Arm Neoverse N1
Cloud-to-Edge Infrastructure SoCs

Andrea Pellegrini
Senior Principal Engineer
Infrastructure Line of Business

Chris Abernathy
Senior Principal Engineer
CPEG

HotChips Conference
Aug 19th 2019
New infrastructure for support of 1T connected devices

Cloud datacenter

Core network

Virtualized provider edge

Virtualized customer premise

1 Trillion Autonomous Machine & Smart Devices

WiFi

Small cell

BTS

V2X

IoT gateways
Arm Limited – Hotchips 2019

+40% perf

30% Higher System Performance per Generation + New Features

16/14nm

Cosmos Platform
A72

Previous

7nm

N1 Platform

Today

7/5nm+

Zeus Platform

Future

5nm

Poseidon Platform

30% Higher System Performance per Generation + New Features
# Neoverse scalable compute platforms

**Arm CPU**
- Neoverse N1

**Arm SoC Backplane**
- Coherent Mesh Network (CMN)
- GIC, MMU Virtualization
- Power Control Kit
- CryptoIsland Secure Enclave

**Memory**
- DDR4
- DDR5
- HBM

**IO**
- PCIe
- CCIX
- GbE

**Common Software Platform, SBSA, SBBR, Arm ServerReady**
- Arm Architecture v8.x-A, AMBA

---

**GIC**: Generic Interrupt Controller  
**MMU**: Memory Management Unit  
**CCIX**: Cache Coherent Interconnect [www.ccixconsortium.org](http://www.ccixconsortium.org)
Neoverse N1 hyperscale reference design

64x-128x Neoverse N1 CPU
8x8 Mesh up to 128MB of cache
128x PCIe Gen4 lanes
8x channels of DDR4 memory
2x-4x CCIX links

Arm provides
- PPA data for 7nm implementation
- IP Configuration parameters
- Register programming
Architected to support custom accelerators

Example CMN-600 mesh design

CML: Coherent Mesh Link

3rd Party PCIe/CCIX combo controller

Customer-specific innovation

Arm provided IP

Third party IP
Scale with Coherent Mesh Network (CMN)

Scalable coherent mesh network with XP for network links and device ports

- Custom sizing from 1x2 up to 8x8

Built for high frequency, non-blocking AMBA CHI transactions with atomics:

- Latency: ~1 cycle/hop at XP
- Max bandwidth per link/direction: 256bits*2.2GHz

Integrated system level cache (SLC) shared by CPUs, Accelerators and IO with cache stashing and RAS capabilities
N1 CPU: Every facet of the design optimized for sustained performance

Infrastructure performance focus

- Hardware I-cache coherency
- 1MB private L2 cache
- Streamlined Direct-Connect to N1 interconnect

Fully Armv8.2 compliant

- Server-class RAS system support
- Infrastructure-specific architecture features

Market leading power efficiency

- $+30\%$ over previous generation Cortex-A72 CPU (iso-process)
Neoverse N1 CPU Power/Performance/Area

Industry-leading power efficiency

- 1.0-1.8W / core+L2 (2.6-3.1GHz)

Industry-leading area efficiency

- 1.15-1.4 mm² core + L2 (512K/1M L2)

World-class 64-core total-system performance

- ~190 SPECint_rate2017 (est.) / 105W SoC power

Data presented here has been collected on models as well as on early samples of the Neoverse N1 Software Development Platform.
Spec CPU2017 silicon performance (4 vCPU)

Neoverse N1 systems performance expected to match or exceed currently available cloud instances

Many hardware improvements including

- Larger private caches
- Four-wide front-end
- Load-Store queue optimizations and data prefetchers
- FP units improvements

Partners can achieve higher performance with core configurations, SoC architectural/silicon improvements and software tuning

Significant Spec CPU2017 int (est.) performance uplift over Cortex-A72

Neoverse N1 Speedup vs. Cortex-A72

Ubuntu cosmic gcc 8.3.0 glibc 2.28 -mcpu=cortex-a57 -O3 -fomit-frame-pointer
Throughput app: characterization of NGINX

An example for a popular high throughput cloud application

NGINX is an example of highly concurrent, high performance web server, reverse proxy, and API gateway

Performance for this kind of workload is directly related to:

1. Memory latency and bandwidth to receive the request and transfer data
2. Overhead to context switch between user and kernel space
3. Efficiency of the CPU front-end to fetch instructions

These stressors are very common with similar applications such as: MemcacheD, HHVM, ...

