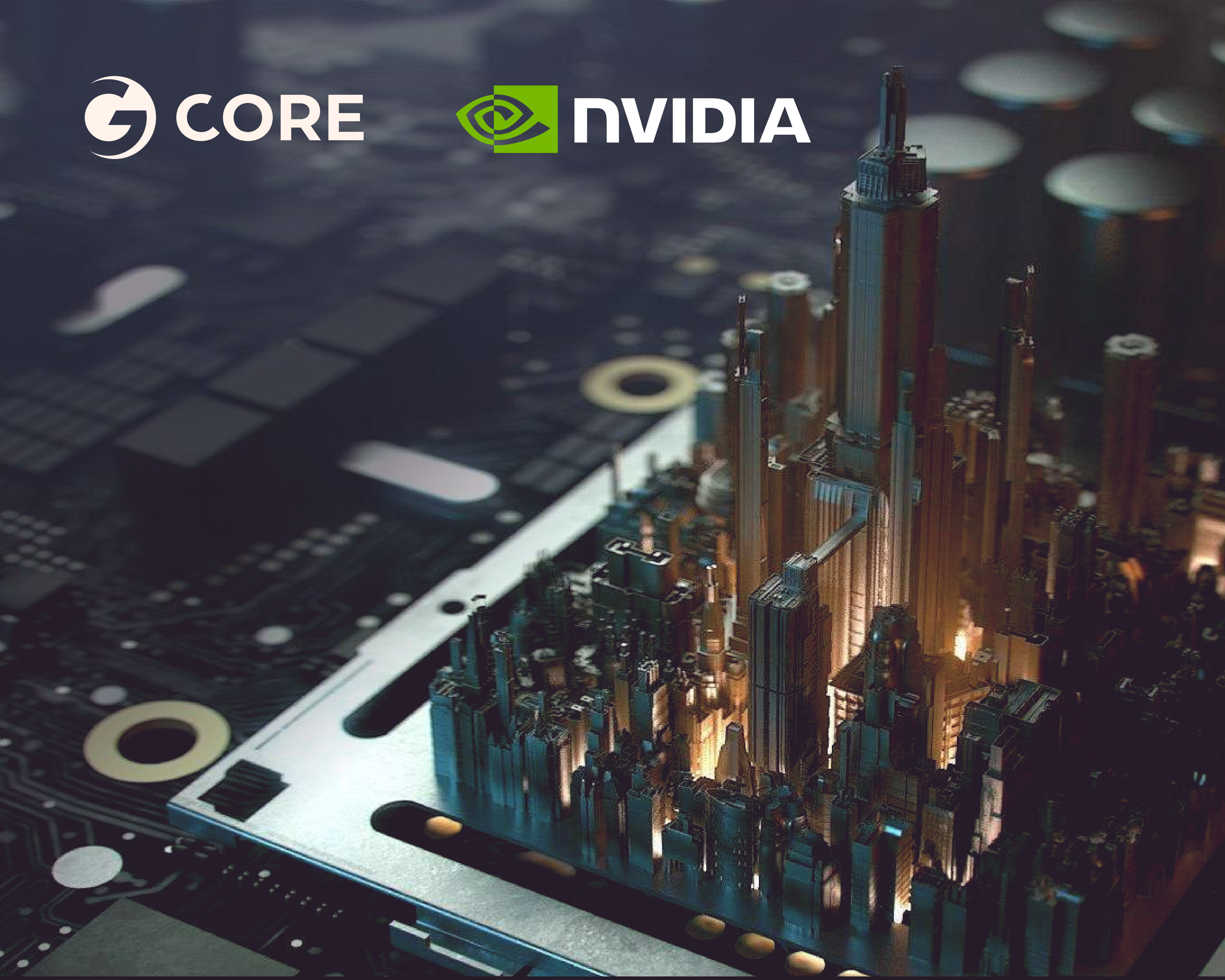

Gcore AI Cloud Infrastructure based on NVIDIA GPU

Increase the efficiency of your AI tasks with the unprecedented power of parallel computing

What is Gcore AI GPU Cloud Infrastructure?

Our GPU cloud infrastructure is based on Bare Metal servers and Virtual Instances, powered by NVIDIA A100 and H100 GPUs. These leading data center-class GPUs accelerate AI/ML training and inference, including those for GenAI and high-performance computing (HPC.)

- **Bare Metal GPU and Virtual GPU Instances provisioned in minutes**
- **Pricing from €1.43 per hour with per-minute billing**
- **AI performance up to 32 PFLOPS**

Why use Gcore AI GPU Cloud Infrastructure?

