TECHNICAL WHITEPAPER

A Deep-Dive into HPE ProLiant RL300 Gen11 Performance Leadership

Ampere-powered HPE RL300 servers with DDR4 memory deliver better efficiency and cost effectiveness than popular Intel and AMD DDR5 platforms

Table of Contents

Introduction

As the industry's first tier one OEM platform powered by Ampere®, the RL300 significantly improves compute density and efficiency in a flexible, and cost-effective 1U server. It gives customers improved compute efficiency to power Cloud Native Processing technology. The ProLiant server series is intuitive, trusted, and optimized, and customers of the RL300 will enjoy all the benefits they've come to know from this flagship server family.

Featuring Ampere® Altra® Cloud Native Processors, the RL300 outshines its competition in cost effectiveness and efficiency across multiple processor generations for many popular cloud native workloads.

Selected for this study were the following CPU configurations. These processors are popular choices among customers for service provider-centric cloud native workloads, and they are sized against core/thread density, power usage and cost. This provides a relevant comparison for a broad set of digital enterprises globally for whom it is not feasible to consider top bin x86 processors given the extreme cost burden and excessive power draw.

Next we outline how Ampere's significant performance efficiency and cost effectiveness allows the RL300 — powered by DDR4 memory — to outshine Intel- and AMD-powered platforms featuring next generation DDR5 memory.

CPU Configuration in SUTs

CPU Manufacturer	Ampere	Intel	AMD
Processor SKU	M128-30 (Altra Max)	6442Y (Sapphire Rapids)	9454 (Genoa)
# CPUs	1	2	1
Cores/threads per CPU	128 / 128	24 / 48	48 / 96
Frequency	3.0 / 3.0	2.6 / 4.0	2.75 / 3.8
TDP	250W	225W ea.	290W

Excellence Across Cloud Native Workloads

Our study excludes findings from commonly cited synthetic benchmarks such as Spec.org's SPECrate® 2017_int_base. We focus here on various popular cloud native workloads behind many modern enterprises and customer services, from storage to web services, video services and beyond.

Rack-level performance is key in overall data center efficiency and cost effectiveness. The performance per rack metric evaluates performance, power consumption, rack density, and overall data center footprint condensed into a single measure that can be scaled linearly for compute installations of all sizes. Ultimately, this kind of analysis guides architects and service operators to build more sustainable data center designs for the modern cloud era.

> **In this light, the RL300 advantage becomes a game changer for customers facing power and cost constraints worldwide. Achieving more performance while using less power and occupying less space generates significant efficiency gains, reduces operational and capital expense, and can help to prevent significant amounts of CO2 emissions.**

Leading in Single Server Performance Efficiency

We performed tests on the HPE ProLiant RL300 Gen11 and two competitive tier one OEM platforms with Intel and AMD processors respectively. The systems were configured with popular CPU, memory, and storage configurations and right-sized to allow an “apples-to-apples” comparison (see Endnotes for details).

In raw single server performance, the RL300 outperforms the x86 competition in most real-world workloads as can be seen in Figure 1. It is expected that the Ampere® Altra® Max platform does not deliver better raw performance in all cases to the higher CPU turbo frequency and DDR5 memory of the x86 platforms compared in this study.

Figure 1: Relative single server performance (higher is better)

Critical to analyzing efficiency is power draw at the individual platform level when under full stress load. Based on this data, we can pair the performance figured with power draw to scale linearly to rack level approximate real-world behavior. Figure 2 outlines the RL300's energy efficiency. Ampere Altra Max generates similar or lower power draw for the entire platform as the single-socket AMD EPYC Genoa.

Compared to the dual-socket Intel Xeon Sapphire Rapids platform, the RL300 consumes significantly less platform power under load (Figure 2) all the while still delivering better or similar performance (Figure 1). Doing more work with half the processors is a simple equation behind Ampere's significant efficiency advantage.

As a result, the RL300 leads in performance per watt – a metric that summarizes the superior energy efficiency of the Ampere Altra Family of processors. Figure 3 shows how Ampere is more than twice as efficient as Intel and up to 1.4x more efficient than the AMD-powered platform.

