MPSoC'13

Scaling Mobile Compute to the Data Centre

Defining the next era of high-performance computing with continued low-power leadership

The Architecture for the Digital World® ARM

Can what's happening in consumer help HPC?

- Mobile devices are transitioning to the only consumer compute platform required for all computing needs
  - Increased performance, rich 64bit instruction set, GPGPU, scalable interconnect
    Creates the central compute unit – adds screens, graphics, sensors, and communications
- Energy and costs forcing new approaches to servers and networks
  - Requirements for security, lower energy, scalability and TCO, lots more data

Connectivity and Sensors

Memory Addressability

Scalability

Security and Reliability

High-Performance and Energy efficiency

Does the compute really vary?

Servers

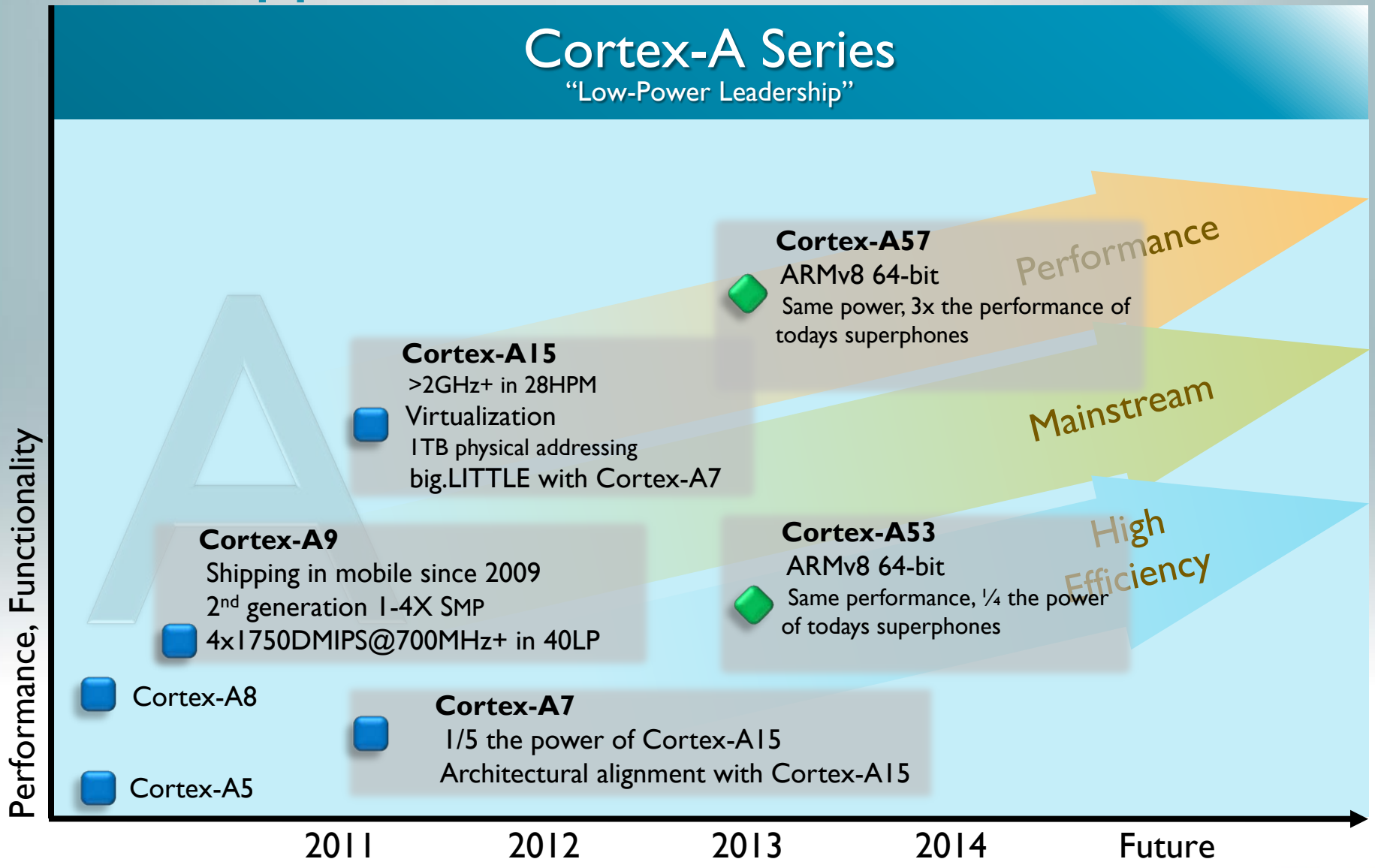
Mobile Computing

Wireless Infrastructure

Connectivity

The Internet of Things

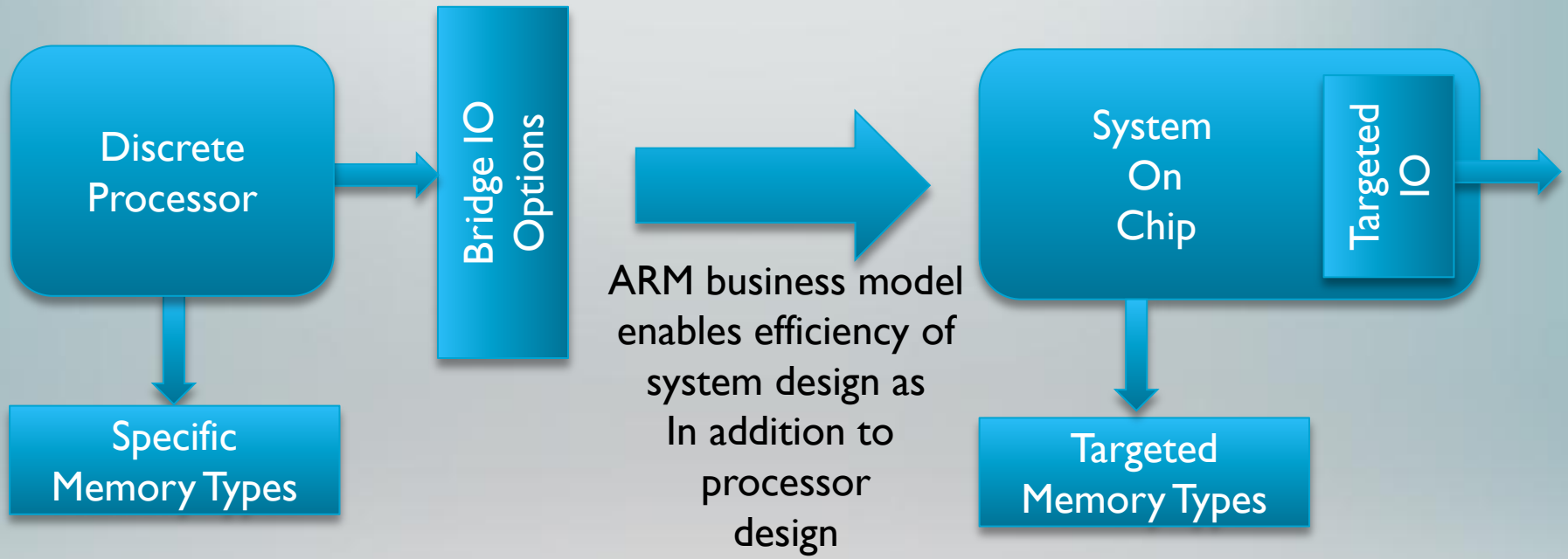
ARM Applications Processors



ARM at the Centre of Compute

- ARM is now delivering three fundamentally different micro architectures supporting the same software Architecture
  - Providing design flexibility, highest performance and lowest power SoC
- Partners construct 100's of different SoC each year based on these processor cores
  - Differing in scalability of compute, IO, memory and peripheral
  - There is a growing number of SoC being designed for server applications
  - A couple of SoC are been designed specifically for HPC
