AWS and NVIDIA Collaborate on Next-Generation Infrastructure for Training Large Machine Learning Models and Building Generative AI Applications

New Amazon EC2 P5 Instances Deployed in EC2 UltraClusters Are Fully Optimized to Harness NVIDIA Hopper GPUs for Accelerating Generative AI Training and Inference at Massive Scale

GTC—Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), and NVIDIA (NASDAQ: NVDA) today announced a multi-part collaboration focused on building out the world's most scalable, on-demand artificial intelligence (AI) infrastructure optimized for training increasingly complex large language models (LLMs) and developing generative AI applications.

The joint work features next-generation Amazon Elastic Compute Cloud (Amazon EC2) P5 instances powered by NVIDIA H100 Tensor Core GPUs and AWS's state-of-the-art networking and scalability that will deliver up to 20 exaFLOPS of compute performance for building and training the largest deep learning models. P5 instances will be the first GPU-based instance to take advantage of AWS's second-generation Elastic Fabric Adapter (EFA) networking, which provides 3,200 Gbps of low-latency, high bandwidth networking throughput, enabling customers to scale up to 20,000 H100 GPUs in EC2 UltraClusters for on-demand access to supercomputer-class performance for AI.

“AWS and NVIDIA have collaborated for more than 12 years to deliver large-scale, cost-effective GPU-based solutions on demand for various applications such as AI/ML, graphics, gaming, and HPC,” said Adam Selipsky, CEO at AWS. “AWS has unmatched experience delivering GPU-based instances that have pushed the scalability envelope with each successive generation, with many customers scaling machine learning training workloads to more than 10,000 GPUs today. With second-generation EFA, customers will be able to scale their P5 instances to over 20,000 NVIDIA H100 GPUs, bringing supercomputer capabilities on demand to customers ranging from startups to large enterprises.”

“Accelerated computing and AI have arrived, and just in time. Accelerated computing provides step-function speed-ups while driving down cost and power as enterprises strive to do more with less. Generative AI has awakened companies to reimagine their products and business models and to be the disruptor and not the disrupted,” said Jensen Huang, founder and CEO of NVIDIA. “AWS is a long-time partner and was the first cloud service provider to offer NVIDIA GPUs. We are thrilled to combine our expertise, scale, and reach to help customers harness accelerated computing and generative AI to engage the enormous opportunities ahead.”

New Supercomputing Clusters

New P5 instances are built on more than a decade of collaboration between AWS and NVIDIA delivering the AI and HPC infrastructure and build on four previous collaborations across P2, P3, P3dn, and P4d(e) instances. P5 instances are the fifth generation of AWS offerings powered by NVIDIA GPUs and come almost 13 years after its initial deployment of NVIDIA GPUs, beginning with CG1 instances.

P5 instances are ideal for training and running inference for increasingly complex LLMs and computer vision models behind the most-demanding and compute-intensive generative AI applications, including question answering, code generation, video and image generation, speech recognition, and more.

Specifically built for both enterprises and startups racing to bring AI-fueled innovation to market in a scalable and secure way, P5 instances feature eight NVIDIA H100 GPUs capable of 16 petaFLOPs of mixed-precision performance, 640 GB of high-bandwidth memory, and 3,200 Gbps networking connectivity (8x more than the previous generation) in a single EC2 instance. The increased performance of P5 instances accelerates the time-to-train machine learning (ML) models by up to 6x (reducing training time from days to hours), and the additional GPU memory helps customers train larger, more complex models. P5 instances are expected to lower the cost to train ML models by up to 40% over the previous generation, providing customers greater efficiency over less flexible cloud offerings or expensive on-premises systems.

Amazon EC2 P5 instances are deployed in hyperscale clusters called EC2 UltraClusters that are comprised of the highest performance compute, networking, and storage in the cloud. Each EC2 UltraCluster is one of the most powerful supercomputers in the world, enabling customers to run their most complex multi-node ML training and distributed HPC workloads. They feature petabit-scale non-blocking networking, powered by AWS EFA, a network interface for Amazon EC2 instances that enables customers to run applications requiring high levels of inter-node communications at scale on AWS. EFA's custom-built operating system (OS) bypass hardware interface and integration with NVIDIA GPUDirect RDMA enhances the performance of inter-instance communications by lowering latency and increasing bandwidth utilization, which is critical to scaling training of deep learning models across hundreds of P5 nodes. With P5 instances and EFA, ML applications can use NVIDIA Collective Communications Library (NCCL) to scale up to 20,000 H100 GPUs. As a result, customers get the application performance of on-premises HPC clusters with the on-demand elasticity and flexibility of AWS. On top of these cutting-edge computing capabilities, customers can use the industry's broadest and deepest portfolio of services such as Amazon S3 for object storage, Amazon FSx for high-performance file systems, and Amazon SageMaker for building, training, and deploying deep learning applications. P5 instances will be available in the coming weeks in limited preview. To request access, visit https://pages.awscloud.com/EC2-P5-Interest.html.

With the new EC2 P5 instances, customers like Anthropic, Cohere, Hugging Face, Pinterest, and Stability AI will be able to build and train the largest ML models at scale. The collaboration through additional generations of EC2 instances will help startups, enterprises, and researchers seamlessly scale to meet their ML needs.

