NEBIUS

NVIDIA Preferred Partner

The global cloud built for production AI

Nebius combines custom hardware, proprietary software and energy-efficient data centers to power modern AI workloads — from the silicon to the app. Whether you're building foundation models, fine-tuning or scaling inference globally, Nebius gives you the performance of a supercomputer with the flexibility of a cloud.

AI infrastructure that works on your terms

Any user

From ML engineers experimenting with models and AI product managers powering theirs apps with AI, to DevOps teams managing hybrid infra. Nebius supports teams at any level of AI maturity.

Any workflow

Train models from scratch, fine-tune and serve existing ones — including open source, by using the workflows that fit your stack. Consume raw GPU compute, easy-to-access ML tooling, or serverless and managed inference, without wasting time on infrastructure or patching things together.

Any workload

Run everything from early experimentation to frontier scale training and global inference. Benefit from industry-specific stacks and deep AI expertise.

Predictable performance. Proven reliability. Lower TCO.

43%
better TCO for fine-tuning vs. AWS.*

112%
better TCO for inference vs. AWS.*

12 min
Mean Time To Resolution.

56.6 hrs
Mean Time Between Failures for 3,000 GPUs.

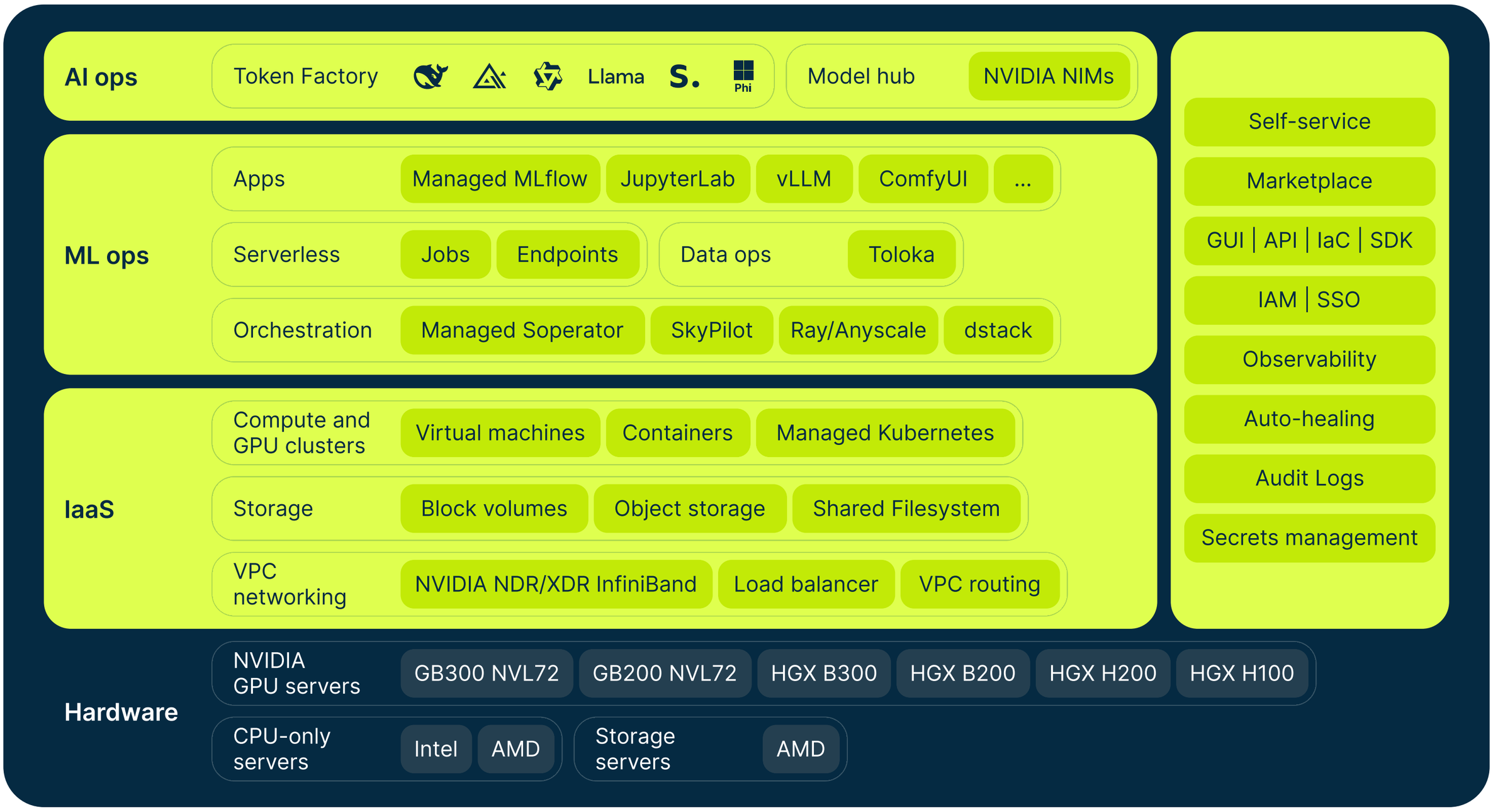
Trusted by leaders

Microsoft · Meta · shopify · World Labs · Black Forest Labs · Tripadvisor · Revolut · PRIME Intellect · brave · Decart · rhoda · VOXEL51 · RoboForce · milestone · JETBRAINS

