



Nutanix and Supermicro Take on Runaway AI Costs with Jointly Engineered System for Private Inference

October 21, 2026

Supermicro servers powered by Nutanix Agentic AI are designed for enterprises to fast track deployments of AI infrastructure, secure critical on-premises GPU resources, reduce AI costs, and scale AI workloads with flexibility and control

News Summary

- Nutanix and Supermicro are expanding their collaboration to help enterprises lower the cost per token of private AI inference and accelerate adoption of open-weight models through a planned engineered, full stack infrastructure solution.
- The solution will enable enterprises to accelerate AI inference workloads and facilitate securing critical on-premises GPU resources.
- The solution will combine Supermicro servers with Nutanix Agentic AI full stack software designed to reduce custom-order, configuration, and integration complexity, giving customers a more predictable way to deploy, manage, and scale on-premises GPU infrastructure.

Why it Matters

For customers, this means AI cost control and a faster path to putting production AI to work because of operational simplicity due to less time that will be spent assembling and integrating multivendor environments.

BERLIN, Oct. 21, 2026 (GLOBE NEWSWIRE) -- [NVIDIA GTC](#) – Nutanix (NASDAQ: NTNX), a hybrid cloud leader and AI platform company, and Super Micro Computer, Inc. (NASDAQ: SMCI), today announced plans for a jointly engineered AI infrastructure solution powered by Nutanix Agentic AI designed to help enterprises take on runaway AI costs with a system for private inference.

Built on the validated Supermicro systems powered by Nutanix Agentic AI, the solution will combine the Nutanix Cloud Platform with Supermicro servers, NVIDIA GPUs, and [NVIDIA Spectrum-X](#) networking. Designed to reduce custom configuration and integration complexity, it will simplify the path from infrastructure deployment to production-ready AI. Nutanix delivers a complete, full-stack AI platform engineered to simplify and accelerate the deployment of enterprise AI workloads -- from underlying infrastructure to model execution and the governance of agentic workloads. The solution will help enable seamless provisioning from bare metal to delivering integrated model-as-a-service and a secure Agent Gateway using Nutanix Kubernetes Platform -- the Cloud Native Computing Foundation certified AI conformant Kubernetes distribution -- and Nutanix Enterprise AI. The solution will be supported by Nutanix Unified Storage (NUS) which provides high-performance NVIDIA Enterprise-certified storage for AI workloads, including context caching and model serving. Together, these capabilities are intended to provide the end-to-end performance, resilience, and orchestration required to support modern AI initiatives. The built-in AI catalog will further simplify deployment by providing access to common AI applications.

As AI requirements grow, the architecture is designed to scale GPU capacity while the dedicated Nutanix control plane and storage services will support resilient operations and high-throughput data workflows. With this engineered offering, customers will be able to more easily deploy, manage, and scale AI inference workloads without having to assemble and connect a complex mix of multivendor technologies.

"Collaborating with Nutanix, we want to accelerate time-to-online for enterprise AI," said **Matthew Thauberger, Chief Revenue Officer, Supermicro**. "Combining Supermicro's engineered servers with Nutanix's full-stack software will deliver a pre-validated, rack-scale solution that bypasses traditional lead-time delays. It will help give organizations predictable access to high-density GPU compute while removing the operational complexity of scaling AI."

"Moving AI from experimentation into production remains a challenge for many enterprises, with infrastructure complexity and hardware procurement often slowing progress. Our expanded collaboration with Supermicro is a direct response to this barrier," said **Thomas Cornely, Executive Vice President, Product Management, Nutanix**. "The Supermicro systems, engineered with our Nutanix Agentic AI full stack software, will help enterprises simplify how AI infrastructure is deployed, operated, and secured, and scaled, enabling them to simply deploy private inference and control the cost per token as AI adoption scales."

Supermicro with Nutanix Cloud Platform: Reducing the Complexity of Enterprise AI Infrastructure

- The engineered solution from Supermicro and Nutanix is intended to provide a hardware foundation that helps eliminate custom-order delays by using the Supermicro server systems which houses eight NVIDIA Blackwell Ultra GPUs per node.
- The resilient single-tenant control plane is expected to feature a dedicated, resilient Nutanix management cluster (NX-8170-G10) to protect against management plane failure without disrupting GPU execution.
- For high-performance storage scaling, an integrated NUS cluster (NX-8150-G10) is optimized for high-bandwidth I/O requirements across AI data ingestion, checkpointing, and cache warming.
- NVIDIA Spectrum-X Ethernet networking combines NVIDIA Spectrum-4 Ethernet switches (SN5610/SN5600) with NVIDIA ConnectX-7/8 SuperNICs for high-performance, predictable GPU-to-GPU scale-out connectivity and NVIDIA BlueField-3 DPUs for infrastructure offloading and in-band management, complementing NVIDIA's scale-in networking architecture.

