NVIDIA
SOLUTION AT-A-GLANCE

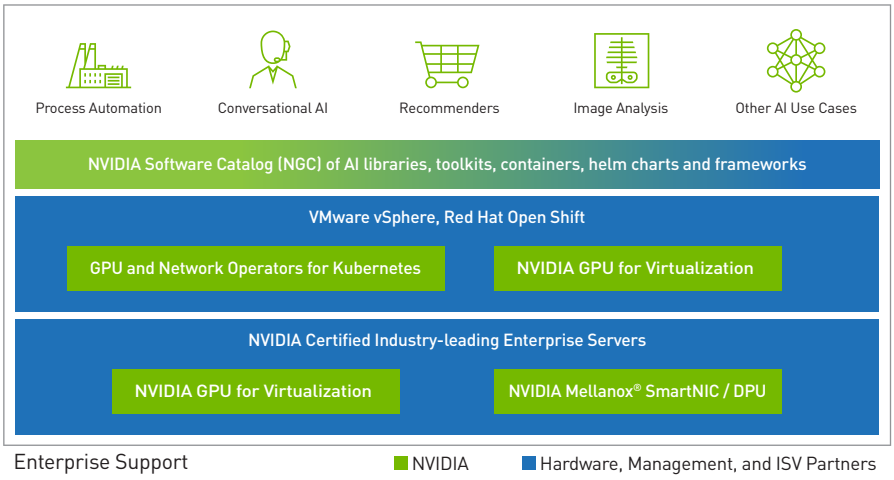
ACCELERATE AI TRAINING AND INFERENCE ON DATA CENTER INFRASTRUCTURE

Artificial Intelligence in the Modern Enterprise

From boosting customer engagement to accelerating health diagnoses to enabling more accurate predictive maintenance, artificial intelligence is transforming every industry. Enterprises know that they need to embrace this change or risk losing out to competitors. But they face challenges with using existing data center infrastructure to power modern AI applications.

IT Challenges for Data Center AI

- The ever-increasing size of AI models, along with the amount of data needed to train them, creates intense resource demands that existing IT compute and networking capabilities cannot meet.
- Modern AI applications are complex, involving many different components that must be orchestrated to work together to obtain useful results. This makes it challenging to manage, monitor, and scale these applications.
- The proliferation of ad-hoc cloud usage or isolated compute silos, especially those based on specialized hardware, results in a lack of standardization that's difficult to manage at scale.



NVIDIA EGX Platform for Data Center AI

The NVIDIA EGX™ platform enables enterprise IT to deliver a complete AI solution on high-performance and cost-effective infrastructure. The platform is based on NVIDIA-Certified Systems-enterprise-class servers that consist of high-performance GPUs and high-speed, secure NVIDIA® Mellanox® networking-built and sold by our partners. The NVIDIA EGX platform allows customers to prepare for the future while driving down costs by standardizing on a single unified architecture for easy management, deployment, operation, and monitoring.

Power your company's AI transformation with an IT-led infrastructure strategy that enables faster ROI, increased productivity, and streamlined manageability.

KEY APPLICATIONS / PLATFORMS

- TensorFlow
- PyTorch
- NVIDIA® TensorRT™
- NVIDIA Triton™ Inference Server
- Industry-specific AI frameworks from NVIDIA

PROOF POINTS

- NVIDIA A100 Tensor Core GPU broke eight AI performance records in the MLPerf Training benchmark and was up to 237X faster than CPU in the MLPerf Inference benchmark.

BENEFITS OF THE EGX PLATFORM

- **Reduced costs:** Record-setting performance and low total cost of ownership (TCO) enable enterprises to achieve their AI results quickly and efficiently.
- **Increased productivity:** Universal acceleration and a rich end-to-end solution stack enable organizations to maximize utilization and insight with a single infrastructure

Key Technologies

GPU

NVIDIA Tensor Core technology has brought dramatic speedups to AI training and inference operations, bringing down training times from weeks to hours and providing massive acceleration to inference.

Networking

NVIDIA Mellanox ConnectX® SmartNICs and NVIDIA BlueField® data processing units (DPUs) provide a host of software-defined hardware engines for accelerating networking and security. These enable the best of both worlds: best-in-class AI training and inference performance, with all the necessary levels of enterprise data privacy, integrity, and reliability.

Multi-instance GPU

Multi-Instance GPU (MIG), available on select GPU models, allows one GPU to be partitioned into multiple independent GPU instances. With MIG, infrastructure managers can standardize their GPU-accelerated infrastructure while ensuring optimal use of their GPU resources for all stages of AI deployment, from development to training to production inference.

vGPU

NVIDIA virtual GPU (vGPU) software products combine the management and security benefits of server and desktop virtualization with the performance benefits of GPU acceleration.

Kubernetes

The NVIDIA GPU Operator and NVIDIA Network Operator standardize and automate the deployment of all the necessary components for provisioning Kubernetes clusters. Using Helm charts, containers, and continuous integration and continuous delivery (CI/CD), organizations can deploy updated AI software effortlessly in minutes.

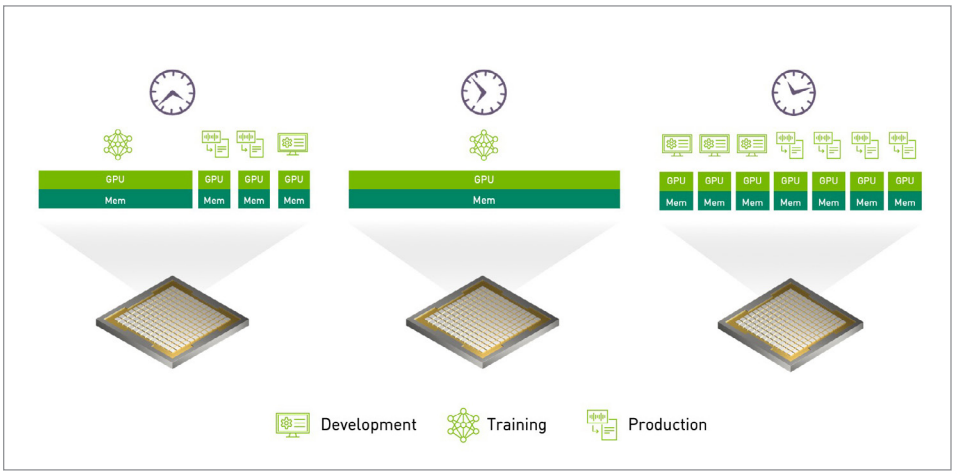
NGC SOFTWARE CATALOG

The NVIDIA NGC™ software catalog is the hub for performance-optimized deep learning and machine learning applications. NGC simplifies building, sharing, and deploying software, so enterprises can gather insights faster and deliver business value sooner.

- **Enterprise management:** IT and DevOps integrations allow AI applications to be operated with existing infrastructure management frameworks.

NVIDIA-CERTIFIED SYSTEMS

- Confidently deploy scalable hardware and software solutions that securely and optimally run accelerated workloads.
- Learn more about accelerated servers at **nvidia.com/certified-systems**



MIG maximizes utilization of GPU resources, allowing a single GPU to be used for all stages of AI deployment.