



NVIDIA Vera Rubin Opens Agentic AI Frontier

Seven New Chips in Full Production to Scale the World's Largest AI Factories With Configurable AI Infrastructure Optimized for Every Phase of AI, From Pretraining, Post-Training and Test-Time Scaling to Agentic Inference

News Summary:

The NVIDIA Vera Rubin platform is opening the next AI frontier with:

- Vera Rubin NVL72 GPU racks
- Vera CPU racks
- NVIDIA Groq 3 LPX inference accelerator racks
- NVIDIA BlueField-4 STX storage racks
- NVIDIA Spectrum-6 SPX Ethernet racks

GTC—NVIDIA today announced the NVIDIA Vera Rubin platform is opening the next frontier of agentic AI, with seven new chips now in full production to scale the world's largest AI factories.

The platform brings together 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, as well as the newly integrated NVIDIA Groq 3 LPU. Designed to operate together as one incredible AI supercomputer, the chips power every phase of AI — from massive-scale pretraining, post-training and test-time scaling to real-time agentic inference.

"Vera Rubin is a generational leap — seven breakthrough chips, five racks, one giant supercomputer — built to power every phase of AI," said Jensen Huang, founder and CEO of NVIDIA. "The agentic AI inflection point has arrived with Vera Rubin kicking off the greatest infrastructure buildout in history."

"Enterprises and developers are using Claude for increasingly complex reasoning, agentic workflows and mission-critical decisions. That demands infrastructure that can keep pace," said Dario Amodei, CEO and cofounder of Anthropic. "NVIDIA's Vera Rubin platform gives us the compute, networking and system design to keep delivering while advancing the safety and reliability our customers depend on."

"NVIDIA infrastructure is the foundation that lets us keep pushing the frontier of AI," said Sam Altman, CEO of OpenAI. "With NVIDIA Vera Rubin, we'll run more powerful models and agents at massive scale and deliver faster, more reliable systems to hundreds of millions of people."

Shift to POD-Scale Systems

AI infrastructure is rapidly evolving — from discrete chips and standalone servers to fully integrated rack-scale systems, POD-scale deployments, AI factories and sovereign AI. These advances are driving dramatic gains in performance, improving cost efficiency for organizations of all sizes and across industries — from startups and mid-sized businesses to public-private institutions and enterprises — while helping democratize access to AI and improving energy efficiency to power the world's most demanding workloads.

Through deep codesign across compute, networking and storage, supported by an ecosystem of more than 80 NVIDIA MGX ecosystem partners with a global supply chain, NVIDIA Vera Rubin offers the most extensive NVIDIA POD-scale platform — a supercomputer where multiple racks purpose-built for AI work together as one massive, coherent system.

NVIDIA Vera Rubin NVL72 Rack

Integrating 72 Rubin GPUs and 36 Vera CPUs connected by NVLink 6, along with ConnectX-9 SuperNICs and BlueField-4 DPUs, Vera Rubin NVL72 delivers breakthrough efficiency — training large mixture-of-experts models with one-fourth the number of GPUs compared with the NVIDIA Blackwell platform and achieving up to 10x higher inference throughput per watt at one-tenth the cost per token.

Designed for hyperscale AI factories worldwide, NVL72 scales seamlessly with NVIDIA Quantum-X800 InfiniBand and Spectrum-X Ethernet to sustain high utilization across massive GPU clusters while reducing time to train and total cost of ownership.

NVIDIA Vera CPU Rack

Reinforcement learning and agentic AI workloads rely on large numbers of CPU-based environments to test and validate the results generated by models running on GPU systems.

The NVIDIA Vera CPU Rack delivers dense, liquid-cooled infrastructure built on NVIDIA MGX, integrating 256 Vera CPUs to provide scalable, energy-efficient capacity with world-class single-threaded performance, unlocking agentic AI at scale.

Integrated with Spectrum-X Ethernet networking, Vera CPU racks keep CPU environments tightly synchronized across the AI factory. Together with GPU compute racks, they provide the CPU foundation for large-scale agentic AI and reinforcement learning — with Vera delivering results twice as efficiently and 50% faster than traditional CPUs.

NVIDIA Groq 3 LPX Rack

NVIDIA Groq 3 LPX marks a milestone in accelerated computing. Designed for the low-latency and large-context demands of agentic systems, LPX and Vera Rubin unite the extreme performance of both processors to deliver up to 35x higher inference throughput per megawatt and up to 10x more revenue opportunity for trillion-parameter models.

