

Accelerate Large-scale AI Model Training, Tuning, and Inference With HPE and AMD

HPE ProLiant Compute XD685 Servers With AMD Instinct™ MI355X GPUs

By Tony Palmer, Practice Director

December 2025

Contents

Introduction	3
AI Infrastructure Challenges	3
HPE ProLiant Compute XD685 With AMD Instinct™ MI355X GPUs	4
Omdia Technical Validation.....	5
Accelerate Large-scale AI With Efficiency and Security With HPE and AMD	5
Omdia Analysis.....	6
Overcoming Infrastructure Challenges in Large-scale AI Deployment With HPE and AMD	7
The Solution: HPE ProLiant Compute XD685 With AMD Instinct™ MI355X.....	7
Performance and Cost of Ownership	9
Conclusion.....	11

Introduction

This Technical Validation from Omdia documents our evaluation of HPE ProLiant Compute XD685 with AMD Instinct™ MI355X graphics processing units (GPUs) , and AMD EPYC™ processors, a key offering from HPE for large-scale artificial intelligence (AI) model building, training, and inference. We evaluated how service providers and enterprises can use the XD685 to deploy the massive AI clusters needed to build and train large AI models efficiently and securely.

AI Infrastructure Challenges

The demand for advanced AI models capable of processing enormous amounts of information, generating insights, and driving innovation at scale is growing rapidly, and the underlying infrastructure is a key to success for the service providers and enterprises that are building, training, and inferencing with large models.

As organizations accelerate their investments in AI, they are increasingly recognizing the critical role of infrastructure across the lifecycle. In 2025, Enterprise Strategy Group (now Omdia) surveyed 350 IT professionals at organizations in North America (U.S. and Canada) involved with or responsible for evaluating, purchasing, managing, and building application infrastructure for on-premises and cloud environments. Respondents overwhelmingly reported that their organization is making or planning to make significant investments in infrastructure to support new AI workloads (91%). Those same respondents identified model development (64%), tuning (67%), inferencing (67%), and training (51%) as areas of targeted support (see Figure 1).¹

Figure 1. Infrastructure Is Key Across the AI Lifecycle

You indicated that your organization is making (or planning to make) significant investments in infrastructure to support new AI workloads. Which of the following areas of the AI lifecycle is your organization planning to support with these new infrastructure investments? (Percent of respondents, N=317, multiple responses accepted)



Source: Omdia

Additionally, organizations called out compute (43%) as a top priority for AI inference infrastructure.² Adequate compute resources ensure that models can execute efficiently and handle large volumes of data and inquiries without latency issues.

¹ Source: Enterprise Strategy Group (now Omdia) Research Report, *IT Transformed: Inside the Convergence of Hybrid Cloud and AI*, July 2025.

² Source: Enterprise Strategy Group (now Omdia) Research Report, *Navigating Build-versus-buy Dynamics for Enterprise-ready AI*, January 2025.

