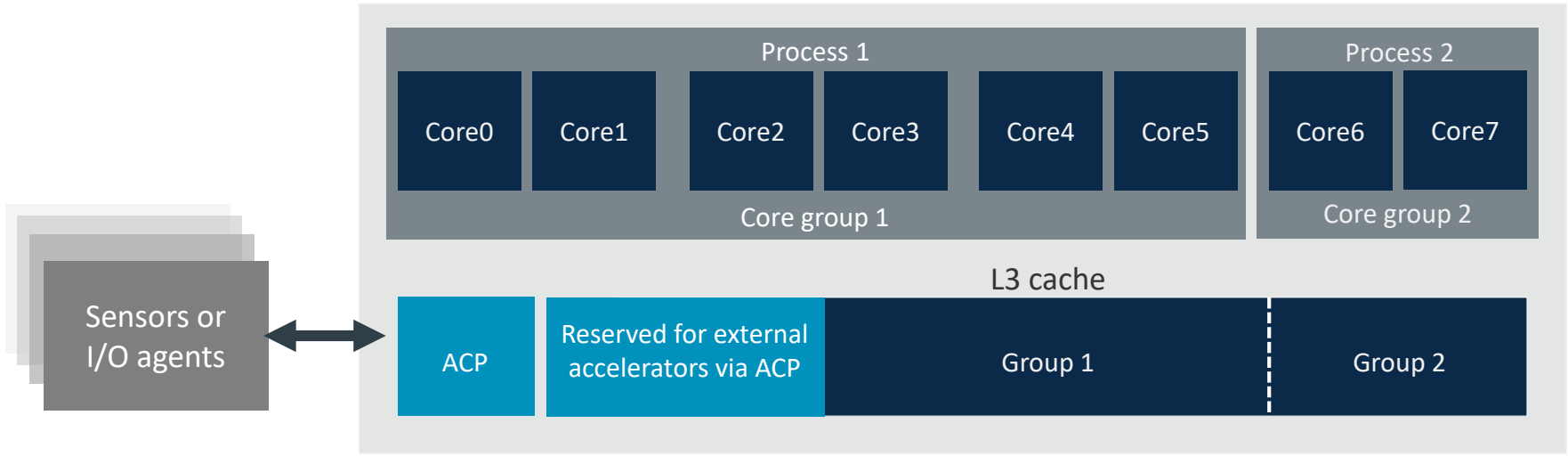
© 2017 Arm Limited
arm

Level 3 cache partition

Infrastructure

- Process 1 = data plane
- Process 2 = control plane
- Packet processing data sent through low latency ACP interface

Example configuration with two Core groups in a DynamIQ cluster



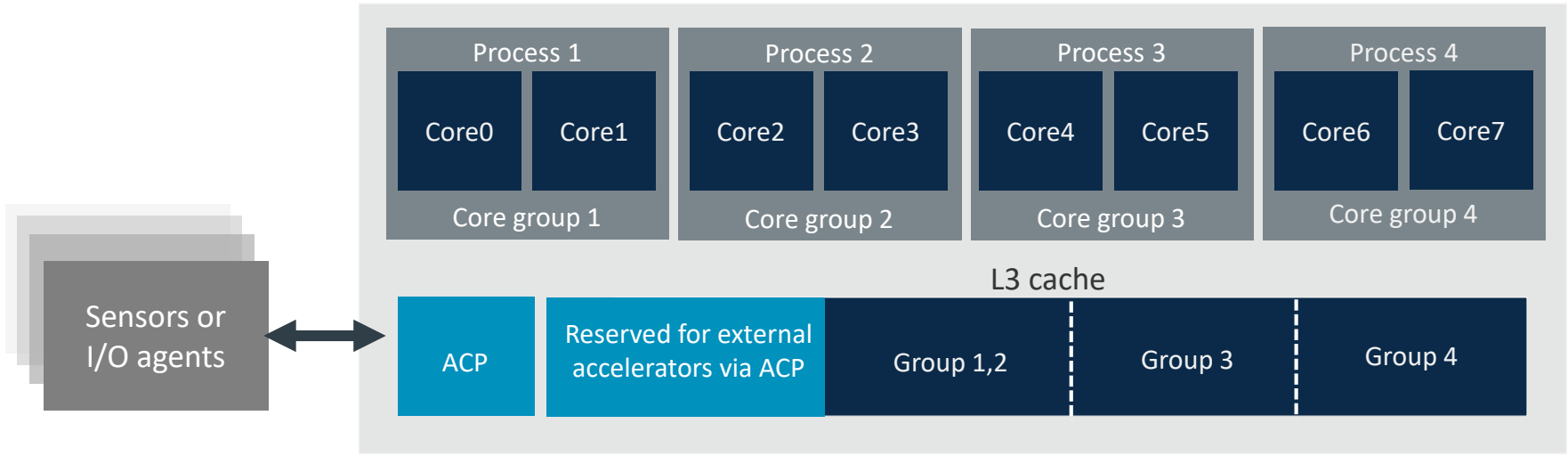
© 2017 Arm Limited
arm

Level 3 cache partition

Automotive

- Each process could represent an independent ADAS algorithm
- Sensors linked through low latency ACP interface