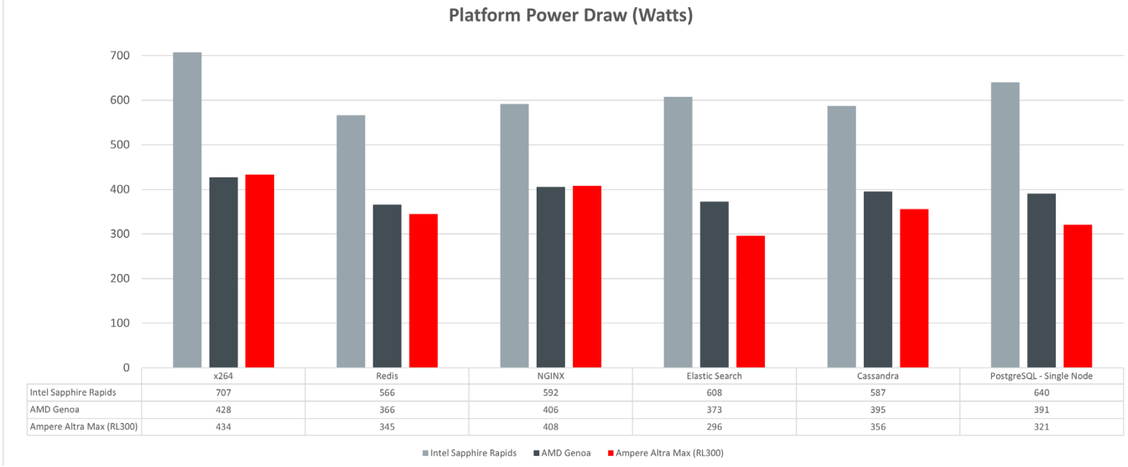


Figure 2: Raw server power draw while under load (lower is better)

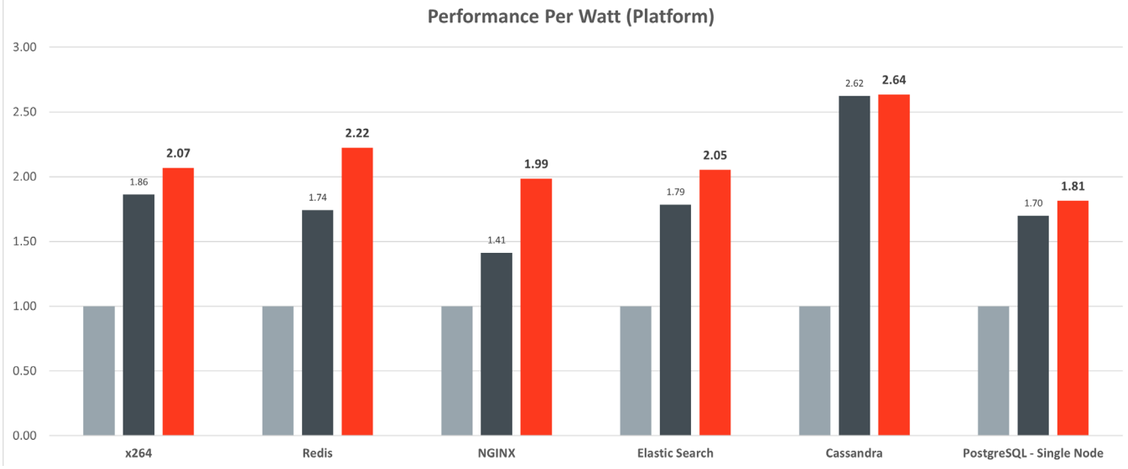


Figure 3: Relative server performance per watt (higher is better)

Leading in Scale-out Performance and Density at the Rack-level

The previous data serves as the baseline to demonstrate Ampere's growing advantage as the scale of infrastructure deployments grows. Racks are constrained by a power budget — 12.8kW in our model — and by the number of Rack Units ("U") of rack space — 38 of which are allocated to compute infrastructure (versus top of rack switches and power equipment, for instance) in our model.

