AWS and NVIDIA Announce Strategic Collaboration to Offer New Supercomputing Infrastructure, Software and Services for Generative AI

- *AWS to offer first cloud AI supercomputer with NVIDIA Grace Hopper Superchip and AWS UltraCluster scalability*
- *NVIDIA DGX Cloud—first to feature NVIDIA GH200 NVL32—coming to AWS*
- *Companies partner on Project Ceiba—the world's fastest GPU-powered AI supercomputer and newest NVIDIA DGX Cloud supercomputer for NVIDIA AI R&D and custom model development*
- *New Amazon EC2 instances powered by NVIDIA GH200, H200, L40S and L4 GPUs supercharge generative AI, HPC, design and simulation workloads*
- *NVIDIA software on AWS—NeMo LLM framework, NeMo Retriever and BioNeMo—to boost generative AI development for custom models, semantic retrieval and drug discovery*

AWS re:Invent—Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), and NVIDIA (NASDAQ: NVDA) today announced an expansion of their strategic collaboration to deliver the most-advanced infrastructure, software and services to power customers' generative artificial intelligence (AI) innovations.

The companies will bring together the best of NVIDIA and AWS technologies—from NVIDIA's newest multi-node systems featuring next-generation GPUs, CPUs and AI software, to AWS Nitro System advanced virtualization and security, Elastic Fabric Adapter (EFA) interconnect, and UltraCluster scalability—that are ideal for training foundation models and building generative AI applications.

The expanded collaboration builds on a longstanding relationship that has fueled the generative AI era by offering early machine learning (ML) pioneers the compute performance required to advance the state-of-the-art in these technologies.

As part of the expanded collaboration to supercharge generative AI across all industries:

- AWS will be the first cloud provider to bring NVIDIA® GH200 Grace Hopper Superchips with new multi-node NVLink™ technology to the cloud. The new NVIDIA GH200 NVL32 multi-node platform connects 32 Grace Hopper Superchips with NVIDIA NVLink and NVSwitch™ technologies into one instance. The platform will be available on Amazon Elastic Compute Cloud (Amazon EC2) instances connected with Amazon's powerful networking (EFA), supported by advanced virtualization (AWS Nitro System), and hyper-scale clustering (Amazon EC2 UltraClusters), enabling joint customers to scale to thousands of GH200 Superchips.
- NVIDIA and AWS will collaborate to host NVIDIA DGX™ Cloud—NVIDIA's AI-training-as-a-service—on AWS. It will be the first DGX Cloud featuring GH200 NVL32, providing developers the largest shared memory in a single instance. DGX Cloud on AWS will accelerate training of cutting-edge generative AI and large language models that can reach beyond 1 trillion parameters.
- NVIDIA and AWS are partnering on Project Ceiba to design the world's fastest GPU-powered AI supercomputer—an at-scale system with GH200 NVL32 and Amazon EFA interconnect hosted by AWS for NVIDIA's own research and development team. This first-of-its-kind supercomputer—featuring 16,384 NVIDIA GH200 Superchips and capable of processing 65 exaflops of AI—will be used by NVIDIA to propel its next wave of generative AI innovation.
- AWS will introduce three additional new Amazon EC2 instances: P5e instances, powered by NVIDIA H200 Tensor Core GPUs, for large-scale and cutting-edge generative AI and HPC workloads, and G6 and G6e instances, powered by NVIDIA L4 GPUs and NVIDIA L40S GPUs, respectively, for a wide set of applications such as AI fine-tuning, inference, graphics and video workloads. G6e instances are particularly suitable for developing 3D workflows, digital twins and other applications using NVIDIA Omniverse™, a platform for connecting and building generative AI-enabled 3D applications.

"AWS and NVIDIA have collaborated for more than 13 years, beginning with the world's first GPU cloud instance. Today, we offer the widest range of NVIDIA GPU solutions for workloads including graphics, gaming, high performance computing, machine learning, and now, generative AI," said Adam Selipsky, CEO at AWS. "We continue to innovate with NVIDIA to make AWS the best place to run GPUs, combining next-gen NVIDIA Grace Hopper Superchips with AWS's EFA powerful networking, EC2 UltraClusters' hyper-scale clustering, and Nitro's advanced virtualization capabilities."

"Generative AI is transforming cloud workloads and putting accelerated computing at the foundation of diverse content generation," said Jensen Huang, founder and CEO of NVIDIA. "Driven by a common mission to deliver cost-effective state-of-the-art generative AI to every customer, NVIDIA and AWS are collaborating across the entire computing stack, spanning AI infrastructure, acceleration libraries, foundation models, to generative AI services."

New Amazon EC2 Instances Combine State-of-the-Art from NVIDIA and AWS
AWS will be the first cloud provider to offer NVIDIA GH200 Grace Hopper Superchips with multi-node NVLink technology.

Each GH200 Superchip combines an Arm-based Grace CPU with an NVIDIA Hopper™ architecture GPU on the same module. A single Amazon EC2 instance with GH200 NVL32 can provide up to 20 TB of shared memory to power terabyte-scale workloads.

These instances will take advantage of AWS's third-generation Elastic Fabric Adapter (EFA) interconnect, providing up to 400 Gbps per Superchip of low-latency, high-bandwidth networking throughput, enabling customers to scale to thousands of GH200 Superchips in EC2 UltraClusters.

AWS instances with GH200 NVL32 will provide customers on-demand access to supercomputer-class performance, which is critical for large-scale AI/ML workloads that need to be distributed across multiple nodes for complex generative AI workloads —spanning FMs, recommender systems, and vector databases.

