Solution Overview

Build Your Models Faster With Serverless AI on NVIDIA DGX Cloud

AI training as a service for the era of generative AI.

The Limitations of Traditional AI Infrastructure

The development of enterprise AI needs accelerated computing to derive insights from oceans of unstructured data. With the rise of generative AI, it's critical to have infrastructure that can support the demands of developers creating this new wave of AI-infused applications. Many are turning to the cloud, but unlike mainstream enterprise workloads, AI places new demands that aren't sufficiently supported by traditional cloud infrastructure.

From CPUs to GPUs to GPU clusters, businesses are finding themselves adopting increasingly complex infrastructure to tame their AI model training. Today's models require an unprecedented level of scale, with multi-node infrastructure being the minimum for those who want to deliver business value from leading-edge AI. But getting the most productivity from this investment requires more than fast GPUs. If you're leading an AI initiative, you might have these concerns:

- How can my developers get the scale of compute resources they need?
- How can they manage their AI workflows and keep their jobs running?
- How can they adapt their work to fit within existing cloud tools?
- What's the business cost if infrastructure hampers developer efficiency?

Many businesses realize too late that, without the right platform, they're seeing marginal returns on their infrastructure investment.

AI Development Requires a Full-Stack Platform

Today's seemingly "low-cost" infrastructure-as-a-service (IaaS) offerings aren't optimized for AI at enterprise scale. The reality for most businesses is that traditional infrastructure can't address the demands of AI initiatives. Enterprise developers are exposed to:

- Greater risk of downtime with increasingly complex infrastructure
- Eroded productivity with effort spent on keeping their jobs running
- Non-linear performance as resources are scaled without optimization
- Inefficient resource-to-job allocation

Key Features of NVIDIA DGX™ Cloud

Software

- Full-stack software suite to streamline AI model development
- Automated infrastructure management
- NVIDIA Base Command™ Platform
- NVIDIA AI Enterprise software suite
- Hybrid-cloud support

Hardware

- Network fabric, purpose-built for multi-node training
- Multi-node capable
- Eight NVIDIA A100 or H100 Tensor Core GPUs per node (640GB total)
- Near limitless access to GPUs
- 10TB of storage per instance

Services

- Access to NVIDIA AI experts
- 24/7 business-critical support
- Technical account manager
- Customer service success manager
- Single-point-of-contact support

The cumulative impact is a risk of failure that grows with the size of the work undertaken.

More than an infrastructure service, developing enterprise AI requires a full-stack platform that includes:

- **Frameworks and tools** for developing and deploying domain-specific, end-to-end AI workflows—from data prep and training to inference and deployment
- **A developer-centric platform** to manage everything from the simplest computer vision application to the most complex large language models (LLMs)
- **An intelligent orchestration layer and scheduler** that delivers right-sized resources to every job, reclaiming and reassigning resources dynamically and efficiently
- **Automated infrastructure management** that maximizes the platform's performance and uptime for worry-free execution of every job
- **An ultra-high-bandwidth, low-latency network fabric** that's purpose-built for multi-node training and can parallelize large AI models over many accelerated compute nodes

Modern AI enterprises need all of these capabilities integrated within their platform, so their developers can focus on creative experimentation with leading-edge AI models instead of wrestling with infrastructure limitations

NVIDIA DGX Cloud: The World's First AI Platform Purpose-Built for Developing Generative AI

NVIDIA DGX Cloud addresses the challenges of today's traditional IaaS offerings. Delivering up to three times the utilization efficiency of traditional infrastructure, it vastly improves training performance and developer productivity.

An AI Platform That Puts Developers First

DGX Cloud integrates **NVIDIA Base Command Platform** for easy-to-use, streamlined AI development. Base Command Platform efficiently configures and manages AI workloads, delivers integrated dataset management, and executes them on right-sized resources ranging from a single GPU to large-scale, multi-node clusters. DGX Cloud also includes **NVIDIA AI Enterprise**, which offers accelerated data science libraries, optimized frameworks, and pretrained models that give developers a faster path to production-ready models..

