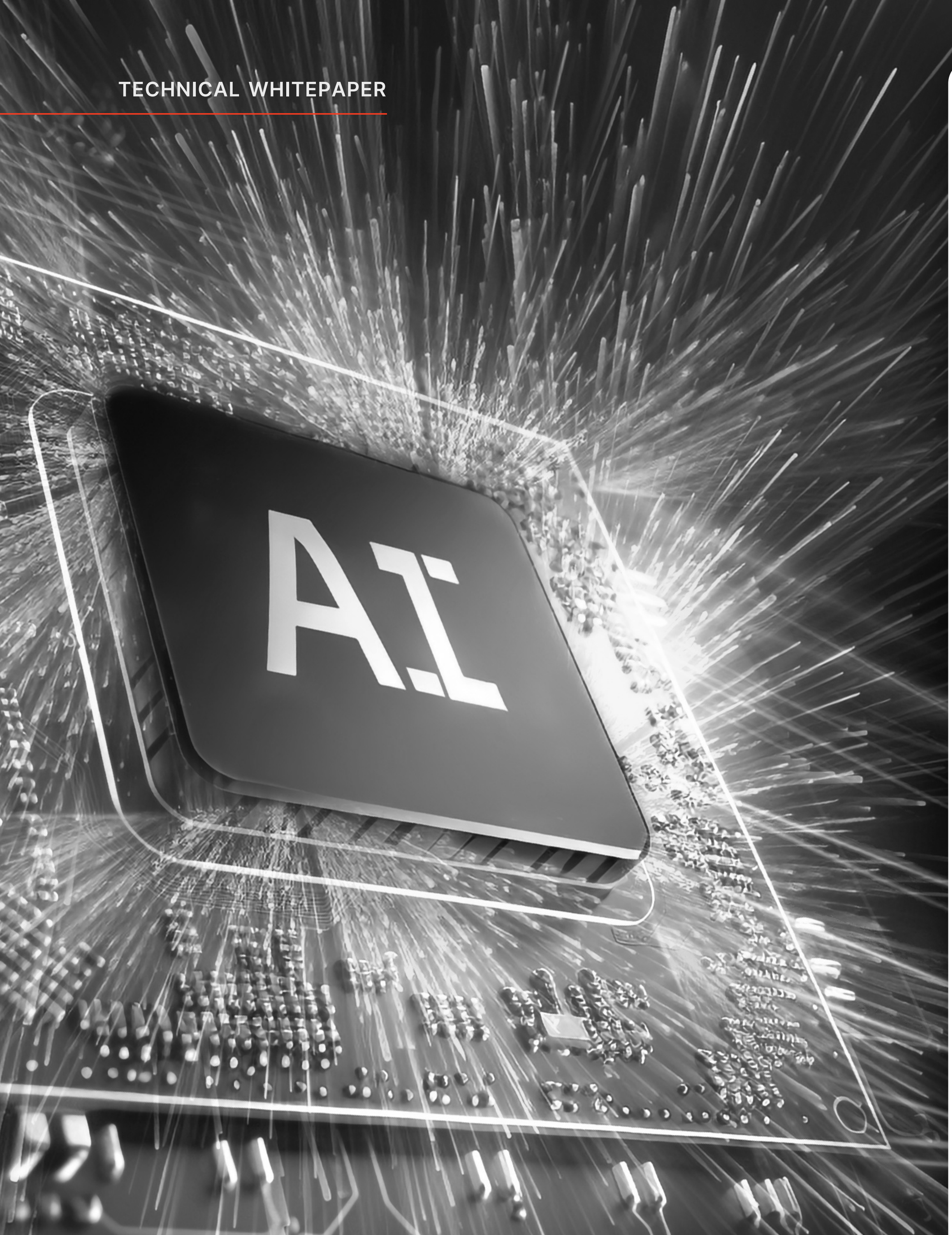
TECHNICAL WHITEPAPER



AMPERE®

A New Era in Data Center Efficiency

AmpereOne®

AMPEREOne®

Executive Summary

This paper delves into the transformative potential of Artificial Intelligence (AI) for enterprises and service organizations worldwide, highlighting the significant opportunities and challenges that come with its adoption. While AI adoption has become a major drain on the world's energy supply, there is a solution to combat this. Existing data centers worldwide can tap into energy-efficient Cloud Native Processors to offset the continuous, growing need for more sustainable compute. AmpereOne® is the newest introduction and leads the way in terms of energy efficiency at scale.

In this paper, you will learn about:

1) The opportunity to adopt AI at scale
2) The challenge of having resource constraints to enable AI adoption
3) The solution of refreshing aging infrastructure to free up space, budget and capacity
4) The options the market provides for infrastructure refreshes
5) The advantage AmpereOne provides data center operators in this journey
6) The roadmap Ampere® is investing in to deliver energy-efficient compute in the future

Woven throughout is an analysis of performance per rack of AmpereOne versus x86 alternatives, along with a deep dive into a real-world AI-enhanced web service model. Read on to learn about how Ampere delivers greater efficiency, better scale-out performance and lower cost of ownership for operators around the world.

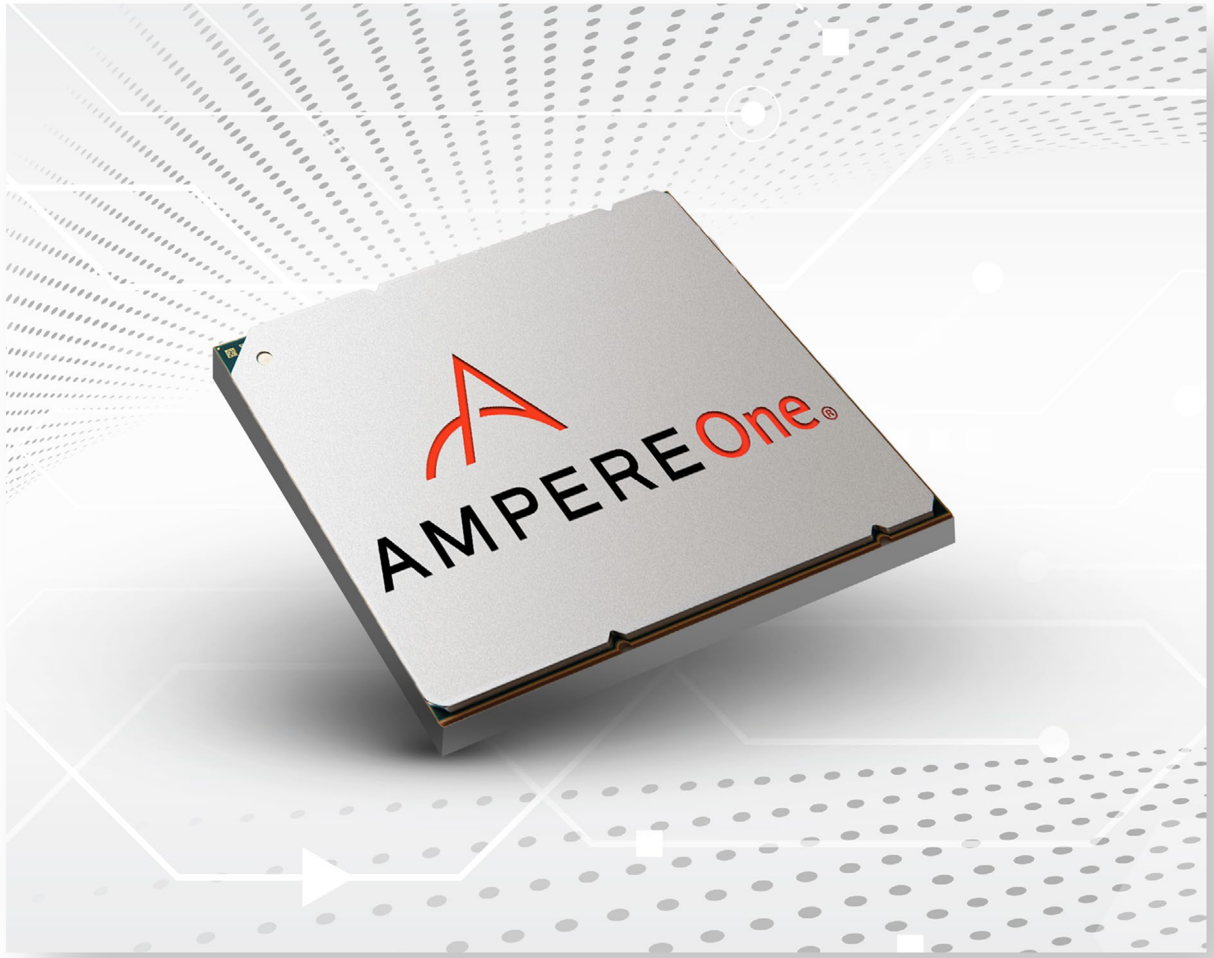


Table of Contents

The Opportunity

Enterprises and service organizations all over the world embrace the value-add of Artificial Intelligence to enhance their offerings, improve user experiences and/or to increase efficiencies. For years, media has been littered with articles on the AI boom and the "where" and "how" it is proliferating our daily lives. There are the obvious ways in which we've all felt this ramp, for example by using AI assistants on our handheld devices or in modern digital commerce platforms that recommend new products according to our taste. Then, there are the much less visible ways in which AI has proliferated our lives, such as the behind-the-scenes activity that generates pricing for airline fares or the near-instant fraud detection algorithms that are in use every time we swipe our credit cards.

> **Generative AI is on track to exceed $1 trillion in industry size by 2032.**

Generative AI, specifically, may grow at double digit rates for the next decade or so, and could reach well over $1T in industry size by 2032.[1] To date, some data suggests that the share of SMB organizations adopting AI — around 30% — is only half that of large enterprise companies.[2] The resource requirements to build and maintain the levels of expertise to run AI at scale have traditionally been high, though the gap is narrowing.

Many readers will have heard about established juggernauts like OpenAI, IBM Watson or newer players such as Hugging Face or Copy.ai that share the goal of making generative AI more accessible. These companies allow small and medium businesses lacking the resources to employ dedicated data science and AI teams to adopt AI at scale to enhance profitability, reduce costs and/or to manage risk. Through pre-trained models, easy-to-use APIs, visual interfaces and other features, these organizations lower the barriers to entry so SMBs can harness the power of generative AI efficiently and cost effectively. This democratizes AI for players of all sizes without losing their competitive edge to large competitors with the staff, the funds and a few years head start in their AI journey.

The Challenge

In part, this rise and democratization of AI has driven the adoption of increasingly power-hungry processors. The GPU has become the quasi currency when it comes to building IT infrastructure for ‘all things AI’. Whether it be Training or Inference workloads, many IT professionals have relied on GPUs, mostly from NVIDIA, to handle their needs. Both their GPUs, as well as those from other makers such as AMD, continue to show significant power consumption increases from one generation to the next.[3] Some hardware now exceeds 700W in TDP per unit.[3] This behavior has significant effects at global scale: total power consumption from data centers may double over the 2022-2026 timespan to over 1,000 terawatt hours — roughly the equivalent to what the country of Japan consumes each year.[4] This creates a major capacity crisis and cause for concern for utility providers and local governments.[5]

> **The average data center rack only has roughly 10kW in available power budget.**

Looking at the capacity challenge from a more practical angle, roughly 4 out of 5 data center operators are power constrained, with no more than 20kW available per rack — the average operator is facing limits as low as 10kW per rack, and operating with Power Usage Effectiveness (PUE) ratios around 1.5.[6,7] For many generations of processor upgrades, we’ve tolerated increasing power envelopes as long as the performance per watt (i.e. the efficiency) of the new processor generations increased. After all, it meant that new servers ‘got better mileage’ than old servers, so investments could be justified. However, focusing on this metric is flawed in that it implicitly assumes unlimited power budget to be available. By now though, many of our customers reached the point where they lament half empty rows of racks because they’ve hit their power ceiling far before they can cater to their business demands for innovation or growth.