Example configuration with four Core groups in a DynamIQ cluster



© 2017 Arm Limited
arm

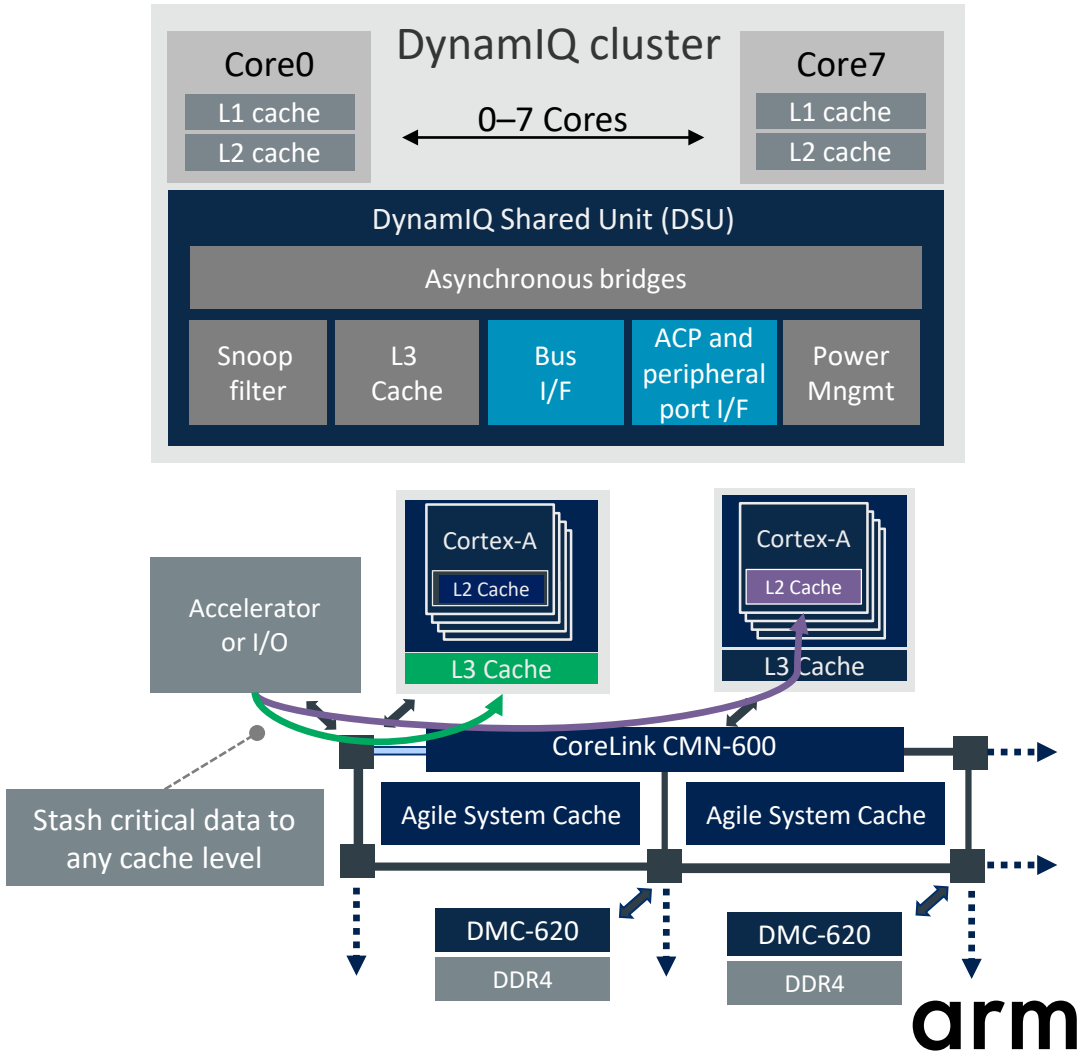
Increasing performance through cache stashing

Enables reads/writes into the shared L3 cache or per-core L2 cache

Allows closely coupled accelerators and I/O agents to gain access to core memory

AMBA 5 CHI and Accelerator Coherency Port (ACP) can be used for cache stashing

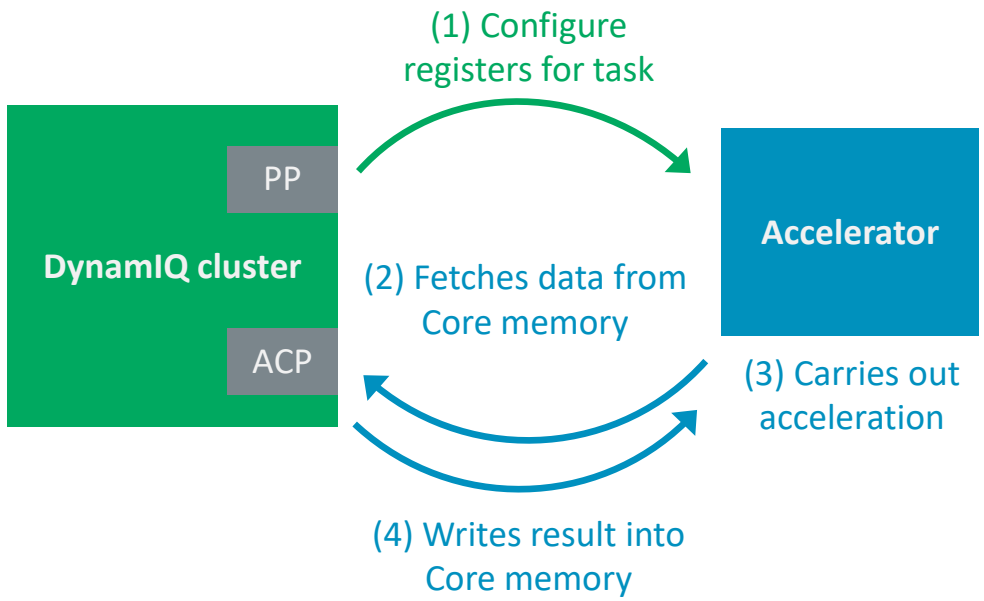
More throughput with Peripheral Port (PP) for acceleration, network, storage use-cases



