NVIDIA s latest AI server chip

Overview

At the 2026 Nvidia GTC conference, Jensen Huang announced an inference-specific chip, the Groq 3 LPU. The LPU will work in concert with the Rubin GPU to accelerate AI workloads. According to TrendForce's latest findings on AI servers, NVIDIA's high-end AI chip shipment mix is expected to change in 2026. This week, over 30,000 people are descending upon San Jose, Calif., to attend Nvidia GTC, the so-called Superbowl of AI—a. Nvidia's Blackwell Ultra chips, the company's next-generation graphics processor for AI, have been commercially deployed at CoreWeave, the companies announced on Thursday. CoreWeave historically has a close relationship with Nvidia, which owns a stake in the cloud provider. CoreWeave went public. The Rubin platform harnesses extreme codesign across hardware and software to deliver up to 10x reduction in inference token cost and 4x reduction in number of GPUs to train MoE models, compared with the NVIDIA Blackwell platform. NVIDIA Spectrum-X Ethernet Photonics switch systems deliver 5x.

Article Content

CoreWeave receives first AI server powered by Nvidia''s latest chip

CoreWeave Inc. said it has received the market''s first artificial intelligence server system based on the newest, high-end Nvidia Corp. chip, a sign of its rising stature in the competitive AI

Nvidia market share in China falls to less than 60% — Chinese chip ...

Tech Industry Nvidia market share in China falls to less than 60% — Chinese chip makers deliver 1.65 million AI GPUs as the government pushes data centers to use domestic chips

Rubin Faces Delay Risks Amid Ongoing Supply Chain Adjustments ...

According to TrendForce''s latest findings on AI servers, NVIDIA''s high-end AI chip shipment mix is expected to change in 2026. The combined share of Hopper and Rubin series in

AI chip startups raise billions to challenge Nvidia''s dominance in ...

The AI boom is entering a new phase, and the shift is happening faster than many expected. The race is no longer just about training bigger models. It''s about running them efficiently

NVIDIA Corporation

New NVIDIA Inference Context Memory Storage Platform with NVIDIA BlueField-4 storage processor to accelerate agentic AI reasoning. Microsoft''s next-generation Fairwater AI superfactories

United States China Tech Rivalry Delays Nvidia AI Chip Exports

The latest developments surrounding Nvidia''s H200 chip sales to China highlight the growing complexity of the technological rivalry between the United States and China.

Prices of Nvidia''s B300 server at $1 million in China on US curbs

Nvidia''s B300 servers have skyrocketed in price to around one million dollars each in China, fueled by an insatiable appetite for AI computing capabilities and increasingly stringent chip

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: https://activa.net.pl
Email: sales@activa.net.pl
Phone: +48 662 748 193
Address: ul. Cybernetyki 7B, 02-677 Warsaw, Poland

This document is for informational purposes only. Specifications subject to change without notice.