Upgrading or expanding rack space is often prohibited by costs, space or regulation. Most existing data centers and rack infrastructure have been designed for air cooling and cannot simply adopt liquid or immersion cooling for large parts of their server fleet.[8] Lastly, many data center racks carry servers more than five years old as refresh cycles have been prolonging across the industry. For operators, that drives up maintenance costs over the years and the aging servers generally become less productive and more vulnerable to security exploits.[9]

The Solution

Life cycle refreshes might be considered mundane IT activity by some, though timely and intentional upgrades can fuel a company's ability to innovate. In fact, we believe the most practical way to solve the capacity challenge that's in the way of AI adoption is by **upgrading existing infrastructure**. Holding off on much-needed upgrades drives up operating costs as infrastructure ages,[10] and it prevents you from reclaiming the valuable space and power budget that could be repurposed for more modern server deployments running AI compute workloads. Similarly, given the typical PUE ratings of most data centers, reducing the power consumption of your server fleet has a multiplying effect of about 1.5X on the total power draw (and spend) of your data center.[6]

The industry is reaching an inflection point and performance per rack is emerging as the primary design criterion for infrastructure upgrades. When upgrading and consolidating infrastructure, using this metric lets operators focus on packing the most compute power into existing power and data center footprint.

By being deliberate and smart about refreshing server infrastructure, companies can bring focus back to building their business, not their data center.

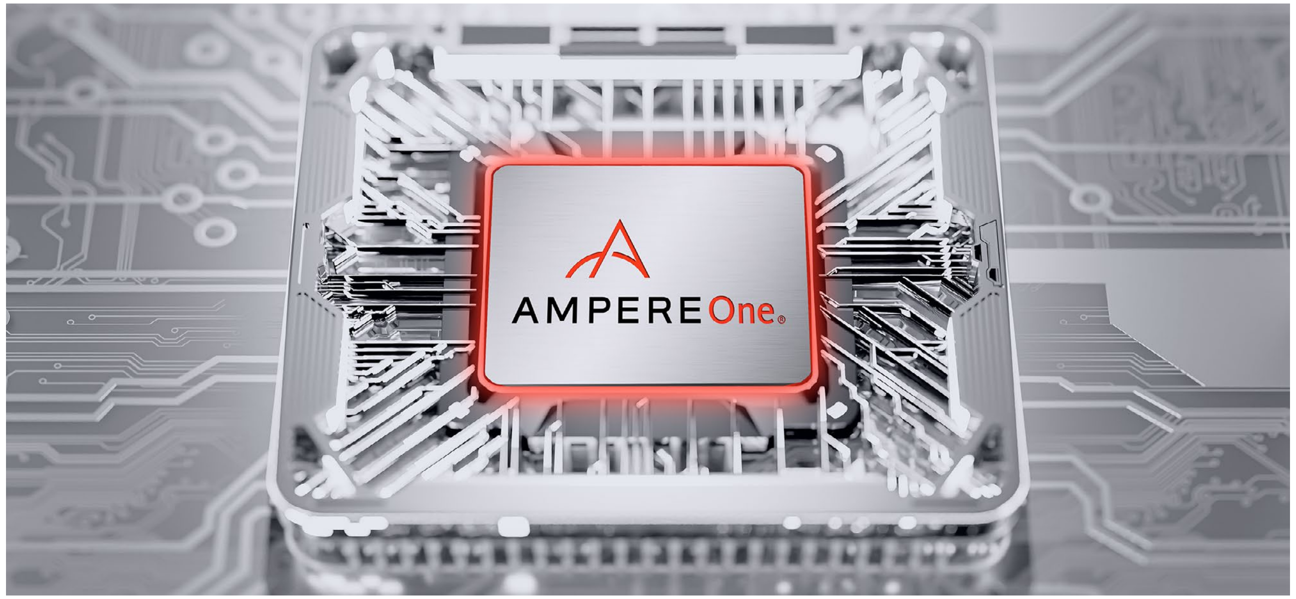
The Options

For server upgrades, operators primarily have the options to upgrade servers with Intel Xeon Scalable Processors or AMD EPYC. These two vendors have been the clear choices for some time.

Intel's most recent product families interesting for refresh are 5th Gen Xeon Scalable Processors (codename "Emerald Rapids") and even 6th Gen Xeon Scalable Processors. The latter offers choices between performance-optimized (P-core) "Granite Rapids" or energy-optimized (E-core) "Sierra Forest" products. AMD, on the other hand, is shipping 4th Gen EPYC 9004 series processors "Genoa" and "Bergamo" into the server segment, featuring up to 128 multi-threaded cores per CPU. Throughout this document, the main body of this writing will generally refer to the processor family codenames, whereas the various figures will reference the top-bin SKU of each of those processor family: Intel Xeon 8592+ for Emerald Rapids, Intel Xeon 6780E for Sierra Forest, AMD EPYC 9654 for Genoa, and lastly AMD EPYC 9754 for Bergamo.

NVIDIA has also established itself as a major player to compete with server platforms — the HGX and DGX series. It builds upon its long history of designing GPUs that have since become the de facto standard to drive the age of AI adoption. These platforms are primarily aimed at AI Training, Deep Learning and HPC workloads, and are functionally much less suitable for those businesses trying to shift to the era of AI Compute, which primarily benefits from AI Inference. NVIDIA's current flagship platforms are HGX H200 and DGX H200. The DGX H200 fully configured platform alone exceeds 10kW in power![10] The HGX H200 platform is offered through multiple OEMs so total platform power consumption may vary. However, the highest density configs feature 8 x GPUs drawing 700W each on top of 2 x host CPUs with 350W each,[11] without even counting other peripherals. As we stated, the average operator has only 10kW to allocate per rack today. So, the challenge deploying NVIDIA at scale for effective consolidation of space, power and budget is obvious.

When weighing options, it is critical to evaluate performance efficiency, and to do so at the performance per rack level. Rack infrastructure and power budgets are finite, and getting the most performance out of each rack is paramount. That is why Ampere offers Cloud Native Processors and has launched AmpereOne as its most efficient processor yet.



The AmpereOne Advantage

Cloud Native Processors from Ampere are designed to deliver scale-out performance and leading efficiency. This helps customers get the best consolidation ratios when upgrading legacy infrastructure by delivering leading performance per rack. The architecture choices behind AmpereOne® and its custom compute core build upon the continued success of the Ampere® Altra® family of processors.

The following section shows how efficiency leadership of a single processor is the foundation to delivering scale-out data center savings to the effect of:

- **Up to 38% less rack space**
- **Up to 37% less power**
- **Up to 49% less server acquisition costs**

... all for the same performance levels.

What does this mean in terms of total cost of ownership (TCO)?

> **Over a period as short as 3 years, AmpereOne can help reduce the total cost of ownership (TCO) by as much as 41% compared to AMD Genoa and 33% compared to AMD Bergamo.**

i. Leading System-on-Chip (SOC) Efficiency

We start our analysis with a review of the popular synthetic benchmark “SPEC CPU®2017 Integer Rate Result” from SPEC.org (Standard Performance Evaluation Corporation). This will serve to inform us of rough consolidation ratios when upgrading from legacy x86 processor generations to AmpereOne.

> **At the time of writing, AmpereOne A192-32X delivers the highest single CPU published SPECrate®2017_int_base score generated via an open-source compiler at 702.**

The result of Ampere’s innovative custom core architecture is particularly interesting when looking at the energy consumption profile during the SPEC CPU®2017 Integer Rate run. Figure 1 below shows that — despite a platform max 400W TDP rating of the A192-32X CPU — the average power consumption (red area in chart) during the test is significantly lower at 284W. That is 29% lower power consumption than max TDP of the part. This equates to a performance per watt of 2.47. All the while, the CPU frequency (black line) remains constant around the rated 3.2 GHz during the entirety of the run (exception: temporary drops as the test suite cycles through its different benchmarks). This characteristic is key to delivering AmpereOne’s energy efficiency advantage.

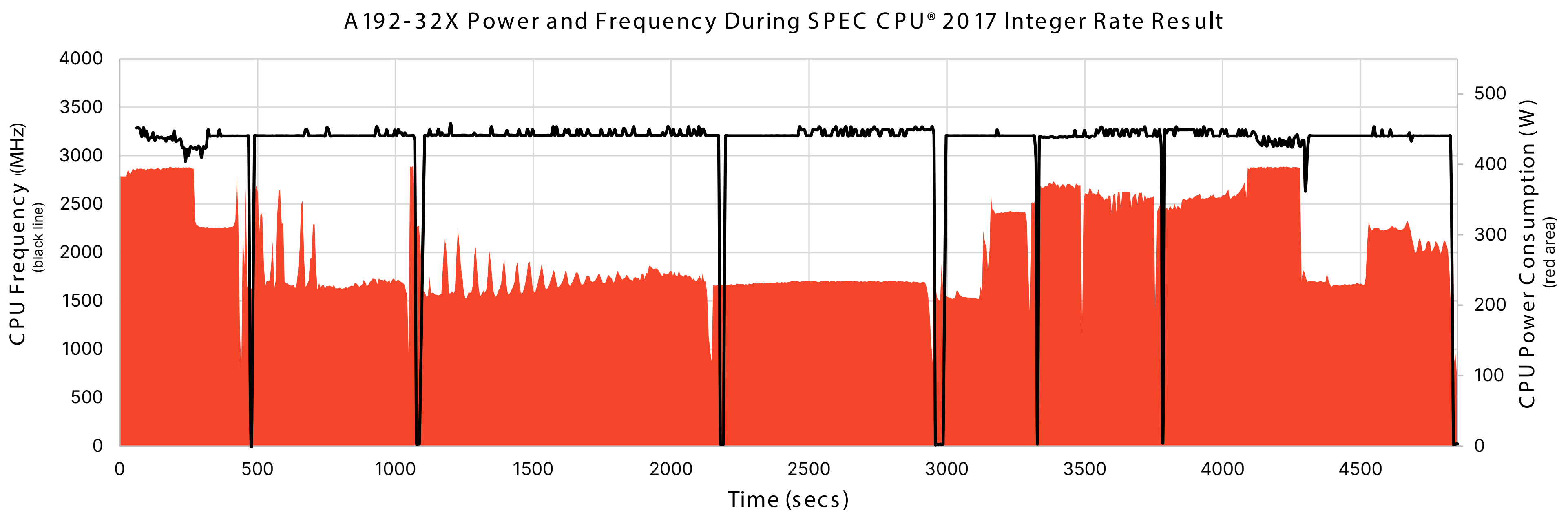


Figure 1: Power and Frequency of an AmpereOne A192-32X Processor Plotted During a SPEC CPU®2017 Integer Rate Result Run

For illustration purposes, Figure 2 shows the implications when compared against leading AMD EPYC Genoa and Bergamo processors as well as against Intel Xeon Emerald Rapids and Sierra Forest. The solid bars show the SPECrate®2017_int_base score, whereas the striped bars refer to usage power (in watts). Above the bars we display the "performance per watt" to show the efficiency advantage that AmpereOne provides. It leads AMD Bergamo by more than 12% and outshines Intel Xeon Emerald Rapids by as much as 89% on efficiency.

Due to the capacity challenge described earlier, it is critical to evolve beyond just a performance per watt analysis at the socket-level. Instead, we use the performance and power draw of the CPU, pair it with system level power draw (for peripherals such as memory, storage, motherboard, fans, etc.) and then extrapolate a rack-level performance figure. The higher power draw of AMD and Intel processors as compared to AmpereOne contributes to fewer servers fitting into a rack's power budget. We then multiply the server performance with the server count and derive a performance per rack figure. Figure 3 shows the relative performance of one rack of AmpereOne A192-32X versus leading x86 competition.

Compared to a rack of AmpereOne-based servers, AMD EPYC falls behind on performance by as much as 22%. More glaringly, Intel Xeon Scalable Processors of the 5th and 6th generation deliver up to 44% less performance than a rack of Ampere's flagship AmpereOne A192-32X processor.