© 2017 Arm Limited

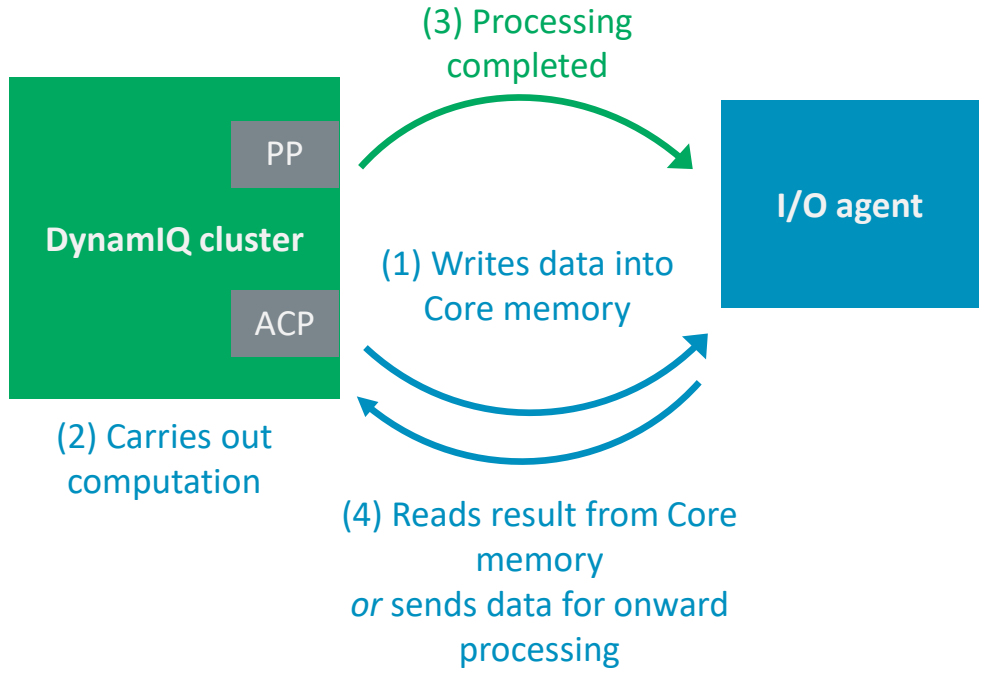
Increasing performance through tight integration

Offload acceleration



Example application:
Offload crypto acceleration

I/O processing



Example application:
Packet processing in network systems

© 2017 Arm Limited
arm

Automotive and industrial safety and reliability

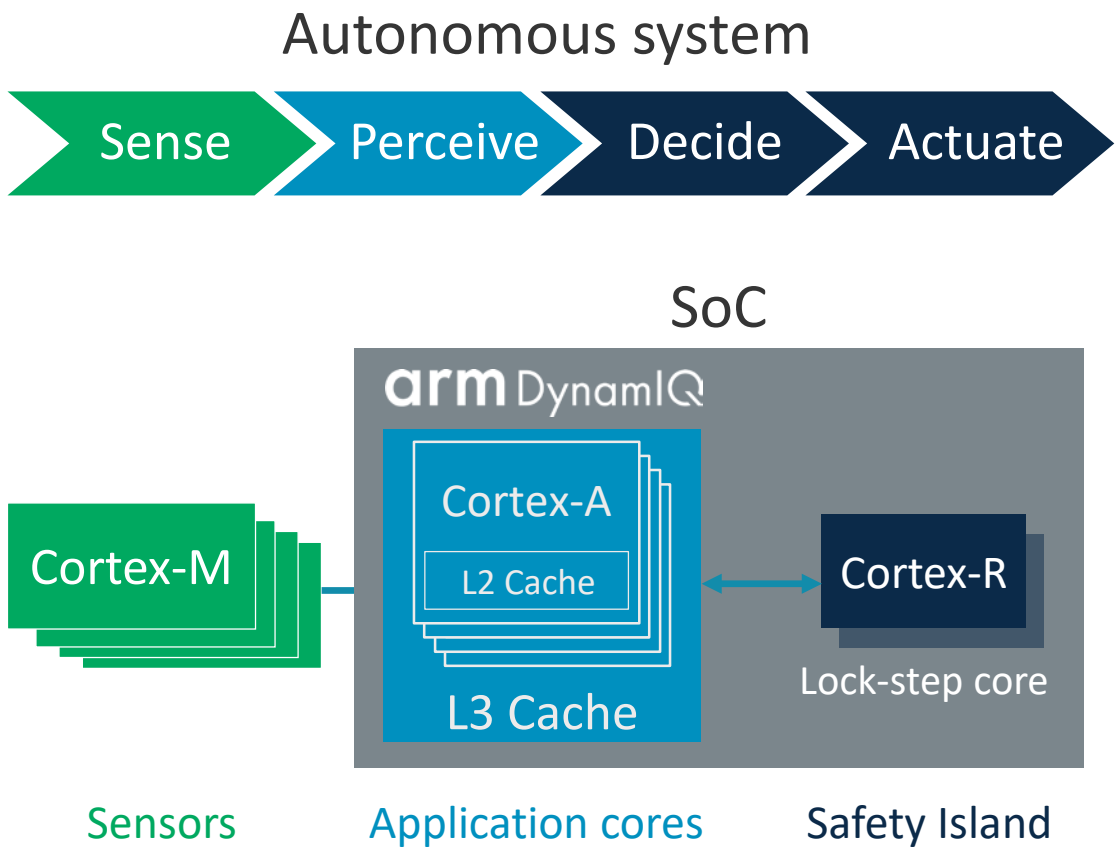
ADAS and IVI compute performance

- DynamIQ provides performance required for autonomous cars
- Faster responsiveness

