

NVIDIA Kicks Off the Next Generation of AI With Rubin — Six New Chips, One Incredible AI Supercomputer

Extreme Codesign Across NVIDIA Vera CPU, Rubin GPU, NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU and Spectrum-6 Ethernet Switch Slashes Training Time and Inference Token Generation Cost

News Summary:

- The Rubin platform harnesses extreme codesign across hardware and software to deliver up to 10x reduction in inference token cost and 4x reduction in number of GPUs to train MoE models, compared with the NVIDIA Blackwell platform.
- NVIDIA Spectrum-X Ethernet Photonics switch systems deliver 5x improved power efficiency and uptime.
- New NVIDIA Inference Context Memory Storage Platform with NVIDIA BlueField-4 storage processor to accelerate agentic AI reasoning.
- Microsoft's next-generation Fairwater AI superfactories — featuring NVIDIA Vera Rubin NVL72 rack-scale systems — will scale to hundreds of thousands of NVIDIA Vera Rubin Superchips.
- CoreWeave among first to offer NVIDIA Rubin, operated through CoreWeave Mission Control for flexibility and performance.
- Expanded collaboration with Red Hat to deliver a complete AI stack optimized for the Rubin platform with Red Hat Enterprise Linux, Red Hat OpenShift and Red Hat AI.

CES—NVIDIA today kickstarted the next generation of AI with the launch of the [NVIDIA Rubin platform](#), comprising six new chips designed to deliver one incredible AI supercomputer. NVIDIA Rubin sets a new standard for building, deploying and securing the world's largest and most advanced AI systems at the lowest cost to accelerate mainstream AI adoption.

The Rubin platform uses extreme codesign across the six chips — the [NVIDIA Vera CPU](#), NVIDIA Rubin GPU, [NVIDIA NVLink™ 6 Switch](#), [NVIDIA ConnectX@-9 SuperNIC](#), [NVIDIA BlueField@-4 DPU](#) and [NVIDIA Spectrum™-6 Ethernet Switch](#) — to slash training time and inference token costs.

“Rubin arrives at exactly the right moment, as AI computing demand for both training and inference is going through the roof,” said Jensen Huang, founder and CEO of NVIDIA. “With our annual cadence of delivering a new generation of AI supercomputers — and extreme codesign across six new chips — Rubin takes a giant leap toward the next frontier of AI.”

Named for Vera Florence Cooper Rubin — the trailblazing American astronomer whose discoveries transformed humanity's understanding of the universe — the Rubin platform features the [NVIDIA Vera Rubin NVL72](#) rack-scale solution and the [NVIDIA HGX Rubin NVL8](#) system.

The Rubin platform introduces five innovations, including the latest generations of NVIDIA NVLink interconnect technology, Transformer Engine, Confidential Computing and RAS Engine, as well as the NVIDIA Vera CPU. These breakthroughs will accelerate agentic AI, advanced reasoning and massive-scale [mixture-of-experts](#) (MoE) model inference at up to 10x lower cost per token of the NVIDIA Blackwell platform. Compared with its predecessor, the NVIDIA Rubin platform trains MoE models with 4x fewer GPUs to accelerate AI adoption.

Broad Ecosystem Support

Among the world's leading AI labs, cloud service providers, computer makers and startups expected to adopt Rubin are Amazon Web Services (AWS), Anthropic, Black Forest Labs, Cisco, Cohere, CoreWeave, Cursor, [Dell Technologies](#), Google, Harvey, [HPE](#), Lambda, [Lenovo](#), Meta, Microsoft, Mistral AI, Nebius, Nscale, OpenAI, OpenEvidence, Oracle Cloud Infrastructure (OCI), Perplexity, Runway, [Supermicro](#), Thinking Machines Lab and xAI.

Sam Altman, CEO of OpenAI: “Intelligence scales with compute. When we add more compute, models get more capable, solve harder problems and make a bigger impact for people. The NVIDIA Rubin platform helps us keep scaling this progress so advanced intelligence benefits everyone.”

