HPE ProLiant Compute

Architecture guide

Unleash accelerated computing with HPE ProLiant Compute and NVIDIA GPUs

Hewlett Packard Enterprise understands that every business is unique with varying workloads. That's why we are committed to providing customers with tailored solutions for your specific business needs. Discover how HPE ProLiant Compute servers and NVIDIA can power any workload with scale, economics, and the security your business demands.

HPE

NVIDIA GPU offerings for every workload

HPE's portfolio includes a range of NVIDIA® GPUs designed to power workloads, from AI and deep learning to professional visualization. Our offerings span from the entry-level L4 to the high performance RTX PRO™ 6000 Blackwell, designed for AI and machine learning, 3D graphics, and scientific simulation. With this comprehensive lineup, we help ensure businesses and developers have the right solution to meet their performance and scalability needs.

NVIDIA L4 24GB

The L4 provides universal acceleration and energy efficiency, making it ideal for AI video applications, virtual workstations, and graphics workloads. Its advanced architecture enhances video processing, ensuring smooth streaming and high-quality rendering.

NVIDIA RTX 4000 Ada

The RTX 4000 Ada is ideal for professional visualization tasks, including 3D modeling, rendering, and simulation in design and engineering. Its capabilities also support content creation workflows, such as video editing and graphic design, providing accelerated performance for creative professionals.

NVIDIA RTX A1000

Built on the NVIDIA Ampere architecture, the NVIDIA RTX A1000 is a compact, low-profile GPU, which is ideal for professionals needing entry-level accelerators to balance performance with power efficiency or requiring the power of GPUs in small spaces. With its real-time ray tracing, AI acceleration, and high-fidelity graphics capabilities, the RTX A1000 is ideal for design, visualization, and digital content creation.

NVIDIA A16 64GB

Designed to accelerate virtual desktop infrastructure (VDI) workloads, the A16 provides high user density and enhanced graphics performance for remote desktop applications.

NVIDIA L40S 48GB

Experience breakthrough multi-workload performance with the NVIDIA L40S GPU. Combining 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—from generative AI and large language model (LLM) inference and training to 3D graphics, rendering, and video.

NVIDIA H100 NVL 94GB

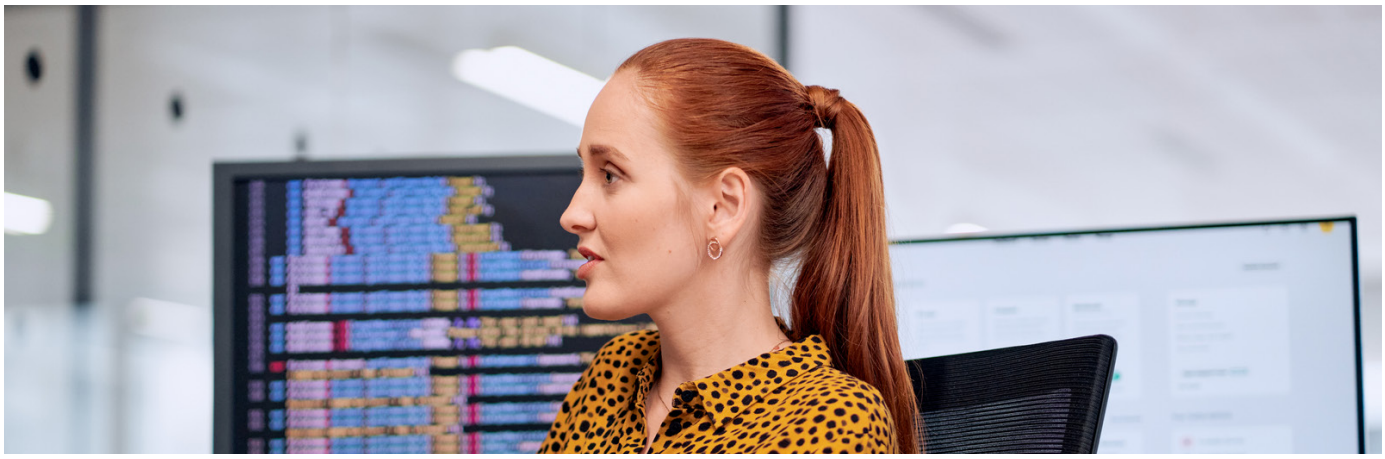
The H100 NVL, optimized for LLM inferencing, supports a broad range of math precisions, providing a single accelerator for many compute workloads. With a PCIe card memory bandwidth of nearly 4,000 GBps, the H100 NVL enables faster processing of large models and massive data sets. In addition, the NVIDIA H100 NVL card features Multi-Instance GPU (MIG), which allows partitioning of the GPU into hardware isolated GPU instances.

