NVIDIA DGX Station A100 System Architecture

The Workgroup Appliance for the Age of AI

Technical White Paper

WP-10454-001_v01 | April 2021

Table of Contents

1 Introduction

Data science teams are looking for ways to improve their workflow and the quality of their models by speeding up iterations-less time spent on each experiment means more experimentation can be done-and by using larger, higher-quality datasets. These experts are at the leading edge of innovation, developing projects that will have a profound impact for their organizations. But they're often left searching for spare AI compute cycles, whether it's with their own individual GPU-enabled laptops and workstations, available GPU cloud instances, or a borrowed portion of a data center AI server.

These teams need a dedicated AI resource that isn't at the mercy of other areas within their organizations: a purpose-built AI system without compromise that can handle all of the jobs that busy data scientists can throw at it, an accelerated AI platform that's fully optimized across hardware and software for maximum performance. These teams need a workgroup server for the age of AI. For these teams, there is NVIDIA DGX Station™ A100.

NVIDIA DGX Station A100 brings AI supercomputing to data science teams, offering data center technology without a data center or additional IT investment. Designed for multiple, simultaneous users, DGX Station A100 leverages server-grade components in an easy-to-place workstation form factor. It's the only system with four fully-interconnected and Multi-Instance GPU (MIG)-capable NVIDIA A100 Tensor Core GPUs with up to 320 gigabytes (GB) of total GPU memory that can plug into a standard power outlet in the office or at home, resulting in a powerful AI appliance that you can place anywhere.

In this white paper, we'll take a look at the design and architecture of DGX Station A100.

2 NVIDIA DGX Station A100 System Architecture

While the outside chassis of DGX Station A100 remains the same as that of the previous DGX Station, the inside has been completely redesigned.

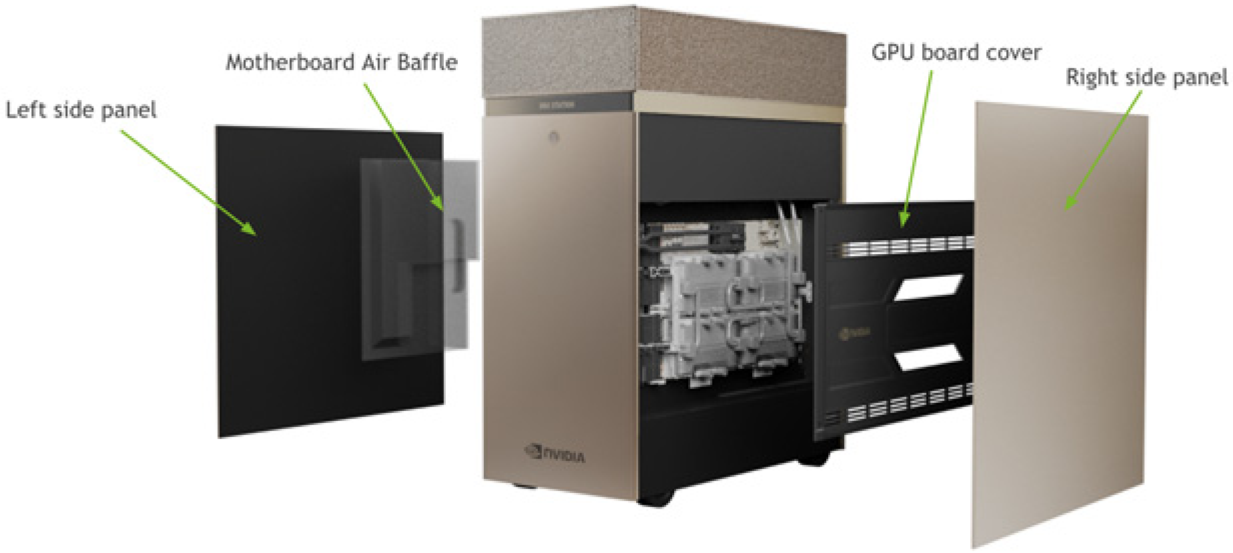


Figure 1. View of the DGX Station A100 chassis and side panels.

2.1 NVIDIA A100 GPU - 8th Generation Data Center GPU for the Age of Elastic Computing

At the core, the NVIDIA DGX Station A100 system leverages the NVIDIA A100 GPU (Figure 2), designed to efficiently accelerate large complex AI workloads as well as several small workloads, including enhancements and new features for increased performance over the NVIDIA V100 GPU. The A100 GPU incorporates 40 gigabytes (GB) of high-bandwidth HBM2 memory, larger and faster caches, and is designed to reduce AI and HPC software and programming complexity.

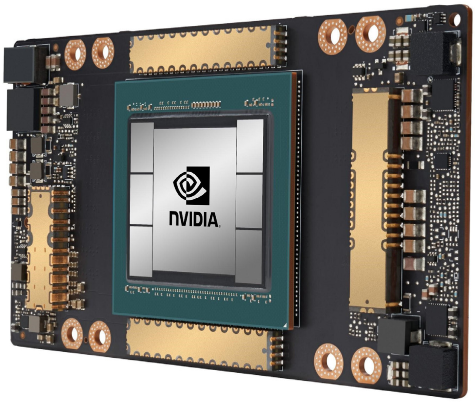


Figure 2. NVIDIA A100 Tensor Core GPU

The NVIDIA A100 GPU includes the following new features to further accelerate AI workload and HPC application performance.

- Third-generation Tensor Cores
- Fine-grained Structured Sparsity
- Multi-Instance GPU

Third-Generation Tensor Cores

The NVIDIA A100 GPU includes new third-generation Tensor Cores. Tensor Cores are specialized high-performance compute cores that perform mixed-precision matrix multiply and accumulate calculations in a single operation, providing accelerated performance for AI workloads and HPC applications.

The first-generation Tensor Cores used in the NVIDIA DGX-1 with NVIDIA V100 provided accelerated performance with mixed-precision matrix multiply in FP16 and FP32. This latest generation in the DGX A100 uses larger matrix sizes, improving efficiency and providing twice the performance of the NVIDIA V100 Tensor Cores along with improved performance for INT4 and binary data types. The A100 Tensor Core GPU also adds the following new data types:

- TF32

- IEEE Compliant FP64
- Bfloat16 (BF16)

  BF16/FP32 mixed-precision Tensor Core operations perform at the same speed as FP16/FP32 mixed-precision Tensor Core operations, providing another choice for deep learning training)

TensorFloat-32 (TF32) Uses Tensor Cores by Default

AI training typically uses FP32 math, without Tensor Core acceleration. The NVIDIA A100 architecture introduces the new TensorFloat-32 (TF32) math operation that uses Tensor Cores by default. The new TF32 operations run 10X faster than the FP32 FMA operations available with the previous generation data center GPU.

The new TensorFloat-32 (TF32) operation performs calculations using an 8-bit exponent (same range as FP32), 10-bit mantissa (same precision as FP16) and 1 sign-bit [Figure 3). In this way, TF32 combines the range of FP32 with the precision of FP16. After performing the calculations, a standard FP32 output is generated.

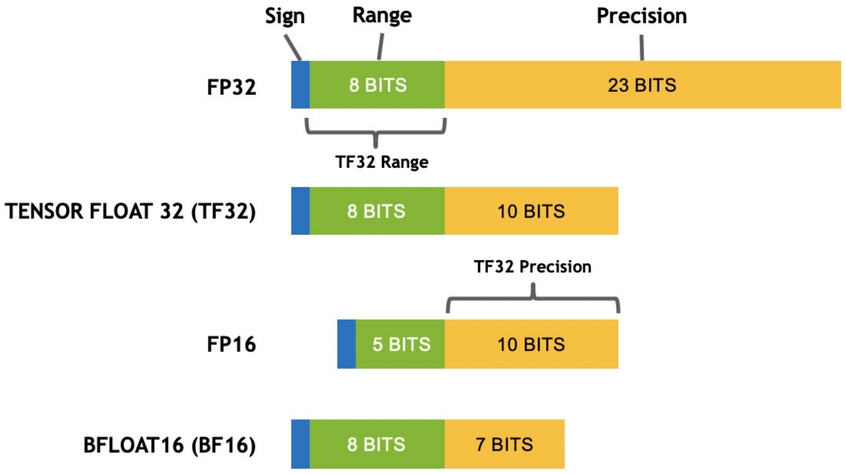


Figure 3. Explanation of Tensor Float 32, FP32, FP16, and BF16

Non-Tensor operations can use the FP32 data path, allowing the NVIDIA A100 to provide TF32-accelerated math along with FP32 data movement.