Dario Amodei, cofounder and CEO of Anthropic: “The efficiency gains in the NVIDIA Rubin platform represent the kind of infrastructure progress that enables longer memory, better reasoning and more reliable outputs. Our collaboration with NVIDIA helps power our safety research and our frontier models.”

Mark Zuckerberg, founder and CEO of Meta: “NVIDIA's Rubin platform promises to deliver the step-change in performance and efficiency required to deploy the most advanced models to billions of people.”

Elon Musk, founder and CEO of xAI: “ NVIDIA Rubin will be a rocket engine for AI. If you want to train and deploy frontier

models at scale, this is the infrastructure you use — and Rubin will remind the world that NVIDIA is the gold standard. ”

Satya Nadella, executive chairman and CEO of Microsoft: “We are building the world’s most powerful AI superfactories to serve any workload, anywhere, with maximum performance and efficiency. With the addition of NVIDIA Vera Rubin GPUs, we will empower developers and organizations to create, reason and scale in entirely new ways.”

Mike Intrator, cofounder and CEO of CoreWeave: “We built CoreWeave to help pioneers accelerate their innovations with the unmatched performance of our purpose-built AI platform, matching the right technology to the right workloads as they evolve. The NVIDIA Rubin platform represents an important advancement for reasoning, agentic and large-scale inference workloads, and we’re excited to add it to our platform. With CoreWeave Mission Control as the operating standard, we can integrate new capabilities quickly and run them reliably at production scale, working in close partnership with NVIDIA.”

Matt Garman, CEO of AWS: “AWS and NVIDIA have been driving cloud AI innovation together for more than 15 years. The NVIDIA Rubin platform on AWS represents our continued commitment to delivering cutting-edge AI infrastructure that gives customers unmatched choice and flexibility. By combining NVIDIA’s advanced AI technology with AWS’s proven scale, security and comprehensive AI services, customers can build, train and deploy their most demanding AI applications faster and more cost effectively — accelerating their path from experimentation to production at any scale.”

Sundar Pichai, CEO of Google and Alphabet: “We are proud of our deep and long-standing relationship with NVIDIA. To meet the substantial customer demand we see for NVIDIA GPUs, we are focused on providing the best possible environment for their hardware on Google Cloud. Our collaboration will continue as we bring the impressive capabilities of the Rubin platform to our customers, offering them the scale and performance needed to advance the boundaries of AI.”

Clay Magouyrk, CEO of Oracle: “Oracle Cloud Infrastructure is a hyperscale cloud built for the highest performance, and together with NVIDIA, we’re pushing the boundaries of what customers can build and scale with AI. With gigascale AI factories powered by the NVIDIA Vera Rubin architecture, OCI is giving customers the infrastructure foundation they need to push the limits of model training, inference and real-world AI impact.”

Michael Dell, chairman and CEO of Dell Technologies: “The NVIDIA Rubin platform represents a major leap forward in AI infrastructure. By integrating Rubin into the Dell AI Factory with NVIDIA, we’re building infrastructure that can handle massive token volumes and multistep reasoning while delivering the performance and resiliency that enterprises and neoclouds need to deploy AI at scale.”

Antonio Neri, president and CEO of HPE: “AI is reshaping not just workloads but the very foundations of IT, requiring us to reimagine every layer of infrastructure from the network to the compute. With the NVIDIA Vera Rubin platform, HPE is building the next generation of secure, AI-native infrastructure, turning data into intelligence and enabling enterprises to become true AI factories.”

Yuanqing Yang, chairman and CEO of Lenovo: “Lenovo is embracing the next-generation NVIDIA Rubin platform, leveraging our Neptune liquid-cooling solution as well as our global scale, manufacturing efficiency and service reach, to help enterprises build AI factories that serve as intelligent, accelerated engines for insight and innovation. Together, we’re architecting an AI-driven future where efficient, secure AI becomes the standard for every organization.”