NVIDIA H200 NVL 141GB

With HBM3E memory and enhanced bandwidth, the H200 NVL enables faster inference and fine-tuning of large language models (LLM). This GPU is also well suited for applications such as seismic analysis, computational fluid dynamics, climate modeling, computer vision, speech AI, and retrieval-augmented generation (RAG).

NVIDIA RTX PRO 6000 Blackwell Server Edition

Built on the groundbreaking NVIDIA Blackwell architecture, the NVIDIA RTX PRO 6000 Blackwell Server Edition combines advanced AI and visual computing capabilities to accelerate enterprise data center workloads. Equipped with 96 GB of ultra-fast GDDR7 memory and FP4 capabilities, the NVIDIA RTX PRO 6000 Blackwell provides unparalleled performance and flexibility to accelerate a broad range of use cases including graphics, visual computing, industrial and physical AI, Enterprise HPC, VDI, and Enterprise AI applications.



Finding the right HPE ProLiant Compute server for your NVIDIA GPU

How do you traverse this rich ecosystem of servers and GPU options? The following tables are designed to make this process easy by showing various NVIDIA GPU compatibilities with each HPE ProLiant Compute server.

Table 1. Quantity of NVIDIA GPUs supported per HPE ProLiant Compute tower server

		L4 24GB	RTX 4000 Ada 20GB	RTX A1000 8GB
		SW	SW	SW
Gen11	HPE ProLiant ML30		1	1
	HPE ProLiant ML110	2	2	
Gen12	HPE ProLiant Compute ML350	4		8

Table 2A. Quantity of NVIDIA GPUs supported per HPE ProLiant Gen11 rack server with Intel® processors

			L4 24GB	RTX A1000 8GB	L40S 48GB	H100 NVL 94GB
		Form factor	SW	SW	DW	DW
Gen11	HPE ProLiant DL20	1U		1		
	HPE ProLiant DL380a	2U	8		4	4

Table 2B. Quantity of NVIDIA GPUs supported per HPE ProLiant Compute Gen12 rack server with Intel processors

			L4 24GB	RTX A1000 8GB	L40S 48GB	H100 NVL 94GB	H200 NVL 141GB	RTX PRO 6000 96GB
		Form factor	SW	SW	DW	DW	DW	DW
Gen12	HPE ProLiant Compute DL320	1U	4		2			
	HPE ProLiant Compute DL340	2U	6		4			
	HPE ProLiant Compute DL360	1U	3					
	HPE ProLiant Compute DL380	2U	8	8	3	3		3
	HPE ProLiant Compute DL380a	4U	16		10	10	10	8