NVIDIA GH200-powered EC2 instances will feature 4.5 TB of HBM3e memory—a 7.2x increase compared to current generation H100-powered EC2 P5d instances—allowing customers to run larger models, while improving training performance. Additionally, CPU-to-GPU memory interconnect provides up to 7x higher bandwidth than PCIe, enabling chip-to-chip communications that extend the total memory available for applications.

AWS instances with GH200 NVL32 will be the first AI infrastructure on AWS to feature liquid cooling to help ensure densely-packed server racks can efficiently operate at maximum performance.

EC2 instances with GH200 NVL32 will also benefit from the AWS Nitro System, the underlying platform for next-generation EC2 instances. The Nitro System offloads I/O for functions from the host CPU/GPU to specialized hardware to deliver more consistent performance, while its enhanced security protects customer code and data during processing.

AWS First to Host NVIDIA DGX Cloud Powered by Grace Hopper
AWS will team up with NVIDIA to host NVIDIA DGX Cloud powered by GH200 NVL32 NVLink infrastructure. NVIDIA DGX Cloud is an AI supercomputing service that gives enterprises fast access to multi-node supercomputing for training the most complex LLM and generative AI models, with integrated NVIDIA AI Enterprise software, and direct access to NVIDIA AI experts.

Massive Project Ceiba Supercomputer to Supercharge NVIDIA's AI Development
The Project Ceiba supercomputer that AWS and NVIDIA are collaborating on will be integrated with AWS services, such as Amazon Virtual Private Cloud (VPC) encrypted networking and Amazon Elastic Block Store high-performance block storage, giving NVIDIA access to a comprehensive set of AWS capabilities.

NVIDIA will use the supercomputer for research and development to advance AI for LLMs, graphics and simulation, digital biology, robotics, self-driving cars, Earth-2 climate prediction and more.

NVIDIA and AWS Supercharge Generative AI, HPC, Design and Simulation
To power the development, training and inference of the largest LLMs, AWS P5e instances will feature NVIDIA's latest H200 GPUs that offer 141 GB of HBM3e GPU memory, which is 1.8x larger and 1.4x faster than H100 GPUs. This boost in GPU memory, along with up to 3,200 Gbps of EFA networking enabled by the AWS Nitro System, will enable customers to continue to build, train and deploy their cutting-edge models on AWS.

To deliver cost-effective, energy-efficient solutions for video, AI and graphics workloads, AWS announced new Amazon EC2 G6e instances featuring NVIDIA L40S GPUs and G6 instances powered by L4 GPUs. The new offerings can help startups, enterprises and researchers meet their AI and high-fidelity graphics needs.

G6e instances are built to handle complex workloads such as generative AI and digital twin applications. Using NVIDIA Omniverse, photorealistic 3D simulations can be developed, contextualized and enhanced using real-time data from services such as AWS IoT TwinMaker, intelligent chatbots, assistants, search and summarization. Amazon Robotics and Amazon Fulfillment Centers will be able to integrate digital twins built with NVIDIA Omniverse and AWS IoT TwinMaker to optimize warehouse design and flow, train more intelligent robot assistants and improve deliveries to customers.

L40S GPUs deliver up to 1.45 petaflops of FP8 performance and feature Ray Tracing cores that offer up to 209 teraflops of ray-tracing performance. L4 GPUs featured in G6 instances will deliver a lower-cost, energy-efficient solution for deploying AI models for natural language processing, language translation, AI video and image analysis, speech recognition, and personalization. L40S GPUs also accelerate graphics workloads, such as creating and rendering real-time, cinematic-quality graphics and game streaming. All three instances will be available in the coming year.

NVIDIA Software on AWS Boosts Generative AI Development
In addition, NVIDIA announced software on AWS to boost generative AI development. NVIDIA NeMo™ Retriever microservice offers new tools to create highly accurate chatbots and summarization tools using accelerated semantic retrieval. NVIDIA BioNeMo™, available on Amazon SageMaker now and coming to AWS on NVIDIA DGX Cloud, enables pharmaceutical companies to speed drug discovery by simplifying and accelerating the training of models using their own data.

NVIDIA software on AWS is helping Amazon bring new innovations to its services and operations. AWS is using the NVIDIA

NeMo framework to train select next-generation Amazon Titan LLMs. Amazon Robotics has begun leveraging NVIDIA Omniverse Isaac to build digital twins for automating, optimizing and planning its autonomous warehouses in virtual environments before deploying them into the real world.

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 the creation of the metaverse. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

About Amazon Web Services
Since 2006, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud. AWS has been continually expanding its services to support virtually any workload, and it now has more than 240 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 102 Availability Zones within 32 geographic regions, with announced plans for 15 more Availability Zones and five more AWS Regions in Canada, Germany, Malaysia, New Zealand, and Thailand. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.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 GH200 Grace Hopper Superchips, NVL32, H200, NeMo Retriever, NVLink, NVSwitch, NVIDIA DGX Cloud, NVIDIA L40S, NVIDIA Omniverse, NVIDIA L4, NVIDIA NeMo, NVIDIA BioNeMo, and NVIDIA AI Enterprise; the benefits and impact of the expanded collaboration between AWS and NVIDIA, including Project Ceiba, and the availability of its services and offerings; and generative AI transforming cloud workloads and putting accelerated computing at the foundation of diverse content generation 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; and 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.

© 2023 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, BioNeMo, DGX, NeMo, NVIDIA Omniverse, NVLink, and NVSwitch 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.

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