- As the number of server devices increase, and the implementation technology costs reduce, more devices suitable for specific HPC applications are likely
  - Some prototypes already happening

Applicability of a Targeted Solution



- One-size can not fit all applications
  - Compromise across all important metric, compute vs memory vs IO
  - Energy overhead from abstractions (eg PCIe) between IO and CPU
- ARM model supports hundreds of companies building hundreds of application optimized System on Chip solutions
  - The challenge is ROI – cost of development vs. market opportunity

Starting point: Instruction Set with ARMv8

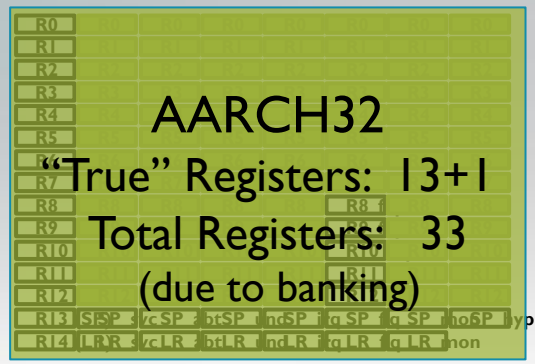
- Fully compatible with existing ARMv7 32-bit code
- Opens ARM to new applications
- AARCH**32:** Evolution of 32-bit
  - Ideal for concurrent programming C11, C++11 Java5
  - More efficient, high-performance thread-safe software
  - Enhanced security and encryption
- AARCH**64:** Efficient 64-bit execution
  - Clean instruction set
  - Modern compiler friendly
  - Reduced complexity for operating systems, hypervisors
  - Designed to maximize reuse of existing hardware

Cortex-A57 supports dual FMAC unit with SIMD offering up to 4 DP FLOPS per clock for HPC use

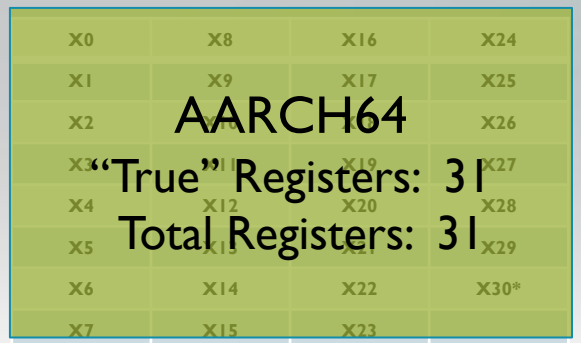
AArch64: Low-Power 64-bit Execution

- 64-bit execution state designed for low-power
  - Fixed length decoding, simplifying and optimizing front end pipeline
  - No legacy allows cleaner architecture implementations

