AWS and NVIDIA Extend Collaboration to Advance Generative AI Innovation

- AWS to offer NVIDIA Grace Blackwell GPU-based Amazon EC2 instances and NVIDIA DGX Cloud to accelerate performance of building and running inference on multi-trillion-parameter LLMs
- Integration of AWS Nitro System, Elastic Fabric Adapter encryption, and AWS Key Management Service with Blackwell encryption provides customers end-to-end control of their training data and model weights to provide even stronger security for customers' AI applications on AWS
- Project Ceiba — an AI supercomputer built exclusively on AWS with DGX Cloud — to feature 20,736 GB200 Superchips capable of 414 exaflops for NVIDIA's own AI R&D
- Amazon SageMaker integration with NVIDIA NIM inference microservices helps customers further optimize price performance of foundation models running on GPUs
- Collaboration between AWS and NVIDIA accelerates AI innovation across healthcare and life sciences

GTC—Amazon Web Services (AWS), an Amazon.com company (NASDAQ: AMZN), and NVIDIA (NASDAQ: NVDA) today announced that the new NVIDIA Blackwell GPU platform — unveiled by NVIDIA at GTC 2024 — is coming to AWS. AWS will offer the NVIDIA GB200 Grace Blackwell Superchip and B100 Tensor Core GPUs, extending the companies' long standing strategic collaboration to deliver the most secure and advanced infrastructure, software, and services to help customers unlock new generative artificial intelligence (AI) capabilities.

NVIDIA and AWS continue to bring together the best of their technologies, including NVIDIA's newest multi-node systems featuring the next-generation NVIDIA Blackwell platform and AI software, AWS's Nitro System and AWS Key Management Service (AWS KMS) advanced security, Elastic Fabric Adapter (EFA) petabit scale networking, and Amazon Elastic Compute Cloud (Amazon EC2) UltraCluster hyper-scale clustering. Together, they deliver the infrastructure and tools that enable customers to build and run real-time inference on multi-trillion parameter large language models (LLMs) faster, at massive scale, and at a lower cost than previous-generation NVIDIA GPUs on Amazon EC2.

"The deep collaboration between our two organizations goes back more than 13 years, when together we launched the world's first GPU cloud instance on AWS, and today we offer the widest range of NVIDIA GPU solutions for customers," said Adam Selipsky, CEO at AWS. "NVIDIA's next-generation Grace Blackwell processor marks a significant step forward in generative AI and GPU computing. When combined with AWS's powerful Elastic Fabric Adapter Networking, Amazon EC2 UltraClusters' hyper-scale clustering, and our unique Nitro system's advanced virtualization and security capabilities, we make it possible for customers to build and run multi-trillion parameter large language models faster, at massive scale, and more securely than anywhere else. Together, we continue to innovate to make AWS the best place to run NVIDIA GPUs in the cloud."

"AI is driving breakthroughs at an unprecedented pace, leading to new applications, business models, and innovation across industries," said Jensen Huang, founder and CEO of NVIDIA. "Our collaboration with AWS is accelerating new generative AI capabilities and providing customers with unprecedented computing power to push the boundaries of what's possible."

Latest innovations from AWS and NVIDIA accelerate training of cutting-edge LLMs that can reach beyond 1 trillion parameters

AWS will offer the NVIDIA Blackwell platform, featuring GB200 NVL72, with 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVIDIA NVLink™. When connected with Amazon's powerful networking (EFA), and supported by advanced virtualization (AWS Nitro System) and hyper-scale clustering (Amazon EC2 UltraClusters), customers can scale to thousands of GB200 Superchips. NVIDIA Blackwell on AWS delivers a massive leap forward in speeding up inference workloads for resource-intensive, multi-trillion-parameter language models.

Based on the success of the NVIDIA H100-powered EC2 P5 instances, which are available to customers for short durations through Amazon EC2 Capacity Blocks for ML, AWS plans to offer EC2 instances featuring the new B100 GPUs deployed in EC2 UltraClusters for accelerating generative AI training and inference at massive scale. GB200s will also be available on NVIDIA DGX™ Cloud, an AI platform co-engineered on AWS, that gives enterprise developers dedicated access to the infrastructure and software needed to build and deploy advanced generative AI models. The Blackwell-powered DGX Cloud instances on AWS will accelerate development of cutting-edge generative AI and LLMs that can reach beyond 1 trillion parameters.

Elevate AI security with AWS Nitro System, AWS KMS, encrypted EFA, and Blackwell encryption

As customers move quickly to implement AI in their organizations, they need to know that their data is being handled securely throughout their training workflow. The security of model weights — the parameters that a model learns during training that are critical for its ability to make predictions — is paramount to protecting customers' intellectual property, preventing tampering with models, and maintaining model integrity.

