NVIDIA AI-Ready Servers From World’s Leading System Manufacturers to Supercharge Generative AI for Enterprises

Servers Featuring NVIDIA L40S GPUs and NVIDIA BlueField Coming Soon From Dell Technologies, Hewlett Packard Enterprise and Lenovo to Support VMware Private AI Foundation with NVIDIA

VMware Explore -- NVIDIA today announced the world’s leading system manufacturers will deliver AI-ready servers that support VMware Private AI Foundation with NVIDIA, announced separately today, to help companies customize and deploy generative AI applications using their proprietary business data.

NVIDIA AI-ready servers will include NVIDIA® L40S GPUs, NVIDIA BlueField®-3 DPUs and NVIDIA AI Enterprise software to enable enterprises to fine-tune generative AI foundation models and deploy generative AI applications like intelligent chatbots, search and summarization tools. These servers also provide NVIDIA-accelerated infrastructure and software to power VMware Private AI Foundation with NVIDIA.

NVIDIA L40S-powered servers from leading global system manufacturers — Dell Technologies, Hewlett Packard Enterprise and Lenovo — will be available by year-end to accelerate enterprise AI.

“A new computing era has begun,” said Jensen Huang, founder and CEO of NVIDIA. “Companies in every industry are racing to adopt generative AI. With our ecosystem of world-leading software and system partners, we are bringing generative AI to the world’s enterprises.”

NVIDIA AI-ready servers are an ideal platform for businesses that will deploy VMware Private AI Foundation with NVIDIA.

“Generative AI is supercharging digital transformation, and enterprises need a fully integrated solution to more securely build applications that enable them to advance their business,” said Raghu Raghuram, CEO of VMware. “Through the combined expertise of VMware, NVIDIA and our server manufacturer partners, businesses will be able to develop and deploy AI with data privacy, security and control.”

Powering Generative AI Transformation in the Enterprise
NVIDIA AI-ready servers are designed to provide full-stack accelerated infrastructure and software for industries racing to adopt generative AI for a broad range of applications, including drug discovery, retail product descriptions, intelligent virtual assistants, manufacturing simulation and fraud detection.

The servers feature NVIDIA AI Enterprise, the operating system of the NVIDIA AI platform. The software provides production-ready enterprise support and security for over 100 frameworks, pretrained models, toolkits and software, including NVIDIA NeMo™ for LLMs, NVIDIA Modulus for simulations, NVIDIA RAPIDS™ for data science and NVIDIA Triton™ Inference Server for production AI.

Built to handle complex AI workloads with billions of parameters, L40S GPUs include fourth-generation Tensor Cores and an FP8 Transformer Engine, delivering over 1.45 petaflops of tensor processing power and up to 1.7x training performance compared with the NVIDIA A100 Tensor Core GPU.

For generative AI applications such as intelligent chatbots, assistants, search and summarization, the NVIDIA L40S enables up to 1.2x more generative AI inference performance than the NVIDIA A100 GPU.

Integrating NVIDIA BlueField DPUs drives further speedups by accelerating, offloading and isolating the tremendous compute load of virtualization, networking, storage, security and other cloud-native AI services.

NVIDIA ConnectX®-7 SmartNICs offer advanced hardware offloads and ultra-low latency, delivering best-in-class, scalable performance for data-intensive generative AI workloads.

Broad Ecosystem to Speed Enterprise Generative AI Deployments
The world’s leading computer makers are building NVIDIA AI-ready servers, including the Dell PowerEdge R760xa, HPE ProLiant Gen11 servers for VMware Private AI Foundation with NVIDIA, and Lenovo ThinkSystem SR675 V3.

"Generative AI is a catalyst for innovation, helping to solve some of the world’s most pressing challenges,” said Michael Dell, chairman and chief executive officer, Dell Technologies. “Dell Generative AI Solutions with NVIDIA AI-ready servers will play a critical role in advancing human progress by driving unprecedented levels of productivity and revolutionizing the way industries operate."

“Generative AI will usher in a new scale of productivity for enterprises, from powering chatbots and digital assistants to helping with the design and development of new solutions,” said Antonio Neri, president and CEO of HPE. “We are pleased to continue working closely with NVIDIA to feature its GPUs and software in a range of enterprise tuning and inference workload solutions that will accelerate deployments of generative AI.”

“Businesses are eager to adopt generative AI to power intelligent transformation,” said Yang Yuanqing, chairman and CEO of Lenovo. “In collaboration with NVIDIA and VMware, Lenovo is further extending our leadership in generative AI and solidifying our unique position in helping customers in their AI journey.”

Availability
NVIDIA AI-ready servers with L40S GPUs and BlueField DPUs will be available by year-end, with instances available from cloud service providers expected in the coming months.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company’s invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features and availability of our products and technologies, including NVIDIA AI-ready servers, NVIDIA L40S GPUs, NVIDIA BlueField-3 DPUs, NVIDIA AI Enterprise, NVIDIA NeMo, NVIDIA Modulus, NVIDIA RAPIDS, NVIDIA Triton Inference Server, NVIDIA Ada Lovelace, NVIDIA A100 Tensor Core GPU, CUDA, and NVIDIA ConnectX-7 SmartNICs; leading system manufacturers delivering NVIDIA AI-ready servers; companies in every industry racing to adopt generative AI; NVIDIA bringing generative AI to the world’s enterprises; leading computer makers building NVIDIA AI-ready servers; generative AI as a catalyst for innovation; generative AI as the next frontier of digital transformation; businesses being eager to adopt generative AI to power innovation, productivity and creativity; and the impact and benefits of NVIDIA’s work with Dell Technologies, Hewlett Packard Enterprise and Lenovo 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; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our 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.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com