



NVIDIA Launches BlueField-4 STX Storage Architecture With Broad Industry Adoption

News Summary:

- New NVIDIA STX reference architecture provides up to 5x token throughput and up to 4x energy efficiency with 2x faster data ingestion.
- Early adopters of STX for context memory storage include CoreWeave, Crusoe, IREN, Lambda, Mistral AI, Nebius, Oracle Cloud Infrastructure (OCI) and Vultr.
- Storage providers and manufacturing partners are building infrastructure using NVIDIA modular reference designs to advance agentic AI, including AIC, Cloudian, DDN, Dell Technologies, Everpure, Hitachi Vantara, HPE, IBM, MinIO, NetApp, Nutanix, Supermicro, Quanta Cloud Technology (QCT), VAST Data and WEKA.

GTC—NVIDIA today announced NVIDIA BlueField-4[®] STX, a modular reference architecture that enables enterprises, cloud and AI providers to easily deploy accelerated storage infrastructure capable of the long-context reasoning required for agentic AI.

Traditional data centers provide high-capacity, general-purpose storage but lack the responsiveness required for seamless interaction with AI agents that work across many steps, tools and sessions. Agentic AI demands real-time access to data and contextual working memory to keep conversations and tasks fast and coherent. As context grows, traditional storage and data paths can slow AI inference and reduce GPU utilization.

[NVIDIA STX](#) allows storage providers to build infrastructure that keeps data close and accessible at scale, so agentic AI factories can deliver higher throughput and responsiveness across inference, training and analytics.

The first rack-scale implementation includes the new [NVIDIA CMX™ context memory storage platform](#), which expands GPU memory with a high-performance context layer for scalable inference and agentic systems — providing up to 5x tokens per second compared with traditional storage.

“Agentic AI is redefining what software can do — and the computing infrastructure behind it must be reinvented to keep pace,” said Jensen Huang, founder and CEO of NVIDIA. “AI systems that reason across massive context and continuously learn require a new class of storage. NVIDIA STX reinvents the storage stack, providing a modular foundation for AI-native infrastructure that keeps AI factories operating at peak performance.”

STX is accelerated by the NVIDIA Vera Rubin platform and harnesses a new, storage-optimized [NVIDIA BlueField-4 processor](#) that combines the [NVIDIA Vera CPU](#) with [NVIDIA ConnectX[®]-9 SuperNIC](#), together with [NVIDIA Spectrum-X™ Ethernet](#) networking, [NVIDIA DOCA™](#) and [NVIDIA AI Enterprise](#) software.

The STX architecture also enables 4x higher energy efficiency compared with traditional CPU architectures for high-performance storage and can ingest 2x more pages per second for enterprise AI data.

Storage providers partners codesigning next-generation AI infrastructure based on NVIDIA STX include [Cloudian](#), [DDN](#), [Dell Technologies](#), [Everpure](#), [Hitachi Vantara](#), [HPE](#), IBM, [MinIO](#), [NetApp](#), [Nutanix](#), [VAST Data](#) and [WEKA](#).

Manufacturing partners building STX-based systems include AIC, [Supermicro](#) and Quanta Cloud Technology (QCT).

Leading AI labs and cloud service providers planning to adopt STX for context memory storage include CoreWeave, Crusoe, IREN, Lambda, Mistral AI, Nebius, OCI and Vultr.

STX-based platforms 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: agentic AI redefining what software can do — providing systems that reason across massive context and continuously build on enterprise knowledge; 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.

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 CMX, NVIDIA DOCA and NVIDIA Spectrum-X 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