- Allows maximized reuse of existing hardware

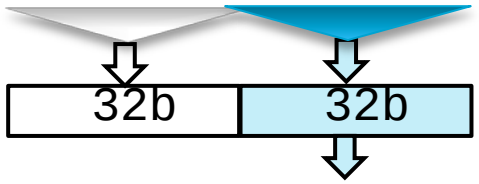


AARCH32
"True" Registers: 13+1
Total Registers: 33
(due to banking)



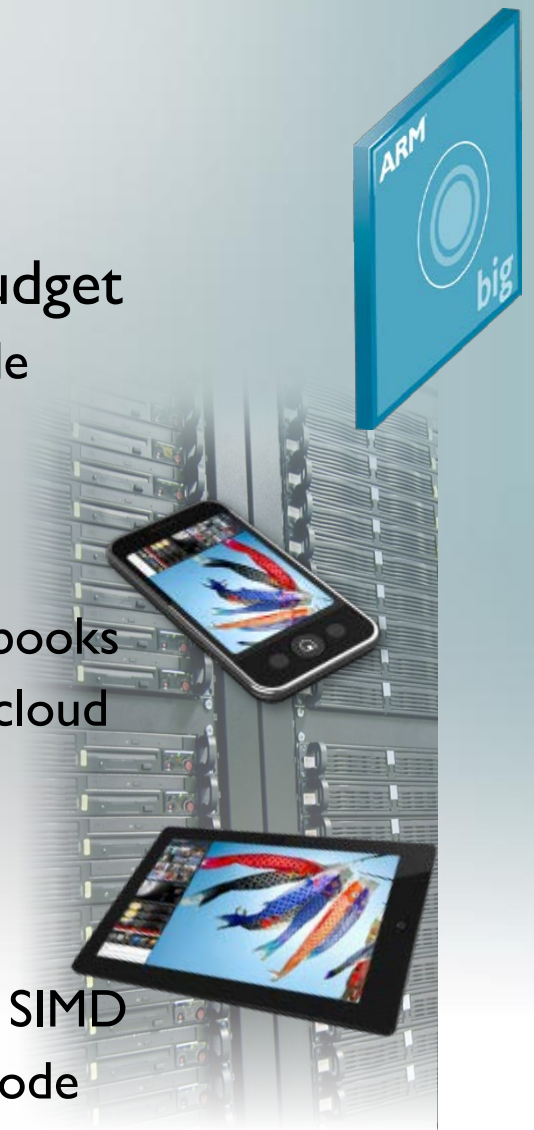
AARCH64
"True" Registers: 31
Total Registers: 31

- Maximization of use of 32-bit values, registers (common)



Cortex-A57: Implementation for High Performance

- Maximum performance in smartphone power budget
  - 3x performance of 2012 super phones, in 32-bit mode
  - Performance per clock comparable to today's PC's
- Driving advanced computing
  - 5x power-efficiency for tomorrow's tablets and notebooks
  - Mobile-level power consumption for enterprise and cloud
- Enhanced capabilities for enterprise and servers
  - 64-bit support for full range of applications
  - High performance IEEE compliant DP floating point / SIMD
  - Solutions already scaling up to a 16x SMP compute node
  - **SoC flexibility** to scale **efficiently** to 100K's or more nodes

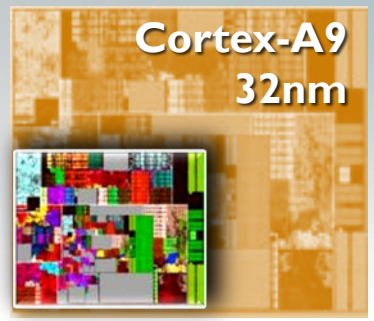


Cortex-A53: Delivering More For Less

- Maximum performance, mass-market cost
  - Cortex-A53 delivers performance of Cortex-A9
  - 40%+ smaller in same process, including 64-bit support

Cortex-A53

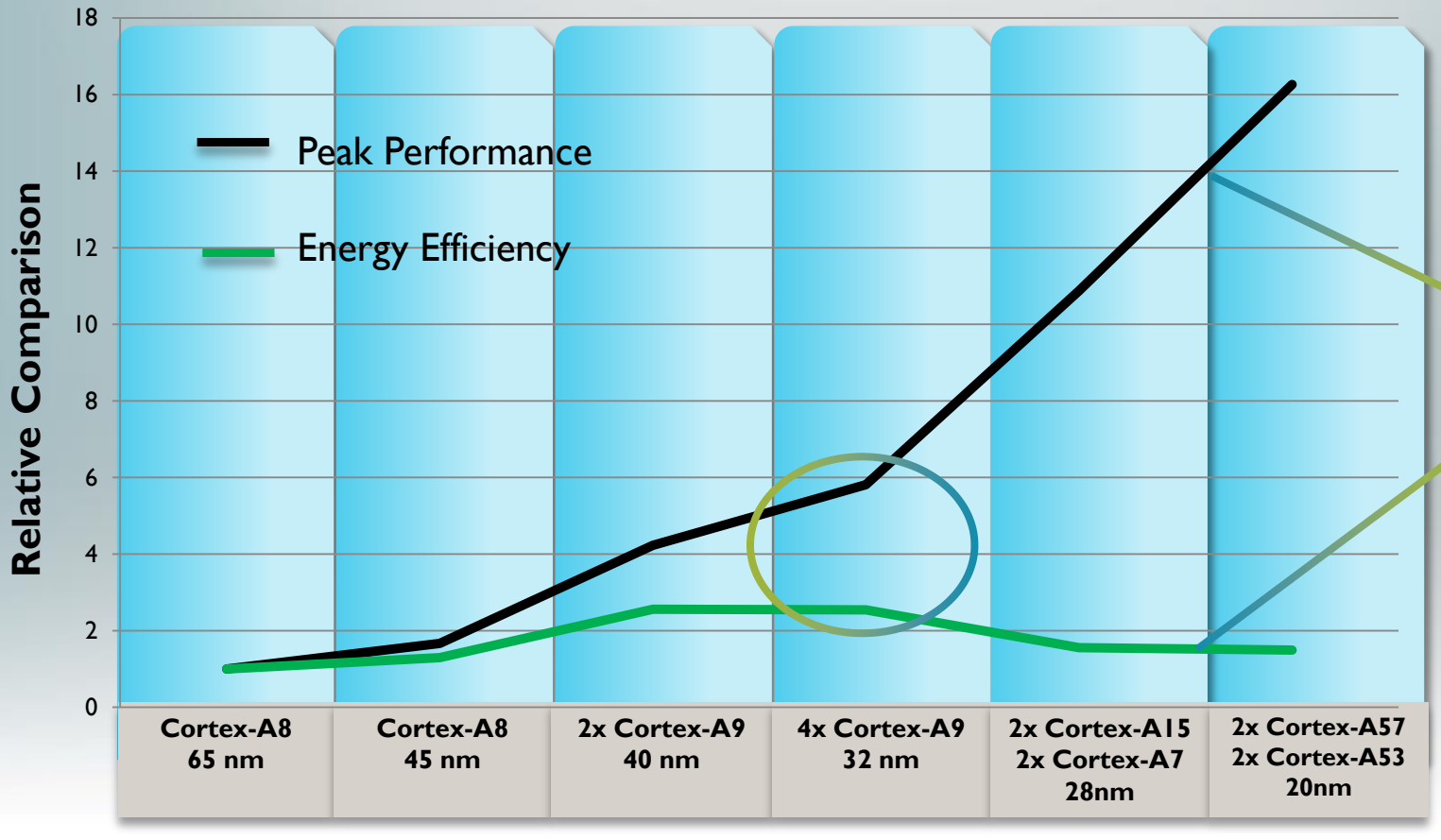
- Outpacing Moore's Law
  - 4x as efficient for matched performance
  - 25% the size of mainstream superphone CPU
  - Exceeds the performance of 1st generation ARM servers



Cortex-A53,20nm

- Enables disruptive server solutions
  - Full 64-bit architecture support, flat GAS up to 48bits
  - Scalable to 16x SMP compute node today
    Also support coherent unified OpenCL accelerated compute (Mali GPGPU)
  - System scalable to millions of nodes

Cortex-A50: High Performance, Lower Power



big.LITTLE Effect

More Performance

Less Energy

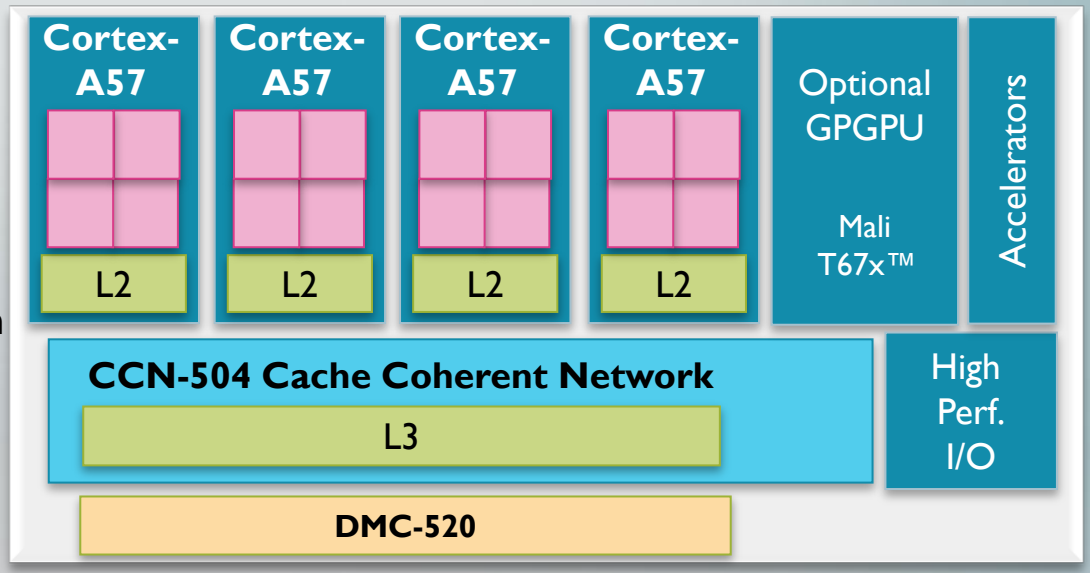
- ARM big.LITTLE enables both peak performance and low power
- Innovation "Beyond-Moores-Law" process technology

