

Nebius to build first NVIDIA GPU cluster in US; expands offices to serve customers across the country

- AI infrastructure provider's first GPU cluster in the US will service customer workloads from early 2025
- GPU cluster in Kansas City has potential capacity to house up to about 35 thousand GPUs after expansion
- Nebius's new customer hubs in San Francisco and Dallas have been operational since September; third office in New York coming by year-end

Amsterdam, November 19, 2024 — Nebius Group N.V. ("Nebius Group", "Nebius" or the "Company"; NASDAQ:NBIS), a leading AI infrastructure company, today announced the launch of its first NVIDIA GPU cluster in the United States with a deployment in Kansas City, MO, bringing its AI-native cloud closer to American customers.

Scheduled to go live in Q1 2025, the Kansas City cluster will house thousands of state-of-the-art NVIDIA GPUs, primarily NVIDIA Hopper GPUs in the initial phase, with the energy-efficient NVIDIA Blackwell platform expected to arrive in 2025. The colocation can be expanded from an initial 5 MW up to 40 MW, or about 35 thousand GPUs, at full potential capacity.

Nebius is actively ramping up its presence in the US as part of its strategy to become a leading provider of AI infrastructure to AI builders globally, and is in advanced discussions for a second, larger-scale GPU cluster in the US, also slated to come online in 2025. The Company has also opened two new customer-facing hubs in San Francisco and Dallas, with a third office set to open in New York later this year.

Arkady Volozh, founder and CEO of Nebius, said:

"Our first GPU cluster in the US and new offices represent a pivotal step in our expansion in the US market. Serving American customers from American facilities means lower latency and maximizes the advantages of our AI-native cloud. We will be building out more GPU clusters across the US to meet exploding demand for high-quality AI infrastructure from US AI developers and enterprises."

Built on top of the latest NVIDIA GPUs with a fleet of H100s already installed and H200s coming onstream this month, Nebius's full-stack AI infrastructure is being purpose-built to meet the demands of the global AI industry and leans on deep technical expertise across hardware and software, cloud engineering and machine learning ("ML").

"Enterprises around the globe need enhanced security, low latency and scalability to advance AI development," said Dave Salvator, director of accelerated computing products at NVIDIA. "With the introduction and deployment of more NVIDIA GPU clusters, Nebius will help U.S. developers and researchers accelerate AI, data processing and high-performance computing tasks."

Publicly announced in October, the [AI-native Nebius cloud](#) is designed to manage the full ML lifecycle – from data processing and training through to fine-tuning and inference – all in one place. The recently launched [Nebius AI Studio](#) inference service expands the Company's offering to app builders, with access to a range of state-of-the-art open-source models in a flexible, user-friendly environment at among the lowest price-per-token on the market.

Nebius has a team of around 400 engineers with decades of knowledge of building world-class tech infrastructure, as well as an in-house large language model ("LLM") R&D team. Listed on Nasdaq, the Company recently announced [investments of more than USD 1 billion](#) in AI infrastructure by mid-2025, enabling Nebius to deploy tens of thousands of NVIDIA GPUs to bring its highly differentiated, energy-efficient, AI-native cloud offering to customers worldwide.

About Nebius

Nebius is a technology company building full-stack infrastructure to service the explosive growth of the global AI industry, including large-scale GPU clusters, cloud platforms, and tools and services for developers. Headquartered in Amsterdam and listed on Nasdaq, the Company has a global footprint with R&D hubs across Europe, North America and Israel.

Nebius's core business is an AI-centric cloud platform built for intensive AI workloads. With proprietary cloud software architecture and hardware designed in-house (including servers, racks and data center design), Nebius gives AI builders the compute, storage, managed services and tools they need to build, tune and run their models.

A Preferred cloud service provider in the NVIDIA Partner Network, Nebius offers high-end infrastructure optimized for AI training and inference. The Company boasts a team of around 400 skilled engineers, delivering a true hyperscale cloud experience tailored for AI builders.

To learn more please visit www.nebius.com

Contacts

For journalists: media@nebius.com

For investors: askIR@nebius.com

--- ENDS ---

Disclaimer

Forward Looking Statements

This press release and the materials referenced herein contain forward-looking statements that involve risks and uncertainties. All statements contained or implied other than statements of historical facts, including, without limitation, statements regarding Nebius's planned GPU cluster expansion, business plans, market opportunities, capital expenditure requirements, financing requirements and projected financial performance, are forward-looking statements. In some cases, these forward-looking statements can be identified by words or phrases such as "may," "will," "expect," "anticipate," "aim," "estimate," "intend," "plan," "believe," "potential," "continue," "is/are likely to" or other similar expressions. In addition, these forward-looking statements reflect Nebius's current views with respect to future events and are not a guarantee of future performance. Actual results may differ materially from the results predicted or implied by such statements, and Nebius's reported results should not be considered as an indication of future performance. The potential risks and uncertainties that could cause actual results to differ from the results predicted or implied by such statements include, among others, Nebius's ability to successfully operate and develop a fundamentally different, early-stage group following the divestment of a significant portion of our historical operations; to implement Nebius's business plans; to conclude property leases or acquisitions on acceptable terms, to continue to successfully capture customers; to continue to successfully obtain required supplies of hardware on acceptable terms; and to obtain any further debt or equity financing that may be necessary to achieve Nebius's objectives. Many of these risks and uncertainties depend on the actions of third parties and are largely outside of Nebius's control. Notwithstanding the completion of the full divestment of the Company's Russian businesses, Nebius also continue to be subject to many of the risks and uncertainties included under the captions "Risk Factors" and "Operating and Financial Review and Prospects" in Nebius's Annual Report on Form 20-F for the year ended December 31, 2023 and "Risk Factors" in a shareholder circular filed as Exhibit 99.2 to a Report on Form 6-K filed with the U.S. Securities and Exchange Commission ("SEC") on February 8, 2024, which are available on the Nebius investor relations website at

<https://group.nebius.com> and on the SEC website at www.sec.gov. All information in this release is as of November 19, 2024, and the Company undertakes no duty to update this information unless required by law.

In addition, statements that “we believe” and similar statements reflect the Company’s beliefs and opinions on the relevant subject. These statements are based upon information available to Nebius as of the date of this press release, and while Nebius believes such information forms a reasonable basis for such statements, such information may be limited or incomplete, and these statements should not be read to indicate that Nebius has conducted an exhaustive inquiry into, or review of, all potentially available relevant information. These statements are inherently uncertain, and investors are cautioned not to unduly rely upon these statements.