Hewlett Packard Enterprise | NVIDIA

HPE ProLiant Gen11 AI optimized solutions

Supporting AI workloads from edge to data center to cloud

Capitalizing on the AI revolution

Based on the groundwork completed in the mid-1950s, artificial intelligence (AI) has graduated from "pioneering science fiction" to mainstream technology used by an ever-growing number of companies across industries. As AI workloads such as machine learning (ML) and deep learning (DL) are extremely compute-hungry, they can be handled by only high-performance, high-density servers utilizing both accelerators and advanced graphical processing units (GPUs). Today, the demand for AI-capable servers is growing at an exponential rate.

To answer your call for powerful servers that can meet your AI requirements, Hewlett Packard Enterprise and NVIDIA® leverage industry-leading server innovation and world-leading AI expertise to deliver HPE ProLiant Gen11 servers with next-generation NVIDIA GPUs.

These performance-intensive, industry-standard servers offer the scalability, efficiency, and performance you need to accelerate business innovation and capitalize on the AI revolution.

> By 2025, 50% of enterprises will have devised artificial intelligence (AI) orchestration platforms to operationalize AI, up from fewer than 10% in 2020."[1]

[1] "Our Top Data and Analytics Predicts for 2021," Gartner blog, January 12, 2021.

HPE ProLiant Gen11 servers — Powering AI today and tomorrow

Why HPE ProLiant Gen11 for AI?

Innovative HPE ProLiant Gen11 servers deliver advanced engineered solutions to resolve today's hybrid cloud infrastructure challenges. They combine the best of on-premises and cloud computing with:

- An **intuitive** cloud operating experience
- HPE **trusted** security by design
- **Optimized** performance for large, complex AI workloads

What's new

The new HPE ProLiant Gen11 server portfolio takes a fresh approach to GPU support — in some cases, moving the GPU to the front of the chassis to improve airflow, increase GPU density, and provide a dedicated power supply to each GPU to improve GPU uptime. This way, HPE ProLiant Gen11 servers can effectively match the requirements for a variety of AI-specific workloads by powering high-wattage GPUs.

The front-end cage design enables select NVIDIA GPUs to communicate directly with each other. This communication allows the available memory of the GPUs to be combined, which in turn increases the speed of data exchange. The result — significant performance increases and the ability to process AI models with 100 billion plus parameters.

In addition, HPE servers are equipped with either 4th Gen Intel® Xeon® Scalable processors or 4th Generation AMD EPYC™ 9004 Series processors. These servers offer a wide range of features for optimizing power and performance and maximizing CPU resources to help your organization achieve its sustainability goals.

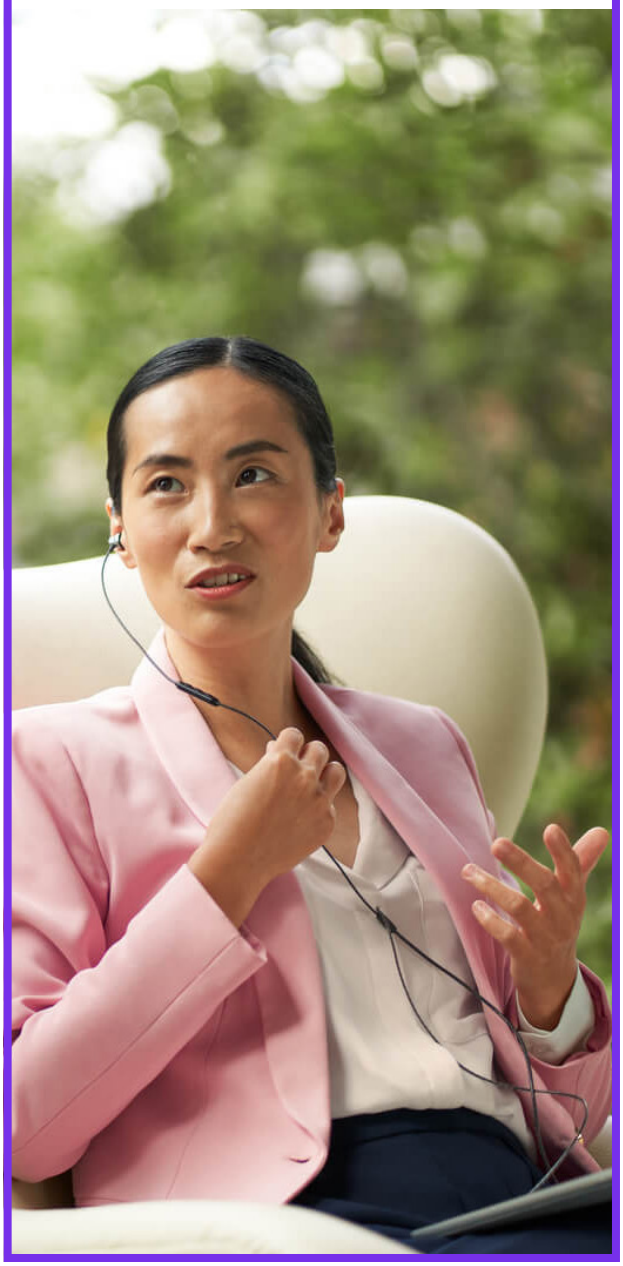
HPE iLO 6 is built into every HPE ProLiant server. HPE iLO 6 enables you to securely configure, monitor, and remotely manage HPE ProLiant servers — from anywhere in the world — and provides valuable asset information about NVIDIA GPUs, such as device inventory, temperature reporting, power management, firmware reporting, health status, and event logs.