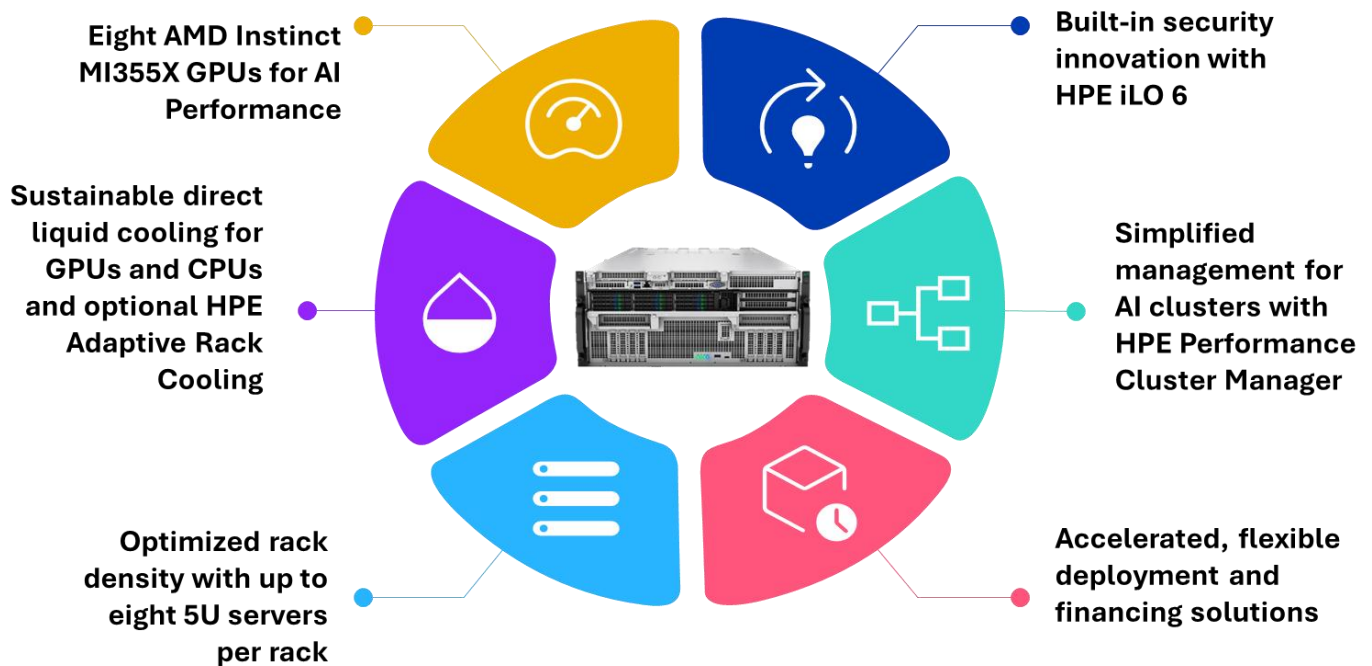
The same respondents identified customized GPUs (28%) and CPUs (25%) as the primary infrastructure approach their organizations take related to processing and compute, the top two responses.³

HPE ProLiant Compute XD685 With AMD Instinct™ MI355X GPUs

The HPE ProLiant Compute XD685 is a high-performance AI system powered by eight AMD Instinct™ MI355X GPUs with AMD EPYC™ 9005 Series CPUs. The XD685 is designed to address core challenges across the entire AI lifecycle at scale.

- **Build and train large models.** The solution enables organizations to accelerate the building and training of new large models using supervised and unsupervised learning, greatly reducing the time to results for large model and large language model (LLM) training.
- **Support global footprints and empower growth.** HPE helps organizations establish and grow their presence anywhere in the world. HPE has deployed massive custom AI clusters worldwide with comprehensive serviceability, enabled by their secure global supply chain.
- **Deploy efficient and secure large AI clusters.** Direct liquid cooling (DLC) brings efficiencies and performance, while HPE iLO management delivers robust built-in security and protection with silicon root of trust.
- **Accelerate AI with a ready software ecosystem.** The AMD ROCm™ platform supports leading AI frameworks and inference engines, enabling efficient training, tuning, and deployment of modern AI models at scale.

Figure 2. HPE ProLiant Compute XD685 With AMD Instinct MI355X GPUs



Source: HPE and Omdia

³ Ibid.

The HPE ProLiant Compute XD685 with AMD Instinct MI355X GPUs is designed with more than just performance in mind.