* Based on the recent TCO study by SemiAnalysis: Calculating the Total Cost of a GPU Cluster, March 2026.

The Nebius end-to-end AI | ML platform

Everything required to build, train and deploy AI — from GPUs clusters to ML tooling.



Built for real-world AI at scale

From idea to AI in minutes

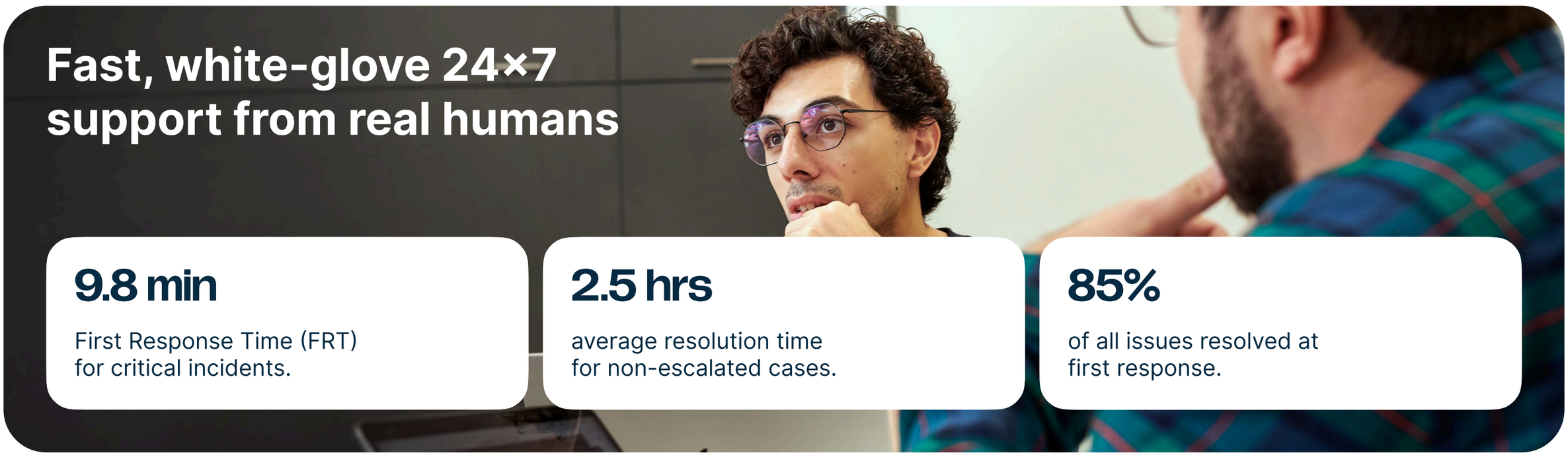
Build and deploy AI-powered applications instantly through simple APIs — without managing GPUs, inference infrastructure or orchestration. Fine-tune open source models in Token Factory, combine with agentic search and access everything self-service.

Run production AI with confidence

Train and run AI on NVIDIA's next generation GPU clusters with InfiniBand networking. Recognized with NVIDIA Exemplar Cloud Partner status and delivered as a true cloud platform with built-in observability, security and compliance — not just a custom bare-metal setup.

Grow at hyper scale

Scale AI workloads on a rapidly expanding global infrastructure built from the ground up for large-scale training and inference. Access high-density optimized GPU clusters and integrated tooling to deliver the capacity and cost-efficiency required to scale your AI in production.



Fast, white-glove 24×7 support from real humans

9.8 min

First Response Time (FRT) for critical incidents.

2.5 hrs

average resolution time for non-escalated cases.

85%

of all issues resolved at first response.

Contact us

nebius.com sales@nebius.com