NVIDIA DGX A100 System Architecture

The Universal System for AI Infrastructure

Technical White Paper

WP-10083-001_v01 | July 2020

Table of Contents

1 Introduction 1
2 System Architecture 2
3 NVIDIA A100 GPU - 8th Generation Data Center GPU for the Age of Elastic Computing 3
    3.1 Third-Generation Tensor Cores 3
    3.2 TensorFloat-32 (TF32) Uses Tensor Cores by Default 4
    3.3 Fine-grained Structured Sparsity 6
    3.4 Multi-Instance GPU (MIG) 7
4 Third-Generation NVLink and NVSwitch to Accelerate Large Complex Workloads 10
5 Highest Networking Throughput with Mellanox ConnectX-6 11
6 First Accelerated System With All PCIe Gen4 12
7 Security 13
    7.1 Self-Encrypted Drives 13
    7.2 Trusted Platform Module (TPM) Technology 13
8 Fully Optimized DGX Software Stack 14
9 Game Changing Performance 16
10 Breaking AI Performance Records for MLPerf v0.7 Training 17
11 Direct Access to NVIDIA DGXperts 18
12 Summary 18
13 Appendix: Graph Details 19
    13.1 Details for Figure 7: Inference Throughput with MIG 19
    13.2 Details for Figure 12: DGX A100 AI Training and Inference Performance 20

1 Introduction

Organizations of all kinds are incorporating AI into their research, development, product, and business processes. This helps them meet and exceed their particular goals, and also helps them gain experience and knowledge to take on even bigger challenges. However, traditional compute infrastructures are not suitable for AI due to slow CPU architectures and varying system requirements for different workloads and project phases. This drives up complexity, increases cost, and limits scale.

To help organizations overcome these obstacles and succeed in a world that desperately needs the power of AI to solve big challenges, NVIDIA designed the world's first family of systems purpose-built for AI—NVIDIA DGX™ systems. By leveraging powerful NVIDIA GPUs and designing from the ground up for multiple GPUs and multi-node deployments with DGX POD™ and DGX SuperPOD™ reference architectures along with optimized AI software from NVIDIA NGC™, DGX systems deliver unprecedented performance and scalability, and eliminate integration complexity.

Built on the brand new NVIDIA A100 Tensor Core GPU, NVIDIA DGX™ A100 is the third generation of DGX systems. Featuring 5 petaFLOPS of AI performance, DGX A100 excels on all AI workloads–analytics, training, and inference–allowing organizations to standardize on a single system that can speed through any type of AI task and dynamically adjust to changing compute needs over time. And with the fastest I/O architecture of any DGX system, NVIDIA DGX A100 is the foundational building block for large AI clusters such as NVIDIA DGX SuperPOD, the enterprise blueprint for scalable AI infrastructure that can scale to hundreds or thousands of nodes to meet the biggest challenges. This unmatched flexibility reduces costs, increases scalability, and makes DGX A100 the universal system for AI infrastructure.

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

2 System Architecture

Figure 1 shows an exploded view of the major components in the NVIDIA DGX A100 system, explained in detail in this white paper.

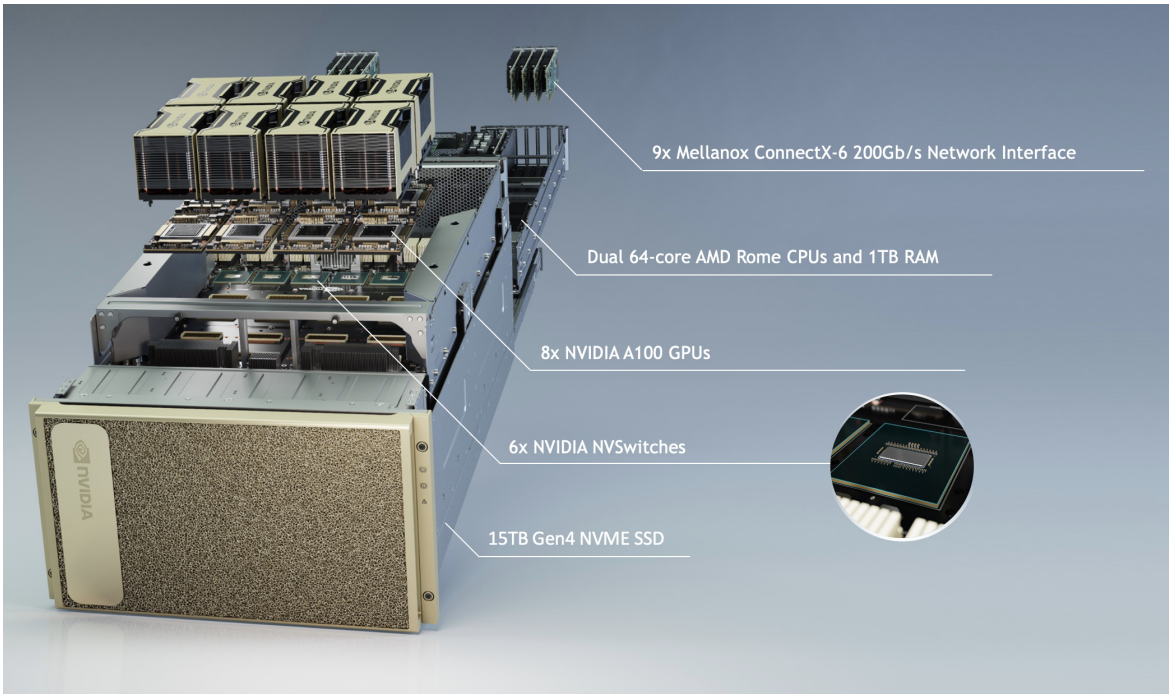


Figure 1. Major components inside NVIDIA DGX A100 system

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

At the core, the NVIDIA DGX 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.

