

**Hewlett Packard
Enterprise**

ActualTech
MEDIA

The Ampere Altra Max Advantage

Ed Tittel

- ✓ With a high core count (up to 128), Ampere Altra is ideal for high-density, high-demand use
- ✓ For NGINX deployments, Ampere Altra Max offers up to 3x throughput, and up to 3.75x performance/watt compared to x86 servers
- ✓ Ampere Altra delivers faster, more reliable performance with less jitter and lower latency when handling HTTP requests

IN THIS PAPER

The Ampere Altra processor family (available in HPE ProLiant RL300 Gen11 servers) offers consistently high and predictable performance, high energy efficiency and a reduced physical footprint that excels at cloud-native applications like NGINX and other intensive workloads.

CONTENTS

- 2 Understanding NGINX
- 3 NGINX on Ampere Altra Max
- 4 Ampere Altra Max: Benefits and Specs
- 5 Comparative Benchmark Setup

The HPE ProLiant RL300 Gen11 product family is a collection of rack-mountable 1U servers designed to support ultra-scalable, high-volume compute capabilities for cloud-native applications. Built around the Ampere family of Altra and Altra Max CPUs, with up to 128 cores per socket, these servers deliver high performance with minimum physical footprint, high performance and superior power efficiency. In particular, the ground-breaking architecture brings predictable high performance, linear scaling, and extreme energy efficiency for cloud computing AI-oriented workloads.

NGINX delivers high performance with massive scaling, and runs at high speeds under heavy loads.

Understanding NGINX

NGINX (pronounced “engine x”) is open source web server software that also covers such functions as reverse proxy, load balancing, email proxy, and HTTP caching. It was designed specifically by its original developer, Igor Sysoev, to handle 10,000-plus concurrent user connections (aka “the C10K problem”). NGINX delivers high performance with massive scaling, and runs at high speeds under heavy loads. Plenty of well-known online players use NGINX to manage heavily visited pages, such as Facebook, LinkedIn, GitLab, Intel, Microsoft, IBM, Google, and Cisco (to name just a few).

Indeed, NGINX is an ideal test case through which to present the capabilities of the HPE ProLiant RL300 Gen11 servers, and their Ampere Altra and Altra Max CPUs. NGINX architecture centers around event-driven handlers to accommodate high-volume incoming resource requests within a low memory footprint and high degrees of concurrency. That's why it has become the most popular web server for high traffic websites since the early 2020s.

NGINX on Ampere Altra Max

The Ampere Altra CPUs—especially the top-of-the-line Max models—are designed to deliver exceptional and consistent performance for cloud-native applications such as NGINX. These Altra CPUs operate at a constant frequency to easily support a single, shared view of time thanks to clock synchronization that keeps activities working well together, supports event sequencing and correlation, and ensures successful data transfers among individual cores.

Such across-the-processor timing consistency also helps with tracking network usage, latency, and pinpointing security events in time. In turn, that helps servers coordinate more efficiently and quickly across network links, to extend the benefits of smooth data exchange and coordinated handling of parallel actions and I/O requests across clusters and racks. Overall, precise timing and coordination is what makes Altra CPUs excel at load workload distribution and management, and improved overall response time when handling web requests.

Overall, precise timing and coordination is what makes Altra CPUs excel at load workload distribution and management, and improved overall response time when handling web requests.

Dealing with Noisy Neighbors

[Noisy neighbor](#) describes a situation where multiple VMs or application threads consume large quantities of shared resources (e.g. CPU, memory, disk I/O, network bandwidth) and operate concurrently on the same server. Unless managed carefully, this can impact performance and efficiency of other applications or VMs on that server. Typical issues include:

- Resource contention
- Performance variation, with a drag on overall server performance
- Latency spikes, slow response times or (worst case) service interruptions

Noisy neighbors are best handled by managing resources carefully, minimizing variation, and dynamic scaling for ever-changing and ever-shifting workloads.

The Ampere Altra Arm processors are designed for maximum single-threaded capability with high core counts and access to cache, memory, and I/O buffers to prevent noisy neighbors from impacting individual cores and the overall runtime environment (all cores). This applies across workload level, including heavily loaded servers where 90%+ cores are busy (for Altra Max, that means 115 of 128 cores, or more).

Ampere Altra Max: Benefits and Specs

Ampere's Altra Max servers confer numerous benefits on those who put them to work. These benefits include the following attributes or capabilities:

1

Cloud native: Altra CPUs are specifically designed for use in cloud data centers or edge locations, for maximum density and minimal provisioning requirements (power, cooling, deployment space).

2

Energy efficiency: Altra CPUs typically consume at least 20% to 30% less power than legacy x86 CPUs, and will often exceed those savings by a wider margin.

3

Lower carbon footprint: Because they consume less power, require less rack space (and volume), organizations can either increase handling capacity in the same space, or reduce overall power, space and cooling requirements for the same capacity as equivalent legacy servers need.

4

Consistency and predictability: Altra CPU cores run single-threaded, at consistent clock speeds yet offer predictable performance even at the heaviest load levels. This means reduced latency, less jitter, smoother and faster operations, and better overall performance.

