



AWS and NVIDIA to Deliver 2 Million Additional GPUs and Next-Generation Infrastructure for Agentic and Physical AI

Companies Deepen Integration Across the AI Stack, Bringing NVIDIA Vera CPUs, Advanced Networking, Nemotron Open Models and Physical AI Technologies to AWS as Customer Demand Accelerates

Amazon Web Services (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), and NVIDIA (NASDAQ: NVDA) today announced a major expansion of their strategic collaboration to meet surging global demand for AI infrastructure as demand continues to accelerate. Building on already-rapid customer adoption of NVIDIA-accelerated compute on AWS, the companies plan to deploy 2 million additional NVIDIA GPUs across AWS's global infrastructure and deepen their work together across AI factories, CPUs, networking, open models, data processing and robotics, delivering co-engineered AI solutions that enable customers to accelerate AI development and deployment at unprecedented scale.

AI workloads are scaling at a swift pace, from how models are trained and run, to how data is processed, indexed and used to power intelligent applications. Customers are moving from pilot to production and scaling workloads across agentic AI, scientific discovery, enterprise automation and robotics. They need broader model choice, faster data pipelines and new capabilities for emerging use cases like physical AI. They also need confidence that the underlying infrastructure can keep pace with their own ability to innovate while maintaining the highest level of security and reliability for mission-critical workloads.

To meet this surging demand from frontier labs, global enterprises, startups and governments, AWS and NVIDIA are building on 16 years of joint innovation to expand AI compute capacity and bring new co-engineered solutions to customers faster. As part of the expanded collaboration, the companies are working to:

- Deploy 2 million additional NVIDIA GPUs across AWS's global infrastructure in 2027-2028
- Bring NVIDIA Vera CPU-based infrastructure to AWS
- Extend NVIDIA NVLink Fusion™ with custom NVIDIA high-bandwidth memory (NVHBM)
- Build AI factories for the U.S. government, including 100,000 GPUs on secure AWS infrastructure for running federal and national-security workloads
- Integrate the NVIDIA platform with the AWS Nitro System and Elastic Fabric Adapter (EFA) for enhanced security and reliability
- Continue to support NVIDIA Nemotron™ open models on Amazon Bedrock and Amazon SageMaker, giving customers more open model choice
- Accelerate data processing and vector indexing on Amazon EMR and Amazon OpenSearch with NVIDIA cuDF and cuVS CUDA-X™ libraries for faster, more cost-efficient analytics and AI applications
- Further advance robotics workloads through Amazon Robotics' adoption of NVIDIA's physical AI platform, speeding innovation in warehouse automation and next-generation robots

"Customers want the freedom to choose the best tools for their AI workloads, and they want confidence that everything works seamlessly together," said Matt Garman, CEO of AWS. "That's why we've invested deeply with NVIDIA to make AWS the best place to run NVIDIA AI technologies, optimizing performance across our infrastructure from networking and security to deployment. This expanded collaboration gives frontier labs, enterprises and governments even more ways to build and deploy AI on AWS."

"NVIDIA and AWS have built one of the great growth engines of the AI era, and demand is running ahead of every forecast," said Jensen Huang, founder and CEO of NVIDIA. "For 16 years, we have scaled NVIDIA computing in the cloud together. Now, we are expanding our partnership across the full stack -- GPUs, CPUs, networking, open models and software -- to make agentic and physical AI real at an unprecedented pace and scale that only AWS and NVIDIA can deliver. This expansion reflects customers' demand for NVIDIA's platform on AWS."

Massive Expansion of AI Compute Capacity

AWS offers the widest range of GPU-based instances of any cloud provider to power a diverse set of AI and machine learning workloads. At NVIDIA GTC 2026, AWS [announced plans](#) to add more than 1 million NVIDIA GPUs starting in 2026. Since then, demand has exceeded those expectations. AWS plans to deploy an additional 2 million NVIDIA Blackwell Ultra, Rubin and Rubin Ultra GPUs in 2027-2028 across AWS Global Infrastructure, including AI factories. This additional capacity will help power customer workloads ranging from agentic AI and scientific discovery to enterprise automation and physical AI. In addition, AWS will expand NVIDIA Blackwell capacity, including NVIDIA RTX PRO™ 4500 Blackwell Server Edition GPUs for [Amazon EC2 G7 instances](#). G7 instances deliver 4.6x AI inference performance and 2.1x graphics performance

compared to previous-generation G6 instances. AWS is the first major cloud provider to offer compute instances accelerated by RTX PRO 4500. AWS and NVIDIA are also collaborating on NVIDIA Spectrum™ networking to further optimize network performance for large-scale AI training workloads across GPU clusters.

Support for NVIDIA Vera CPUs on AWS

AWS and NVIDIA are working to bring Vera CPU-based infrastructure to AWS, providing an additional option to support agentic AI workloads that require high-performance CPU compute alongside accelerated infrastructure. Purpose-built for the next generation of AI, Vera complements AWS's strategy to offer the broadest choice of compute -- from AWS custom silicon to the latest accelerators and CPUs from partners.