SW: Single-wide DW: Double-wide

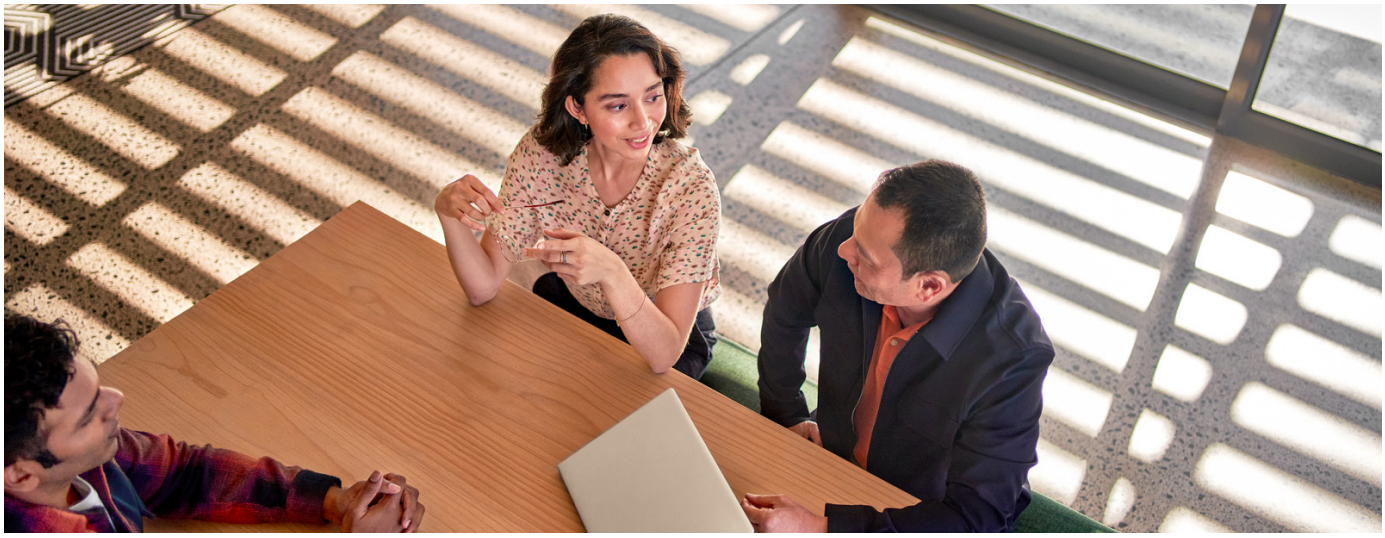


Table 3A. Quantity of NVIDIA GPUs supported per 1U HPE ProLiant Compute rack server with AMD processors

			L4 24GB	A16 64GB
		Form factor	SW	DW
Gen11	HPE ProLiant DL365	1U	3	2
Gen12	HPE ProLiant DL325	1U	4	

Table 3B. Quantity of NVIDIA GPUs supported per 2U HPE ProLiant Compute rack server with AMD processors

			L4 24GB	A16 64GB	L40 48GB	L40S 48GB	H100 NVL 94GB	RTX PRO 6000 96GB
		Form factor	SW	DW	DW	DW	DW	DW
Gen11	HPE ProLiant DL145	2U	3			1		
	HPE ProLiant DL385	2U	8	4	4	4	4	2
Gen12	HPE ProLiant DL345	2U	6			4		

