Hewlett Packard Enterprise

ActualTech MEDIA

HPE ProLiant RL300 Gen11 Servers Deliver Disruptive Value

Ed Tittel

- Up to 3.75x lower energy costs for the same workloads
- Faster, more consistent and predictable performance makes for improved user experiences, smoother networking, and better handling under heavy loads
- Make the most of cloud-native applications and services without needing GPU support

IN THIS PAPER

Learn how the HPE ProLiant RL300 Gen11 server family, with its Ampere Altra (and Altra Max) CPUs can decrease energy usage and increase data center space utilization. In fact, it handles cloud-native applications and services well enough to disrupt typical cloud compute scenarios for competitive advantage.

CONTENTS

2 The Heart of the Beast: Ampere Altra Max

3 What Makes Predictable Performance Disruptive?

4 What Makes Power Efficiency Disruptive?

4 What Makes Server Scalability Disruptive?

5 Exploring HPE ProLiant RL300 Gen11 Servers with Ampere Altra Max

HPE ProLiant RL300 Gen11 servers come in small 1U rack-mountable form. But lurking inside this compact, standard package is a beast of a device that can shake up the way organizations use cloud computing and its many advanced and widely distributed services. With nearly four-fifths of all businesses already using multiple public or private clouds (78% use more than three public clouds), making the most of such usage pays off in all kinds of interesting ways. Delivering and using the cloud much, much better than others can, however, is what makes such servers potentially disruptive in terms of the financial returns and competitive advantages they can deliver.

The Heart of the Beast: Ampere Altra Max

As compared to other leading blade and 1U Server processors—e.g., x86-based servers—Ampere Altra Max delivers a huge win-win-win situation. Those various wins come in the following forms:

1

Predictable performance: Up to 3.2x (318%) more throughput following strict P99 latency SLA (see "Understanding the P99 Latency SLA" for more info on this service level peg) as compared to other x86-based CPUs.

2

Power Efficiency: Consumes up to 73% less power to deliver the same performance as identical workloads running on the aforementioned x86-based CPUs (a 2.7 multiplier).

3

Scalability: Only 31% as many Altra Max processors are needed to process 1 million HTTP requests as compared those other x86-based CPUs (a 3.2 multiplier).

For more information about the benchmarks that produced these results, see the Ampere briefs on NGINX cited at the end of this article.

Understanding the P99 Latency SLA

In statistics, the 99th percentile represents some value below which 99% of data points in a set of measurements or results fall. Thus, a P99 latency value as an SLA sets forth a latency guarantee where all but 1% of requests come in at or under that latency interval. In other words, 99% of incoming requests should be serviced faster than the P99 SLA value so it can serve as the floor for "acceptable network delivery times" for all kinds of requests such as HTTP page loads, read or write request for I/O, database lookups or commits, and so forth. The benchmarks used to produce the speed comparisons between Ampera Altra and other x86 processors here all use a 10ms value for HTTP requests, as is typical in many Web-based scenarios.

What Makes Predictable Performance Disruptive?

One of the key features in the Ampere Altra processor family is a fixed 3.0 GHz clock that operates only at that speed. This produces consistent, constant clock timing for all cores, and for all on- and off-chip actions and transfers. That approach supports predictable performance for high-volume compute- and I/O-intensive applications and services. These include such things as NGINX, compute- and I/O-intense workloads such as CAD/CAM, 3D Modeling, complex simulations, and more.

One of the key features in the Ampere Altra processor family is a fixed 3.0 GHz clock that operates only at that speed.

The benefits that come with consistent, constant timing support simplified and faster inter-process communication, synchronization, network and I/O access, and a whole lot more. Ultimately, a shared notion of time and easy interaction also does the following:

1

Provides a solid, reliable foundation for easy server, cluster, and rack-level scale-up and scale-out

2

Enables organizations to deliver excellent user experiences even for the highest-traffic websites (consistent performance limits variation and jitter at speeds that meet or exceed timings needed to deliver positive user experiences)

3

Lets organizations do more, show more, and deliver more to keep up with increasing demand and ongoing innovation. Even when tackling "greenfield opportunities"—Ampere Altra processors help organizations handle such workloads quickly, consistently, and efficiently.

What Makes Power Efficiency Disruptive?