DynamIQ: Functional Safety

- Following ASIL D systematic flow
- Provides higher safety integrity

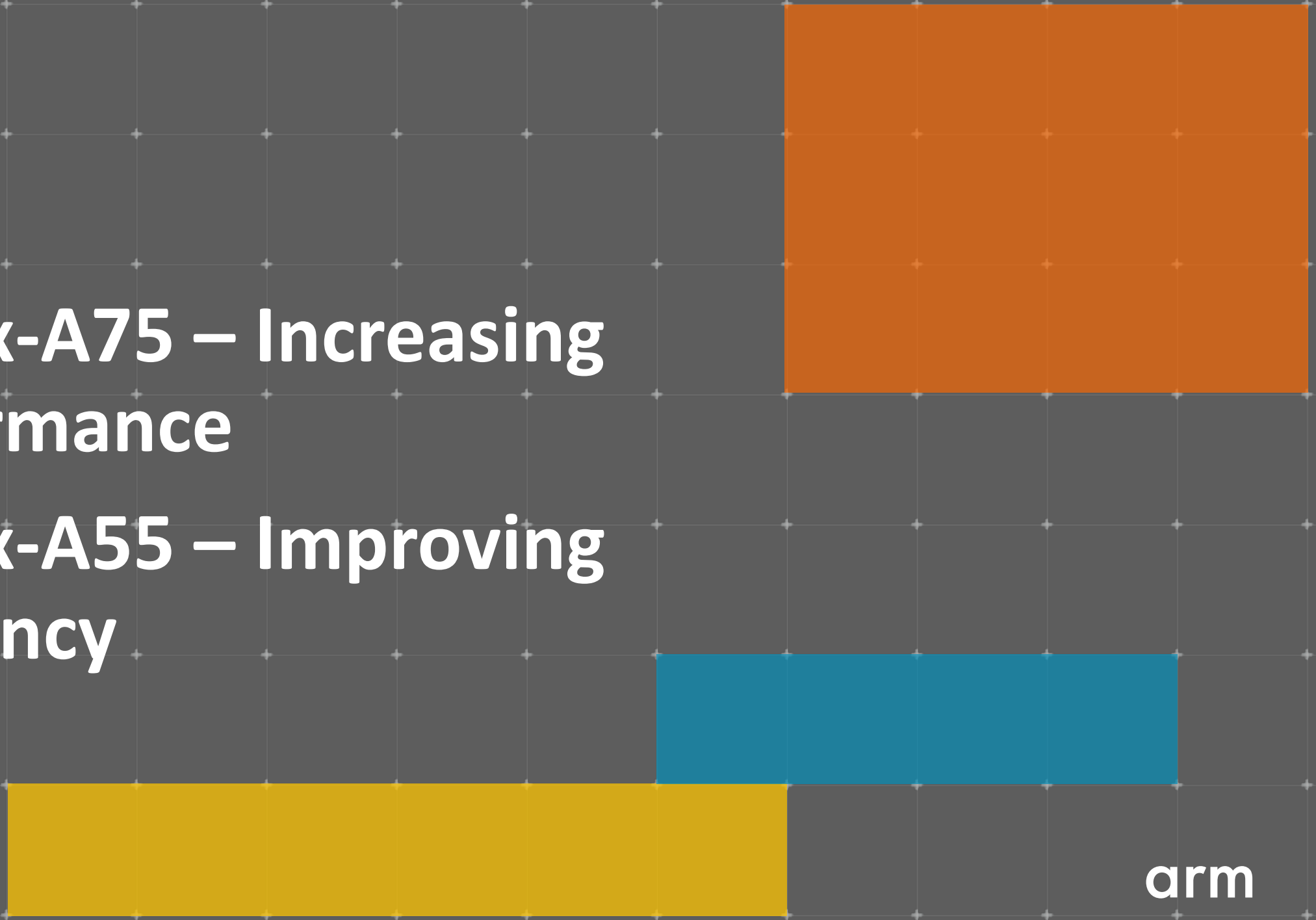
Industry's broadest functional safety capable CPU portfolio