TF32 is the default mode for TensorFlow, PyTorch and MXNet, starting with NGC Deep Learning Container 20.06 Release. For TensorFlow 1.15, the source code and pip wheels have also been released. These deep learning frameworks require no code change. Compared to FP32 on V100, TF32 on A100 provides over 6X speedup for training the BERT-Large model, one of the most demanding conversational AI models.

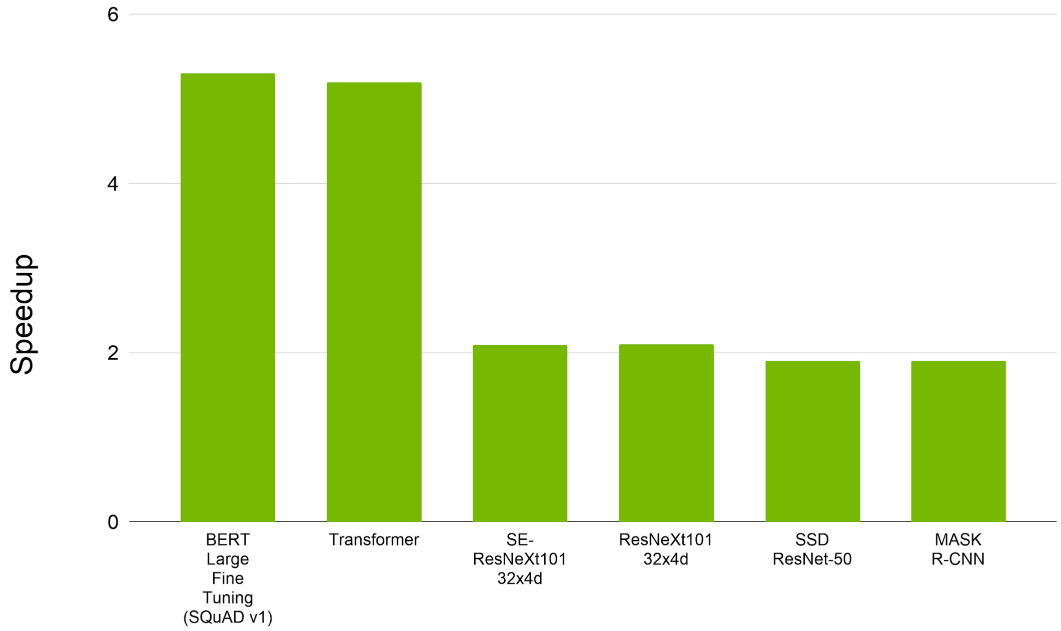


Figure 4. **TF32 can provide over 5X speedup compared to FP32**, *PyTorch 1.6 in NGC pytorch:20.06-py3 container, training on BERT-Large model. Results on DGX A100 (8x A100 GPUs). All model scripts can be found in the* Deep Learning Examples repository.

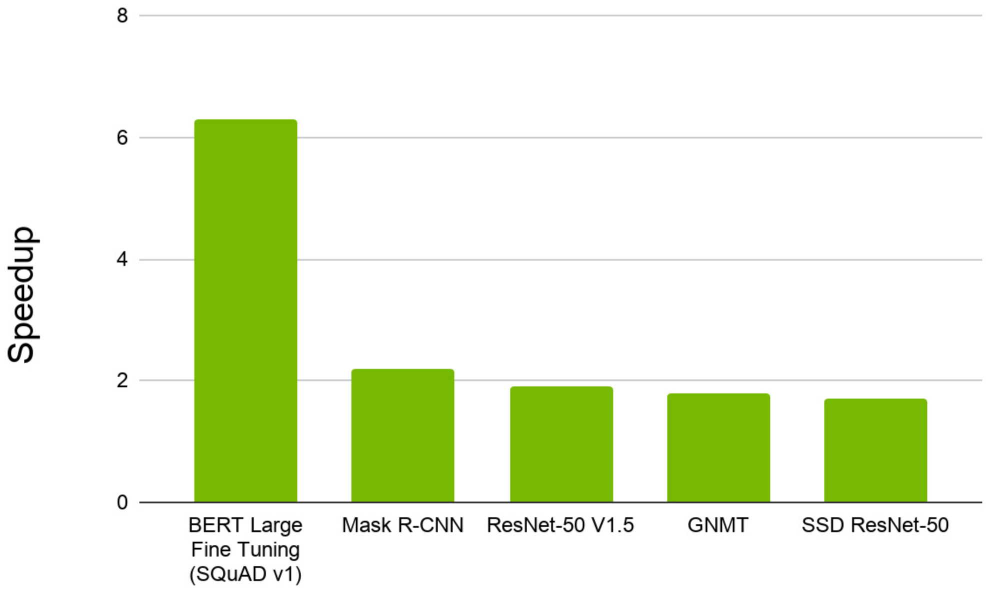


Figure 5. **TF32 can provide over 6X speedup compared to FP32**, *TensorFlow 1.15 in NGC tensorflow:20.06-tf1-py3 container, training on BERT-Large model. Results on DGX A100 (8x A100 GPUs). All model scripts can be found in the* Deep Learning Examples repository

2.2 Fine-grained Structured Sparsity

The NVIDIA A100 GPU supports fine-grained structured sparsity to accelerate simplified neural networks without harming accuracy. Sparsity often comes from pruning - the technique of removing weights that contribute little to the accuracy of the network. Typically, this involves "zeroing out" and removing weights that have zero or near-zero values. In this way, pruning can convert a dense network into a sparse network that delivers the same level of accuracy with reduced compute, memory, and energy requirements. Until now, though, this type of fine-grained sparsity did not deliver on its promises of reduced model size and faster performance.

With fine-grained structured sparsity and the 2:4 pattern supported by A100 (Figure 6), each node in a sparse network performs the same amount of memory accesses and computations, which results in a balanced workload distribution and even utilization of compute nodes. Additionally, structured sparse matrices can be efficiently compressed, and their structure leads to doubled throughput of matrix multiply-accumulate operations with hardware support in the form of Sparse Tensor Cores.

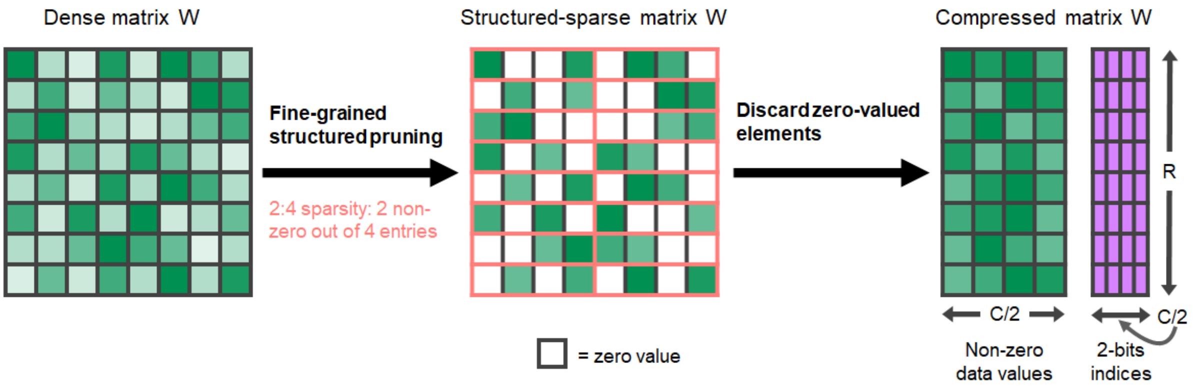

Dense matrix W
Fine-grained structured pruning
2:4 sparsity: 2 non-zero out of 4 entries
Structured-sparse matrix W
= zero value
Discard zero-valued elements
Compressed matrix W
R
C/2
C/2
Non-zero data values
2-bits indices

Figure 6. NVIDIA A100 GPU supports fine-grained structured sparsity with an efficient compressed format and 2X instruction throughput.

The result is accelerated Tensor Core computation across a variety of AI networks and increased inference performance. With fine-grained structured sparsity, INT8 Tensor Core operations on A100 offer 20X more performance than on V100, and FP16 Tensor Core operations are 5X faster than on V100.

2.3 Multi-Instance GPU (MIG)

The NVIDIA A100 GPU incorporates a new partitioning capability called Multi-Instance GPU (MIG) for increased GPU utilization. MIG uses spatial partitioning to carve the physical resources of a single A100 GPU into as many as seven independent GPU instances. With MIG, the NVIDIA A100 GPU can deliver guaranteed quality of service at up to 7 times higher throughput than V100 with simultaneous instances per GPU (Figure 7).

