

NVIDIA s latest AI server chip



AI Overview

Nvidia's latest AI server platform, Rubin, integrates six next-generation chips—including Vera CPUs and Rubin GPUs—designed for large-scale AI training and inference with unprecedented efficiency and performance. Rubin Platform Overview Nvidia's Rubin platform is a next-generation AI supercomputing solution that combines six specialized chips into a tightly integrated system for AI workloads. The platform is designed for agentic AI, advanced reasoning, and mixture-of-experts (MoE) models, enabling multi-step inference over extremely long contexts with high efficiency. Key components include: NVIDIA Vera CPU: 88 custom Olympus cores optimized for AI data orchestration, Arm-compatible, and designed to manage data movement, memory, and control flow at AI factory scale. NVIDIA Rubin GPU: High-performance AI compute with HBM4 memory and the new Transformer Engine for accelerated model training and inference. NVIDIA NVLink 6 Switch: Sixth-generation interconnect providing 3.6 TB/s GPU-to-GPU bandwidth for high-speed communication. NVIDIA ConnectX-9 SuperNIC: Low-latency, high-throughput networking interface for scale-out AI deployments. NVIDIA BlueField-4 DPU: Infrastructure offload and security processor to manage AI workloads efficiently. NVIDIA Spectrum-6 Ethernet Switch: Co-packaged optics for reliable, scale-out connectivity. Performance and Efficiency The Rubin platform achieves up to 10x lower cost per token compared to the previous Blackwell platform and can train MoE models with 4x fewer GPUs, significantly reducing infrastructure requirements. The NVL72 rack-scale system combines 72 GPUs into a single server, and multiple NVL72s can be linked to form a DGX SuperPOD, creating massive AI supercomputers suitable for hyperscalers. Industry Adoption Major companies are rapidly adopting Rubin systems. Meta plans to deploy millions of Nvidia chips, including standalone Vera CPUs and next-generation GPUs, across its AI data centers to support AI features on platforms like WhatsApp. Micros...

Article Content

2026-07-06 Pre-Market Financial News Brief | Intellectia

Amazon's AI Chip Strategy Amazon's AWS has raised its third-quarter shipment target for Trainium 3 servers by 20-30%, indicating strong demand for AI capabilities.

NVIDIA Vera Rubin Opens Agentic AI Frontier

NVIDIA today announced the NVIDIA Vera Rubin platform is opening the next frontier of agentic AI, with seven new chips now in full production to

DGX Vera Rubin: Nvidia shows next AI server

At his CES keynote, Nvidia unveiled its next-generation AI server, DGX Vera Rubin. It consists of six specific chips, all of which are already running

Computex 2026 Roundup: NVIDIA, Intel and AMD's AI Showdown

Computex 2026 recap: NVIDIA's RTX Spark PC chip, Intel's 480GB Crescent Island GPU, Vera Rubin in full production, Cosmos 3, AMD's \$10B Taiwan bet and more.

China's DeepSeek Is Making Its Own AI Chip — Nvidia Shares

DeepSeek is building its own AI inference chip, sources say. Nvidia shares slipped 2%. Here's what we know about the early-stage effort.

Latest NVIDIA Stock News | NASDAQ:NVDA | Benzinga

Get the latest news on NVIDIA stock with updates on market trends, financials, and analysis from Benzinga. Stay informed on NVDA performance

Top 25+ AI Chip Makers: NVIDIA & Its Competitors

AI chips enable parallel computing capabilities are increasingly in demand. This article will information to you on 10 popular AI chip makers.

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

Nvidia's Kyber rack system delayed to 2028 over manufacturing

Kyber is a server cabinet that packs 144 of Nvidia's most powerful chips into a single unit so they can work together as one giant computer, providing the horsepower AI companies need to train

Everything NVIDIA Announced at GTC 2026, Explained | The Neuron

Everything NVIDIA Just Announced at GTC 2026: Seven Chips, Five Racks, One Giant Bet on Agentic AI Here's every major announcement from GTC 2026: Jensen Huang unveiled the

NVIDIA Kicks Off the Next Generation of AI With Rubin

NVIDIA today kickstarted the next generation of AI with the launch of the NVIDIA Rubin platform, comprising six new chips designed to deliver one

DGX Vera Rubin: Nvidia shows next AI server generation with in-house ...

At his CES keynote, Nvidia unveiled its next-generation AI server, DGX Vera Rubin. It consists of six specific chips, all of which

NVIDIA unveils Rubin six-chip system for next-gen AI at CES 2026

NVIDIA says its Rubin platform is now in full production, delivering up to 50 petaflops and powering the next wave of agentic AI systems.

NVIDIA GTC 2026 Highlights: Recap on Everything You

NVIDIA GTC 2026 Highlights Recap: \$1T AI orders, Vera Rubin platform, OpenClaw & NemoClaw agentic AI, physical robotics breakthroughs,

Massive NVIDIA-Intel Partnership Ignites Future of AI-Ready PCs

NVIDIA buys a 5% Intel stake for \$5 billion and partners on AI PCs and data center chips, integrating NVLink and x86 CPUs to launch next-gen AI products in 2026.

Top 25+ AI Chip Makers: NVIDIA & Its Competitors

Announced in May 2025 at Computex, NVIDIA introduced RTX PRO Servers powered by RTX PRO 6000 Blackwell Server Edition GPUs, designed for enterprise AI factories.

Nvidia's New Hedge Against Chip Competitors? Partner with Them

Nvidia has a plan to deal with the growing list of AI server chip competitors: partner with them. In the latest example, Nvidia and AI chip startup d-Matrix are combining their respective

Qualcomm announces AI chips to compete with AMD

Qualcomm announced that it will release new AI accelerator chips. Nvidia has dominated the market for AI chips, with AMD seen as the second

NVIDIA GTC 2026: Feynman AI Chip, TSMC 1.6nm A16 Process

A comprehensive analysis of NVIDIA GTC 2026. Discover the breakthrough Feynman AI chip featuring TSMC's A16 (1.6nm) process, Backside Power Delivery technology, and the shift

NVDA Stock Swings In Premarket: Report Says Nvidia's Next-Gen AI

Nvidia's Kyber rack architecture, designed to house a large number of Rubin Ultra chips, is delayed to 2028, according to SemiAnalysis.

NVIDIA Vera Rubin NVL72: Full Specs & Platform Breakdown

NVIDIA announced its most ambitious AI platform yet at GTC 2026. Here's the full spec breakdown on every chip, rack, and configuration.

NVIDIA Corporation

Meta's AI Roadmap Supported by Large-Scale Deployment of NVIDIA CPUs, Networking and Millions of NVIDIA Blackwell and Rubin GPUs

NVIDIA Corporation

NVIDIA Vera Rubin NVL72 offers a unified, secure system that combines 72 NVIDIA Rubin GPUs, 36 NVIDIA Vera CPUs, NVIDIA NVLink 6, NVIDIA ConnectX-9 SuperNICs and NVIDIA

Contact Us

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

Website: <https://dknconsulting.co.za>

Email: info@dknconsulting.co.za

Phone: +31 6 2948 7351

Address: Herengracht 123, 1015 BT Amsterdam, Netherlands

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