Anthropic builds reliable, interpretable, and steerable AI systems that will have many opportunities to create value commercially and for public benefit. "At Anthropic, we are working to build reliable, interpretable, and steerable AI systems. While the large, general AI systems of today can have significant benefits, they can also be unpredictable, unreliable, and opaque. Our goal is to make progress on these issues and deploy systems that people find useful," said Tom Brown, co-founder of Anthropic. "Our organization is one of the few in the world that is building foundational models in deep learning research. These models are highly complex, and to develop and train these cutting-edge models, we need to distribute them efficiently across large clusters of GPUs. We are using Amazon EC2 P4 instances extensively today, and we are excited about the upcoming launch of P5 instances. We expect them to deliver substantial price-performance benefits over P4d instances, and they'll be available at the massive scale required for building next-generation large language models and related products."

Cohere, a leading pioneer in language AI, empowers every developer and enterprise to build incredible products with world-leading natural language processing (NLP) technology while keeping their data private and secure. "Cohere leads the charge in helping every enterprise harness the power of language AI to explore, generate, search for, and act upon information in a natural and intuitive manner, deploying across multiple cloud platforms in the data environment that works best for each customer," said Aidan Gomez, CEO at Cohere. "NVIDIA H100-powered Amazon EC2 P5 instances will unleash the ability of businesses to create, grow, and scale faster with its computing power combined with Cohere's state-of-the-art LLM and generative AI capabilities."

Hugging Face is on a mission to democratize good machine learning. "As the fastest growing open source community for machine learning, we now provide over 150,000 pre-trained models and 25,000 datasets on our platform for NLP, computer vision, biology, reinforcement learning, and more," said Julien Chaumond, CTO and co-founder at Hugging Face. "With significant advances in large language models and generative AI, we're working with AWS to build and contribute the open source models of tomorrow. We're looking forward to using Amazon EC2 P5 instances via Amazon SageMaker at scale in UltraClusters with EFA to accelerate the delivery of new foundation AI models for everyone."

Today, more than 450 million people around the world use Pinterest as a visual inspiration platform to shop for products personalized to their taste, find ideas to do offline, and discover the most inspiring creators. "We use deep learning extensively across our platform for use-cases such as labeling and categorizing billions of photos that are uploaded to our platform, and visual search that provides our users the ability to go from inspiration to action," said David Chaiken, Chief Architect at Pinterest. "We have built and deployed these use-cases by leveraging AWS GPU instances such as P3 and the latest P4d instances. We are looking forward to using Amazon EC2 P5 instances featuring H100 GPUs, EFA and Ultraclusters to accelerate our product development and bring new Empathetic AI-based experiences to our customers."

As the leader in multimodal, open-source AI model development and deployment, Stability AI collaborates with public- and private-sector partners to bring this next-generation infrastructure to a global audience. "At Stability AI, our goal is to maximize the accessibility of modern AI to inspire global creativity and innovation," said Emad Mostaque, CEO of Stability AI. "We initially partnered with AWS in 2021 to build Stable Diffusion, a latent text-to-image diffusion model, using Amazon EC2 P4d instances that we employed at scale to accelerate model training time from months to weeks. As we work on our next generation of open-source generative AI models and expand into new modalities, we are excited to use Amazon EC2 P5 instances in second-generation EC2 UltraClusters. We expect P5 instances will further improve our model training time by up to 4x, enabling us to deliver breakthrough AI more quickly and at a lower cost."

New Server Designs for Scalable, Efficient AI

Leading up to the release of H100, NVIDIA and AWS engineering teams with expertise in thermal, electrical, and mechanical fields have collaborated to design servers to harness GPUs to deliver AI at scale, with a focus on energy efficiency in AWS infrastructure. GPUs are typically 20x more energy efficient than CPUs for certain AI workloads, with the H100 up to 300x more efficient for LLMs than CPUs.

The joint work has included developing a system thermal design, integrated security and system management, security with the AWS Nitro hardware accelerated hypervisor, and NVIDIA GPUDirect™ optimizations for AWS custom-EFA network fabric.

Building on AWS and NVIDIA's work focused on server optimization, the companies have begun collaborating on future server designs to increase the scaling efficiency with subsequent-generation system designs, cooling technologies, and network scalability.

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 200 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 99 Availability Zones within 31 geographic regions, with announced plans for 15 more Availability Zones and five more AWS Regions in Canada, Israel, 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, availability and features of NVIDIA's collaboration with AWS and Amazon EC2 P5 instances; the benefits, impact, performance, features and availability of NVIDIA's products and technologies, including NVIDIA Hopper GPUs, NVIDIA H100 Tensor Core GPUs, NVIDIA GPUDirect RDMA, NVIDIA Collective Communications Library and NVIDIA GPUDirect optimizations; the benefits, impact and performance of Amazon EC2 P5 instances as used by third parties, including Pinterest, Stability AI, Cohere and Hugging Face; NVIDIA and AWS collaborating on future server designs to increase the scaling efficiency with subsequent-generation system designs, cooling technologies, and network scalability 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.

Allie Courtney
NVIDIA Corporation
+1-408-706-8995
acourtney@nvidia.com
Amazon Media Hotline
Amazon.com, Inc.
amazon-pr@amazon.com