On an NVIDIA A100 GPU with MIG enabled, parallel compute workloads can access isolated GPU memory and physical GPU resources as each GPU instance has its own memory, cache, and streaming multiprocessor. This allows multiple users to share the same GPU and run all instances simultaneously, maximizing GPU efficiency.

MIG can be enabled selectively on any number of GPUs in the DGX Station A100 system - not all GPUs need to be MIG-enabled. However, if all GPUs in a DGX Station A100 system are MIG enabled, up to 28 users can simultaneously and independently take advantage of GPU acceleration.

In fact, DGX Station A100 is the only system in a workstation form-factor that supports MIG. Typical uses cases that can benefit from a MIG-enabled DGX Station A100 are

- Evaluating multiple inference jobs with batch sizes of one that involve small, low-latency models and that don't require all the performance of a full GPU before deploying it into production on a DGX A100 server.
- Jupyter notebooks for model exploration.
- Resource sharing of the GPU among multiple users, such as students or members of data science teams at large organizations.

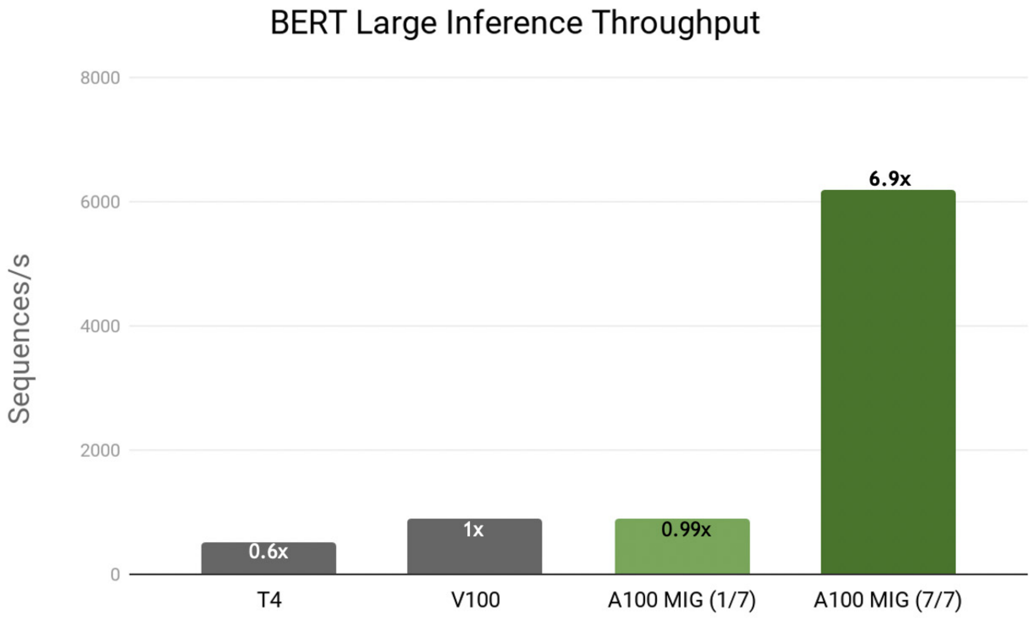


Figure 7. Up to 7X Higher Inference throughput with Multi-Instance GPU (MIG). Above results were performed on a DGX A100. BERT Large Inference (with Sequence Length=128)

- *T4: TRT 7.1, Precision = INT8, Batch Size =256,*
- *V100: TRT 7.1, Precision = FP16, Batch Size =256*
- *A100 with 7 MIG instances of 1g.5gb. TensorRT Release Candidate, Batch Size =94, Precision = INT8 with Sparsity (1g.5gb is the smallest instance of the A100 which specifies 1/7 of the compute and 5 GB of total memory)*

Taking it further on DGX A100 with 8 A100 GPUs, users can configure different GPUs for vastly different workloads, as shown in the following example (Figure 8):

- 4 GPUs for AI training
- 2 GPUs for HPC or data analytics
- 2 GPUs in MIG mode, partitioned into 14 MIG instances, each one running inference

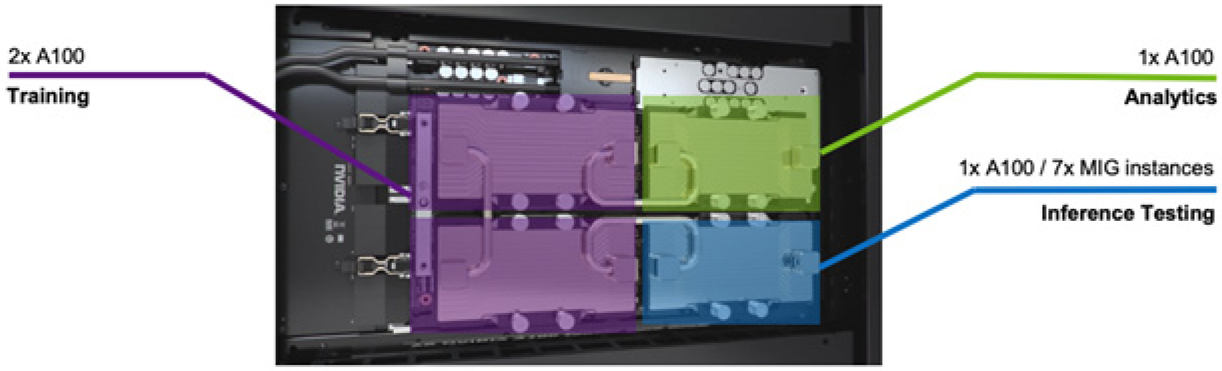


Figure 8. Different workloads on different GPUs in NVIDIA DGX Station A100 system.

MIG supports a variety of deployment options, allowing users to run CUDA applications on bare-metal or containers. MIG support is available using the NVIDIA Container Toolkit (previously known as nvidia-docker2) for Docker, allowing users to run CUDA accelerated containers on GPU instances. Refer to Running CUDA Applications as Containers for more information.

2.4 Four NVIDIA A100 Tensor Core GPUs in DGX Station A100

One of the most unique features of DGX Station A100 is the incorporation of the 4-way NVIDIA HGX™ GPU boards. Designed for high performance accelerated data center servers, the NVIDIA HGX™ platform brings together the full power of NVIDIA GPUs and a fully optimized NVIDIA AI and HPC software stack from NGC™ to provide highest application performance. With its end-to-end performance and flexibility, NVIDIA HGX enables researchers and scientists to combine simulation, data analytics, and AI to advance scientific progress.

With four A100 Tensor Core GPUs, with a total of 160GB of HBM2 or 320GB HBM2e memory, and an aggregate bandwidth of 2.4TB/s, data scientists can get access to data center technology and performance to wherever they choose to put their DGX Station A100.

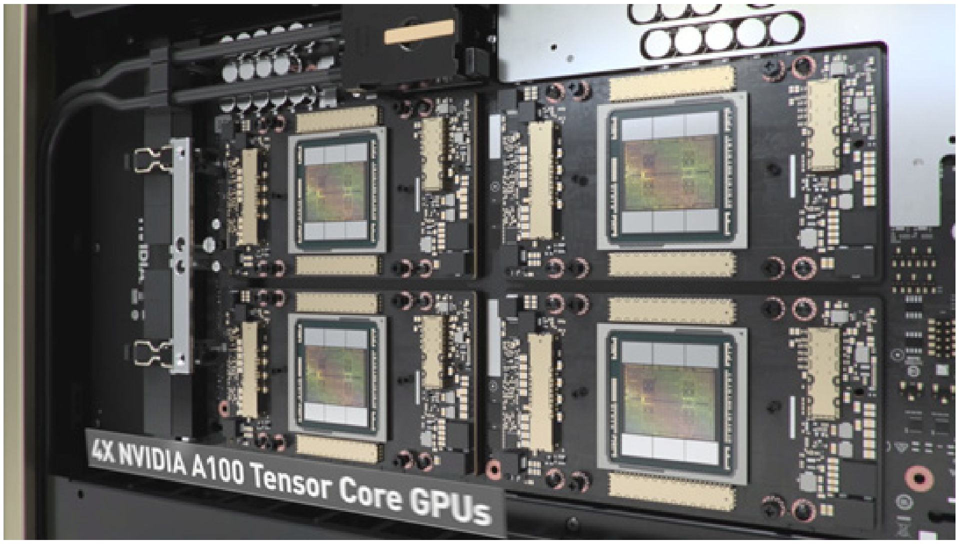


Figure 9. View of the 4-way NVIDIA HGX board inside NVIDIA DGX Station A100.