- **Scalability:** The AMD Instinct MI355X GPU delivers performance, efficiency, and memory capacity for generative AI, training, and HPC. With 288GB of HBM3E and support for up to 520B parameter models on a single GPU, AMD Instinct MI355X GPUs can power demanding workloads. Expanded datatype support, including FP4 and FP6, enables optimal compute density and energy efficiency across inference and training. Scalability is delivered via HPE's deep expertise building large scale clusters.
- **Efficiency:** The HPE ProLiant Compute XD685 with AMD Instinct MI355X GPUs, housed in a 5U DLC chassis, utilizes sustainable DLC engineered for efficiency and performance. HPE has nearly five decades of DLC experience, with a dedicated thermodynamics research lab, and has deployed some of the world's largest liquid-cooled IT environments. HPE Performance Cluster Manager is integrated system management software that can automate and accelerate setup and operations of complex systems.
- **Security:** HPE iLO has been the basis of HPE ProLiant server management for 20 years and enables customers to securely configure, monitor, and update HPE servers from anywhere. 5th generation AMD EPYC processors introduce Secure Encrypted Virtualization (SEV) Trusted IO, a new SEV technology, to the Infinity Guard™ feature set,⁴ extending data security integrity across external trusted devices. Additionally, AMD Instinct MI350 Series GPUs are designed to enhance reliability, scalability, and data security for AI and cloud workloads by providing trusted firmware, verifying hardware integrity, and encrypting GPU communication.
- **Services:** HPE has demonstrated expertise in designing, installing, and supporting AI environments at scale globally with serviceability. HPE offers numerous professional, advisory, and operational services to guide organizations toward achieving their specific business goals and deploying the ideal solution for their initiatives. HPE Advisory and Professional Services experts work closely with organizations to build a foundation for successfully realizing their vision.
- **Sovereign AI:** Nation-states, public entities, and organizations operating internationally need a purpose-built sovereign solution that delivers the scale, performance, and governance to enable growth, accelerate AI innovation, and drive a positive societal impact while protecting data and intellectual property. HPE and AMD enable organizations to build a robust platform combining the most advanced AMD AI technology with HPE's efficient infrastructure in a truly sovereign environment.

Omdia Technical Validation

Omdia validated how AI service providers and large model builders can benefit from using HPE ProLiant Compute XD685 servers with AMD Instinct™ MI355X GPUs for building, training, and inference of large AI models.

[Accelerate Large-scale AI With Efficiency and Security With HPE and AMD](#)

AI is a transformative force, spurring innovation and driving growth. Organizations want to use AI to make themselves more competitive, delivering more accurate insights ever faster and enabling better business results. The challenges of scaling AI can quickly overwhelm the capabilities of traditional compute environments.

⁴ Learn more about Infinity Guard at <http://www.amd.com/en/products/processors/server/epyc/infinity-guard.html>.

In addition, the increasing power demands of AI systems have led to corresponding rises in rack density and cooling requirements. While higher rack density allows for increased data and processing capacity, it simultaneously generates greater energy consumption and heat output.

Organizations providing AI services and developing large-scale models face pressure to rapidly establish extensive AI clusters to meet expanding customer demands and maintain market competitiveness. These entities require effective strategies for power utilization and rack cooling management. Their operations rely on integrated, reliable, and secure computing environments that support quick deployment and enable large-scale AI performance optimization.

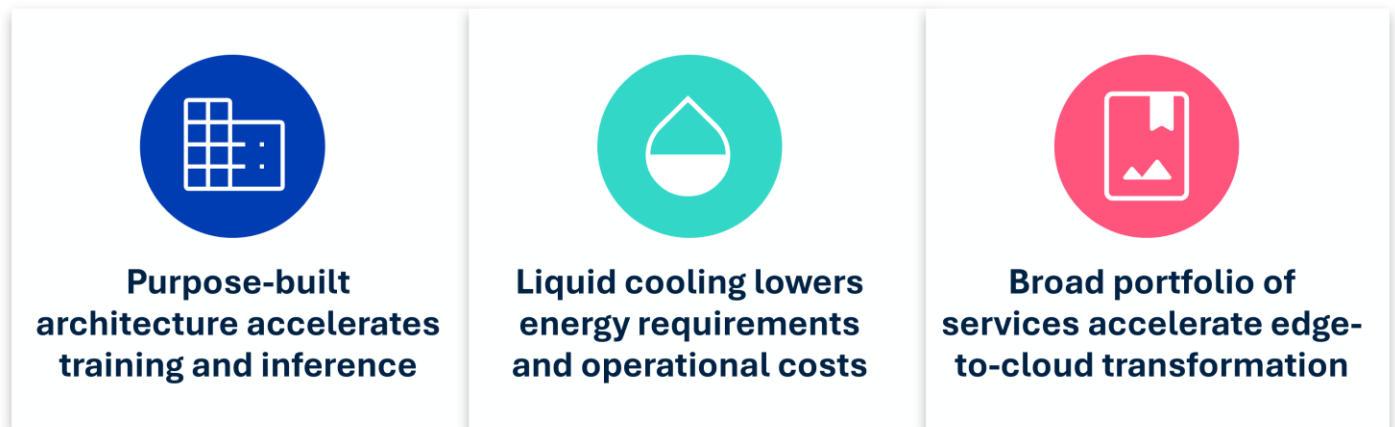
Omdia Analysis

Training LLMs requires strong scaling and massively parallel compute capabilities. HPE and AMD have collaborated to extend their technologies and services to support a growing market of large-scale AI model builders and accelerate scientific and engineering breakthroughs across industries. Their global presence and deep expertise help them empower AI service providers and enterprises to build and train large AI models and deploy AI clusters across the globe.