Key Features of NVIDIA Base Command Platform

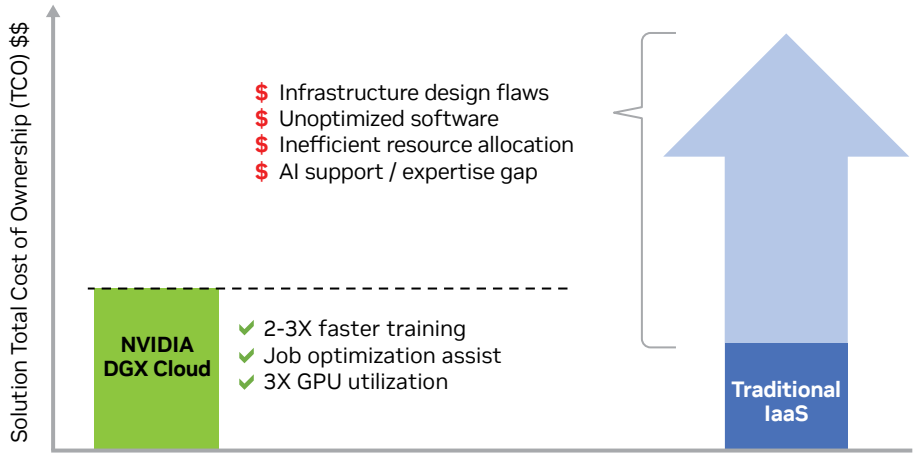
- Role-based access control (RBAC)
- Teams and projects sharing
- Showback and chargeback
- Utilization and resource management
- Quick Start Jupyter access
- Development tooling
- Data management
- Model management
- MLOps workflows
- Hyperparameter optimization
- Job checkpointing and pre-emption

Your Own Dedicated Serverless AI Factory

Data scientists and developers need a serverless AI development platform that includes frictionless scale of GPU resources optimized for multi-node training. Unlike traditional IaaS offerings, DGX Cloud lets developers onboard quickly and experiment rapidly without having to worry about the underlying hardware or refactoring their code. DGX Cloud is based on the leading-edge **NVIDIA DGX technology**, a solution that accelerates innovation in every industry.

Fastest ROI for Your AI Endeavors

With DGX Cloud, your team can focus on innovating with AI instead of struggling with infrastructure, resulting in faster time to insights. You'll experience the benefits of reliable, rapid allocation of resources, fully optimized workloads with error-free execution, and less developer idle time, resulting in a lower overall total cost of ownership (TCO) compared to traditional infrastructure services.



Total Cost of Ownership: DGX Cloud vs. IaaS

Expert Services Accelerate Your AI Journey

Going beyond traditional IaaS, DGX Cloud includes proactive support from NVIDIA AI experts to help you get results faster. You can improve your team's efficiency and utilization with direct guidance, support, and best practices from the industry's leading practitioners. These AI experts include a dedicated technical account manager (TAM), solution architects, a customer service manager, NVIDIA Training instructors, and NVIDIA Professional Services, all backed by 24/7 **NVIDIA Enterprise Support**.

DGX Cloud Compared With Alternatives

A typical scenario—building a large language model to power a customer service chatbot—illustrates the differences between developing generative AI on traditional services and DGX Cloud.

Assume that:

- The model is of modest complexity, for example GPT-3 with 40 billion parameters and 300 billion tokens
- Training the model in approximately one month requires 160 NVIDIA A100 GPUs

> "NVIDIA DGX Cloud is the gold standard for single- and multi-node training. Distributed training was more than twice as fast as other leading services that don't feature NVIDIA networking."
>
> **Michael Royzen**,
> Co-founder and CEO, Phind

> "The responsiveness of the NVIDIA AI experts who supported getting our codes running efficiently on their platform was key. Instead of going to forums, we got answers about our infrastructure and tooling in real time. Because of this, my team can focus on modeling, not software engineering."
>
> **Christopher James Langmead**,
> Director of Digital Biologics Discovery, **Amgen**

	With DGX Cloud	With IaaS
GPU infrastructure	20 DGX Cloud instances with NVIDIA A100 GPUs	20 eight-GPU instances
Resource onboarding and training readiness	Less than one week	Weeks to months, depending on provider
Cluster design	Prebuilt, multi-node architecture with high-bandwidth, low-latency NVIDIA networking at predictable best-in-class performance	Instances may not be co-resident or interconnected over NVIDIA networking, risking potentially 30–50% lower performance
Storage	Included	Not included
LLM software	NVIDIA AI Enterprise with NVIDIA NeMo™	Not included
Software engineering, code adaptation, or code refactoring	None required	Weeks to months, depending on provider
Access to NVIDIA Enterprise Support and premium TAM	Included and Enterprise Support	Not available; community/forum support
DGX Cloud training speedup	**2–3X over traditional IaaS**	Assume 40–60 more eight-GPU instances to match DGX Cloud
Summary		
Time to production-ready model	Approximately one month (can be reduced to 10 days with DGX Cloud H100 instances)	**Over three months** (includes time lost on cluster design, procurement, software engineering, training troubleshooting)
Estimated savings	Approximately one-third the cost of IaaS	Approximately 3X the cost of DGX Cloud

DGX Cloud can easily result in having a production-ready AI model in one month instead of three by sparing you from validating and optimizing your IaaS provider's infrastructure design, engineering your software stack, and scouring open-source community forums for answers as training issues arise.

Ready to Get Started?

To start building mission-critical AI now and deliver AI-fueled insights with a lower TCO, visit: **nvidia.com/dgx-cloud**

Contact Sales at: **nvidia.com/dgx-cloud/trial**

NVIDIA