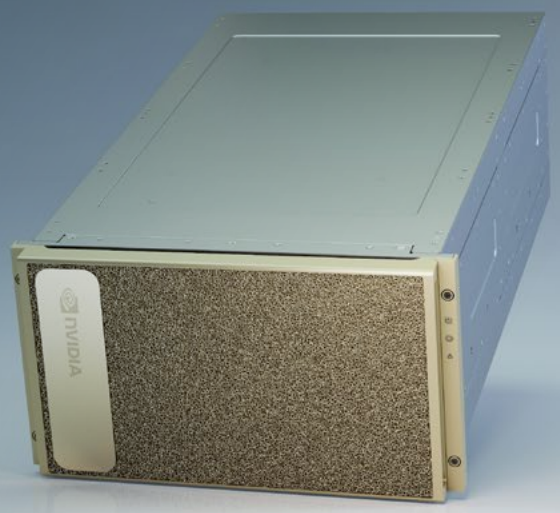
DATASHEET

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 a fine-grained allocation of computing power, using the Multi-Instance GPU (MIG) capability in the NVIDIA A100 Tensor Core GPU. This 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 NVIDIA 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 Bright Cluster Manager.

Unmatched Level of Support and Expertise

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

SYSTEM SPECIFICATIONS

	NVIDIA DGX A100 640GB	
GPUs	**8x NVIDIA A100 80GB Tensor Core GPUs**	
GPU Memory	**640GB 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	**2TB**	
Networking	**Up to 8x Single-Port NVIDIA ConnectX-7** **200 Gb/s InfiniBand** **Up to 2x Dual-Port NVIDIA ConnectX-7 VPI** **10/25/50/100/200 Gb/s Ethernet**	**Up to 8x Single-Port NVIDIA ConnectX-6 VPI** **200 Gb/s InfiniBand** **Up to 2x Dual-Port NVIDIA ConnectX-6 VPI** **10/25/50/100/200 Gb/s Ethernet**
Storage	**OS: 2x 1.92TB M.2 NVME drives** **Internal Storage: 30TB (8x 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	**5ºC to 30ºC (41ºF to 86ºF)**	

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 Tensor Float 32 (TF32) precision, the default precision format for both TensorFlow and PyTorch AI frameworks. This 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.

The A100 80GB GPU increases GPU memory bandwidth 30 percent over the A100 40GB GPU, making it the world's first with 2 terabytes per second (TB/s). It also has significantly more on-chip memory than the previous-generation NVIDIA GPU, including a 40 megabyte (MB) level 2 cache that's nearly 7X larger, maximizing compute performance. 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.

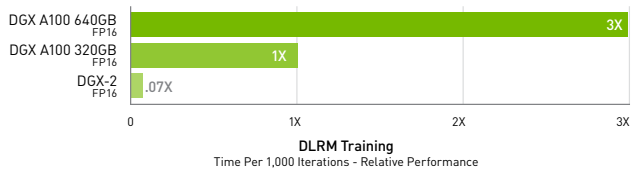
The World's Most Secure AI System for Enterprise

NVIDIA DGX A100 delivers robust security posture for the 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, and self-encrypted drives, DGX A100 has security built in, allowing IT to focus on operationalizing AI rather than spending time on threat assessment and mitigation.

Unparalleled Data Center Scalability with NVIDIA Networking

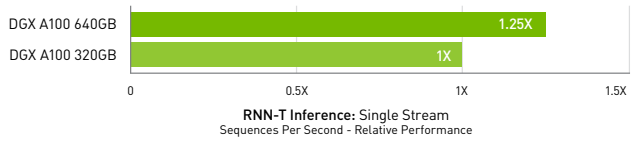
With the fastest I/O architecture of any DGX system, NVIDIA DGX A100 is the foundational building block for large AI clusters like NVIDIA DGX SuperPOD™, the enterprise blueprint for scalable AI infrastructure. DGX A100 features up to eight single-port NVIDIA® ConnectX®-6 or ConnectX-7 adapters for clustering and up to two dual-port ConnectX-6 or ConnectX-7 adapters for storage and networking, all capable of 200Gb/s. With ConnectX-7 connectivity to the NVIDIA Quantum-2 InfiniBand switches, DGX SuperPOD can be built with fewer switches and cables, saving CAPEX and

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



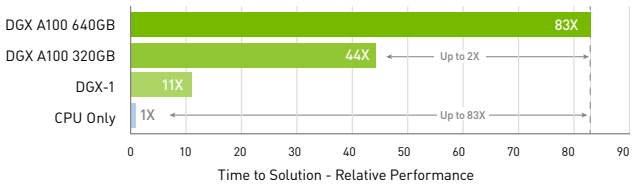
DLRM on HugeCTR framework, precision = FP16 | 1x DGX A100 640GB batch size = 48 | 2x DGX A100 320GB batch size = 32 | 1x DGX-2 (16x V100 32GB) batch size = 32. Speedups normalized to number of GPUs.

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.

OPEX on the data center infrastructure. 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.

Proven Infrastructure Solutions Built with Trusted Data Center Leaders

In combination with leading storage and networking technology providers, a portfolio of infrastructure solutions is available that incorporates 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.

Ready to Get Started?

To learn more about NVIDIA DGX A100, visit **www.nvidia.com/dgxa100**