NVIDIA H100 GPUs Now Available on AWS Cloud

New Amazon EC2 P5 instance uses NVIDIA’s latest accelerators to deliver industry-leading performance for generative AI and more.

Author: Dave Salvator

AWS users can now access the leading performance demonstrated in industry benchmarks of AI training and inference .

The cloud giant officially switched on a new Amazon EC2 P5 instance powered by NVIDIA H100 Tensor Core GPUs . The service lets users scale generative AI , high performance computing (HPC) and other applications with a click from a browser.

The news comes in the wake of AI’s iPhone moment. Developers and researchers are using large language models (LLMs) to uncover new applications for AI almost daily. Bringing these new use cases to market requires the efficiency of accelerated computing .

The NVIDIA H100 GPU delivers supercomputing-class performance through architectural innovations including fourth-generation Tensor Cores , a new Transformer Engine for accelerating LLMs and the latest NVLink technology that lets GPUs talk to each other at 900GB/sec.

Amazon EC2 P5 instances are ideal for training and running inference for increasingly complex LLMs and computer vision models. These neural networks drive the most demanding and compute-intensive generative AI applications, including question answering, code generation, video and image generation, speech recognition and more.

P5 instances can be deployed in hyperscale clusters, called EC2 UltraClusters, made up of high-performance compute, networking and storage in the cloud. Each EC2 UltraCluster is a powerful supercomputer, enabling customers to run their most complex AI training and distributed HPC workloads across multiple systems.

So customers can run at scale applications that require high levels of communications between compute nodes, the P5 instance sports petabit-scale non-blocking networks, powered by AWS EFA, a 3,200 Gbps network interface for Amazon EC2 instances.

With P5 instances, machine learning applications can use the NVIDIA Collective Communications Library to employ as many as 20,000 H100 GPUs.

NVIDIA AI Enterprise helps users make the most of P5 instances with a full-stack suite of software that includes more than 100 frameworks, pretrained models , AI workflows and tools to tune AI infrastructure.

Designed to streamline the development and deployment of AI applications, NVIDIA AI Enterprise addresses the complexities of building and maintaining a high-performance, secure, cloud-native AI software platform. Available in the AWS Marketplace , it offers continuous security monitoring, regular and timely patching of common vulnerabilities and exposures, API stability, and enterprise support as well as access to NVIDIA AI experts.

NVIDIA and AWS have collaborated for more than a dozen years to bring GPU acceleration to the cloud. The new P5 instances, the latest example of that collaboration, represents a major step forward to deliver the cutting-edge performance that enables developers to invent the next generation of AI.

Here are some examples of what customers are already saying:

Anthropic builds reliable, interpretable and steerable AI systems that will have many opportunities to create value commercially and for public benefit.

“While the large, general AI systems of today can have significant benefits, they can also be unpredictable, unreliable and opaque, so our goal is to make progress on these issues and deploy systems that people find useful,” said Tom Brown, co-founder of Anthropic. “We expect P5 instances to deliver substantial price-performance benefits over P4d instances, and they’ll be available at the massive scale required for building next-generation LLMs and related products.”

Cohere, a leading pioneer in language AI, empowers every developer and enterprise to build 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 of 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.”

For its part, 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 pretrained models and 25,000 datasets on our platform for NLP, computer vision, biology, reinforcement learning and more,” said Julien Chaumond, chief technology officer and co-founder of Hugging Face. “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 and discover 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’re looking forward to using Amazon EC2 P5 instances featuring NVIDIA H100 GPUs, AWS EFA and UltraClusters to accelerate our product development and bring new empathetic AI-based experiences to our customers.”

Learn more about new AWS P5 instances powered by NVIDIA H100 .

Original URL: https://blogs.nvidia.com/blog/2023/07/26/aws-cloud-h100/