© 2017 Arm Limited
arm

Cortex-A75 – Increasing Performance

Cortex-A55 – Improving Efficiency



© 2017 Arm Limited

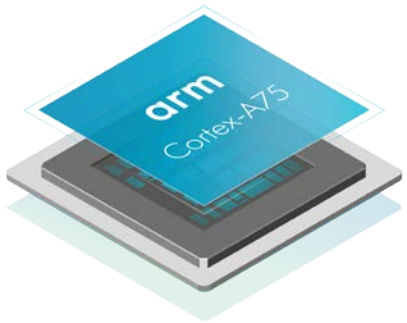
arm

New levels of performance for smart solutions

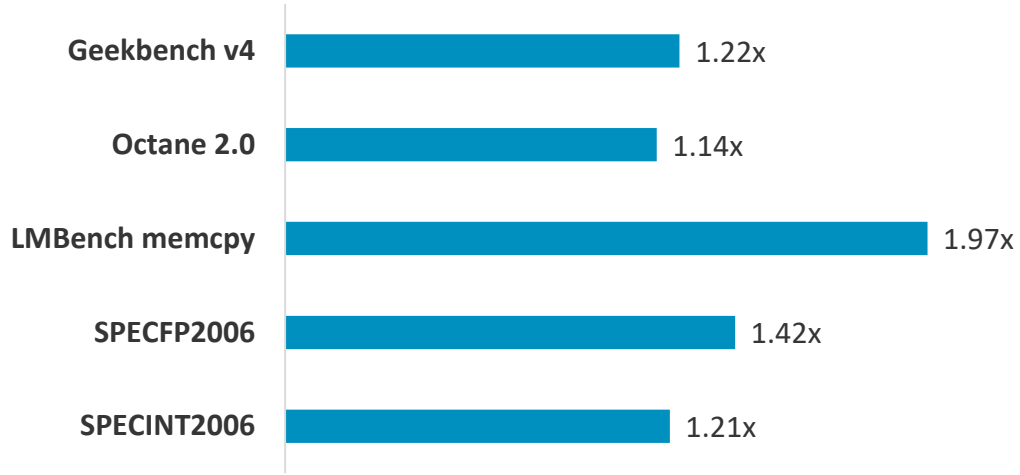
Cortex-A75

Baseline to Cortex-A73

All comparisons at ISO process and frequency

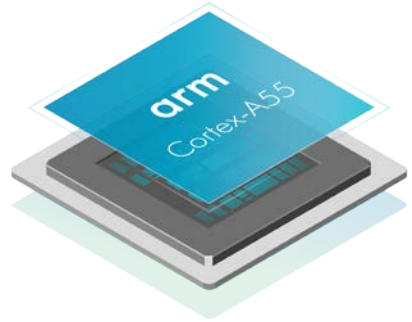


Cortex-A55



Baseline to Cortex-A53

All comparisons at ISO process and frequency



© 2017 Arm Limited

arm

Architecture and Pipelines

Common features

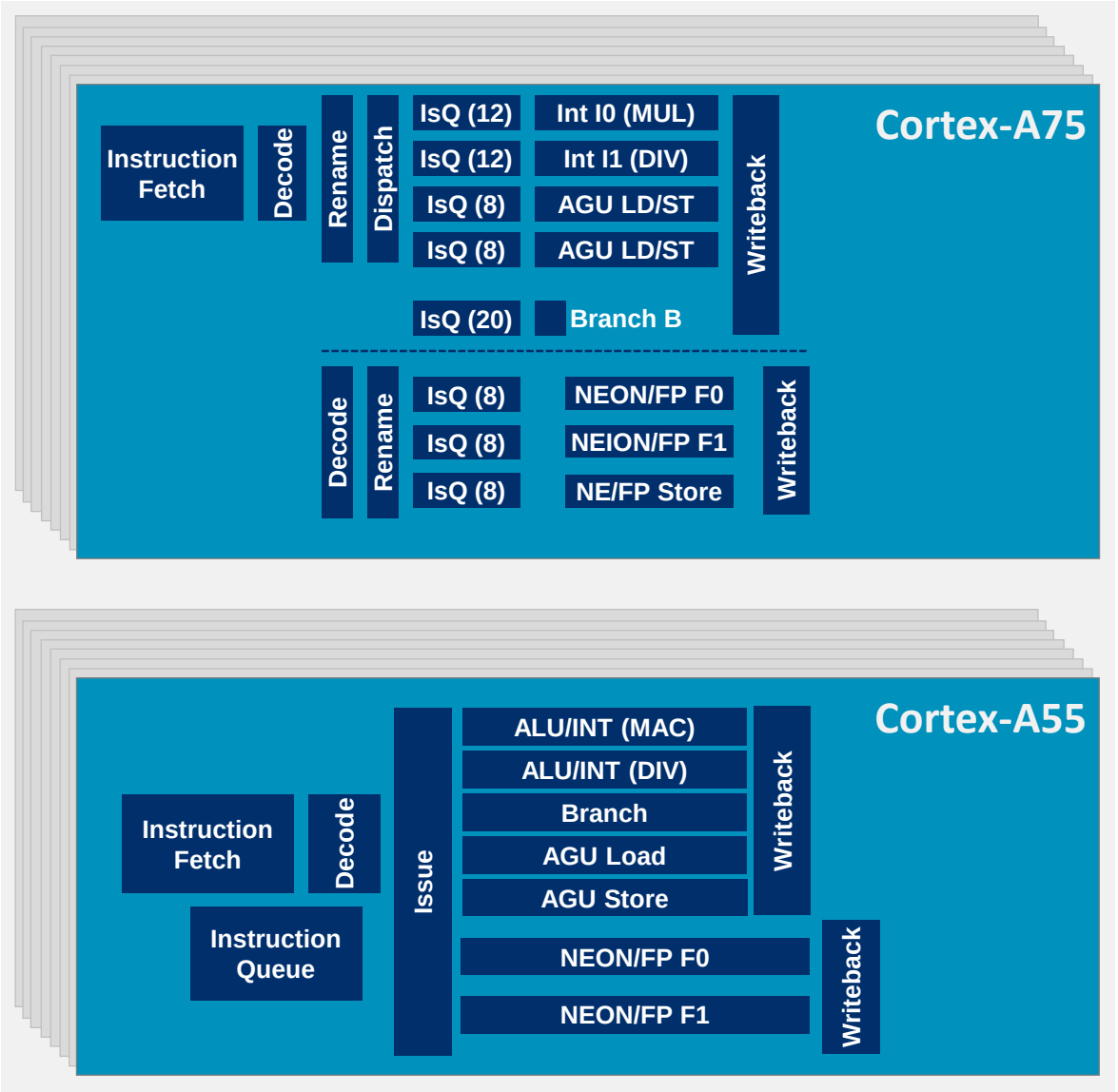
- Armv8.2-A Architecture
- DynamIQ big.LITTLE

