Arm Neoverse N2 Platform: A Significant Uplift in Cloud-to-Edge Performance Efficiency

arm

Solution Brief

AT A GLANCE

To meet the rising demands of today's customers, cloud service providers, carriers, system designers and others are increasingly turning to Arm Neoverse to provide the foundation of their digital infrastructure. Arm Neoverse brings market-leading performance and scalability to cloud data centers, enterprise networking, edge deployments and 5G networks while dramatically reducing power consumption and total cost of ownership.

WHY NEOVERSE N2 PLATFORM?

- Leading per-thread performance for cloud workloads.
- 2-3.5x better performance per watt over traditional processors in clouds and 5G networks
- Processor diversity with CPUs ranging from 8 to 128+ cores.
- Enables performant, heterogeneous compute and fanless solutions for edge and 5G.

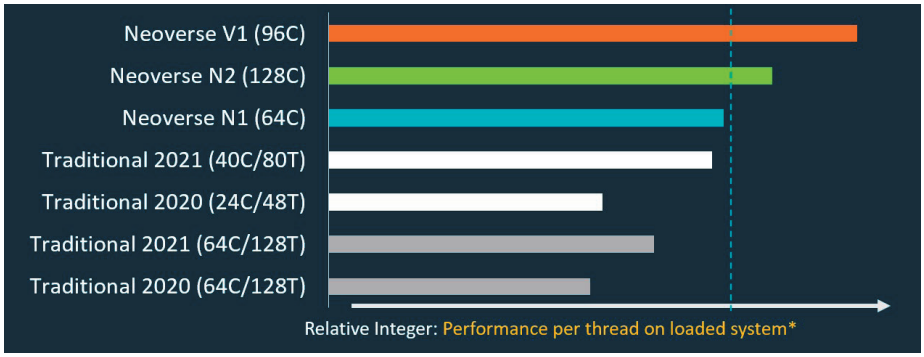
Performance Efficiency with Improved Performance Per Rack and Per Watt

Datacenter workloads are expected to grow by more than 10 times in the coming decade. At the same time, carriers and cloud providers are being tasked with managing more complex and compute-intensive applications, such as telemedicine, predictive maintenance, and smart building management. To fulfill the promise of digital transformation, cloud service providers, carriers, and independent datacenters need to rethink their infrastructure architecture to scale rapidly and economically.

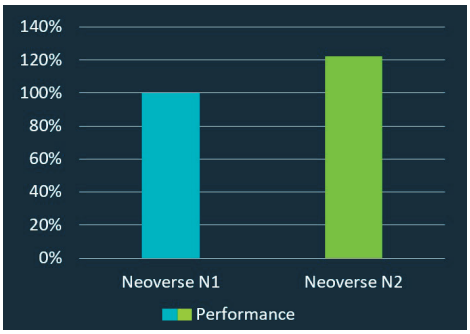
Arm Neoverse is a diversified portfolio of processor IP designed for tomorrow's infrastructure. Infrastructure silicon designed around Neoverse processor IP can be optimized for specific tasks, such as application processing or networking, or performance characteristics, including floating point performance.

Neoverse N2 combines performance and power efficiency for building scalable hyperscale datacenters, edge infrastructure, or 5G networks. Neoverse N2 is Arm's leading solution for optimizing for application servers, edge devices and wireless equipment. It provides a 40 percent IPC uplift over its predecessor Neoverse N1, a 30 percent advantage in cores per rack advantage over performance-optimized Neoverse V1, and an even greater advantage in performance and power over conventional processors. The end result is datacenters, edge equipment, and networks that can meet demanding KPIs while staying inside sustainable operating envelopes.