At scale, a fleet of LPUs function as a giant single processor for fast, deterministic inference acceleration. The LPX rack with 256 LPU processors features 128GB of on-chip SRAM and 640 TB/s of scale-up bandwidth. Deployed with Vera Rubin NVL72, Rubin GPUs and LPUs boost decode by jointly computing every layer of the AI model for every output token.

Optimized for trillion-parameter models and million-token context, the codesigned LPX architecture pairs with Vera Rubin to maximize efficiency across power, memory and compute. The additional throughput per watt and token performance unlocks a new tier of ultra-premium, trillion-parameter, million-context inference, expanding revenue opportunity for all AI providers. Fully liquid cooled and built on MGX infrastructure, LPX integrates seamlessly into next-generation Vera Rubin AI factories to be available in the second half of this year.

NVIDIA BlueField-4 STX Storage Rack

The [NVIDIA BlueField-4 STX](#) rack-scale system is an AI-native storage infrastructure that extends GPU memory seamlessly across the POD. Powered by BlueField-4 — combining the NVIDIA Vera CPU and NVIDIA ConnectX-9 SuperNIC — STX delivers a high-bandwidth shared layer optimized for storing and retrieving the massive key-value cache data generated by large language models and agentic AI workflows.

NVIDIA DOCA Memos™ — a new DOCA framework that supercharges BlueField-4 storage — enables dedicated KV cache storage processing to boost inference throughput by up to 5x while significantly improving power efficiency compared with general-purpose storage architectures. The result is POD-wide context that delivers faster multi-turn interactions with AI agents, more scalable AI services and higher overall infrastructure utilization.

“The NVIDIA BlueField-4 STX rack-scale context memory storage system will enable a critical performance boost needed to exponentially scale our agentic AI efforts,” said Timothée Lacroix, cofounder and chief technology officer of Mistral AI. “By delivering a new storage tier purpose-built for AI agents memory, STX is ideally positioned to ensure that our models can maintain coherence and speed when reasoning across massive datasets.”

NVIDIA Spectrum-6 SPX Ethernet Rack

Spectrum-6 SPX Ethernet is engineered to accelerate east-west traffic across AI factories. Configurable with either Spectrum-X Ethernet or NVIDIA Quantum-X800 InfiniBand switches, it delivers low-latency, high-throughput rack-to-rack connectivity at scale.

Spectrum-X Ethernet Photonics with co-packaged optics achieves up to 5x greater optical power efficiency and 10x higher resiliency compared with traditional pluggable transceivers.

Improving Resiliency and Energy Efficiency

NVIDIA, along with over 200 data center infrastructure partners, announced the NVIDIA DSX platform for Vera Rubin. This includes DSX Max-Q to enable dynamic power provisioning across the entire AI factory, resulting in the deployment of 30% more AI infrastructure within a fixed-power data center. The new DSX Flex software enables AI factories to be grid-flexible assets, unlocking 100 gigawatts of stranded grid power.

NVIDIA also today released the [Vera Rubin DSX AI Factory reference design](#), a blueprint for codesigned AI infrastructure that maximizes tokens per watt and overall goodput, improving system resiliency and accelerating time to first production.

By tightly integrating compute, networking, storage, power and cooling, the architecture increases energy efficiency and ensures AI factories can scale reliably under continuous, high-intensity workloads with maximum uptime.

Broad Ecosystem Support

Vera Rubin-based products will be available from partners starting the second half of this year. This includes leading cloud providers Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure, along with NVIDIA Cloud Partners CoreWeave, [Crusoe](#), Lambda, [Nebius](#), [Nscale](#) and Together AI.

Global system manufacturers Cisco, [Dell Technologies](#), [HPE](#), Lenovo and Supermicro are expected to deliver a wide range of servers based on Vera Rubin products, as well as [Aivres](#), [ASUS](#), Foxconn, GIGABYTE, Inventec, Pegatron, Quanta Cloud Technology (QCT), Wistron and [Wiwynn](#).

AI labs and frontier model developers including Anthropic, Meta, Mistral AI and OpenAI are looking to use the NVIDIA Vera 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.

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; Vera Rubin being a generational leap built to power every phase of AI; Vera Rubin kicking off the greatest infrastructure buildout in history; 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, CMX, ConnectX, DOCA Memos, MGX, 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
press@nvidia.com