The focus of this next section is to maximize the performance of a given workload when a full rack of servers is dedicated that specific workload. This analysis assumes that all total available power in a rack is used to operate servers running the given workload, and the desired outcome is to maximize total aggregate performance of all servers in said rack.

Latest generation x86 platforms can draw tremendous amounts of power (Figure 2), and thus result in infrastructure operators maxing out power budget far earlier than maxing out physical space in the rack. As an example, the dual-socket Intel server consumes 707 watts of power under full load of x264 media transcoding. Given a 12.8 kilowatt maximum power budget for servers, only 18 Intel-based 1U servers could be installed and operated in the rack (18 * 707 W = 12,726W); thus, rendering exactly 20 U of rack space unusable to operate servers. Figure 4 shows this waste of physical rack space and how it results in datacenter architects being forced to build out second and even third racks to fit within the rack power budget constraints. Each additional rack would require additional infrastructure (top of rack switches, cooling, power supplies, etc.) that add extra costs for the infrastructure buildout.

Ampere's density and energy efficiency allow more servers to be installed and operated under full load in a given rack than with competing x86 architecture platforms. Paired with similar or better performance per server (as outlined earlier), this results in far superior results at the rack level than with Intel and AMD.

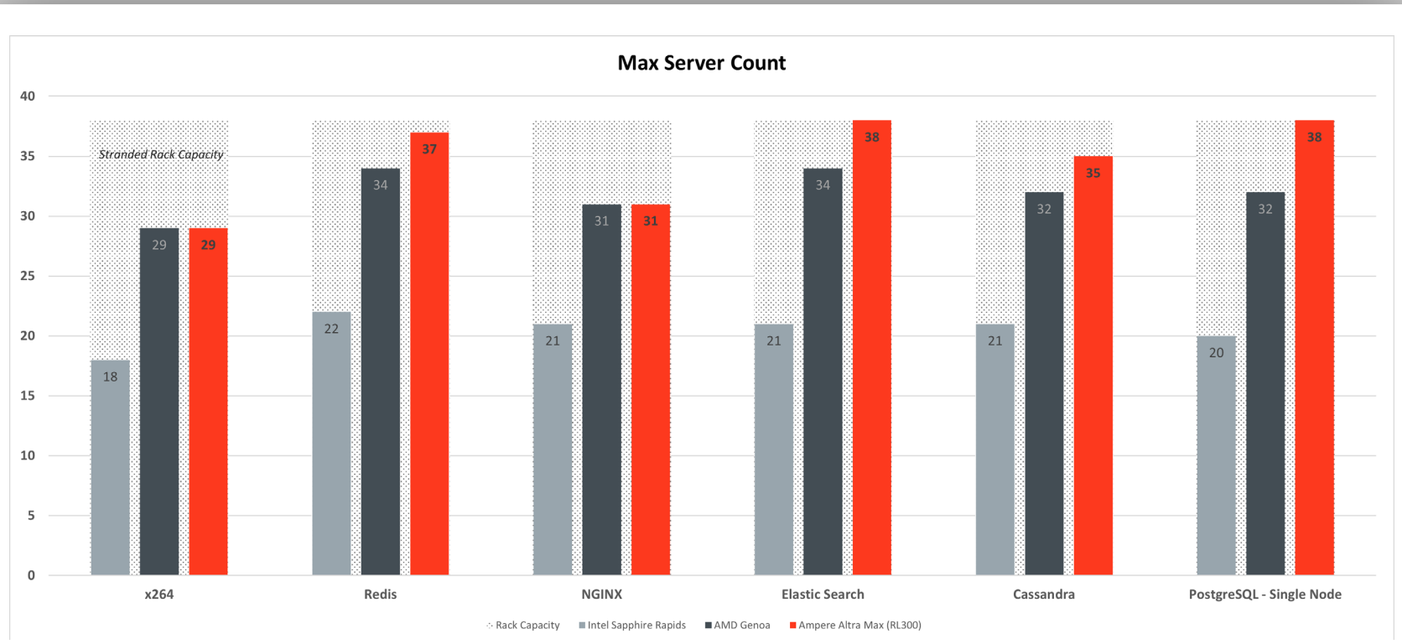


Figure 4: Maximum number of servers that can be populated in 38U rack space before hitting the power ceiling (higher is better)