Gcore has a robust global infrastructure:

150+ PoPs

Across six continents

110+ Tbps

Total network capacity

30 ms

Average response time

11,000+

Peering partners

The Gcore AI GPU Cloud combines Bare Metal servers and Virtual Machines with the most advanced NVIDIA GPUs designed specifically for AI workloads. Our infrastructure eliminates the need to deploy on-premises GPU hardware and build an AI platform from scratch. Simply provision powerful GPU instances as needed and only pay for what you use.

Gcore's GenAI cluster in Europe

We deployed a generative AI cluster based on NVIDIA A100 and H100 GPUs in our Luxembourg data center. This is the first GenAI cluster on this scale in Europe. It offers the significant boost in performance necessary for training large AI/ML models, including those for generative AI.

The cluster consists of 45 A100 GPU-based servers and 128 H100 GPU-based servers. They are unified by InfiniBand interfaces that provide direct GPU connections, ideal for scaling generative AI workloads.

NVIDIA A100 and H100 GPU Features

NVIDIA A100 and H100 Tensor Core GPUs deliver industry-leading performance for a broad range of AI and HPC tasks. The A100 and H100 win the latest MLPerf industry benchmarks for AI training and inference, ahead of platforms like Google TPUv4, Habana Gaudi2, and Intel Xenon Platinum 8480+.[1,2] MLPerf benchmarks test eight workloads across different use cases, including computer vision, large language models, and recommender systems.

NVIDIA GPUs are the perfect solution for achieving business goals with deep learning, HPC, graphics, and virtualization in the data center or at the edge. With 100x higher performance in AI training than CPUs,[3] GPUs also enable faster tuning and training of deep learning models, saving money on compute resource rentals.

A100 Tensor Core GPU

- Up to 100x higher AI training performance over CPUs
- Ampere Tensor Core (3rd generation)
- Up to 80 GB of HBM2e memory
- 600 GB/s NVLink interconnect

H100 Tensor Core GPU

- Up to 4x higher AI training performance than the A100
- Hopper Tensor Core (4th generation)
- Up to 100 GB of HBM3 memory
- 900 GB/s NVLink interconnect

[1] https://blogs.nvidia.com/blog/2022/06/29/nvidia-partners-ai-mlperf/
[2] https://blogs.nvidia.com/blog/2023/06/27/generative-ai-debut-mlperf/
[3] https://www.nvidia.com/en-us/data-center/dgx-a100/

Benefits for key industries

Gcore’s GPU-based AI Cloud is designed to help businesses accelerate deep learning (DL) and deploy compute-intensive workloads across various sectors, including financial services, healthcare, scientific research, and game development.

Healthcare

- DL for diagnostics
- Medical imaging
- Drug discovery
- Real-time patient data monitoring
- Robot-assisted surgery
- Personalized medicine

Scientific research

- Simulations and modeling
- Astrophysics and cosmology
- Genomic sequencing
- Climate modeling
- Molecular dynamics simulations
- Protein folding

Game development

- 3D rendering and animation
- High frame rates
- Physics simulation
- Advanced shader effects
- Post-processing effects
- VR and AR

Financial services

- Risk assessment
- Fraud detection
- Trading acceleration
- Regulatory compliance
- Complex data analysis
- Option pricing

With the Gore AI GPU Cloud, you can quickly deploy GPUs and pay only for the resources you consume. Sensitive data is reliably protected: Our infrastructure complies with PCI DSS, ISO/EIC 27001, and GDPR requirements.

More Gcore AI GPU Cloud Infrastructure Features

- Build, train, and deploy ready-to-use DL models via the control panel, API, or Terraform
- Flexible IaaS capabilities, including direct connect for multicloud and on-premises
- Dataset management and integration with S3/NFS storage
- Secure trusted cloud platform, complied with ISO/EIC 27001, PCI DSS, and GDPR requirements
- Multiple European data centers
- SLA 99.9% guaranteed uptime
- Highly skilled 24/7 technical support

Gcore products are trusted by

WARGAMING.NET LET'S BATTLE · RedFox Games · WARPCACHE · SHOPCADA

SANDBOX INTERACTIVE · BANDAI NAMCO · avast · SynEdge

ZUMIDIAN · NANOBIT · AGENCE eSanté LUXEMBOURG · momento solutions

JSDELIVR · SABER AN EMBRACER GROUP COMPANY · api.video · StageAudioWorks™ TECHNOLOGY & ENGINEERING GROUP

Contact us and go global faster

Gcore is an international leader in public cloud and edge computing, content delivery, hosting, and security solutions.

We manage a global infrastructure that provides enterprise-level businesses with first-class edge and cloud-based services.

+352 208 80 507 | sales@gcore.com | gcore.com