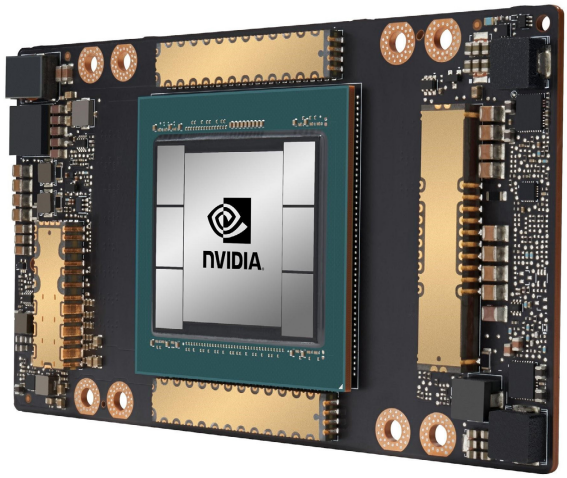


Figure 2. NVIDIA A100 Tensor Core GPU

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

3.1 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:

- **TensorFloat-32 (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]

3.2 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.

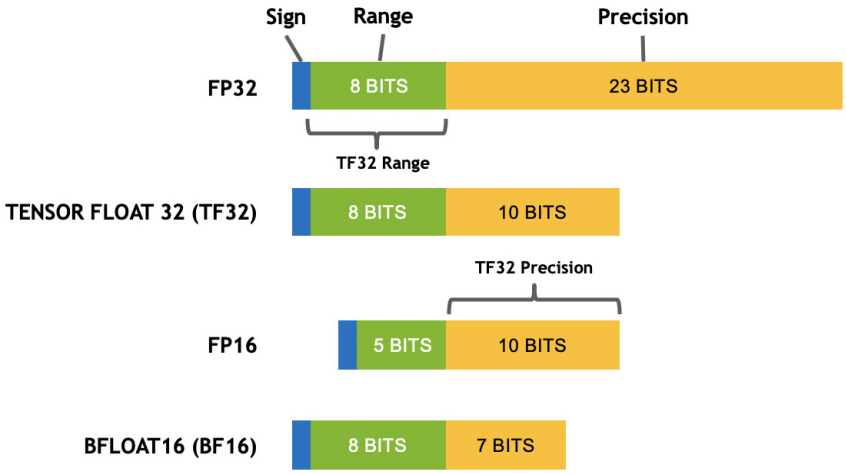


Figure 3. Explanation of Tensor Float 32, FP32, FP16, and BF16

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.

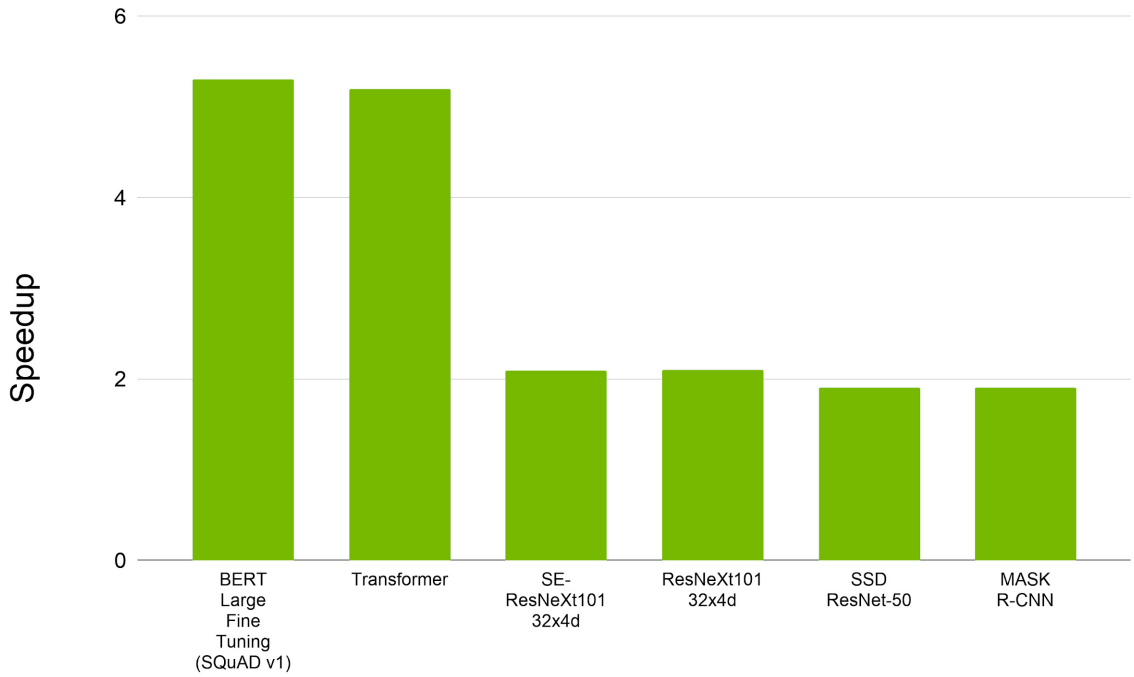


Figure 4. TF32 can provide over 5X speedup compared to FP32, PyTorch 1.6 in NGC pytorch:20.06-py3 container, training on BERT-Large model. Results on DGX A100 (8x A100 GPUs). All model scripts can be found in the Deep Learning Examples repository