Cores Per Rack

The RL300 with its 128 cores Ampere Altra Max delivers multiple times the number of CPU cores than popular x86 competitor platforms. Its lower power consumption also allows customers to populate more servers into an individual rack before hitting the power budget threshold of the rack. All in all, having both (1) denser core count per platform, and (2) more platforms per rack leads to a significant Ampere advantage. Figure 5 shows this drastically — **4 to 5 times more cores per rack with Ampere.**

Service providers and digital enterprises can leverage the greater core count of an Ampere-powered rack by increasing the number of containers or virtual machine hosted in each rack of compute. This allows expansion of service capacity without the need to expand into a second or third rack.

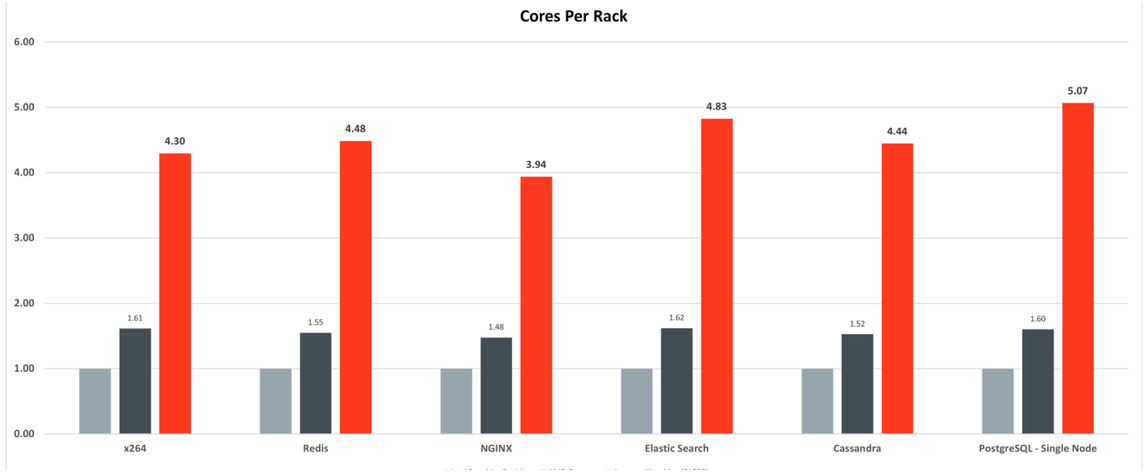


Figure 5: Relative total physical core count per rack (higher is better)

Performance Per Rack

Generally, service architects want to maximize the total performance generated by a server rack. For providers, generating more output per rack, and thus serving more customers per rack, is a direct measure of performance efficiency.

The RL300 more than doubles the rack-level performance that over an Intel Sapphire Rapids system for many popular cloud native workloads. Even against AMD EPYC Genoa, Ampere delivers up to 1.4x more performance per rack, as seen in Figure 6.

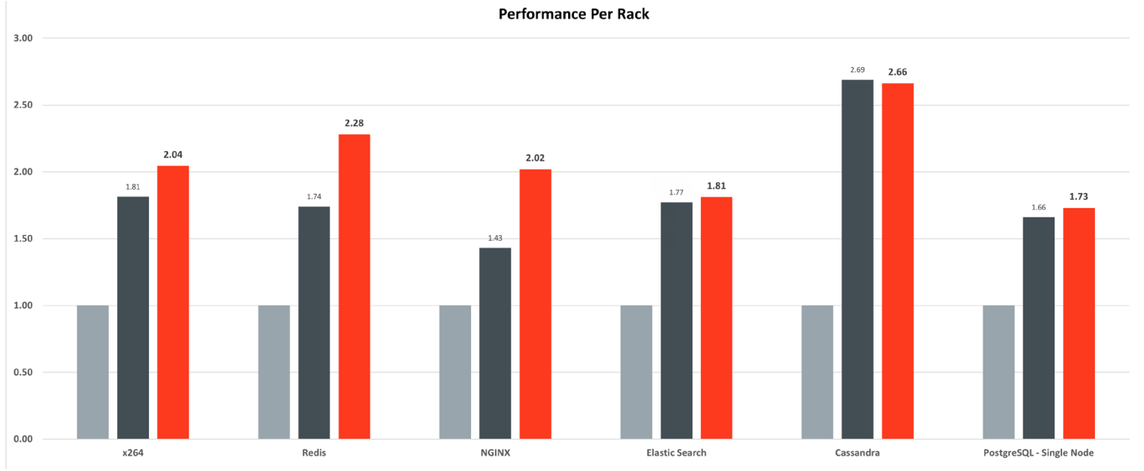


Figure 6: Relative total performance per rack (higher is better)

Reducing Power Costs for Equivalent Performance

Next, we take the performance generated by a full rack of Intel-powered servers and then determine *'how much more efficiently'* that same performance could be delivered on x86. We describe how the RL300 delivers lower operating cost and may help significantly reduce carbon footprint. This is critical for infrastructure operators motivated to consolidate footprint or minimize the impact of adding additional servers due to rapidly encroaching rack- or datacenter-level power constraints.

Here, we evaluate the cost of electricity required to power the rack of servers. Other factors such as Power Distribution Unit (PDU) efficiency and cooling infrastructure are assumed in this model to be equivalent for all platforms under comparison.

We first look at the number of servers required to match the performance of an entire rack of Intel Sapphire Rapids servers. The high power draw of the x86 platforms generally causes rack power budget to be exhausted before the physical rack capacity is reached.

In Figure 7, we list the maximum number of Intel servers that a rack can support under full load before exceeding power budget. We then calculate (based Figure 1 data) how many AMD and Ampere servers would be required to meet the performance of the rack of Intel servers. As indicated in Figure 1, there are some workloads in which an Ampere Altra Max processor does not deliver quite the same performance as an x86 processor. In the case of PostgreSQL, this means more Ampere servers are needed to deliver the same performance level as a rack of dual socket Intel servers. For all other workloads, equivalent performance to Intel can be achieved with the same or with fewer servers.

This study does not include an analysis of acquisition cost (CapEx). However, we consider power costs as the main operational cost factor (OpEx) to demonstrate the Ampere advantage. As shown in Figure 2 above, the Ampere Altra Max-based RL300 draws significantly less power than the Intel-based systems, and it draws less or similar power than the AMD-based ones. We extrapolate power costs per rack per year from these numbers in Figure 8 assuming $0.20 per KWh.

The RL300 saves a staggering $11,000 per year per rack on average for popular cloud-native workloads over Intel. Service providers may enjoy multi-million dollar energy savings even with relatively small deployments of 100 racks, thus providing an immense opportunity to invest those savings to modernize their business by introducing AI or other innovative technology stacks.

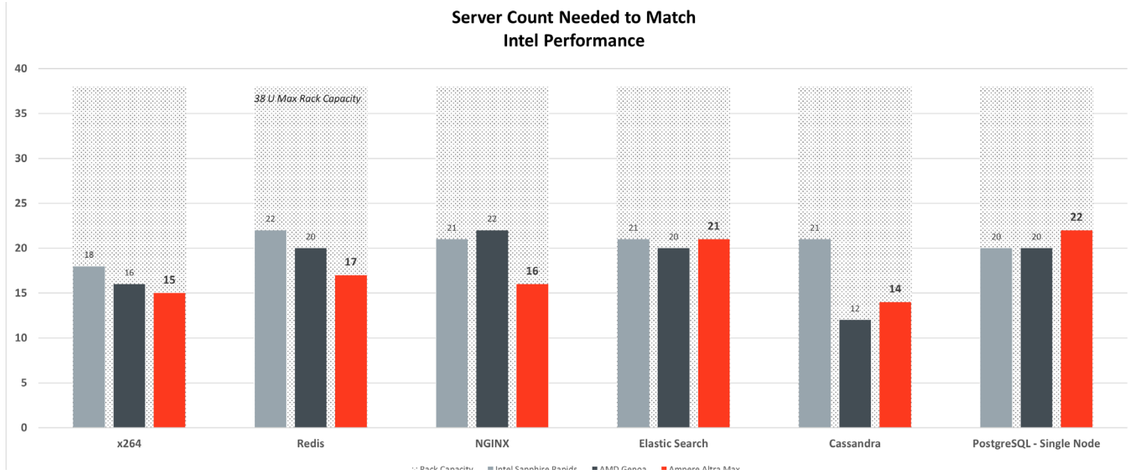
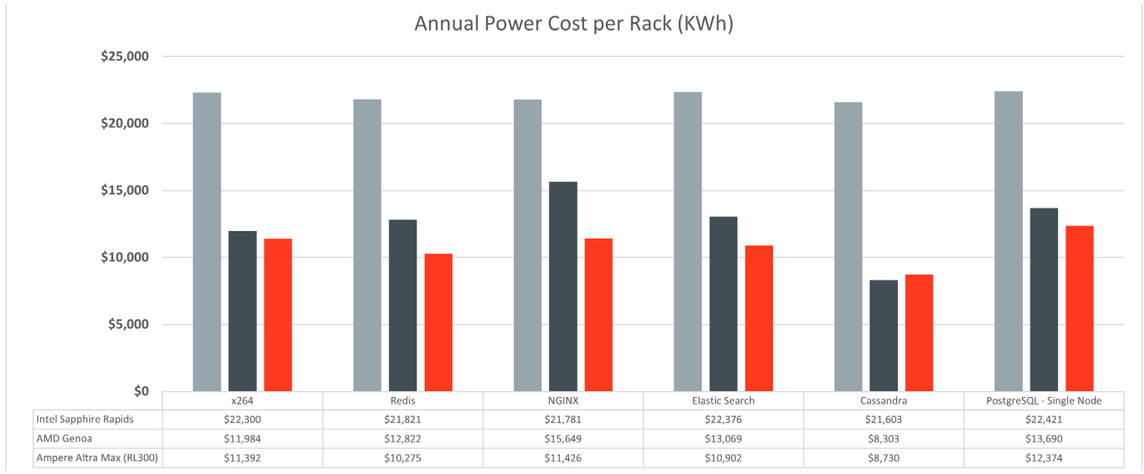


Figure 7: Number of servers needed to match Intel rack-level performance (lower is better)



	x264	Redis	NGINX	Elastic Search	Cassandra	PostgreSQL - Single Node
Intel Sapphire Rapids	$22,300	$21,821	$21,781	$22,376	$21,603	$22,421
AMD Genoa	$11,984	$12,822	$15,649	$13,069	$8,303	$13,690
Ampere Altra Max (RL300)	$11,392	$10,275	$11,426	$10,902	$8,730	$12,374

Figure 8: Power cost per rack per year at $0.20/KWh (lower is better)

In short, the RL300 is a breakthrough on the quest to introduce game-changing compute efficiency and density for the modern service provider or digital enterprise.

For the cloud native workloads we observed and the specific SKUs we tested, Intel's Sapphire Rapids and AMD's Genoa families of processors cannot deliver the same performance, performance efficiency and cost performance as Ampere's Altra Family of processors does in the HPE ProLiant RL300 Gen11 server.

