NEBIUS

The ultimate cloud for AI innovators

The Nebius AI Cloud brings powerful full-stack infrastructure for AI developers and practitioners across startups, enterprises and science institutes to build and deploy generative AI applications and rapidly deliver scientific breakthroughs by training and running ML models within a secure, scalable, and cost-optimized cloud environment.

We provide access to AI solutions accelerated by NVIDIA and more

- Large-scale GPU clusters
- On-demand GPU instances
- MLOps apps and services
- Latest models, production-ready API

Effective vertical integration

Proprietary hardware design and optimized operations **reduce our TCO by up to 20%** compared to leading market players, enabling highly competitive pricing

Why choose Nebius for your AI journey?

- Immediate access to cutting-edge NVIDIA GPUs
- Unparalleled GPU performance and power efficiency
- Free high-touch customer support for everyone
- Cost-effective for any AI workload
- Predictable and secure cloud experience
- Sustainable data centers in Europe and the U.S.

Worldwide locations

Data centers

We are expanding our data center network across the US and Europe, with more locations planned for 2025

Mäntsälä, Finland **Kansas City, US** **Paris, France** **New Jersey, US** **Keflavík, Iceland**

R&D and customer-facing hubs

Amsterdam (HQ), Dallas, London, San Francisco, Tel Aviv

We serve global AI innovators across industries

JETBRAINS Higgsfield Recraft Genesis Therapeutics CentML وتد Watad

Converge London Institute for Mathematical Sciences PRISMA LABS seqera Unum krisp

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

Recraft

“Nebius stands out when compared to other clouds, especially if you're looking for flexibility and quick support. 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. Nebius offers more control, better support, and reliable scalability compared to some other clouds.”