Heterogeneous AI Infrastructure Using NVIDIA NVLink Fusion With NVHBM

At re:Invent 2025, AWS announced support for NVIDIA NVLink Fusion high-speed chip interconnect technology in next-generation Trainium chips. NVIDIA and Amazon's Annapurna Labs are expanding that support to work on [NVIDIA's new custom high-bandwidth memory \(NVHBM\) technology](#), in partnership with memory suppliers, which would give Trainium access to faster, more power-efficient memory. Combined with NVLink Fusion, Annapurna Labs can now tap NVIDIA's custom memory technology and scale-up architecture to enhance performance and efficiency for AI workloads while seamlessly integrating Trainium and GPUs within a common rack-scale architecture.

Powering Federal AI at the Highest Levels of Security

Government agencies need secure AI infrastructure to keep pace with the demands of national security. AWS and NVIDIA plan to build AI factories for the U.S. government, delivering NVIDIA's AI stack, including plans to deliver 100,000 GPUs on AWS's secure infrastructure for federal and national-security workloads. This collaboration puts AWS and NVIDIA at the center of federal AI advancement for national security, enabling government agencies to deploy AI at scale for workloads classified at Impact Level 6 (IL6) and above.

These new commitments build on a foundation of deep technical integrations between AWS and NVIDIA that are already delivering results for customers today, including:

- **Enhanced security and reliability with AWS Nitro System and EFA** -- Across this expanded collaboration, all NVIDIA GPU-based and Trainium-based EC2 instances -- including those leveraging NVLink Fusion -- are built on the AWS Nitro System and interconnected through EFA. Both GPU-accelerated and Trainium-based EC2 instances will continue to be built on the Nitro System and scaled out through EFA. Together, Nitro and EFA help ensure that as AWS expands its NVIDIA GPU fleet and integrates new interconnect technologies, customers retain the security, reliability and network performance they depend on for production AI workloads at scale.
- **NVIDIA Nemotron models on AWS** -- As part of AWS's commitment to offering customers the broadest choice of AI models, NVIDIA's Nemotron family of open models is available on Amazon Bedrock as fully managed, serverless models and on Amazon SageMaker for customers who want to deploy and fine-tune on their own infrastructure. This integration gives customers access to NVIDIA's latest open models with the security, scalability and operational tooling of AWS.
- **GPU-accelerated data processing and vector indexing** -- As data volumes grow, workloads such as feature engineering, large-scale ETL and real-time analytics require increasingly faster processing. AWS and NVIDIA are collaborating to deliver GPU-accelerated data processing on Amazon EMR using Amazon EC2 G7 instances and the NVIDIA cuDF library, delivering up to 3.7x faster processing speeds and a 30% better price performance compared to CPU-based configurations. Separately, as AI applications, retrieval-augmented generation pipelines and semantic search push vector databases to billions of records, index building and tuning becomes a bottleneck. GPU-accelerated vector indexing on Amazon OpenSearch Service offloads index construction onto dedicated GPUs, delivering up to 9x faster vector indexing at a quarter of the cost -- available across both managed clusters and Amazon OpenSearch Serverless.
- **Physical AI for robotics** -- Amazon Robotics is collaborating with NVIDIA to accelerate the development of next-generation robots integrating NVIDIA's full-stack physical AI platform, including the NVIDIA Jetson™ platform, NVIDIA Omniverse™ libraries and the NVIDIA Isaac™ open robotics development platform. The collaboration spans simulation, synthetic data generation, robot training, route optimization, functional safety and real-to-sim validation -- all running on GPU-accelerated Amazon EC2 instances. Together, AWS and NVIDIA are helping advance the capabilities that robotics workloads require at scale: massive simulation, diverse training data and continuous real-world validation.

About Amazon Web Services

Amazon Web Services (AWS) is guided by customer obsession, pace of innovation, commitment to operational excellence, and long-term thinking. By democratizing technology for nearly two decades and making cloud computing and generative AI accessible to organizations of every size and industry, AWS has built one of the fastest-growing enterprise technology businesses in history. Millions of customers trust AWS to accelerate innovation, transform their businesses, and shape the future. With the most comprehensive AI capabilities and global infrastructure footprint, AWS empowers builders to turn big ideas into reality. Learn more at aws.amazon.com and follow @AWSNewsroom.

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

NVIDIA Forward-Looking Statements

Certain statements in this press release including, but not limited to, statements as to: NVIDIA and AWS expanding partnership across the full stack—GPUs, CPUs, networking, open models and software—to make agentic and physical AI real at an unprecedented pace and scale that only AWS and NVIDIA can deliver; expectations with respect to growth, performance, availability, demand, and benefits of NVIDIA's products, services and technologies, and related trends and drivers; expectations with respect to NVIDIA's third party arrangements, including with AWS; expectations with respect to technology developments, and related trends and drivers; projected market growth and trends; expectations with respect to AI and related industries; and other statements that are not historical facts are forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, which are subject to the "safe harbor" created by those sections based on management's beliefs and assumptions and on information currently available to management and 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 and political 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 products and technologies; market acceptance of NVIDIA's products or NVIDIA's 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; NVIDIA's ability to realize the potential benefits of business investments or acquisitions; and changes in applicable laws and regulations, 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.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA-X, Nemotron, NVIDIA Isaac, NVIDIA Jetson, NVIDIA Omniverse, NVIDIA RTX PRO, NVIDIA Spectrum and NVLink Fusion 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.

Corporate Communications

NVIDIA Corporation

press@nvidia.com

Media Hotline

Amazon.com, Inc.

amazon-pr@amazon.com