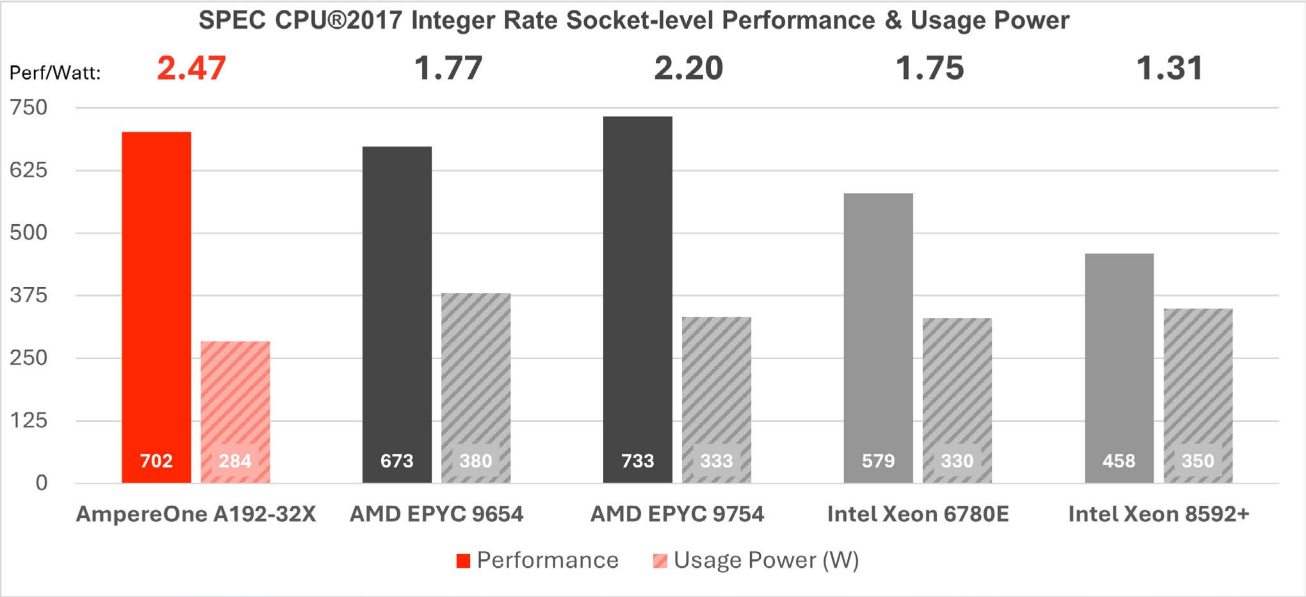


Figure 2: SPECrate®2017_int_base score (estimated) and usage power of a SPEC CPU®2017 Integer Rate Result run using GCC13 compiler

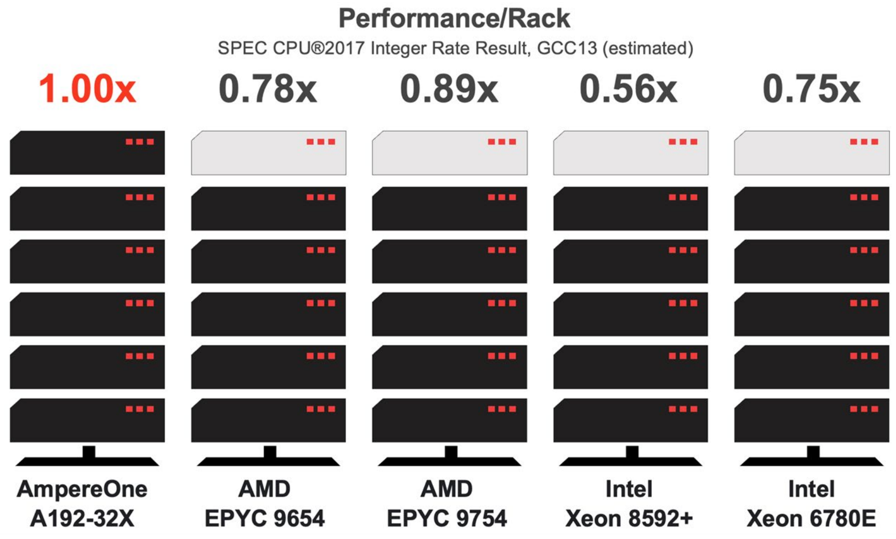


Figure 3: Rack-level performance based on SPECrate®2017_int_base score (estimated) and related usage power.

ii. Consolidating Legacy Infrastructure

An average server experiences rising costs for annual IT support and unplanned downtime productivity costs from year to year during its life.[9] That said, even once fully depreciated, an aging server fleet carries rising upkeep costs all the while underperforming compared to the latest generally available server processors.

Given the average refresh cycle for servers now exceeding 5 years, let us assume an operator is aiming to upgrade their existing 1st or 2nd generations of either Intel Xeon Scalable Processors ("Skylake" or "Cascade Lake") or AMD EPYC ("Naples" or "Rome"). Those processor generations first entered the market between mid 2017 and mid 2019.

Given the outstanding rack-level performance of AmpereOne, somewhere between 3.6 and 5.4 racks of 1st generation Intel Xeon Scalable Processors (Skylake) or AMD EPYC (Naples) processors could be consolidated into a single rack. Similarly, between 2.1 and 3.6 racks of the 2nd generation Intel (Cascade Lake) and AMD (Rome) could be replaced with a single rack of AmpereOne based servers. These consolidation estimations are based on published SPEC CPU®2017 Integer Rate Results across the SKU stack and assume dual socket systems (see end notes for details). Actual results are dependent on specific deployment configurations, production workloads being run, and other factors.

Let us assume AmpereOne's consolidation ratio is 5:1 for going from 1st generation Intel/AMD and it is 3:1 when upgrading from 2nd generation Intel/AMD. What used to take multiple racks of equipment, space and power now can be done in 1 rack.

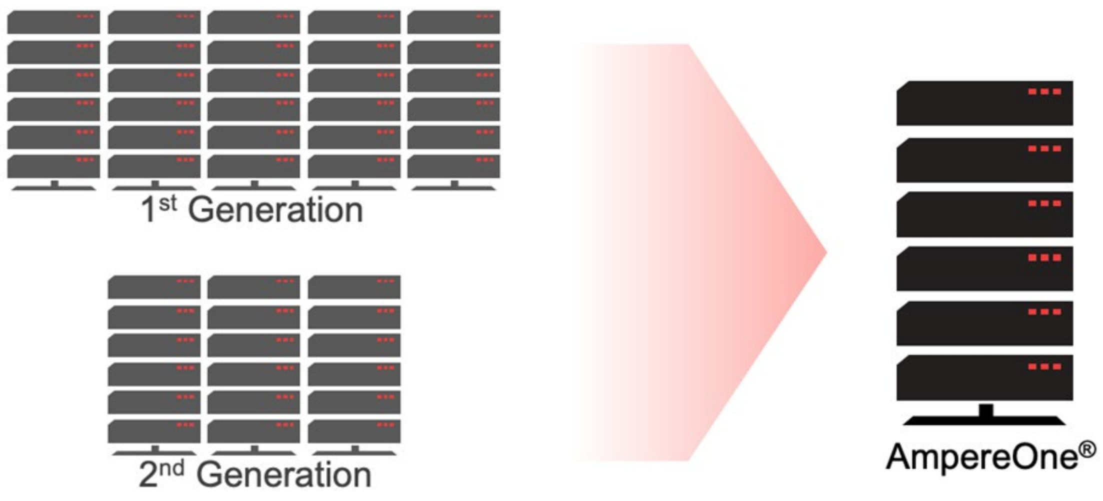


Figure 4: Rack consolidation illustration for migration from 1st or 2nd generation Intel Xeon Scalable Processors or AMD EPYC.

For example, let us look at a modest data center deployment with 20 racks of Intel Skylake servers, the value of which has been completely depreciated. Consolidating this setup at a 5:1 ratio would shrink the footprint to 4 racks. These 4 racks of AmpereOne would be delivering the same performance as the initial 20 rack Intel setup. This comes at an upfront investment (CapEx). However, the tradeoff for operators are the monthly (OpEx) savings by eliminating the need for 16 out of 20 racks. This drastically cuts space, power, server admins, and PUE overhead just to name a few obvious cost factors. Table 1 is a simplified view in that it ignores the increasing monthly costs for aging servers (as described above).

Options	CapEx (one-time)	OpEx (monthly)	TCO (3 years)
Staying on Intel Skylake (20 racks)	$0	$134,471	**$4,840,965**
Upgrade to AmpereOne (4 racks)	**$928,620**	**$28,094**	**$1,940,013**

Table 1: CapEx and OpEx implications of either staying on legacy infrastructure or of upgrading to AmpereOne.

Figure 5 illustrates how the CapEx (~$930k, all accounted for in Q1 Y1) to populate 4 racks with AmpereOne is recouped in as little as 9 months due the staggering 79% reduction in OpEx. After as little as 3 years, total TCO savings could amount to $2.9M (approximately 60%) with AmpereOne. See the end notes for further details.

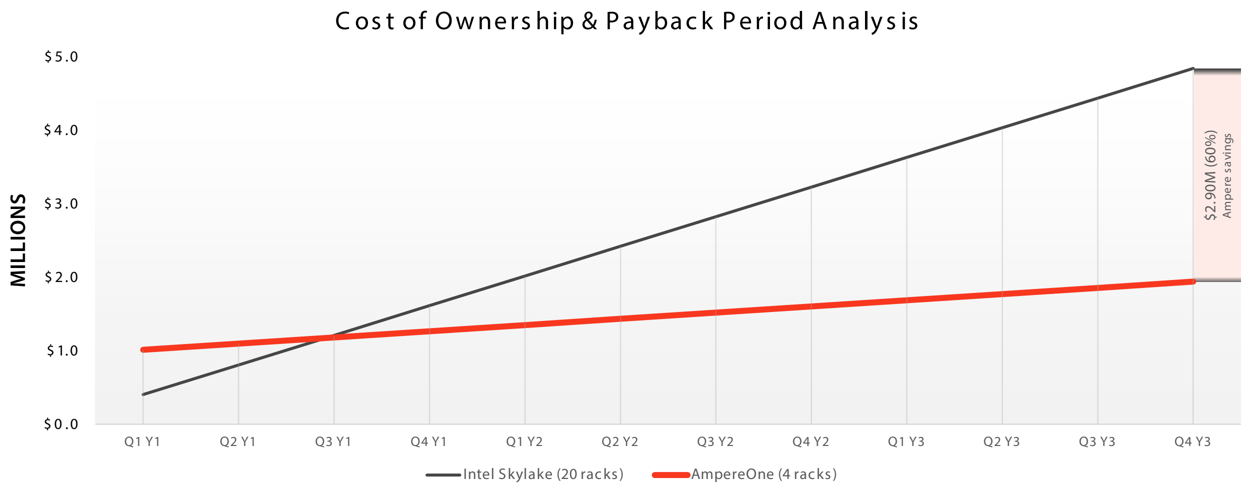


Figure 5: Payback period for capital investment & cost of ownership to consolidate to AmpereOne-based servers.

AmpereOne is an excellent choice to refresh legacy servers. Its innovative design delivers high density and energy efficiency that allows operators to consolidate legacy architecture and thus create real space, power and budget savings. TCO can be cut in half and savings then can be reallocated to deploy AI Compute infrastructure to keep up with the rapid adoption of AI across virtually every sector.

iii. Real-World Workloads

While SPECrate®2017_int_base rigorously measures a CPU's performance for a variety of compute-intensive, integer-based workloads on a single server, it misses the larger context of running real world services at scale. Hence, we focus on various popular cloud native workloads behind many modern enterprises and customer services, from web services and video services to storage services and beyond. Here, too, we will assess the impacts that AmpereOne's CPU-level performance and efficiency has when deployed at scale (rack-level and beyond).

Based on Ampere lab test results, Figure 6 on the next page indicates how AmpereOne A192-32X with 192 cores at 3.2 GHz delivers leading performance and efficiency against AMD's 4th generation EPYC 9004 Series processors for several popular cloud native workloads. The Appendix contains information on the test methodology. AmpereOne's raw performance advantage against AMD Genoa reaches up to 28%, whereas the compute efficiency advantage with AmpereOne is as large as 86%.