Anna Veronika Dorogush,
Founder and CEO at Recraft

nebius.com sales@nebius.com

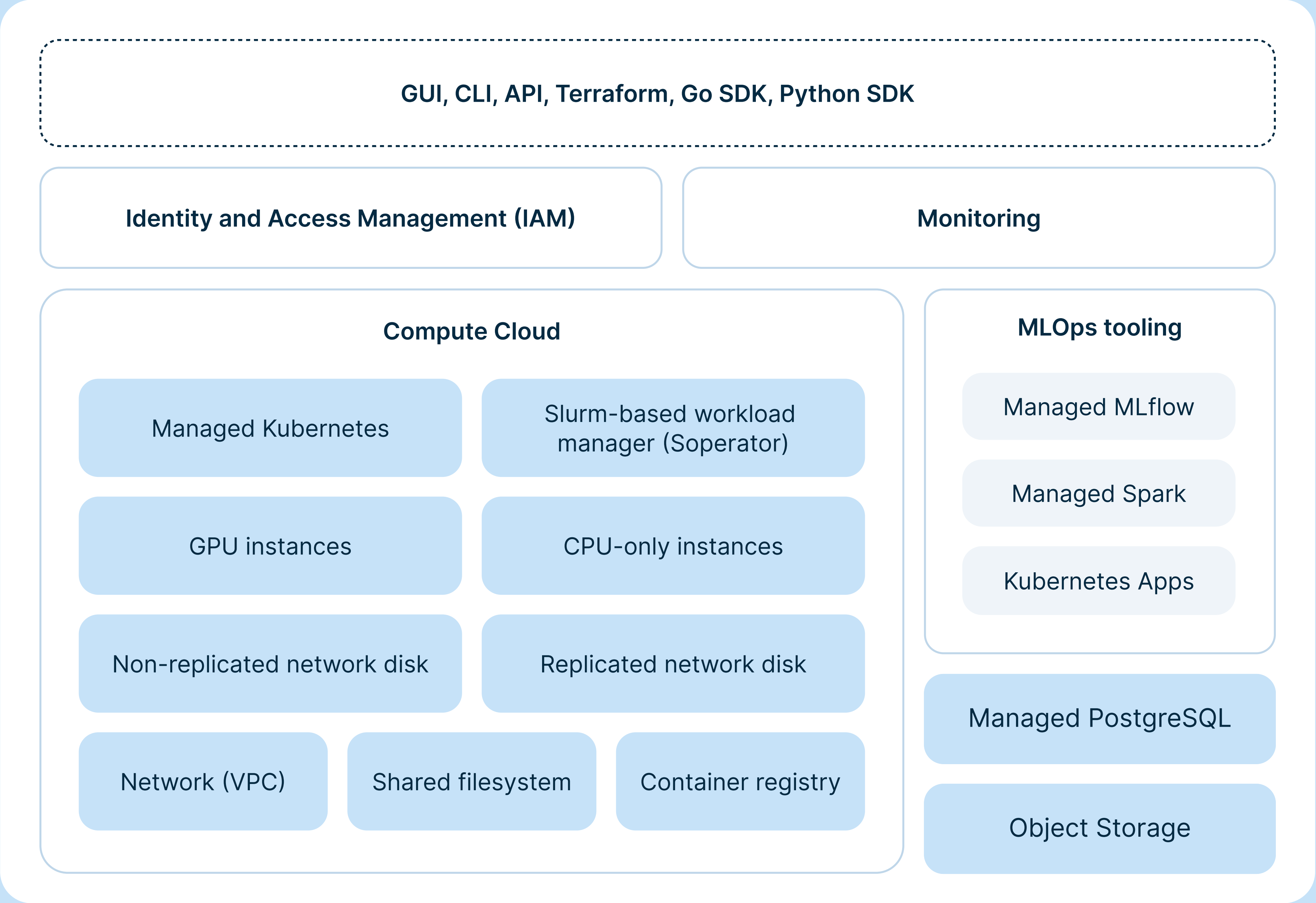
NVIDIA Preferred Partner

Products and services

Explore a unified platform experience for model builders, tuners and AI application developers

Nebius AI Cloud

An AI-native cloud purpose-built for generative and multimodal model development, featuring cutting-edge NVIDIA accelerated servers and MLOps toolset.



GUI, CLI, API, Terraform, Go SDK, Python SDK

Identity and Access Management (IAM)

Monitoring

Compute Cloud

- Managed Kubernetes
- Slurm-based workload manager (Soperator)
- GPU instances
- CPU-only instances
- Non-replicated network disk
- Replicated network disk
- Network (VPC)
- Shared filesystem
- Container registry

MLOps tooling

- Managed MLflow
- Managed Spark
- Kubernetes Apps

Managed PostgreSQL

Object Storage

NVIDIA® GPU lineup

from $0.80/h
NVIDIA
L40S GPU

from $2.00/h
NVIDIA
H100 GPU

from $2.30/h
NVIDIA
H200 GPU

available soon
NVIDIA
Blackwell GPUs

Nebius AI Studio

A complete Inference-as-a-Service solution powering production apps with leading open-source models, unlimited scale, and unmatched cost efficiency. Build with a single API.

Interfaces

API
OpenAI-compatible

Playground
Model comparison

Documentation
Integration guides

Support
Technical assistance

Professional services
Custom solutions

Modalities

General purpose
Llama 3, Mistral, Qwen

Vision
Qwen-VL, LLaVa

Guard
Llama Guard models

Embedding models
BGE, e5-mistral

LoRA adapters

Text-to-image
State-of-the-art image generation

Capabilities

Base and fast flavor
Performance option

Structured output
JSON, markdown

AI tools
RAG, prompting

Batch inference
5M requests/file

Rate limits
10M+ tokens/minute

Model inference service by AI Studio

from $0.13
per 1 million tokens
Llama-3.3-70B-Instruct

from $0.13
per 1 million tokens
Qwen2-VL-72B

from $0.10
per 1 million tokens
Phi-3.5-MoE

And much more!

nebius.com **sales@nebius.com**

Contact us via request form