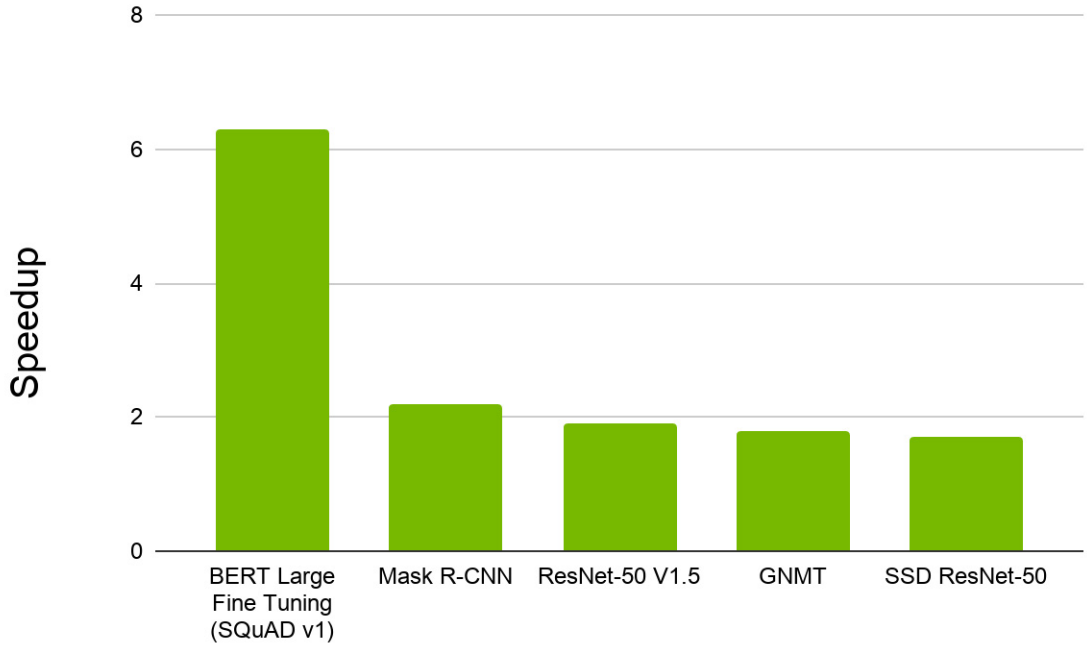


Figure 5. TF32 can provide over 6X speedup compared to FP32, TensorFlow 1.15 in NGC tensorflow:20.06-tf1-py3 container, training on BERT-Large model. Results on DGX A100 (8x A100 GPUs). All model scripts can be found in the Deep Learning Examples repository

3.3 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.

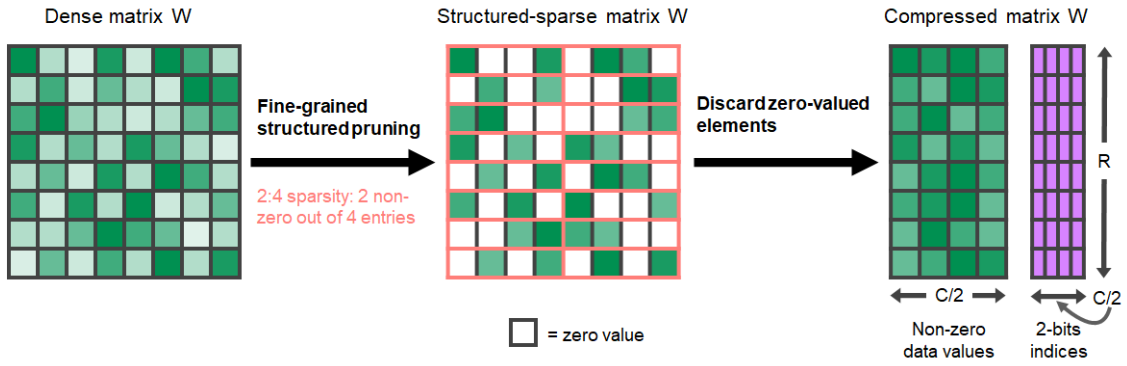


Figure 6. NVIDIA A100 GPU supports fine-grained structured sparsity with an efficient compressed format and 2X instruction throughput.

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

3.4 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 A100 system - not all GPUs need to be MIG-enabled. However, if all GPUs in a DGX A100 system are MIG enabled, up to 56 users can simultaneously and independently take advantage of GPU acceleration.

