



NVIDIA Launches Vera CPU, Purpose-Built for Agentic AI

NVIDIA Vera CPU Delivers the Highest Performance and Energy Efficiency for Data Processing, AI Training and Agentic Inference at Scale

News Summary:

- The NVIDIA Vera CPU delivers results with twice the efficiency and 50% faster than traditional CPUs.
- Customers collaborating with NVIDIA to deploy Vera CPU include Alibaba Cloud, ByteDance, Meta and Oracle Cloud Infrastructure, along with CoreWeave, Lambda, Nebius and Nscale.
- Manufacturing partners already adopting the Vera CPU include Dell Technologies, HPE, Lenovo and Supermicro, along with ASUS, Compal, Foxconn, GIGABYTE, Pegatron, Quanta Cloud Technology (QCT), Wistron and Wiwynn.

GTC—NVIDIA today launched the NVIDIA Vera CPU, the world's first processor purpose-built for the age of agentic AI and reinforcement learning — delivering results with twice the efficiency and 50% faster than traditional rack-scale CPUs.

As reasoning and agentic AI advances, scale, performance and cost are increasingly driven by the infrastructure supporting the models that plan tasks, run tools, interact with data, run code and validate results.

The [NVIDIA Vera CPU](#) builds on the success of the [NVIDIA Grace™ CPU](#), enabling organizations of all sizes and across industries to build AI factories that unlock agentic AI at scale. With the highest single-thread performance and bandwidth per core, Vera is a new class of CPU that delivers higher AI throughput, responsiveness and efficiency for large-scale AI services such as coding assistants, as well as consumer and enterprise agents.

Leading hyperscalers collaborating with NVIDIA to deploy Vera include Alibaba Cloud, CoreWeave, Meta and Oracle Cloud Infrastructure, as well as global system makers [Dell Technologies](#), HPE, Lenovo, Supermicro and others. This broad adoption establishes Vera as the new CPU standard for the AI workloads that matter most for developers, startups, public-private institutions and enterprises — helping democratize access to AI and accelerating innovation.

“Vera is arriving at a turning point for AI. As intelligence becomes agentic — capable of reasoning and acting — the importance of the systems orchestrating that work is elevated,” said Jensen Huang, founder and CEO of NVIDIA. “The CPU is no longer simply supporting the model; it’s driving it. With breakthrough performance and energy efficiency, Vera unlocks AI systems that think faster and scale further.”

Configurable for Every Data Center

NVIDIA announced a new Vera CPU rack integrating 256 liquid-cooled Vera CPUs to sustain more than 22,500 concurrent CPU environments, each running independently at full performance. AI factories can quickly deploy and scale to tens of thousands of simultaneous instances and agentic tools in a single rack.

The new Vera rack is built using the [NVIDIA MGX™](#) modular reference architecture, supported by 80 ecosystem partners worldwide.

As part of the [NVIDIA Vera Rubin NVL72](#) platform, Vera CPUs are paired with NVIDIA GPUs through NVIDIA NVLink™-C2C interconnect technology, with 1.8 TB/s of coherent bandwidth — 7x the bandwidth of PCIe Gen 6 — for high-speed data sharing between CPUs and GPUs. Additionally, NVIDIA introduced new reference designs that use Vera as the host CPU for NVIDIA HGX™ Rubin NVL8 systems, coordinating data movement and system control for GPU-accelerated workloads.

Vera systems partners are providing both dual and single-socket CPU server configurations, optimal for workloads such as reinforcement learning, agentic inference, data processing, orchestration, storage management, cloud applications and high-performance computing.

Across all configurations, Vera systems integrate [NVIDIA ConnectX® SuperNIC cards](#) and [NVIDIA BlueField®-4](#) DPUs for accelerated networking, storage and security, which are critical for agentic AI. This enables customers to optimize for their specific workloads while maintaining a single software stack across the NVIDIA platform.

Designed for Agentic Scaling

By combining high-performance, energy-efficient CPU cores, a high-bandwidth memory subsystem and the second-generation NVIDIA Scalable Coherency Fabric, Vera enables faster agentic responses under the extreme utilization conditions common for agentic AI and reinforcement learning.

Vera features 88 custom NVIDIA-designed Olympus cores, delivering high performance for compilers, runtime engines, analytics pipelines, agentic tooling and orchestration services. Each core can run two tasks, using NVIDIA Spatial Multithreading, to deliver consistent, predictable performance — ideal for multi-tenant AI factories running many jobs at once.

To further enhance energy efficiency, Vera introduces the second generation of NVIDIA's low-power memory subsystem, now built on LPDDR5X memory and delivering up to 1.2 TB/s of bandwidth — twice the bandwidth and at half the power compared with general-purpose CPUs.

Widespread Ecosystem Support

Cursor, an innovator in AI-native software development, is adopting NVIDIA Vera to boost performance for its AI coding agents.

"We're excited to use NVIDIA Vera CPUs to improve overall throughput and efficiency so we can deliver faster, more responsive coding agent experiences for our customers," said Michael Truell, cofounder and CEO of Cursor.

Redpanda, a leading streaming data and AI platform, is using Vera to dramatically boost performance.

"Redpanda recently tested NVIDIA Vera running Apache Kafka-compatible workloads and saw dramatically better performance than other systems we've benchmarked, delivering up to 5.5x lower latency," said Alex Gallego, founder and CEO of Redpanda. "Vera represents a new direction in CPU architecture, with more memory and less overhead per core, enabling our customers to scale real-time streaming workloads further than ever and unlock new AI and agentic applications."

National laboratories planning to deploy Vera CPUs include Leibniz Supercomputing Centre, Los Alamos National Laboratory, Lawrence Berkeley National Laboratory's National Energy Research Scientific Computing Center and the Texas Advanced Computing Center (TACC).

"At TACC, we recently tested NVIDIA's Vera CPU platform as we prepare for deployment in our upcoming Horizon system — and running six of our scientific applications, we saw impressive early results," said John Cazes, director of high-performance computing at TACC. "Vera's per-core performance and memory bandwidth represent a giant step forward for scientific computing, and we look forward to bringing Vera-based nodes to our CPU users on Horizon later this year."

Leading cloud service providers planning to deploy Vera CPUs include Alibaba Cloud, ByteDance, Cloudflare, CoreWeave, Crusoe, Lambda, Nebius, Nscale, Oracle Cloud Infrastructure, Together.AI and Vultr.

Leading infrastructure providers adopting Vera CPUs include [Aivres](#), ASRock Rack, ASUS, [Compal](#), Cisco, Dell, Foxconn, GIGABYTE, HPE, Hyve, Inventec, Lenovo, [MiTAC Computing](#), [MSI](#), Pegatron, Quanta Cloud Technology (QCT), Supermicro, Wistron and Wiyynn.

Availability

NVIDIA Vera is in full production and will be available from partners in the second half of this year.

Watch the [GTC keynote](#) from Huang and explore [sessions](#).

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: the CPU driving the model; 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; 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 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; 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, BlueField, ConnectX, NVIDIA Grace, NVIDIA HGX, NVIDIA MGX, 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.

Alex Shapiro
Enterprise Networking
press@nvidia.com