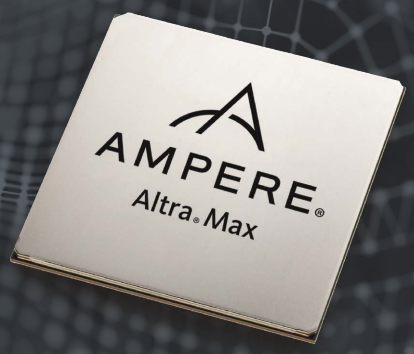
Workload Brief

NGINX Web server on Ampere® Altra® Max



Ampere®—Empowering What's Next

The Ampere® Altra® and Ampere® Altra® Max processors are complete system-on-chip (SOC) solutions built for cloud native applications. Ampere Altra supports up to 80 cores and Ampere Altra Max supports up to 128 AArch64 cores. In addition to incorporating a large number of high-performance cores, the innovative architecture delivers predictable high performance, linear scaling and high energy efficiency.

NGINX is an open source, high performance HTTP server and reverse proxy with many other web service-related features bundled. It is often used as load balancer in the cloud. NGINX implements event driven architecture to handle incoming requests. It is built to offer low memory footprint and high concurrency. NGINX is the most popular web server among high-traffic websites as of 2021.

In this workload brief, we compare Ampere Altra Max M128- 30 to the Intel® Xeon® 8380 and the AMD EPYC™ 7763 running NGINX while measuring the throughput and latencies on each of these processors.

NGINX on Ampere® Altra® Max

Ampere Altra Max is designed to deliver exceptional performance for cloud native applications like NGINX. This is achieved by using an innovative architectural design, operating at consistent frequencies, and using single-threaded cores that make applications more resistant to noisy neighbor issues. This allows workloads to run in a predictable manner with minimal variance under increasing loads.

Ampere processors are designed to deliver exceptional energy efficiency. This translates to industry leading performance/watt capability and a lower carbon footprint.

Benefits of running NGINX on Ampere Altra Max

- **Cloud Native:** Designed for maximum performance and scalability for cloud customers, Ampere Altra Max can deliver up to 3.19x higher performance than the best x86 servers for cloud native workloads like NGINX.
- **Energy Efficiency:** With up to 128 energy-efficient Arm cores, Ampere Altra Max can consume 16% lower power while delivering 3x better raw performance than the best x86 servers
- **Lower Carbon Footprint:** Industry-leading performance and high energy efficiency result in Ampere Altra Max demonstrating up to 3.76x higher Performance/Watt, leading to lower TCO and a smaller carbon footprint
- **Consistency & Predictability:** Single threaded cores running at fixed maximum frequencies ensure linear scaling under stringent SLAs and at high loads while running auto scaling NGINX instances

Benchmarking Configuration

An open-source http benchmarking tool, wrk, is used as the load generator. The wrk application running on the client system generates simultaneous HTTP requests over HTTPS connections to NGINX running on the target system. The test was configured to run with multiple threads and connections.

On the server side, NGINX is configured to serve static HTML file over https protocol using Brottli compression (quality 5). The server redirects the incoming request URL using LuaJIT, which does all the regex processing. In order to support the additional functionalities, NGINX is compiled with HTTP SSL module, openssl, brottli compression and NGINX-lua modules. The source code version of NGINX 1.15.4 and wrk v4.10 versions are used for this benchmarking.

The load generator, wrk, was configured to run with 96 clients and connections increasing from 100 to 500 in steps of 100 to observe the impact on throughput and p99 latency. We measure throughput, Requests Per Second (RPS) under SLA of 10ms p99 latency. Each test was run for 120 seconds and repeated at least 3 times for taking the average RPS and p99 across multiple runs. We observed little to no run-to-run variations in RPS or p99 latencies.

The workload was run on Ampere Altra Max M128-30, AMD EPYC 7763 and Intel Xeon 8380 (refer to the chart below for results). The same client system was used as load generator across all the platforms.