Typical uses cases that can benefit from MIG are

- 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
- Jupyter notebooks for model exploration
- Resource sharing of the GPU among multiple users

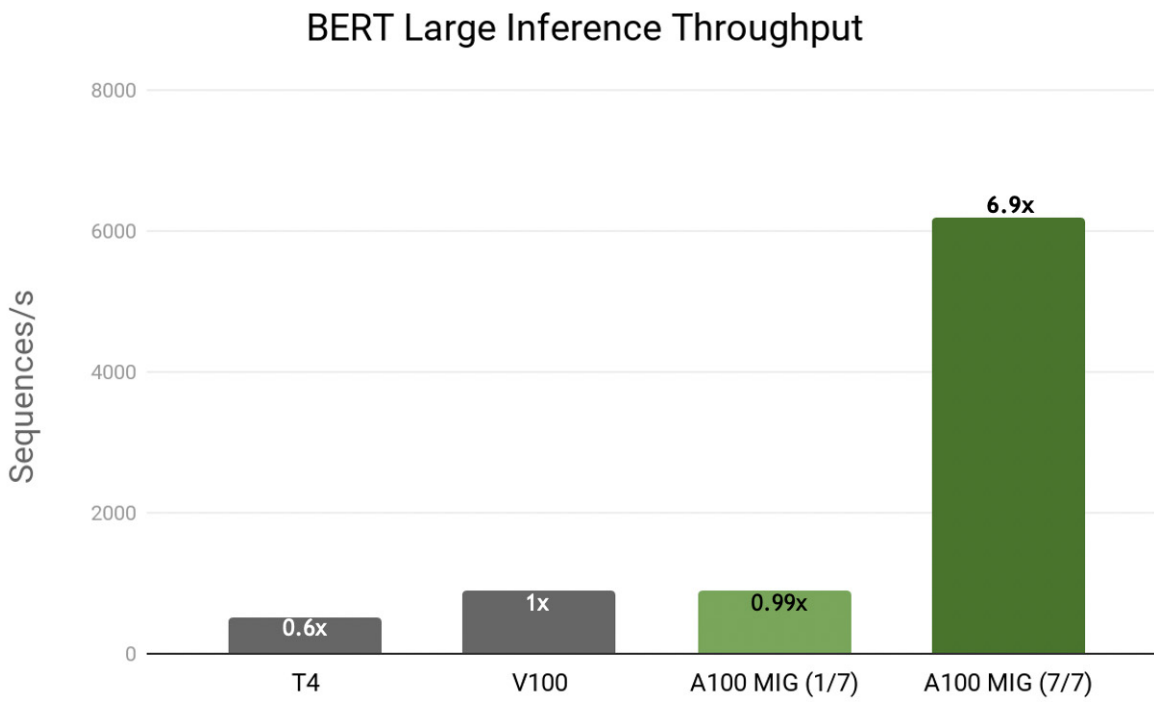


Figure 7. Up to 7X Higher Inference throughput with Multi-Instance GPU (MIG)[1]

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

1.Refer to Details for Figure 7: Inference Throughput with MIG" on page 19 for more details.

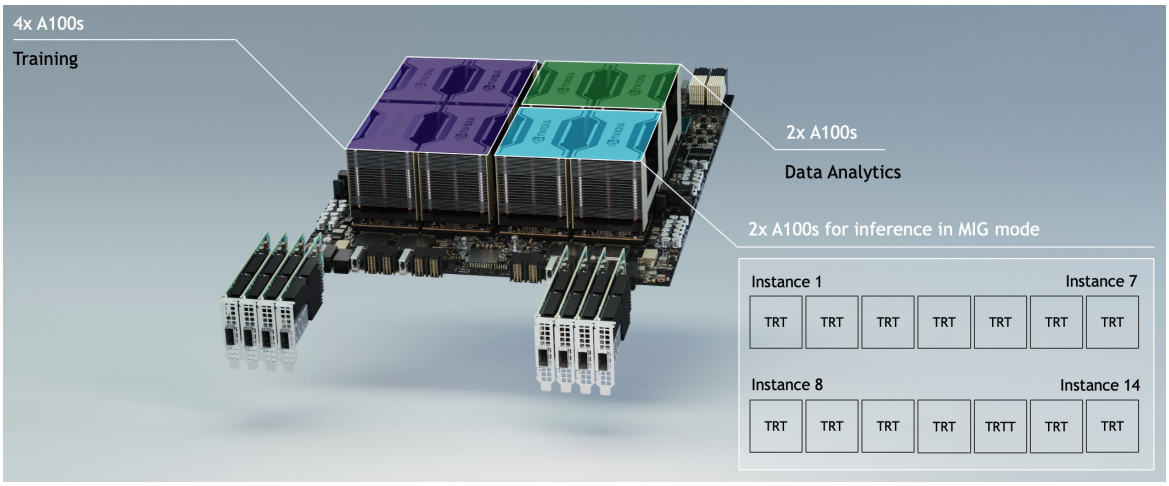


Figure 8. Different workloads on different GPUs in NVIDIA DGX A100 system

MIG supports a variety of deployment options, allowing users to run CUDA applications on bare-metal, containers or scale out with the Kubernetes container management platform. 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. More information is available here.

