



Hewlett Packard Enterprise and NVIDIA Announce ‘NVIDIA AI Computing by HPE’ to Accelerate Generative AI Industrial Revolution

New Lineup Features First-of-Its-Kind Turnkey, Private-Cloud AI Solution Including Sustainable Accelerated Computing with Full Lifecycle Services to Streamline Time to Value with AI

HPE Discover 2024—[Hewlett Packard Enterprise](#) (NYSE: HPE) and NVIDIA today announced [NVIDIA AI Computing by HPE](#), a portfolio of co-developed AI solutions and joint go-to-market integrations that enable enterprises to accelerate adoption of generative AI.

Among the portfolio’s key offerings is [HPE Private Cloud AI](#), a first-of-its-kind solution that provides the deepest integration to date of NVIDIA AI computing, networking and software with HPE’s AI storage, compute and the HPE GreenLake cloud. The offering enables enterprises of every size to gain an energy-efficient, fast and flexible path for sustainably developing and deploying generative AI applications. Powered by the new OpsRamp AI copilot that helps IT operations improve workload and IT efficiency, HPE Private Cloud AI includes a self-service cloud experience with full lifecycle management and is available in four right-sized configurations to support a broad range of AI workloads and use cases.

All NVIDIA AI Computing by HPE offerings and services will be available through a joint go-to-market strategy that spans sales teams and channel partners, training and a global network of system integrators — including Deloitte, HCLTech, Infosys, TCS and Wipro — that can help enterprises across a variety of industries run complex AI workloads.

Announced during the HPE Discover keynote by HPE President and CEO Antonio Neri, who was joined by NVIDIA founder and CEO Jensen Huang, NVIDIA AI Computing by HPE marks the expansion of a decades-long partnership and reflects the substantial commitment of time and resources from each company.

“Generative AI holds immense potential for enterprise transformation, but the complexities of fragmented AI technology contain too many risks and barriers that hamper large-scale enterprise adoption and can jeopardize a company’s most valuable asset — its proprietary data,” said Neri. “To unleash the immense potential of generative AI in the enterprise, HPE and NVIDIA co-developed a turnkey private cloud for AI that will enable enterprises to focus their resources on developing new AI use cases that can boost productivity and unlock new revenue streams.”

“Generative AI and accelerated computing are fueling a fundamental transformation as every industry races to join the industrial revolution,” said Huang. “Never before have NVIDIA and HPE integrated our technologies so deeply — combining the entire NVIDIA AI computing stack along with HPE’s private cloud technology — to equip enterprise clients and AI professionals with the most advanced computing infrastructure and services to expand the frontier of AI.”

HPE and NVIDIA co-developed Private Cloud AI portfolio

HPE Private Cloud AI delivers a unique, cloud-based experience to accelerate innovation and return on investment while managing enterprise risk from AI. The solution offers:

- Support for inference, fine-tuning and RAG AI workloads that utilize proprietary data.
- Enterprise control for data privacy, security, transparency and governance requirements.
- Cloud experience with ITOps and AIOps capabilities to increase productivity.
- Fast path to consume flexibly to meet future AI opportunities and growth.

Curated AI and data software stack in HPE Private Cloud AI

The foundation of the AI and data software stack starts with the [NVIDIA AI Enterprise software](#) platform, which includes [NVIDIA NIM™ inference microservices](#).

NVIDIA AI Enterprise accelerates data science pipelines and streamlines development and deployment of production-grade copilots and other GenAI applications. Included with NVIDIA AI Enterprise, NVIDIA NIM delivers easy-to-use microservices for optimized AI model inferencing offering a smooth transition from prototype to secure deployment of AI models in a variety of use cases.

Complementing NVIDIA AI Enterprise and NVIDIA NIM, HPE AI Essentials software delivers a ready to run set of curated AI and data foundation tools with a unified control plane that provide adaptable solutions, ongoing enterprise support and

trusted AI services, such as data and model compliance and extensible features that ensure AI pipelines are in compliance, explainable and reproducible throughout the AI lifecycle.

To deliver optimal performance for the AI and data software stack, HPE Private Cloud AI delivers a fully integrated AI infrastructure stack that includes [NVIDIA Spectrum-X™ Ethernet networking](#), HPE GreenLake for File Storage and HPE ProLiant servers with support for [NVIDIA L40S](#), [NVIDIA H100 NVL Tensor Core GPUs](#) and the [NVIDIA GH200 NVL2 platform](#).

Cloud experience enabled by HPE GreenLake cloud

HPE Private Cloud AI offers a self-service cloud experience enabled by HPE GreenLake cloud. Through a single, platform-based control plane, HPE GreenLake cloud services provide manageability and observability to automate, orchestrate and manage endpoints, workloads and data across hybrid environments. This includes sustainability metrics for workloads and endpoints.