Ampere Altra Max

- 128 Armv8.2+ 64-bit cores at 3.0GHz
- 64KB i-Cache, 64KB d-Cache per core
- 1MB L2 Cache
- 16MB-32MB System Level Cache
- Coherent mesh-based interconnect

Memory

- 8x72 bit DDR4-3200 channels
- ECC and DDR4 RAS
- Up to 16 DIMMs (2 DPC) and 4TB/socket

Connectivity

- 128 lanes of PCIe Gen4
- Coherent multi-socket support
- 4x16 CCIX lanes

System

- Armv8.2+, SBSA Level 4
- Advanced Power Management

Performance

- SPECrate®2017_int_base:350

Benchmarking Results and Conclusion

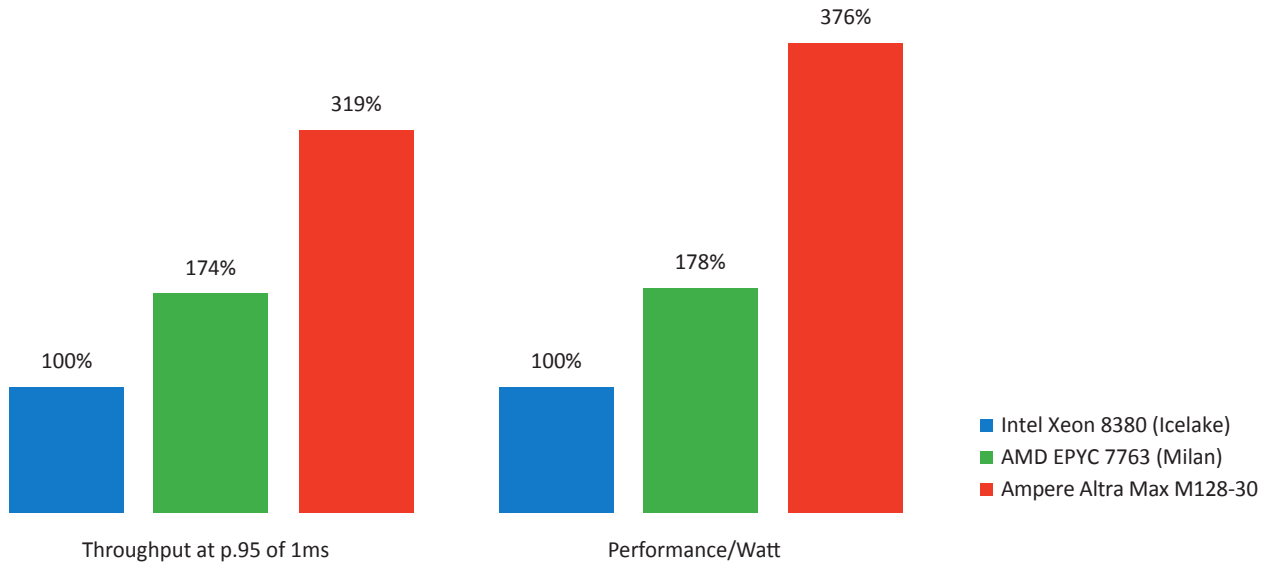
In the chart below, results are shown relative to Intel Xeon. Ampere Altra Max M128-30 has more than 3x higher performance compared to Intel Xeon 8380 and 1.74x better than the highest end AMD EPYC series.

For large-scale cloud deployments, performance/Watt (i.e. energy efficiency) is an important metric in addition to raw performance. Ampere Altra Max processors lead in performance with significant perf/watt advantage as shown the graph, up to 3.76x better efficiency

NGINX is one of the popular web servers in the cloud today. It scales well with compute resources that Ampere Altra Max series offers. Ampere Altra Max series is designed to deliver exceptional performance and energy efficiency for cloud native applications. In Ampere's testing, these processors demonstrated up to 3.19x performance improvements and they achieved up to 3.76x energy efficiency improvements.

For more information on this workload or other workloads our engineers have been working on, please visit: https://developer.amperecomputing.com/.

Ampere Altra Max M128-30 Industry-leading Performance and Energy Efficiency on NGINX



Ampere Computing® / 4655 Great America Parkway, Suite 601 / Santa Clara, CA 95054 / amperecomputing.com