

NVIDIA Unveils Vera, the CPU for Agents

1.8x Faster Than x86 Processors to Drive Diverse Workloads Across Industries, Generating More Data Center Token Revenue

News Summary:

- NVIDIA launches high-performance, energy-efficient NVIDIA Vera CPUs to drive diverse workloads across industries, including agentic AI, reinforcement learning and data processing.
- Vera serves as the CPU powering standalone Vera servers, NVIDIA Vera Rubin systems and Vera BlueField-4 STX AI storage platforms.
- Global AI labs planning to adopt Vera to transform their AI factories include Anthropic, OpenAI and SpaceXAI, and hyperscalers ByteDance, CoreWeave and Oracle Cloud Infrastructure.
- Manufacturers building standalone Vera CPU systems at scale include Dell Technologies, HPE, Lenovo and Supermicro, along with ASUS, Compal, Foxconn, GIGABYTE, Pegatron, Quanta Cloud Technology (QCT), Wistron and Wiwynn.

NVIDIA GTC Taipei--NVIDIA today announced that the world's technology leaders are planning to adopt [NVIDIA Vera](#), the first CPU built for AI agents.

Now in full production, NVIDIA Vera is a new class of processor enabling 1.8x faster task completion compared with x86 CPUs to drive diverse workloads across industries -- including agentic AI, reinforcement learning and data processing -- generating more data center token revenue.

Building on the success of NVIDIA Grace™ CPUs, which have nearly 2.5 million shipments to date, Vera takes CPU performance and energy efficiency to new levels for the most demanding AI workloads in modern data centers -- where agents move from answering basic questions to taking actions, running code, using tools and evaluating results.

Customers exploring the Vera CPU include finance leader NYSE, global AI labs Anthropic, OpenAI and SpaceXAI, and hyperscalers ByteDance, CoreWeave, Lambda, Nebius, Nscale and Oracle Cloud Infrastructure (OCI). Vera is also being integrated into AI infrastructure from world-leading system manufacturers such as Dell Technologies, HPE, Lenovo and Supermicro, along with Taiwan system builders.

"AI agents will be the largest users of computing," said Jensen Huang, founder and CEO of NVIDIA. "Vera is the first CPU designed for that future -- built to run agentic AI at hyperscale with extraordinary performance, efficiency and programmability."

"At the NYSE, our focus is to optimize the latency, throughput and reliability of the systems underpinning our unrivaled infrastructure," said Lynn Martin, president of NYSE Group. "The NYSE processes more than 1.1 trillion messages per day, and in collaboration with Redpanda and HPE, using NVIDIA Vera CPUs, we will be scaling our capacity while further optimizing latency to power a high-performance, resilient and AI-ready market infrastructure."

Anthropic, the AI innovator behind Claude, is evaluating adding Vera to scale CPU-intensive agentic workloads.

"Scaling compute is an important accelerant for the growth of models," said James Bradbury, head of compute at Anthropic. "We're excited to see Vera emerge as a promising part of the ecosystem when solving for agentic workloads."

OCI Supercluster powered by NVIDIA Vera represents the next frontier in hyperscale AI supercomputing.

"Oracle Cloud Infrastructure is rapidly scaling AI infrastructure to meet surging demand for training, inference and agentic AI," said Mahesh Thiagarajan, executive vice president of Oracle Cloud Infrastructure. "By deploying NVIDIA Vera CPUs, OCI will support high-throughput reasoning and data processing workloads across next-generation AI environments."

According to Phoronix, which offers a comprehensive, open source benchmarking suite, NVIDIA Vera [delivered](#) the fastest overall performance across agentic workloads including code compilation, Python, Java and database processing. These workloads sit on the critical path of modern AI factories, including for agent tool use and sandbox execution, where faster CPU performance delivers higher agent throughput and interactivity.

A Custom CPU for the Agentic Era

AI factory economics are shifting from cores per dollar to tokens per dollar, requiring CPUs that complete agentic, data-processing and orchestration work faster and more efficiently.

Vera is powered by Olympus, a custom NVIDIA CPU core engineered for the CPU work behind that shift, from Python

runtimes and sandboxed code execution to orchestration logic and analytics pipelines.

Vera is built to process more instructions, anticipate application behavior and move data across large numbers of concurrent environments, queries and data processing tasks -- featuring 88 Olympus cores, Spatial Multithreading and a LPDDR5X memory subsystem that delivers up to 1.2TB/s of bandwidth. This helps agents spend less time waiting on CPU-bound steps and lets AI factories keep accelerators moving.

The Vera CPU can also be deployed across the full AI factory -- from the standalone CPU infrastructure to tightly coupled accelerated systems. Vera helps AI factories deliver higher end-to-end throughput and faster time to solution for users, improving responsiveness and efficiency across training, inference and agentic execution.

Vera serves as the host CPU for NVIDIA Vera Rubin platforms through second-generation NVIDIA NVLink™-C2C interconnect technology, which provides up to 1.8TB/s of coherent bandwidth between CPU and GPU. It extends NVIDIA Confidential Computing at rack scale, protecting agentic workloads.

The [NVIDIA Vera BlueField-4 STX](#) processor integrates Vera with high-performance networking, storage acceleration and in-silicon security to create secure-by-design AI-native data platforms.

Extensive Ecosystem Support

Vera CPUs are available in dense, liquid-cooled racks for large-scale agentic AI and reinforcement learning environments, as well as flexible two-socket air-cooled systems for enterprise, cloud, data processing and AI factory deployments.

Leading infrastructure providers offering Vera CPU-based systems include Aivres, ASRock Rack, [ASUS](#), Compal, Dell, Foxconn, GIGABYTE, HPE, Hyve Solutions, Inventec, Lenovo, MiTAC Computing, [MSI](#), Pegatron, Quanta Cloud Technology (QCT), Supermicro, Wistron and Wiwynn. Major original equipment manufacturers -- Dell, HPE, Lenovo and Supermicro -- will be offering Vera in standalone CPU server configurations, the first standard CPU option beyond x86.

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

Availability

Vera systems will be available from system builders and cloud partners starting this fall.

Watch Huang's [keynote](#) and learn more at [NVIDIA GTC Taipei](#).

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: AI agents will be the largest users of computing; expectations with respect to growth, performance, availability, 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 its collaborators and partners; 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, NVIDIA Grace and NVLink 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.

Alex Shapiro
Corporate Communications
NVIDIA Corporation