Under Kubernetes, GPU support has traditionally been achieved using the Device Plugin API and the NVIDIA Device Plugin. The NVIDIA device plugin for Kubernetes is a Daemonset that allows GPUs to be advertised on each of the nodes in the cluster and users to request devices (GPUs) in their job specification. The NVIDIA device plugin has been extended to enumerate the MIG devices, so that these resources can be requested as regular GPUs. NVIDIA also provides Helm charts for easily deploying the device plugin into a Kubernetes cluster. The GitHub repo includes information on getting started.

4 Third-Generation NVLink and NVSwitch to Accelerate Large Complex Workloads

The DGX A100 system contains six second-generation NVIDIA® NVSwitch™ fabrics that interconnect the A100 GPUs using third-generation NVIDIA NVLink® high-speed interconnects. Each A100 GPU uses twelve NVLink interconnects to communicate with all six NVSwitches, which means there are two links from each GPU to each switch. This provides a maximum amount of bandwidth to communicate across GPUs over the links.

The second-generation NVSwitch [Figure 9] is two times faster than the previous version, which was first introduced in NVIDIA DGX-2 system. The combination of six NVSwitches and third-generation NVLinks enables individual GPU to GPU communication to peak at 600 GB/s, which means that if all GPUs are communicating with each other, the total amount of data transferred peaks at 4.8 TB/s for both directions.

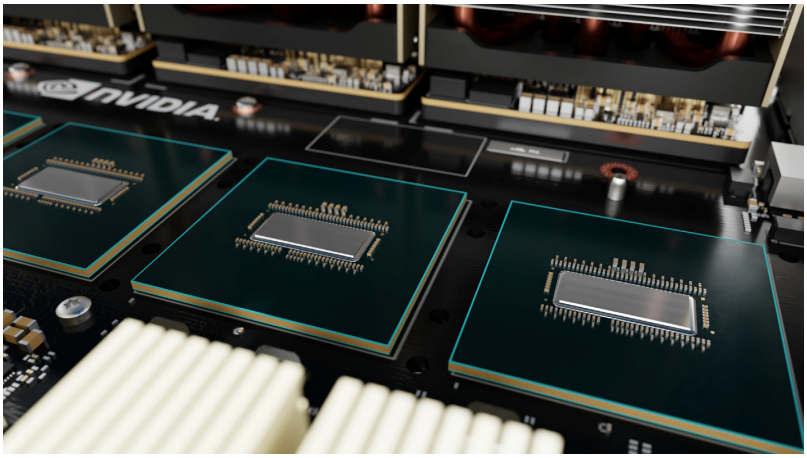


Figure 9. Second-generation NVIDIA NVSwitches in NVIDIA DGX A100 system

5 Highest Networking Throughput with Mellanox ConnectX-6

Multi-system scaling of AI deep learning and HPC computational workloads requires strong communications between GPUs in multiple systems to match the significant GPU performance of each system. In addition to NVLink for high-speed communication internally between GPUs, the DGX A100 is purpose-built for multi-system AI scaling with eight single-port Mellanox ConnectX-6 200Gb/s HDR InfiniBand ports (also configurable as 200Gb/s Ethernet ports) providing 3.2 Tb/s of peak bandwidth from a single system that can be used to immediately build a high-speed cluster of DGX A100 systems such as NVIDIA DGX SuperPOD.

The most common methods of moving data to and from the GPU involve leveraging the on-board storage and using the Mellanox ConnectX-6 network adapters through Remote Direct Memory Access (RDMA). The DGX A100 incorporates a one-to-one relationship between the IO cards and the GPUs, which means each GPU can communicate directly with external sources without blocking other GPUs' access to the network.

The Mellanox ConnectX-6 I/O cards offer flexible connectivity as they can be configured as HDR InfiniBand or 200Gb/s Ethernet. This allows the NVIDIA DGX A100 to be clustered with other nodes to run HPC and AI workloads using low latency, high bandwidth InfiniBand, or RDMA over Converged Ethernet (RoCE).

The DGX A100 includes an additional dual-port ConnectX-6 card that can be used for high-speed connection to external storage. The flexibility in I/O configuration also allows connectivity to a variety of high-speed networked storage options.

Figure 10. Mellanox Single-port ConnectX-6 VPI card for highest network throughput

The ConnectX-6 VPI cards (Figure 10) in DGX A100 provide:

- 200 Gb/s per port (4 data lanes operating at 50 Gb/s or 200 Gb/s total)
- Both IBTA RDMA (Remote Data Memory Access) and RoCE (RDMA over Converged Ethernet) technologies;

- Low-latency communication and built-in primitives and collectives to accelerate large computations across multiple systems;
- High performance network topology support to enable data transfer between multiple systems simultaneously with minimal contention;
- NVIDIA GPUDirect RDMA across InfiniBand for direct transfers between GPUs in multiple systems.

The latest DGX A100 multi-system clusters use a network based on a fat tree topology using advanced Mellanox adaptive routing and Sharp collective technologies to provide well-routed, predictable, contention-free communication from each system to every other system. A fat tree is a tree-structured network topology with systems at the leaves that connect up through multiple switch levels to a central top-level switch. Each level in a fat tree has the same number of links providing equal non-blocking bandwidth. The fat tree topology ensures the highest communication bisection bandwidth and lowest latency for all-to-all or all-gather type collectives that are common in computational and deep learning applications.

