ARM Neoverse N1

Aishik Ghosh, Anant Kandikuppa, Zane Fink

CS 433 Mini Project

Outline

1. Introduction
2. Core Microarchitecture
3. Memory Hierarchy
4. Parallelism Support
5. Special Features for Infrastructure and Security

Introduction

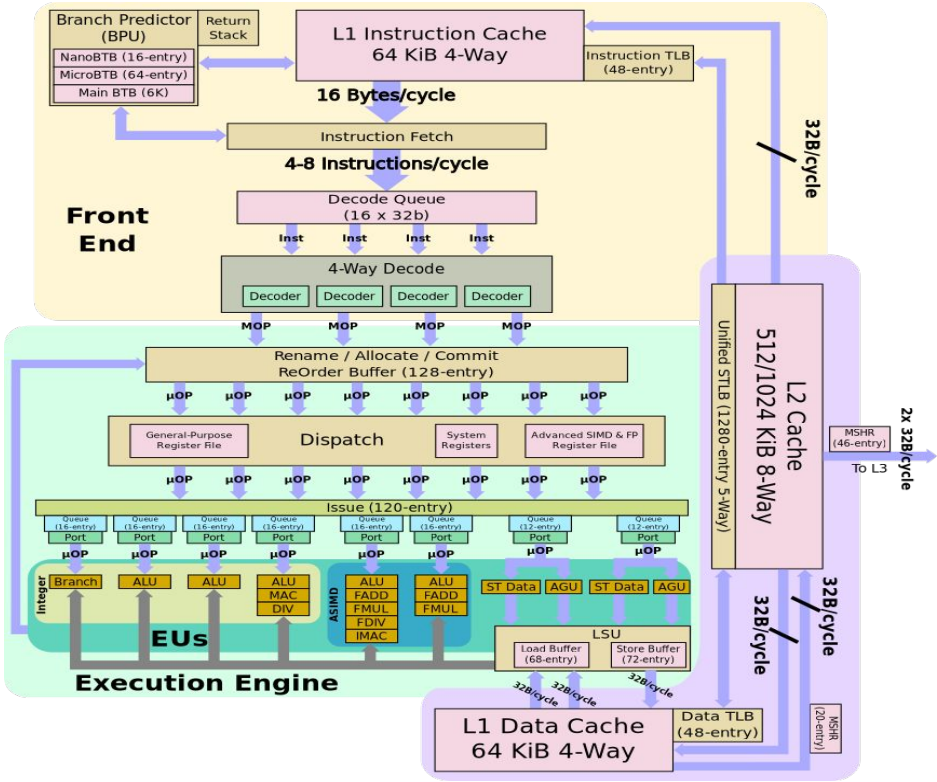
- ARM CPUs are RISC processors designed by Arm Holdings
- ARM Neoverse N1 is an infrastructure-focused chip that implements the Arm v8.2-A instruction set
- Hardware features make the Neoverse N1 especially suited to cloud and infrastructure applications

Introduction: Infrastructure

- Workloads commonly found in cloud environments such as AWS, Microsoft Azure, etc.
- Codes characterized by:
  - Complex branching behavior
  - JIT compiled
  - Object management
  - Garbage collection
- Concretely: web servers, datacenter applications