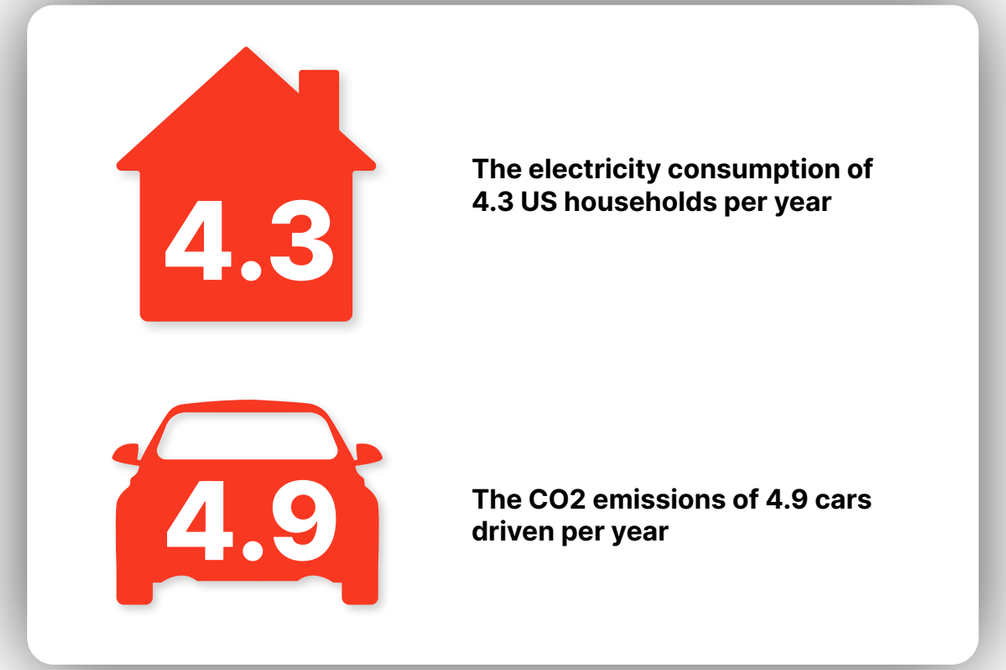
And There is More...

Ampere Altra Processors Built for Sustainable Computing

As shown, popular cloud native workloads powered by Ampere Altra Max can achieve the same performance levels with up to seven fewer servers needed per rack, and do so with around 6.5kW less in power usage. Service providers and digital enterprises can reclaim stranded rack capacity (Figure 4) and save significant cost for rack power (Figure 8). This reduction of wasted rack and datacenter space is inherently more sustainable than legacy x86 platforms. Preventing the spill into additional racks also reduces physical hardware (metal, cables, peripherals) and cooling infrastructure which carries an inherent carbon footprint. However, those spillover effects are not analyzed in this study.

The carbon footprint savings associated with the energy consumption in Figure 8 alone are staggering. Ampere-based compute can save over 60 MWh of power draw per rack per year.

According to the US Energy Information Administration (see endnotes), the benefits of deploying the RL300 in a rack of compute can reduce annual CO2 emissions by up to 22 to 25 metric tons. That is roughly equivalent to:



Competitive for Forthcoming Generations

The x86 server processor ecosystem continues to evolve with Intel's Sierra Forest and AMD's Turin processors slated as the next notable introductions. Both devices deliver increased core count and remain on DDR5 memory. At time of writing, Ampere was unable to acquire pre-release samples of these platforms to perform competitive benchmarking. Early estimates indicate performance and density gains of these new x86 CPUs come at the cost of increased power of 350W for Intel Sierra Forest and 400W for AMD Turin. Prohibitive modifications to cooling infrastructure will likely be required. This will be especially true for platforms which cannot be air cooled due to the extreme CPU power draw. This impact on rack and datacenter infrastructure poses a significant adoption challenge for service providers and digital enterprises looking to modernize.

> **Ampere Altra Max will remain competitive and likely continue to exceed competition from Intel Sierra Forest and AMD Turin in terms of efficiency. The HPE ProLiant RL300 Gen11 is the optimum compute choice for space and power constrained environments looking to refresh servers and avoid prohibitive modifications to existing infrastructure.**

Conclusion

Cloud Native Processors are designed for efficient scale-out workloads. The HPE ProLiant RL300 delivers this new class of compute, allowing operators to decrease power consumption and cost of ownership for various applications from edge to cloud. Ampere's groundbreaking processor architecture leads in performance efficiency and cost effectiveness even against the leading commercially available generations of x86 processors with the latest generation of system memory.

The HPE ProLiant RL300 Gen11 server can deliver more than 2.5X the performance as x86 in the same rack and power constraints and cut energy costs significantly. This provides an excellent opportunity for consolidation and reallocation of saved resources for service providers and digital enterprises globally.