With the fastest I/O architecture of any DGX system, NVIDIA DGX A100 is ideally suited for large AI clusters such as NVIDIA DGX SuperPOD.

6 First Accelerated System 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 Mellanox ConnectX-6, the NVMe storage, and the CPUs.

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.

Each DGX A100 system comes with dual 1.92 TB NVMe M.2 boot OS SSDs configured in a RAID 1 volume, and four 3.84 TB PCIe gen4 NVMe U.2 cache SSDs configured in a RAID 0 volume. The base RAID 0 volume has a total capacity of 15 TB, but an additional 4 SSDs can be added to the system for a total capacity of 30 TB. These drives use CacheFS to increase the speed at which workloads access data and to reduce network data transfers.

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 A100 system comes with two of these CPUs for boot, storage management, and deep learning framework scheduling and coordination. Each CPU runs at a maximum speed of 3.4GHz, has 64 cores with 2 threads per core.

For I/O, the AMD Epyc processor offers 128 PCIe Gen4 links per processor, for a total of 256 in the system. This provides the system with maximum bandwidth from the processor as well as flexibility on the number of devices that can be connected directly to it. In the case of the DGX A100, the PCI lanes are used for socket-to-socket communication, direct access to a number of PCI switches that extend to eight GPUs, InfiniBand interconnects and high-speed storage.

The CPU provides extensive memory capacity and bandwidth. Each has 8 memory channels for an aggregate of 204.8 GB/s of memory bandwidth per CPU. Memory capacity on the DGX A100 is 1TB standard with 16 DIMM slots populated, expandable to 2TB with all 32 DIMM slots populated.

Similar to previous DGX systems, DGX A100 is designed to be air cooled in a data center with operating temperature ranging from 5°C - 30°C.

7 Security

The NVIDIA DGX A100 system supports self-encrypted drives and Trusted Platform Module (TPM) technology for added security.

7.1 Self-Encrypted Drives

The NVIDIA DGX™ OS software supports managing self-encrypting drives (SEDs). SEDs encrypt data on the drives automatically without user intervention. For additional security, DGX allows configuration of an Authentication Key so that drives lock on shut down, requiring the key to be entered at power on to unlock the drives for use.

7.2 Trusted Platform Module (TPM) Technology

The NVIDIA DGX A100 system includes a secure cryptoprocessor which conforms to the Trusted Platform Module (TPM 2.0)[2] 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.

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

8 Fully Optimized DGX Software Stack

The DGX 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 A100 with minimal setup effort. The design of the platform software is centered around a minimal OS and driver install on the server, and provisioning of all application and SDK software available through the NGC Private Registry.

The NGC Private Registry provides 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). This software has been developed, tested, and tuned on DGX systems, and is compatible with all DGX products: DGX-1, DGX-2, DGX Station, and DGX A100. The NGC Private Registry also provides a secure space for storing custom containers, models, model scripts, and Helm charts that can be shared with others within the enterprise. Learn more about the NGC Private Registry in this blog post.

Figure 11 shows how all these pieces fit together as part of the DGX software stack.

