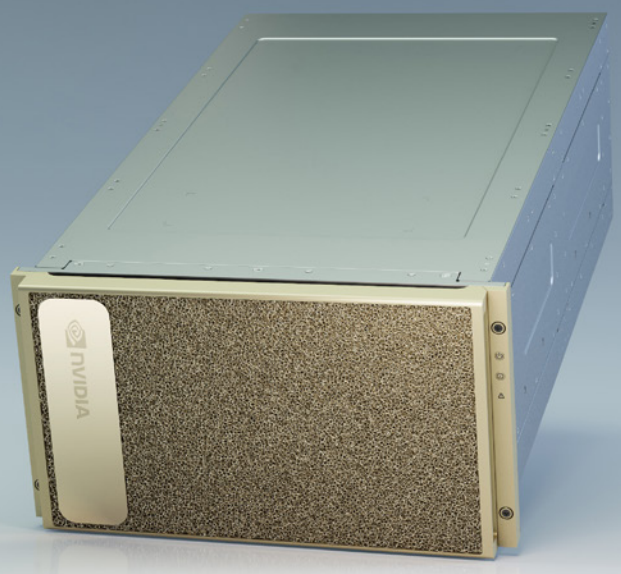
NVIDIA DGX A100

THE UNIVERSAL SYSTEM FOR AI INFRASTRUCTURE



The Challenge of Scaling Enterprise AI

Every business needs to transform using artificial intelligence (AI), not only to survive, but to thrive in challenging times. However, the enterprise requires a platform for AI infrastructure that improves upon traditional approaches, which historically involved slow compute architectures that were siloed by analytics, training, and inference workloads. The old approach created complexity, drove up costs, constrained speed of scale, and was not ready for modern AI. Enterprises, developers, data scientists, and researchers need a new platform that unifies all AI workloads, simplifying infrastructure and accelerating ROI.

The Universal System for Every AI Workload

NVIDIA DGX™ A100 is the universal system for all AI workloads—from analytics to training to inference. DGX A100 sets a new bar for compute density, packing 5 petaFLOPS of AI performance into a 6U form factor, replacing legacy compute infrastructure with a single, unified system. DGX A100 also offers the unprecedented ability to deliver fine-grained allocation of computing power, using the Multi-Instance GPU (MIG) capability in the NVIDIA A100 Tensor Core GPU, which enables administrators to assign resources that are right-sized for specific workloads. Available with up to 640 gigabytes (GB) of total GPU memory, which increases performance in large-scale training jobs up to 3X and doubles the size of MIG instances, DGX A100 can tackle the largest and most complex jobs, along with the simplest and smallest. Running the DGX software stack with optimized software from NGC, the combination of dense compute power and complete workload flexibility make DGX A100 an ideal choice for both single node deployments and large scale Slurm and Kubernetes clusters deployed with NVIDIA DeepOps.

Direct Access to NVIDIA DGXperts

NVIDIA DGX A100 is more than a server. It's a complete hardware and software platform built upon the knowledge gained from the world's largest DGX proving ground—NVIDIA DGX SATURNV—and backed by thousands of DGXperts at NVIDIA. DGXperts are AI-fluent practitioners who offer prescriptive guidance and design expertise to help fastrack AI transformation. They've built a wealth

SYSTEM SPECIFICATIONS

	NVIDIA DGX A100 640GB	NVIDIA DGX A100 320GB
GPUs	8x NVIDIA A100 80 GB GPUs	8x NVIDIA A100 40 GB GPUs
GPU Memory	640 GB total	320 GB total
Performance	5 petaFLOPS AI 10 petaOPS INT8	
NVIDIA NVSwitches	6	
System Power Usage	6.5 kW max	
CPU	Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)	
System Memory	2 TB	1 TB
Networking	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 2x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet
Storage	OS: 2x 1.92 TB M.2 NVME drives Internal Storage: 30 TB (8x 3.84 TB) U.2 NVMe drives	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15 TB (4x 3.84 TB) U.2 NVMe drives
Software	Ubuntu Linux OS Also supports: Red Hat Enterprise Linux CentOS	
System Weight	271.5 lbs (123.16 kgs) max	
Packaged System Weight	359.7 lbs (163.16 kgs) max	
System Dimensions	Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) max Length: 35.3 in (897.1 mm) max	
Operating Temperature Range	5–30 °C (41–86 °F)	

of know-how and experience over the last decade to help maximize the value of your DGX investment. DGXperts help ensure that critical applications get up and running quickly, and stay running smoothly, for dramatically-improved time to insights.

Fastest Time to Solution

NVIDIA DGX A100 features eight NVIDIA A100 Tensor Core GPUs, which deliver unmatched acceleration, and is fully optimized for NVIDIA CUDA-X™ software and the end-to-end NVIDIA data center solution stack. NVIDIA A100 GPUs bring a new precision, Tensor Float 32 (TF32), which works just like FP32 but provides 20X higher floating operations per second (FLOPS) for AI compared to the previous generation. Best of all, no code changes are required to achieve this speedup. And when using NVIDIA's automatic mixed precision with FP16, A100 offers an additional 2X boost to performance with just one additional line of code.