Core Microarchitecture

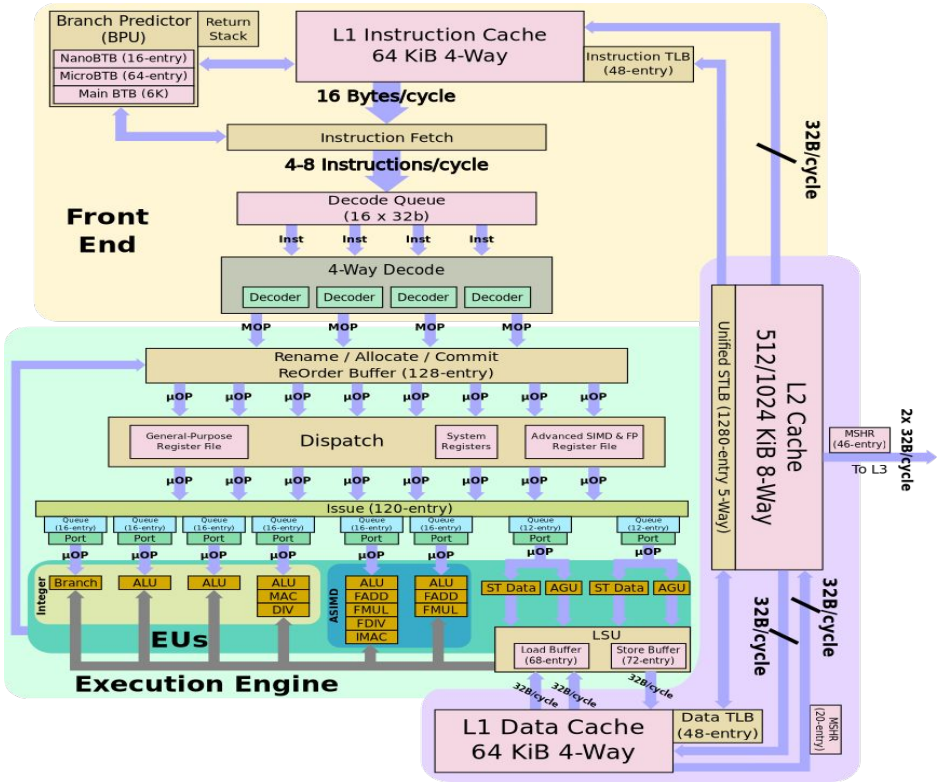
- Superscalar processor
- 11-stage out of order accordion pipeline
  - Can drop to 9-stages



[WikiChip - Neoverse N1 Microarchitecture]

Core Microarchitecture

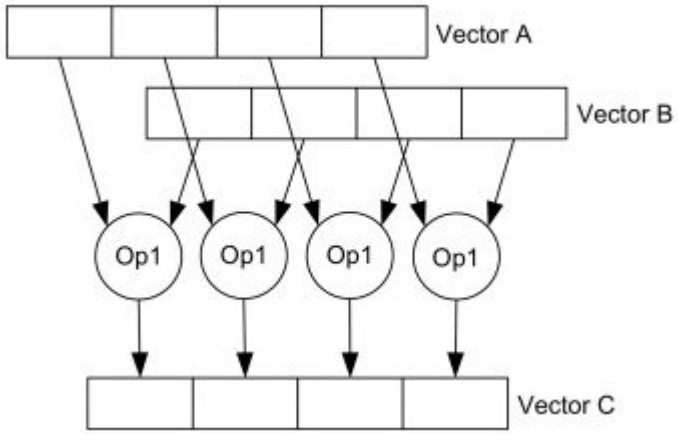
- Core has
  - 4 way decode
  - 3 ALUs
  - 1 branch exec unit
  - 2 adv SIMD units
  - 2 load/store exec unit



[WikiChip - Neoverse N1 Microarchitecture]

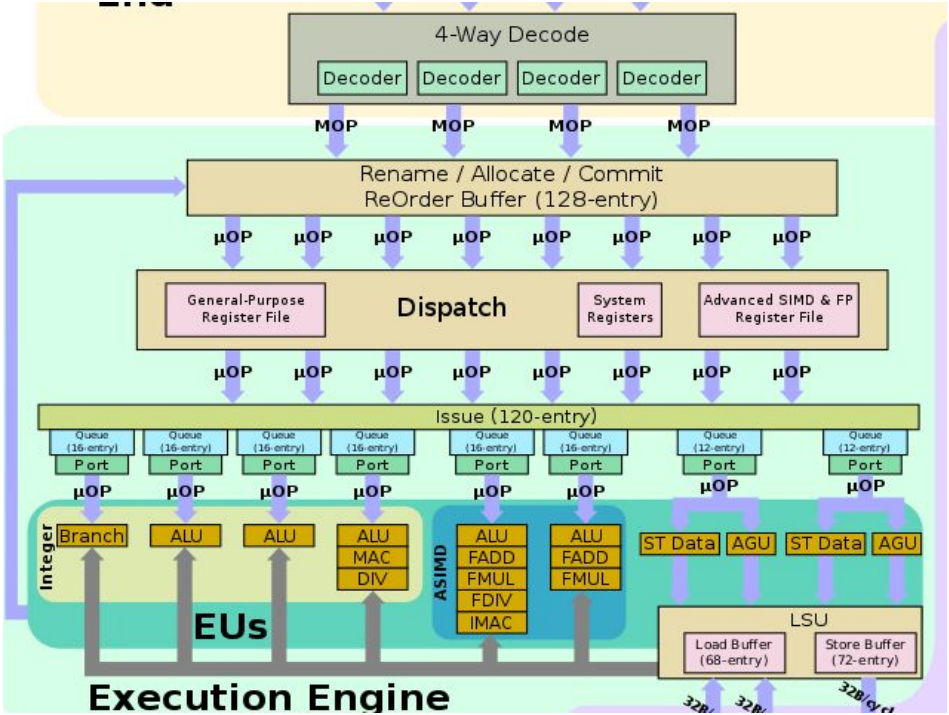
SIMD

- Single instruction multiple data
- Adv SIMD performance
  - native quad-word, dual 128-bit units fed by separate issue queues
  - Latencies
    - Floating point add - 2 cycles
    - Floating point multiply - 3 cycles
    - Floating point multiply add - 4 cycles