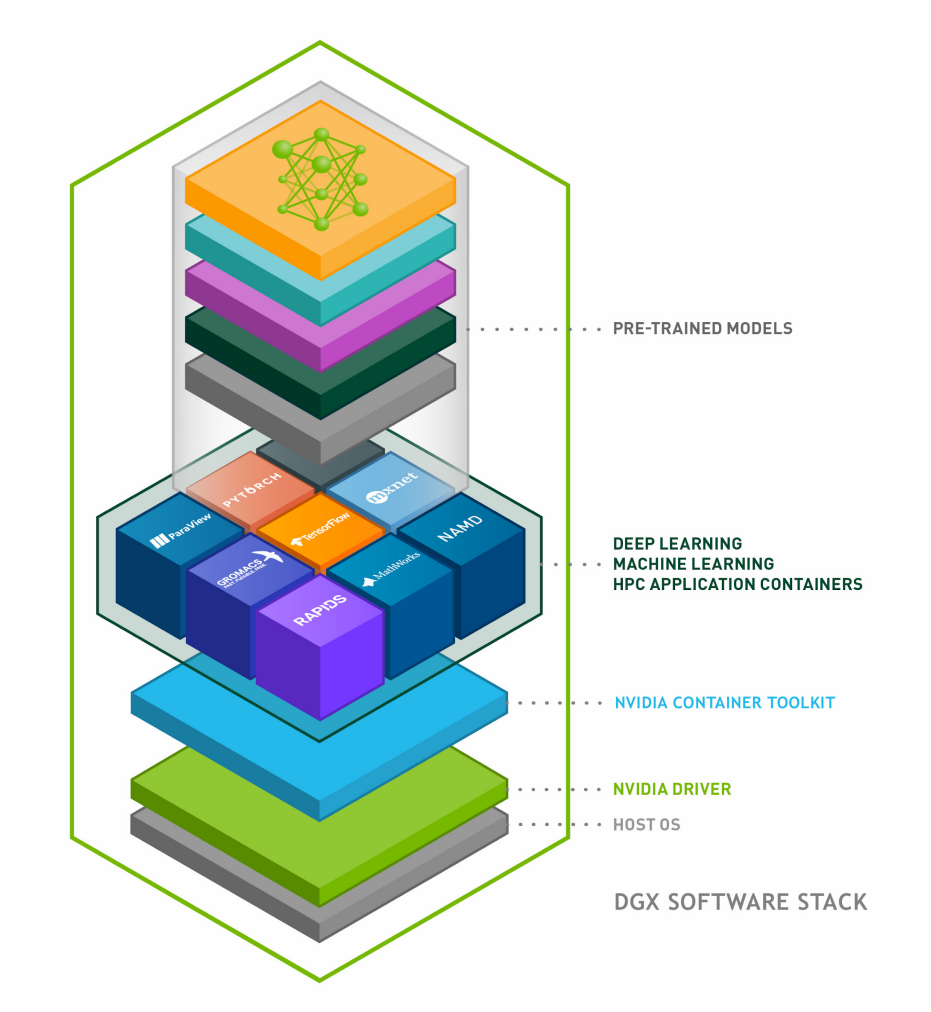


Figure 11. NVIDIA DGX Software Stack

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.

9 Game Changing Performance

Packed with innovative features and a balanced system design, the DGX A100 delivers unprecedented performance in Deep Learning Training and Inference. The graph in Figure 12 demonstrates the following performance results:

- DGX A100 delivers **6X** the training performance of an 8x V100/16GB GPU system, such as the DGX-1. DGX A100 using TF32 precision achieves **1823 sequences per second** compared to 308 sequences per second on DGX-1 using FP32.
- DGX A100 offers inference performance that is **172X** the performance of a CPU server. DGX A100 using INT8 with structured sparsity achieves inference peak compute of **10 PetaOPS**, where a Peta equates to 1000 trillion, compared to 58 trillion operations per second (TOPS) on 2x Intel Platinum 8280 CPU server using INT8.

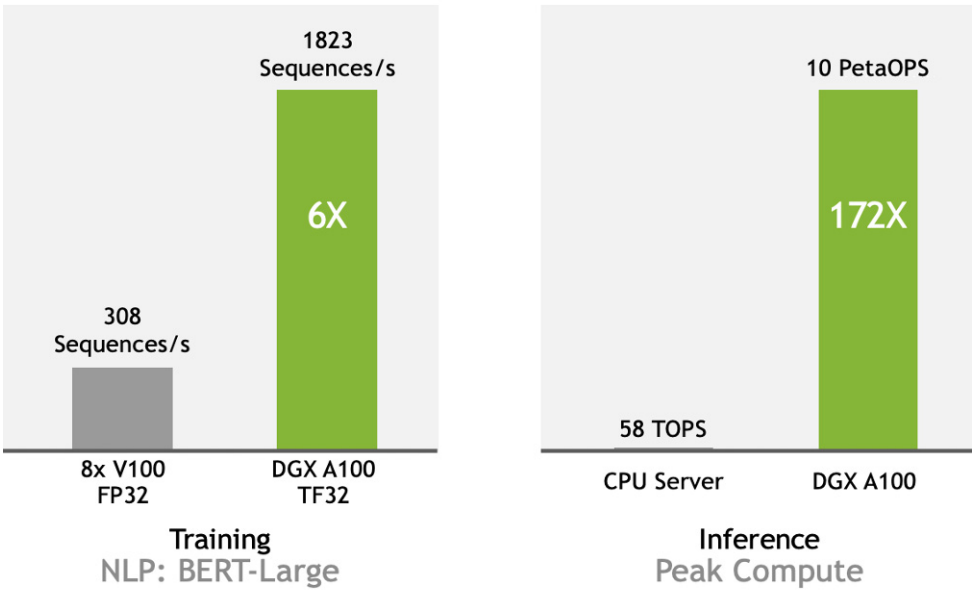


Figure 12. DGX A100 delivers unprecedented AI performance for training and inference[3]

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

A single DGX 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 A100 provides 320 GB of high bandwidth GPU memory. NVIDIA interconnect technologies like NVLink, NVSwitch and Mellanox networking bring all GPUs together to work as one on large AI models with high-bandwidth communication for efficient scaling.

3.Refer to Details for Figure 12: DGX A100 AI Training and Inference Performance" on page 20 for additional information.

With Tensor Core acceleration of INT8 and fine-grained Structured Sparsity in NVIDIA A100 GPU, the DGX A100 sets a new bar for inference workloads. Using the Multi-Instance GPU capability in A100 GPU, users can assign resources that are right-sized for specific workloads on all 8 GPUs in DGX A100 system.

10 Breaking AI Performance Records for MLPerf v0.7 Training

In MLPerf, the industry-standard benchmark for machine learning, NVIDIA DGX A100 as a single node and DGX SuperPOD, cluster of DGX A100 systems, demonstrated world-class performance and versatility. NVIDIA DGX SuperPOD, combining multiple NVIDIA DGX A100 systems with NVIDIA Mellanox network fabric, gives businesses a proven formula to shorten their design and deployment cycles.

NVIDIA DGX SuperPOD with DGX A100 systems set world records for fastest time to solution in all 8 of the at scale benchmarks in MLPerf v0.7 Training and the NVIDIA A100 GPU also demonstrated fastest per accelerator performance overall in the commercially available systems category. The commercially available systems category submissions represent the state of the art of hardware and software that is available for the industry to benefit from today.