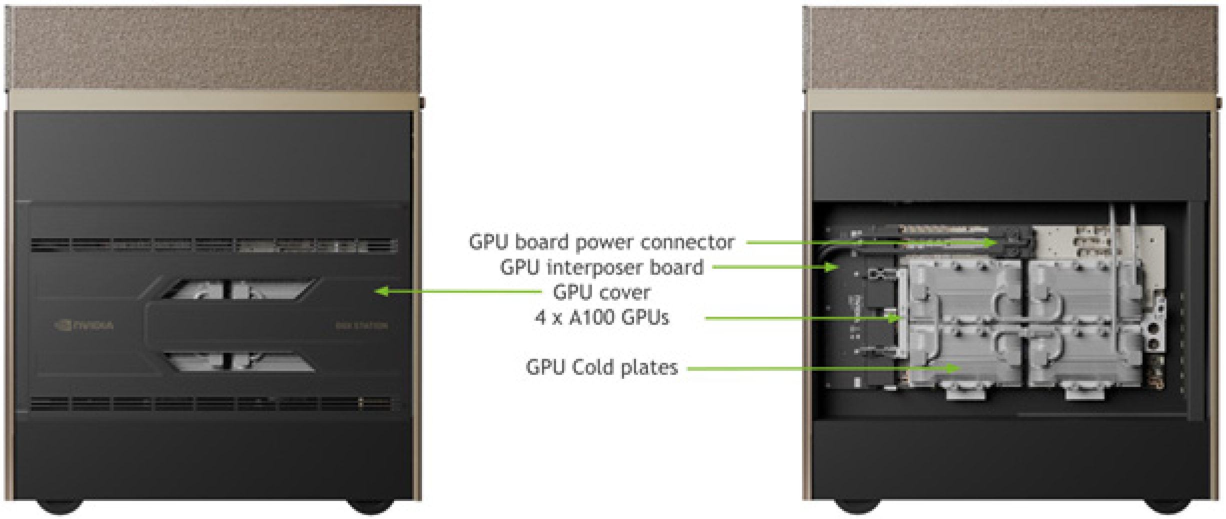


Figure 10. View of NVIDIA DGX Station A100, GPU side, with GPU cold plates.

Third-Generation NVLink to Accelerate Large Complex Workloads

NVIDIA® NVLink® is a high-speed, direct GPU-to-GPU interconnect.

The four A100 GPUs on the HGX GPU baseboard are directly connected with third-generation NVLink, enabling full connectivity. Any A100 GPU can access any other A100 GPU's memory using high-speed NVLink ports. The A100-to-A100 peer bandwidth is 300 GB/s bi-directional, which is more than 3X faster than the fastest PCIe Gen4 x16 bus.

See Figure 11 for a topology diagram.

Accelerated Performance With All PCIe Gen4

The NVIDIA A100 GPUs are connected to the PCI switch infrastructure over x16 PCI Express Gen 4 (PCIe Gen4) buses that provide 31.5 Gb/s each for a total of 252 Gb/s, doubling the bandwidth of PCIe 3.0/3.1. These are the links that provide access to the DGX Dis-play Adaptor, the NVMe storage, and the CPU.

Training workloads commonly involve reading the same datasets many times to improve accuracy. Rather than use up all the network bandwidth to transfer this data over and over, high performance local storage is implemented with NVMe drives to cache this data. This increases the speed at which the data is read into memory, and it also reduces network and storage system congestion.

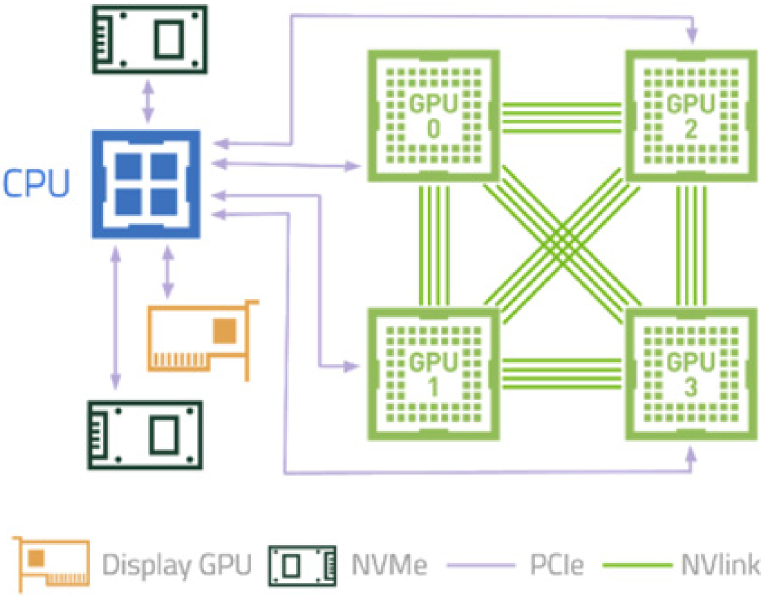


Figure 11. View of the DGX Station A100 devices topology

2.5 A Better Way To Cool Down

Similar to previous generation DGX Station, the DGX Station A100 is designed to be operated quietly in an office environment within a nominal operating temperature ranging from 5oC - 30oC. However, unlike the previous generation it now features a new and improved refrigerant based cooling system which can not only handle higher GPU/CPU component temperatures, it can do this completely maintenance free. This means no more water level checking and refilling. No chance of system damage should a leak develop in the cooling system and it's completely environmentally safe and non-toxic with no user serviceable parts to worry about.

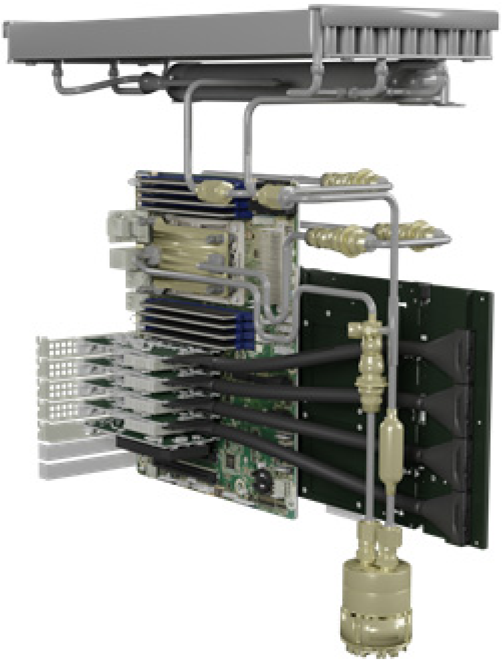


Figure 12. Figure 10. View of NVIDIA DGX Station A100 cooling system.

The refrigerant system consists of a single circulation pump, cold plates which are mounted to the GPUs and system CPU, plumbing to interconnect the various system components, and heat exchanger unit which includes a refrigerant reservoir to evacuate the heat. Three low speed fans provide the airflow to the condenser to whisper quietly (<37dBm) and displace the heat collected into the surrounding ambient air.

2.6 A Server-Class CPU For A Server-Class AI Appliance

DGX Station A100 features the latest AMD Epyc 7742 enterprise-class server processor based on the Zen 2 micro architecture. Using the latest TSMC 7nm manufacturing process, the AMD Epyc 7742 processor offers the highest performance for HPC and AI workloads as has been demonstrated by numerous world records and benchmarks. The DGX Station A100 system includes one of these CPUs for boot, storage management, and deep learning framework scheduling and coordination. The CPU runs at a maximum speed of 3.4GHz of boost, has 64 cores with 2 threads per core.

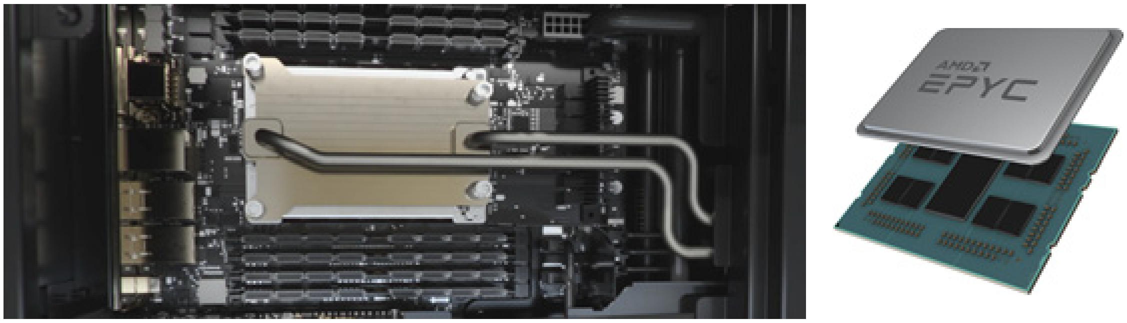


Figure 13. View of the AMD Epyc 7742 processor, covered by its cold plate.