Relentless Improvement on Workloads: Socket Performance Throughput



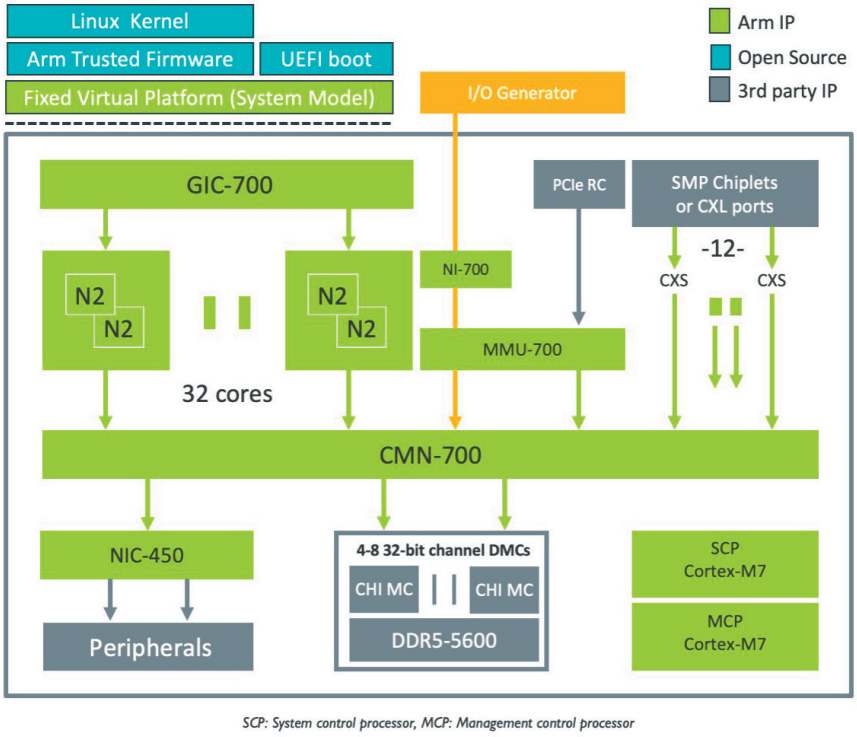
DPDK L3fwd



Uplifts 1.3x or higher on Neoverse N2

Source: Arm

The Foundation for Scalable Infrastructure



Neoverse N2 Block Diagram

Release: 2021
Manufacturing node: 7nm/5nm
CCIX: 2.0
CXL 2.0
PCI3 gen 5
DDR5
HBR3
SVE 2 x128b
Bfloat16

A Leader in Performance Per Watt

Neoverse N2 can offer a 40 percent IPC performance uplift over Neoverse N1 within the same power envelope. Neoverse N1 already offers 2 times to 3.5 times the performance per watt over contemporary traditional processors.

The balance of performance and power efficiency makes single-threaded Neoverse N2 the leading choice for tasks such as cloud computing, network acceleration, general purpose application processing, security, and packet processing.

Diversity in Processor Implementations

While Arm specifies the Neoverse architecture, our semiconductor partners determine the ultimate design and features of Neoverse processors. Neoverse N2 processors range in core count (which currently ranges from 8 to 128 cores), core type (i.e. integrated GPUs, NPUs and other heterogeneous accelerators) speed, cache size, I/O, and other features.

A Broad Ecosystem of Hardware and Software

Neoverse N2-based equipment is targeted for cloud and hyperscale servers, fanless edge servers, sealed (but upgradeable) universal customer premises equipment, general purpose servers, virtualized 5G equipment and cloud-based 5G deployments. Software developers meanwhile have ported a range of operating systems, middleware, applications, and tools to Neoverse.

Scalable Vector Extension 2

Vector processing is one of the most forceful tools in computing, but traditional processors only handle vectors of specific widths. Scalable Vector Extension 2 (SVE 2), a feature of Neoverse N2, allows developers to deploy and run code containing ranging from 128 bits to 2048 bits on Neoverse N2 processors without recompiling, enabling faster time to market and greater choice for customers.

- With traditional architectures, every time a new vector length is introduced in hardware, code has to be rebuilt and optimized to take advantage of the additional vector bandwidth.
- SVE and SVE2 are vector length agnostic SIMD instruction sets that allow users to write/optimize code once, compile once and be able to run on a diverse set of hardware while taking full advantage of the vector available bandwidth. And as process scaling allows us to build larger vector machines, code written and compiled today using SVE/SVE2 will automatically scale to these large machines.
- SVE and SVE2 simpler programming model enables compliers to much more easily auto-vectorize your code. This will allow programmers to take advantage of vectorization without doing anything special!
- SVE2 builds on the foundations of SVE to bring scalable SIMD vector performance and advanced auto-vectorization capabilities to a wider range of software, including ML, DSP, Regular expressions, and 5G RAN
- Neoverse N2 continues to fully support NEON for preexisting/pre-optimized code, however, focusing new development/optimizations efforts on SVE2 will greatly increase the useful life of the SW and ROI on the development efforts.

CMN-700

CMN (Coherent Mesh Network) 700 is the high-speed interconnect connecting all CPU, caches, memory, IO, and accelerator elements. To support high core count systems with good performance, CMN-700 includes several software controlled features to dynamically control resources. One such feature is MPAM (Memory Partitioning and Monitoring) that enables fair-sharing of system-level shared resources like SCL cache and DRAM memory bandwidth. CMN-700 supports both the CCIX and CXL protocols, for chiplet, multichip and multisocket support. CMN-700 is architected to support PCIe Gen5 transfer rates along with a low-latency path to DDR5 and CXL-attached pools of memory.

Bfloat16

Neoverse N2 supports Bfloat16, which helps accelerate the ML training and inference without having to translate data format. Bfloat16 also offers 4x the per-core vector throughput of Neoverse N1.

For more information, please visit www.arm.com/products/silicon-ip-cpu/neoverse/neoverse-n2.

* Traditional current data and Neoverse N1 data is measured by Arm. Traditional next data is projected by Arm. Arm Neoverse performance data is estimated by Arm pre-silicon, on Arm Neoverse V1 and N2 Reference Designs:
Neoverse V1: 96 cores, 2.6GHz, 10xDDR5-4800
Neoverse N2: 128 cores, 3.0GHz, 8xDDR5-4800