



**Power large-scale AI
with purpose-built HPE
AI Servers for training,
tuning, and inferencing**

Power large-scale AI with HPE AI Servers

The rapid acceleration of generative AI, multimodal models, and agentic AI is reshaping infrastructure requirements across service providers, emerging neoclouds, sovereign AI initiatives, and advanced enterprises. Large language models and foundation models now demand massive parallelism, dense GPU acceleration, and predictable performance at scale, while also introducing new constraints around energy consumption, supply chain assurance, and data sovereignty. GPU-accelerated compute has become the foundation of modern AI innovation, but deploying and operating these environments remains complex, capital-intensive, and operationally demanding. Organizations must balance speed of innovation with long-term sustainability, security, and compliance. Power density, cooling, and operational efficiency have emerged as critical challenges, as AI data centers consume dramatically more energy per square foot than traditional facilities, and these requirements continue to grow. At the same time, organizations face pressure to move from pilot projects to full production AI environments quickly—without overinvesting in infrastructure that could become stranded or underutilized. Therefore, flexible deployment models that can scale from a single rack to thousands of nodes are often paramount for these ambitious “first AI adopters.”

Organizations must balance speed of innovation with long-term sustainability, security, and compliance. Power density, cooling, and operational efficiency have emerged as critical challenges.

For sovereign AI entities and regulated industries in particular, infrastructure decisions also carry strategic weight. The ability to control where data resides, how models are trained, and how systems are secured is as important as raw performance. These organizations require AI platforms that deliver not only leadership-class performance, but also offer a strong security foundation and transparent supply chains.



A purpose-built 8-GPU server portfolio designed for AI at industrial scale

HPE addresses these challenges with a portfolio of purpose-built servers engineered specifically for large-scale AI training, tuning, and inference, each supporting eight high-end GPUs in dense, optimized form factors. The HPE Cray XD670, HPE ProLiant Compute XD685, and HPE Compute XD690 are designed to serve different operational models and performance objectives, while sharing a common philosophy: deliver predictable AI performance at scale, enable efficient power and cooling strategies, and accelerate and simplify deployment and lifecycle management through integrated software and services.

These systems are not general purpose servers adapted for AI. They are architected from the ground up to support GPU-intensive workloads such as large language model training, natural language processing, and multimodal AI. Each platform integrates balanced CPU-to-GPU architectures, support for air or direct liquid cooling, and high bandwidth interconnects, allowing organizations to match infrastructure design to data center realities and sustainability goals.

Together, this portfolio allows customers to standardize on an 8-GPU building block while choosing the system that best aligns with their deployment environment—whether that is a sovereign AI facility, a hyperscale neocloud, or an enterprise AI factory. This modularity reduces architectural risk, accelerates time to deployment, and enables consistent operational practices across AI environments.

The HPE Cray XD670, HPE ProLiant Compute XD685, and HPE Compute XD690 are designed to serve different operational models and performance objectives.



HPE Cray XD670: Proven performance for demanding AI training workloads

The HPE Cray XD670 server is designed for organizations that require extreme AI training performance and proven scalability. Built in a dense 5U form factor, the system supports eight NVIDIA® H200 Tensor Core GPUs, paired with two 5th Generation Intel® Xeon® Scalable processors, delivering leadership performance for large scale model training and deep learning workloads. The platform is optimized for GPU intensive applications that rely on heavy parallelism, such as large language model training and advanced multimodal AI, and can scale from a single system to large clusters supporting thousands of GPUs.

HPE Cray XD670 offers optional direct liquid cooling, enabling organizations to improve power efficiency and advance sustainability goals, where supported by data center infrastructure.



