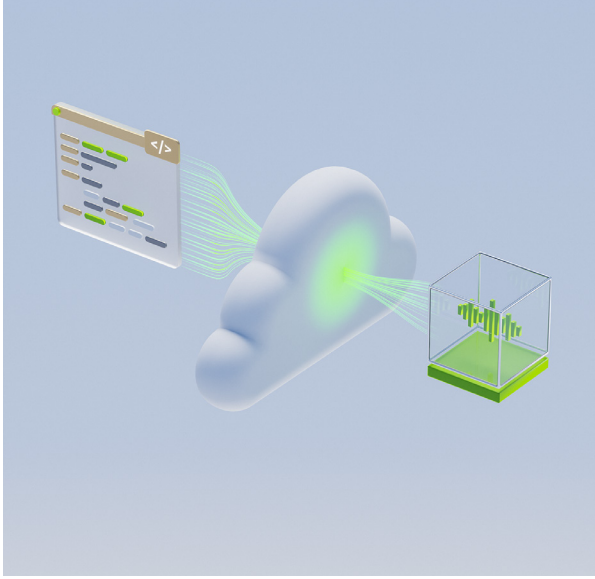
Solution Overview

NVIDIA DGX Cloud

A fully-managed AI platform for training and fine-tuning.

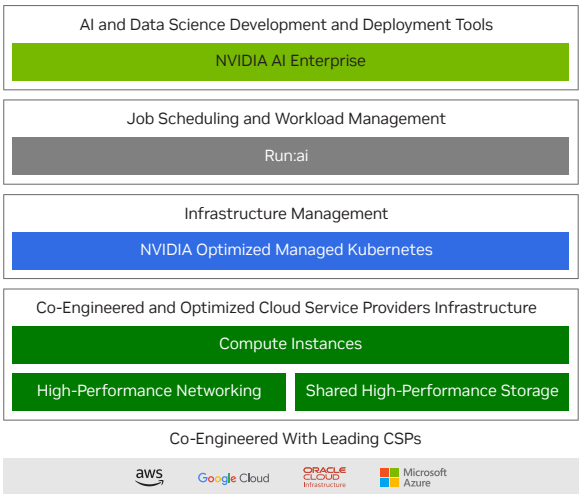


The Many Barriers to Realizing True Generative AI ROI

Enterprises looking to deploy generative AI applications face a number of challenges, including data management, model transparency, legacy IT infrastructure, compute resources, technical debt, and talent shortage. Among the barriers to digital infrastructure investment for generative AI, 33.2% of enterprises identify a lack of necessary IT staff and skills, 30.93% indicate insufficient IT automation and observability, and 21.88% cite tech debt from mission-critical AI workloads.[1] However, organizations that successfully manage AI infrastructure, optimize compute resources, leverage proprietary data for AI model customization, and have access to the right AI expertise are achieving the desired generative AI outcomes that speed time to value.

DGX Cloud: The NVIDIA-Optimized AI Cloud

NVIDIA DGX Cloud is a high-performance, fully managed AI platform designed to deliver productivity from day one. Designed to accelerate training at every layer, DGX Cloud lets enterprises use the latest NVIDIA AI architecture and software on any leading cloud, with flexible, short-term durations, seamless multi-cloud portability, enterprise support, and access to NVIDIA experts for maximum ROI.



Key Features of NVIDIA DGX™ Cloud

Hardware*

- Contiguous clusters to decrease latency
- Multi-node enabled
- Eight NVIDIA H100 Tensor Core GPUs per node
  - up to 3,200 gigabits per second of node-to-node bandwidth
  - 10TB minimum of storage per node included
- Seventy-two GPUs per NVIDIA GB200 NVL72 domain (coming soon)

Services

- Scalable compute clusters with flexible term lengths: 1–12 months
- Access to NVIDIA experts
- 24/7 business-critical support
- Designated technical account manager (TAM)
- Single-point-of-contact support

*GPU availability may vary per cloud service provider (CSP).

Unlock Flexibility and Support Multi-Cloud Strategies With an Open Platform

Many enterprises use different platforms and cloud services to train and deploy models, adding to the complexity of AI model portability and data movement. The absence of AI platforms, insufficient compute resources, and the unpredictability of AI training can drive up costs and limit the flexibility needed to bring models into production more quickly.

NVIDIA DGX Cloud is an open platform that gives you the flexibility to deploy and run your machine learning models and pipelines across a variety of environments. Whether you're working within a single cloud or leveraging a multi-cloud strategy, DGX Cloud supports effortless migration and integration, giving enterprises AI model portability to any compute environment.

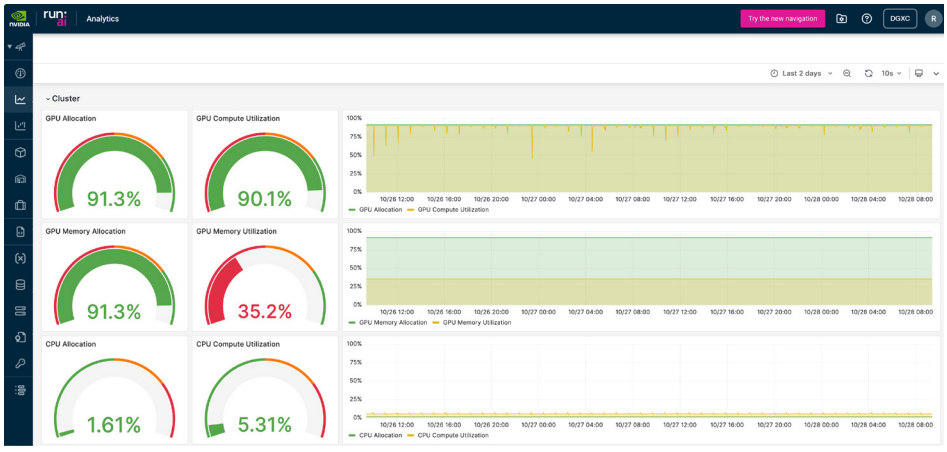
DGX Cloud also brings a vast ecosystem of machine learning tools to cloud customers, helping you leverage a broad range of cutting-edge technologies without being confined to a single platform. As enterprise AI workloads and model sizes grow, DGX Cloud also provides flexible, short-term access to larger high-performance compute clusters.