Management has been transformed for next-generation HPE ProLiant Gen11 servers. Leveraging the HPE GreenLake edge-to-cloud platform — including its built-in management application and architecture — the intuitive cloud operating experience streamlines and secures operations from edge to cloud. By automating key lifecycle tasks for onboarding, updating, managing, and monitoring, HPE servers bring agility and efficiencies to wherever compute devices reside via a unified single browser-based interface.

Data-first GPU modernization

HPE ProLiant Gen11 servers with NVIDIA next-generation GPUs power today's AI, ML, and DL workloads.

- **NVIDIA L40S Tensor Core GPU:** Multi-workload performance for powerful AI compute with best-in-class graphics and media acceleration, the L40S GPU is built to power the next generation of data center workloads.
- **NVIDIA L4 Tensor Core GPU:** Cost-effective, low-profile, low-power, single-wide form factor for high throughput and low latency in any server.
- **NVIDIA L40 GPU:** Double-wide accelerator designed for visual computing and graphics-intense data processing.



Get to know HPE ProLiant Gen11 servers for AI

Table 1. HPE ProLiant Gen11 servers for AI at a glance

			
	HPE ProLiant DL320 Gen11 Server	**HPE ProLiant DL380a Gen11 Server**	**HPE ProLiant DL385 Gen11 Server**
Form factor	1U, single socket	2U, dual socket	2U, dual socket
Processor	4th Gen Intel Xeon Scalable processors	4th Gen Intel Xeon Scalable processors	4th Generation AMD EPYC 9004 Series processors
Memory	2 TB DDR5	3 TB DDR5	6 TB DDR5
Storage	10 SFF SSDs	8 EDSFF E3.S 1T NVMe SSDs	36 EDSFF E3.S 1T NVMe SSDs
GPUs	Up to 2 double-wide or 4 single-wide GPUs	Up to 4 double-wide or 8 single-wide GPUs	Up to 4 double-wide or 8 single-wide GPUs
Connectivity	ConnectX-7 family of networking	ConnectX-7 family of networking	ConnectX-7 family of networking
Management	HPE GreenLake for Compute Operations Management (subscription)		

The perfect pairings — at a glance

With so many server and GPU options available in today's data-first world, data center managers are constantly searching for the next best way to modernize their IT infrastructure. Managers must consider the complexity of sophisticated AI algorithms and other data-intensive computing demands operating at the edge and in the cloud. The decision-making process becomes more nuanced when trying to match workload requirements with a server/GPU solution. Together, HPE and NVIDIA can help you navigate your AI journey, wherever it leads, with recommended product pairings.

Taking your decision-making to the next level, HPE systems that are NVIDIA-Certified offer you a reference design for building scalable units of accelerated computing. The NVIDIA certification ensures optimum performance, reliability, and scale for a diverse range of AI workloads.