Cortex-A75 – performance focussed

- Out-of-Order, 11-13 stage integer pipeline

Cortex-A55 – efficiency focussed

- In-order, 8 stage integer pipeline



© 2017 Arm Limited
arm

Instruction fetch

Common features

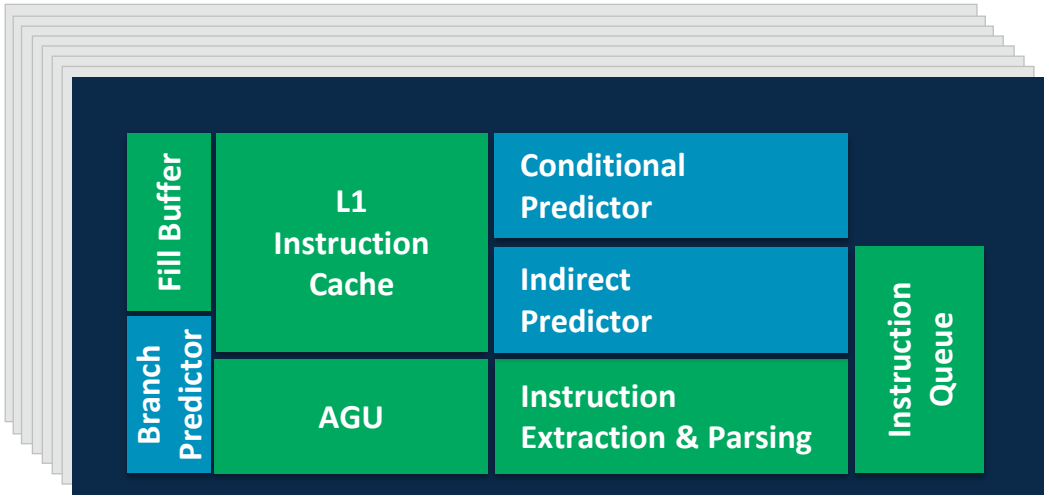
- 4-way set associative
- Virtually indexed, physically tagged (VIPT)
- Decoupled from Cores thru instruction queue

Cortex-A75

- 64KB
- 4-wide instruction fetch

Cortex-A55

- 16KB / 32KB / 64KB
- 2-wide instruction fetch



© 2017 Arm Limited
arm

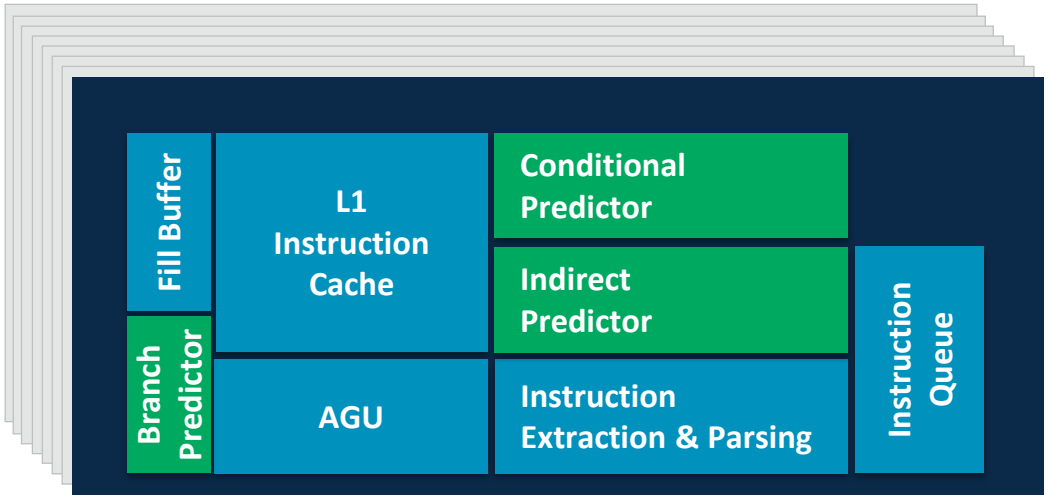
Branch prediction

Cortex-A75

- Fine-tuned 0-cycle prediction
- State of the art, mobile focussed, table based conditional prediction

Cortex-A55

- Brand new 0-cycle predictors
- New main conditional predictor - Neural network based
- New loop predictors



© 2017 Arm Limited
arm

Cortex-A75: Datapaths

3-way superscalar high-performance pipeline

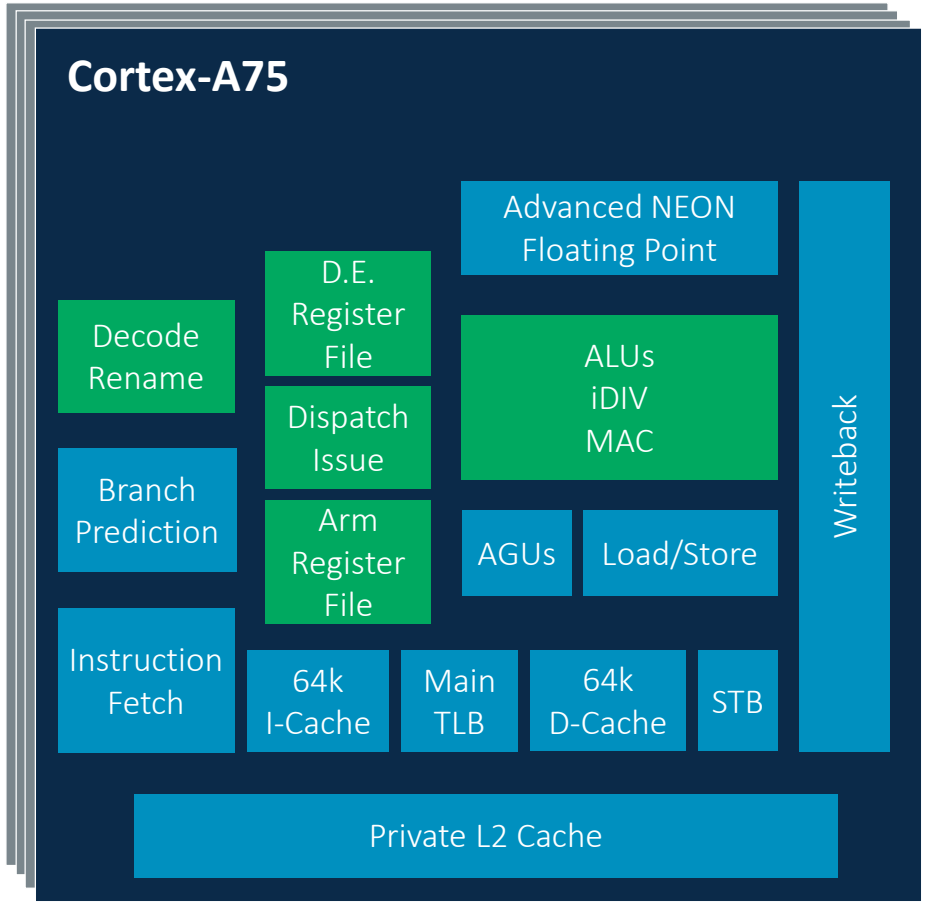
- Single-cycle decode with instruction fusing and micro-ops

7 independent high-performance issue queues

- 2x Load/Store, 2x NEON/FPU, 1x Branch and 2x Integer core

Increased capacity to sustain operation under L1 miss / L2 hit

- 12 entries for integer core to maximise on in-flight instructions and out-of-order capabilities
- 8 entries for Load/Store and NEON/FPU