SIMD

Core Backend



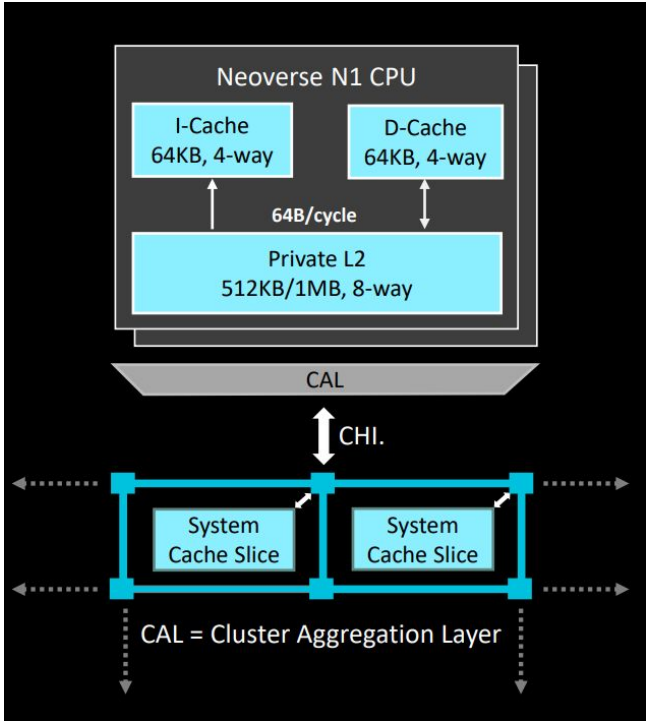
[WikiChip - Neoverse N1 Microarchitecture]

Branch Prediction

- 6k-entry main branch target buffer
- 3-cycle access latency to retrieve branch target addresses (without I-cache)
- Once a prediction is made, the predicted address is stored into a 12-entry fetch queue which tracks future fetch transactions.
- Decoupled Branch Prediction

Memory Hierarchy Introduction

- Deep Cache Hierarchy L1, L2, optional shared L3, and an optional system-level cache.
- Strict inclusivity between L1 data cache and L2 cache, non-inclusivity between L1 instruction cache and L2 cache



[Arm Neoverse N1 Cloud-to-Edge Infrastructure SoCs]

Memory Hierarchy Overview

- **Instruction Cache (Private):**
  - 64 KB
  - 4-way set-associative
  - 16B of instructions per cycle
  - 4-cycle LD-use latency.
- **L1 Data Cache (Private):**
  - 64 KB
  - 4-way set-associative
  - 32B per cycle
  - 4-cycle LD-use latency

Memory Hierarchy Overview

- **L2 Data Cache (Private):**
  - Configurable, 256KB-1MB in size
  - 8-way set associative
  - 64B per cycle between L1/L2 caches
  - 9-11-cycle LD-use latency

Memory Hierarchy Overview

- **(Optional) L3 Cluster Data Cache:**
  - Multiple cores can be configured in a cluster
  - Up to 2MB, 28-33 cycle LD-use latency
  - Modified Exclusive Shared Invalid (MESI) coherence protocol
  - Coherency managed through snoop filter
- **(Optional) System level Cache:**
  - Up to 128MB
  - Coherency managed through snoop filter
  - 22ns LD-use Latency

Memory Hierarchy Overview

- Instruction TLB
  - 48-entry, fully associative
  - Support for 4KB, 16KB, 64KB, 2MB, and 32MB page sizes
- Data TLB
  - 48-entry, fully associative
  - Support for 4 KB, 16KB, 64KB, 2MB, and 512MB page sizes
- Unified (instruction/data) L2 TLB
  - 1280-entry 5-way set associative