Engineered to Scale Intelligence

Agentic AI and reasoning models, along with state-of-the-art video generation workloads, are redefining the limits of computation. Multistep problem-solving requires models to process, reason and act across long sequences of tokens. Designed to serve the demands of complex AI workloads, the Rubin platform’s five groundbreaking technologies include:

- **Sixth-Generation NVIDIA NVLink:** Delivers the fast, seamless GPU-to-GPU communication required for today’s massive MoE models. Each GPU offers 3.6TB/s of bandwidth, while the Vera Rubin NVL72 rack provides 260TB/s — more bandwidth than the entire internet. With built-in, in-network compute to speed collective operations, as well as new features for enhanced serviceability and resiliency, NVIDIA NVLink 6 switch enables faster, more efficient AI training and inference at scale.
- **NVIDIA Vera CPU:** Designed for agentic reasoning, NVIDIA Vera is the most power-efficient CPU for large-scale AI factories. The NVIDIA CPU is built with 88 NVIDIA custom Olympus cores, full Armv9.2 compatibility and ultrafast NVLink-C2C connectivity. Vera delivers exceptional performance, bandwidth and industry-leading efficiency to support a full range of modern data center workloads.
- **NVIDIA Rubin GPU:** Featuring a third-generation Transformer Engine with hardware-accelerated adaptive compression, Rubin GPU delivers 50 petaflops of NVFP4 compute for AI inference.
- **Third-Generation NVIDIA Confidential Computing:** Vera Rubin NVL72 is the first rack-scale platform to deliver [NVIDIA Confidential Computing](#) — which maintains data security across CPU, GPU and NVLink domains — protecting the world’s largest proprietary models, training and inference workloads.
- **Second-Generation RAS Engine:** The Rubin platform — spanning GPU, CPU and NVLink — features real-time health checks, fault tolerance and proactive maintenance to maximize system productivity. The rack’s modular, cable-free tray design enables up to 18x faster assembly and servicing than Blackwell.

AI-Native Storage and Secure, Software-Defined Infrastructure

NVIDIA Rubin introduces [NVIDIA Inference Context Memory Storage Platform](#), a new class of AI-native storage infrastructure designed to scale inference context at gigascale.

Powered by NVIDIA BlueField-4, the platform enables efficient sharing and reuse of key-value cache data across AI infrastructure, improving responsiveness and throughput while enabling predictable, power-efficient scaling of agentic AI.

As AI factories increasingly adopt bare-metal and multi-tenant deployment models, maintaining strong infrastructure control and isolation becomes essential.

BlueField-4 also introduces Advanced Secure Trusted Resource Architecture, or ASTRA, a system-level trust architecture that gives AI infrastructure builders a single, trusted control point to securely provision, isolate and operate large-scale AI environments without compromising performance.

With AI applications evolving toward multi-turn agentic reasoning, AI-native organizations must manage and share far larger volumes of inference context across users, sessions and services.

Different Forms for Different Workloads

NVIDIA Vera Rubin NVL72 offers a unified, secure system that combines 72 NVIDIA Rubin GPUs, 36 NVIDIA Vera CPUs, NVIDIA NVLink 6, [NVIDIA ConnectX-9 SuperNICs](#) and [NVIDIA BlueField-4 DPUs](#).

NVIDIA will also offer the NVIDIA HGX Rubin NVL8 platform, a server board that links eight Rubin GPUs through NVLink to support x86-based generative AI platforms. The HGX Rubin NVL8 platform accelerates training, inference and scientific computing for AI and high-performance computing workloads.

[NVIDIA DGX SuperPOD™](#) serves as a reference for deploying Rubin-based systems at scale, integrating either [NVIDIA DGX Vera Rubin NVL72](#) or [DGX Rubin NVL8 systems](#) with NVIDIA BlueField-4 DPUs, NVIDIA ConnectX-9 SuperNICs, NVIDIA InfiniBand networking and [NVIDIA Mission Control™](#) software.

Next-Generation Ethernet Networking

Advanced Ethernet networking and storage are components of AI infrastructure critical to keeping data centers running at full speed, improving performance and efficiency, and lowering costs.

[NVIDIA Spectrum-6 Ethernet](#) is the next generation of Ethernet for AI networking, built to scale Rubin-based AI factories with higher efficiency and greater resilience, and enabled by 200G SerDes communication circuitry, co-packaged optics and AI-optimized fabrics.

