SOLUTION BRIEF

NVIDIA DGX SYSTEMS

Purpose-Built for the AI Enterprise

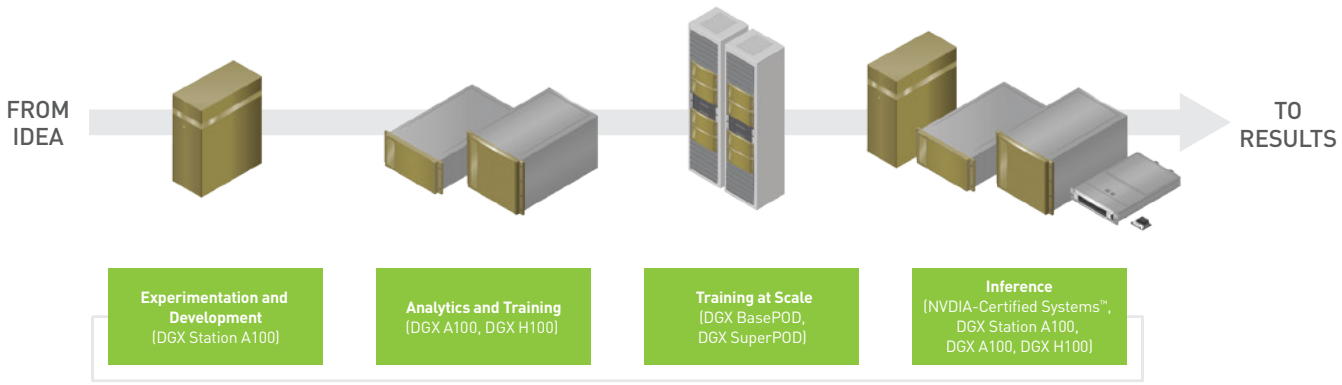
Thousands of Leading Companies Deploy NVIDIA DGX Systems

9	OF THE TOP 10 **GLOBAL UNIVERSITIES**	7	OF THE TOP 10 **US HOSPITALS**	6	OF THE TOP 10 **US BANKS**	7	OF THE TOP 10 **GLOBAL CAR MANUFACTURERS**	8	OF THE TOP 10 **GLOBAL TELCOS**
10	OF THE TOP 10 **US GOVERNMENT INSTITUTIONS**	7	OF THE TOP 10 **CONSUMER INTERNET COMPANIES**	10	OF THE TOP 10 **GLOBAL AEROSPACE AND DEFENSE COMPANIES**				

Companies Strategically Scaling AI Experience Nearly 2X the Success Rate and 3X the Return[1]

Today's enterprise needs an end-to-end strategy for AI innovation to accelerate time to insights and reveal new business frontiers. To stay ahead of the competition, they also need to construct a streamlined AI development workflow that supports fast prototyping, frequent iteration, and continuous feedback, as well as a robust infrastructure that can scale in an enterprise production setting.

NVIDIA DGX™ systems are purpose-built to meet the demands of enterprise AI and data science, delivering the fastest start in AI development, effortless productivity, and revolutionary performance—for insights in hours instead of months.



[1] Accenture. (2019). AI: Built to Scale from Experimental to Exponential. Retrieved from **https://www.accenture.com/_acnmedia/Thought-Leadership-Assets/PDF-2/Accenture-Built-to-Scale-PDF-Report.pdf**

A Purpose-Built Portfolio for End-to-End AI Development

- **NVIDIA DGX Station™ A100** is the world's fastest workstation for data science teams. With four NVIDIA A100 Tensor Core GPUs, fully interconnected with NVIDIA® NVLink® architecture, DGX Station A100 delivers 2.5 petaFLOPS of AI performance, bringing the power of a data center to the convenience of your office.
- **NVIDIA DGX H100** is the world's most complete AI platform—a powerhouse that features eight groundbreaking NVIDIA H100 Tensor Core GPUs. The increased performance, faster networking, and scalability makes it ideal for the largest workloads, including natural language processing (NLP) and deep learning recommendation models.
- **NVIDIA DGX™ A100** is the universal system for all AI workloads. It integrates eight NVIDIA A100 Tensor Core GPUs, delivering a 5 petaFLOPS AI system. Now enterprises can create a complete workflow—from data preparation and analytics to training and inference—using one easy-to-deploy AI infrastructure.
- **NVIDIA DGX BasePOD™** is a reference architecture that incorporates best practices for AI scale, combining compute, networking, storage, power, cooling, and more in an integrated AI infrastructure design built on NVIDIA DGX. DGX BasePOD is available as a turnkey solution, uniting the world's leading providers of data center storage and networking—all backed by single-point-of-contact support.
- **NVIDIA DGX SuperPOD™** delivers a turnkey AI data center solution that delivers a full-service experience with best-of-breed computing, software tools, expertise, and continuous innovation delivered seamlessly. DGX SuperPOD provides high-performance infrastructure with compute foundation built on either DGX A100 or DGX H100.