The A100 80GB GPU doubles the high-bandwidth memory from 40 GB (HBM) to 80GB (HBM2e) and increases GPU memory bandwidth 30 percent over the A100 40 GB GPU to be the world's first with over 2 terabytes per second (TB/s). DGX A100 also debuts the third generation of NVIDIA® NVLink®, which doubles the GPU-to-GPU direct bandwidth to 600 gigabytes per second (GB/s), almost 10X higher than PCIe Gen 4, and a new NVIDIA NVSwitch™ that's 2X faster than the last generation. This unprecedented power delivers the fastest time to solution, allowing users to tackle challenges that weren't possible or practical before.

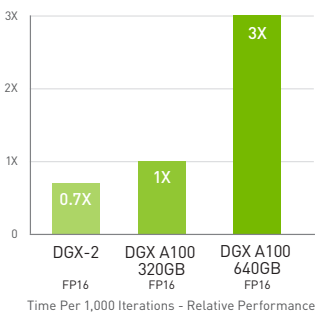
The World's Most Secure AI System for Enterprise

NVIDIA DGX A100 delivers the most robust security posture for your AI enterprise, with a multi-layered approach that secures all major hardware and software components. Stretching across the baseboard management controller (BMC), CPU board, GPU board, self-encrypted drives, and secure boot, DGX A100 has security built in, allowing IT to focus on operationalizing AI rather than spending time on threat assessment and mitigation.

Unmatched Data Center Scalability with NVIDIA Mellanox

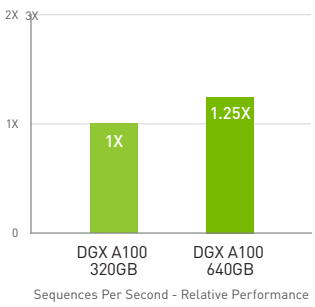
With the fastest I/O architecture of any DGX system, NVIDIA DGX A100 is the foundational building block for large AI clusters like NVIDIA DGX SuperPOD™, the enterprise blueprint for scalable AI infrastructure. DGX A100 features eight single-port NVIDIA Mellanox® ConnectX®-6 VPI HDR InfiniBand adapters for clustering and up to two dual-port ConnectX-6 VPI Ethernet adapters for storage and networking, all capable of 200 Gb/s. The combination of massive GPU-accelerated compute with state-of-the-art networking hardware and software optimizations means DGX A100 can scale to hundreds or thousands of nodes to meet the biggest challenges, such as conversational AI and large-scale image classification.

Up to 3X Higher Throughput for AI Training on Largest Models



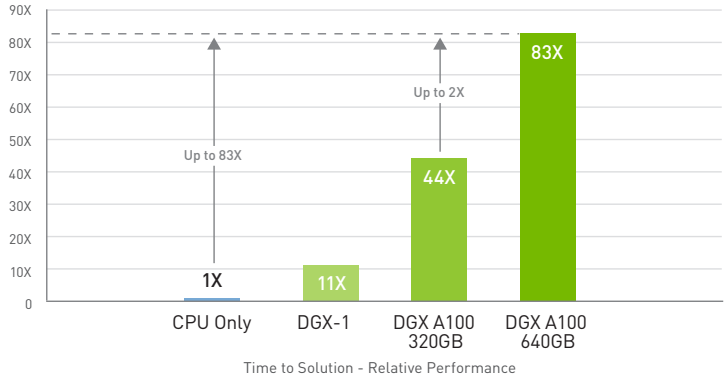
DLRM on HugeCTR framework, precision = FP16 | 1x DGX A100 640GB batch size = 48 | 2x DGX A100 320GB batch size = 32 | 1x DGX-2 (16x V100 32GB) batch size = 32. Speedups Normalized to Number of GPUs.

Up to 1.25X Higher Throughput for AI Inference



MLPerf 0.7 RNN-T measured with (1/7) MIG slices. Framework: TensorRT 7.2, dataset = LibriSpeech, precision = FP16.

Up to 83X Higher Throughput than CPU, 2X Higher Throughput than DGX A100 320GB on Big Data Analytics Benchmark



Big data analytics benchmark | 30 analytical retail queries, ETL, ML, NLP on 10TB dataset | CPU: 19x Intel Xeon Gold 6252 2.10 GHz, Hadoop | 16x DGX-1 (8x V100 32GB each), RAPIDS/Dask | 12x DGX A100 320GB and 6x DGX A100 640GB, RAPIDS/Dask/BlazingSQL. Speedups Normalized to Number of GPUs.

Proven Infrastructure Solutions Built with Trusted Data Center Leaders

In combination with leading storage and networking technology providers, a portfolio of infrastructure solutions are available that incorporate the best of the NVIDIA DGX POD™ reference architecture. Delivered as fully integrated, ready-to-deploy offerings through our NVIDIA Partner Network (NPN), these solutions simplify and accelerate data center AI deployments.

To learn more about NVIDIA DGX A100, visit **www.nvidia.com/DGXA100**

© 2020 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA-X, DGX A100, DGX POD, DGX SuperPOD, Mellanox, NVLink, and NVSwitch are trademarks and/or registered trademarks of NVIDIA Corporation. All company and product names are trademarks or registered trademarks of the respective owners with which they are associated. Features, pricing, availability, and specifications are all subject to change without notice. NOV20

NVIDIA