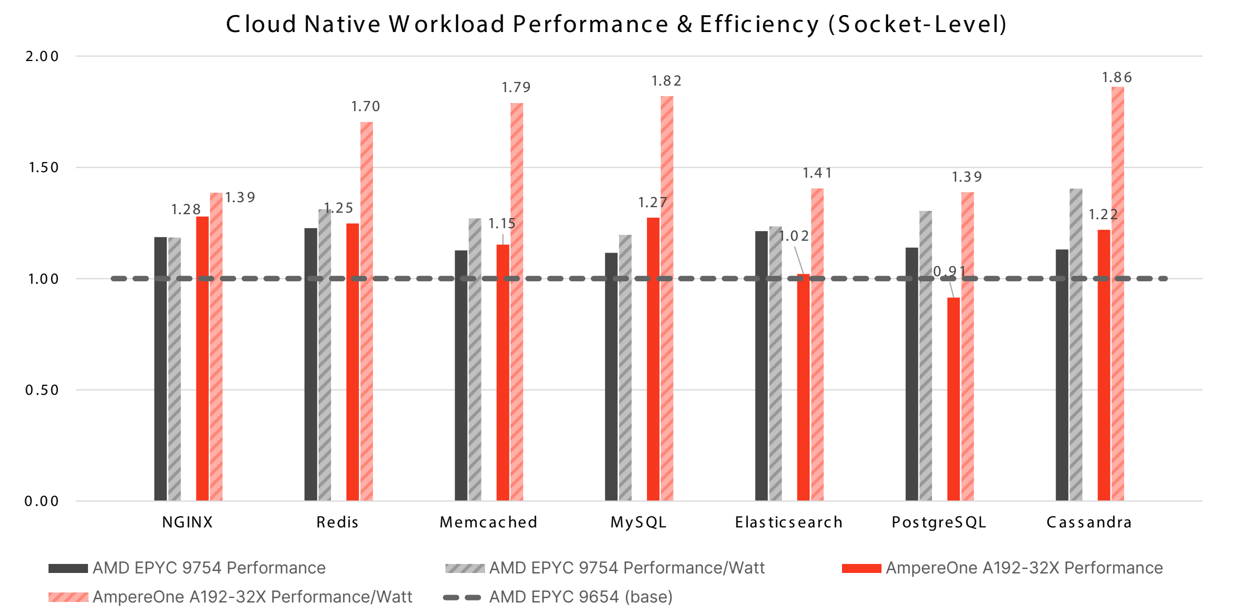


Figure 6: Raw CPU performance and performance per watt measurements. Displayed for various popular cloud native workloads.

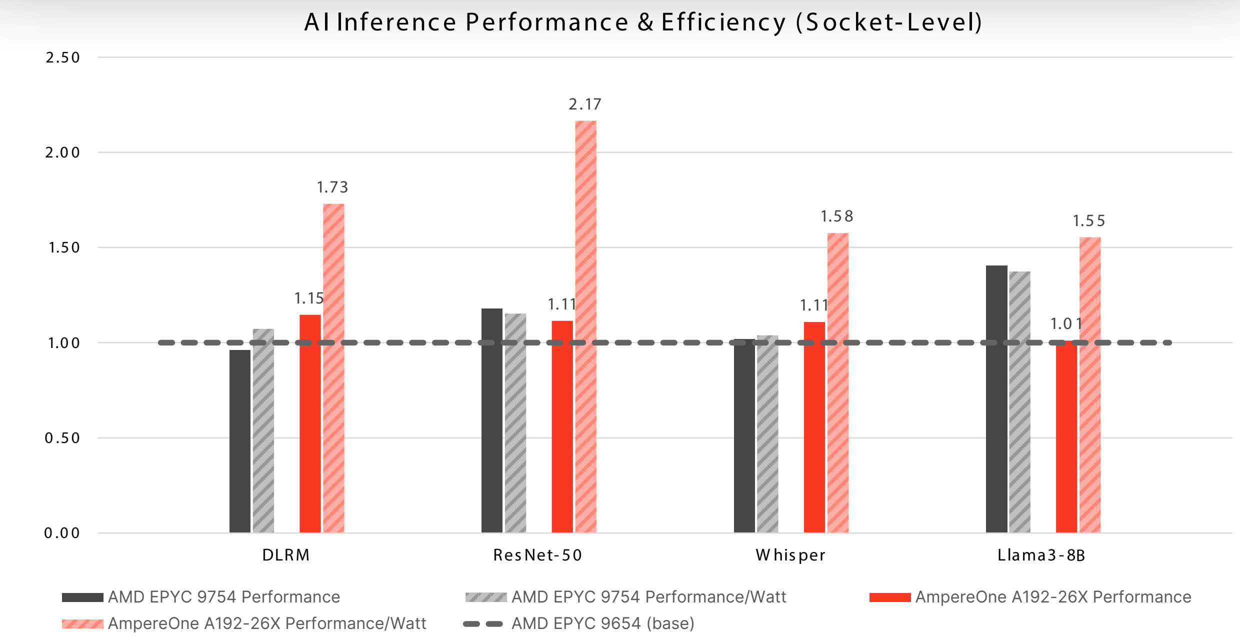


Figure 7: Raw CPU performance and performance per watt measurements. Displayed for various popular AI inference workloads.

In the realm of AI Inference, the AmpereOne A192-26X with 192 cores at 2.6 GHz is the ideal combination of performance and efficiency for various popular workloads spanning recommender engines, object classification, speech-to-text conversion and generative AI. Compared to the top bin A192-32X, this part runs 600MHz lower, thus reducing SOC power consumption. Many AI inference workloads are not as clock-rate bound as they are core and memory-bound. Thus, still having 192 compute cores (and thus same number of SIMD vector units) and system memory improves efficiency without sacrificing much performance. Like the analysis around the cloud native workloads, we see that the AmpereOne efficiency advantage over AMD EPYC 9654 is more than 50% in all cases, and as high as 2.17X for ResNet-50 object classification, as shown in Figure 7 below. The Appendix contains information on the test methodology.

AmpereOne clearly delivers high performance and excellent CPU-level efficiency for relevant, modern workloads. Next, we will incorporate these findings into server- and rack-level analysis.

iv. Leading Rack-Level Efficiency for Real-World Workloads

As alluded to earlier, rack-level performance has emerged as the key factor that can drive up data center efficiency. The metric combines performance, power consumption, rack density, and overall data center footprint into a single measure that can be scaled linearly for compute installations of all sizes.

Based on processor-level data (Figures 6 and 7) and platform power draw assumptions, we estimate the throughput generated by an entire rack of servers.

Said rack is populated with as many servers as possible until the available power budget is maxed out, and server throughput is scaled linearly to the total rack-level to approximate real-world behavior. The results in Figures 8 and 9 show AmpereOne's excellent performance per rack advantages. It exceeds AMD EPYC 9004 Series processors across cloud native and AI inference applications as much as 67% in performance per rack.

v. AI Compute – Build your Business with AI

The AI hype cycle has led to continual transformation with both general purpose and AI workloads converging to drive product innovation and appeal in virtually any industry. We call this AI Compute. The AI market is overly focused on GPUs, which were originally built for non-AI tasks and are incorrectly balanced for most AI tasks. While GPUs excel in training, 85% of AI silicon is used for inference,[12] which has different compute requirements. We believe those inference workloads should be managed without using GPUs — dependent on the use case — either by deploying Ampere CPUs, or by deploying domain-specific accelerators.

As described above, making smart refresh choices can yield significant space, power and cost savings for operators. Hence, we want to assess a composite AI Compute application stack such as a modern web service and then examine how the freed-up space and power budgets could be repurposed.

Modern digital services commonly contain 3-4 service layers and are enhanced by AI-enabled features such as recommender engines and chatbots. In our analysis, each traditional web service layer is weighted at 20% with both AI

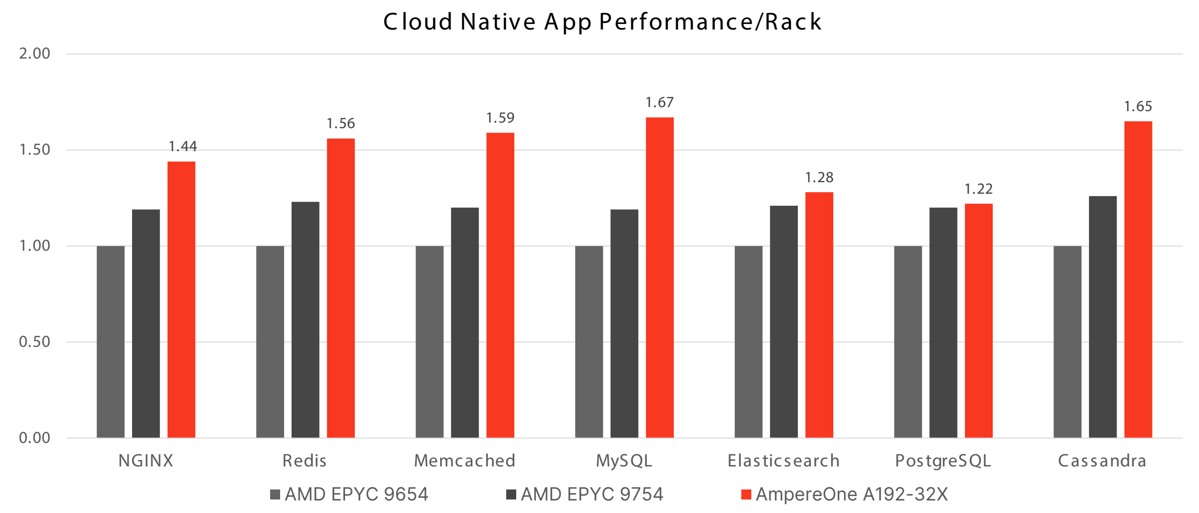


Figure 8: Rack-level performance projections. Displayed for various popular cloud native workloads.

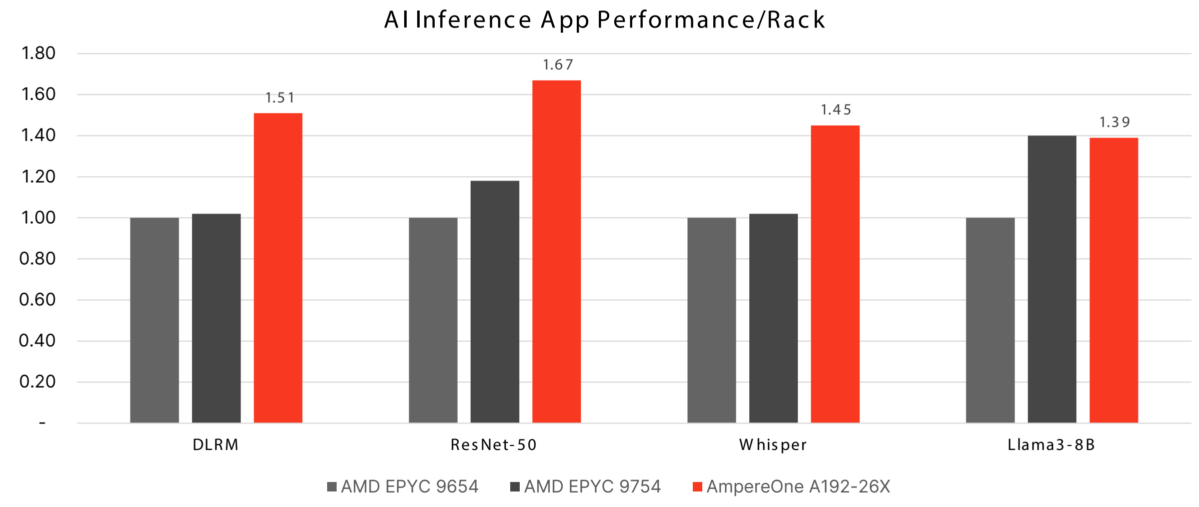


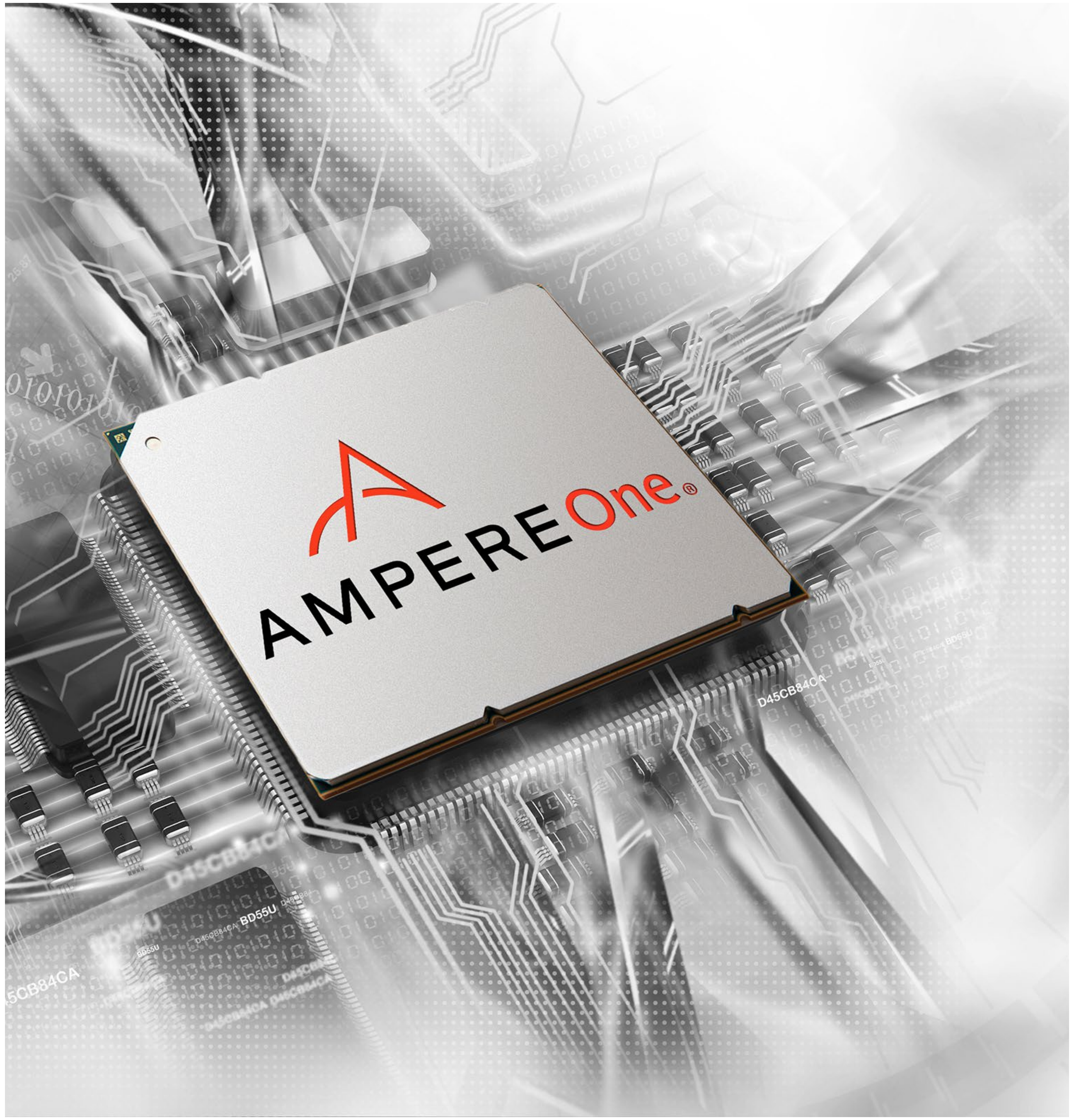
Figure 9: Rack-level performance projections. Displayed for various popular AI inference workloads.

applications weighted at 10% each. Configurations of real customer applications will vary dependent on the individual service implementation and business needs. Though, our below example serves as a proxy for modern web products such as e-commerce, content libraries, travel booking, online banking or social networks.

Service Layer	Application	Weight	Rackspace
Front End / Web Serving	NGINX	20%	1 Rack (40U)
In-Memory Caching	Memcached	20%	1 Rack (40U)
Key Value Store	Redis	20%	1 Rack (40U)
Relational Database	MySQL	20%	1 Rack (40U)
Recommender Engine	DLRM	10%	0.5 Racks (20U)
Chatbot	Llama3-8B	10%	0.5 Racks (20U)

Table 2: Relative application weight of proxy modern AI compute web service stack.

In this example, the infrastructure consists of 5 fully utilized server racks populated with AmpereOne. We take the total output generated by these five racks and contrast it with the server count, rack count, power draw and CapEx that would be required to match the performance if using 4th gen AMD EPYC processors instead.

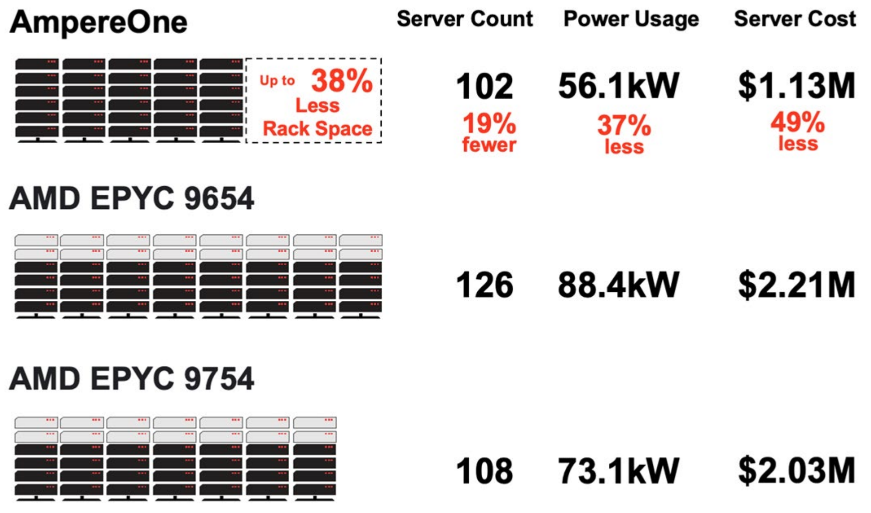


As illustrated in Figure 10, the individual rack-level performance and efficiency advantages of AmpereOne over leading AMD EPYC processors are magnified when analyzed in such way.

AmpereOne clearly shines through its cost effectiveness and performance efficiency. To run a modern AI-enhanced web service powered by 5 racks of AmpereOne, leading AMD EPYC processors would require upwards of $1M (96% more) in additional upfront investment, occupy up to 3 additional (60% more) racks, and draw about 32kW (58% more) extra power.

Clearly, repurposing budget, space and power reclaimed by refreshing aging legacy servers is best done with Ampere. Figure 11 shows the growing impact of making the smart refresh choice by deploying AmpereOne. In as little as 3 years, running the example web stack mentioned above would save nearly $1.6M in total cost of ownership (TCO), equivalent to 41% cost savings as compared to AMD Genoa. Even compared to AMD's Bergamo top bin processor, an AmpereOne deployment saves $1.1M, or as much as 33%. See end notes for additional details.

> **When building new AI compute infrastructure, AmpereOne can help operators save up to 41% in TCO over three years compared to AMD EPYC 9004 series processors.**



	Server Count	Power Usage	Server Cost
AmpereOne Up to 38% Less Rack Space	102 19% fewer	56.1kW 37% less	$1.13M 49% less
AMD EPYC 9654	126	88.4kW	$2.21M
AMD EPYC 9754	108	73.1kW	$2.03M

Figure 10: Scale-out projections for proxy AI Compute web service including rack and server count along with power consumption and cost estimates.

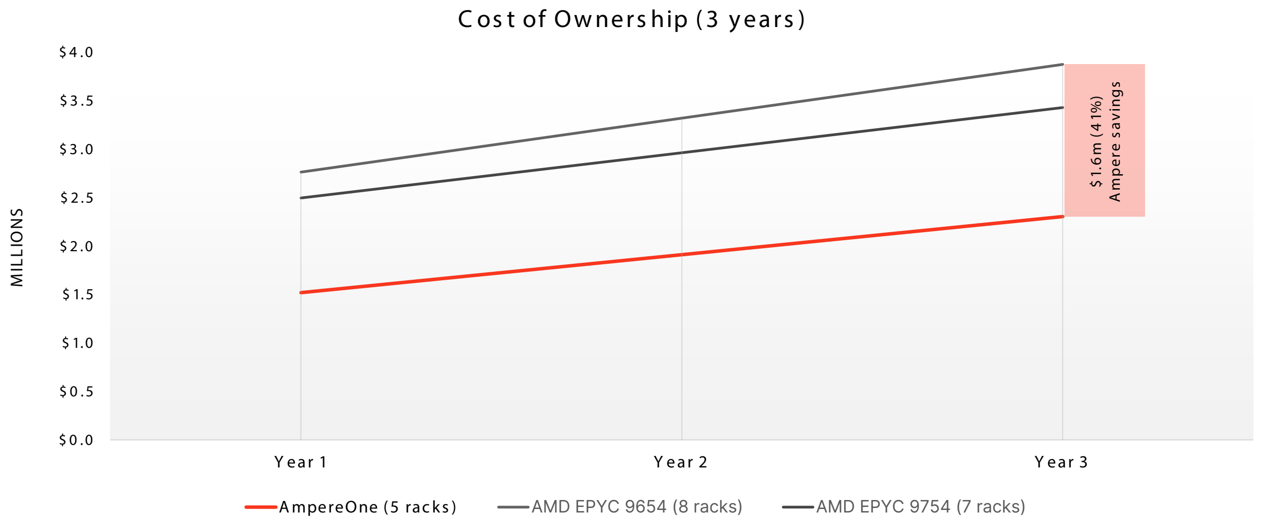


Figure 11: Three-year TCO model for equivalent performance deployments based on AmpereOne, AMD Genoa and AMD Bergamo.

The Ampere Roadmap

Ampere's roadmap beyond the 192 core AmpereOne processor family continues to focus on high-efficiency server processors. Our roadmap is meticulously designed to address the evolving demands of data centers and cloud computing environments, especially with the ramp of AI inference applications in virtually every domain. Our Cloud Native Processors offer a compelling combination of performance, scalability, and power efficiency, making them ideal for modern workloads.

Starting with AmpereOne, we have introduced disaggregation into our product design, strategically matching the processor's subsystems with the best manufacturing processes available. This strategy focuses on separating key components of server infrastructure such as compute and memory into distinct, scalable units (dies) to better serve cloud native environments. To connect the various dies together, we have developed a custom interconnect with up to 2.8TB/s aggregate bandwidth in each direction, and we've developed our own single, unified mesh. This modular approach aligns well with modern cloud workloads and allows dynamic resource scaling to provide flexibility for cloud native applications. For more information, visit our "Breaking Boundaries: AmpereOne's Disaggregation Strategy for the Next-Gen Cloud" blog.

In late 2024, we'll begin shipping the 12-channel DDR5 AmpereOne product with up to 192 cores. On top of that, we will release our 256 core AmpereOne "MX" and our 512 core AmpereOne Aurora products in the future, which continue to push the envelope in terms of core density and energy efficiency. With the release of AmpereOne Aurora, Ampere will bring its own integrated AI silicon to market targeting air-cooled server platforms. This focus allows us to further democratize AI adoption by delivering solutions that virtually any data center operator can rack and stack in their existing space. Tying back to our custom die-to-die interconnect, we are rapidly advancing toward SoC-level integration using UCIe. This allows us to rapidly integrate customer IP and customize I/O to different applications and use cases.

Our roadmap at Ampere is the culmination of years of engineering work and innovation to create the critical building blocks and integration technologies for an AI Compute world. As AI and general purpose workloads are rapidly converging in the cloud, we focus on providing the right mix of flexibility and efficiency in our SoCs. Data center operators around the world can confidently invest in Ampere as an accelerant to their digitization and AI adoption strategies.

The Bottom Line

The rapid rise of AI has driven the adoption of power-intensive processors, creating urgent capacity challenges for data centers. Alarmingly, 77% of operators have just 20kW of power per rack, with upgrades often constrained by costs, space, or regulations.[7] Global energy needs are skyrocketing, and the global climate is in disarray. The time to act is now. Cloud Native Processor technology can replace outdated, inefficient servers, freeing up critical rack space and resources to adopt AI, all the while preserving our planet's finite resources.

