



NVIDIA BlueField-4 Powers New Class of AI-Native Storage Infrastructure for the Next Frontier of AI

News Summary:

- NVIDIA BlueField-4 powers NVIDIA Inference Context Memory Storage Platform, a new kind of AI-native storage infrastructure designed for gigascale inference, to accelerate and scale agentic AI.
- The new storage processor platform is built for long-context-processing agentic AI systems with lightning-fast long- and short-term memory.
- Inference Context Memory Storage Platform extends AI agents' long-term memory and enables high-bandwidth sharing of context across clusters of rack-scale AI systems -- boosting tokens per seconds and power efficiency by up to 5x.
- Enabled by NVIDIA Spectrum-X Ethernet, extended context memory for multi-turn AI agents improves responsiveness, increases throughput per GPU and supports efficient scaling of agentic inference.

CES--NVIDIA today announced that the NVIDIA BlueField®-4 data processor, part of the full-stack [NVIDIA BlueField](#) platform, powers NVIDIA Inference Context Memory Storage Platform, a new class of AI-native storage infrastructure for the next frontier of AI.

As AI models scale to trillions of parameters and multistep reasoning, they generate vast amounts of context data -- represented by a key-value (KV) cache, critical for accuracy, user experience and continuity.

A KV cache cannot be stored on GPUs long term, as this would create a bottleneck for real-time inference in multi-agent systems. AI-native applications require a new kind of scalable infrastructure to store and share this data.

NVIDIA Inference Context Memory Storage Platform provides the infrastructure for context memory by extending GPU memory capacity, enabling high-speed sharing across nodes, boosting tokens per seconds by up to 5x and delivering up to 5x greater power efficiency compared with traditional storage.

"AI is revolutionizing the entire computing stack -- and now, storage," said Jensen Huang, founder and CEO of NVIDIA. "AI is no longer about one-shot chatbots but intelligent collaborators that understand the physical world, reason over long horizons, stay grounded in facts, use tools to do real work, and retain both short- and long-term memory. With BlueField-4, NVIDIA and our software and hardware partners are reinventing the storage stack for the next frontier of AI."

NVIDIA Inference Context Memory Storage Platform boosts KV cache capacity and accelerates the sharing of context across clusters of rack-scale AI systems, while persistent context for multi-turn AI agents improves responsiveness, increases AI factory throughput and supports efficient scaling of long-context, multi-agent inference.

Key capabilities of the NVIDIA BlueField-4-powered platform include:

- [NVIDIA Rubin](#) cluster-level KV cache capacity, delivering the scale and efficiency required for long-context, multi-turn agentic inference.
- Up to 5x greater power efficiency than traditional storage.
- Smart, accelerated sharing of KV cache across AI nodes, enabled by the NVIDIA DOCA™ framework and tightly integrated with the NVIDIA NIXL library and NVIDIA Dynamo software to maximize tokens per second, reduce time to first token and improve multi-turn responsiveness.
- Hardware-accelerated KV cache placement managed by NVIDIA BlueField-4 eliminates metadata overhead, reduces data movement and ensures secure, isolated access from the GPU nodes.
- Efficient data sharing and retrieval enabled by [NVIDIA Spectrum-X™ Ethernet](#) serves as the high-performance network fabric for RDMA-based access to AI-native KV cache.

Storage innovators including [AIC](#), [Cloudian](#), [DDN](#), [Dell Technologies](#), HPE, Hitachi Vantara, [IBM](#), [Nutanix](#), Pure Storage, Supermicro, [VAST Data](#) and [WEKA](#) are among the first building next-generation AI storage platforms with BlueField-4, which will be available in the second half of 2026.

Learn more by watching [NVIDIA Live at CES](#).

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 revolutionizing the entire computing stack and storage; AI no longer about one-shot chatbots but intelligent collaborators that understand the physical world, reason over long horizons, stay grounded in facts, use tools to do real work, and retain both short- and long-term memory;

with BlueField-4, NVIDIA and its software and hardware partners reinventing the storage stack for 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.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, 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
Corporate Communications
NVIDIA Corporation
press@nvidia.com