AWS AI infrastructure and services already have security features in place to give customers control over their data and ensure that it is not shared with third-party model providers. The combination of the AWS Nitro System and the NVIDIA GB200 takes AI security even further by preventing unauthorized individuals from accessing model weights. The GB200 allows inline encryption of the NVLink connections between GPUs and encrypts data transfers, while EFA encrypts data across servers for distributed training and inference. The GB200 will also benefit from the AWS Nitro System, which offloads I/O for functions from the host CPU/GPU to specialized AWS hardware to deliver more consistent performance, while its enhanced security protects customer code and data during processing — on both the customer side and AWS side. This capability — available only on AWS — has been independently verified by NCC Group, a leading cybersecurity firm.

With the GB200 on Amazon EC2, AWS will enable customers to create a trusted execution environment alongside their EC2 instance, using AWS Nitro Enclaves and AWS KMS. Nitro Enclaves allow customers to encrypt their training data and weights with KMS, using key material under their control. The enclave can be loaded from within the GB200 instance and can communicate directly with the GB200 Superchip. This enables KMS to communicate directly with the enclave and pass key material to it in a cryptographically secure way. The enclave can then pass that material to the GB200, protected from the customer instance and preventing AWS operators from ever accessing the key or decrypting the training data or model weights, giving customers unparalleled control over their data.

Project Ceiba taps Blackwell to propel NVIDIA's future generative AI innovation on AWS
Announced at AWS re:Invent 2023, Project Ceiba is a collaboration between NVIDIA and AWS to build one of the world's fastest AI supercomputers. Hosted exclusively on AWS, the supercomputer is available for NVIDIA's own research and development. This first-of-its-kind supercomputer with 20,736 B200 GPUs is being built using the new NVIDIA GB200 NVL72, a system featuring fifth-generation NVLink, that scales to 20,736 B200 GPUs connected to 10,368 NVIDIA Grace CPUs. The system scales out using fourth-generation EFA networking, providing up to 800 Gbps per Superchip of low-latency, high-bandwidth networking throughput — capable of processing a massive 414 exaflops of AI — a 6x performance increase over earlier plans to build Ceiba on the Hopper architecture. NVIDIA research and development teams will use Ceiba to advance AI for LLMs, graphics (image/video/3D generation) and simulation, digital biology, robotics, self-driving cars, NVIDIA Earth-2 climate prediction, and more to help NVIDIA propel future generative AI innovation.

AWS and NVIDIA collaboration accelerates development of generative AI applications and advance use cases in healthcare and life sciences
AWS and NVIDIA have joined forces to offer high-performance, low-cost inference for generative AI with Amazon SageMaker integration with NVIDIA NIM™ inference microservices, available with NVIDIA AI Enterprise. Customers can use this combination to quickly deploy FMs that are pre-compiled and optimized to run on NVIDIA GPUs to SageMaker, reducing the time-to-market for generative AI applications.

AWS and NVIDIA have teamed up to expand computer-aided drug discovery with new NVIDIA BioNeMo™ FMs for generative chemistry, protein structure prediction, and understanding how drug molecules interact with targets. These new models will soon be available on AWS HealthOmics, a purpose-built service that helps healthcare and life sciences organizations store, query, and analyze genomic, transcriptomic, and other omics data.

AWS HealthOmics and NVIDIA Healthcare teams are also working together to launch generative AI microservices to advance drug discovery, medtech, and digital health — delivering a new catalog of GPU-accelerated cloud endpoints for biology, chemistry, imaging and healthcare data so healthcare enterprises can take advantage of the latest advances in generative AI on AWS.

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 infrastructure 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, media, and application development, deployment, and management from 105 Availability Zones within 33 geographic regions, with announced plans for 18 more Availability Zones and six more AWS Regions in Malaysia, Mexico, New Zealand, the Kingdom of Saudi Arabia, Thailand, and the AWS European Sovereign Cloud. 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.

About Amazon
Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking. Amazon strives to be Earth's Most Customer-Centric Company, Earth's Best Employer, and Earth's Safest Place to Work. Customer reviews, 1-Click shopping, personalized recommendations, Prime, Fulfillment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa,

Just Walk Out technology, Amazon Studios, and The Climate Pledge are some of the things pioneered by Amazon. For more information, visit amazon.com/about and follow @AmazonNews.

NVIDIA Forward-Looking Statements

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 Grace Blackwell Superchip, NVIDIA DGX Cloud, NVIDIA Omniverse Cloud APIs, NVIDIA AI and Accelerated Computing Platforms, and NVIDIA Generative AI Microservices; the benefits and impact of NVIDIA's collaboration with Microsoft, and the features and availability of its services and offerings; AI transforming our daily lives, the way we work and opening up a world of new opportunities; and building a future that unlocks the promise of AI for customers and brings transformative solutions to the world through NVIDIA's continued collaboration with Microsoft 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 NVIDIA's 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 NVIDIA 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.

Amazon Media Hotline
Amazon.com, Inc.
amazon-pr@amazon.com
Natalie Hereth
NVIDIA Corporation
nhereth@nvidia.com