8x NVIDIA H200 air or liquid cooled

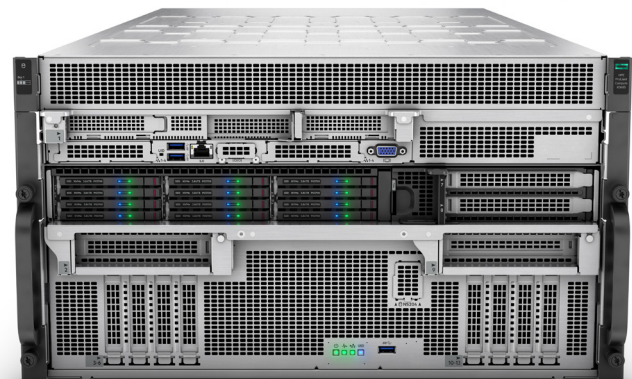
The platform has demonstrated top results in industry benchmarks, including six #1 results in the recent MLPerf™ Inference v5.1 tests in Computer vision, LLM—chat Q&A, LLM—text generation and other workloads. These leading results on a range of AI benchmarks are proof of the superior HPE Cray XD670 performance, reinforcing its role as a foundation for large scale AI environments where time-to-results is critical.

HPE ProLiant Compute XD685: Flexible, secure, and sustainable AI at scale

The HPE ProLiant Compute XD685 extends HPE's AI server portfolio with a system optimized for flexibility, sustainability, and global deployment at scale. Designed in a modular 5U or 6U chassis, depending on the cooling method, it supports eight NVIDIA or AMD Instinct™ GPUs, paired with two 5th Generation AMD EPYC™ processors, enabling customers to align accelerator choices with workload requirements and ecosystem preferences. This flexibility makes the platform well suited for organizations building diverse AI services or supporting multiple model architectures.

Optional direct liquid cooling² helps organizations manage escalating power demands while advancing sustainability objectives, particularly in high-density AI clusters. Support for HPE iLO provides built in security features, including Silicon Root of Trust, to help protect firmware integrity and reduce the risk of supply chain attacks. These capabilities are especially relevant for sovereign AI initiatives and regulated industries that require strong security assurances.

A versatile AI training and tuning platform, the HPE ProLiant Compute XD685 supports large-scale deployments across geographies.



- **8x NVIDIA H200 (air or DLC)**
- **8x NVIDIA B200 (DLC)**
- **8x NVIDIA B300 (DLC)**
- **8x AMD Instinct™ MI355X (DLC)**

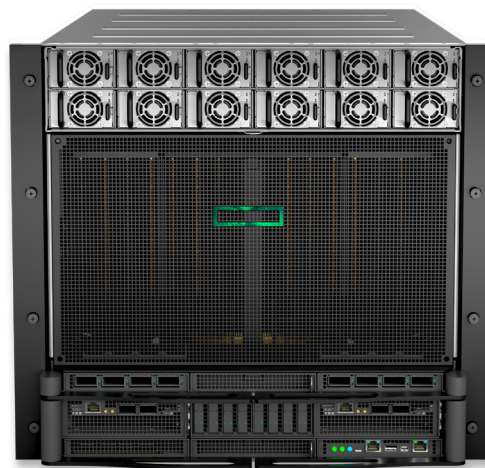
¹ ML Commons, [MLPerf Inference: Datacenter](#); HPE, [HPE delivers several world records in latest MLPerf Inference benchmarks](#), September 2025.

² Direct liquid cooling is optional for NVIDIA H200 Tensor Core, and mandatory for AMD Instinct MI355X, NVIDIA Blackwell Ultra (B300 HGX) and NVIDIA Blackwell (B200).

HPE Compute XD690: Next-generation performance for the escalating demands of AI

HPE Compute XD690 represents HPE's next-generation approach to large-scale AI infrastructure. Part of the NVIDIA AI Computing by HPE portfolio, HPE Compute XD690 supports eight NVIDIA Blackwell Ultra (B300 HGX) GPUs, delivering exceptional performance for organizations pushing the limits of model size and complexity. The system comes with two Intel Xeon 6 processors in a 10U chassis, and it is air-cooled, enabling deployment in standard data centers while still supporting high-power GPU configurations.