> **AmpereOne delivers up to 67% more performance per rack compared to AMD EPYC.**

Historically, infrastructure operators prioritized per-core performance when planning for expansion. More recently, performance per watt (i.e. compute efficiency) has also risen in importance. Though, increasing power consumption of recent generations of x86 processors is pushing most operators to their capacity limits and looking to maximize performance per rack has become paramount. As demonstrated above, high-performance, low-power processors like Ampere's offer up to 67% more performance per rack for specific Cloud Native workloads compared to 4th generation AMD EPYC alternatives. More so, operators maximize their return on investment with AmpereOne as it can help reduce server TCO by over 41% over a 3-year span, even for modest size deployments. This empowers businesses to fully utilize existing space, drive AI adoption, and fuel the next wave of innovation without the need for costly data center expansion or relocation.

> **AmpereOne can help reduce server TCO by over 41% over 3 years.**

Our world stands at the edge of an exciting transformation, where today's challenges are simply stepping stones to greater opportunities fueled by Artificial Intelligence. At Ampere, we empower those ready to push the boundaries, and we offer them the most energy-efficient, high-performance processors to upgrade legacy infrastructure.

The right choice today sets the stage for tomorrow's success. **Choose Cloud Native Processors from Ampere.**

The Next Steps

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

Where to Buy
https://amperecomputing.com/where-to-buy

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

Disclaimer

All data and information contained in or disclosed by this document are for informational purposes only and are subject to change. 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.

This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere.

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.

End Notes

1) Raw Performance Claims

Results shown throughout this white paper are estimates and actual results may vary. Product and company names are for informational purposes only and may be trademarks of their respective owners.

Performance per Rack: Rack is based on 42U rack with 12.5kW power budget. 2U and 1.0kW allocated as buffer for networking, management and PDU. Total performance per rack calculated by multiplying the performance per server with the maximum number of servers that fit in a rack (until space or power constraints are reached).

All server-level performance and socket-level power draw claims are based on Ampere Computing LLC internal lab testing. In order to calculate server usage power, platform power draw is added on top of CPU power draw. Platform power assumptions informed by three leading OEM server power calculator tools. Results shown here are estimates and actual results may vary. Product and company names are for informational purposes only and may be trademarks of their respective owners.

Platform Power Assumptions

Component	Description	Total Power Draw (W)
Storage	4 x NVMe (10W ea)	40
Networking	1 × 1GbE OCP NIC, 1 × 10/25GbE NIC, 1 × 100GbE NIC	40
Other	Motherboard, Fans, Misc	96
Memory (per socket)	6 ch DDR4, 1DPC	42
	8 ch DDR4, 1DPC	56
	8 ch DDR5, 1DPC	80
	12 ch DDR5, 1DPC	120

System Hardware & Software Configurations: Cloud Native Workloads & SPEC CPU®2017 Integer Rate:

AmpereOne

- HW: 1 x AmpereOne A192-32X (192c/192t, 3.2 GHz), 8 × 64 GiB DDR5 5200 MHz
- OS: Fedora 38
- Kernel: 6.8.9-100.fc38.aarch64
- Page size: 4K

AMD EPYC Genoa

- HW: 1 x AMD EPYC 9654 (96c/192t, 2.4/3.55 GHz), 12 × 64 GiB DDR5 4800 MHz
- OS: Ubuntu 22.04
- Kernel: 6.4.13-200.fc38.x86_64
- Page size: 4K
- SMT enabled

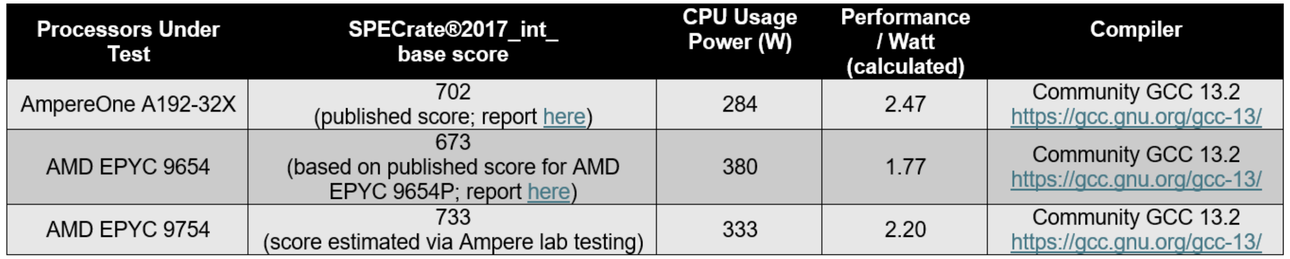
AMD EPYC Bergamo

- HW: 1 x AMD EPYC 9754 (128c/256t, 2.25/3.1 GHz), 12 × 64 GiB DDR5 4800 MHz
- OS: Ubuntu 22.04
- Kernel: 6.4.13-200.fc38.x86_64
- Page size: 4K
- SMT enabled

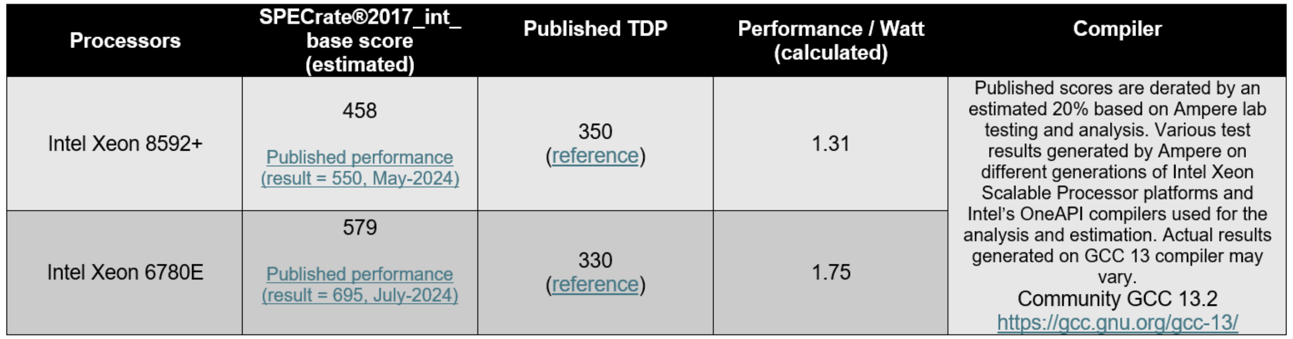
Socket-Level Power Draw & Performance: SPEC CPU®2017 Integer Rate

SPEC CPU®2017 Integer Rate Results: All SPECrate®2017_int_base performance estimates for AMD and Ampere platforms are based on GCC (version 13 compiler). See details below. Rack-level estimates based on 1U server height and 1 socket platforms. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

Server Usage Power: All CPU power draw figures are based on Ampere-performed lab tests under load (for each referenced application). To calculate server usage power, platform power draw is added on top of CPU power draw. Platform power assumptions informed by three leading OEM server power calculator tools. For Intel Xeon 8592+ and Intel Xeon 6780E, manufacturer-published CPU TDP was used instead.



Processors Under Test	SPECrate®2017_int_ base score	CPU Usage Power (W)	Performance / Watt (calculated)	Compiler
AmpereOne A192-32X	702 (published score; report here)	284	2.47	Community GCC 13.2 https://gcc.gnu.org/gcc-13/
AMD EPYC 9654	673 (based on published score for AMD EPYC 9654P; report here)	380	1.77	Community GCC 13.2 https://gcc.gnu.org/gcc-13/
AMD EPYC 9754	733 (score estimated via Ampere lab testing)	333	2.20	Community GCC 13.2 https://gcc.gnu.org/gcc-13/



Processors	SPECrate®2017_int_ base score (estimated)	Published TDP	Performance / Watt (calculated)	Compiler
Intel Xeon 8592+	458 Published performance (result = 550, May-2024)	350 (reference)	1.31	Published scores are derated by an estimated 20% based on Ampere lab testing and analysis. Various test results generated by Ampere on different generations of Intel Xeon Scalable Processor platforms and Intel's OneAPI compilers used for the analysis and estimation. Actual results generated on GCC 13 compiler may vary. Community GCC 13.2 https://gcc.gnu.org/gcc-13/
Intel Xeon 6780E	579 Published performance (result = 695, July-2024)	330 (reference)	1.75	

Relative AmpereOne efficiency advantage is calculated as follows:

2.47 performance/watt of AmpereOne A192-32X is 12% higher than 2.20 performance/watt of AMD EPYC 9754: (2.47 / 2.20) / 2.20 = 12.295%

2.47 performance/watt of AmpereOne A192-32X is 40% higher than 1.77 performance/watt of AMD EPYC 9654: (2.47 / 1.77) / 1.77 = 39.568%

2.47 performance/watt of AmpereOne A192-32X is 89% higher than 1.31 performance/watt of Intel Xeon 8592+: (2.47 / 1.31) / 1.31 = 88.758%

2.47 performance/watt of AmpereOne A192-32X is 41% higher than 1.75 performance/watt of Intel Xeon 6780E: (2.47 / 1.75) / 1.75 = 40.882%

Socket-Level Power Draw (W): Cloud Native Workloads

Measured CPU Usage Power (W)	NGINX	Redis	Memcached	MySQL	Elasticsearch	PostgreSQL	Cassandra
AmpereOne A192-32X	378	300	264	280	297	218	239
AMD EPYC 9654	410	410	410	401	409	331	365
AMD EPYC 9754	410	383	363	374	402	289	294

Socket-Level Performance: Cloud Native Workloads

Processor Under Test	NGINX (req/sec)	Redis (ops/sec)	Memcached (ops/sec)	MySQL (queries/sec)
AmpereOne A192-32X	174,344	178,597,186	105,806,379	408,141
AMD EPYC 9654	136,298	143,131,802	91,770,575	320,436
AMD EPYC 9754	161,693	175,553,748	103,372,044	357,586

Processor Under Test	Elasticsearch (docs/sec)	PostgreSQL (NOPM)	Cassandra (kQps/sec)
AmpereOne A192-32X	4,619,524	6,463,425	1,674
AMD EPYC 9654	4,526,171	7,068,980	1,373
AMD EPYC 9754	5,491,127	8,054,091	1,533

System Hardware & Software Configurations: AI Inference Workloads:

AmpereOne

- HW: 1 x AmpereOne A192-26X (192c/192t, 2.6 GHz), 8 × 64 GiB DDR5 5200 MHz
- OS: Fedora 38
- Kernel: 6.8.9-100.fc38.aarch64

AMD EPYC Genoa

- HW: 1 x AMD EPYC 9654 (96c/192t, 2.4/3.55 GHz), 12 × 64 GiB DDR5 4800 MHz
- OS: Ubuntu 22.04
- Kernel: 6.4.13-200.fc38.x86_64

SMT disabled

AMD EPYC Bergamo

- HW: 1 x AMD EPYC 9754 (128c/256t, 2.25/3.1 GHz), 12 × 64 GiB DDR5 4800 MHz
- OS: Ubuntu 22.04
- Kernel: 6.4.13-200.fc38.x86_64
- SMT enabled

Socket-Level Power Draw (W) & Performance: AI Inference Workloads

Processor Under Test Application Weight	DLRM (dlrm_torchbench)		RESNET-50 (resnet_50_v1.5)		Whisper (whisper_medium.en_hf)		Llama3 8B	
	Performance (ips)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)
AmpereOne A192-26X	1,489,031	270	1,744	207	453,710	288	252	252
AMD EPYC 9654	1,298,523	407	1,565	401	409,364	410	249	387
AMD EPYC 9754	1,246,828	365	1,843	410	417,070	402	349	396

Performance units of measure:

tps = tokens per second

ips = inferences per second

fps = frames per second

sps = samples per second (16KHz audio)

2) Composite Web Service Performance Claims

Rack-Level performance and Power Draw by Workload

AmpereOne application rack performance used as baseline to calculate the # of AMD Genoa and AMD Bergamo systems and power required to match AmpereOne rack performance. Total power draw (all applications combined) used to calculate the total required rack count for AMD Geno and AMD Bergamo.

Web Service Composite		AmpereOne				AMD EPYC 9654		AMD EPYC 9754	
Application	Weight	# Racks	Max # Servers / Rack	Power Draw (W)	Rack Performance	# Servers to match performance	Power Draw (W)	# Servers to match performance	Power Draw (W)
NGINX	20%	1	18	11,415	3,138,192	24	16,948	20	14,123
Redis	20%	1	20	11,123	3,571,943,720	25	17,654	21	14,262
Memcached	20%	1	22	11,444	2,327,740,338	26	18,360	23	15,161
MySQL	20%	1	21	11,259	8,570,961	27	18,823	24	16,084
DLRM	10%	0.5	10	5,262	14,890,307	12	8,440	12	7,930
LLAMA3-8B	10%	0.5	11	5,589	2,767	12	8,201	8	5,536
		5 total	102 total	56,092 total		126 total	88,426 total	108 total	73,096 total

Rack count for AMD EPYC systems calculated by dividing total server power draw by 11,500 (individual rack power budget for servers).