Efficiently Scaling Upwards: The SMP node

Maximum per-thread performance

- Highest peak performance
- High-performance IO
- Multi-cluster solutions: 16+ cores
- Process technology: 28nm down to 14nm

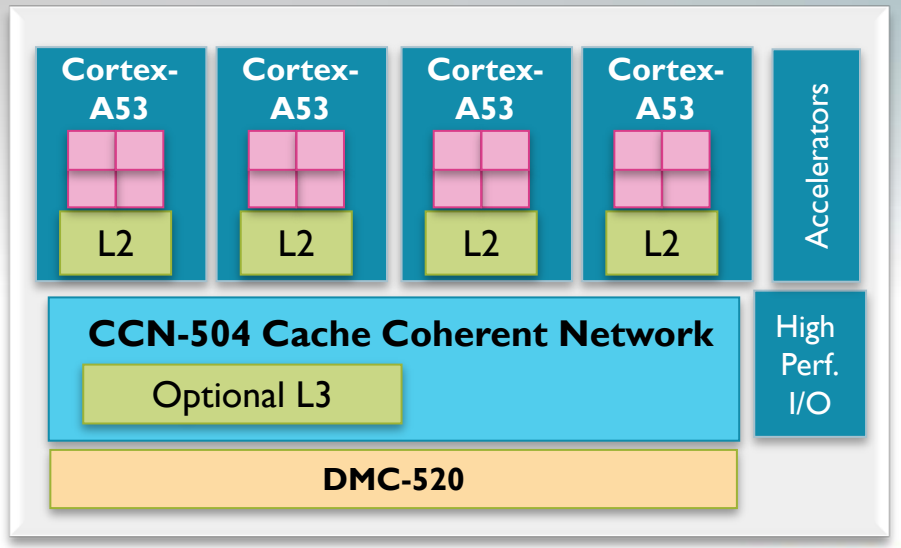
Key applications: macro-basestations, servers, compute heavy HPC



Throughput optimized

- Lower total power and size
- High aggregate performance
- Maximizing throughput/mm2, mW
- 16 cores, scaling to sea-of-cores

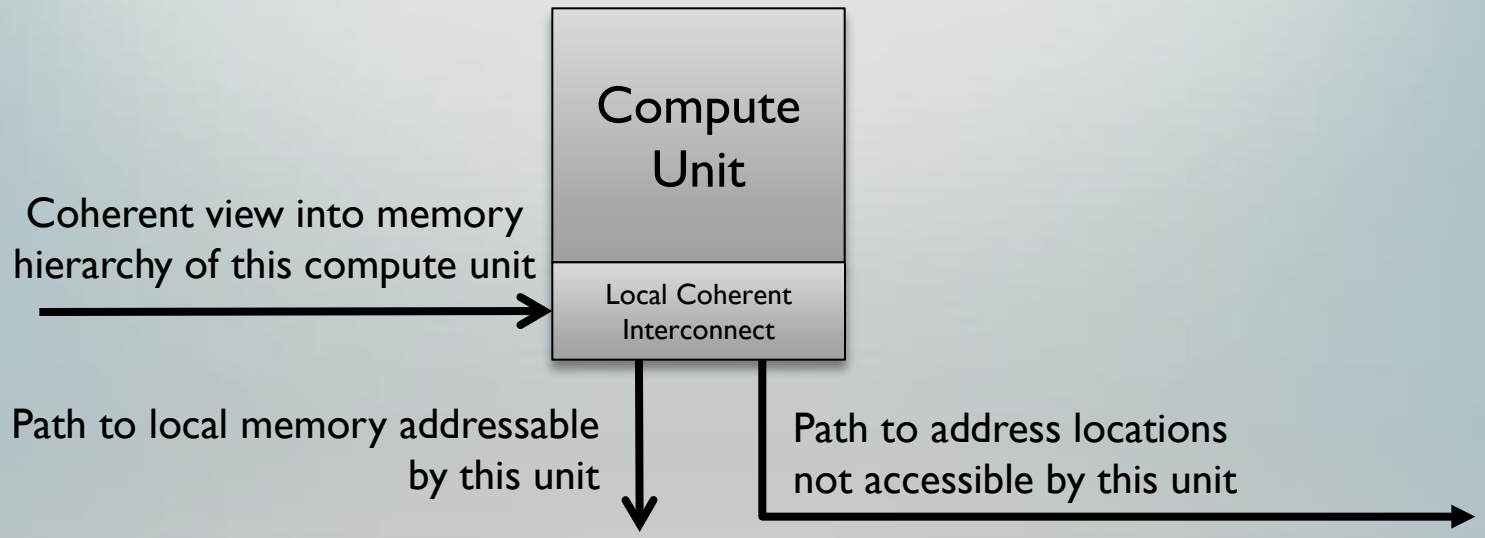
Key applications: cell basestations, dataplane compute, scale-out servers, control and search heavy HPC