SW: Single-wide DW: Double-wide

Supercharged performance: Unleashing GPU power with NVIDIA NVLink

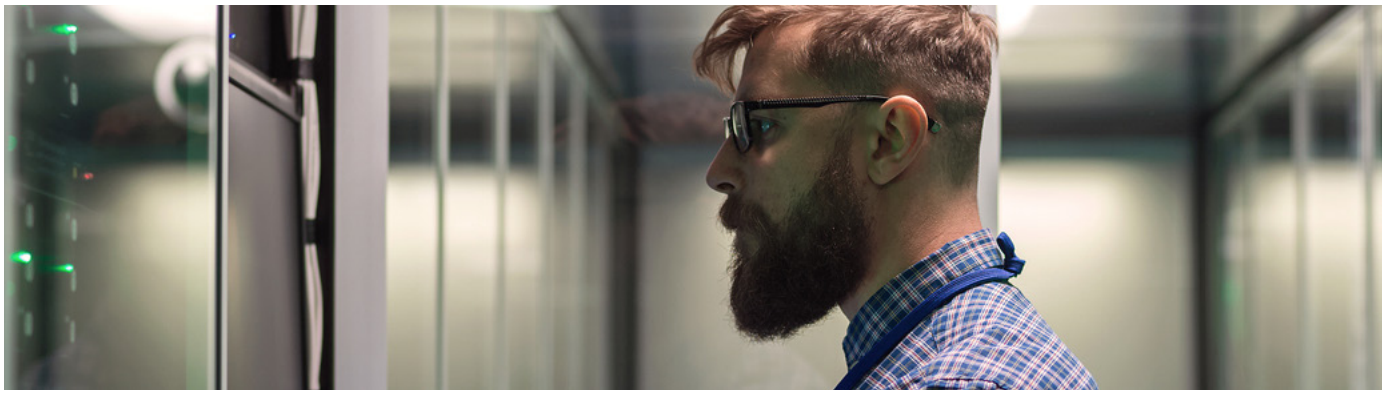
NVLink is a high-speed interconnect technology that enables faster, more efficient communication between NVIDIA GPUs, unleashing performance that would be impossible with traditional interconnect technologies. It is a cornerstone for HPC, AI workloads, and large-scale data processing for the following reasons:

1. **High bandwidth communication:** NVLink enables much faster data transfer rates between NVIDIA GPUs compared to traditional PCIe connections. For example, NVLink can provide bandwidths of up to 900 GB/s, compared to PCIe Gen4's 64 GB/s or Gen5's 128 GB/s. This high bandwidth significantly reduces the bottleneck when transferring large datasets between NVIDIA GPUs, which is essential for applications such as deep learning and HPC.
2. **Efficient GPU-to-GPU communication:** NVLink allows NVIDIA GPUs to communicate directly with each other without going through the CPU. This direct communication leads to lower latency and increased efficiency in multi-GPU setups. This is particularly beneficial for workloads such as AI model training, where multiple GPUs need to share large amounts of data in real time.
3. **Unified memory space:** NVLink offers significant benefits for multi-GPU systems by enabling faster GPU-to-GPU communication and a unified memory pool, leading to improved performance, especially in data-intensive tasks like AI and scientific simulations. Specifically, it provides higher bandwidth and lower latency compared to traditional PCIe connections, allowing for more efficient data sharing and synchronization between GPUs.
4. **Optimized for AI and HPC:** Many AI and HPC frameworks, such as TensorFlow™, PyTorch, and CUDA, are optimized to take advantage of NVLink, allowing developers to unleash its full potential.

HPE ProLiant Compute DL380a Gen12

HPE supports NVIDIA NVLink, which enhances GPU-to-GPU communication for improved performance in AI, HPC, and visualization workloads. The 2-way NVLink, connects two GPUs for increased memory pooling and data transfer speeds, and 4-way NVLink, links four GPUs for even greater scalability and parallel processing power.

- Both the **HPE ProLiant DL380a Gen11 and DL385 Gen11 Servers** support 2-way NVLink when using NVIDIA H100 NVL GPUs
- The **HPE ProLiant DL385 Gen11 Server** supports the 2-way NVLink when using NVIDIA H200 NVL GPUs
- The **HPE ProLiant Compute DL380a Gen12 Server** supports both 2-way and 4-way NVLink with NVIDIA H200 NVL GPUs



Accelerate your AI Journey

Choosing a server and GPU combination for your business can be a daunting process, but it doesn't have to be. HPE and NVIDIA understand the importance of keeping you informed in your compute decisions. We remain committed to delivering codeveloped enterprise AI solutions and joint go-to-market integrations to assist businesses in streamlining their deployments to accelerate AI applications.

Learn more at

How HPE and NVIDIA unlock AI

HPE ProLiant Compute DL380a Gen12

HPE ProLiant DL385 Gen11

NVIDIA Accelerators for HPE

HPE ProLiant Compute

Visit HPE.com

Chat now

a00150148ENW, Rev. 2

HEWLETT PACKARD ENTERPRISE

hpe.com