NVIDIA DGX SYSTEMS

PURPOSE-BUILT FOR THE AI ENTERPRISE

Thousands of Leading Companies Deploy NVIDIA DGX Systems

8 OF THE TOP 10 **US UNIVERSITIES**	7 OF THE TOP 10 **US HOSPITALS**	6 OF THE TOP 10 **US banks**	7 OF THE TOP 10 **GLOBAL CAR MANUFACTURERS**	7 OF THE TOP 10 **GLOBAL TELCOS**	9 OF THE TOP 10 **US GOVERNMENT INSTITUTIONS**

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.

AI-fluent practitioners, or DGXperts, come with every DGX system. With their extensive track record of field-proven deployments, they offer prescriptive planning, deployment, and optimization expertise to help fast-track your AI transformation.

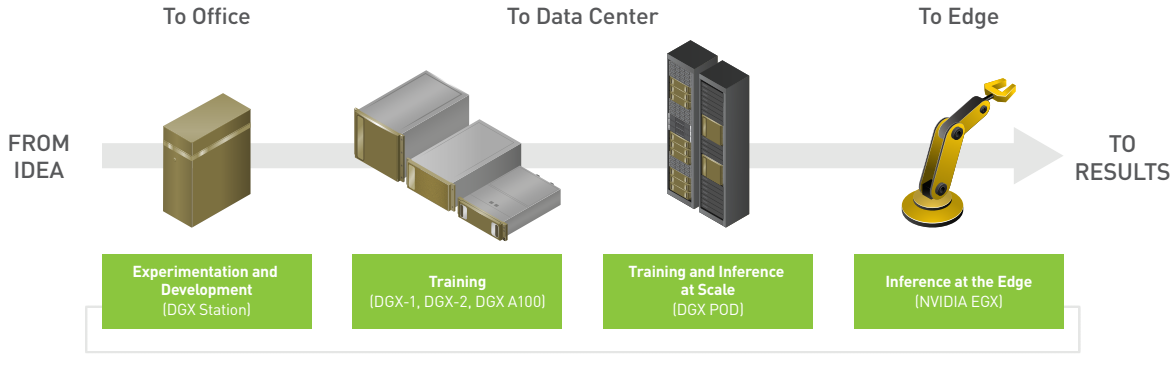

Figure 1: DGX systems—from idea to results.

[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™** is the world's fastest workstation for data science teams. With four NVIDIA V100 Tensor Core GPUs, fully-connected with four-way NVIDIA NVLink™ architecture, DGX Station delivers 500 teraFLOPS of performance, bringing the power of an AI data center to the convenience of your office, no data center required.
- **NVIDIA DGX-1™** is the first AI system purpose-built for enterprise AI in the data center. It integrates eight NVIDIA V100 Tensor Core GPUs, using NVLink technology, delivering 1 petaFLOPs of AI performance.
- **NVIDIA DGX-2™** is the AI system built for complex AI challenges, integrating sixteen NVIDIA V100 Tensor Core GPUs connected by NVLink and NVSwitch, for 2 petaFLOPS of AI performance.
- **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 a single point of contact support.

Combined with innovative GPU-optimized software and simplified management tools, these fully-integrated systems are designed to give data scientists the most powerful tools for AI development—from the office to the data center. Experiment sooner, train the largest models faster, and get insights from data—starting from day one.

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 NGC for additional assets like pre-trained models, model scripts, and industry solutions.

The simplified deployment, revolutionary performance, and enterprise-grade quality of DGX systems insulate enterprises from open-source churn, delivering effortless productivity for data scientists and developers. Unlike off-the-shelf commodity hardware, DGX systems incorporate ongoing software stack innovation, available in containerized versions for DGX customers. This ensures continual performance improvement over time, representing a savings of hundreds of thousands of dollars in software engineering OpEx.

> *"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 ResNet50. 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, 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-1, DGX-2, and 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 enable 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.

> *"Strong infrastructure is fundamental to large-scale computer vision problems. We rely heavily on it to test different architectures and different hyperparameter configurations. Moving from DGX-1 to DGX-2 gave us up to 86% of productivity gain. Having a reference architecture that works gave us the peace of mind we needed to deliver AI products at scale."*
>
> – Ludwig Gamache, Head of IT, Element.AI

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 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

Our **NVIDIA DGXperts** are with you every step of the way. 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 a global team of AI-fluent practitioners that offer prescriptive guidance and design expertise to help fastrack AI transformation with know-how and experience from more than a decade of NVIDIA AI leadership. This ensures mission-critical applications get up and running quickly and stay running smoothly, dramatically improving time to insights.

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

BIOS IT

NVIDIA