AMD EPYC 9654 Rack Count: 88,426W / 11,500W = 7.69 → 8 Racks

AMD EPYC 9754 Rack Count: 73,096W / 11,500W = 6.36 → 7 Racks

102 servers AmpereOne is 19% lower server count than 126 servers of AMD EPYC 9654:
1 - (102 servers / 126 servers) = 19.048%

126 servers AMD EPYC 9654 is 24% higher server count than 102 servers of AmpereOne:
1 - (126 servers / 102 servers) = 23.529%

102 servers AmpereOne is 6% lower server count than 108 servers of AMD EPYC 9754:
1 - (102 servers / 108 servers) = 5.556%

108 servers AMD EPYC 9754 is 6% higher server count than 102 servers of AmpereOne:
1 - (108 servers / 102 servers) = -5.882%

5 racks AmpereOne is 38% lower rack count than 8 racks of AMD EPYC 9654:
1 - (5 racks / 8 racks) = 37.500%

8 racks AMD EPYC 9654 is 60% higher rack count than 5 racks of AmpereOne:
1 - (8 racks / 5 racks) = -60.000%

5 racks AmpereOne is 29% lower rack count than 7 racks of AMD EPYC 9754:
1 - (5 racks / 7 racks) = 28.571%

7 racks AMD EPYC 9754 is 40% higher rack count than 5 racks of AmpereOne:
1 - (7 racks / 5 racks) = -40.000%

88,426W power draw for AMD EPYC 9654 is 58% higher than the 56,092W power draw for AmpereOne: 1 - (88,426W / 56,092W) = 57.645%

56,092W power draw for AmpereOne is 37% lower than the 88,426W power draw for AMD EPYC 9654: (88,426W - 56,092W) / 88,426W = -36.566%

73,096W power draw for AMD EPYC 9754 is 30% higher than the 56,092W power draw for AmpereOne: 1 - (88,426W / 73,096W) = 30.314%

56,092W power draw for AmpereOne is 23% lower than the 73,096W power draw for AMD EPYC 9754: (73,096W - 56,092W) / 73,096W = -23.262%

3) Consolidation Ratios & Cost of Ownership Claims

CapEx: Processor and Platform Cost Assumptions

Server cost estimates include the same estimated costs for all AMD and Ampere platforms for the following components:

- $2,300 per server for a 1U chassis
- $400 for internal storage
- $350 for 1 DIMM of 64GB DDR5 memory for a total of $4,200 for AMD EPYC 9654 and AMD EPYC 9754 (12 memory DIMMs each) and $2,800 for AmpereOne A192-32X (8 memory DIMMs) based on actual test configuration

AMD EPYC server cost estimates include processor cost based on published 1KU pricing as of September 27, 2024:

- $10,625 for AMD EPYC 9654P
- https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9654p.html
- $11,900 for AMD EPYC 9754
- https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9754.html

AmpereOne server cost estimates include processor cost based on published suggested base volume (SBV) price published as of September 27, 2024:

- $5,555 for AmpereOne A192-32X
- https://amperecomputing.com/briefs/ampereone-family-product-brief

Individual Server Cost Assumption:

Summary	AmpereOne A192-32X	AMD EPYC 9654	AMD EPYC 9754
Base	$2,300	$2,300	$2,300
MEM	$2,800	$4,200	$4,200
SSD	$400	$400	$400
CPU	$5,555	$10,625	$11,900
Total	**$11,055**	**$17,525**	**$18,800**

Total Server Acquisition Costs in Support of Figure 10:

Summary	AmpereOne A192-32X	AMD EPYC 9654	AMD EPYC 9754
Single Server Cost	$11,055	$17,525	$18,800
Server Count	102	126	108
Total CapEx	$ 1,127,610	$ 2,208,150	$ 2,030,400

OpEx: Ongoing Cost Assumptions

OpEx costs include electricity costs, server administrator cost, data center provisioning costs. The time period for TCO analysis is 36 months. Electricity cost assumed at $0.20 per kilowatt hour (kWh). Total power draw per server assumes PUE factor of 1.5 and power load factor (utilization) of 100.00%.

Data center provisioning costs assumed at $0.083 per kilowatt hour (kWh). Data center provisioning cost is the ongoing cost of the physical infrastructure necessary to house, power, cool, and operate data center server hardware. It is a function of the proportion of total available power consumed by the computing infrastructure.

Server administrator costs consist of an assumed $100,000.00 annual salary and a 35% salary overhead for a total annual labor cost of $135,000.00 per server administrator. Each administrator is assumed to maintain 75 servers. The annual server administrator cost per server is calculated as $135,000.00/75 = $1,800.00.

1st & 2nd Gen Intel Xeon Scalable and AMD EPYC Consolidation Assumptions

Processor Generations	SPECrate®2017_int_ base score (estimated)	CPU Power	Qty Systems per Rack	Performance per Rack	Consolidation Ratio (versus AmpereOne)
Intel Xeon Scalable 1st Gen: “Skylake”	143	150	19	2,718	5.4
Intel Xeon Scalable 2nd Gen: “Cascade Lake”	202	140	20	4,044	3.6
AMD EPYC 7001 “Naples”	203	155	20	4,069	3.62
AMD EPYC 7002 “Rome”	385	190	18	6,925	2.13

Ampere Performance per rack: 21 systems x 702 SPECint®2017_int_base score = 14,742. Intel and AMD performance assumptions based on export of published SPEC CPU®2017 Integer Rate Result reports from https://www.spec.org/cgi-bin/osgresults?conf=rint2017 as of September 3, 2024. Results filtered by “# Chips” = 2 indicating the assumption of each server having two CPU sockets. For Platform power and rack configuration assumptions, see above section “I. Performance Claims” within “End Notes”. An additional 15W is added to account for the second socket on each motherboard.

Intel Skylake single rack annual cost calculation:

12,500W total rack power * 1.5 PUE * 100% Power Load Factor * 24 Hours * 365 Days / 1000 = 164,250 kWh annual power draw

(164,250 kWh * $0.20 electricity cost) + (164,250 kWh * $0.083 data center provisioning cost) + ($1,800 server admin cost * 19 servers per rack) = $80,682.75 annual cost per rack

AmpereOne single rack annual cost calculation:

12,500W total rack power * 1.5 PUE * 100% Power Load Factor * 24 Hours * 365 Days / 1000 = 164,250 kWh annual power draw

(164,250 kWh * $0.20 electricity cost) + (164,250 kWh * $0.083 data center provisioning cost) + ($1,800 server admin cost * 21 servers per rack) = $84,282.75 annual cost per rack

Options	CapEx (one-time)	OpEx (monthly)	TCO (3 years)
Staying on Intel Skylake (20 racks)	$0	$134,471	$4,840,965
Upgrade to AmpereOne (4 racks)	$928,620	$28,094	$1,940,013

Detailed 3 year cost schedule:

Quarterly	Q1 Y1	Q2 Y1	Q3 Y1	Q4 Y1	Q1 Y2	Q2 Y2	Q3 Y2	Q4 Y2	Q1 Y3	Q2 Y3	Q3 Y3	Q4 Y3
Intel (20 racks)	$403,414	$403,414	$403,414	$403,414	$ 403,414	$403,414	$403,414	$403,414	$403,414	$ 403,414	$403,414	$403,414
Ampere (4 racks)	$1,012,903	$84,283	$84,283	$ 84,283	$ 84,283	$84,283	$ 84,283	$84,283	$84,283	$ 84,283	$84,283	$84,283

Composite AI Compute Web Stack Cost Assumptions

See "Rack-Level performance and Power Draw by Workload" notes above for details on calculating total server footprint.

Total Server CapEx:

Summary	AmpereOne A192-32X	AMD Genoa 9654	AMD Bergamo 9754
Server Cost	$11,055	$17,525	$18,800
Server Quantity	102	126	108
TOTAL CapEx / Yr	**$1,127,610**	**$2,208,150**	**$2,030,400**

Total Server OpEx (per year):

Summary	AmpereOne A192-32X	AMD Genoa 9654	AMD Bergamo 9754
Total Server Power / Yr	737,049 kWh	1,161,922 kWh	960,478 kWh
Power Cost / Yr	$147,410	$232,384	$192,096
DC provisioning cost / yr	$61,175	$96,440	$79,720
Admin cost / yr	$183,600	$226,800	$194,400
Total OpEx / Yr	**$392,185**	**$555,624**	**$466,215**

Detailed 3 year cost schedule in support of Figure 11:

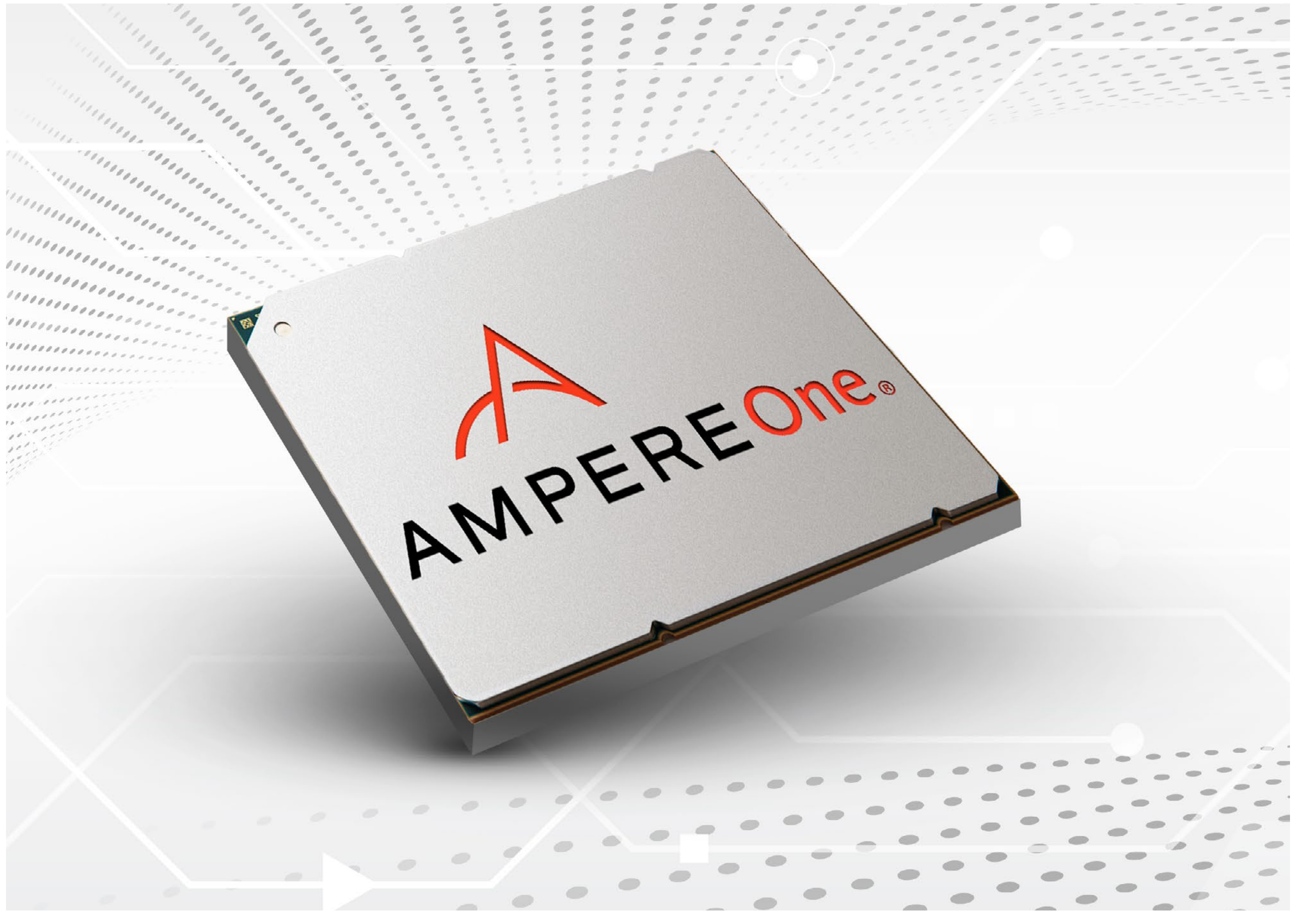
Quarterly	Q1 Y1	Q2 Y1	Q3 Y1	Q4 Y1	Q1 Y2	Q2 Y2	Q3 Y2	Q4 Y2	Q1 Y3	Q2 Y3	Q3 Y3	Q4 Y3	TOTAL
AmpereOne A192-32X	$1,225,656	$98,046	$98,046	$98,046	$98,046	$98,046	$98,046	$98,046	$98,046	$98,046	$98,046	$98,046	**$2,304,164**
AMD Genoa 9654	$2,347,056	$138,906	$138,906	$138,906	$138,906	$138,906	$138,906	$138,906	$138,906	$138,906	$138,906	$138,906	**$3,875,022**
AMD Bergamo 9754	$2,146,954	$116,554	$116,554	$116,554	$116,554	$116,554	$116,554	$116,554	$116,554	$116,554	$116,554	$116,554	**$3,429,046**