Built on the Spectrum-6 architecture, [Spectrum-X Ethernet Photonics co-packaged optical switch systems](#) deliver 10x greater reliability and 5x longer uptime for AI applications while achieving 5x better power efficiency, maximizing performance per watt compared with traditional methods. [Spectrum-XGS Ethernet technology](#), part of the [Spectrum-X Ethernet platform](#), enables facilities separated by hundreds of kilometers and more to function as a single AI environment.

Together, these innovations define the next generation of the NVIDIA Spectrum-X Ethernet platform, engineered with extreme codesign for Rubin to enable massive-scale AI factories and pave the way for future million-GPU environments.

Rubin Readiness

NVIDIA Rubin is in full production, and Rubin-based products will be available from partners the second half of 2026.

Among the first cloud providers to deploy Vera Rubin-based instances in 2026 will be AWS, Google Cloud, Microsoft and OCI, as well as [NVIDIA Cloud Partners](#) CoreWeave, Lambda, Nebius and Nscale.

[Microsoft](#) will deploy NVIDIA Vera Rubin NVL72 rack-scale systems as part of next-generation AI data centers, including future Fairwater AI superfactory sites.

Designed to deliver unprecedented efficiency and performance for training and inference workloads, the Rubin platform will provide the foundation for Microsoft's next-generation cloud AI capabilities. Microsoft Azure will offer a tightly optimized platform enabling customers to accelerate innovation across enterprise, research and consumer applications.

CoreWeave will integrate NVIDIA Rubin-based systems into its AI cloud platform beginning in the second half of 2026. CoreWeave is built to operate multiple architectures side by side, enabling customers to bring Rubin into their environments, where it will deliver the greatest impact across training, inference and agentic workloads.

Together with NVIDIA, CoreWeave will help AI pioneers take advantage of Rubin's advancements in reasoning and MoE models, while continuing to deliver the performance, operational reliability and scale required for production AI across the full lifecycle with [CoreWeave Mission Control](#).

In addition, Cisco, Dell, HPE, Lenovo and Supermicro are expected to deliver a wide range of servers based on Rubin products.

AI labs including Anthropic, Black Forest, Cohere, Cursor, Harvey, Meta, Mistral AI, OpenAI, OpenEvidence, Perplexity,

Runway, Thinking Machines Lab and xAI are looking to the NVIDIA Rubin platform to train larger, more capable models and to serve long-context, multimodal systems at lower latency and cost than with prior GPU generations.

Infrastructure software and storage partners AIC, [Canonical](#), Cloudian, [DDN](#), Dell, HPE, Hitachi Vantara, IBM, NetApp, [Nutanix](#), Pure Storage, Supermicro, [SUSE](#), VAST Data and [WEKA](#) are working with NVIDIA to design next-generation platforms for Rubin infrastructure.

The Rubin platform marks NVIDIA's third-generation rack-scale architecture, with more than 80 NVIDIA MGX™ ecosystem partners.

To unlock this density, Red Hat today [announced](#) an expanded collaboration with NVIDIA to deliver a complete AI stack optimized for the NVIDIA Rubin platform with Red Hat's hybrid cloud portfolio, including Red Hat Enterprise Linux, Red Hat OpenShift and Red Hat AI. These solutions are used by the vast majority of Fortune Global 500 companies.

Learn more by watching [NVIDIA Live at CES](#) and reading the ["Inside Vera Rubin" technical blog](#).

About NVIDIA

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

Certain statements in this press release including, but not limited to, statements as to: Rubin arriving at exactly the right moment; with our annual cadence of delivering a new generation of AI supercomputers — and extreme codesign across six new chips — Rubin taking a giant leap toward the next frontier of AI; the benefits, impact, performance, and availability of NVIDIA's products, services, and technologies; expectations with respect to NVIDIA's third party arrangements, including with its collaborators and partners; expectations with respect to technology developments; 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 product 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; 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.

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.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, ConnectX, NVIDIA DGX SuperPOD, NVIDIA MGX, NVIDIA Mission Control, NVIDIA Spectrum, NVIDIA Spectrum-X and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and 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