Scaling Compute: Unit of Compute Concept

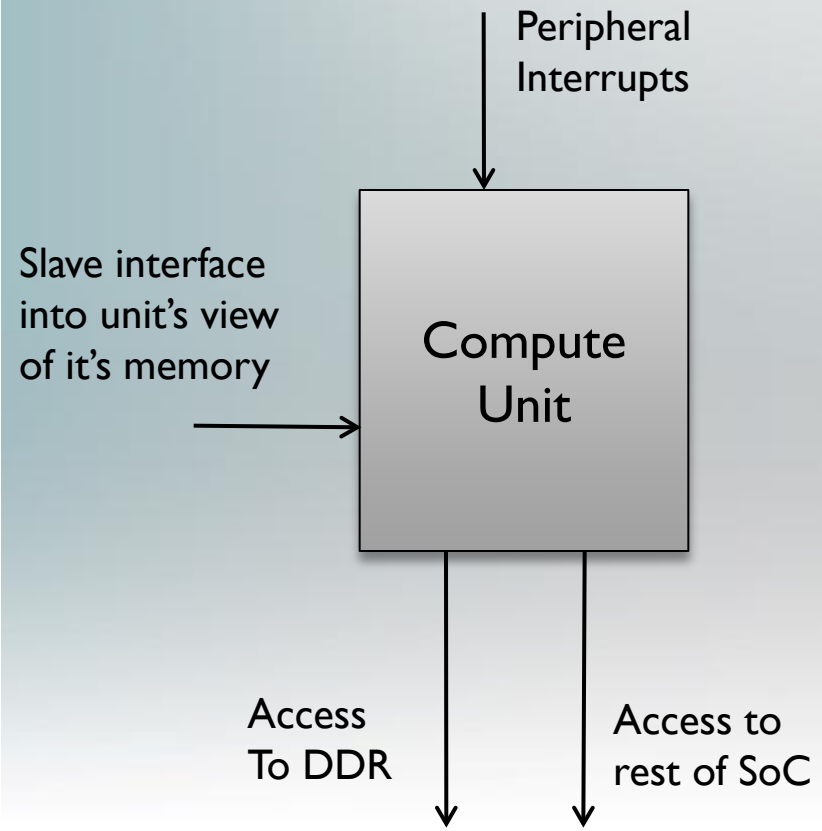
- A Unit of Compute is managed by a single (SMP) operating system operating within an coherent region of memory
  - Therefore manages all associated peripherals, accellerators
  - (For clarity, ignoring Numa and distributed OS – for now)
- It can range from a single CPU core up to a multi-core, multi-cluster design
  - Its size is defined by the target application's ability to utilize the SMP resources the unit manages
- Conceptually a Compute Node provides to a SoC:
  - The Unit of Compute (mix of integer, float, SIMD, GPGPU, and accelerators)
  - Access to its local memory (limited by device pin count)
  - Access to the rest of the remote memory (limited by system interconnect)
  - A path into the unit that allows other unit to have view into its memory

Logical Structure of a Compute Unit



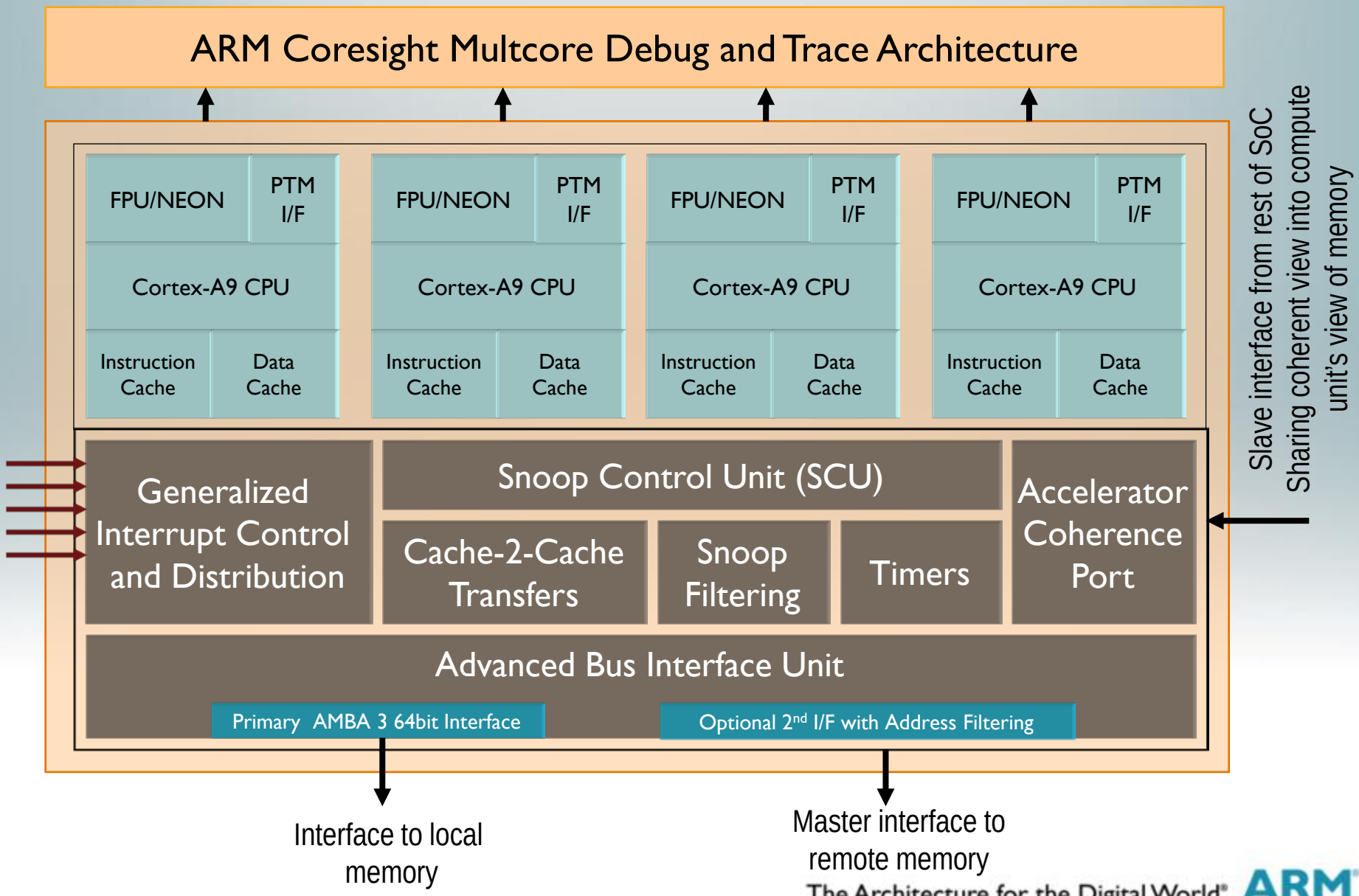
- Unit can include any number of compute resources
  - Potentially both general purpose and other local accelerators
- Provides coherent and symmetric access across local resources
  - Enabling a SMP capable operating system and resource sharing
- Each Compute Unit is assigned a partition within a system's global address space (GAS)
  - Any unit can coherently access any location in the GAS
  - DMA can master transfer between units