Hence, supporting this composite web service with AmpereOne would cost 41% less over 3 years than supporting it with AMD EPYC 9654:

($3,875,022 - $2,304,164) / ($3,875,022) = 40.538%

Alternatively, supporting this composite web service with AmpereOne would cost 33% less over 3 years than supporting it with AMD EPYC 9754:

($3,429,046 - $2,304,164) / ($3,429,046) = 32.805%



4) Industry and Market Claims

1. Generative AI growth projection claims in accordance with Bloomberg "Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds".

Full article accessible here:
https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

2. Claims around AI adoption among SMB and enterprises in accordance with Colorwhistle "AI Statistics for Small Business" and Forbest "SMBs Flourish: Embracing AI for Efficiency and Engagement".

Full articles accessible here:
https://colorwhistle.com/artificial-intelligence-statistics-for-small-business/

https://www.forbes.com/sites/garydrenik/2024/06/25/smbs-flourish-embracing-ai-for-efficiency--engagement/

3. Claims around power consumption increases across different generations of GPUs in accordance with Forbes "AI Power Consumption: Rapidly Becoming Mission-Critical".

Full article accessible here: https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

4. Global Data Center Energy Consumption and growth figures in accordance with International Energy Agency (IEA) "Electricity 2024 – Analysis and forecast to 2024" report and corresponding Data Center Frontier article "IEA Study Sees AI, Cryptocurrency Doubling Data Center Energy Consumption by 2026".

Full IEA report accessible here:
https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf

5. Claim around utility provider concern in accordance with CNBC "Failure to meet surging data center energy demand will jeopardize economic growth, utility execs warn".

Full article accessible here:
https://www.cnbc.com/2024/06/30/failure-to-meet-surging-energy-demand-will-jeopardize-economic-growth-utility-execs-warn.html

6. Claims around data center PUE in accordance with Uptime Institute "Large data centers are mostly more efficient, analysis confirms".

Full report accessible here:
https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/

7. Claims around rack-level power constraints and PUE assumptions in accordance with Uptime Institute "Global Data Center Survey Results 2023".

Full report accessible here:
https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2023

8. Claims around data center challenges to adopt new infrastructure cooling technologies in accordance with Equinix "Data Center Cooling Continues to Evolve for Efficiency and Density".

Full report accessible here:
https://blog.equinix.com/blog/2023/12/11/data-center-cooling-continues-to-evolve-for-efficiency-and-density/

9. Claims around server refresh cycles, unplanned downtime and maintenance cost trends of aging infrastructure accordance with "Global Data Center Survey 2023", IDC Research "Adopting a Technology Rotation Program from Dell Improves Operational and Cost Efficiencies for Servers", and Grassroots IT "The Hidden Costs of Aging Technology Infrastructure".

Executive summary accessible here:
https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2023

Full articles accessible here:
https://www.delltechnologies.com/asset/en-us/solutions/financing-and-payment-solutions/industry-market/idc-business-value-tr-servers.pdf

https://www.grassrootsit.com.au/blog/the-hidden-costs-of-aging-technology-infrastructure/

10. Claims around operational cost trends of aging infrastructure in accordance with VentureBeat "Data center modernization: The heavy — and rising cost — of doing nothing".

Full article accessible here:
https://venturebeat.com/data-infrastructure/data-center-modernization-the-heavy-and-rising-cost-of-doing-nothing/

11. Statements about NVIDIA HGX H200 and DGX H200 power consumption and thermal design power based on NVIDIA and Quanta Computer data sheets located here:

https://resources.nvidia.com/en-us-dgx-systems/dgx-h200-datasheet

https://www.nvidia.com/en-us/data-center/h200/

https://www.qct.io/product/index/Server/rackmount-server/GPGPU-Xeon-Phi/QuantaGrid-D74H-7U#specifications

12. Claims around the percentage make-up of AI compute cycles based on the "AI Semis Market Landscape" study by D2D Advisory Inc.

Full article accessible here:
https://digitstodollars.com/2023/05/17/ai-semis-market-landscape/

Appendix

Appendix A — Cloud Native Workload Test Methodologies

Software configurations selected for each workload and each processor optimized for maximum socket-level throughput.

NGINX (req/sec)

Community GCC 13.2

Throughput measured at p.99 latency of 1ms

Client:

- Wrk HTTP/S Generation Tool generates millions of HTTPS requests over hundreds of connections

Target:

- NGINX Web Server responds to HTTPS requests for file
- 1 NGINX Worker Process per Core

Additional NGINX-side Processing (for 50KB file):

- SSL/TLS decrypt/encrypt: HTTPS protocol
- Lua: Requested file redirected for processing with Lua — remove all line breaks and spaces, add timestamp
- Compression: Compress processed file

CPU intensive: ~100% core utilization

REDIS & Memcached (ops/sec)

Community GCC 13.2

Throughput (total number of get/set ops) measured at p.99 latency of 1ms

Server Side: redis (version 7.2.0), Memcached (version 1.6.21)

Each instance is a single process and is allocated 2.0 GB of memory and configured to hold 13,600,000 keys/data before eviction will happen.

Client Side: memtier_benchmark (version 1.3.0)

Configuration

- We first populate the cache, so we get a 95% hits rate → 5% of key/value requests will come as not found
- 1:10 set/get ratio → 1 key/value write, and 10 key/value read
- 64 bytes payload (average) → 56% 16 Bytes, 30% 64 Bytes, 12% 128 Bytes, 2% 1024 Bytes
- Data sent to servers is randomized
- Clients per thread and concurrent pipelined requests → we pick the best configuration for each platform

Each run is a 2 minutes-long test (after the cache is populated).

Some run-to-run variability, so we perform 5 runs and get the median value.

MySQL (queries/sec)

Community GCC 13.2

Throughput measured at p.95 latency of 1ms

Sysbench multi-threaded benchmark tool

OLTP load SQL queries: POINT_SELECT, SELECT_SIMPLE_RANGE, SELECT_SUM_RANGES, SELECT_ORDER_RANGES, SELECT_DISTINCT_RANGES, UPDATE_KEY, UPDATE_NO_KEY

Tests based on SQL queries:

- sb11-OLTP_RO_10M_8tab-uniform-dst_ranges1-notrx : 5
- sb11-OLTP_RO_10M_8tab-uniform-notrx : 1+2+3+4+5

- sb11-OLTP_RO_10M_8tab-uniform-p_sel1-notrx : 1
- sb11-OLTP_RO_10M_8tab-uniform-s_ranges1-notrx : 2
- sb11-OLTP_RW_10M_8tab-uniform-notrx : 1+2+3+4+5+6+7
- sb11-OLTP_RW_10M_8tab-uniform-upd_idx1-notrx : 6

Elasticsearch (docs/sec)

Community GCC 13.2

Throughput measured at p.99 latency of 1s

Elasticsearch: v8.0.0, **Rally benchmark tool:** v2.7.0, Docker: v23.0.2, JDK: v17

Optimization Configurations

1) Bare-metal machine benchmark, 1P mode, 100GbE networking interface, 250GB+ Memory, NVMe SSDs, Ext4 filesystem
2) Pull the official Docker Elasticsearch v8.0.0 image (docker.elastic.co/elasticsearch/elasticsearch:8.0.0)
3) Use the recommended bundled JVM(Java) version in Docker container Elasticsearch (OpenJDK 64-Bit 17.0.1)
4) Set minimum and maximum JVM heap size 8GB (ES_JAVA_OPTS, -Xms8g –Xmx8g)
5) JVM Garbage Collector: G1GC
6) Rally as load generator: One 100GbE network interfaces, Rally Track HTTP_LOGS (HTTP server log data)

Esrally http_logs command, e.g.: esrally race --track=http_logs --target-hosts=10.16.105.199:9200 --challenge=append-no-conflicts-index-only --pipeline=benchmark-only --track-params='bulk_indexing_clients:16' --on-error=abort --kill-running-processes

PostgreSQL (NOPM)

Community GCC 13.2

Setup

- PostgreSQL DBs created on SUT
- All DBs exist on external drive (1 DB per partition)
- Ideally, number of DBs on system determined by core count of system, each DB gets 16 cores
- Not the case for AMD, which can't support the NVME requirement for this test case
- Genoa – 8 DBs tested with 24 VUs
- Bergamo – 10 DBs tested with 23 VUs

Run

- HammerDB is used to test the DBs using it's TPCC like workload

Results

- Metric for throughput is New Orders Per Minute (NOPM)

Cassandra (kQps/sec)

Community GCC 13.2

Throughput measured at p.99 latency of 10 ms

TLP-Stress tool

- Number of server nodes: 8
- Server heap per node: 32G
- Server commit/data log: 4 NVMe (2 partitions per NVMe)
- Number of clients used: 1
- Client tests: RandomPartitionAccess
- Number of Procs per client: 8
- Client test time: 5m

Appendix B — AI Inference Workload Test Methodologies

Software configurations and data format selected for each workload and each processor optimized for maximum socket-level throughput. LLAMA3 configuration optimized for token-generation throughput.

DLRM

AmpereOne: data format = FP16, Framework: amperecomputingai/pytorch:1.10.0 (docker image), batch size = 2,048

AMD Genoa: data format = BF16, Framework: Ubuntu, 22.04 (docker image + torch==2.3.1 from PyPI (python 3.10)), batch size = 4,096

AMD Bergamo: data format = BF16, Framework: Ubuntu, 22.04 (docker image + torch==2.3.1 from PyPI (python 3.10)), batch size = 4,096

ResNet-50

AmpereOne: data format = FP16, Framework: amperecomputingai/pytorch:1.10.0 (docker image), batch size = 128

AMD Genoa: data format = BF16, Framework: Ubuntu, 22.04 (docker image + torch==2.3.1 from PyPI (python 3.10)), batch size = 4

AMD Bergamo: data format = BF16, Framework: Ubuntu, 22.04 (docker image + torch==2.3.1 from PyPI (python 3.10)), batch size = 2

Whisper

AmpereOne: data format = FP16, Framework: amperecomputingai/pytorch:1.10.0 (docker image), batch size = 2

AMD Genoa: data format = FP32, Framework: Ubuntu, 22.04 (docker image + torch==2.3.1 from PyPI (python 3.10)), batch size = 4

AMD Bergamo: data format = FP32, Framework: Ubuntu, 22.04 (docker image + torch==2.3.1 from PyPI (python 3.10)), batch size = 2

LLAMA3 8B

AmpereOne: Q4, pp128, batch size = 16

llama.aio - r1.2.6, Meta-Llama-3-8B-Instruct.Q4_K_4.gguf

AMD Genoa: Q4_K_M, pp128, batch size = 8, llama.cpp – b3452, Meta-Llama-3-8B-Instruct.Q4_K_M.gguf

AMD Bergamo: Q4_K_M, pp128, batch size = 8, llama.cpp – b3452, Meta-Llama-3-8B-Instruct.Q4_K_M.gguf