Memory Hierarchy Key Features

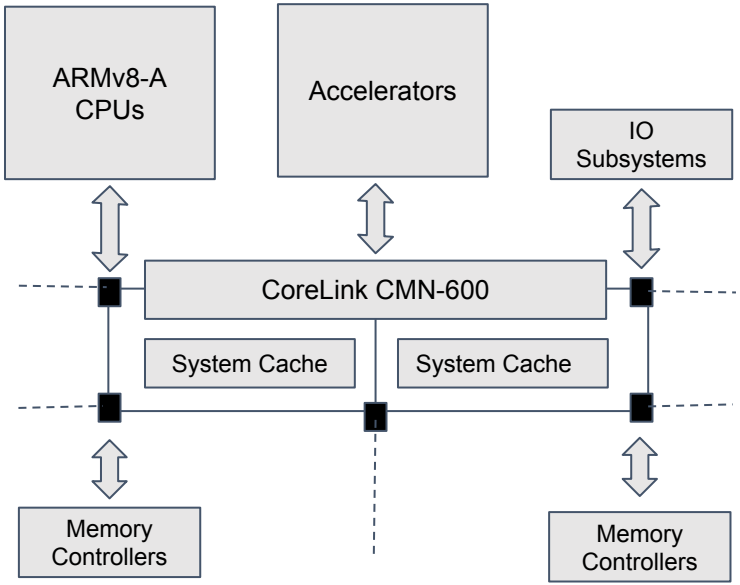
- Cache miss predictor: Bypasses cache hierarchy to avoid cache access times.
- Cache prefetcher that can detect complicated access patterns.
- Fetch-aware cache replacement policies
- Supports maximum physical memory of 256 TB
- 68 in-flight loads, 72 in-flight stores

Parallelism Support

- 4-128 single threaded cores per chip
- Private L1 & L2 caches
- Optional shared L3 cache
- Multiple clusters supported by CoreLink CMN-600 mesh interconnect
- Allows specialized cores and accelerators to work together

Corelink CMN-600 Mesh Interconnect

- Provides high-frequency, non-blocking access to shared memory resources
- Supports upto 32 coherent clusters of CPUs / Accelerators
- Provides upto 128MB of shared system cache (SC)
- Optimized for low memory latency; ~1TB/s bandwidth



CMN-600 Overview

Corelink CMN-600 Mesh Interconnect - SC

- Integrated Snoop Filters
  - Reduces number of snoop requests
  - Lowers power consumption
- Cache Stashing
  - Allows external agents to directly place data into L3/L2 core caches
- Direct Memory Transfer
  - Allows memory controllers to directly send data to the requestor

Infrastructure Specific Features

- Enhanced Virtualization Support
  - VMID extended to 16 bits
  - Hardware update of access/dirty bits in page tables
- Improved Side Channel Protection
- Supports Armv8-A architecture extensions
  - CRC32 instruction to accelerate checksum computation
  - FP16 and int8 dot product support for ML inference
- Commercial deployment - AWS Graviton 2 processor

Thank You!

Questions?

References

[1] R. Christy, S. Riches, S. Kottekkat, P. Gopinath, K. Sawant, A. Kona, and R. Harrison. 2020. 8.3 A 3GHz ARM Neoverse N1 CPU in 7nm FinFET for Infrastructure Applications. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 148–150. DOI:https://doi.org/10.1109/ISSCC19947.2020.9062889

[2] A. Pellegrini and C. Abernathy. 2019. Arm Neoverse N1 Cloud-to-Edge Infrastructure SoCs. In *2019 IEEE Hot Chips 31 Symposium (HCS)*, 1–21. DOI:https://doi.org/10.1109/HOTCHIPS.2019.8875640

[3] A. Pellegrini, N. Stephens, M. Bruce, Y. Ishii, J. Pusdesris, A. Raja, C. Abernathy, J. Koppanalil, T. Ringe, A. Tummala, J. Jalal, M. Werkheiser, and A. Kona. 2020. The Arm Neoverse N1 Platform: Building Blocks for the Next-Gen Cloud-to-Edge Infrastructure SoC. *IEEE Micro* 40, 2 (March 2020), 53–62. DOI:https://doi.org/10.1109/MM.2020.2972222

[4] Neoverse N1 - Microarchitectures - ARM - WikiChip. Retrieved December 2, 2020 from https://en.wikichip.org/wiki/arm_holdings/microarchitectures/neoverse_n1