ARM Units of Compute for 2013

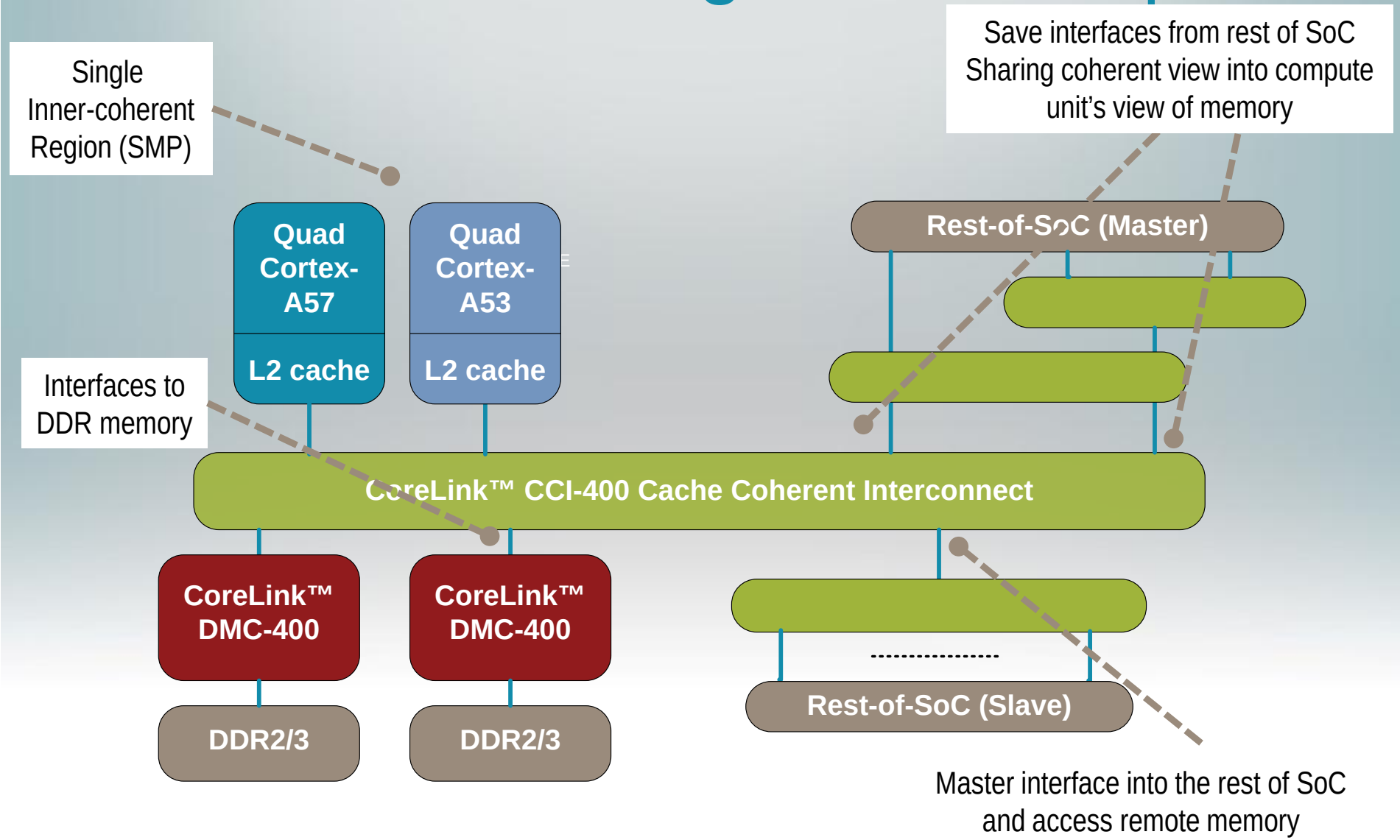


- In a system, various ARM IP blocks enable the construction of a Compute Unit from a single Cortex-A5 up to a 16-way Cortex-A57
- Such systems can support multiple software models due to supporting a coherent, virtually mappable memory across a global address space
  - Eg Shared memory MPI
  - Capable of remote (off chip) memory access latencies under 100ns