NVIDIA-Certified systems

You can purchase an AI solution from HPE with confidence, knowing that every system has been certified by HPE and provides:

- Reference design guides for building accelerated systems
- NVIDIA networking to ensure super-fast data transfers from storage to GPU computing
- The ability for every system to be managed no matter where it is located — at the edge, in colocations, or in data centers — using HPE GreenLake for Compute Ops Management

Matching technology with purpose

The following examples of paired technology solutions can help you choose the best server and GPU combination for a particular AI-related business purpose/workload. The highlights below provide deeper insight into each pairing's features and functions.

Table 2. HPE and NVIDIA paired technology solutions

Combination	Purpose/Workload	Highlights
Computer vision AI at the edge HPE ProLiant DL320 Gen11 Server NVIDIA L4 GPU	• Computer vision & Intelligent Video Analytics • Loss prevention • Smart spaces	• Real-time computer vision inference ideal for the edge • Compact 1U form factor of HPE ProLiant DL320 • Up to 4 NVIDIA L4 GPUs • Enabled by NVIDIA Metropolis • Perfect for retail, hospitality, and manufacturing
Generative visual AI HPE ProLiant DL380a Gen11 Server NVIDIA L40 GPU	• 3D animation and rendering • Video content creation • Image generation	• HPE ProLiant DL380a supports 4 NVIDIA L40 GPUs • NVIDIA AI Enterprise enables market-leading software • Enhances AI video performance, enabling significant gains to personalize content • Ideal for Media & Entertainment, healthcare, and manufacturing
Natural language processing AI HPE ProLiant DL380a Gen11 Server NVIDIA L40S GPU	• NLP • Speech AI • Fraud detection • Predictive maintenance	• AI-optimized HPE ProLiant DL380a with 4 L40S GPUs • Designed to power advanced NLP for use-cases like fraud detection and speech AI • Develop and deploy fine-tuned models • Ideal for FSI, manufacturing, customer service

Note: This is by no means a complete list. Contact your HPE or NVIDIA representative to discuss your specific AI requirements.

Accelerated software from NVIDIA

NVIDIA AI Enterprise software accelerates the data science pipeline and streamlines development of production-level AI including generative AI, computer vision, speech AI, and more.

With more than 100 frameworks, pretrained models, and development tools, NVIDIA AI Enterprise is designed to accelerate your enterprise to the leading edge of AI while also simplifying AI to make it accessible to every enterprise. When combined with HPE systems that are NVIDIA-Certified, NVIDIA AI Enterprise ensures you get the right level of performance, scalability, and enterprise-grade support.

