Arm Neoverse V2 platform: Leadership Performance and Power Efficiency for Next-Generation Cloud Computing, ML and HPC Workloads

Hot Chips 2023

Magnus Bruce, Lead CPU Architect and Fellow, Arm

August 28th, 2023

Arm Technology is Defining the Future of Computing

A semiconductor design and software platform company

250+ Billion

Arm-based chips shipped since inception

30.6 Billion

Arm-based chips reported shipped in FYE 2023

650+

Active licensees, growing by 50+ every year.

The global leader in the development of licensable compute technology

R&D excellence for semiconductor companies and large OEMs.

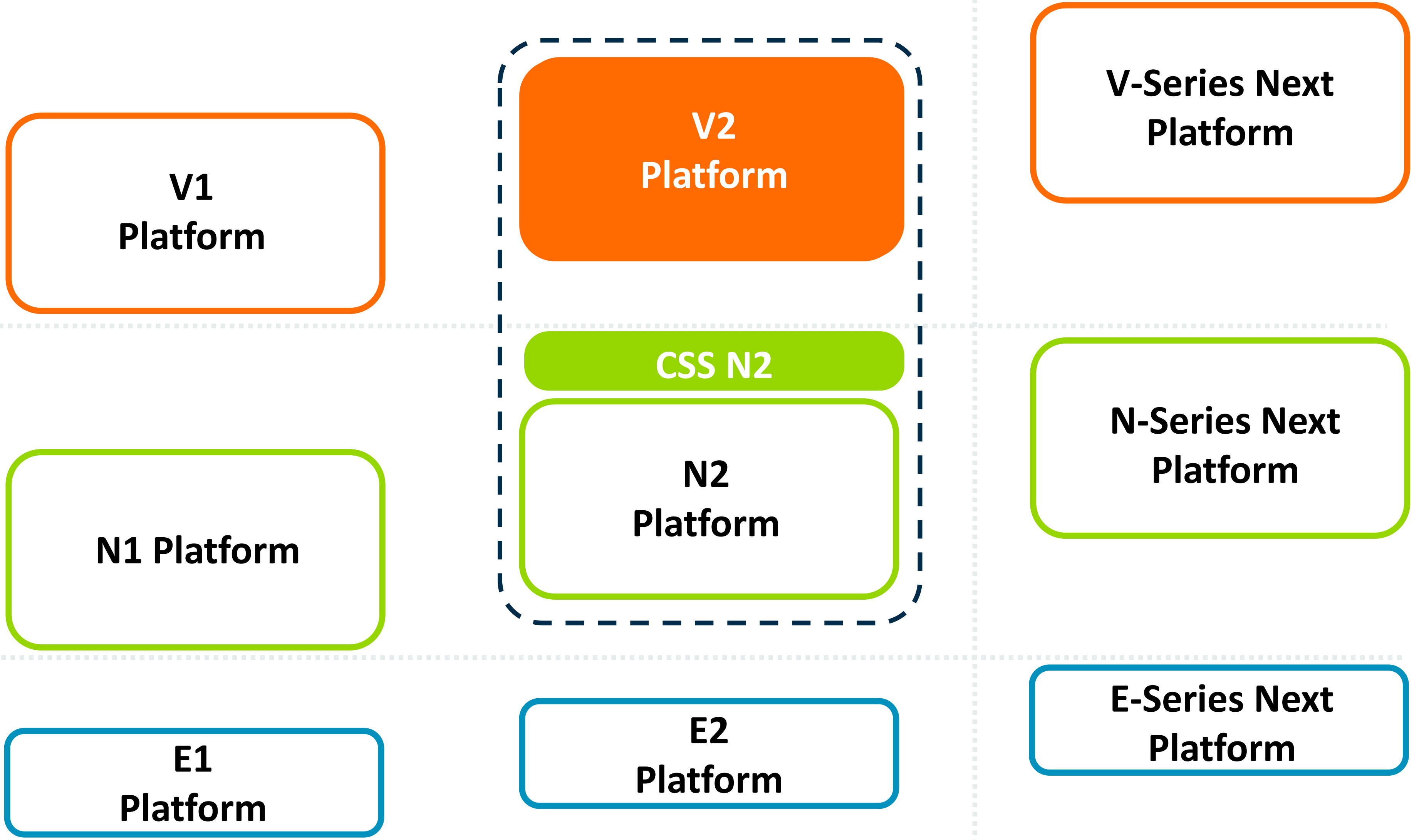
Arm's energy-efficient processor designs and software platforms enable advanced computing

Our technologies securely power products from the sensor to the smartphone and the supercomputer.

Arm delivers the foundational building blocks for trust in the digital world

Arm provides enhanced system-level security technologies such as Arm TrustZone and Arm Confidential Compute Architecture (CCA).

Arm Neoverse Roadmap and Product Positioning



Neoverse V-series Maximum Performance and Optimal TCO Cloud, HPC, AI/ML	V1 Platform	V2 Platform	V-Series Next Platform
Neoverse N-series Efficient Performance Cloud, Networking, DPU, 5G	N1 Platform	CSS N2 N2 Platform	N-Series Next Platform
Neoverse E-series Throughput Efficiency Networking, 5G	E1 Platform	E2 Platform	E-Series Next Platform

Arm Neoverse V2 Design Principles

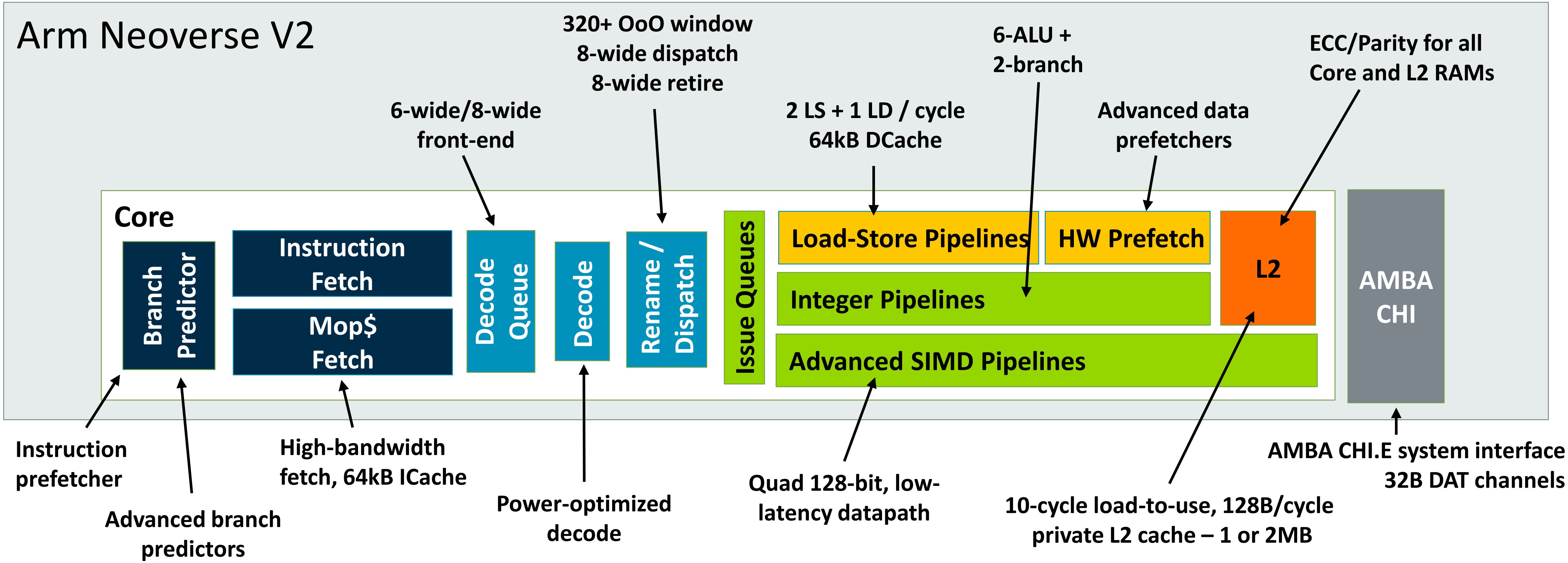
Performance Leadership in Cloud, HPC and AI/ML

- Run-ahead branch prediction pipeline
  - Decouples branch from fetch
  - Tolerates a relatively small L1 instruction cache
  - Large BTBs to avoid redirection later in pipeline
  - Predicts direct branches during fetch
- Physical register files, read after issue
- High bandwidth, low-latency L1 and private L2 caches
- Store-to-load forwarding at L1 hit latency
- Advanced prefetchers with timeliness and accuracy monitoring
- Dynamic feedback mechanisms to adapt to system conditions
- Push 'width' and 'depth' higher
  - Maintain short pipelines for quick branch mispredict recovery

Continue to deliver the highest single-thread performance
in the lowest power and area footprint

