GCORE GRAPHCORE

Gcore AI Cloud infrastructure based on Graphcore IPUs

Flexible, scalable, and cost-effective solution for accelerating results in ML tasks.

What is Gcore AI Cloud infrastructure?

Our cloud infrastructure is based on Graphcore IPUs for powerful AI computing. With its ready-made AI infrastructure, customers can quickly accelerate ML, train and compare models, or train custom code.

- **AI cluster deployment in minutes**
- **Pricing from €5.10 per hour**
- **Performance level from 4 to 89.6 petaflops of AI compute**

Why use Gcore AI Cloud infrastructure?

Gcore has a robust worldwide cloud infrastructure.

140+ PoPs

on six continents

110+ Tbps

Total network capacity

30 ms

Average response time

11000+

Peering partners

To help you accelerate ML, we built AI Cloud on top of the infrastructure. The AI Cloud combines servers with Graphcore IPU AI compute engines. This means there is no need to deploy on-premises hardware and create AI infrastructure from scratch. Easily add state-of-the-art machine intelligence computing on demand.

Graphcore IPU features

The IPU is an entirely new massively parallel processor designed from the ground up with the Poplar® SDK to accelerate machine intelligence. Support for industry standard ML frameworks such as PyTorch and TensorFlow facilitates new ML model development and porting of existing models. The Graphcore model garden provides a wide variety of models across a range of ML domains including NLP, CV, and GNNs.

IPUs show significantly higher performance for different AI workloads than their alternatives, as shown below in Natural Language processing (NLP) which is currently the most important field of machine learning, and Graph Neural Networks (GNNs) which is the fastest growing area of machine learning. Many more performance results and customer case studies are available at www.graphcore.ai.

Natural Language Processing (NLP)	**BERT-Large Training**	Bow Pod16 is over 2x faster than a DGX A100
	BERT-Large Inference	Bow-Pod4 delivers 2x higher throughput within 5ms latency threshold compared to the A100
Graph Neural Networks (GNNs)	**TGN (Temporal Graph Networks)**	“Up to an order of magnitude speedup when comparing a single IPU processor to an NVIDIA A100 GPU” - M. Bronstein (DeepMind Professor of AI, University of Oxford) [1]
	Open-Graph Benchmarks – Large Scale Challenge (OGB-LSC)	Double-first position in the Open Graph Benchmark Large-Scale Challenge - the AI industry’s leading test of graph network model capability [2]

Gcore AI Cloud server clusters are already available in Luxembourg and Amsterdam. You can choose from various configurations based on Graphcore Bow Pod & IPU-POD Classic available on a pay-as-you-go basis.

[1] https://www.graphcore.ai/posts/accelerating-and-scaling-temporal-graph-networks-on-the-graphcore-ipu

[2] https://www.graphcore.ai/posts/graphcore-claims-double-win-in-open-graph-benchmark-challenge

Benefits for key industries

Gcore's IPU-based AI Cloud is designed to help businesses across a broad range of sectors, including Healthcare, Financial Services, Scientific Research, Consumer Internet & Media, and Manufacturing.

Accelerating chatbot solutions & delivering faster insights with NLP

Gcore IPU-based AI Cloud delivers unique benefits for production-level transformer workloads for NLP:

- Provide faster answers with lower latency
- Reduce costs with attractive pricing combined with superior performance
- Get up & running quickly with model garden of commonly used transformer models

See how the IPU has helped accelerate NLP solutions:

- Pienso & Graphcore empower business with deeper, faster insights
- Aleph Alpha use IPUs to demonstrate sparsified chatbot model

Benefits for scientific researchers and acceleration of drug discovery in healthcare using GNNs

The IPU's unique architecture enables innovation in the fast-growing domain of GNNs across many sectors:

- Accelerate drug discovery with protein-protein interaction modelling
- Advance healthcare research with molecular property prediction
- Accelerate dynamic graph models for social network analysis

See how the IPU is already enabling innovation using GNNs:

- Improving journey time predictions with GNNs on the Graphcore IPU
- Accelerating & Scaling Temporal Graph Networks on the Graphcore IPU

With the Gcore AI Cloud, you can quickly connect IPUs and pay only for the resources you consume. Sensitive data is reliably protected: our platform complies with PCI DSS, ISO/IEC 27001 and GDPR requirements.

More Gcore AI Cloud infrastructure features

- Build, train and deploy ready-to-use ML models via control panel, API, or Terraform
- Flexible IaaS capabilities including direct connect for multi-cloud & on-prem
- Dataset management and integration with S3/NFS storage
- Secure Trusted Cloud platform, complied with ISO/IEC 27001, PCI DSS and GDPR requirements.
- Multiple European data centres
- SLA 99.9% guaranteed uptime and highly skilled technical support 24/7

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