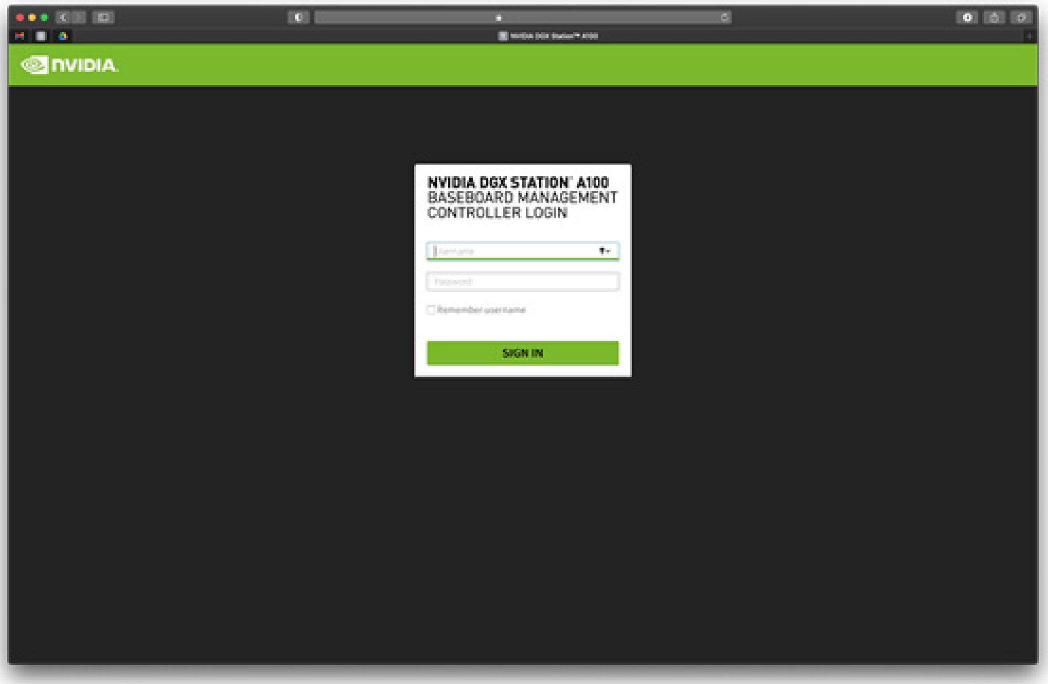
The CPU provides extensive memory capacity and bandwidth, and features 8 memory channels for an aggregate of 204.8 GB/s of memory bandwidth. Memory capacity on the DGX Station A100 is 512GB standard with 8 DIMM slots populated with 64GB DDR4-3200 ECC RDIMM memory.

For I/O, the AMD Epyc 7742 processor offers 128 PCIe Gen4 links. This provides the system with maximum bandwidth from the processor which supports high speed connectivity to the GPUs and other IO devices. Each DGX Station A100 system comes with 1.92 TB NVMe M.2 boot OS SSDs, and one 7.68 TB PCIe gen4 NVMe U.2 cache SSDs.

2.7 Added Flexibility With Remote Management (BMC)

Remote management provides the flexibility to share the compute resources in a DGX Station A100 across multiple researchers or teams, while allowing IT services to manage the system. Remote management also provides the option to install the DGX Station A100 in a central location with other IT infrastructure as well under a desk in a cubicle.

The remote management capabilities are implemented through a full-featured Baseboard Management Controller (BMC) embedded in the motherboard of the DGX Station A100. It provides a web-based user interface to monitor and manage the system remotely, as well as IPMI and RedFish interfaces that allow existing infrastructure tools to manage and monitor the system.



NVIDIA DGX STATION™ A100
BASEBOARD MANAGEMENT
CONTROLLER LOGIN

Figure 14. Screenshot of the DGX Station A100 BMC login page.

The web-based user interface provides a secure way to read all the sensors and review the system logs through an easy to use menu and multiple screens that provide details on all the components. The interface provides temperature monitoring of GPUs, memory DIMMs, CPU, display card, and motherboard. Fan speeds, power consumption, and system voltages are also monitored and displayed with historic graphs and current readings.

All of these features are also available through the IPMI interfaces, so monitoring software that collects logs, statistics and sensor readings can get the information automatically without user intervention. The IPMI interface also features a Serial Over LAN (SOL) interface to access the system's serial console to manage the system BIOS settings, or the installed operating system.

In addition, the web-based interface also provides the remote Keyboard, Video, Mouse (KVM) capability which allows the user to see exactly what is being displayed on the monitor as well as manage the system from a distance. The KVM functionality also features virtual storage capabilities which enables mounting remote volume, and enables the remote DGX Station A100 to be re-installed or booted from an ISO image.

A useful feature available on DGX Station A100 is to connect the remote management network interface and the regular system LAN cable through a single network connection. This allows a single network drop in a cubicle to be used for both network functions. This is easily configurable through the BMC and it leverages a technology known as Network Controller Sideband Interface (NCSI).

The BMC also provides control of an LED that can identify the system remotely by illuminating a button available in the rear panel of the system. This LED can also be controlled by the button it is mounted on. This is a great tool to coordinate remote teams managing the system with local teams that need to maintain the system.

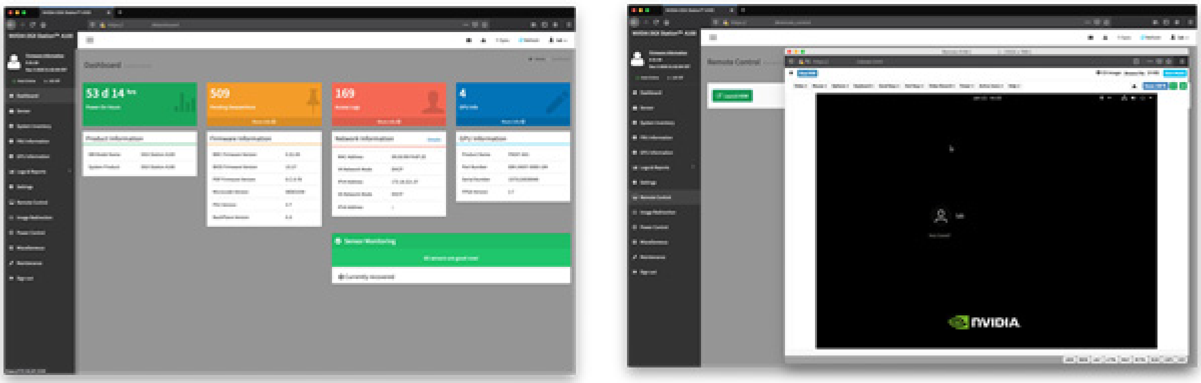


Figure 15. Example screenshots of the DGX Station A100 BMC web UI.

2.8 Graphics When You Need Them

The DGX Station A100 now features its own dedicated workstation grade graphics adapter, the NVIDIA DGX Display Adapter powered by NVIDIA Turing^{TM} graphics processing unit (GPU) technology and 4GB of GDDR6 memory. The flexible single slot and low profile form factor card is powerful enough to meet the most demanding workflows. It utilizes the ultimate 3D performance in a compact and power-efficient form factor.

Key performance features and capabilities include:

- 4 GB GDDR6 graphics memory allows data scientists and researchers to render results at unprecedented scale, complexity, and richness.
- 896 streaming multiprocessor (SMX) cores deliver incredible performance with a peak power consumption of 50W in a cool and quiet single slot form factor.
- Supports four simultaneous displays and up to 5K resolution with VESA DisplayPort 1.4 Mini-DP outputs.

In addition to the DGX Display Adapter the BMC also has its own PCIe-based 2D/VGA display adapter capability. This adapter will support a standard (DB-15) analog VGA display up to 1920x1200@60Hz resolution. The BMC 2D/VGA display can be enabled/disabled in the BIOS; by default it is enabled. Alternatively, the BMC 2D/VGA display adapter can be set as the default X-Windows display by changing the "Display Adapter" settings in the BIOS to "On-Board" mode.