Accelerate Innovation and Realize AI Impact Sooner

In today's fast-paced market, the speed at which you can innovate and deploy AI models is crucial. DGX Cloud lets you accelerate your AI initiatives and unlock business value faster, all while optimizing infrastructure and resources.

Day-one productivity: DGX Cloud provides a fully managed, preconfigured environment ready for use from day one. Co-engineered with leading clouds and optimized across every layer—compute, storage, networking, and managed Kubernetes—DGX Cloud accelerates training and fine-tuning by eliminating the undifferentiated heavy lifting of building and maintaining large-scale compute clusters.

Maximum GPU utilization: DGX Cloud leverages Run:ai software, which includes features for optimizing GPU resources. This maximizes throughput, reduces idle time, and improves training efficiency, while delivering predictable costs.



The DGX Cloud UI provides advanced telemetry and cluster observability to achieve high GPU utilization and productivity on day one.

Faster training and lower latency: With contiguous clusters designed for high-performance workloads, DGX Cloud reduces training time and lowers latency, enabling rapid iteration and accelerated training for your AI models.

Key Features of the DGX Cloud Software Stack

- Workload scheduling, prioritization and preemption for maximum resource utilization
- Hierarchical quota management
- User and team management
- Graphical user interface (GUI) and command-line interface (CLI) capabilities
- Ability to launch and manage interactive workloads and connect to JupyterLab
- Built-in telemetry, including cluster and workload observability
- Ability to deploy and integrate cloud-native AI ecosystem tools and applications

Data proximity with CSPs: Bring faster training closer to your data by leveraging CSP environments while minimizing data transfer expenses. By locating compute resources near your data, you ensure lower latency and improved performance for data-heavy training processes.

Optimized for industry-specific use cases: Tailored NVIDIA NIM microservices and reference AI workflows ensure that DGX Cloud meets the unique requirements of your industry. Whether you're in healthcare, finance, or manufacturing, our platform is built to accelerate the deployment and impact of your AI models with minimal customization.

With DGX Cloud, you can bring innovation to market faster, reduce training times, and realize the transformative potential of AI more quickly—all while ensuring the best performance for your enterprise.

The Best Way to Experience NVIDIA AI in the Cloud

Easily access the power of NVIDIA AI in the cloud: DGX Cloud provides an optimized, end-to-end platform that integrates NVIDIA best-in-class technologies, enabling you to fully unlock the potential of AI faster and more efficiently.

Faster access to the latest NVIDIA architecture: Stay ahead of the curve with early access to the latest NVIDIA GPU architectures and AI innovations. Our platform is built to integrate with new NVIDIA technologies, ensuring your AI workloads are always powered by the most advanced computing solutions available.

Full-stack platform, from compute to software: DGX Cloud delivers a comprehensive, fully integrated stack—from the underlying infrastructure to the software layer. This all-in-one platform ensures that your AI initiatives are built on a reliable, high-performance foundation, reducing complexity and streamlining deployment.

Enterprise-grade software, workflows, and AI models: Access the entire NVIDIA AI Enterprise software suite, including reference AI workflows and pretrained models. Whether you're developing custom models, integrating existing solutions, or deploying an NVIDIA NIM microservice for accelerated AI inference, the NVIDIA AI Enterprise platform provides the tools and resources to accelerate your AI projects.

NVIDIA expertise: With DGX Cloud, you gain more than just access to NVIDIA technology—you also gain access to NVIDIA professionals. From initial setup to ongoing support, our team provides the guidance and support necessary to maximize the impact of your AI initiatives.

With the latest NVIDIA technology, flexible term lengths, and access to NVIDIA experts, DGX Cloud is a high-performance, fully managed AI platform that provides optimized clusters on any cloud to accelerate your AI initiatives. Spend less time managing infrastructure and waiting for compute resources, and start training on day one.

Key Features of NVIDIA AI Enterprise

- NVIDIA NIM™ is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing.
- NVIDIA NeMo™ is an end-to-end framework for building, customizing, and deploying enterprise-grade generative AI models.
- SDKs and frameworks support AI applications across many domains.
- Tools and libraries accelerate data analytics and AI model training and customization.

> "We benefited from NVIDIA's end-to-end support, ranging from platform assistance for multi-node training setup and container updates to application-level guidance, leveraging their extensive expertise in healthcare frameworks and models to optimize our AI models effectively."
>
> **Dan Ferrante**,
> AI Leader for Innovation and R&D,
> Deloitte Consulting LLP

Ready to Get Started?

To learn more about NVIDIA DGX Cloud, visit:
www.nvidia.com/dgx-cloud

To learn how to train your custom AI model on NVIDIA AI Foundry powered by DGX Cloud, visit:
www.nvidia.com/foundry

1. Source: Future of Enterprise Resiliency and Spending Survey Wave 7, IDC, August 2023.

NVIDIA