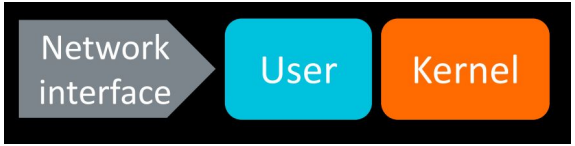
[5] Single Data Multiple Instruction. Retrieved December 6, 2020 from https://www.sciencedirect.com/topics/computer-science/single-instruction-multiple-data

[6] Corelink CMN-600 - ARM - Retrieved December 6, 2020 from https://developer.arm.com/ip-products/system-ip/corelink-interconnect/corelink-coherent-mesh-network-family/corelink-cmn-600

Appendix

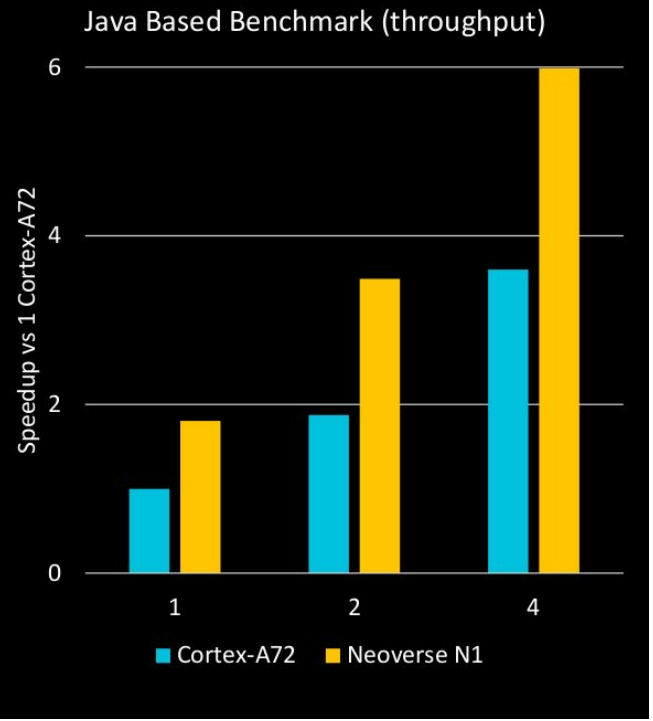
Concrete Example Application

- Nginx: an open-source high-performance webserver, proxy, and load balancer
- Goal: Reduce latency of a single request, maximize number of concurrent requests that can be serviced
- Quickly servicing a request requires:
  - Low memory latency and high bandwidth to process the request and transfer the response
  - Fast context switches between user and kernel space
  - Fast instruction fetching on the CPU front end



Memory Hierarchy Performance

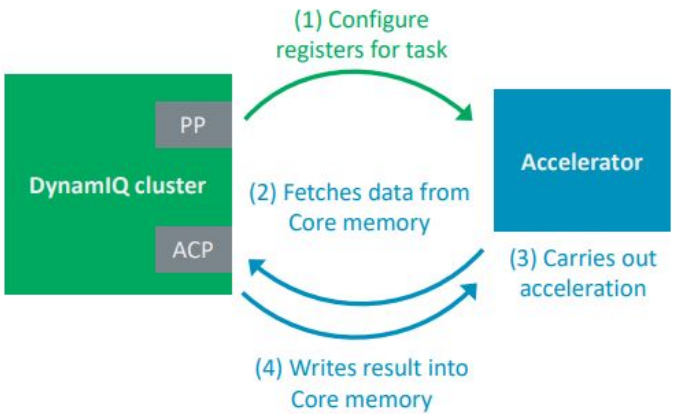
Workload stressor	Neoverse N1 Features	N1 improvement over Cortex-A72
Object management	Memory allocations	***2.4x faster***
	Object/array initializations	***5x faster***
	Copy chars	***1.6x faster***
	Smart HW handling of SW barriers (DMBs)	***Memory barriers elided if unnecessary***
Instruction footprint	i-cache miss rate and branch mispredicts	***Reduced by 1.4x***
	L2 accesses	***Reduced by 2.25x***
	Fully HW coherent Icache	***Accelerates VM bring up by up to 20x***
Garbage collection	Locking throughput w/ V8.2 Arch Atomic Instructions	***Improved by 2x***



[Arm Neoverse N1 Cloud-to-Edge Infrastructure SoCs]

