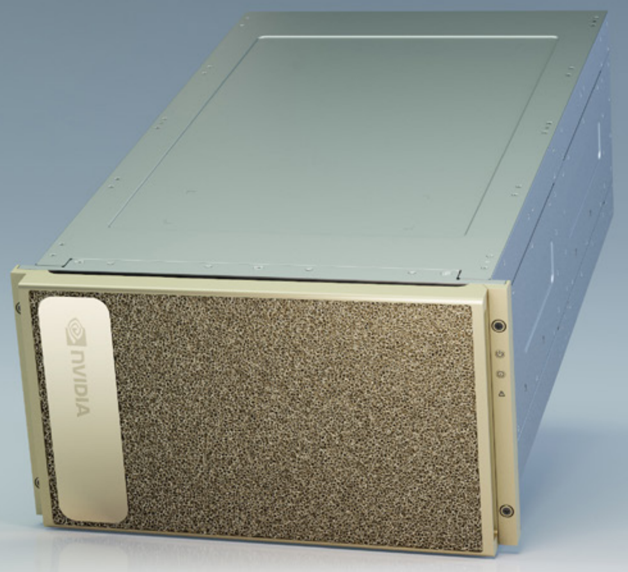


NVIDIA DGX A100
THE UNIVERSAL SYSTEM FOR AI INFRASTRUCTURE

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 capability in the NVIDIA A100 Tensor Core GPU, which enables administrators to assign resources that are right-sized for specific workloads. This ensures that the largest and most complex jobs are supported, 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 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.

SYSTEM SPECIFICATIONS

GPUs	**8x NVIDIA A100 Tensor Core GPUs**
GPU Memory	**320 GB total**
Performance	**5 petaFLOPS AI** **10 petaOPS INT8**
NVIDIA NVSwitches	**6**
System Power Usage	**6.5kW max**
CPU	**Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)**
System Memory	**1TB**
Networking	**8x Single-Port Mellanox ConnectX-6 VPI** **200Gb/s HDR InfiniBand** **1x Dual-Port Mellanox ConnectX-6 VPI** **10/25/50/100/200Gb/s Ethernet**
Storage	**OS: 2x 1.92TB M.2 NVME drives** **Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives**
Software	**Ubuntu Linux OS**
System Weight	**271 lbs (123 kgs)**
Packaged System Weight	**315 lbs (143kgs)**
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°C to 30°C (41°F to 86°F)**

Fastest Time to Solution

NVIDIA DGX A100 features eight NVIDIA A100 Tensor Core GPUs, providing users with 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, TF32, which works just like FP32 while providing 20X higher FLOPS for AI vs. the previous generation, and best of all, no code changes are required to get this speedup. And when using NVIDIA's automatic mixed precision, A100 offers an additional 2X boost to performance with just one additional line of code using FP16 precision. The A100 GPU also has a class-leading 1.6 terabytes per second (TB/s) of memory bandwidth, a greater than 70% increase over the last generation. Additionally, the A100 GPU has significantly more on-chip memory, including a 40MB Level 2 cache that is nearly 7X larger than the previous generation, maximizing compute performance. DGX A100 also debuts the next 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 is 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.

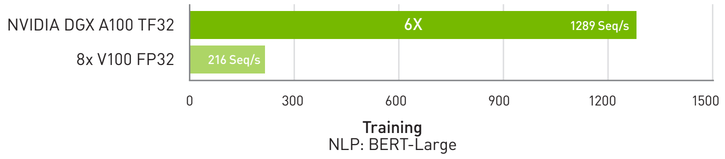
The World's Most Secure AI System for Enterprise

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 Mellanox

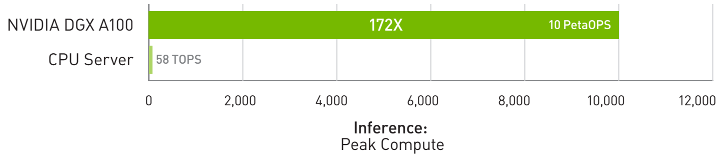
With the fastest I/O architecture of any DGX system, NVIDIA DGX A100 is the foundational building block for large AI clusters like NVIDIA DGX SuperPOD™, the enterprise blueprint for scalable AI infrastructure. DGX A100 features eight single-port Mellanox ConnectX-6 VPI HDR InfiniBand adapters for clustering and 1 dual-port ConnectX-6 VPI Ethernet adapter for storage and networking, all capable of 200Gb/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.

DGX A100 Delivers 6 Times The Training Performance



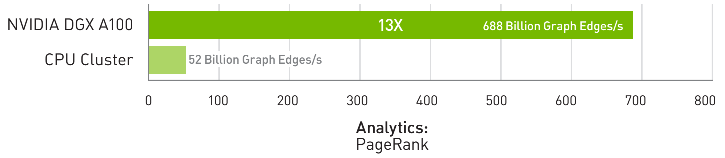
BERT Pre-Training Throughput using PyTorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 | V100: DGX-1 with 8x V100 using FP32 precision | DGX A100: DGX A100 with 8x A100 using TF32 precision

DGX A100 Delivers 172 Times The Inference Performance



CPU Server: 2x Intel Platinum 8280 using INT8 | DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity

DGX A100 Delivers 13 Times The Data Analytics Performance



3000x CPU Servers vs. 4x DGX A100 | Published Common Crawl Data Set: 128B Edges, 2.6TB Graph

Proven Infrastructure Solutions Built with Trusted Data Center Leaders

In combination with leading storage and networking technology providers, we offer a portfolio of infrastructure solutions that incorporate the best of the NVIDIA DGX POD™ reference architecture. Delivered as fully integrated, ready-to-deploy offerings through our NVIDIA Partner Network, these solutions make data center AI deployments simpler and faster for IT.

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