High-Level Microarchitecture

- Every aspect of the microarchitecture optimized for performance & TCO


Arm Neoverse V2

6-wide/8-wide front-end

320+ OoO window
8-wide dispatch
8-wide retire

2 LS + 1 LD / cycle
64kB DCache

6-ALU +
2-branch

Advanced data prefetchers

ECC/Parity for all Core and L2 RAMs

Instruction prefetcher

Advanced branch predictors

High-bandwidth fetch, 64kB ICache

Power-optimized decode

Quad 128-bit, low-latency datapath

10-cycle load-to-use, 128B/cycle private L2 cache – 1 or 2MB

AMBA CHI.E system interface
32B DAT channels

Ultra-High Performance Armv9 CPU

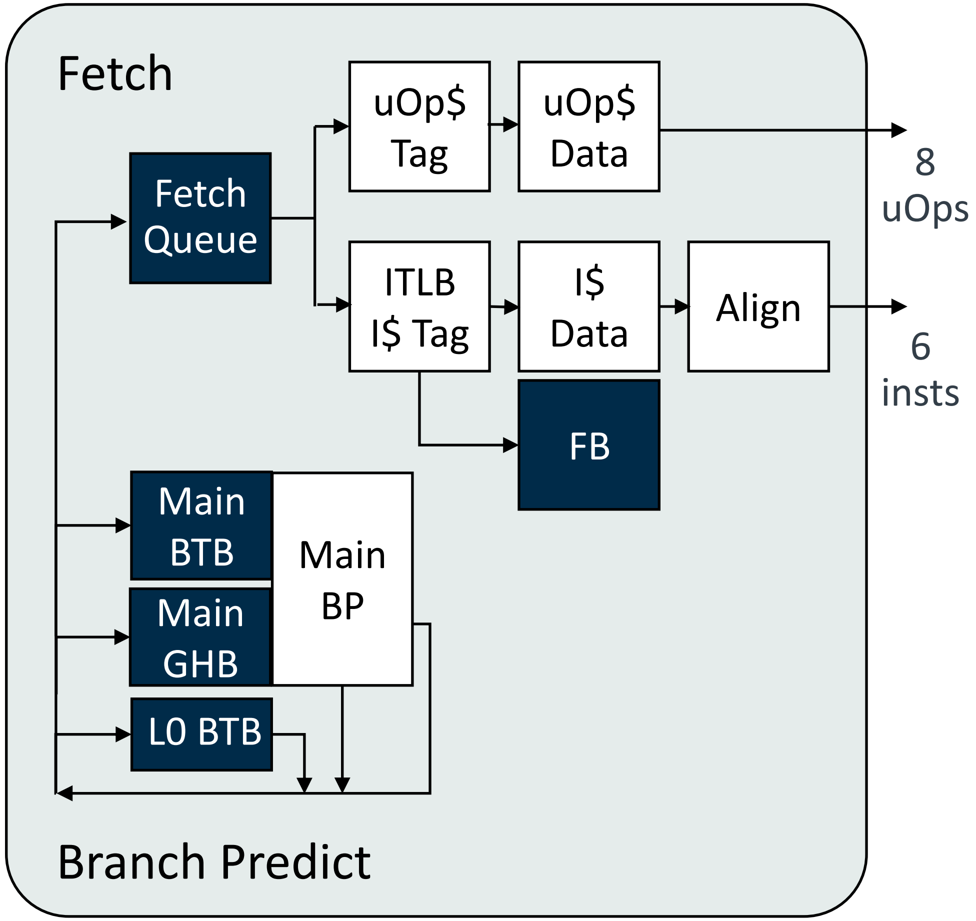
Branch Predict/Fetch/ICache

uArch features shared with Neoverse V1

- Decoupled predict/fetch pipelines
  - Predict runs ahead to avoid bubbles and cover cache misses
- Two predicted branches per cycle
- Predictor acts as ICache prefetcher
- 64kB, 4-way set-associative L1 instruction cache
- Two-level Branch Target Buffer
- 8 table TAGE direction predictor with staged output

uArch features **new with Neoverse V2**

Branch Target Buffer	10x larger nanoBTB Split main BTB into two levels with 50% more entries
TAGE	2x larger tables with 2-way associativity Longer history
Indirect branches	Dedicated predictor
Fetch bandwidth	Doubled instruction TLB and cache BW
Fetch Queue	Doubled from 16 to 32 entries
Fill Buffer	Increased size from 12 to 16 entries
uOp cache	Reduced size for efficiency



+2.9% SPEC CPU® 2017 Integer[1]

[1] Performance is estimated for SPEC CPU® 2017

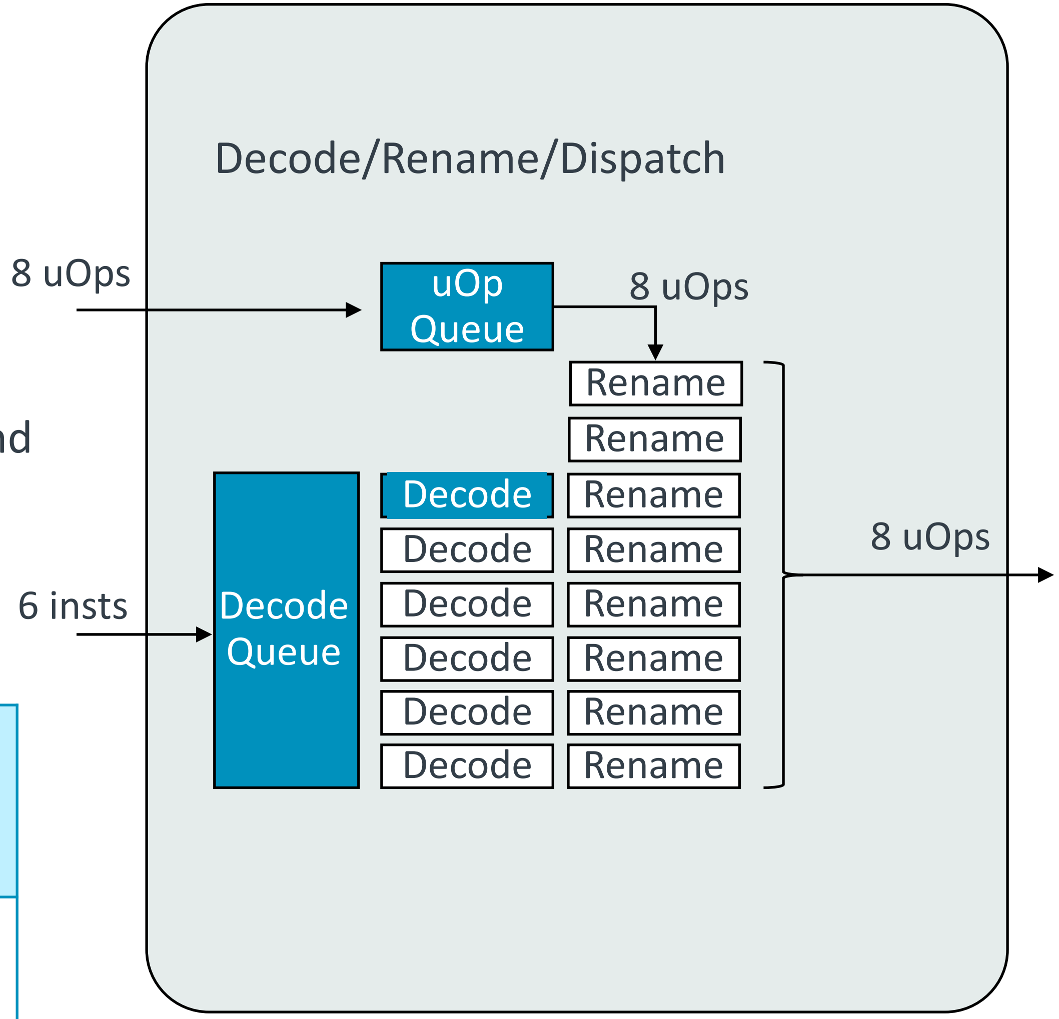
Decode/Rename/Dispatch

uArch features shared with Neoverse V1

- Partially decoded instructions from I$ feed parallel decoders
- Fully decoded uOps bypass decode with higher width
- Decode handles simple instruction fusion
- Rename manages physical register files with both architected and speculative state using mapping tables and free list

uArch features **new with Neoverse V2**

Decode bandwidth	Increased decoder lanes from 5 to 6 Increased Decode Queue from 16 to 24 entries
Rename checkpoints	Increased from 5 to 6 total checkpoints Increased from 3 to 5 vector checkpoints
Rename rebuild	Improved rebuild flows for more efficient recovery after pipeline flush



+2.8% SPEC CPU® 2017 Integer[1]

[1] Performance is estimated for SPEC CPU® 2017

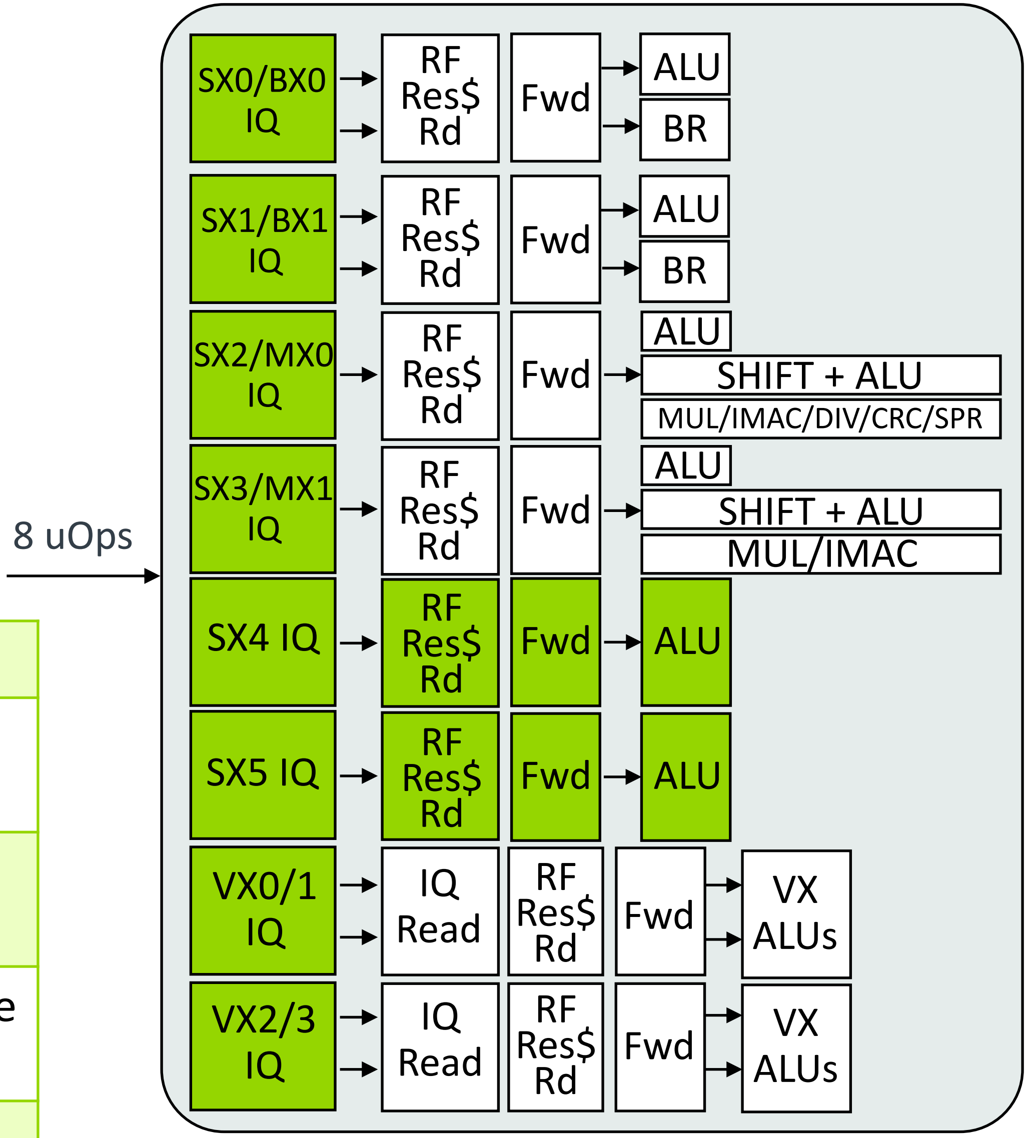
Issue/Execute

uArch features shared with Neoverse V1

- Multiple independent Issue Queues, some with dual pickers
- Late read of physical register file – no data in IQs
- Result caches with lazy writeback

uArch features **new with Neoverse V2**

ALU bandwidth	Added two more single-cycle ALUs
Larger Issue Queues	SX/MX: Increased from 20 to 22 entries VX: Increased from 20 to 28 entries
Predicate operations	Doubled predicate bandwidth
Zero latency MOV	Subset of register-register and immediate move operations execute with zero latency
Instruction fusion	More fusion cases, including CMP + CSEL/CSET



+3.3% SPEC CPU® 2017 Integer[1]

[1] Performance is estimated for SPEC CPU® 2017

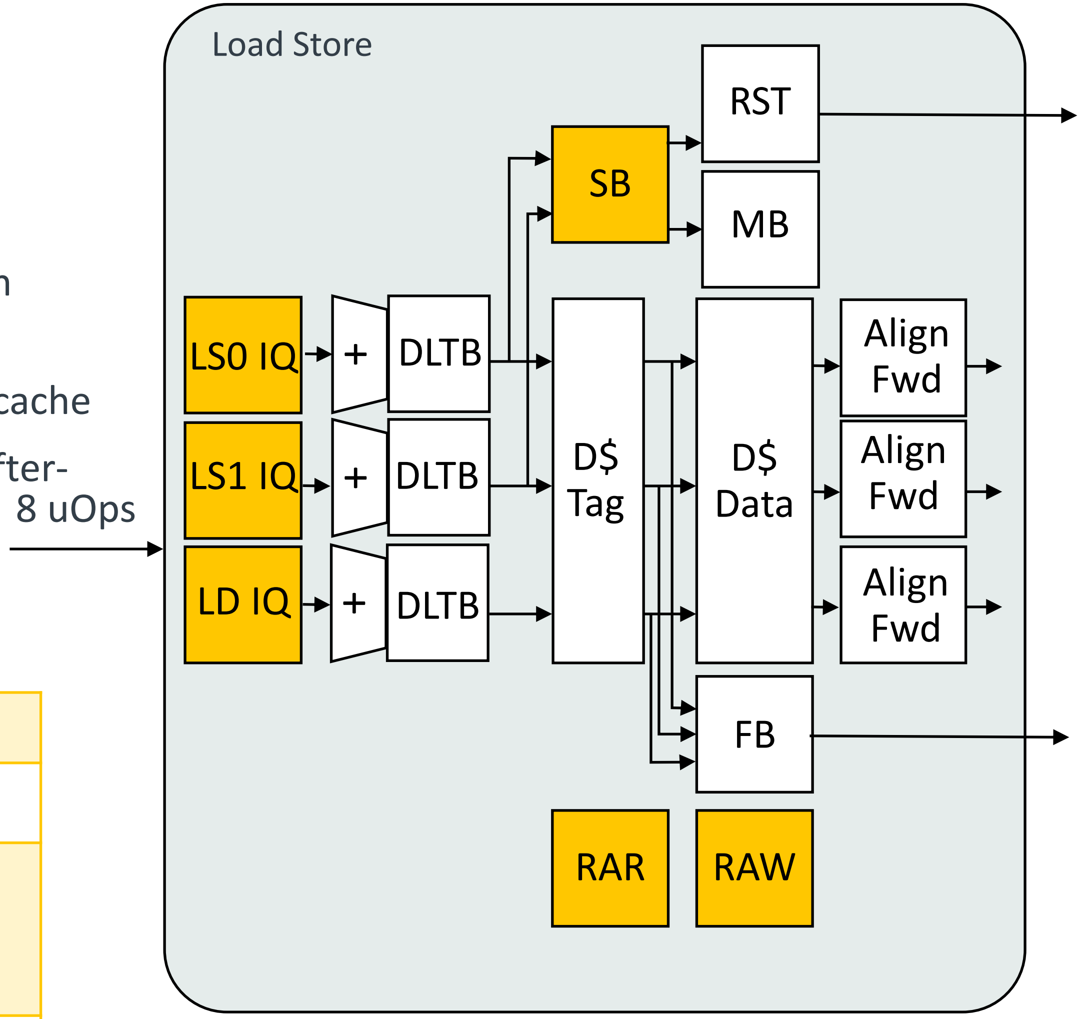
LoadStore/DCache

uArch features shared with Neoverse V1

- Two load/store pipes + one load pipe
- 4 x 8B result busses (integer)
- 3 x 16B result busses (FP, SVE, Neon)
- ST to LD forwarding at L1 hit latency
- RST and MB to reduce tag and data accesses
- Fully-associative L1 DTLB with multiple page sizes
- 64kB 4-way set associative Dcache
- Read-After-Read and Read-After-Write hazard detection

uArch features **new with Neoverse V2**

TLB	Increased from 40 to 48 entries
Replacement policy	Changed from PLRU to dynamic RRIP
Larger Queues	Store Buffer ReadAfterRead ReadAfterWrite
Efficiency	VA hash based store to load forwarding
Reduced flushes	RAR hazards tracked through L2 cache lifetime



+3.0% SPEC CPU® 2017 Integer[1]

[1] 1. Performance is estimated for SPEC CPU® 2017

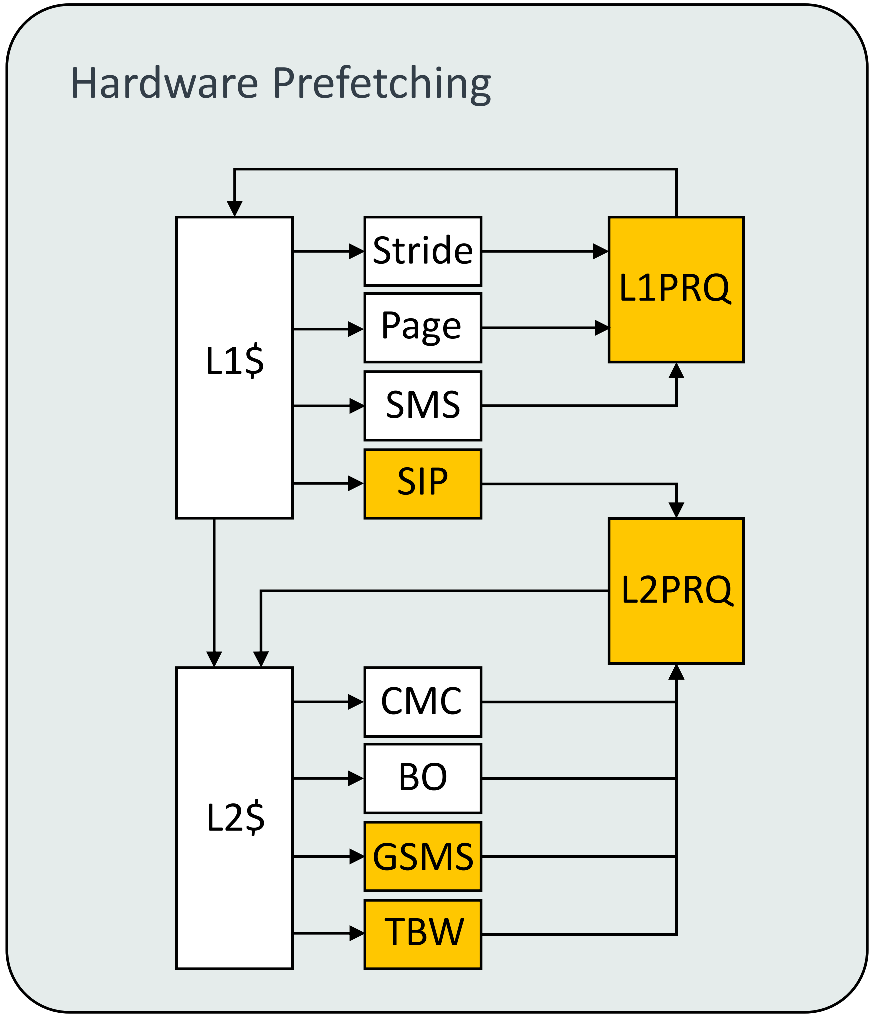
Hardware Prefetching

uArch features shared with Neoverse V1

- Multiple prefetching engines training on L1 and L2 accesses
  - Spatial Memory Streaming
  - Stride
  - Page
  - Best Offset
  - Correlated Miss Cache
- Prefetch into L1 and L2
- Virtual address to allow page crossing and TLB prefetching
- Adaptive distance based on accuracy and timeliness

uArch features **new with Neoverse V2**

Training	Refined filtering of transactions used for training
Accuracy	Apply Program Counter to L2 GSMS training
New PF engines	Global SMS – larger offsets than SMS Sampling Indirect Prefetch – pointer dereference TableWalk – Page Table Entries
Differentiated QoS	Lower QoS value for prefetches than demand for reduced loaded latency


Hardware Prefetching

+5.3% SPEC CPU® 2017 Integer[1]

[1] Performance is estimated for SPEC CPU® 2017

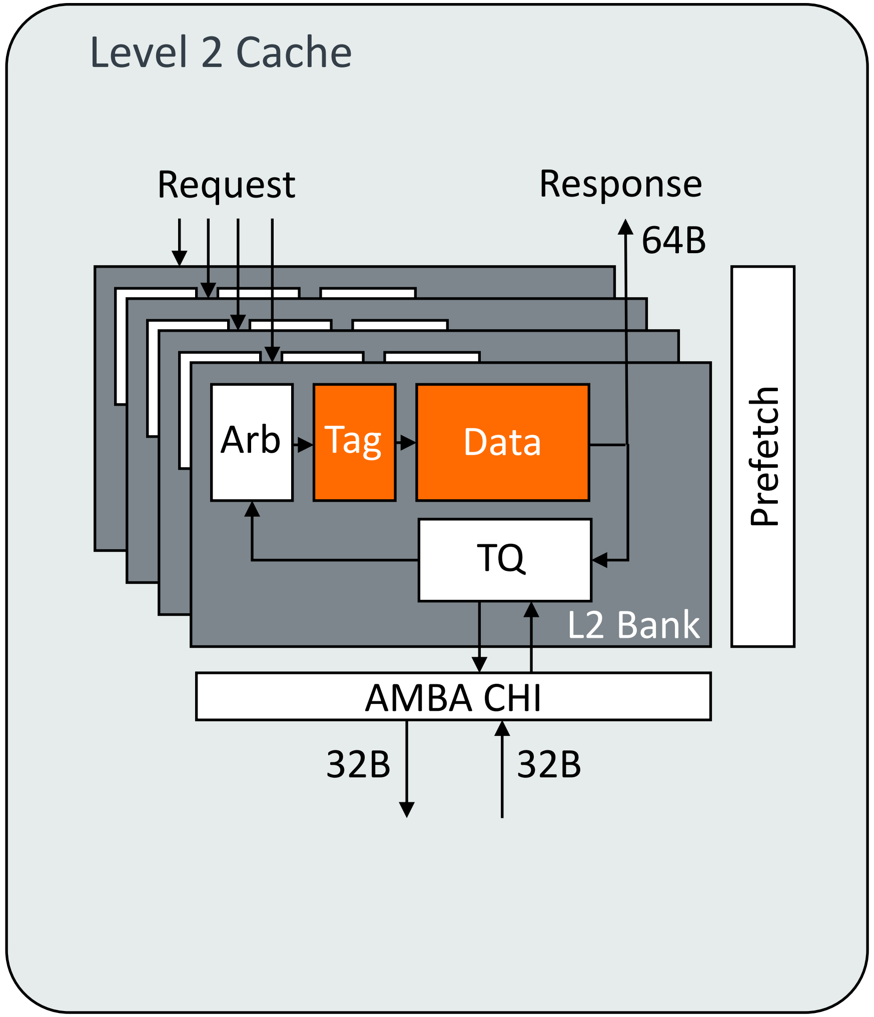
Level 2 Cache

uArch features shared with Neoverse V1

- Private unified Level 2 cache, 8-way SA, 4 independent banks
- 64B read or write per 2 cycles per bank = 128B/cycle total
- 96-entry Transaction Queue
- Inclusive with L1 caches for efficient data and instruction coherency
- Inline SECDED ECC in Tag, Data, and TQ RAMs
- AMBA CHI interface with 256b DAT channels

uArch features **new with Neoverse V2**

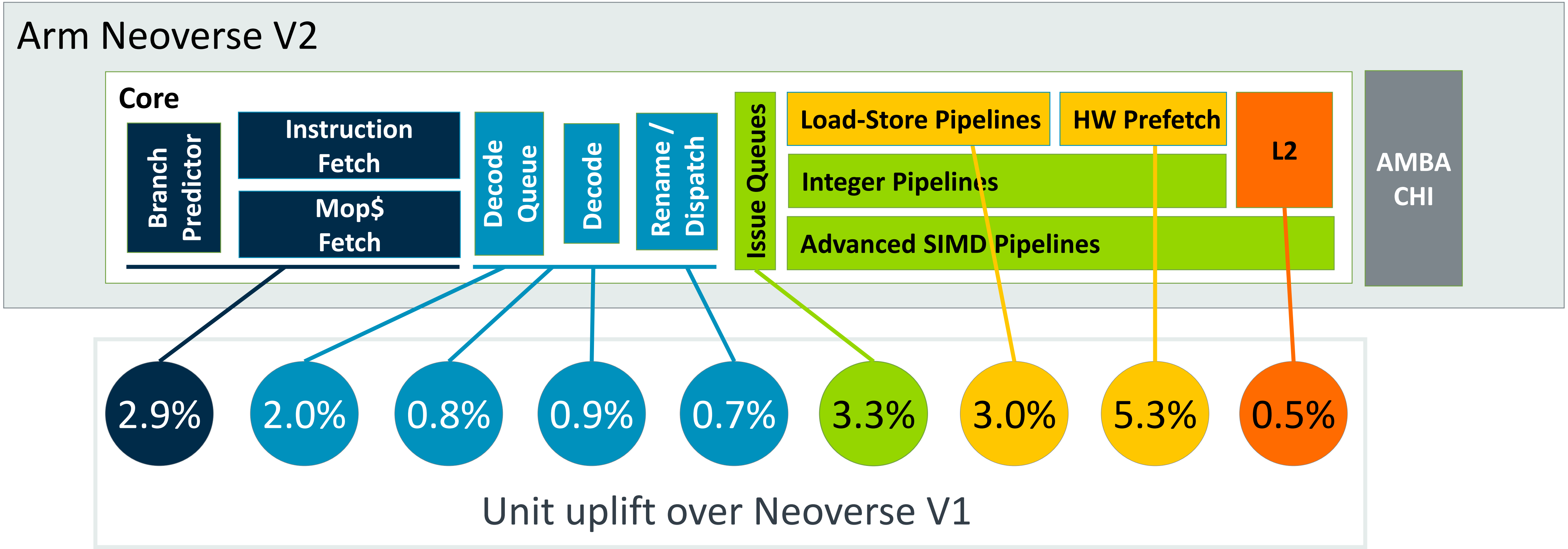
Capacity	2MB/8-way with latency of 1MB (10-cycle ld-to-use)
Replacement policy	6-state RRIP (up from 4)
Dead-block prediction	Separate tracking of used-once and used-multiple data
Replacement	L2 copybacks transfer replacement hints to SLC
CHI revision E interconnect	Improved store-hit-shared flow (MakeReadUnique) Combined Write/Cache Maintenance transactions Write*Zero transactions for memset