2.9 Other System Features

The system provides plenty of IO connectivity with two on-board 10 Gigabit Ethernet interfaces for fast data transfers. These interfaces can be used to mount network shared storage to load data for HPC or AI applications with low latency and high bandwidth. As mentioned earlier, one of the ports also has NCSI functionality, so it offers the ability to consolidate network cables, reducing the number of ports needed next to the system.

The system is easy to service. In case of a failure, this system can be repaired directly by the customer because components are clearly labeled and they can be replaced with few tools, or none. Among the items that can be serviced include DIMMs, NVMe solid state drives, Display Card, the Trusted Platform Module, and the battery.

The system weighs 105 lbs, so wheels were incorporated to move it from its packaging to the floor, and to easily maneuver it around the office or laboratory. The wheels come with a set of locks that can be installed to prevent the unit from moving. These locks can also be very handy when the unit is prominently displayed atop a desk or a pedestal to prevent it from rolling.

For security, a Kensington lock was added in case the DGX Station A100 is installed in an area where it could be moved inadvertently.

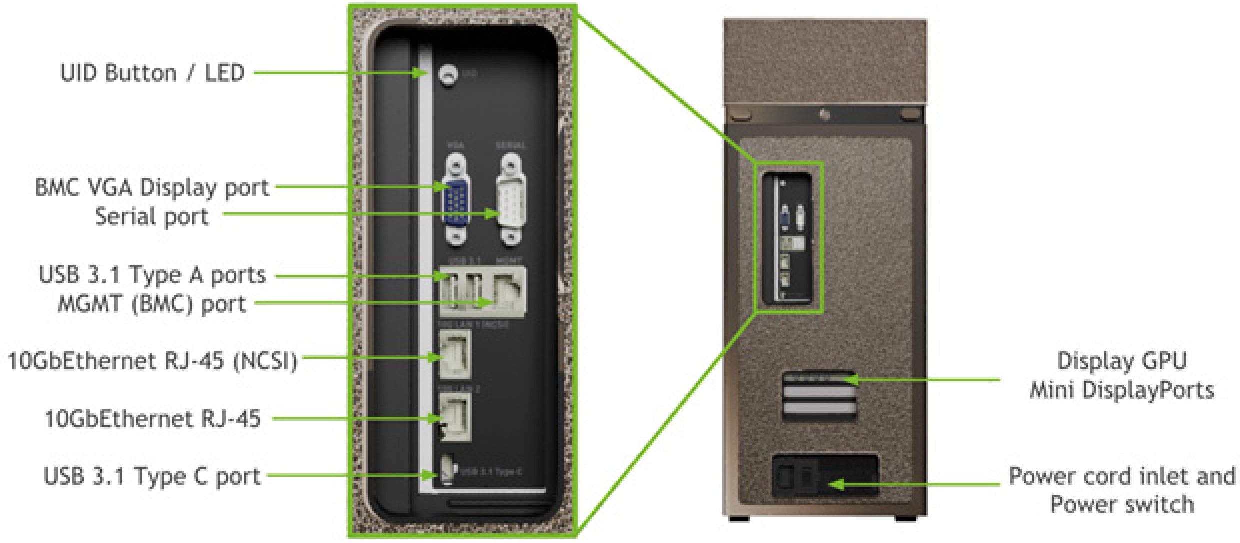


Figure 16. View of the DGX Station A100 back panel, including all ports.

3 Security

The NVIDIA DGX Station A100 system supports a number of advanced security features.

3.1 Drive Encryption

The NVIDIA DGX™ OS software supports filesystem-level encryption of the system partition and full drive encryption of the data drives using optional self-encrypting drives (SEDs). SEDs encrypt data on the drives automatically on the fly without performance impact. The drives can be automatically unlocked from a key stored in the TPM module or by using a centralized key server.

Encrypting the system partition is implemented in software. It requires unlocking of the filesystem on every boot, either manually by entering a passphrase or automatically by using a centralized key server.

3.2 System Memory Encryption

The NVIDIA DGX Station A100's 64-core AMD Eypc CPU provides customers additional protection against unwanted attack with its Secure Memory Encryption (SME) and Secure Encrypted Virtualization (SEV) capabilities which are enabled in the system BIOS by default. The CPU provides hardware accelerated memory encryption for data-in-use protection and takes advantage of new security components available in the processor:

- AES-128 encryption engine - Embedded in the memory controller, it automatically encrypts and decrypts data in main memory when an appropriate key is provided.
- AMD Secure Processor - Provides cryptographic functionality for secure key generation and key management.

More information on AMD's SME and SEV capabilities can be found at https://developer.amd.com/sev/.

3.3 Trusted Platform Module (TPM) Technology

The NVIDIA DGX Station A100 system includes a secure cryptoprocessor which conforms to the Trusted Platform Module (TPM 2.0)[1] industry standard. The cryptoprocessor is the foundation of the security subsystem in the DGX A100, securing hardware via cryptographic operations.

When enabled, The TPM ensures the integrity of the boot process until the DGX OS has fully booted and applications are running.

The TPM is also used with the self-encrypting drives and the drive encryption tools for secure storage of the vault and SED authentication keys.

[1] See the Trusted Platform Module white paper from the Trusted Computing group https://trustedcomputinggroup.org/resource/trusted-platform-module-tpm-summary/

4 Fully Optimized DGX Software Stack

The DGX Station A100 software has been built to run AI workloads at scale. A key goal is to enable practitioners to deploy deep learning frameworks, data analytics, and HPC applications on the DGX Station A100 with minimal setup effort. The design of DGX OS, the platform software, is centered around an optimized OS based on Ubuntu with all required drivers, libraries, and tools preinstalled on the system. Customers can be productive immediately without having to identify and install matching software drivers and libraries.

DGX OS is already pre-installed on all DGX systems for providing a turn-key experience. The software is also available from repositories hosted by Canonical and NVIDIA for installing additional software and upgrading existing software with the latest security patches and bug fixes. The repositories include additional NVIDIA driver and CUDA Toolkit versions, which allow customers to painlessly upgrade the software stack when new features are required.

DGX OS also provides the necessary tools and libraries for using containerized applications and SDK software and includes support for GPUs. Using containerized software further expedites access to well-tested deep learning frameworks and other applications. Practitioners can retrieve GPU-optimized containers for deep learning (DL), machine learning (ML), and high-performance computing (HPC) applications, along with pretrained models, model scripts, Helm charts, and software development kits (SDKs) from the NGC Registry. All containers have been developed, tested, and tuned on DGX systems, and is compatible with all DGX products: DGX-1, DGX-2, DGX A100, DGX Station and DGX Station A100.

DGX customers also have access to private registries, which provide a secure storage location for custom containers, models, model scripts, and Helm charts that can be shared with others within the enterprise or organization. Learn more about the NGC Private Registry in this blog post.

Figure 17 shows how all these pieces fit together as part of the DGX software stack.