Featuring eight of the latest AMD Instinct™ MI355X GPUs and two 5th Generation AMD EPYC™ processors, the HPE ProLiant Compute XD685 is optimized for intensive AI workloads like natural language processing, LLM training, and multimodal training. DLC increases energy efficiency and lowers operational costs. For a seamless end-to-end operating experience, these systems include built-in security features delivered via HPE iLO as well as native capabilities in AMD CPUs and GPUs. HPE Services has decades of expertise helping organizations deploy some of the largest computing clusters in the world.

The AMD ROCm™ platform enables efficient training, tuning, and inference with leading AI frameworks to help organizations fully leverage the capabilities of MI355X-based clusters in a robust software ecosystem.

Figure 3. End-to-end AI With HPE ProLiant Compute XD685 and AMD



Source: Omdia

Why This Matters

AI is reshaping competitive dynamics across industries, driving organizations to make substantial infrastructure investments. The success of these investments will determine both IT and broader business outcomes. As AI transforms how organizations interact with applications and data, it fundamentally alters infrastructure architecture and operational investment strategies.

Omdia validated that HPE and AMD collaborate to deliver purpose-built solutions to meet the demands of AI and evolve for future requirements. Whether you are an AI service provider or an ambitious enterprise building a large AI model, HPE and AMD can transform your environments with flexible, high-performance solutions that can be brought to market quickly.

Overcoming Infrastructure Challenges in Large-scale AI Deployment With HPE and AMD

Omdia explored how a hypothetical rapidly growing AI service provider, faced with critical infrastructure challenges while scaling its large model training operations, could address power density, cooling, and deployment obstacles using HPE ProLiant Compute XD685 servers powered by AMD Instinct™ MI355X GPUs and 5th generation EPYC processors.

For this example, we modeled a midsize AI service provider specializing in custom LLM development for enterprise clients across healthcare, finance, and scientific research sectors. With a growing customer base demanding increasingly sophisticated AI models, the company needed to rapidly expand its training infrastructure while maintaining competitive operational costs.

This company's existing data center infrastructure was struggling with the sharp growth in power needs driven by increasing processing requirements. As processing demands by AI increased and server density followed suit, power requirements quickly escalated as well, creating an urgent need for more efficient solutions.

Greater rack density enables greater processing power and data capacity, but it also results in higher energy use and heat generation. Traditional air-cooled systems were becoming inadequate for their expanding AI workloads, leading to thermal throttling during intensive training sessions, increased operational costs due to inefficient cooling, and limited scalability for future growth plans.

The company faced significant challenges in quickly setting up large AI clusters to serve the needs of its growing customer base and stay competitive in a dynamic market. Its existing self-built infrastructure lacked an integrated, stable, and secure compute environment that could be deployed quickly and accelerate AI performance on a massive scale.

The Solution: HPE ProLiant Compute XD685 With AMD Instinct™ MI355X

The company implemented HPE ProLiant Compute XD685 servers, featuring eight AMD Instinct™ MI355X accelerators per system and dual 5th Generation AMD EPYC™ 9005 series processors. The XD685 architecture is optimized for intensive AI workloads including natural language processing, LLM training, and multimodal training, delivered in a compact 5U package configured in an eight node per rack arrangement for maximum density.

The AMD Instinct MI355X GPUs delivered significantly improved AI performance compared to previous generations, enabling the provider to run the largest AI models with fewer GPUs, significantly reducing its overall hardware footprint and associated costs.

The implementation of DLC transformed operational efficiency. Energy efficiency increased dramatically compared to traditional air cooling, and operational costs were lowered through reduced power consumption, helping to advance the company's sustainability goals.

The trusted ecosystem of HPE Services helped the company build, integrate, validate, and customize AI environments at the largest scale, facilitating quicker on-site deployment and shorter time to insight. This comprehensive support system enabled faster time to market for new AI services, reduced deployment complexity, and enhanced operational reliability. The AMD ROCm™ software ecosystem further streamlined deployment by enabling developers to use familiar AI frameworks and tools, helping accelerate model development and operational readiness.

The XD685 platform's built-in security features include AMD Infinity Guard™ secure processor technology; HPE iLO embedded server management and silicon root of trust; and AMD Instinct MI355 GPUs trusted firmware, verified hardware integrity, secure multi-tenant GPU sharing, and encrypted GPU communication, which all offer protection against increasingly complex threats.

The provider simplified management using HPE Performance Cluster Manager software, which provided automated setup from bare-metal and detailed telemetry and GPU stress tests for continued resilience. HPE iLO 6 provided seamless configuration, monitoring, and updates. The HPE AI Performance Engineering team worked with the service provider to assist the customer with sizing AI workloads and tuning individual applications with the goal of optimizing performance throughout the life of the systems.

HPE and AMD enabled the service provider to successfully establish its presence as a competitive AI service provider, leveraging the same technologies used to power some of the world's fastest and most sustainable supercomputers. The solution enabled the company to make AI work effectively for its business and its customers while maintaining operational sustainability.

Why This Matters

Omdia validated that the HPE ProLiant Compute XD685 with AMD Instinct MI355X GPUs provide the foundation for a comprehensive solution that can address customers' critical infrastructure challenges. By combining performance, energy efficiency, advanced cooling, and integrated security with expert services, the solution enabled rapid scaling while maintaining competitive operational costs.

This collaboration between HPE and AMD demonstrates how purpose-built AI infrastructure can transform organizational capabilities, enabling service providers and large model builders to meet the growing demands of AI while advancing sustainability goals and maintaining security standards.

Performance and Cost of Ownership

First, Omdia examined AMD-submitted MLPerf performance testing of multiple generations of AMD Instinct™ GPUs to gauge generational performance improvements. Testing compared the MI300X and MI325X with the current-generation MI350X and the MI355X found in the HPE ProLiant Compute XD685.⁵

The AMD Instinct MI355X GPU is designed to deliver efficient compute performance for the latest generation of AI models and intended for organizations aiming to advance their AI model training and inference capabilities. The MI355X GPU is integrated with the AMD ROCm™ software ecosystem, providing developers with a robust, open platform for creating and scaling AI applications.

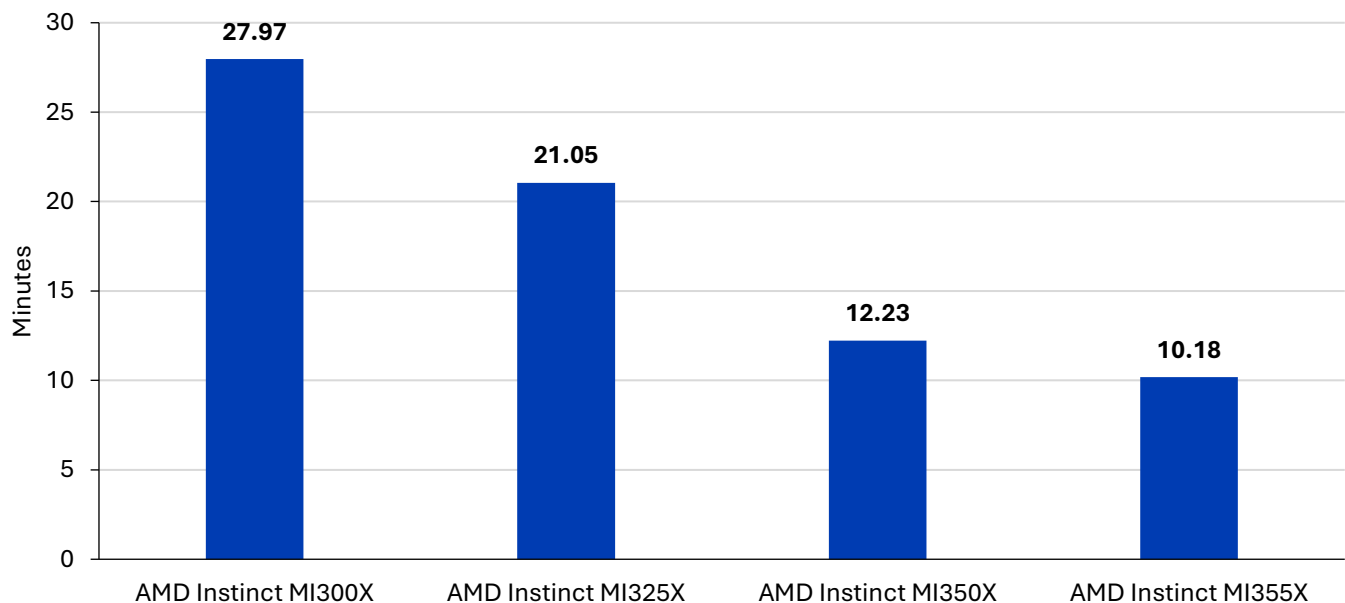
MI350 Series GPUs have native support for FP4 (4-bit floating-point) precision, offering up to 20 petaflops of FP4 performance combined with 288 GB of high-bandwidth HBM3e memory with a bandwidth of 8TB/s to minimize memory and computational overhead without sacrificing accuracy.

The Llama 2 70B LoRA benchmark focuses on the efficient fine-tuning of a large-scale language model using advanced parameter-efficient training techniques. Llama 2 70B LoRA fine-tuning was submitted on four AMD Instinct GPUs with results shown in Figure 4. All results use the same training code, and the only difference between them is the configuration file that is customized for each GPU, ensuring that the benchmark reflects a fair comparison of compute capabilities across three generations of AMD Instinct GPUs.

As seen in Figure 4, the AMD Instinct MI355X GPUs show an impressive 2.8x improvement in training time over the MI300X GPU and a 2.1x improvement over the MI325X GPU.

Figure 4. MLPerf Inference v5.1 Llama 2 70B LoRA Fine-tuning

Time to complete (lower is better)

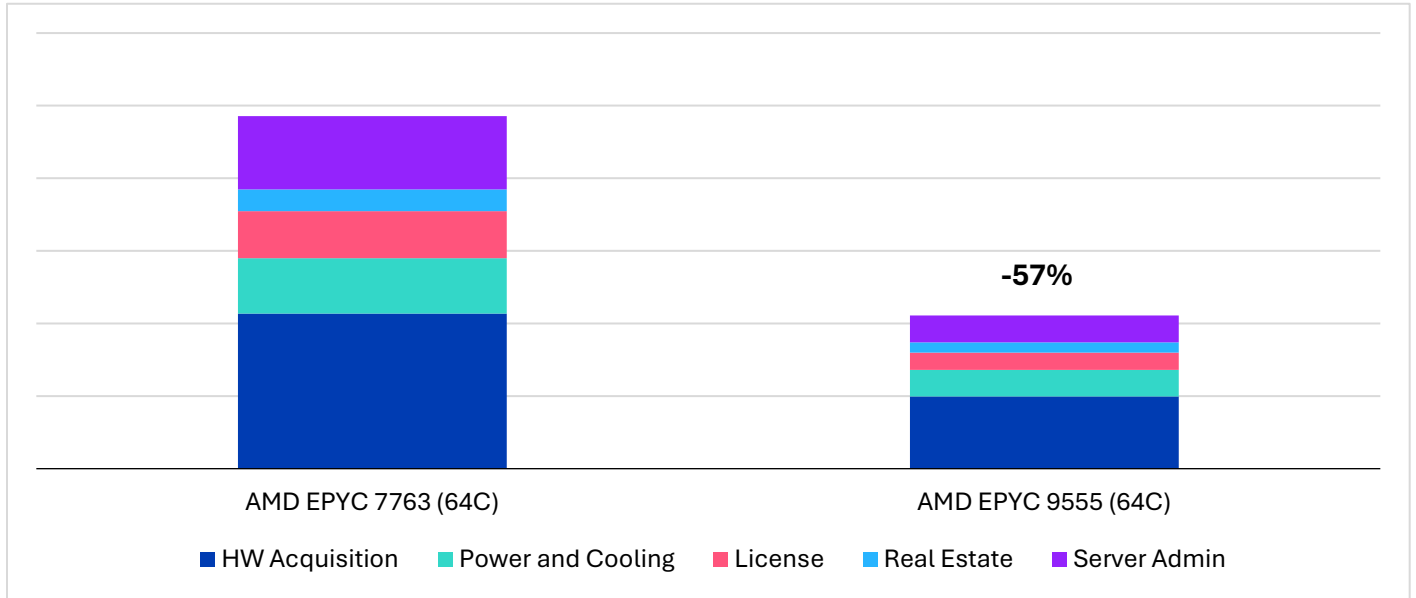


Source: HPE and Omdia

⁵ For full benchmark configuration and parameters, visit: [Reproducing AMD MLPerf Training v5.1 Submission Result](#).

Next, we modeled a simple cost of ownership analysis using data provided by AMD and HPE, along with publicly available data for costs associated with power, cooling, real estate, and IT administration. Metrics included in the analysis were cost metrics for hardware acquisition, power and cooling, software licensing, data center real estate, and server admin and maintenance.

Figure 5. Cost of Ownership Compared to Previous-generation AMD EPYC CPUs



Source: HPE and Omdia

For a large-scale AI environment, we found that the HPE ProLiant Compute XD685 server, powered by AMD EPYC™ 9555 processors, could provide a significant cost of ownership improvement of 57%, delivering equivalent compute power with 36% of the server footprint of previous generation AMD EPYC 7763 processors with similar core counts. The reduction in server footprint drove all the other cost of ownership savings: power and cooling (-52%), software licensing (-64%), data center real estate (-52%), and server maintenance and administration (-64%).

Why This Matters

Service providers and large model builders need a scalable, performant platform that can meet performance requirements for demanding AI workloads while optimizing costs.

Omdia’s analysis of GPU and CPU performance benchmarks and cost showed that the HPE ProLiant Compute XD685 server with AMD Instinct™ MI355X GPUs and EPYC CPUs offer superior performance to previous generations, delivering significantly lower cost of ownership.

Conclusion

