Solution overview

Accelerate your HPC and AI initiatives with NVIDIA GB200 NVL4 by HPE

The next gen rack-scale solution for converged HPC and AI

Hewlett Packard Enterprise and NVIDIA are delivering a next gen rack-scale solution designed for converged HPC and AI workloads. NVIDIA GB200 NVL4 by HPE combines GPU acceleration and double-precision performance in a dense form factor, enabling seamless scaling for scientific simulations and AI model training. With advanced liquid cooling and simplified management, this offering supports rapid deployment and expansion for research labs, universities, and sovereign entities wanting to operate in Arm®-based environments.

Unlocking value with HPC and AI convergence

As high performance computing (HPC) and artificial intelligence (AI) continue to advance, powerful new infrastructures are needed to unleash insights, knowledge, and potential. Leveraging the latest technologies is critical to stay competitive, particularly as data volumes rise and workloads grow larger and more complex. Traditional HPC users are looking to develop their codebases in efficient Arm environments as well as experiment with AI workloads and AI factories to augment HPC. Bringing AI into the core of their operations is a strategic shift to fuel breakthroughs in science, business, and technology.

74.9% of organizations plan to moderately or significantly expand the use of AI to support HPC workloads.[1] However, challenges surrounding the convergence of HPC/AI have slowed progress.

The sheer volume of data generated by converged HPC/AI workloads creates bandwidth bottlenecks and memory limitations. According to McKinsey, AI-ready data center capacity is expected to rise 33% annually through 2030.[2] Traditional IT infrastructures are not built to support the immense computational demands, often requiring costly re-engineering efforts. Organizations lacking experience in integrating HPC systems and modern AI frameworks often struggle to balance the need for traditional, centralized computing power with the distributed, diverse data demands of AI.

Universities, research labs, and sovereign entities are leading the search for a solution to bridge the gap between HPC and AI. The ideal system must deliver revolutionary performance for scientific computing, AI training, and inference as well as scale for the future.

Powering HPC workloads in Arm-based environments

HPE and NVIDIA are leaders in the new era of AI. Our partnership comprises decades of experience, collaborating with customers to enhance their HPC capabilities and delivering purpose-built solutions that accelerate the adoption of AI.

NVIDIA AI Computing by HPE empowers organizations across many industries to accelerate success with AI. The partnership brings together technology, people, and economics to create a faster, more flexible path to developing and deploying AI applications. The NVIDIA AI Computing by HPE portfolio simplifies AI with tailored infrastructure, software, networking, services, and support for enterprises, model builders, service providers, and sovereign entities. Regardless of your size or expertise, a full-stack approach ensures that your organization can achieve optimal performance, efficiency, and scale to gain important insights faster.

NVIDIA GB200 NVL4 by HPE is a new offering from the NVIDIA AI Computing by HPE portfolio. This dense, rack-scale solution is an entry point for AI and is designed for converged HPC and AI workloads. The NVIDIA GB200 Grace™ Blackwell superchip combines four NVIDIA Blackwell GPUs and two NVIDIA Grace Arm-based CPUs on a single board. The NVL4 architecture contains two NVIDIA Grace Blackwell superchips with a PCB hard bridge that connects the four GPUs to form a single NVLink domain, sharing resources and compute for massive HPC workloads and AI models. Combined with advanced liquid cooling and simplified management, the system provides up to 2X application performance over the prior generation and enables rapid deployment and scaling as data and model sizes grow.[3]

NVIDIA GB200 NVL4 by HPE gives organizations a high performance and highly flexible Arm-based environment to develop their codebase, allowing teams to increasingly augment their HPC efforts with AI.

[1] AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results, Hyperion Research 2025 hyperionresearch.com/olp_document/ai-investments-for-hpc-and-advanced-computing-continue-to-expand-and-begin-to-show-roi-results/
[2] AI power: Expanding data center capacity to meet growing demand, McKinsey 2024 mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand
[3] NVIDIA Blackwell Data Sheet, NVIDIA 2025 nvdam.widen.net/s/wwnsxrhm2w/blackwell-datasheet-3384703

Extending NVIDIA AI Computing by HPE with a dense solution for converged HPC/AI workloads

Delivering exceptional density

HPC customers depend on the latest technologies to keep pace with rapid data growth, power advanced applications, and accelerate demanding mixed workloads (including HPC, data analytics, AI post-training, and inferencing) for use cases such as climate change, disease research, and financial forecasting. Many of these organizations are experimenting with AI workloads to move their businesses forward, which creates new demands for high accuracy and precision in a limited IT footprint.

NVIDIA GB200 NVL4 by HPE is designed to meet these requirements, delivering the right combination of GPU acceleration and double-precision FP64 performance in a dense form factor. With up to 136 GPUs per rack, the system delivers 1.9x the density of comparable NVL72 solutions and 2.1x that of traditional 8-GPU servers. Organizations can start with a recommended minimum of 12 1U compute trays and expand to as many as 34 trays per rack to meet new requirements without disruptive infrastructure changes.

Organizations that adopt NVIDIA GB200 NVL4 by HPE benefit from lower cost and complexity compared to deploying individual servers or NVL72 configurations. The balanced compute engine enables seamless scaling for mixed workloads that require both scientific simulation capabilities and high performance AI acceleration. Benchmarking tests show 2.2x faster simulation and 1.8x faster training and inference compared to results from the previous generation.[4]

NVIDIA GB200 NVL4 by HPE

Solution components:

- Up to 34 1U compute trays in an NVIDIA 48U MGX™ rack
- Each compute tray has:
- 4 NVIDIA Blackwell (B200) GPUs (max 136 GPUs per rack)
- 2 NVIDIA Grace Arm-based CPUs
- PCB interconnect
- Direct liquid cooling from HPE
- HPE Performance Cluster Manager
- HPE Services

Capabilities:

- 40/80 PFLOPs (dense/sparse) in FP4 operations
- 40 TFLOPS in FP64 operations
- 186GB HBM3E memory at 8TB/s

[4] NVIDIA GB200 versus NVIDIA GH200 NVL4 | Scientific Computing – MILC | GNN Training – GraphCast 37M parameters, BF16 | Inference – OpenFold evoformer block, FP16, GenSLM Llama2-7B, FP16

Unleashing higher efficiency

Creating operational efficiencies is a critical challenge of AI. The use of data-hungry GPUs and CPUs has amplified the need for cooling options that boost efficiencies while also meeting demanding power requirements. HPC customers need systems that deliver sustainable high performance and simplify deployment and management so teams can reduce operational costs, better utilize resources, and focus on productivity.

Efficiency is a core feature of HPE and NVIDIA technology. NVIDIA GB200 NVL4 by HPE is a liquid-cooled environment that is easy to manage. Featuring NVIDIA Grace Arm-based CPUs, the system doubles data center performance for the same power envelope as competitive x86 processors, making it a sustainable choice for energy-conscious organizations.

HPE has decades of experience deploying the world's largest liquid-cooled environments. Today, we leverage our market leadership to bring direct liquid cooling (DLC) to HPC customers everywhere. DLC infrastructure is designed for high energy efficiency to optimize compute-intensive workloads and remain future-proof as HPC and AI continue to scale.

HPE Performance Cluster Manager (HPCM) ensures that deployment and ongoing operations are streamlined and cost-effective. HPCM offers integrated tools for managing system setup, hardware, software updates, cluster health, power and liquid cooling, and IT integration. Organizations can maintain high performance to mitigate costly downtime and increase productivity for a growing variety of workloads. Even the largest configurations can be provisioned, monitored, and managed through this flexible, easy-to-use solution (from clusters with dozens of nodes to exascale-level systems).

Accelerating deployment for faster time to value

Choosing the right technology partner is critical to succeed with AI. HPC customers around the world trust HPE and NVIDIA to accelerate their strategic HPC and AI initiatives. We specialize in designing, testing, installing, and supporting AI environments that remain future-proof for years. Our proven ecosystem offers the expertise and skills needed to rapidly deploy and scale AI infrastructure, achieving operational excellence from Day 1.

HPE Services includes a variety of advisory, professional, financial, and educational services:

- **HPC/AI custom support services:** Proactive support from dedicated remote and resident engineers with several levels of SLA coverage, tailored to meet your specific requirements.
- **On-site engineering resources:** Comprehensive on-site support from highly trained resident engineers who work with you to ensure system availability and performance.
- **Performance and benchmarking engagements:** Industry-leading performance engineering and benchmarking team available to fine tune your solution throughout its lifetime.
- **Sustainable IT services:** Energy and emissions reporting, sustainability workshops, and resource monitoring to reduce environmental impact.
- **HPE asset upcycling services:** Conversion of owned IT assets into an incremental cash source in a secure, globally consistent, and environmentally responsible way.

Conclusion

The convergence of HPC/AI is paving the way to a new era of productivity. HPC users at the forefront of this shift have the opportunity to unlock broader, faster insights for solving some of the world's greatest challenges and driving innovation on a global scale.

To accelerate enterprise AI, HPE and NVIDIA have joined forces to deliver AI factories integrating NVIDIA's accelerated computing with secure, scalable infrastructure from HPE. The NVIDIA AI Computing by HPE portfolio enables organizations to move beyond experiments and build robust, production-ready AI systems. Solutions include turnkey private cloud AI for fast starts, scalable AI factories for model builders, and sovereign AI factories for specialized data protection. These offerings feature validated technology stacks, multi-tenant support, air-gapped and federated architectures, predictable pricing, and rapid deployment. Fully integrated infrastructure and services allow IT teams to focus on developing AI use cases.

HPE and NVIDIA are proven leaders in accelerated computing, empowering organizations to succeed in this era of intelligence and the next. NVIDIA GB200 NVL4 by HPE provides exceptional density, high efficiency, and rapid scaling to harness the combined powers HPC and AI. Let us help you discover your full potential.

Learn more at

hpe.com/AI

Visit HPE.com

Chat now

a00155083ENW

HEWLETT PACKARD ENTERPRISE

hpe.com