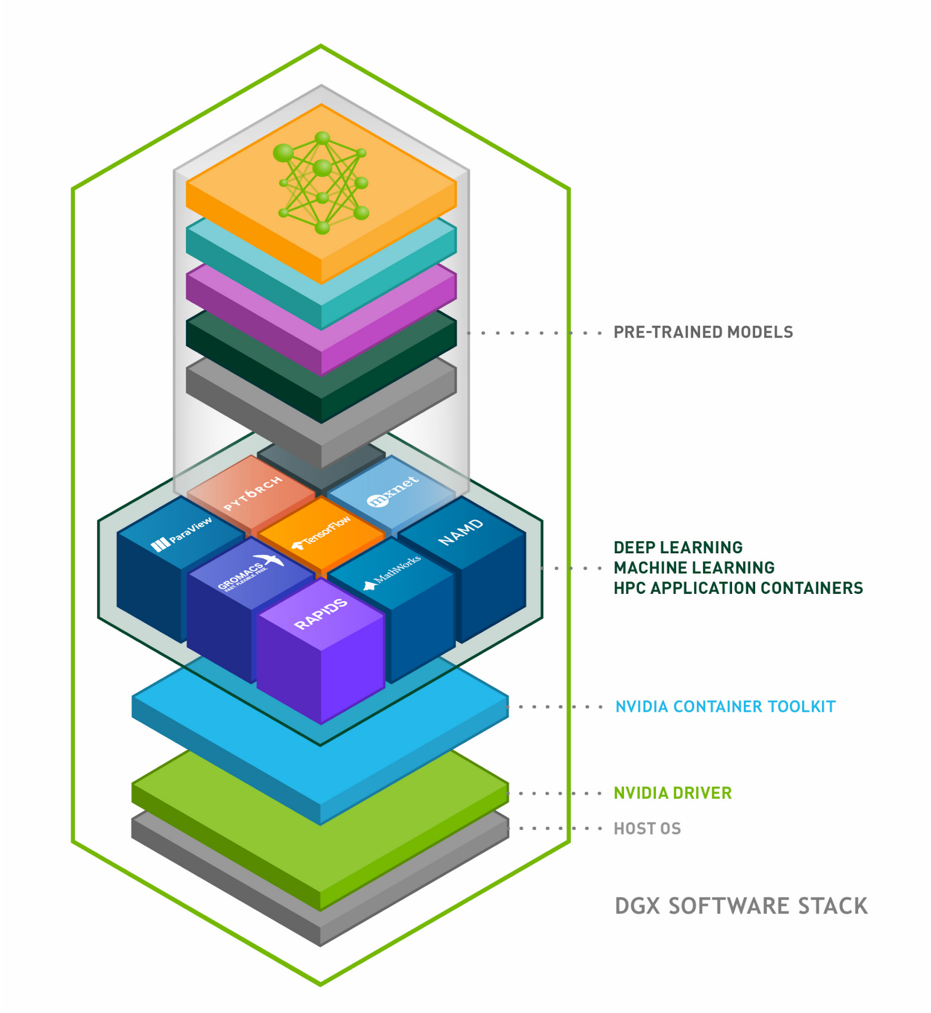


Figure 17. NVIDIA DGX Software Stack

The DGX software stack includes the following major components.

- **The NVIDIA Container Toolkit** allows users to build and run GPU accelerated Docker containers. The toolkit includes a container runtime library and utilities to automatically configure containers to leverage NVIDIA GPUs.
- **GPU-accelerated containers** feature software to support
  - Deep learning frameworks for training, such as PyTorch, MXNet, and TensorFlow
  - Inference platforms, such as TensorRT
  - Data analytics, such as RAPIDS, the suite of software libraries for executing end-to-end data science and analytics pipelines entirely on GPUs.
  - High-Performance Computing (HPC), such as CUDA-X HPC, OpenACC, and CUDA®.
- **The NVIDIA CUDA Toolkit**, incorporated within each GPU-accelerated container, is the development environment for creating high performance GPU-accelerated applications.

CUDA 11 enables software developers and DevOps engineers to reap the benefits of the major innovations in the new NVIDIA A100 GPU, including the following:

- Support for new input data type formats, Tensor Cores and performance optimizations in CUDA libraries for linear algebra, FFTs, and matrix multiplication
- Configuration and management of MIG instances on Linux operating systems, part of the DGX software stack
- Programming and APIs for task graphs, asynchronous data movement, fine-grained synchronization, and L2 cache residency control

Read more about what's new in the CUDA 11 Features Revealed Devblog.

5 Game Changing Performance

5.1 Deep Learning Training and Inference

Packed with innovative features, data center-grade technology, and a balanced system design, the DGX Station A100 delivers unprecedented performance in deep learning training and inference in such a form factor.

The combination of the groundbreaking A100 GPUs with massive computing power and high-bandwidth access to large DRAM, and fast local storage, makes the NVIDIA DGX Station A100 system optimal for dramatically accelerating complex networks like BERT.

A single DGX Station A100 system features 5 petaFLOPs of AI computing capability to process complex models. The large model size of BERT requires a huge amount of memory, and each DGX Station A100 provides 160 GB or 320 GB of high bandwidth GPU memory. The NVIDIA interconnect technology NVLink brings all GPUs together to work as one on large AI models with high-bandwidth communication for efficient scaling.

For example, compared to DGX Station (with four NVIDIA V100 GPUs), the speed up for BERT Large, Pre-Training Phase 1, is significant:

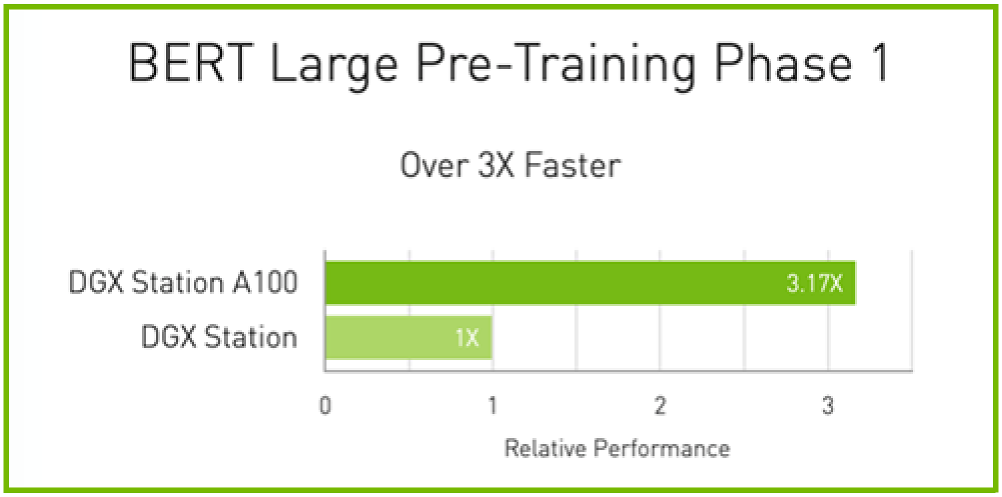


Figure 18. BERT Large Pre-Training, Speedup of DGX Station A100 relative to DGX Station. Batch Size=64; Mixed Precision; With AMP; Real Data; Sequence Length=128

While DGX Station A100 is not typically used for deep learning inference use cases, it does include four A100 Tensor Core GPUs on a NVIDIA HGX baseboard, and hence supports Multi-Instance GPU (MIG) technology. As described earlier, slicing GPUs into smaller instances to perform inference workloads can make sense in certain circumstances, and performance of these improved tremendously in this generation:

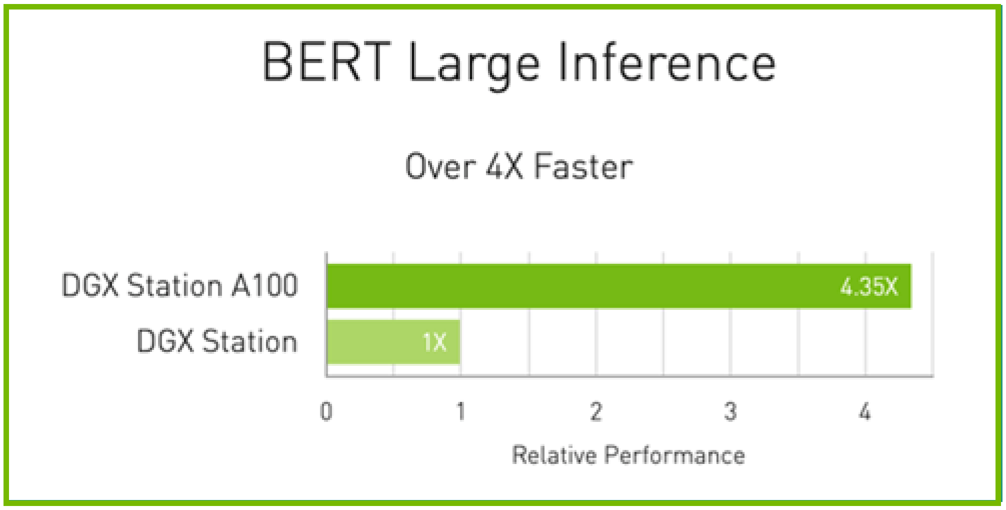


Figure 19. BERT Large Inference, Speedup of DGX Station A100 relative to DGX Station. Batch Size=256; INT8 Precision; Synthetic Data; Sequence Length=128, cuDNN 8.0.4

As machine learning and deep learning workloads get bigger and bigger, multi-GPU training becomes more important. Almost-linear scalability across 1, 2, and 4 GPUs can be achieved with DGX Station A100. Here are some examples of typical image related deep learning training workloads:

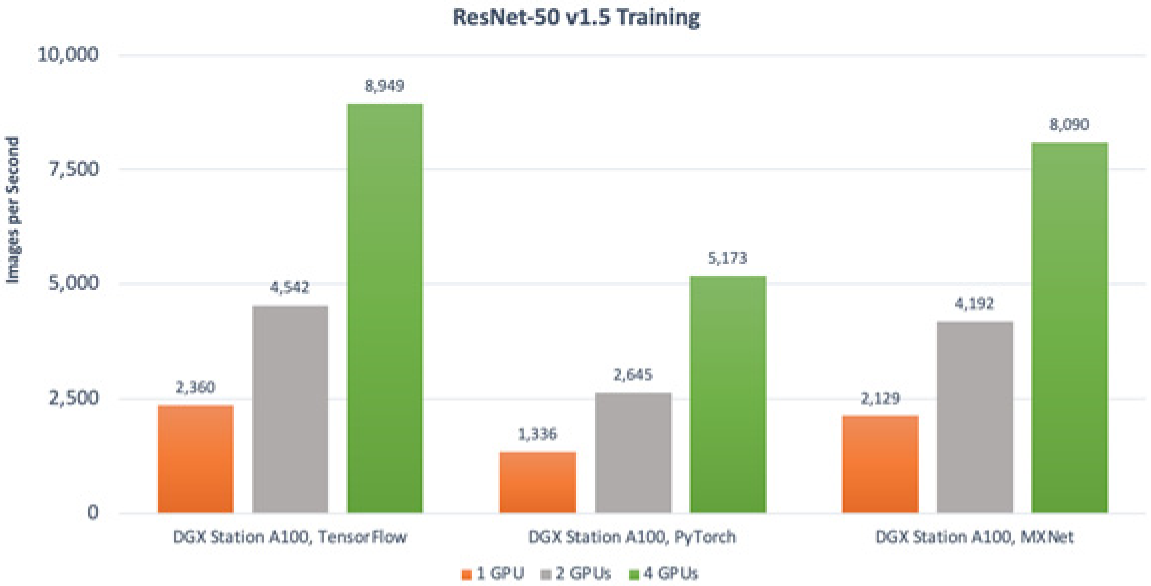


Figure 20. DGX Station A100 (4x A100 SXM 80GB), ResNet-50 V1.5 training, Mixed Precision, TensorFlow (21.02-tf1_py3, bs 256), PyTorch (21.02_py3, bs 256), MXNet (21.02_py3, bs 192)

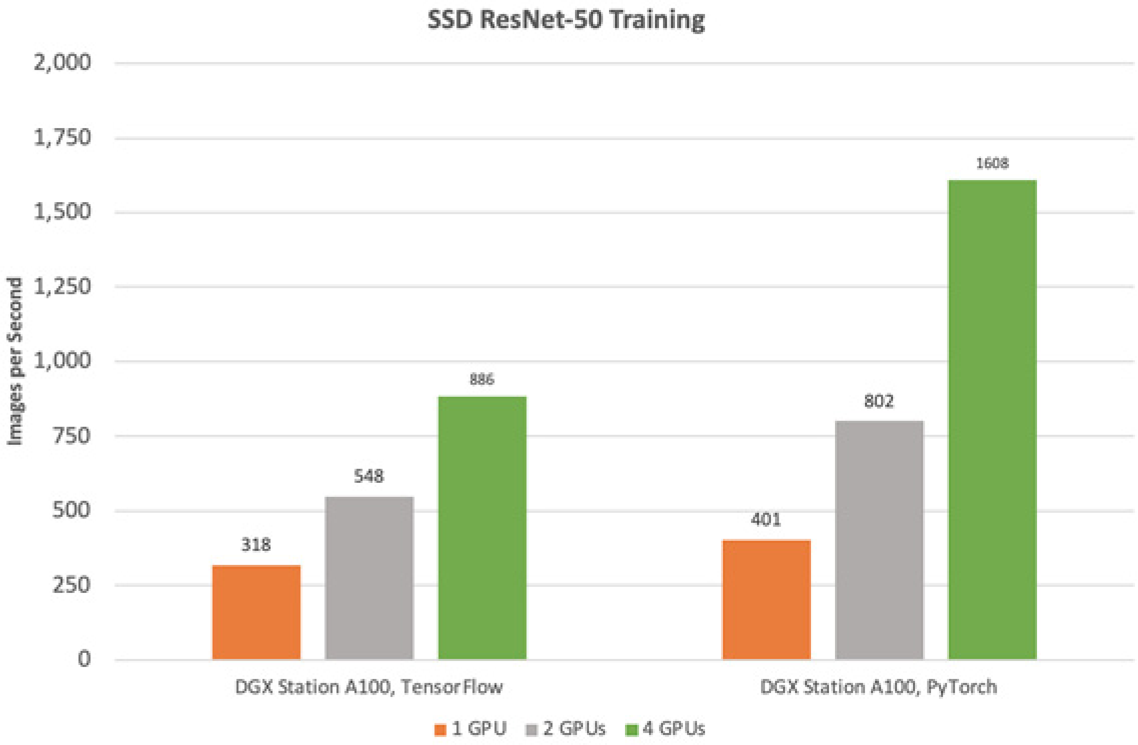


Figure 21. DGX Station A100 (4x A100 SXM 80GB), SSD ResNet-50 training, Mixed Precision, TensorFlow (21.02-tf1_py3 Container, bs 32), PyTorch (21.02_py3, bs 128)

5.2 Machine Learning and Data Science

Near Real-Time Experimentation with DGX Station A100 + RAPIDS

To test our products, NVIDIA has built a standard benchmark, the GPU Big Data Benchmark (GPU-BDB), to mimic real operations at a typical large retail or finance company. It consists of 30 queries requiring large-scale extract, transform, and load (ETL) operations, natural language processing, and machine learning using a mixture of structured and unstructured data at 1TB or 10TB scale. The benchmark is evaluated "end-to-end", meaning data starts and ends on disk, and everything in between is measured.

Performing the 1TB benchmark on DGX Station A100 with RAPIDS, each query was completed in an average of 13.7 seconds with a total runtime of 5 minutes. Running the same workflow on NVIDIA DGX-1® server, the average query-completion time was 68.7 seconds and total runtime was 35 minutes. When compared with DGX-1 server, DGX Station A100 is 5X more performant.

Beyond delivering server-grade performance in a compact, office floor-friendly package, DGX Station A100 offers near real-time interactivity for data science teams solving challenging business problems. This level of interactivity makes it easy for teams to get more done in less time and helps enterprises get the most value out of their data.

6 Direct Access to NVIDIA DGXperts

NVIDIA DGXperts are a global team of over 16,000 AI-fluent professionals who have gained the experience of thousands of DGX system deployments and who have expertise in full-stack AI development. Their skill set includes system design and planning, data center design, workload testing, job scheduling, resource management, and on-going optimizations.

Owning an NVIDIA DGX A100 or any other DGX system gives you direct access to these experts as part of NVIDIA Enterprise Support Services. NVIDIA DGXperts complement your in-house AI expertise and let you combine an enterprise-grade platform with augmented AI-fluent talent to achieve your organization's AI project goals.

7 Summary

The innovations in NVIDIA DGX Station A100 system make it possible for developers, researchers, IT managers, business leaders, and more to push the boundaries of what's possible and realize the full benefits of AI in their projects and across their organizations.

It's a powerful system that can be shared by whole teams of data scientists. Designed as a workgroup server for the age of AI, it's capable of running training, inference, and analytics workloads in parallel, and with MIG, it can provide up to 28 separate GPU devices to individual users and jobs so that activity is contained and doesn't impact performance across the system.

These data science teams might work in corporate offices, labs, research facilities, or even at home. Whereas installing large-scale AI infrastructure requires significant IT investment and large data centers with industrial-strength power and cooling, DGX Station A100 simply plugs into any standard wall outlet, wherever your workspace may be. And its innovative, refrigeration-based design means that it stays cool to the touch

NVIDIA DGX Station A100 provides a data center-class AI server in a workstation form factor, suitable for use in a standard office environment without specialized power and cooling. Its design includes four ultra-powerful NVIDIA A100 Tensor Core GPUs, a top-of-the-line server-grade CPU, super-fast NVMe storage, and leading-edge PCIe Gen4 buses. DGX Station A100 also includes the same Baseboard Management Controller (BMC) as NVIDIA DGX A100, allowing system administrators to perform any required tasks over a remote connection. DGX Station A100 is the most powerful AI system for an office environment, providing data center technology without the data center.

To learn more, visit:

- NVIDIA DGX Station A100 web page
- NVIDIA DGX Station A100 data sheet
- NVIDIA Ampere Architecture In-Depth DevBlog

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, DGX, DGX-1.DGX-2,.DGX Station, NVIDIA Turing, and NVIDIA Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright