Solution overview

# Achieve AI at scale with future-ready systems for modern data centers

## HPE ProLiant Compute XD685, built for the Age of AI Reasoning

Hewlett Packard Enterprise and NVIDIA have engineered a complete stack solution for the Age of AI Reasoning. HPE ProLiant Compute XD685 powered by NVIDIA® Blackwell Ultra GPUs (NVIDIA HGX B300) —a new GPU option—is an accelerated platform for scalable, secure, and high-throughput AI performance. This future-ready solution features direct liquid cooling technology, intuitive management tools, and proven services for operationalizing AI at scale.

## Unlocking the next era of AI

AI has become a cornerstone of business, science, and technology. Widespread adoption of generative AI (genAI), coupled with the rise of agentic AI, is fueling business growth and competition worldwide as organizations allocate more resources and budget to operate complex models. AI service providers and enterprises building their own AI models need technology environments that are engineered for the demands of AI, delivering exceptional capacity and performance to meet customer requirements and accelerate innovation. Those that invest in scalable, secure infrastructure will be best prepared for the future of AI.

GPU-accelerated compute has emerged as a critical enabler of AI innovation. Despite the benefits, these robust technologies can be difficult to deploy and manage. Many organizations face supply chain disruptions which can complicate project plans, while others lack in-house expertise to integrate new technologies into existing environments. Furthermore, AI data centers require dramatically more energy per square foot than traditional data centers—and their power demands are growing exponentially. Deloitte projects AI data center power to increase 30x by 2035.[1] Overinvesting in data center infrastructure risks stranding assets, while underinvesting means falling behind. Organizations must develop a strategy that makes AI not only attainable today but sustainable for years to come.

HPE and NVIDIA are collaborating to deliver AI factory solutions to help accelerate the adoption of AI. NVIDIA offers optimized performance with advanced AI technologies, while HPE connects these technologies into the core of the business.

## The race for competitive advantage with AI

NVIDIA AI Computing by HPE is leading the next wave of AI innovation. HPE and NVIDIA have decades of combined experience supporting AI service providers and enterprises. Our strategic partnership equips these organizations for AI success by delivering flexible infrastructure in conjunction with a wide set of AI capabilities for developing, training, and tuning AI models. The NVIDIA AI Computing by HPE portfolio enables AI factories for all use cases. Based on NVIDIA Enterprise AI Factory validated design, these offerings include cutting-edge HPE systems with NVIDIA accelerated compute, networking, and software as well as expert services.

HPE ProLiant Compute XD685 servers powered by NVIDIA HGX B300 GPUs balance performance, agility, and efficiency to operationalize AI at scale. These next gen technologies are part of a complete stack engineered solution, combined with AI services that help to integrate into existing infrastructure. As a result, organizations can accelerate time to deployment for new technologies and utilize AI with confidence.

[1] Deloitte AI Infrastructure Survey 2025 deloitte.com/us/en/insights/industry/power-and-utilities/data-center-infrastructure-artificial-intelligence.html#endnote-20

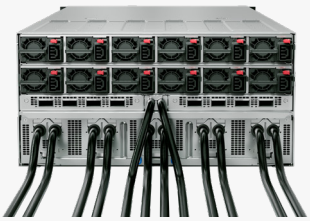
# Comprehensive AI infrastructure, services, and security for the Age of AI Reasoning

## Speeding time to solution

AI service providers and enterprises building their own AI models need to quickly stand up high performance AI clusters that utilize the latest GPU technology, to service the needs of their growing customer base, and stay competitive in a dynamic market.

The NVIDIA AI Computing by HPE portfolio offers co-developed AI factory solutions, enabling organizations to scale AI-driven capabilities with confidence. HPE ProLiant Compute XD685 is an essential offering within this portfolio. With this cutting-edge system, organizations can have the choice of 8x NVIDIA H200 GPUs (air or liquid-cooled), NVIDIA B200 GPUs (liquid-cooled), or new NVIDIA B300 (liquid-cooled) GPUs. Liquid-cooled systems feature a compact arrangement of 8 nodes per rack for maximum rack density in an 8-way GPU system. To achieve peak AI acceleration, NVIDIA HGX B300 GPUs deliver 4x training performance and 11x inference performance over the previous generation, all in a new form factor designed to fit seamlessly into modern data centers.[2] NVIDIA accelerated computing platforms with robust NVIDIA GPUs and embedded Connect-X SuperNICs drive unprecedented high throughput for

**HPE ProLiant Compute XD685 with NVIDIA GPUs**



Choice of 8x GPUs:

— NVIDIA H200 GPUs (air or liquid-cooled)

— NVIDIA B200 GPUs (liquid-cooled)

— NVIDIA B300 GPUs (liquid-cooled)

5U direct liquid-cooled or 6U air-cooled

2x 5th Generation AMD EPYC™ Processors

HPE Integrated Lights-Out

HPE Performance Cluster Manager

NVIDIA AI Enterprise

HPE Services

[2] Accelerating AI for Every Enterprise, NVIDIA 2025 nvidia.com/en-us/data-center/dgx-b300/

AI data centers. Designed for network-intensive, massively parallel computing, Connect-X SuperNICs deliver powerful networking capabilities for increased GPU-to-GPU communication, which ensures high-bandwidth data transfers and enhances AI performance.