BASIC AMPERE ALTRA MAX SPECIFICATIONS

	<ul style="list-style-type: none">■ 128 64-bit cores run at 3.0GHz■ 64KB instruction and data caches per core<ul style="list-style-type: none">■ 1MB L2 cache per core
Memory	<ul style="list-style-type: none">■ 8x72-bit DDR4-3200 channels■ ECC & DDR RAS■ Up to 16 DIMMs (2 DPC) with 4TB addressable RAM
Connectivity	<ul style="list-style-type: none">■ 128 lanes of PCIe Gen4■ Coherent multi-socket support■ 4x16 CCIX lanes
System	<ul style="list-style-type: none">■ Armv8.2+, SBSA Level 4■ Advanced Power Management
Performance	<ul style="list-style-type: none">■ SPECrate 2017Integer Est: 350

Comparative Benchmark Setup

In lab testing, HPE used the [open source HTTP benchmarking tool wrk](#) to generate an HTTP benchmarking load. On a client system, *wrk* creates simultaneous HTTP requests via HTTPS connection to NGINX on a target server, configured to run with multiple threads and connections. On the server side, NGINX serves up static HTML files through secure HTTP (HTTPS) to redirect and load balance parallel requests. As **FIGURE 1** shows, this lets Ampere Altra Max outperform one x86 CPU (CPU 1) by a throughput factor of 3.19, and another (CPU 2) by a factor of 1.83. On a performance per watt basis, the Ampere to CPU 1 ratio is 3.76, and the Ampere to CPU 2 ratio is 2.11. That's considerable!

For large-scale cloud computing scenarios, performance/watt may even be more important than performance, because energy outlays represent outright costs, while performance differentials may or may not translate into equivalent productivity gains.

For large-scale cloud computing scenarios, performance/watt may even be more important than performance, because energy outlays represent outright costs, while performance differentials may or may not translate into equivalent productivity gains.

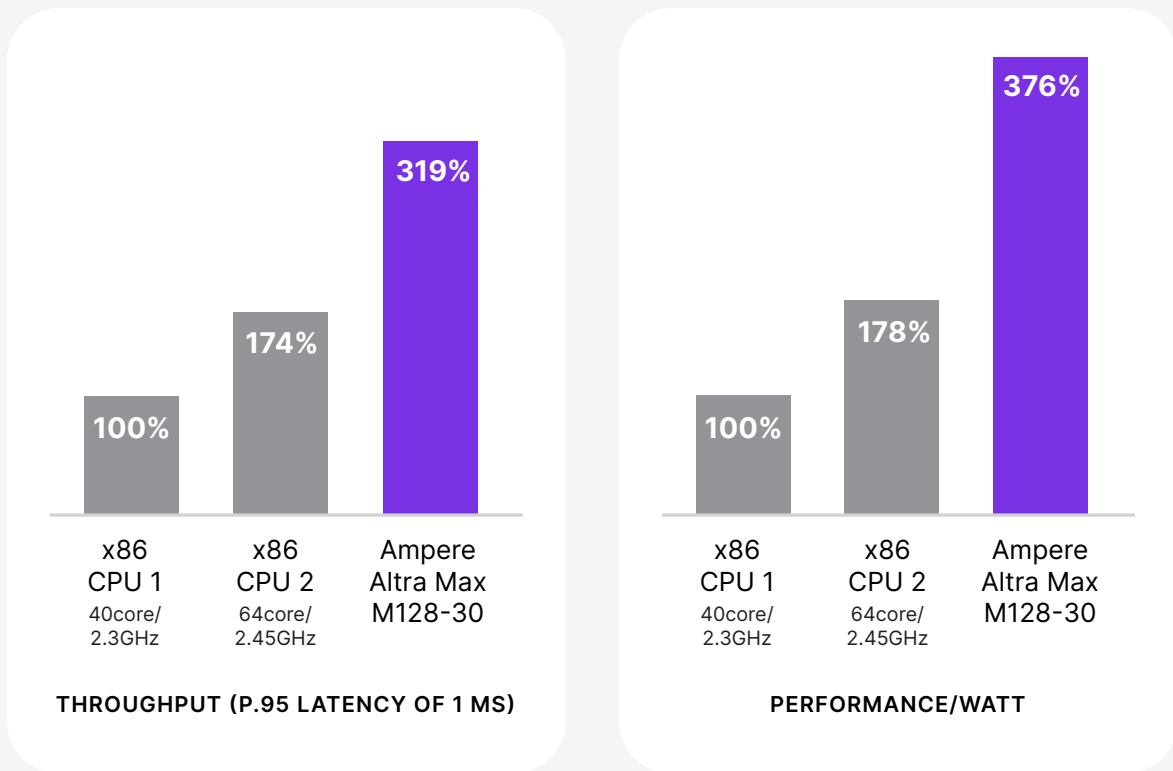


FIGURE 1: Throughput left, performance/watt right, Ampere Altra Max outpaces both x86 CPUs

LEARN MORE

No matter how you look at the Ampere Altra CPUs—especially the Mac M128-30 used for the various benchmarks mentioned in **FIGURE 1**—the HPE ProLiant RL300 Gen11 servers deliver massive advantages in both performance and energy consumption to their owners. Then, too, a 3x reduction in physical footprint means buyers can either reduce their space requirements or triple their workload-handling capabilities. These factors are well worth considering when refreshing existing NGINX deployments (and other compute-intensive workloads) or costing out new ones. Visit the [HPE ProLiant RL300 Gen11](#) product pages, and Ampere’s [NGINX Workload](#) brief for more information.

**Hewlett Packard
Enterprise**