Direct power consumption is an ongoing and recurring operational cost when providing computing services of any kind. Likewise, cooling costs typically represent 30% to 55% of a data center's overall energy consumption. By reducing power consumption significantly, Ampere Altra processors not only deliver cloud computing services with lower direct energy costs, they also result in lower cooling costs. That provides a multiplicative effect whereby Ampere Altra Max delivers nearly 4x performance/watt compared to other x86-based server CPUs. Indeed, doubling or quadrupling power efficiency for the same workloads greatly reduces the energy costs involved.

Such savings opens doors that can disrupt existing competition. On the one hand, companies and organizations can deliver the same workloads at lower costs for energy, cooling, and space. This lets them compete on price. On the other hand, players with Ampere Altra servers can use the same space and power to deliver up to 2 to 3 times as much compute and service capability as they could before. This lets them compete on volume, capacity, and scale-up terms.

Even when tackling "greenfield opportunities"—Ampere Altra processors help organizations handle such workloads quickly, consistently, and efficiently.

What Makes Server Scalability Disruptive?

Ampere Altra Max processors are disruptive because fewer of them can handle the same 1 million HTTP requests (roughly one-third, according to internal benchmarks). As in the preceding discussion, that means either the number of processors can stay the same and the capacity and capability can approximately triple. Or the capacity and capability can stay the same and the footprint shrinks by around two-thirds. That means astounding savings on space, equipment, energy, and cooling. In turn, that translates into a greatly reduced TCO or much higher scaling and efficiency. It also opens the door to change-ups, should the provider wish to alter its operations and offer its users more opportunities to scale up or down as and when they might like.

Players with Ampere Altra servers can use the same space and power to deliver up to 2 to 3 times as much compute and service capability as they could before. This lets them compete on volume, capacity, and scale-up terms.

Exploring HPE ProLiant RL300 Gen11 Servers with Ampere Altra Max

HPE offers many cores per socket (up to 128) with cloud-native capabilities thanks to its Ampere Altra (and Altra Max) CPUs. These are engineered to provide consistent, predictable, and energy efficient performance even when heavily loaded. This compact 1U rack mount format for such servers makes them easy to install, and requires only minimal rack space to accommodate. Providers can either choose to eliminate under-used rack space, or expand their compute and services capabilities without increasing space and energy budgets and stay within sustainability goals.

As with other members of the HPE ProLiant Gen11 server family, these HPE ProLiant RL300 Gen11 servers support Integrated Lights Out (iLO) management and firmware for robust reliability and security, to make best use of HPE's Silicon Root of Trust technology. There are two kinds of HPE ProLiant Gen11 server models available depending on the Ampere Altra processor selected:

- **Ampere Altra Max M128-30 models** include 128 cores, run at 3.0 GHz, support up to 4TB of RAM, consume an average of 178 watts under heavy load[1] (the device has a 250W TDP), with 64KB of L1 Cache and 1MB of L2 Cache per core, plus 16MB of system level cache (SLC).
- **Ampere Altra Q80-30 models** include 80 cores, with 161 watts of average power consumption under heavy load, (210W TDP), but otherwise the same specifications as the M128-30 model preceding.
- **Buyers can configure the various HPE ProLiant RL300 Gen11 models with multiple DDR4 smart ECC memory modules**, across a total of 16 DIMM slots. Maximum capacity with RDIMMs is 1TB (16×64GB), or 4TB with LRDIMMs (16×256GB).
- **All HPE ProLiant RL300 Gen11 servers include two PCIe 4.0 expansion slots** (primary and secondary risers) that can accommodate a full-height, full-length x16 device in the primary, and a full-height, half-length x16 device in the secondary slot.
- **To learn more about the HPE ProLiant RL300 Gen11 offerings**, please visit the HPE ProLiant RL300 Gen11 product page and the accompanying data sheets (these will help with understanding memory, storage and networking options).

[1] Actual watts consumed varies according to the workload specifics, but the averages 178 for the Ampere Altra Max, and 161 for the Ampere Altra, are based on a broad cross-section of benchmark data from repeat testing.

LEARN MORE

Visit the HPE ProLiant RL300 Gen11 product pages, Ampere's Altra family product brief, and the NGINX and benchmarking briefs for more information.

Hewlett Packard Enterprise