Cortex-A9 MPCore: The first "Compute Unit"



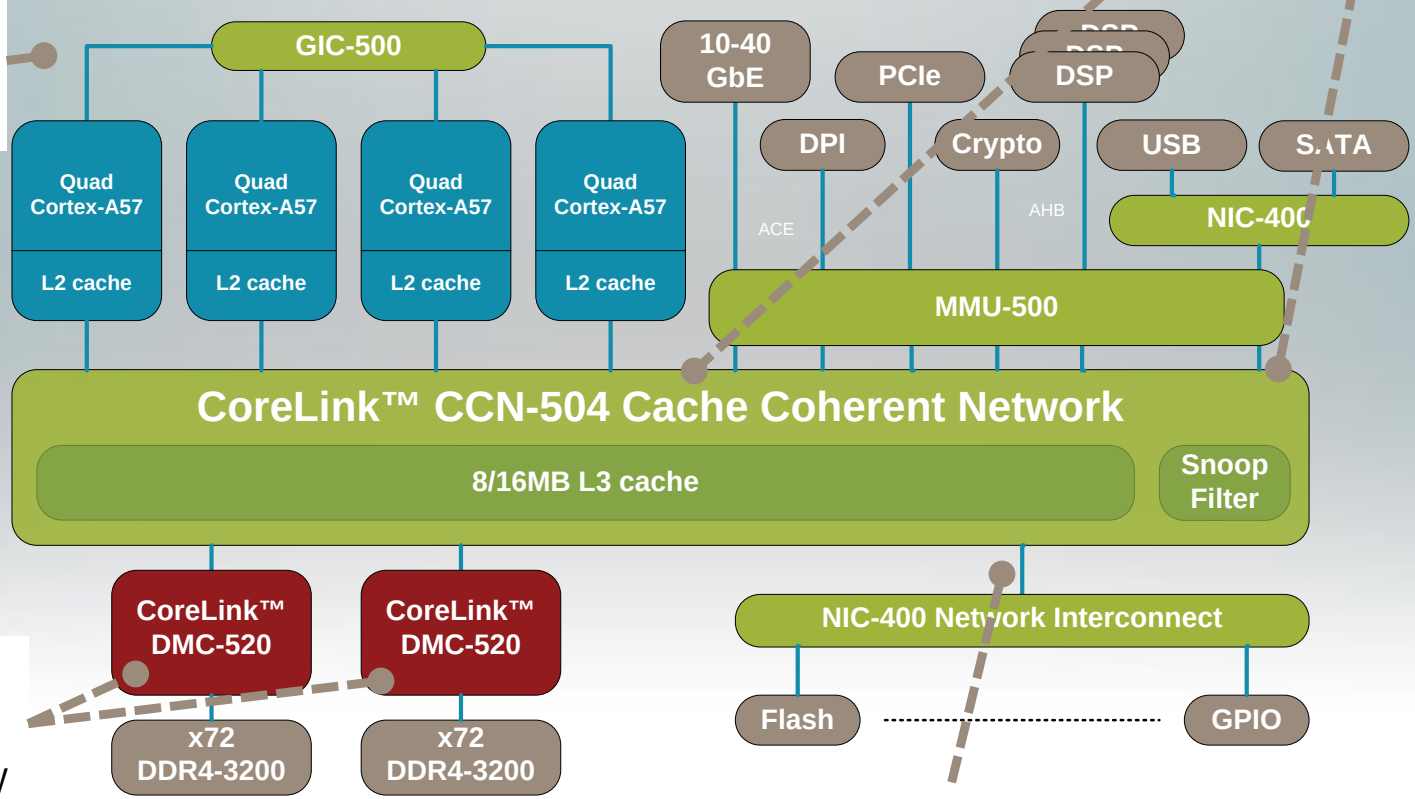
CoreLink CCI-400: big.LITTLE Compute



CoreLink CCN-504: Server Unit of Compute

Slave interfaces from rest of SoC
Sharing coherent view into compute unit's view of memory

Single Inner-coherent Region (SMP)



Interface to Local DDR memory

Master interface to the rest of SoC and access remote memory

Summary

- ARM does not build chips – we design the instruction set architecture and a few processor implementations (CPU/GPU)
  - Wide range of different micro architectures allowing device manufactures maximum flexibility in building application-targeted solutions
  - Architecture licensees (eg nvidia) can also build ARM compatible solutions
- The flexibility of SoC can remove interface abstractions and integrate key interface and accelerator components on chip
  - Removing the power over heads, interface bottlenecks and unnecessary abstraction associated with general purpose devices
- ARM compute units open a reusable and scalable approach to building power optimized, high performance solutions
  - Quickly gaining traction in servers
  - Supporting the ARMv8 Architecture suitable for HPC applications