The benefits of NVIDIA HGX B300 GPUs are significant:

— 11x faster AI performance for inference on LLMs[3]

— 4x larger memory[3]

— 7x more compute[3]

Purpose-built solutions from HPE and NVIDIA are bringing organizations the agility and capacity they need to focus less on preparing for AI and more on core business activities.

## Optimizing AI infrastructure

As genAI and agentic AI require more power, data centers are increasingly focused on reliable, energy-efficient technologies. AI service providers and enterprises building their own AI models are reevaluating their infrastructure design in order to create sustainable AI data centers.

HPE ProLiant Compute X685 provides highly scalable, secure, and efficient infrastructure designed specifically for AI workloads. These environments are optimized for AI performance and simplify cluster management through integrated tools for system monitoring, security, and lifecycle operations.

Direct Liquid Cooling (DLC) from HPE is a key capability amid the AI boom, which dramatically improves power efficiency and lowers total cost of ownership. Cooling can account for 40% of data center electricity usage, especially in AI data centers housing heat-intensive GPUs.[1] DLC technology reduces power demand, supports denser rack configurations, and minimizes physical footprint while maintaining peak levels of performance. HPE ProLiant Compute XD685 with NVIDIA B300 GPUs is direct liquid-cooled. With more than 500 patents and the world's largest liquid-cooled environments, HPE is a leader in deploying DLC at scale.

For workload deployment and AI cluster scaling, HPE Performance Cluster Manager (HPCM) delivers comprehensive software and hardware management tools to streamline provisioning, support GPU and liquid-cooled systems, and provide extensive resource monitoring and performance optimization. HPCM shortens initial cluster setup and software deployment to empower complex AI environments.

Organizations can securely set up and manage infrastructure from anywhere with the help of HPE Integrated Lights-Out (iLO) and the Silicon Root of Trust, a multi-layered security strategy. HPE ProLiant Compute XD685 features HPE iLO 6, which provides dashboards and real time insights to enhance single server management. Organizations can protect against attacks and prepare for the future with built-in security and lifecycle protection across their entire AI infrastructure, from assembly in highly secure facilities to end-of-life decommissioning.

NVIDIA AI Enterprise is a cloud-native suite of software tools, libraries, and frameworks that streamline the development, deployment, and scaling of AI applications. Organizations of all sizes can deploy AI systems anywhere (across clouds, data centers, or at the edge) leveraging the extensive partner ecosystem. The software is designed to accelerate time to market and reduce infrastructure costs while ensuring fast. reliable, secure AI operations.

## Powering AI globally

When expanding the use of AI and offering consistent AI services, organizations must deliver to and provide support in their own geographies. As operations scale globally, they need technology partners with a strong

[3] NVIDIA Blackwell Ultra, NVIDIA 2025 nvidianews.nvidia.com/news/nvidia-blackwell-ultra-ai-factory-platform-paves-way-for-age-of-ai-reasoning

worldwide presence. To manage these large, complex environments effectively, financial and deployment flexibility are essential to drive efficiency while enabling rapid response to customer needs.

HPE offers technical and advisory support to empower AI environments around the globe. HPE Services experts collaborate at each stage of innovation—from planning, to operating, and beyond—to help organizations optimize solution deployments and maximize IT investments. Every solution is protected by HPE Trusted Supply Chain services, which provide a new first line of defense against cyberthreats with HPE ProLiant Compute XD685 servers built to the world's toughest security standards in secured facilities. This approach brings together enterprise-grade security, tools, and experts to deliver massive custom AI clusters anywhere in the world with superior serviceability.

HPE experts include the AI performance engineering team, holder of a multitude of MLPerf™ benchmark world records. They can assist organizations with AI workload sizing and individual application tuning with the goal of optimizing performance throughout the life of the systems.

To ease transformation, HPE offers specialized financing and management services for AI projects at scale. Organizations benefit from simpler, faster, and cost-effective access to genAI and agentic AI. HPE Accelerated Migration Services allow organizations to convert aging infrastructure into an incremental cash source while converting use of AI systems to an HPE GreenLake as-a-service model. To recover value from any HPE or third-party data center solutions, HPE Asset Upcycling is available to reuse or recycle legacy technologies.

These collective offerings give organizations peace of mind to operationalize AI in a secure, globally consistent, and environmentally responsible way.

## Preparing for the future of AI

HPE and NVIDIA are trusted partners for empowering the AI journey. This alliance is built on the development of cutting-edge solutions that redefine compute, unleash massive GPU acceleration, and push the boundaries of AI at scale. Discover what is possible for your business. Explore HPE ProLiant Compute XD685 today, or learn more about NVIDIA AI Computing by HPE.

## Learn more at

hpe.com/us/en/compute/proliant-compute-xd685.html

Visit HPE.com

Chat now

© Copyright 2025 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for 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. NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks 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.

a00154393ENW

HEWLETT PACKARD ENTERPRISE

hpe.com