NVIDIA DGX SYSTEMS

PURPOSE-BUILT FOR THE AI ENTERPRISE

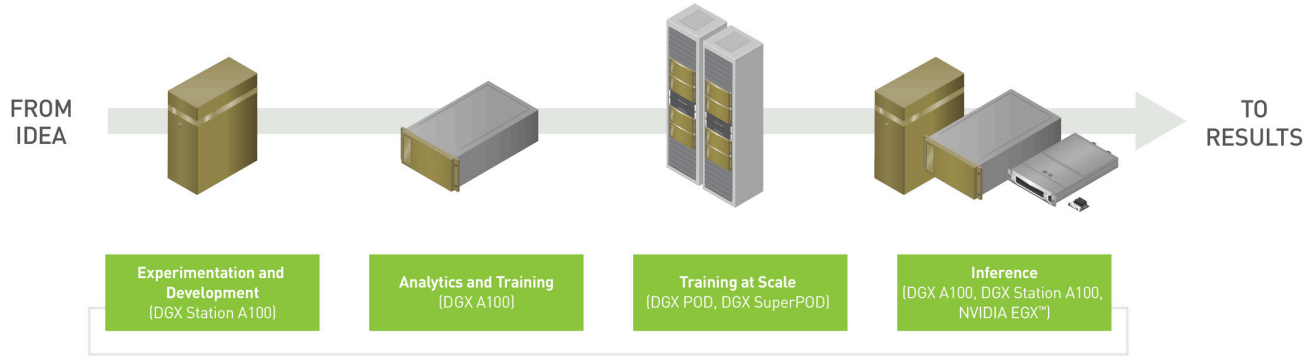
Thousands of Leading Companies Deploy NVIDIA DGX Systems

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 AI of performance, bringing the power of a data center to the convenience of your office, no data center required.
- **NVIDIA DGX™ A100** is the universal system for all AI workloads. It integrates eight of the world's most advanced NVIDIA A100 Tensor Core GPUs, delivering the very first 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 POD™** 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 POD 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.

Powered by NVIDIA DGX Software Stack

End-to-end AI development productivity and performance is enabled by the NVIDIA DGX software stack powering each DGX system. This full-stack suite of pre-optimized AI software includes a DGX-optimized OS, drivers, libraries, and containers and access to NVIDIA NGC™ Catalog for additional assets like pre-trained models, model scripts, and industry solutions for effortless productivity.

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.

> *"NVIDIA-optimized software allowed us to do more. We saw 1.5X faster training on DGX-optimized TensorFlow. Compared with 1,680 images per second on our home-grown 'optimized' TensorFlow software stack, we were seeing 2,600 images per second on the NVIDIA DGX-optimized stack, using ResNet-50. Two years later, with the latest software optimizations from NVIDIA, we saw 4X additional improvement in performance on the same hardware. Impressive work!"*
> – Global Stock Photography Company

Effortless Productivity—From Prototype to Production

Your developers need a fast start with easy access to powerful compute that just works, without being tethered to infrastructure. Start quickly by experimenting and developing on DGX Station, a server-class system for your data science teams that doesn't require a data center. Train models on DGX A100 when you need the fastest time to solution. Train AI at scale leveraging a turnkey solution with DGX POD.

As your AI development journey progresses, each of these solutions enables the effortless mobility of your most important work from one system to the next, without changing any code along the way, so that you can right-size resources for the task at hand.

DGX POD: 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 modular reference architecture, DGX POD, which takes insights from SATURNV and learnings from thousands of customer deployments, customers can easily build their own world-class computing cluster. Turnkey DGX POD offerings from our ecosystem of trusted IT solutions providers are also available to give customers maximum choice and flexibility.

For large-scale, multi-node deployments, NVIDIA DGX SuperPOD™ incorporates the best practices and know-how gained from the world's largest AI deployments. For organizations needing to operationalize AI at scale, 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.

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 from improved infrastructure utilization with the flexibility to handle multiple, parallel workloads by multiple users.

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.

Learn how to accelerate AI on NVIDIA DGX™ systems, powered by NVIDIA A100 Tensor Core GPUs and second-generation AMD EPYC™ CPUs, at > www.nvidia.com/dgx

NVIDIA