The NVIDIA Blackwell Ultra GPU delivers 50% higher performance in FP8 training and FP4 inferencing than its predecessor, making HPE Compute XD690 ideal to power highly demanding AI training, tuning, and inference workloads. Service providers and organizations that want to unlock the power of large scale AI without redesigning data center infrastructure can benefit from the air-cooled design of this platform.



8x NVIDIA B300 air-cooled

Purpose-built servers for large AI model training, tuning, and inferencing

Table 1. A table visualizing purpose-built servers for large AI model training, tuning, and inferencing. 8-way GPU servers for accelerated AI.

	HPE Cray XD670	HPE ProLiant Compute XD685	HPE Compute XD690
GPU	<ul style="list-style-type: none"> • 8x NVIDIA H200 air- or liquid-cooled 	<ul style="list-style-type: none"> • 8x NVIDIA H200 (air or DLC) • 8x NVIDIA B200 (DLC) • 8x NVIDIA B300 (DLC) • 8x AMD Instinct™ MI355X (DLC) 	<ul style="list-style-type: none"> • 8x NVIDIA B300 air-cooled
CPU	2x 5th Gen Intel Xeon Scalable processors	2x 5th Gen AMD EPYC™ processors	2x Intel Xeon 6 processors
Dimensions	5U	5U DLC, 6U air-cooled	10U
Management	<ul style="list-style-type: none"> • Server: BMC, Redfish APIs, 1GbE LAN • Cluster: HPCM 	<ul style="list-style-type: none"> • Server: HPE iLO 6 • Cluster: HPCM 	<ul style="list-style-type: none"> • Server: BMC, Redfish APIs, 1GbE LAN • Cluster: HPCM
Proven ecosystem of HPE Services: advisory, professional, operational, and financial services			

Services and expertise that turn any AI ambition into operational reality



A strong technology foundation is not sufficient to operationalize AI at scale.

HPE complements its 8-GPU server portfolio with a broad and proven portfolio of end-to-end services designed to accelerate deployment, reduce risk, and support long term success. HPE Services include factory integration, validation, and testing, enabling accelerated on-site deployment. This approach reduces time to value and minimizes disruption during large scale rollouts.



HPE has a proven ability to support custom AI clusters worldwide.

Backed by advanced technical support for AI environments that scale to thousands of nodes, HPE offers superior serviceability to help organizations optimize AI use and operate predictably at global scale.



HPE provides installation, startup, and lifecycle services tailored to AI clusters.

HPE draws on decades of experience deploying some of the world's largest and most complex computing environments—including the world's largest liquid-cooled supercomputers³—to support AI cluster installation, startup, and ongoing operations. For service providers and sovereign AI organizations, this expertise helps address challenges related to power, cooling, and operational readiness. HPE's collaboration with NVIDIA and AMD further strengthens this portfolio, a foundational building block for the integrated AI factory at scale and sovereign AI factory solutions, part of the HPE AI Factory portfolio.