Powered by NVIDIA Base Command

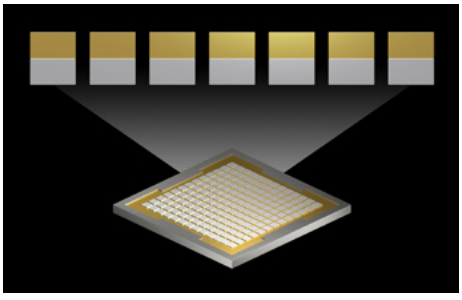
NVIDIA Base Command™ powers every DGX system, enabling organizations to leverage the best of NVIDIA software innovation. Enterprises can unleash the full potential of their DGX investment with a proven platform that includes enterprise-grade orchestration and cluster management, libraries that accelerate compute, storage and network infrastructure, and an operating system optimized for AI workloads. Additionally, DGX systems include NVIDIA AI Enterprise, offering a suite of software optimized to streamline AI development and deployment.

And with ongoing software stack innovation, DGX customers experience continual performance improvement over time, representing a savings of hundreds of thousands of dollars in software engineering OpEx.

Even as you scale to large AI deployments, you can ensure data scientist productivity and optimal utilization of AI infrastructure with **NVIDIA DGX-Ready Software solutions**, which are certified for use on clusters of DGX systems. Enterprises can industrialize AI by taking an MLOps approach, which brings data scientists and DevOps together, using these proven software solutions.

Flexible AI Infrastructure That Adapts to Your Needs

Traditional approaches to AI infrastructure involve slow compute architectures that are siloed by analytics, training, and inference workloads, creating complexity, driving up cost, and constraining speed of scale. NVIDIA DGX A100 unifies all of these AI workloads into a consolidated system with optimized software that is the foundational building block for AI infrastructure. DGX A100 further lowers total cost of ownership (TCO), not only by offering the highest performance, but also by improving infrastructure utilization with the flexibility to handle multiple parallel workloads by multiple users.



Expand the performance and value of NVIDIA H100 and A100 Tensor Core GPUs with the ability to right-size GPUs for multiple workloads

The Proven Choice for Enterprise AI

As enterprise AI initiatives grow, so do their AI infrastructure needs. The DGX H100 system is the fourth generation of the world's first purpose-built AI infrastructure, designed for the evolved AI enterprise that requires the most powerful compute building blocks. The system is created for the singular purpose of maximizing AI throughput, providing enterprises with a highly refined, systemized, and scalable platform to help them achieve breakthroughs in natural language processing, recommender systems, data analytics, and much more.

Towards an AI Center of Excellence

With the growth in AI and its use in day-to-day operations by companies, many are now starting to recognize the importance of developing AI Centers of Excellence (CoE).

For large-scale, multi-node deployments, NVIDIA DGX SuperPOD incorporates the best practices and know-how gained from the world's largest AI deployments. It includes **NVIDIA Base Command Platform** to manage the end-to-end lifecycle of AI training, including workload management, resource sharing, and integrated monitoring and reporting dashboards. For organizations that need to operationalize AI at scale, the NVIDIA DGX SuperPOD Solution for Enterprise takes NVIDIA's industry-leading reference architecture and wraps it in a comprehensive solution and services offering, all backed by NVIDIA.

A Blueprint for Scaling AI
NVIDIA powers its own critical AI research and development with DGX SATURNV, the world's largest proving ground for AI, built on more than 2,000 DGX nodes. Using NVIDIA's applied AI insights and DGX SuperPOD or DGX BasePOD architecture customers can easily build their own world-class computing cluster.

Trusted AI Experts for the Most Challenging Problems

More than a server or workstation, a DGX system is a complete hardware and software platform backed by thousands of AI experts at NVIDIA. Owning a DGX system gives you direct access to **NVIDIA DGXperts**, a global team of AI-fluent practitioners that offer prescriptive guidance and design expertise to help fast-track AI transformation. This ensures mission-critical applications get up and running quickly and stay running smoothly, dramatically improving time to insights.

Ready to Get Started?

To learn more about NVIDIA DGX systems, visit:
www.nvidia.com/dgx

NVIDIA