Nebius B.V.,
Gustav Mahlerlaan 308,
1082 ME Amsterdam

Nebius to offer NVIDIA Vera Rubin NVL72 in US and Europe from H2 2026

- Nebius will be among the first NVIDIA Cloud Partners to bring the next-generation accelerated computing platform to market

Amsterdam, January 5, 2026 — Nebius (NASDAQ: NBIS) will deploy the NVIDIA Rubin platform through Nebius AI Cloud and Nebius Token Factory, unlocking next-generation reasoning and agentic AI capabilities for customers starting H2 2026.

Nebius, an NVIDIA Cloud Partner, will be among the first AI cloud providers to offer NVIDIA Vera Rubin NVL72. Nebius will integrate Vera Rubin NVL72 across its full-stack infrastructure at data centers in the US and Europe, enabling customers to build next-generation AI applications with regional availability and control.

Launched at CES 2026, Vera Rubin NVL72 is engineered to serve the demands of complex AI workloads, including agentic, advanced reasoning, and massive-scale mixture-of-experts (MoE) models that push computational limits across long sequences of tokens for multistep problem-solving with the lowest cost per token.

Arkady Volozh, founder and CEO of Nebius, said:

“We are proud to be one of the first on the market to offer Vera Rubin GPUs as we fuel the next wave of AI innovation. By integrating Vera Rubin into Nebius AI Cloud and our inference platform Nebius Token Factory, we're giving AI innovators and enterprises the infrastructure they need to develop agentic and reasoning AI systems faster and more efficiently.”

Dave Salvator, director of accelerated computing products, NVIDIA, said:

“Leading in the era of agentic AI requires infrastructure that is purpose-built for scale, performance, reliability and cost efficiency. Nebius’s AI-native infrastructure will enable customers to deploy NVIDIA Rubin–powered AI applications in production with confidence.”

The Rubin accelerated computing platform will be available through Nebius AI Cloud and will serve as the computational layer in Nebius Token Factory, complementing existing NVIDIA GB200 NVL72 and NVIDIA Grace Blackwell Ultra NV72 capacity and expanding customers' choice of platforms optimized for different AI workload profiles.

Through Nebius AI Cloud, customers get direct infrastructure access with the same benchmark-validated, bare-metal performance Nebius delivers on existing NVIDIA platforms. Through Nebius Token Factory — an enterprise-ready inference and post-training platform — they can train, distill, and serve open-source models with predictable latency, performance, and cost.

Nebius consistently delivers benchmark-aligned performance at scale — and as an NVIDIA Exemplar Cloud Partner, Nebius' infrastructure is validated against NVIDIA reference architectures and benchmarks. These validations give customers confidence that Rubin-based workloads, whether accessed directly through AI Cloud or via Token Factory, will perform as designed from first availability.

About Nebius

Nebius is a technology company building full-stack cloud infrastructure for the global AI industry. Headquartered in Amsterdam and listed on Nasdaq (NASDAQ: NBIS), the company has a global footprint with R&D hubs across Europe, North America, and Israel.

Nebius AI Cloud has been built from the ground up for intensive AI workloads. With proprietary software and hardware designed in-house, Nebius gives AI builders the compute, storage, managed services, and tools they need to build, tune, and run their models.

Contacts

Investor Relations: askIR@nebius.com

Media Relations: media@nebius.com