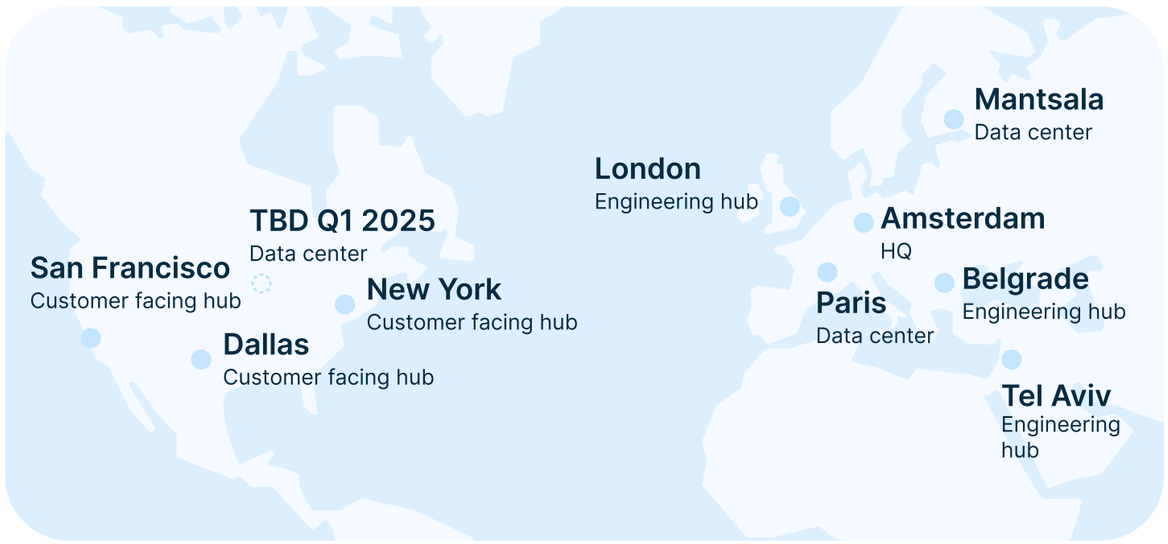
NEBIUS

The ultimate cloud for AI practitioners



An AI-centric cloud platform building large, cost-efficient GPU clusters to service the explosive growth of the global AI industry

Data centers
Improved unit economics through higher energy efficiency reduces data center costs and enables scalability: 250+ MW planned capacity mid-term

Hardware
In-house hardware R&D further improves business unit economics by providing resilient servers and reducing power consumption

Partnership with NVIDIA
Long-standing collaboration on hardware and cloud as a Cloud and OEM partner

AI expertise
Dogfooding by in-house LLM team helps ensure hardware and platform are tuned to the real needs of ML/AI practitioners

Experienced team
500+ employees including ~400 engineers with experience in creating cloud business from scratch, including code development, data center operations and business

Cloud platform
A proprietary cloud platform is essential for creating large GPU clusters optimized for extensive AI training and inferencing without performance bottlenecks

What we offer

Available capacity

- Additional 12k H100/H200 GPUs arriving by the end of 2024
- NVIDIA Preferred partner status gives Nebius early access to new GPU generations, with H200 available in November and GB200 early spring 2025
- Short lead time for customized offers for significant deployments

Custom offerings

- H100 starting from $2.1 per hour for reserve, or on-demand if you need flexibility
- Flexible capacity planning: reserve, on-demand, spots, burst
- Exclusive offer for your portfolio companies

What our customers say: testimonials and recognition

Genesis Therapeutics

“We are grateful to be working with Nebius for our GPU infrastructure needs. Nebius' team is responsive, they clearly communicate capacity / availability, and technical set up has been straightforward. We consider them a strategic partner who can accelerate our state-of-the-art AI research at Genesis Therapeutics.”

Carl Tilbury, BD & Strategy at Genesis Therapeutics

MISTRAL AI_

“Nebius compute and network infrastructure proved to be in the top tiers of what we tried in terms of stability and allowed us to train our models efficiently.”

Timothee Lacroix, CTO of Mistral.ai

Recraft

”In Recraft we are serving image generation models to more than 1 million users. We have our own model trained from scratch, and run training and data processing all the time. We have been using Nebius for those purposes for about a year. Nebius stands out when compared to other clouds, especially if you're looking for flexibility and quick support. With Nebius, you get your own managed Kubernetes cluster, with full admin control to deploy whatever you need without restrictions. Plus, their support is top-notch — super fast, and you get to talk directly with cloud architects, which is a huge plus.

On the scalability front, they always have GPUs available, so no frustrating delays waiting for quotas. The storage speed on Nebius is also higher in comparison to some other clouds.

Overall, Nebius offers more control, better support, and reliable scalability compared to some other clouds. Apart from that, we are the first company to test experimental setups in Nebuis. Most problems in those have been fixed in days. This shows the pace of development in the company.”

Anna Veronika Dorogush, Founder and CEO at Recraft

Our value proposition

Scale your resources

We are a partner you can trust your scaling plans, minimizing risks of dependency on one cloud provider.

Start your growth with us

You can start with smaller clusters (starting from 1 GPU) and scale them to thousands. We support both training and inference needs with flexible pricing plans. Our team is also ready to help with the setup of your first deployment.

Optimize your costs

Architect team provides you with dedicated support with complex deployment for multi cloud configurations. You can continue with your non-GPU workloads on you current Hyperscaler while optimizing your GPU costs on our cloud.

Training scenarios

MISTRAL AI_

Large language models and multi-modal models

- Clusters of thousands GPUs
- Slurm orchestration provided by Nebius AI
- Shared file storage optimized for data streaming and check pointing (reading speed of 100GB per second)

Genesis Therapeutics

AI-powered drug discovery

- Training AI models for drug discovery
- Managed K8s cluster with advanced auto-scaling reduces costs by automatically releasing idle GPUs when not in use, maximizing resource efficiency

JETBRAINS

LLM-powered copilot model

- Cluster of hundreds of H100 GPUs with Infiniband
- NVIDIA Preferred partner status gives Nebius early access to new GPU generations, with H200 available in November and GB200 early spring 2025
- Using Nebius AI API and Terraform support

Inference scenarios

LUMA AI

Multi-modal models for video generation

- Largest supplier of GPU capacity for inference
- Clusters of thousands H100s, support for both single- and multi-host inference
- Managed K8s cluster for inference orchestration

Recraft

Generative AI design tool

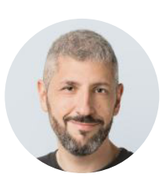
- Nebius AI is main supplier for the inference for their text-to-image 20B model
- Using managed K8s as core of their inference engine
- Storing generated images in Cloudflare R2 via direct 400 Gbps peering between Nebius AI and Cloudflare

Higgsfield

Multi-modal models for video generation

- Clusters of hundreds L40s GPUs optimized for inference
- Several isolated cloud environments for production and pre-prod workloads
- Managed K8s service with GPU auto-scaling to handle consumption spikes

Interested? Let’s get in touch!



Roman Chernin
Chief Business Officer
LinkedIn

Rashid Ivaev
Business Development Director
LinkedIn