+0.5% SPEC CPU® 2017 Integer[1]
8.2% reduction in SLC misses

[1] Performance is estimated for SPEC CPU® 2017

Neoverse V2 Performance Uplift over Neoverse V1



13% increase in SPEC CPU® 2017 Integer performance[1],
while seeing a 10.5% reduction in SLC misses

[1] Performance is estimated for SPEC CPU® 2017

Arm Neoverse V2 Performance, Power, Area (PPA)

Neoverse V1 with 1 MB L2

Typical 7nm implementation

2.52 mm^2 with 1 MB L2

1.2 W*

*SIR at 2.8 GHz, 0.75 V, H280, 16LM

Neoverse V2 with 2 MB L2

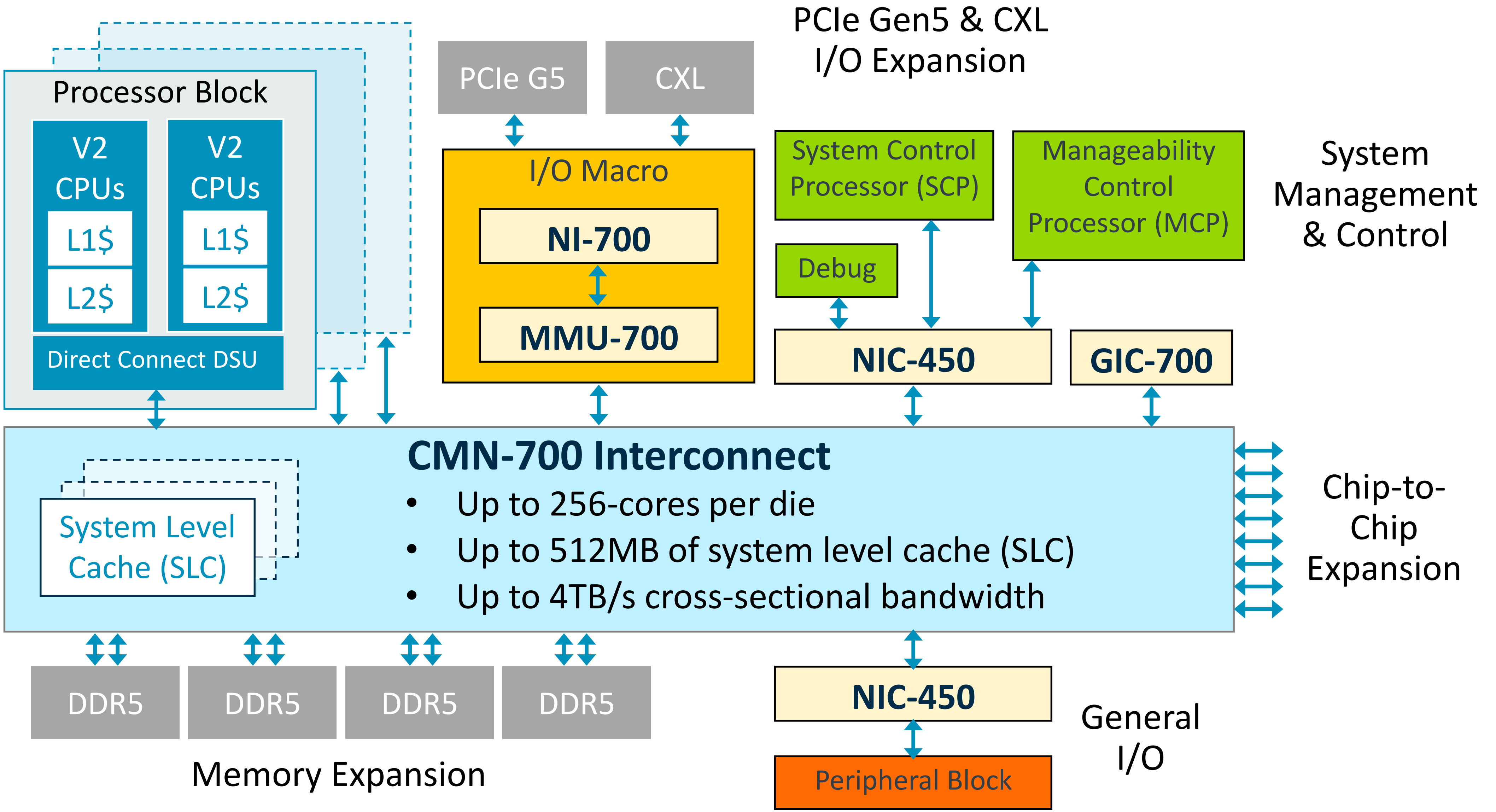
Typical 5nm implementation

2.50 mm^2 with 2 MB L2

1.4 W*

*SIR at 2.8 GHz, 0.75 V, H280, 17LM

Arm Neoverse V2 Platform IP



PCIe Gen5 & CXL
I/O Expansion

PCIe G5

CXL

Processor Block

V2
CPUs

L1$

L2$

V2
CPUs

L1$

L2$

Direct Connect DSU

I/O Macro

NI-700

MMU-700

System Control Processor (SCP)

Manageability Control Processor (MCP)

System
Management
& Control

Debug

NIC-450

GIC-700

CMN-700 Interconnect

- Up to 256-cores per die
- Up to 512MB of system level cache (SLC)
- Up to 4TB/s cross-sectional bandwidth

System Level Cache (SLC)

Chip-to-Chip Expansion

DDR5

DDR5

DDR5

DDR5

Memory Expansion

NIC-450

Peripheral Block

General I/O

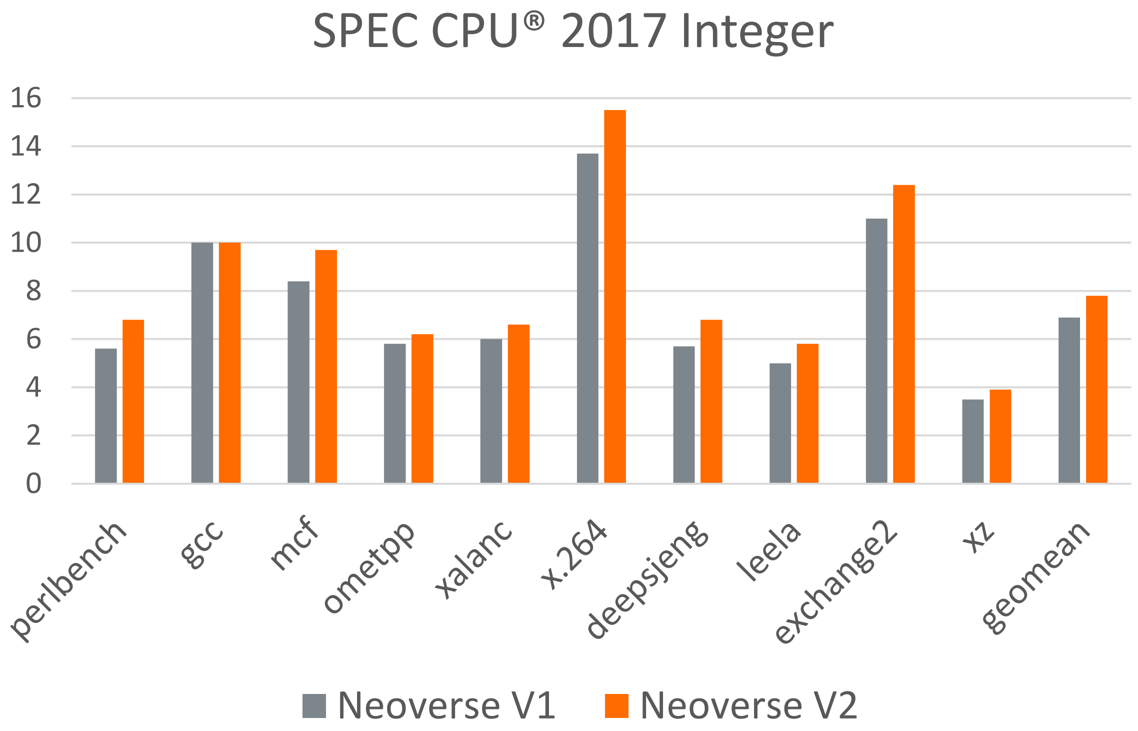
Next: Performance Analysis of Neoverse V2 compared to Neoverse V1

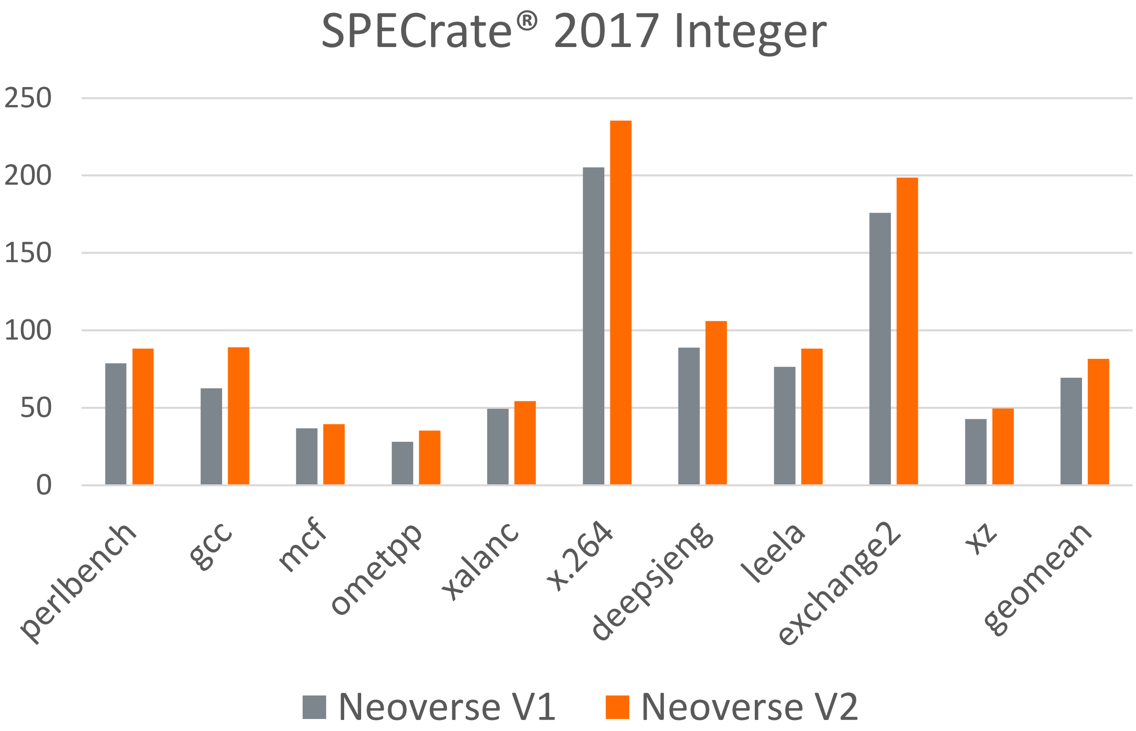
- Neoverse V1 and Neoverse V2 performance comparisons are derived from equivalent systems in an emulation environment
- 32 CPU cores @ 3 GHz
- Neoverse V1 with 1MB L2, Neoverse V2 with 2MB L2
- CMN-700 interconnect @ 2GHz with 32MB System Level Cache
- Four DDR-5600 memory controllers, 40-bit memory interfaces – 89.6 GB/s max BW

- SPEC CPU®2017 scores are estimated using reduced benchmarks

- GCC 10 with standard compile options – no special optimizations

General Performance: SPEC CPU® 2017 Integer



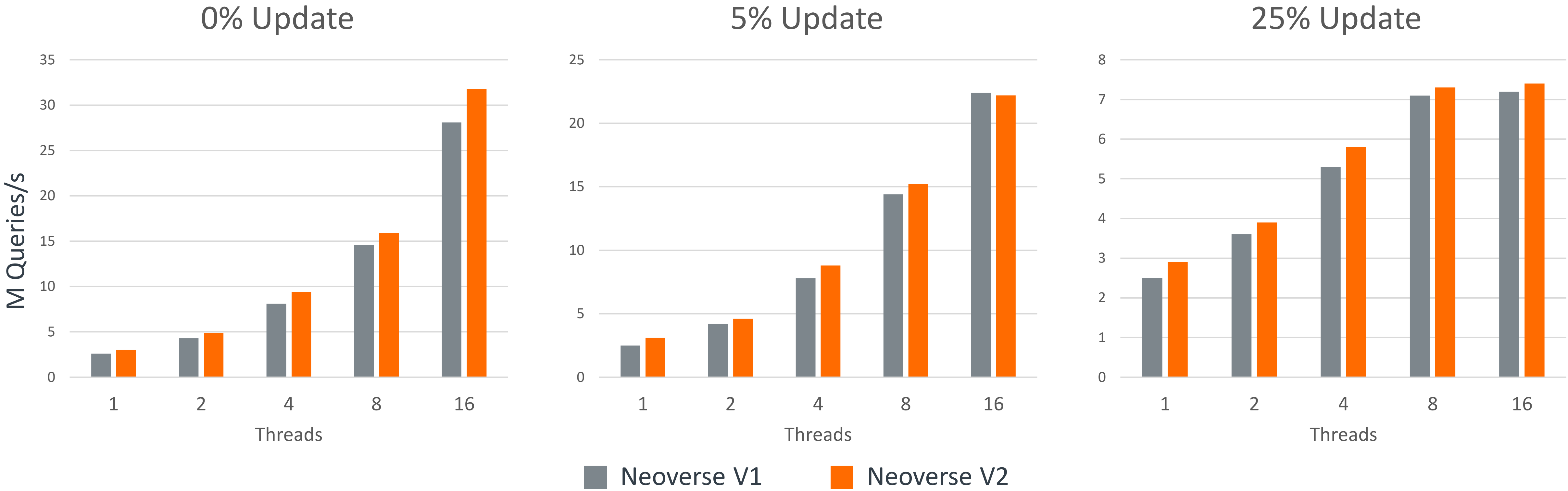


On SPEC CPU® 2017 Integer, Neoverse V2 shows a 13% improvement over Neoverse V1[1]

On SPECrate® 2017 Integer, Neoverse V2 shows a 17.3% improvement over Neoverse V1[1]

[1] Performance is estimated for SPEC CPU® 2017

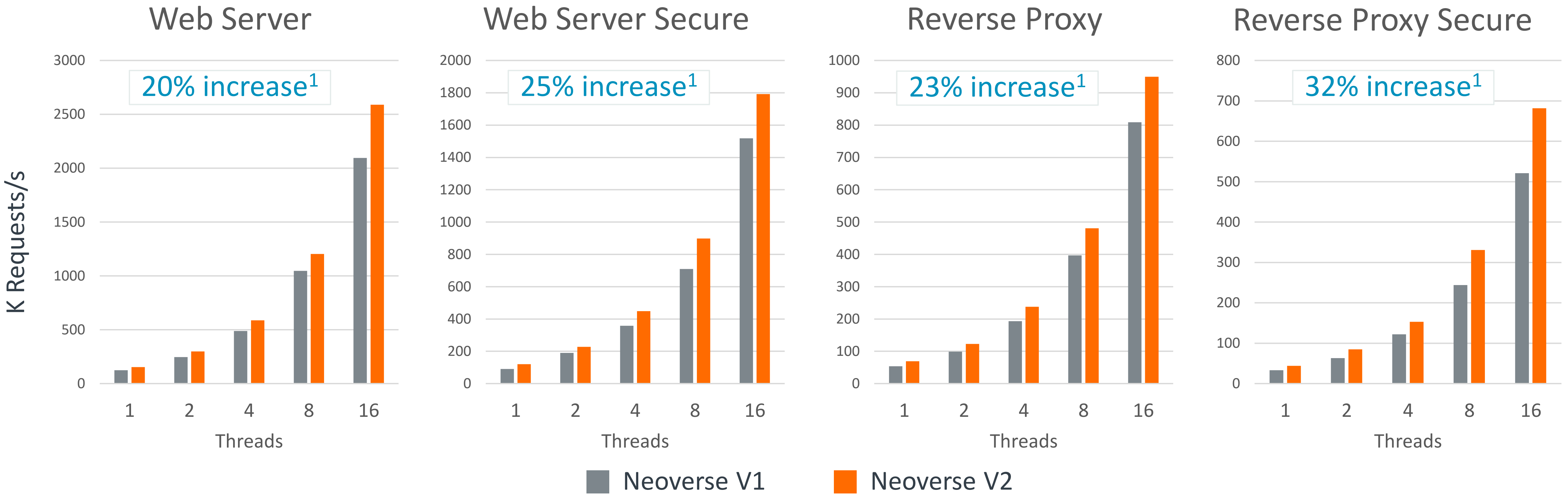
Caching Tier Performance: MemCacheD



On Memcached, Neoverse V2 shows a 13-15% improvement over Neoverse V1

Performance scaling becomes system limited as the update percentage increases

Web and Proxy Server Performance: NGINX

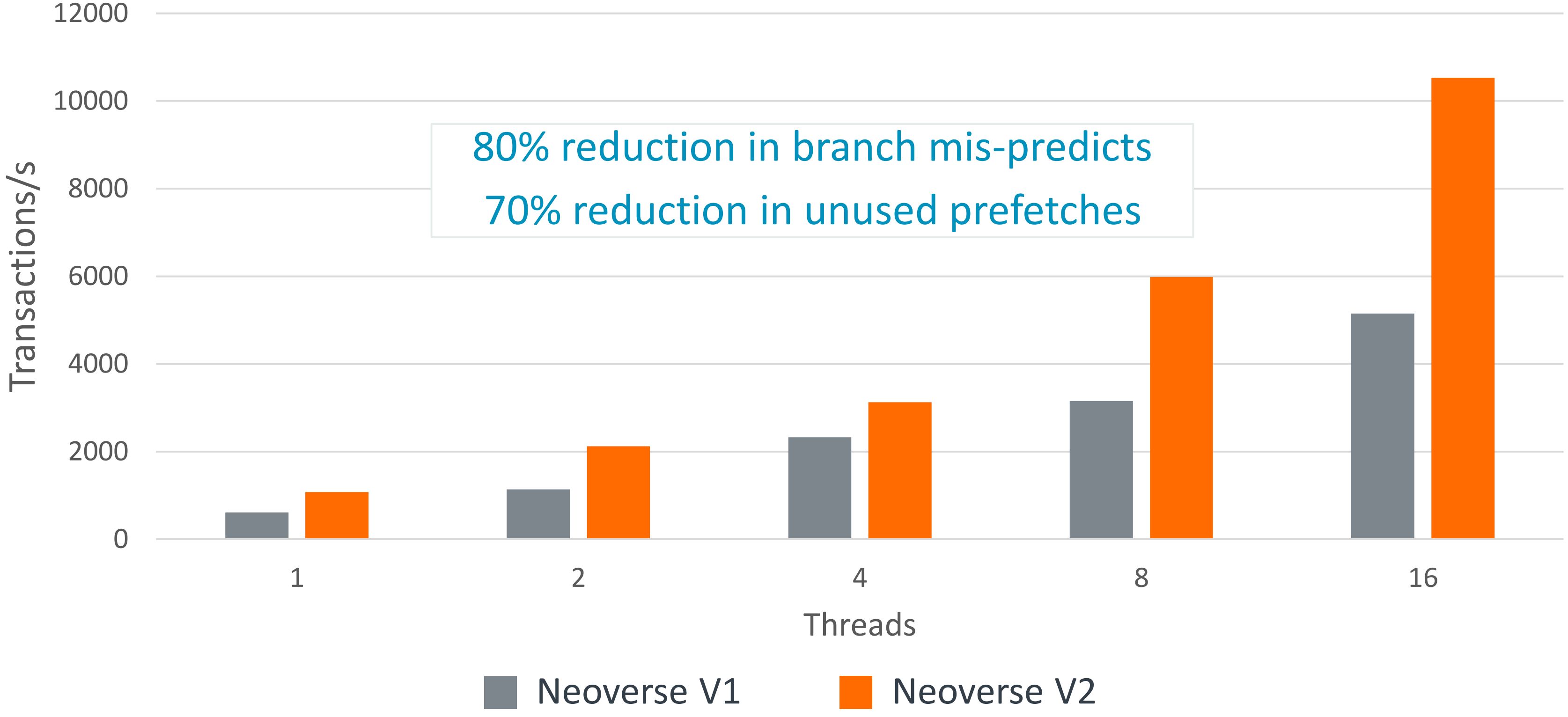


On NGINX, Neoverse V2 shows a 20-32% performance improvement over Neoverse V1

Performance scaling improves with higher thread count

[1] Average performance improvement across thread count

Database Performance: Percona MySQL Server



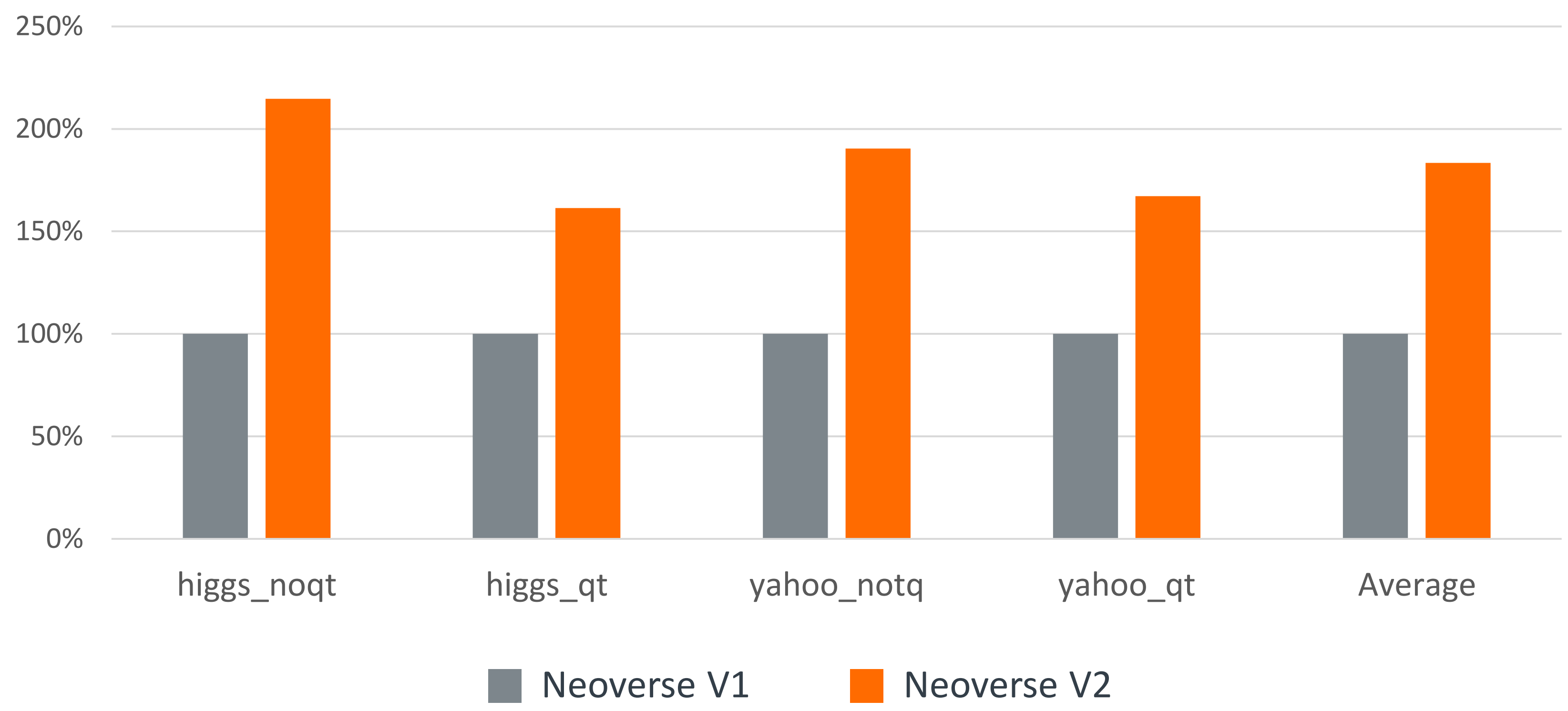
80% reduction in branch mis-predicts

70% reduction in unused prefetches

On Percona MySQL, Neoverse V2 shows a 35-104% performance improvement over Neoverse V1

Significant gains from improvements in branch prediction, fetch, and hardware prefetching

ML Performance: XGBoost

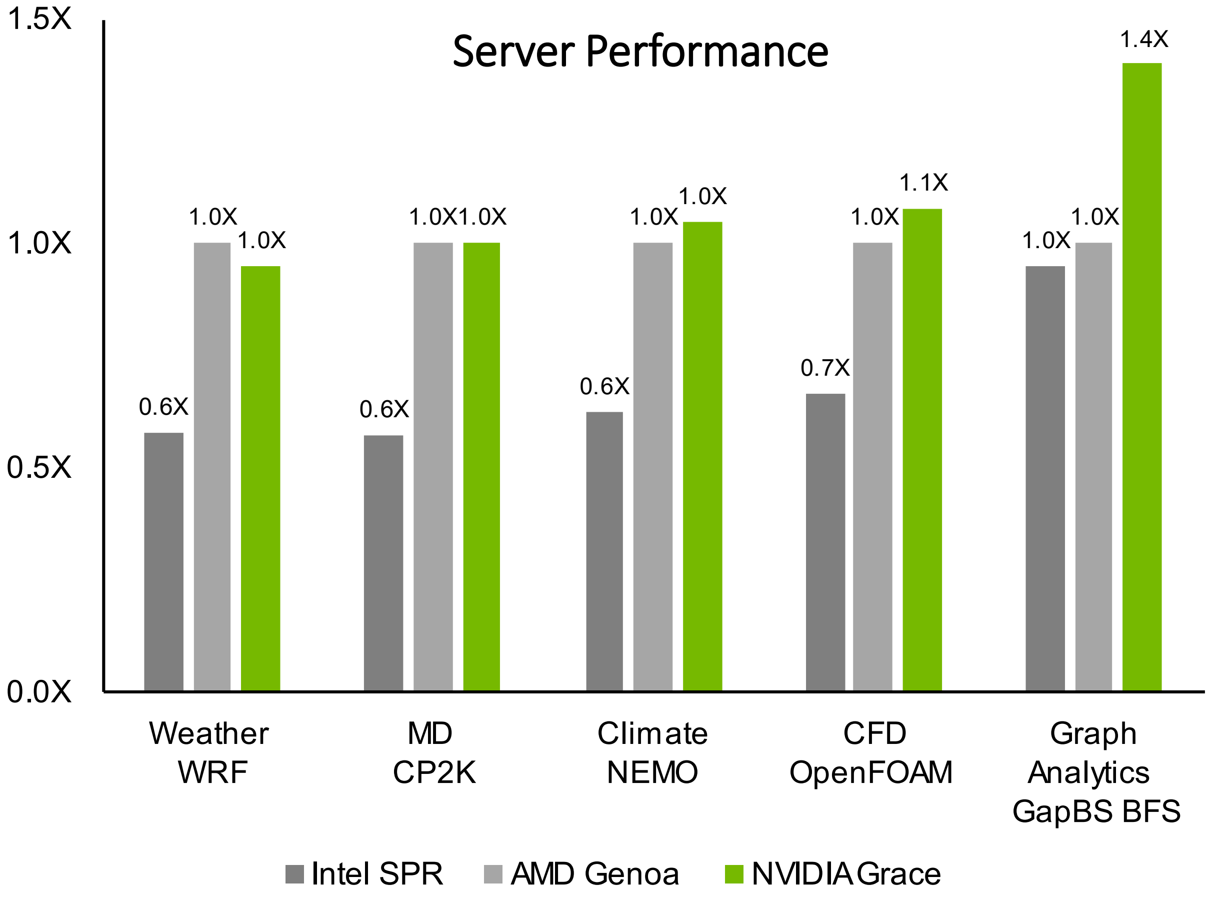


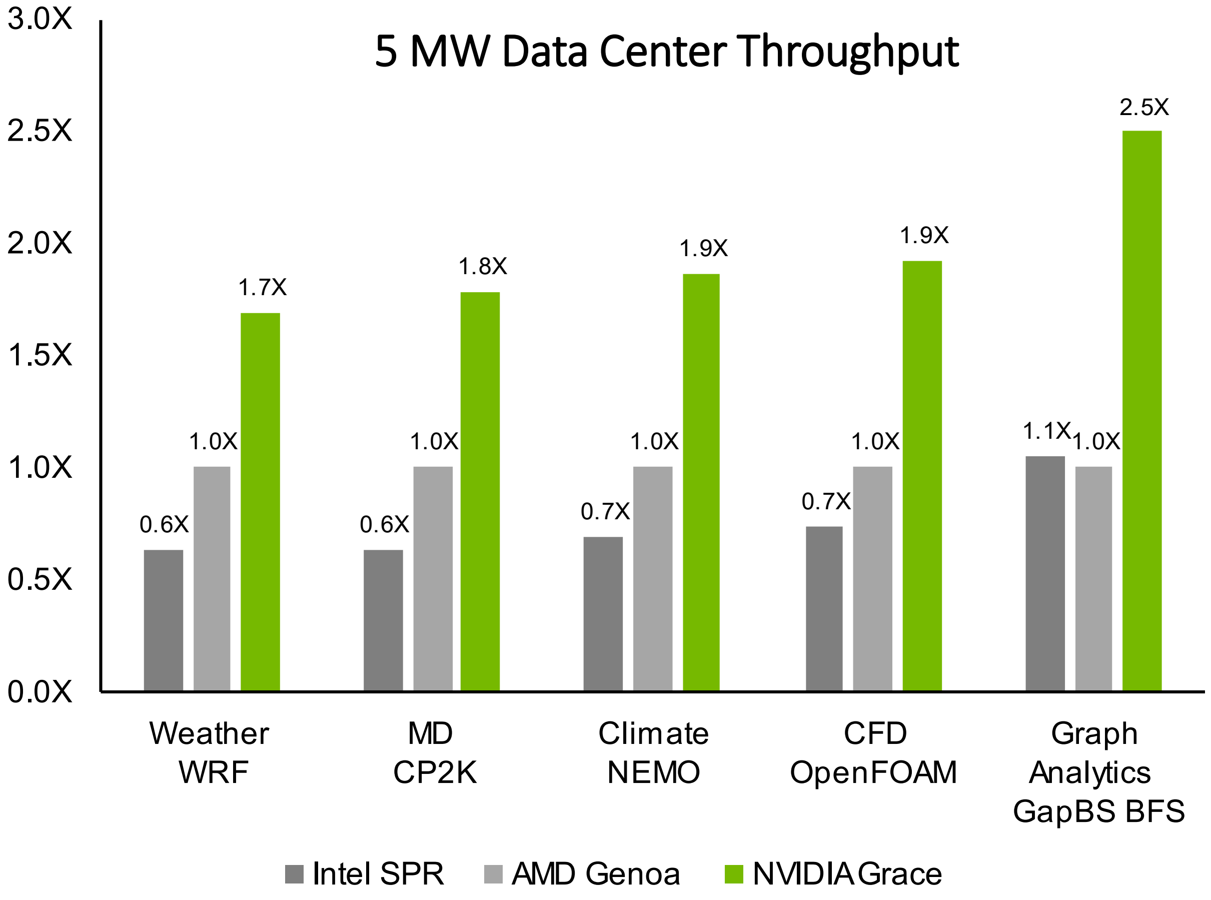
On XGBoost, Neoverse V2 shows a 67-114% performance improvement over Neoverse V1

Branch prediction and fetch improvements enable large performance gains

NVIDIA Grace CPU Delivers 2X Throughput at the Same Power

Powered by Neoverse V2 Core and High-Speed NVIDIA-Designed Scalable Coherency Fabric with LPDDR5X Memory





Data provided by NVIDIA

5 MW Data Center level projection based on server measurement of NVIDIA Grace Superchip vs x86 2-socket data center systems (AMD Epyc 9654 and Intel Xeon 8480+).
Weather: WRF CONUS12, 24 hr simulation 4.4.2 MD: CP2K RPA 2023.1 Climate: NEMO Gyre_Pisces v4.2.0 CFD: OpenFOAM Motorbike | Large v2212 Graph Analytics: The Gap Benchmarks Suite BFS NVIDIA Grace Superchip performance based on engineering measurements. Results subject to change.

NVIDIA

arm NEOVERSE

Arm Neoverse V2 Platform Summary

- Designed for cloud performance leadership
  - Double-digit gains over Neoverse V1 on cloud infrastructure workloads
    - 13% uplift on SPEC CPU® 2017 Integer[1]
    - 15% to 100% uplift across a range of server workloads (caching, web, database)

- Designed for HPC and AI/ML performance leadership
  - Up to 2x the performance of Neoverse V1 on HPC and ML workloads
    - Up to 114% uplift on XGBoost (83% average)
    - Meets or exceeds leading x86-CPUs on performance with up to 2x the performance efficiency

- Available today in NVIDIA Grace CPU Superchip
- Additional partner silicon expected

NVIDIA

[1] Performance is estimated for SPEC CPU® 2017

Performance and benchmark disclaimer

- This benchmark presentation made by Arm Ltd and its subsidiaries (Arm) contains forward-looking statements and information. The information contained herein is therefore provided by Arm on an "as-is" basis without warranty or liability of any kind. While Arm has made every attempt to ensure that the information contained in the benchmark presentation is accurate and reliable at the time of its publication, it cannot accept responsibility for any errors, omissions or inaccuracies or for the results obtained from the use of such information and should be used for guidance purposes only and is not intended to replace discussions with a duly appointed representative of Arm. Any results or comparisons shown are for general information purposes only and any particular data or analysis should not be interpreted as demonstrating a cause and effect relationship. Comparable performance on any performance indicator does not guarantee comparable performance on any other performance indicator.
- Any forward-looking statements involve known and unknown risks, uncertainties and other factors which may cause Arm’s stated results and performance to be materially different from any future results or performance expressed or implied by the forward-looking statements.
- Arm does not undertake any obligation to revise or update any forward-looking statements to reflect any event or circumstance that may arise after the date of this benchmark presentation and Arm reserves the right to revise our product offerings at any time for any reason without notice.
- Any third-party statements included in the presentation are not made by Arm, but instead by such third parties themselves and Arm does not have any responsibility in connection therewith.

End Notes

Slide Title: Branch Predict/Fetch/ICache

Slide Title: Decode/Rename/Dispatch

Slide Title: Issue/Execute

Slide Title: LoadStore/DCache

Slide Title: Hardware Prefetching

Slide Title: Level 2 Cache

Slide Title: Neoverse V2 Performance Uplift over Neoverse V1

Slide Title: General Performance: SPEC CPU® 2017 Integer

Slide Title: Arm Neoverse V2 Platform Summary

- SPEC CPU®2017 scores are estimated using reduced benchmarks
  - Neoverse V1 and Neoverse V2 performance comparisons are derived from equivalent systems in an emulation environment, 32 CPU cores @ 3 GHz, Neoverse V1 with 1MB L2, Neoverse V2 with 2MB L2, CMN-700 interconnect @ 2GHz with 32MB system level cache, four DDR-5600 memory controllers, 40-bit memory interfaces – 89.6 GB/s max BW
  - GCC 10 with standard compile options – no special optimizations

Thank You

Danke

Gracias

Grazie

谢谢

ありがとう

Asante

Merci

감사합니다

धन्यवाद

Kiitos

شكرًا

ধন্যবাদ

תודה