NVIDIA, Global Data Center System Manufacturers to Supercharge Generative AI and Industrial Digitalization

OVX Servers Feature New NVIDIA GPUs to Accelerate Training and Inference, Graphics-Intensive Workloads; Coming Soon From Dell Technologies, Hewlett Packard Enterprise, Lenovo, Supermicro and More

SIGGRAPH—NVIDIA today announced NVIDIA OVX™ servers featuring the new NVIDIA® L40S GPU, a powerful, universal data center processor designed to accelerate the most compute-intensive, complex applications, including AI training and inference, 3D design and visualization, video processing and industrial digitalization with the NVIDIA Omniverse™ platform.

The new GPU powers accelerated computing workloads for generative AI, which is transforming workflows and services across industries, including text, image and video generation, chatbots, game development, product design and healthcare.

“As generative AI transforms every industry, enterprises are increasingly seeking large-scale compute resources in the data center,” said Bob Pette, vice president of professional visualization at NVIDIA. “OVX systems with NVIDIA L40S GPUs accelerate AI, graphics and video processing workloads, and meet the demanding performance requirements of an ever-increasing set of complex and diverse applications.”

Powerful Performance for AI and Graphics
NVIDIA OVX systems will enable up to eight NVIDIA L40S GPUs per server, each equipped with 48GB of memory. Based on the NVIDIA Ada Lovelace GPU architecture, the L40S includes fourth-generation Tensor Cores and an FP8 Transformer Engine, delivering over 1.45 petaflops of tensor processing power. For complex AI workloads with billions of parameters and multiple data modalities — such as text and video — L40S enables up to 1.2x more generative AI inference performance and up to 1.7x training performance compared with the NVIDIA A100 Tensor Core GPU.

To power high-fidelity professional visualization workflows like real-time rendering, product design and 3D content creation, the NVIDIA L40S GPU includes 142 third-generation RT Cores that deliver 212 teraflops of ray-tracing performance. This enables creative professionals to create immersive visual experiences and photorealistic content.

For computationally demanding workflows, such as engineering and scientific simulations, the NVIDIA L40S includes 18,176 CUDA® cores, delivering nearly 5x the single-precision floating-point (FP32) performance of the NVIDIA A100 GPU to accelerate complex calculations and data-intensive analyses.

Early Adoption
Among the first cloud service providers to offer L40S instances is CoreWeave, which specializes in large-scale, GPU-accelerated workloads.

“With the explosion of generative AI, our customers across industries are seeking powerful compute offerings and scale to match the complexity of any workload — from interactive video to AI design and automation,” said Brian Venturo, chief technology officer at CoreWeave. “NVIDIA L40S GPUs will further expand our broad portfolio of NVIDIA solutions, making CoreWeave the first specialized cloud provider to offer these new resources for fast, efficient and cost-effective accelerated computing to power the next wave of generative AI applications.”

Software to Boost AI
Enterprises deploying L40S GPUs can benefit from NVIDIA AI Enterprise software, which announced a major update today. The software provides production-ready enterprise support and security for over 100 frameworks, pretrained models, toolkits and software, including NVIDIA Modulus for simulations, NVIDIA RAPIDS™ for data science and NVIDIA Triton™ Inference Server for production AI.

Omniverse Expands
NVIDIA also announced major updates to the Omniverse platform, introducing capabilities and platform enhancements that enable developers to accelerate and advance OpenUSD pipelines and industrial digitalization applications with the power of generative AI. The next generation of NVIDIA OVX systems powering Omniverse Cloud will feature L40S GPUs to deliver the AI and graphics performance needed to supercharge generative AI pipelines and Omniverse workloads.

Availability
The NVIDIA L40S will be available starting this fall. Global system builders, including ASUS, Dell Technologies, GIGABYTE, HPE, Lenovo, QCT and Supermicro, will soon offer OVX systems that include the NVIDIA L40S GPUs. These servers will help professionals worldwide advance AI and bring generative AI applications like intelligent chatbots, search and summarization tools to users across industries.

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, services and technologies, including NVIDIA OVX servers, NVIDIA Omniverse, NVIDIA Ada Lovelace, NVIDIA A100 Tensor Core GPU, NVIDIA L40S GPU, NVIDIA AI Enterprise software, NVIDIA Modulus, NVIDIA RAPIDS and NVIDIA Triton Inference Server; generative AI transforming workflows and services across industries; generative AI transforming every industry; enterprises increasingly seeking large-scale compute resources in the data center; OVX systems meeting the demand performance requirements of complex and diverse applications; the next generation of NVIDIA OVX systems powering Omniverse Cloud featuring L40S GPUs; and the builders offering OVX systems that include L40S GPUs 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.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com
Kasia Johnston
+1-415-813-8859
kasiaj@nvidia.com