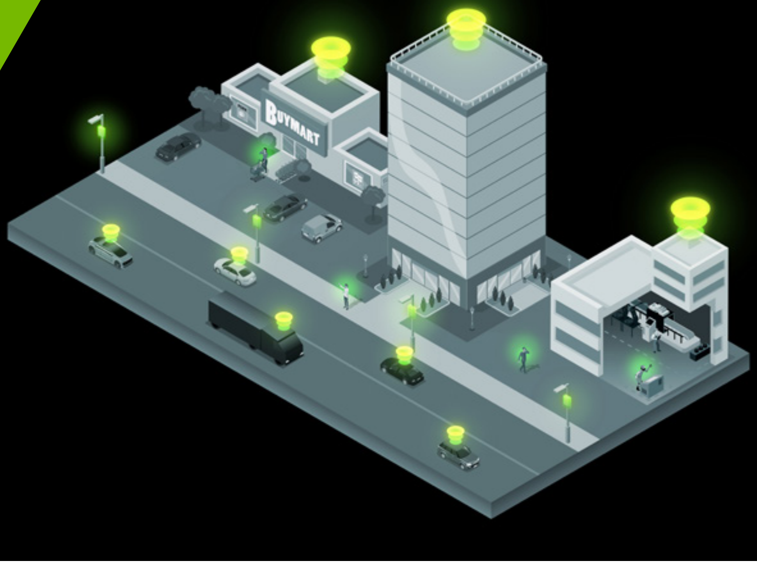


NVIDIA DGX A100

THE UNIVERSAL SYSTEM FOR AI IN TELECOMMUNICATIONS

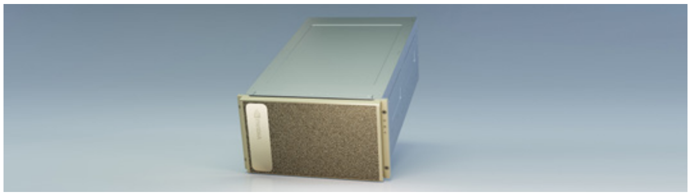
Leading telecommunications companies are using AI, machine learning, and deep learning to optimize network performance and data security, enhance customer service, and enable new edge revenue opportunities. NVIDIA DGX™ A100 offers unmatched performance and versatility, powering AI applications and analytics workloads that can accelerate ROI, minimize risk, and drive innovation.

Telco Industry Use Cases

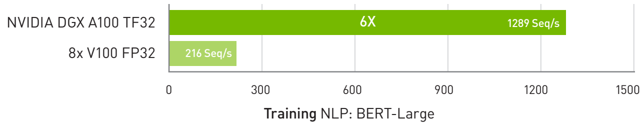
- Anomaly detection for network performance improvement, security threat detection, and predictive performance
- Conversational AI for automated customer service, search and retrieval of technical data, automated speech recognition, and AI assistants
- Research into new AI-enabled edge compute use cases

Telco R&D teams are experimenting with larger datasets and deep learning. Leveraging them for network anomaly detection effectively increases detection rates and reduces false positives. They're also essential for conversational AI model development. With the launch of the NVIDIA Jarvis SDK, telcos can accelerate the creation of multimodal conversational AI applications, leading to increased customer satisfaction and loyalty. At the heart of conversational AI are deep learning models that require significant computing power to train chatbots that communicate in telco-specific language. DGX A100 delivers the fastest time to solution on the most complex models for superhuman levels of language understanding.

Research teams can leverage the power of the DGX A100 to create new revenue opportunities. Whether developing new edge compute applications or offering DGX as a service to customers, the DGX A100 offers the most complete platform for all training and inference needs.

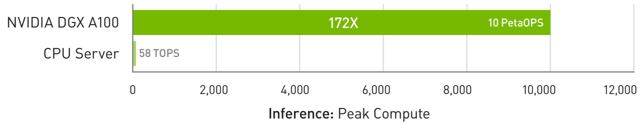


DGX A100 Delivers 6 Times The Training Performance



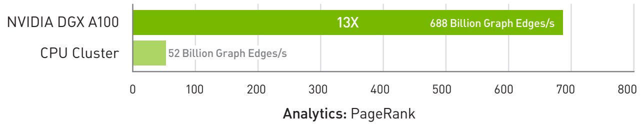
BERT Pre-Training Throughput using PyTorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 | V100: DGX-1 with 8x V100 using FP32 precision | DGX A100: DGX A100 with 8x A100 using TF32 precision

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

The Universal System for Every Data Science Workload

NVIDIA DGX A100 is the universal system for all **data science workloads**—from analytics to machine learning training to AI 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. This ensures that the largest and most complex jobs are supported, along with the simplest and smallest—on the same hardware. Running the DGX software stack with optimized, containerized 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 clusters.

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, 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 the 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 GPU also has a class-leading 1.6 terabytes per second (TB/s) of memory bandwidth, a greater than 70 percent increase over the last generation. It also has significantly more on-chip memory, including a 40 megabyte (MB) level 2 cache that's nearly 7X larger than the previous generation, 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. Additionally, DGX A100 includes the second generation of 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 Telecommunications

NVIDIA DGX A100 delivers the most robust security posture for R&D and production environments 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 one dual- port ConnectX-6 VPI Ethernet adapter for storage and networking, all capable of 200 gigabits per second (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.

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	**6,500 W 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** **200 Gb/s HDR InfiniBand** **1x Dual-Port Mellanox ConnectX-6 VPI** **10/25/50/100/200 Gb/s Ethernet**
Storage	**Operating system: 2x 1.92 TB M.2 NVME drives** **Internal storage: 15 TB (4x 3.84 TB) 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	**5º–30ºC (41º–86ºF)**

Proven Infrastructure Solutions Built with Trusted Data Center Leaders

In combination with leading storage and networking technology providers, 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 the NVIDIA Partner Network, these solutions simplify and accelerate data center AI deployments.

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 fast-track AI transformation. They've built a wealth of knowledge and experience over the last decade to help maximize the value of DGX investments. DGXperts help ensure that critical applications get up and running quickly, and stay running smoothly, for dramatically improved time to insights.

To learn more about NVIDIA DGX A100, visit **www.nvidia.com/dgx-a100**

To learn about NVIDIA technology's top use cases for telecommunications, visit **www.nvidia.com/telco**