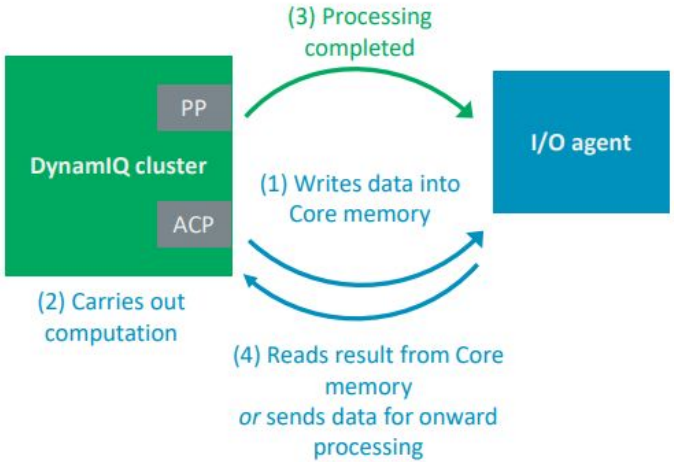
Cache Stashing

Offload acceleration



Example application:
Offload crypto acceleration

I/O processing



Example application:
Packet processing in network systems

[Arm DynamIQ HotChips Presentation]

Direct Memory Transfer

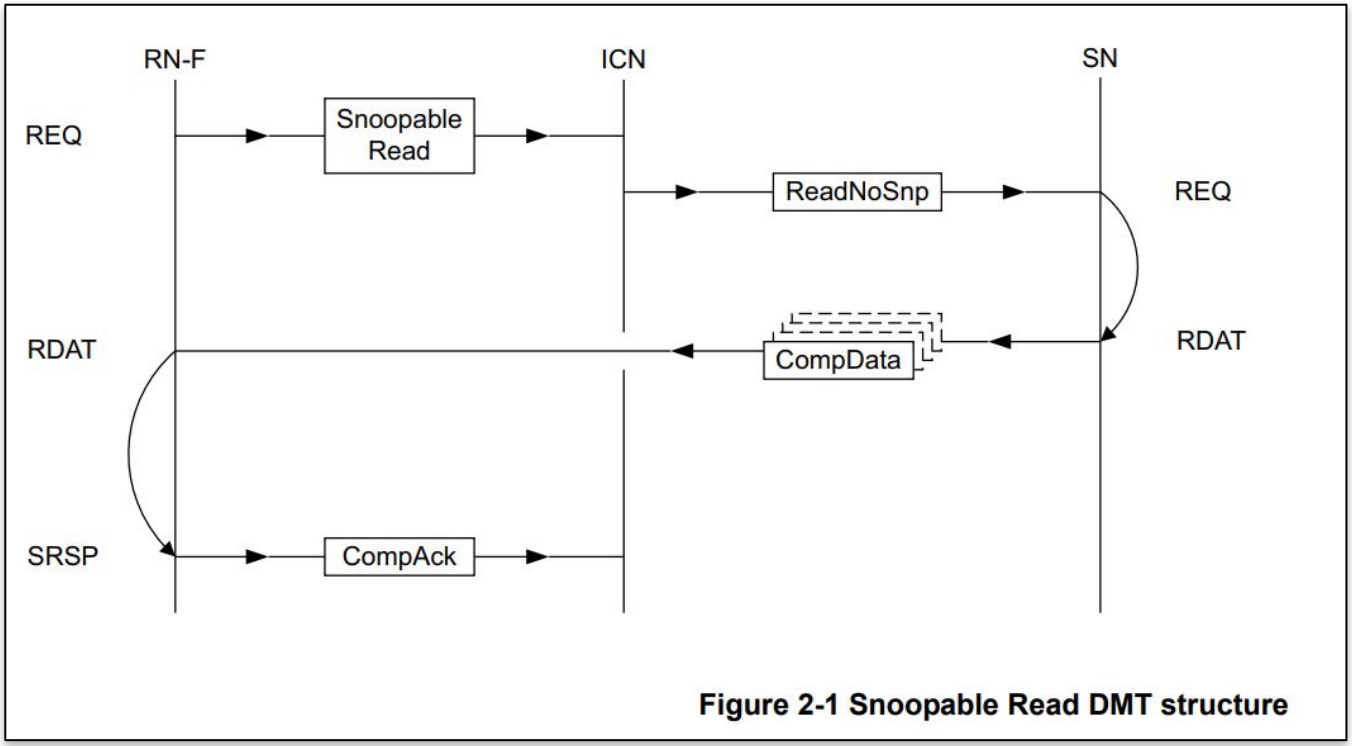


Figure 2-1 Snoopable Read DMT structure

AMBA 5 CHI Architecture Specification