Analyst Commentary

"The AI imperative is shifting from simply building raw compute capacity to optimizing the cost and efficiency of every inference delivered. Hardware, memory, networking, model optimization, and orchestration all need to be managed as a system if organizations are going to scale inference efficiently and translate AI investment into real business outcomes."

- **Matt Kimball, Vice President & Principal Analyst, Moor Insights & Strategy**

Availability

The Supermicro systems powered by Nutanix Agentic AI will be available early 2027 from Supermicro and its channel partners.

About Nutanix

Nutanix is a hybrid cloud leader and AI platform company, offering organizations a unified infrastructure software platform to safely run applications, data, and AI anywhere. Trusted by customers worldwide, Nutanix empowers more than 50% of the Global 2000 to innovate faster with AI, while modernizing infrastructure, simplifying operations, and controlling costs. Learn more at www.nutanix.com or follow us on social media.

About Super Micro Computer, Inc.

Supermicro (NASDAQ: SMCI) is a global leader in Application-Optimized Total IT Solutions. Founded and operating in San Jose, California, Supermicro is committed to delivering first-to-market innovation for Enterprise, Cloud, AI, and 5G Telco/Edge IT Infrastructure. We are a Total IT Solutions provider with server, AI, storage, IoT, switch systems, software, and support services. Supermicro's motherboard, power, and chassis design expertise further enables our development and production, enabling next-generation innovation from cloud to edge for our global customers. Our products are designed and manufactured in-house (in the US, Taiwan, and the Netherlands), leveraging global operations for scale and efficiency and optimized to improve TCO and reduce environmental impact (Green Computing). The award-winning portfolio of Server Building Block Solutions® allows customers to optimize for their exact workload and application by selecting from a broad family of systems built from our flexible and reusable building blocks that support a comprehensive set of form factors, processors, memory, GPUs, storage, networking, power, and cooling solutions (air-conditioned, free air cooling or liquid cooling).

Nutanix Forward-Looking Statements

©2026 Nutanix, Inc. All rights reserved. Nutanix, the Nutanix logo and all Nutanix product and service names mentioned are registered trademarks or trademarks of Nutanix, Inc. in the United States and other countries.

This press release contains express and implied forward-looking statements, including but not limited to statements regarding the expanded collaboration between Nutanix and Supermicro; the planned jointly engineered AI infrastructure solution; anticipated technical integrations; future products, services, features and offerings; anticipated customer benefits; planned technical collaboration and go-to-market activities; customer adoption; the timing, availability and performance of future solutions; and the parties' ability to successfully finalize and implement commercial arrangements supporting the planned solution relating to the expanded collaboration. These forward-looking statements are based on Nutanix's current expectations, estimates, assumptions and projections and involve risks and uncertainties that could cause actual results to differ materially. Actual results may differ materially due to a number of factors, including the parties' ability to finalize and implement commercial arrangements supporting the planned solution, complete anticipated integration efforts, develop and deliver contemplated functionality, successfully execute collaborative activities, achieve customer adoption, and realize the anticipated benefits of the planned solution, as well as other risks and uncertainties described in Nutanix's filings with the Securities and Exchange Commission, including its Annual Report on Form 10-K and subsequent Quarterly Reports on Form 10-Q and other filings. These forward-looking statements speak only as of the date of this press release, and Nutanix undertakes no obligation to update or revise any forward-looking statements, whether as a result of new information, future events or otherwise, except as required by law. Many of the anticipated products, services, integrations, offerings, features and functionalities described herein remain in various stages of planning, development, testing and implementation and will be offered on a when-and-if-available basis. The development, release and timing of any such products, features or functionalities are subject to change. Nutanix will not have any liability arising from reliance on this press release for any failure to deliver, or delay in the delivery of, any such products, features or functionalities. Any future product or product feature information is intended to outline general product direction and should not be relied upon in making a purchasing decision. Such information is not a commitment, promise or legal obligation to deliver any functionality, offering or integration. The expanded collaboration, planned solution, offerings and related activities described in this press release remain subject to ongoing development efforts and, where applicable, the finalization and implementation of related commercial arrangements between the parties.

Media Contacts: Nutanix Ken Lotich pr@nutanix.com Supermicro pr@supermicro.com