HPE GreenLake cloud and OpsRamp AI infrastructure observability and copilot assistant

[OpsRamp's](#) IT operations are integrated with HPE GreenLake cloud to deliver observability and AIOps to all HPE products and services. OpsRamp now provides observability for the end-to-end NVIDIA accelerated computing stack, including NVIDIA NIM and AI software, NVIDIA Tensor Core GPUs and AI clusters as well as [NVIDIA Quantum InfiniBand](#) and NVIDIA Spectrum Ethernet switches. IT administrators can gain insights to identify anomalies and monitor their AI infrastructure and workloads across hybrid, multi-cloud environments.

The new OpsRamp operations copilot utilizes NVIDIA's accelerated computing platform to analyze large datasets for insights with a conversational assistant, boosting productivity for operations management. OpsRamp will also integrate with CrowdStrike APIs so customers can see a unified service map view of endpoint security across their entire infrastructure and applications.

Accelerate time to value with AI — expanded collaboration with global system integrators

To advance the time to value for enterprises to develop industry-focused AI solutions and use cases with clear business benefits, [Deloitte](#), [HCLTech](#), [Infosys](#), [TCS](#) and [Wipro](#) announced their support of the NVIDIA AI Computing by HPE portfolio and HPE Private Cloud AI as part of their strategic AI solutions and services.

HPE adds support for NVIDIA's latest GPUs, CPUs and Superchips

- HPE Cray XD670 supports eight NVIDIA H200 NVL Tensor Core GPUs and is ideal for LLM builders.
- HPE ProLiant DL384 Gen12 server with NVIDIA GH200 NVL2 is ideal for LLM consumers using larger models or RAG.
- HPE ProLiant DL380a Gen12 server support for up to eight NVIDIA H200 NVL Tensor Core GPUs is ideal for LLM users looking for flexibility to scale their GenAI workloads.
- HPE will be time-to-market to support the NVIDIA GB200 NVL72 / NVL2, as well as the new NVIDIA Blackwell, NVIDIA Rubin and NVIDIA Vera architectures.

High-density file storage certified for NVIDIA DGX BasePOD and NVIDIA OVX systems

[HPE GreenLake for File Storage](#) has achieved [NVIDIA DGX BasePOD](#) certification and NVIDIA OVX™ storage validation, providing customers with a proven enterprise file storage solution for accelerating AI, GenAI and GPU-intensive workloads at scale. HPE will be a time-to-market partner on upcoming NVIDIA reference architecture storage certification programs.

Availability

- HPE Private Cloud AI is expected to be generally available in the fall.
- HPE ProLiant DL380a Gen12 server with NVIDIA H200 NVL Tensor Core GPUs is expected to be generally available in the fall.
- HPE ProLiant DL384 Gen12 server with dual NVIDIA GH200 NVL2 is expected to be generally available in the fall.
- HPE Cray XD670 server with NVIDIA H200 NVL is expected to be generally available in the summer.

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in accelerated computing.

About Hewlett Packard Enterprise

Hewlett Packard Enterprise (NYSE: HPE) is the global edge-to-cloud company that helps organizations accelerate outcomes by unlocking value from all of their data, everywhere. Built on decades of reimagining the future and innovating to advance the way people live and work, HPE delivers unique, open and intelligent technology solutions as a service. With offerings spanning Cloud Services, Compute, High Performance Computing & AI, Intelligent Edge, Software, and Storage, HPE provides a consistent experience across all clouds and edges, helping customers develop new business models, engage in new ways and increase operational performance. For more information, visit: www.hpe.com.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features, and availability of NVIDIA's products and technologies, including NVIDIA AI Enterprise software platform, NVIDIA NIM inference microservices, NVIDIA Spectrum-X Ethernet networking, NVIDIA L40S, NVIDIA H100 NVL Tensor Core GPUs, the NVIDIA GH200 NVL2 platform, NVIDIA Quantum InfiniBand, and NVIDIA Spectrum Ethernet switches; the benefits and impact of NVIDIA's expanded partnership with HPE, and the features and availability of its services and offerings; third parties using or adopting NVIDIA's products or technologies and the benefits thereof; generative AI and accelerated computing fueling a fundamental transformation as every industry races to join the industrial revolution; and NVIDIA and HPE combining the entire NVIDIA AI computing stack along with HPE's private cloud technology to equip enterprise clients and AI professionals with the most advanced computing infrastructure and services to expand the frontier of AI are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; NVIDIA's reliance on third parties to manufacture, assemble, package and test its products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or its partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

© 2024 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA DGX BasePOD, NVIDIA OVX and NVIDIA Spectrum-X are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and/or other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability, and specifications are subject to change without notice.

Pearlina Boc
NVIDIA Corporation
+1-562-275-5781
pboc@nvidia.com
John Choi
HPE
john.choi@hpe.com