Read More About HPE And Ampere
https://amperecomputing.com/products/partners/hewlett-packard-enterprise

Contact Us
https://amperecomputing.com/company/contact-sales

Developer Access Programs
https://amperecomputing.com/where-to-try

Ampere Efficiency Leadership and Footnotes
https://amperecomputing.com/home/increase-data-center-efficiency

https://amperecomputing.com/home/efficiency-footnotes

Endnotes

SUT Configurations

Base Platform	HPE ProLiant RL300	AMD Platform	Intel Platform
# Sockets	1	2	2
# CPUs populated	1 x Ampere Altra Max M128-30 Cores/Threads: 128 / 128 TDP: 250W	1 x AMD Genoa 9454 Cores/Threads: 48 / 96 TDP: 290W	2 x Intel Xeon Sapphire Rapids 6442Y Cores/Threads: 24 / 48 TDP: 225W
MEM	256GB (32GB x 8) DDR4 3200 MT/s	256GB (32GB x 8) DDR5 4800 MT/s	256GB (32GB x 8) DDR5 4800 MT/s
NIC	1 x ConnectX-6 MCX6231AS-CDAT 100GE 2P NIC	1 x ConnectX-6 MCX6231AS-CDAT 100GE 2P NIC	1 x ConnectX-6 MCX6231AS-CDAT 100GE 2P NIC
SSD	OS: 1 × 480GB NVMe Storage: 4 × 3.84TB NVMe	OS: 1 × 1.6TB NVMe Storage: 4 × 3.84TB NVMe	OS: 1 × 1.6TB NVMe Storage: 4 × 3.84TB NVMe
OS	Red Hat Enterprise Linux 9.2	Red Hat Enterprise Linux 9.2	Red Hat Enterprise Linux 9.2
Kernel	5.14.0-284.18.1.el9_2.aarch64	5.14.0-284.18.1.el9_2.x86_64	5.14.0-284.18.1.el9_2.x86_64
GCC version	12.2.1	12.2.1	12.2.1

Rack Level Assumptions

One full server rack consists of 38U useful rack space and 12.8KW of power for servers.

OpEx Assumptions and Notes:

- Includes only direct power (utility) cost
- Cost of $0.20 per kilowatt hour (KWh)
- 100% utilization of all servers at all times during useful life

Carbon Footprint Assumptions

To calculate the amount of CO2 generated by powering compute servers based on power consumption, the following source was used:

- U.S. Energy Information Administration (EIA)
  https://www.eia.gov/tools/faqs/faq.php?id=74&t=11

All calculations in this analysis assume Natural Gas as the source for electricity generation.

To further convert the mass (metric tons) of Carbon Dioxide to equivalencies, the following source was used:

- U.S. Environmental Protection Agency (EPA)
  https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results

Competitive for Forthcoming Generation of x86

At the time of writing, the early estimation is based on the published document by Intel and AMD.

- Intel
  https://www.intel.com/content/www/us/en/products/sku/232380/intel-xeon-gold-6442y-processor-60m-cache-2-60-ghz/specifications.html

- AMD
  https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9454.html

Disclaimer

All data and information contained in or disclosed by this document are for informational purposes only and are subject to change. Your results may differ. This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere®.

This document may contain technical inaccuracies, omissions and typographical errors, and Ampere Computing LLC, and its affiliates ("Ampere"), is under no obligation to update or otherwise correct this information. Ampere makes no representations or warranties of any kind, including express or implied guarantees of noninfringement, merchantability or fitness for a particular purpose, regarding the information contained in this document and assumes no liability of any kind. Ampere® is not responsible for any errors or omissions in this information or for the results obtained from the use of this information. All information in this presentation is provided "as is", with no guarantee of completeness, accuracy, or timeliness.

This document is not an offer or a binding commitment by Ampere®. Use of the products and services contemplated herein requires the subsequent negotiation and execution of a definitive agreement or is subject to Ampere's Terms and Conditions for the Sale of Goods.

The technical data contained herein may be subject to U.S. and international export, re-export, or transfer laws, including "deemed export" laws. Use of these materials contrary to U.S. and international law is strictly prohibited.