# Nvidia launches DGX Cloud Lepton to unify global GPU infrastructure for AI



Nvidia has unveiled DGX Cloud Lepton, a global platform designed to unify thousands of distributed GPUs into a single next-generation AI infrastructure. The launch follows Nvidia’s acquisition of Chinese startup Lepton AI earlier this year, co-founded by Yancing Jia, former Vice President of Technology at Alibaba, and Junjie Bai. Both have joined Nvidia to lead the project.

DGX Cloud Lepton offers AI developers a unified entry point into a global ecosystem of GPUs, removing geographical and provider limitations. It mirrors the early cloud revolution in storage by offering seamless and scalable access to compute power, regardless of underlying infrastructure.

The move comes as the cloud GPU market expands rapidly. Valued at $3.17 billion in 2023, it is expected to reach nearly $47 billion by 2032, with GPU as a Service forecast to grow from $5 billion in 2025 to $32 billion by 2034. Nvidia’s platform consolidates compute from major providers including AWS, CoreWeave, Lambda and Microsoft Azure, all accessible through a single software interface built on Nvidia’s AI stack.

This stack integrates key tools: NVIDIA Inference Microservices for rapid inference; NeMo for building and fine-tuning large models; Blueprints for assembling applications; and Cloud Functions for deploying modular AI tasks. Together, they offer a predictable and flexible environment for AI development.

DGX Cloud Lepton supports three workflows: Dev Pods for experimentation with tools like Jupyter Notebooks and VS Code; Batch Jobs for training and dataset generation across nodes with real-time monitoring; and Inference Endpoints for scalable API deployment with automatic resource scaling.

The platform also offers advanced features including observability tools, automated scaling, workspace customisation and regulatory compliance, making it suitable for startups, enterprises and government projects requiring secure and scalable infrastructure.

Nvidia is positioning DGX Cloud Lepton not as a rival to hyperscale providers, but as an integrative layer unifying their GPU resources into an open AI compute marketplace. Key partnerships with AWS, Microsoft Azure, CoreWeave and Hugging Face support this approach. Hugging Face provides training clusters for open model development, while venture-backed startups receive GPU credits and technical support to accelerate innovation.

DGX Cloud Lepton also responds to regional regulatory demands, offering localised GPU access to support data sovereignty and compliance, particularly in Europe and other regulated markets.

By consolidating global GPU capacity under a single interface, DGX Cloud Lepton could reshape AI infrastructure access. It supports ambitions to position the UK and other regions as leaders in responsible AI by enabling developers to scale efficiently without infrastructure barriers.

With demand for AI compute power growing, Nvidia’s platform signals a shift towards open, compliant and globally connected infrastructure—one that could accelerate the next wave of AI innovation.

Created by [Amplify](https://www.hbmadvisory.com/amplify): AI-augmented, human-curated content.

## Bibliography

1. <https://www.computerra.ru/317059/nvidia-zapuskaet-globalnuyu-platformu-ii-infrastruktury-dgx-cloud-lepton/> - Please view link - unable to able to access data
2. <https://nvidianews.nvidia.com/news/nvidia-announces-dgx-cloud-lepton-to-connect-developers-to-nvidias-global-compute-ecosystem> - NVIDIA has introduced DGX Cloud Lepton, an AI platform that connects developers to a global network of GPU compute resources. This platform integrates with NVIDIA's software stack, including NIM and NeMo microservices, Blueprints, and Cloud Functions, to streamline the development and deployment of AI applications. Cloud providers participating in DGX Cloud Lepton offer management software that delivers real-time GPU health diagnostics and automates root-cause analysis, reducing manual operations and downtime. The platform aims to improve productivity and flexibility for developers by offering a unified experience across development, training, and inference.
3. <https://www.nvidia.com/en-us/data-center/dgx-cloud-lepton/> - NVIDIA's DGX Cloud Lepton is an AI platform that connects developers to a global network of GPU compute resources. It offers a unified experience across development, training, and inference, allowing developers to access GPU capacity directly from participating cloud providers through the marketplace or bring their own compute clusters. The platform enables deployment of AI applications across multi-cloud and hybrid environments with minimal operational burden, using integrated services for inference, testing, and training workloads. It also provides access to GPU resources in specific regions, supporting strategic and sovereign AI operational requirements.
4. <https://www.reuters.com/business/media-telecom/nvidia-software-aims-create-marketplace-ai-computing-power-2025-05-19/> - NVIDIA has announced Lepton, a new software platform designed to create a centralized marketplace for cloud-based AI chip capacity. The platform aims to streamline the process of accessing NVIDIA's GPUs, which are widely used for training AI models. Companies such as CoreWeave, Nebius Group, Crusoe, Foxconn, and SoftBank have joined Lepton, enabling them to offer GPU capacity to developers more efficiently. Lepton is expected to replace the current manual system of finding available compute resources and make NVIDIA’s extensive developer ecosystem more accessible across both established and emerging cloud providers.
5. <https://www.nasdaq.com/press-release/nvidia-dgx-cloud-lepton-connects-europes-developers-global-nvidia-compute-ecosystem> - NVIDIA's DGX Cloud Lepton connects Europe's developers to the global NVIDIA compute ecosystem. The platform integrates with NVIDIA's software stack, including NIM and NeMo microservices, Blueprints, and Cloud Functions, to accelerate and simplify the development and deployment of AI applications. Cloud providers participating in DGX Cloud Lepton offer management software that delivers real-time GPU health diagnostics and automates root-cause analysis, reducing manual operations and downtime. The platform aims to improve productivity and flexibility for developers by offering a unified experience across development, training, and inference.
6. <https://www.engineering.com/nvidia-unveils-dgx-cloud-lepton-for-developer-access/> - NVIDIA has unveiled DGX Cloud Lepton, a platform that connects developers to a global network of GPU compute resources. The platform integrates with NVIDIA's software stack, including NIM and NeMo microservices, Blueprints, and Cloud Functions, to accelerate and simplify the development and deployment of AI applications. Cloud providers participating in DGX Cloud Lepton offer management software that delivers real-time GPU health diagnostics and automates root-cause analysis, reducing manual operations and downtime. The platform aims to improve productivity and flexibility for developers by offering a unified experience across development, training, and inference.
7. <https://docs.nvidia.com/dgx-cloud/lepton/guides/get-started/quickstart/index.html> - The NVIDIA DGX Cloud Lepton Quickstart guide provides instructions for setting up the DGX Cloud Lepton platform. After creating a workspace and activating it, users can request a node group to create workloads. The guide also outlines next steps, including creating a development pod and submitting a training job. For further assistance, users are encouraged to reach out to their Technical Account Manager (TAM).