Growing demand for advanced AI models is driving rapid infrastructure investment among service providers and large enterprises. Organizations recognize infrastructure as critical throughout the AI lifecycle, from model development to training, tuning, and inferencing. Compute power is a top priority for AI inference infrastructure, as it ensures efficient model execution and handling of large data volumes without latency.

The joint solution offered by HPE and AMD combines their technologies and services to support AI service providers and large model builders, accelerating scientific, engineering, and business breakthroughs across industries. Their global presence and deep expertise empower AI service providers and enterprises to build and train large AI models and deploy robust AI clusters across the globe.

Omdia validated that the HPE ProLiant Compute XD685 with AMD Instinct™ MI355X GPUs enables organizations to scale AI innovation effectively while maintaining performance standards with a purpose-built solution designed to meet the demands of AI and evolve for future requirements. With the XD685, HPE and AMD provide the foundation for a comprehensive solution that can address customer's critical infrastructure challenges, combining performance, efficiency, advanced cooling, and integrated security to enable rapid scaling while maintaining competitive operational costs. AMD ROCm™ contributes to ease of adoption by enabling teams to work with standard AI frameworks and development practices, reducing friction as they bring new models and services into production.

If your organization aims to deploy a highly performant, scalable, efficient, and secure AI environment with an end-to-end optimized solution that can support its global footprint and empower you to scale for growth, Omdia recommends you evaluate and consider partners like HPE and AMD that have the experience and technology to deliver sustainable AI at worldwide scale and solutions like the HPE ProLiant Compute XD685 with AMD Instinct MI355X GPUs.

Copyright notice and disclaimer

The Omdia research, data, and information referenced herein (the “Omdia Materials”) are the copyrighted property of TechTarget, Inc. and its subsidiaries or affiliates (together “Informa TechTarget”) or its third-party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice, and Informa TechTarget does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa TechTarget and its affiliates, officers, directors, employees, agents, and third-party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa TechTarget will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.

Get in touch: www.omdia.com askananalyst@omdia.com