© 2017 Arm Limited
arm

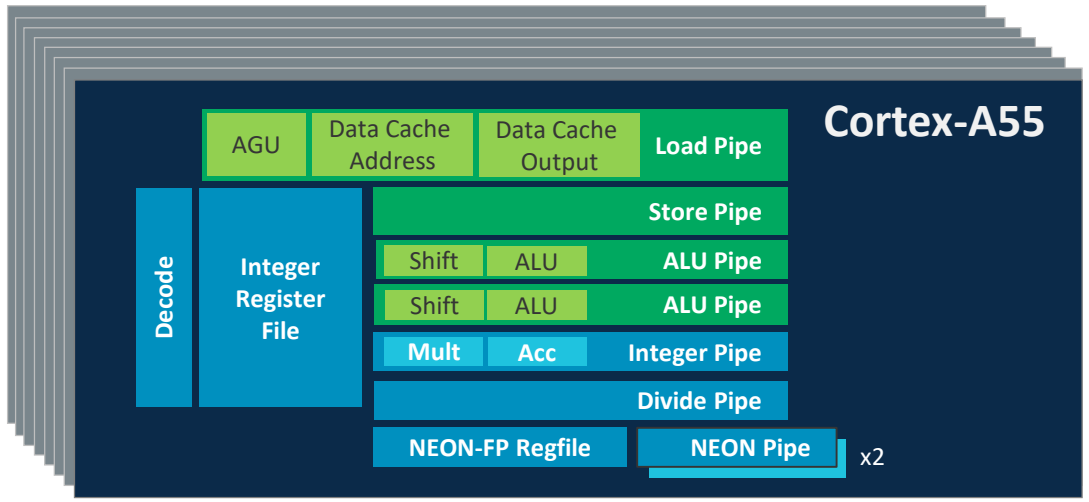
Cortex-A55: Datapaths

Dual issue of loads and stores

Improved latency for forwarding ALU results to the AGU

- Reduced by one cycle for many common ALU operations

Reduced L1 cache load-to-use latency for pointer chasing to two cycles



© 2017 Arm Limited
arm

L1 memory system

Common features

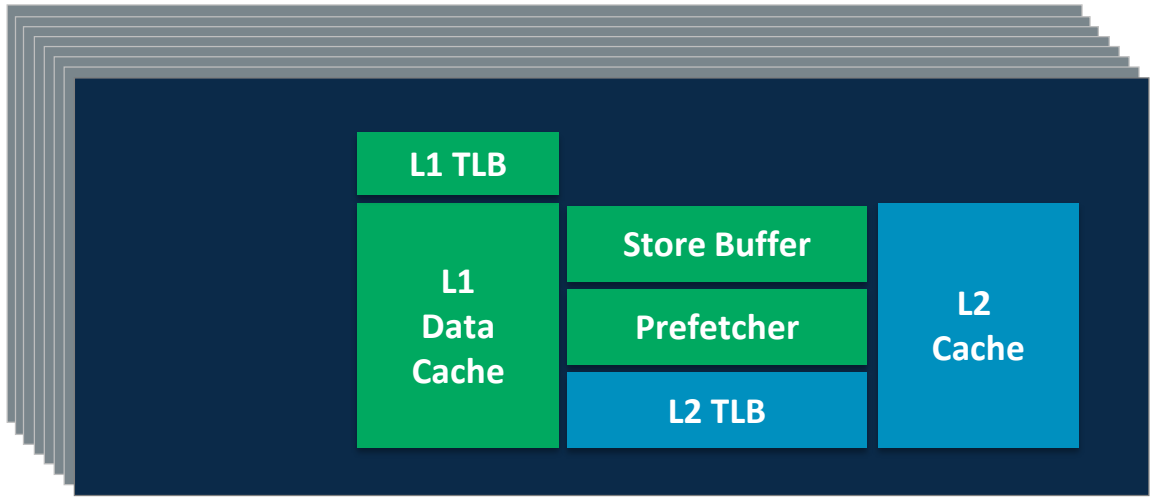
- 4-way set associative
- VIPT with PIPT programmer's view
- Improved prefetchers

Cortex-A75

- 64KB
- Wider load-store than Cortex-A73
- Support Read-after-Write OoO with filtering

Cortex-A55

- 16KB / 32KB / 64KB
- Improved store buffer bandwidth to L1
- Larger 16-entry L1-TLB



© 2017 Arm Limited
arm

L2 memory system

Common features

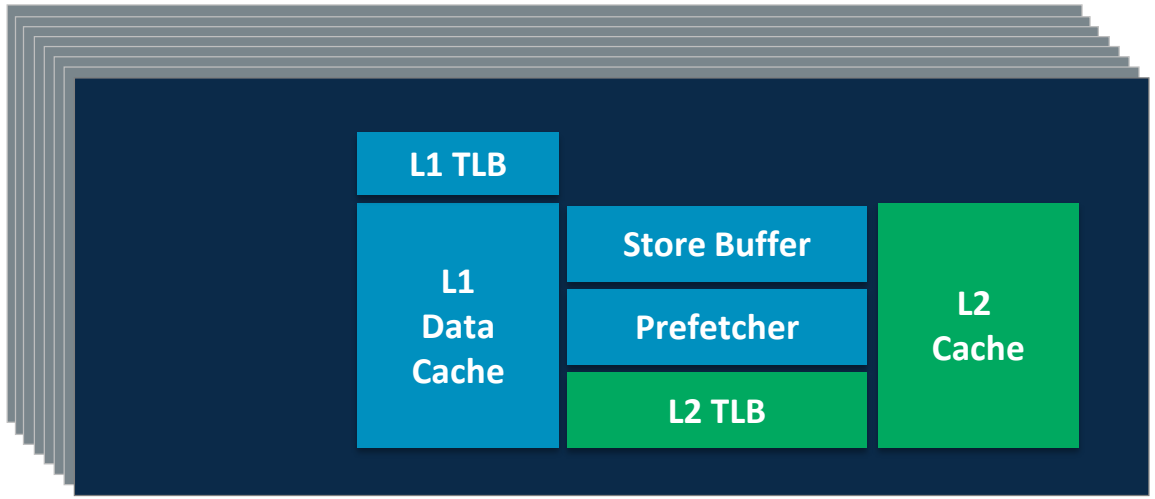
- Private L2 cache in each Core
- Running at Core speed
- Exclusive data cache
- Cache stashing into the L2
- Non-blocking 1024-entry TLB for hit-under-miss

Cortex-A75

- 256KB / 512 KB

Cortex-A55

- 0KB / 64KB / 128KB / 256KB



© 2017 Arm Limited
arm

Next-generation features

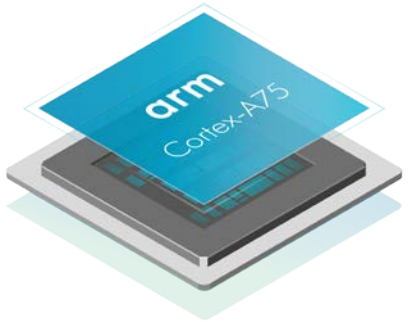
Dot product and half-precision float for AI/ML processing

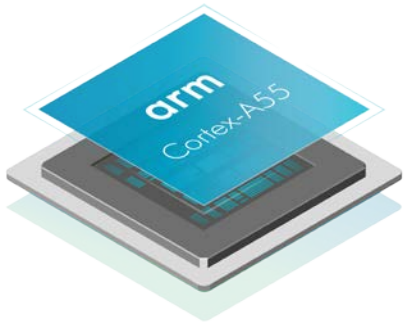
Virtualized Host Extensions (VHE) offering Type-2 hypervisor (KVM) performance improvements

Cache stashing and atomic operations improves multicore networking performance and improves latency

Cache clean to persistence to support storage class memory

Infrastructure class RAS enhancement including data poisoning and improved error management





© 2017 Arm Limited
arm

Innovating for the scalable future

2013-2017: The nature of compute is changing the landscape

Expanding Arm technologies for broad market applicability

New cluster design with new DynamIQ cores:

- Cortex-A75: Breakthrough performance
- Cortex-A55: Efficiency redefined

Functional safety for industrial and automotive applications

New features expanding microarchitecture capabilities:

- DynamIQ Shared Unit , new cache features, new branch prediction

© 2017 Arm Limited

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

© 2017 Arm Limited

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

 © 2017 Arm Limited