Prepare for AI success

Decision-making considerations

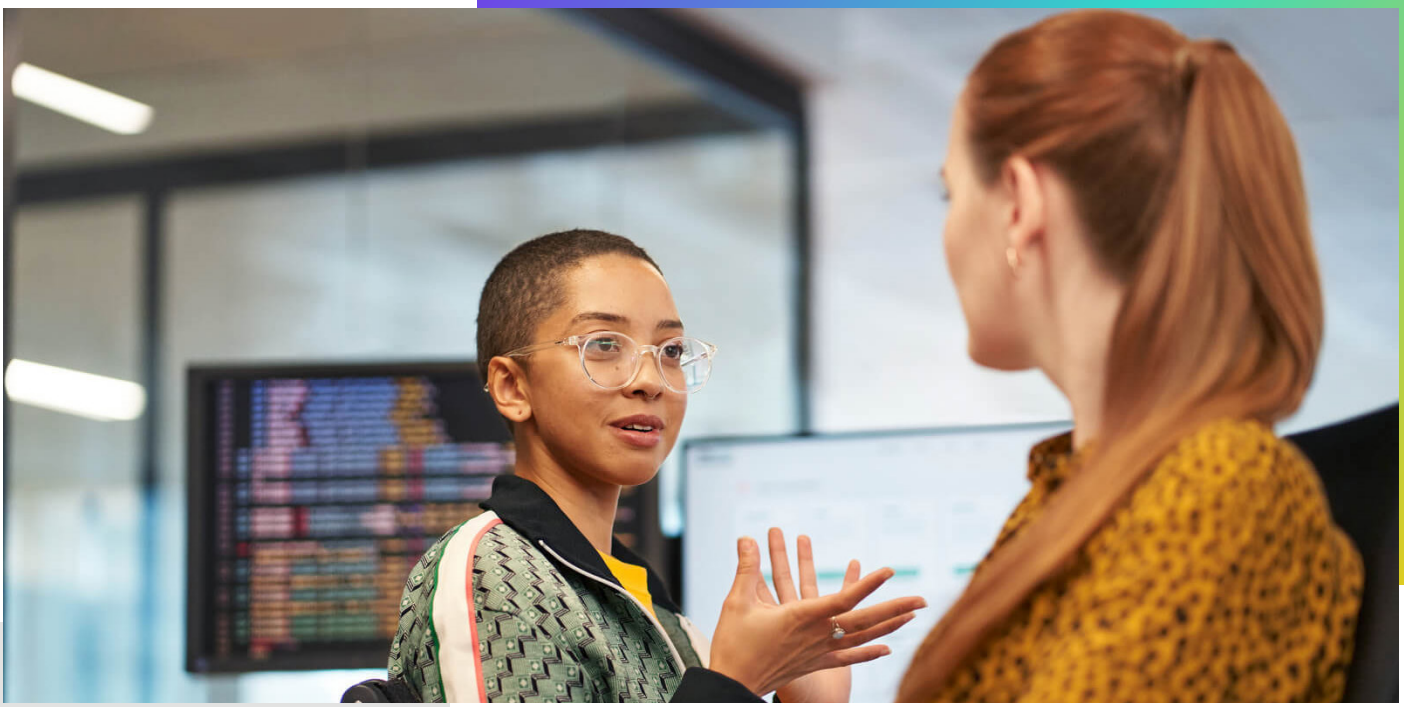
Joint solutions from HPE and NVIDIA create a pathway to AI success — enabling you to derive the most value from your IT modernization projects. Regardless of the individual HPE server, NVIDIA GPU, or paired combination, every AI infrastructure decision should consider the following factors:

- **Scalability:** Choose easy-to-upgrade servers that allow you to add or remove GPUs as your AI processing needs change. Deploy a scalable solution that meets your needs from edge to data center to cloud.
- **Performance:** Select a solution capable of running large, complex models and tackling increasingly challenging AI tasks efficiently.
- **Connectivity:** Implement a solution that offers a high-bandwidth connection for smooth data transfer between GPUs to support parallel processing and reduce the time required to train AI models using multiple GPUs.
- **Management:** Included with every HPE ProLiant Gen11 server is HPE GreenLake for Compute Ops Management — designed to simplify and automate operations across the server lifecycle, no matter where your compute infrastructure lives.

Leveraging HPE GreenLake cloud services

HPE GreenLake platform delivers a trusted, enterprise-grade cloud experience to accelerate data modernization initiatives, on-premises, in your data center, or at a colocation. Take control of your data while gaining insights and optimizing operations by scaling as needed with continuous monitoring and a flexible architecture that enables right-sized capacity bursting on demand.

To learn more about HPE GreenLake, visit hpe.com/us/en/greenlake.html.



Reimagining the future

Video analytics at the edge, large language models (LLMs), NLP, AI-driven recommendations, and instantaneous searches — including ChatGPT and Microsoft Bing — are already a reality in our digital ecosystem. On the horizon, you can see a future where self-driving vehicles, robots, smart cities, and smart homes will become commonplace occurrences, thanks to breakthrough AI innovations — and we're only just getting started.

