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

© 2023 Arm
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

**Neoverse N-series**
- Efficient Performance
- Cloud, Networking, DPU, 5G

**Neoverse E-series**
- Throughput Efficiency
- Networking, 5G

**Platforms**
- **V1 Platform**
- **N1 Platform**
- **E1 Platform**
- **V2 Platform**
- **N2 Platform**
- **E2 Platform**

**Next Platforms**
- **V-Series Next Platform**
- **N-Series Next 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
  - Push ‘width’ and ‘depth’ higher
    - Maintain short pipelines for quick branch mispredict recovery

- Store-to-load forwarding at L1 hit latency

- Advanced prefetchers with timeliness and accuracy monitoring

- Dynamic feedback mechanisms to adapt to system conditions

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

- Core
  - Instruction Fetch
  - Mop$ Fetch

- Decode Queue
  - Connects to Decode

- Rename / Dispatch
  - Connects to Issue Queues

- Issue Queues
  - Connects to Load-Store Pipelines

- Load-Store Pipelines
  - Connects to HW Prefetch

- HW Prefetch
  - Connects to L2

- L2
  - Connects to AMBA CHI

- AMBA CHI
  - System interface 32B DAT channels

- 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

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**

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

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

1. Performance is estimated for SPEC CPU® 2017

+2.9% SPEC CPU® 2017 Integer
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**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decode bandwidth</strong></td>
<td>Increased decoder lanes from 5 to 6</td>
</tr>
<tr>
<td></td>
<td>Increased Decode Queue from 16 to 24 entries</td>
</tr>
<tr>
<td><strong>Rename checkpoints</strong></td>
<td>Increased from 5 to 6 total checkpoints</td>
</tr>
<tr>
<td></td>
<td>Increased from 3 to 5 vector checkpoints</td>
</tr>
<tr>
<td><strong>Rename rebuild</strong></td>
<td>Improved rebuild flows for more efficient</td>
</tr>
<tr>
<td></td>
<td>recovery after pipeline flush</td>
</tr>
</tbody>
</table>

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

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

+3.3% SPEC CPU® 2017 Integer

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

uArch features new with Neoverse V2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLB</td>
<td>Increased from 40 to 48 entries</td>
</tr>
<tr>
<td>Replacement policy</td>
<td>Changed from PLRU to dynamic RRIP</td>
</tr>
<tr>
<td>Larger Queues</td>
<td>Store Buffer</td>
</tr>
<tr>
<td></td>
<td>ReadAfterRead</td>
</tr>
<tr>
<td></td>
<td>ReadAfterWrite</td>
</tr>
<tr>
<td>Efficiency</td>
<td>VA hash based store to load forwarding</td>
</tr>
<tr>
<td>Reduced flushes</td>
<td>RAR hazards tracked through L2 cache lifetime</td>
</tr>
</tbody>
</table>

+3.0% SPEC CPU® 2017 Integer

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 + Best Offset
  + Stride + Correlated Miss Cache
  + Page
+ 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

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

1. Performance is estimated for SPEC CPU® 2017 Integer

+5.3% SPEC CPU® 2017 Integer

© 2023 Arm
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**

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

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

1. Performance is estimated for SPEC CPU® 2017
Neoverse V2 Performance Uplift over Neoverse V1

Arm Neoverse V2

Core
- Branch Predictor
- Instruction Fetch
- Mop$ Fetch
- Decode Queue
- Decode
- Rename / Dispatch
- Issue Queues

Load-Store Pipelines
HW Prefetch
L2
AMBA
CHI

Integer Pipelines
Advanced SIMD Pipelines

Unit uplift over Neoverse V1

- 2.9%
- 2.0%
- 0.8%
- 0.9%
- 0.7%
- 3.3%
- 3.0%
- 5.3%
- 0.5%

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 mm² with 1 MB L2
1.2 W*

Neoverse V2 with 2 MB L2
Typical 5nm implementation
2.50 mm² with 2 MB L2
1.4 W*

*SiR at 2.8 GHz, 0.75 V, H280, 16LM, 17LM
Arm Neoverse V2 Platform IP

- Processor Block
  - V2 CPUs
    - L1$
    - L2$
  - Direct Connect DSU
  - I/O Macro
    - NI-700
    - MMU-700

- PCle G5
- CXL

- PCIe Gen5 & CXL
  - I/O Expansion
  - System Control Processor (SCP)
  - Manageability Control Processor (MCP)
  - Debug

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

- Memory Expansion
  - DDR5

- Chip-to-Chip Expansion

- Peripheral Block
- NIC-450
- GIC-700
- 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
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.
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

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.

80% reduction in branch mis-predicts
70% reduction in unused prefetches
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

Server Performance

5 MW Data Center Throughput

Data provided by NVIDIA
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 Integer¹
  - 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

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

- 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
ありがとう
Asante
Merci
감사합니다
धन्यवाद
धन्यवाद
شكرًا
多谢
谢谢