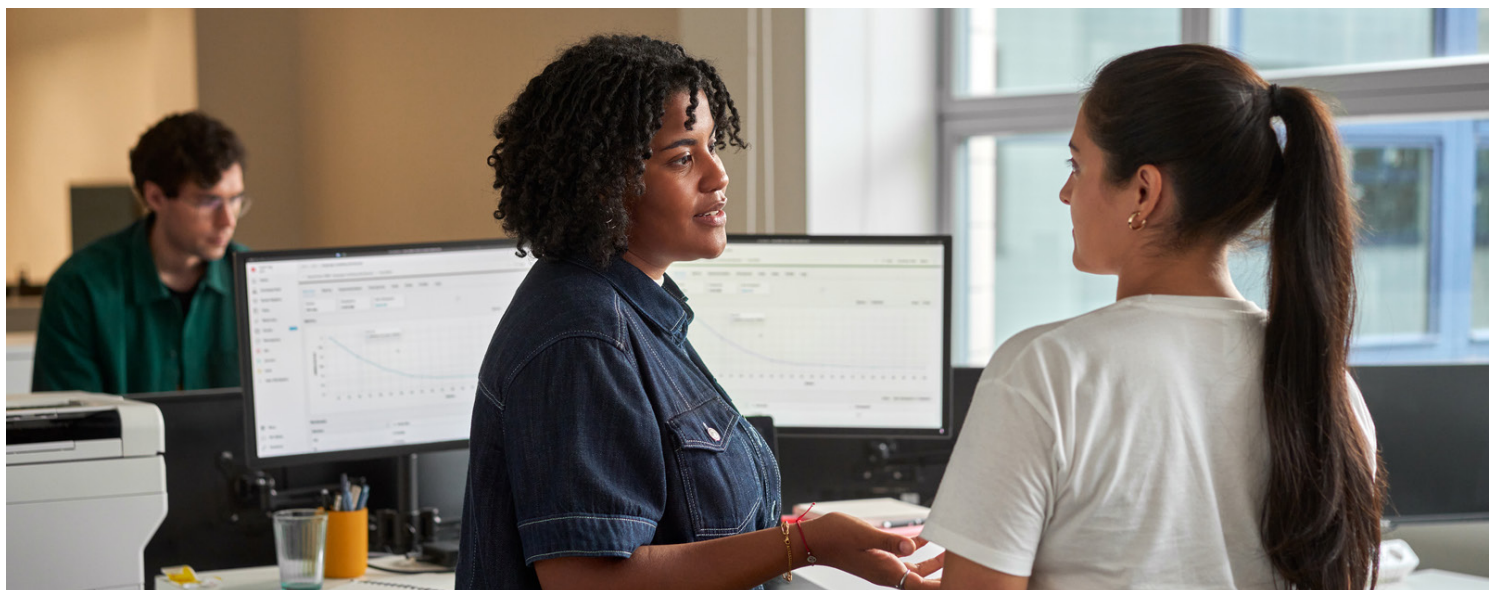
By combining infrastructure, software, and services, **HPE enables customers to focus on model innovation and business outcomes**, rather than infrastructure integration. This holistic approach is particularly valuable for organizations moving from AI experimentation to sustained, production-scale operations.



Build your AI future on a foundation designed to scale with confidence

As AI models continue to grow in size and importance, infrastructure decisions made today will shape competitiveness for years to come. The HPE portfolio of 8-GPU AI servers—HPE Cray XD670, HPE ProLiant Compute XD685, and HPE Compute XD690—offers a flexible, future-ready foundation for organizations that demand performance, sovereignty, and sustainability at scale. Backed by HPE Services and deep ecosystem partnerships, these platforms enable faster deployment, predictable scaling, and operational confidence.

Whether you are building a sovereign AI capability, establishing or growing a neocloud environment, or training next-generation models in-house, HPE provides the building blocks to move from ambition to execution. Engage with HPE to design an AI infrastructure strategy that aligns with your performance goals, operational realities, and long term vision—and turn AI potential into measurable impact.



Visit [HPE.com](https://www.hpe.com)

Learn more at

[HPE.com/ai/servers](https://www.hpe.com/ai/servers)

[Chat now](#)

© Copyright 2026 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

AMD is a trademark of Advanced Micro Devices, Inc. Intel Xeon is a trademark of Intel Corporation or its subsidiaries in the U.S. and/or other countries. NVIDIA is a trademark and/or registered trademark of NVIDIA Corporation in the U.S. and other countries. MLPERF™ is a trademark and service mark of MLCommons Association in the United States and other countries. All third-party marks are property of their respective owners.

a00156977enw

HEWLETT PACKARD ENTERPRISE

[hpe.com](https://www.hpe.com)