Neoverse N1 CPU

Armv8.2-A
32b/64b CPU

AdvSIMD™ SIMD engine
Crypto extensions
64K I-Cache w/parity
64K D-Cache w/ECC
Private L2 cache (512KB/1MB) w/ECC

Direct-Connect to CMN-600 Mesh CHI

Arm CoreSight™ Multicore Debug and Trace
Microarchitecture optimized for infrastructure

Caching structures sized for large, branch-heavy infrastructure workloads

- 64K I$/D$, 512K/1M low-latency private L2, 64-bank 128M system-level cache (SLC)
- 6K Branch Target Buffer, 8K bimodal +5k history-based predictors
- High-capacity hybrid indirect-branch predictor
- 48-entry ITLB/DTLB, 1280-entry L2 TLB with flexible leaf(descriptor caching/acceleration and page-aggregation

Optimized for heavy OS/Hypervisor activity

- Low-latency high-bandwidth context save/restore that minimize context-synchronization serialization
Throughput app: Improvements vs Cortex-A72 on NGINX

We expect Neoverse N1 to deliver competitive levels of performance for this kind of application.

Neoverse N1 significantly improves:

<table>
<thead>
<tr>
<th>Workload stressor</th>
<th>Neoverse N1 Features</th>
<th>N1 improvement over Cortex-A72</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory latency and bandwidth</td>
<td>Cache stashing</td>
<td>Not available on A72</td>
</tr>
<tr>
<td></td>
<td>Memcpy bandwidth</td>
<td>2x increase</td>
</tr>
<tr>
<td></td>
<td>Larger and faster caches</td>
<td>L2: up to 4x larger, 66% faster access</td>
</tr>
<tr>
<td>Context switching</td>
<td>Context switching</td>
<td>2.5x faster</td>
</tr>
<tr>
<td>CPU front-end</td>
<td>Branch mispredicts</td>
<td>7x reduction</td>
</tr>
<tr>
<td></td>
<td>L1 cache misses</td>
<td>2x reduction</td>
</tr>
<tr>
<td></td>
<td>TLB misses</td>
<td>3x reduction</td>
</tr>
</tbody>
</table>

![Graph showing Nginx throughput in rps](image)
Runtime environments workloads: characterization

Object management:

- Almost everything is an object, and each object needs to be allocated, initialized, prefetched, and garbage collected

Managing the instruction footprint:

- Jitted code can put a tremendous pressure on caches, TLBs, and core front-end

Garbage collection:

- Requires both low memory latency and high bandwidth, as well as fast synchronization between the cores
Microarchitecture optimized for infrastructure

Memory hierarchy design for low latency, high bandwidth, extreme MLP*, and scalability

- Predict-directed-fetch front-end (Iside MLP)
- 46 outstanding system-transactions, 32 outstanding non-prefetch transactions
- 68 in-flight LDs, 72 in-flight STs

Support for both near and far Arm Atomic instructions

Extensive system-level co-optimization

- Designed for high bandwidth, low latency, and high-core-count scalable systems

*Memory Level Parallelism
Neoverse N1 cache hierarchy

First Arm CPU with coherent 64K I-cache
- Removes bottlenecks from high core count systems that use a lot of memory (e.g. VM setup/teardown)

64K DL1 with 4-cycle LD-use

Private, per-core L2 cache
- 512KB-1MB private L2 with 9-11 cycle LD-use latency
- Adaptable to system latency/bandwidth characteristics

Direct-connect to shared system cache (SLC)
- High-bandwidth / multi-banked / high-capacity
- 22ns LD-use in 64-core, 32-bank, 64MB SLC system
## Runtime environments workloads: Improvements vs Cortex-A72

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

![Java Based Benchmark (throughput)](image)
Scalability from edge to hyperscale configurations

**Neoverse N1 Hyperscale Reference Design (single die)**

64x N1 CPUs (64T) for hyperscale performance

- 8-ch DDR4
- ~105 W SOC power
- < 70 W Total CPU budget
- ~190 SpecInt-Rate2017 (est)
- @2.6 GHz 7nm
- ~70 W
- @2.6 GHz 7nm
- >10Gbps/core
- ~20 SpecInt-Rate2017 (est)

**Neoverse N1 Edge Reference Design**

8x N1 CPUs for Control Plane and Application processing

- CMN-600
- 2-ch DDR4
- <25 W SOC power
- <10 W Total CPU budget
- ~4G LTE Backhaul LAN traffic
- ~20 SpecInt-Rate2017 (est)
Conclusions

Neoverse N1 offers comparable or better performance to popular cloud instances at a fraction of the power and silicon area

Demonstrates Arm’s commitment to the infrastructure market

This is just the beginning...

• Partners have many opportunities to deliver above and beyond these figures
• SW optimizations and workload tuning are still in progress
• Arm delivering multi-generation roadmap targeted for the infrastructure market
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.
The Cloud to Edge Infrastructure Foundation for a World of 1T Intelligent Devices

Thank You!