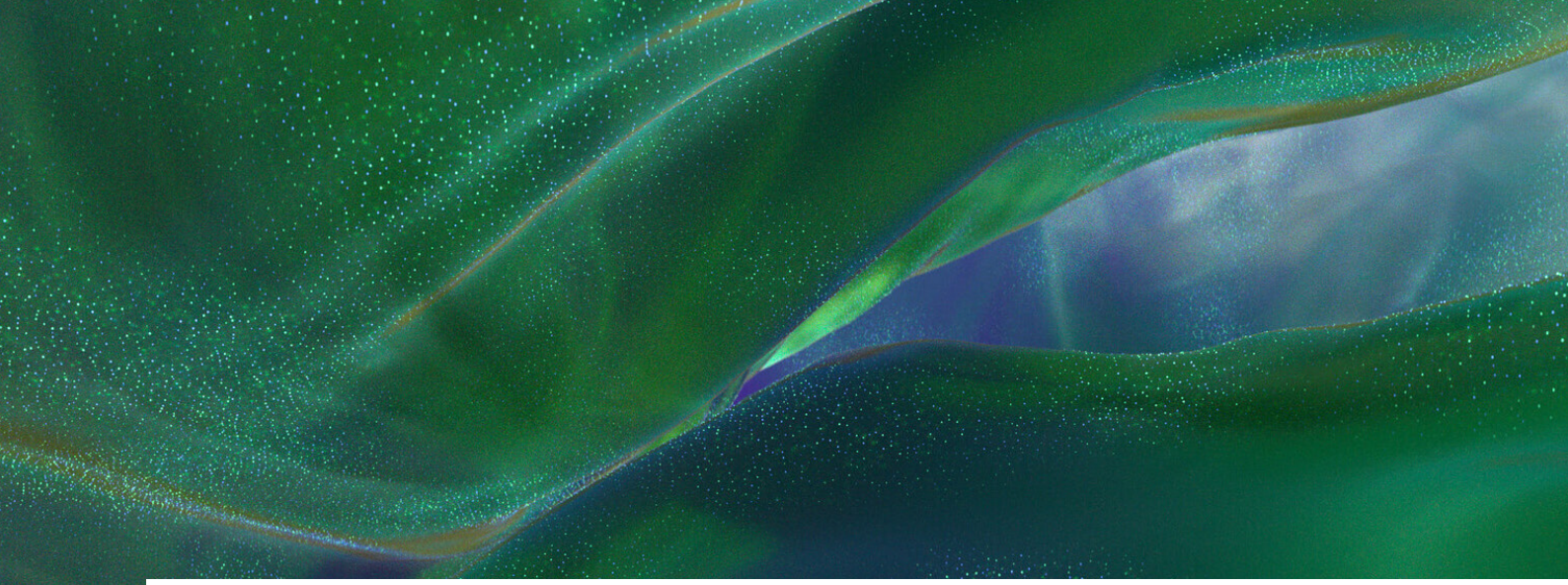
AI advancements drive change

Use case examples

Today, a growing number of industries use AI and other emerging technologies such as DL and ML to answer questions and reveal insights that were considered unsolvable or impossible just a few years ago. Increasingly, AI is the technology engine that drives advancements that improve our everyday lives — at work, at home, and at play.

To explore some of the areas where HPE and NVIDIA are making a significant impact on the world of AI, please visit the following webpages:

- Clinical research
- Financial services
- Healthcare and life sciences
- Manufacturing
- Sports stadiums



Advance beyond the competition with HPE and NVIDIA

HPE ProLiant Gen11 servers and NVIDIA next-generation GPUs stand at the forefront of AI technology — delivering the right solutions at the right time to drive the wheels of progress forward. Through continued research and collaboration, our compute products and solutions will continue to evolve in lockstep with AI and whatever comes next.

Now is the time to future-proof your IT infrastructure with industry-leading, NVIDIA qualified HPE ProLiant Gen11 servers. These innovative systems can help your organization capitalize on the AI revolution by quickening development and deployment of AI models and high-performance data analytics. No matter how complex or demanding your AI workloads are, you can find the right-fit server/GPU solution from HPE and NVIDIA.

Learn more at the NVIDIA collaboration page on hpe.com, where you can find out how to:

- **Tap into the right expertise.** For custom NVIDIA-based AI and edge deployments, work with experienced advisory and professional services teams to leverage the HPE Cloud Adoption Framework.
- **Optimize IT.** Quickly spin up containerized AI and ML environments and offload management of your data science ecosystem so that you can deploy resources and capacity to reach better business outcomes.
- **Achieve faster time to value.** Streamline AI workloads across business units to improve performance — helping data scientists, developers, IT teams, and researchers spend more time building solutions, gathering insights, and accelerating time to value.
- **Pay only for what you need.** Stay ahead of the variability of AI and ML workloads and reduce infrastructure costs with active capacity management based on a pay-per-use* model.

*May be subject to minimums or reserve capacity may apply

Learn more at

HPE ProLiant Servers

HPE ProLiant Servers for AI

Chat now (sales)

a50008742ENW, Rev. 1

Hewlett Packard Enterprise