The MLPerf v0.7 training benchmark suite includes vision, language, recommenders, and reinforcement learning workloads.

Table 1 NVIDIA MLPerf AI Records

Benchmark	Fastest Time to Solution (DGX SuperPOD)	Per Accelerator Performance (Results based on one GPU in a DGX A100 system)
Recommendation : DLRM	3.33 Min	0.44 Hrs
NLP : BERT	0.81 Min	6.53 Hrs
Reinforcement Learning : MiniGo	17.07 Min	39.96 Hrs
Translation (Non-recurrent): Transformer	0.62 Min	1.05 Hrs
Translation (Recurrent) : GNMT	0.71 Min	1.04 Hrs
Object Detection (Heavy Weight) : Mask R-CNN	10.46 Min	10.95 Hrs
Object Detection (Light Weight) SSD	0.82 Min	1.36 Hrs

Table 1 NVIDIA MLPerf AI Records (Continued)

Benchmark	Fastest Time to Solution (DGX SuperPOD)	Per Accelerator Performance (Results based on one GPU in a DGX A100 system)
Image Classification (ResNet-50 v1.5)	0.76 Min	5.30 Hrs

Details:

Per Chip Performance arrived at by comparing performance at the same scale when possible. Per Accelerator comparison using reported performance for MLPerf 0.7 NVIDIA A100 (8 A100s). MLPerf ID DLRM: 0.7-17, ResNet50 v1.5: 0.7-18, 0.7-15 BERT, GNMT, Mask R-CNN, SSD, Transformer: 07-19, MiniGo: 0.7-20.

Max Scale: All results from MLPerf v0.7 using NVIDIA DGX A100 (8xA100s)\. MLPerf ID Max Scale: ResNet50 v1.5: 0.7-37, Mask R-CNN: 0.7-28, SSD: 0.7-33, GNMT: 0.7-34, Transformer: 0.7-30, MiniGo: 0.7-36, BERT: 0.7-38, DLRM: 0.7-17.

MLPerf name and logo are trademarks. See www.mlperf.org for more information

11 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.

12 Summary

The innovations in the NVIDIA DGX 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.

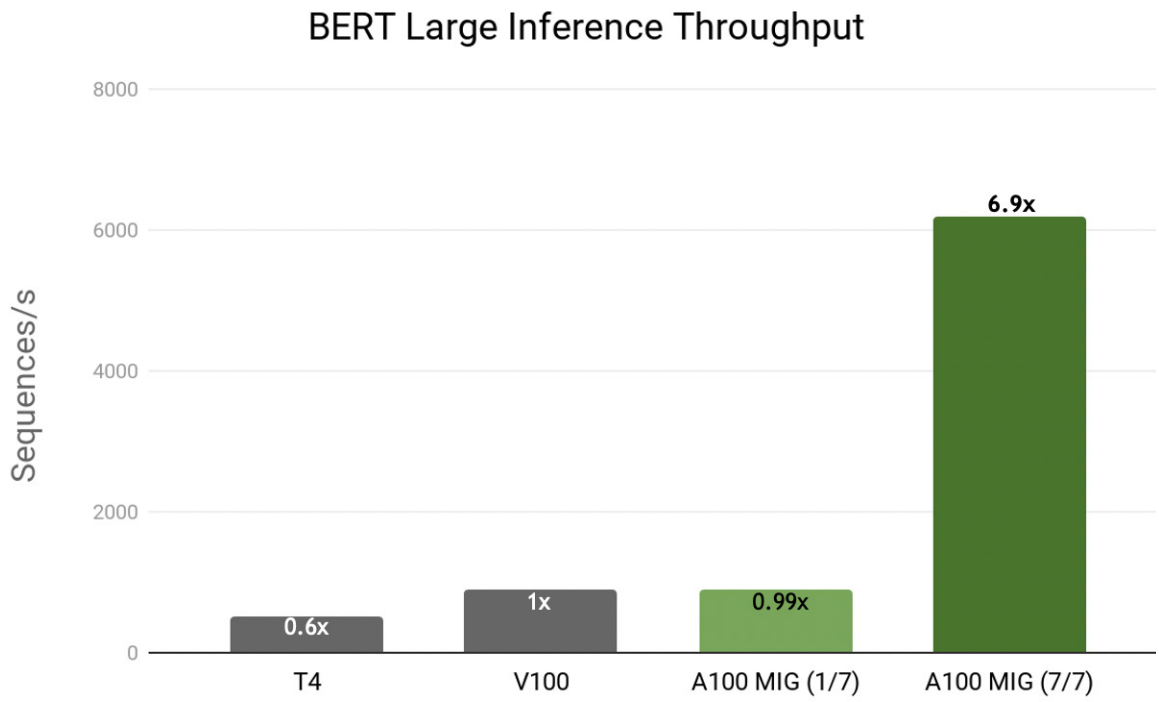
To learn more, visit:

- NVIDIA DGX A100 web page
- NVIDIA DGX A100 data sheet
- NVIDIA DGX SuperPOD reference architecture
- NVIDIA Ampere Architecture In-Depth DevBlog

13 Appendix: Graph Details

This appendix provides supplemental information for the performance graphs presented in this white paper.

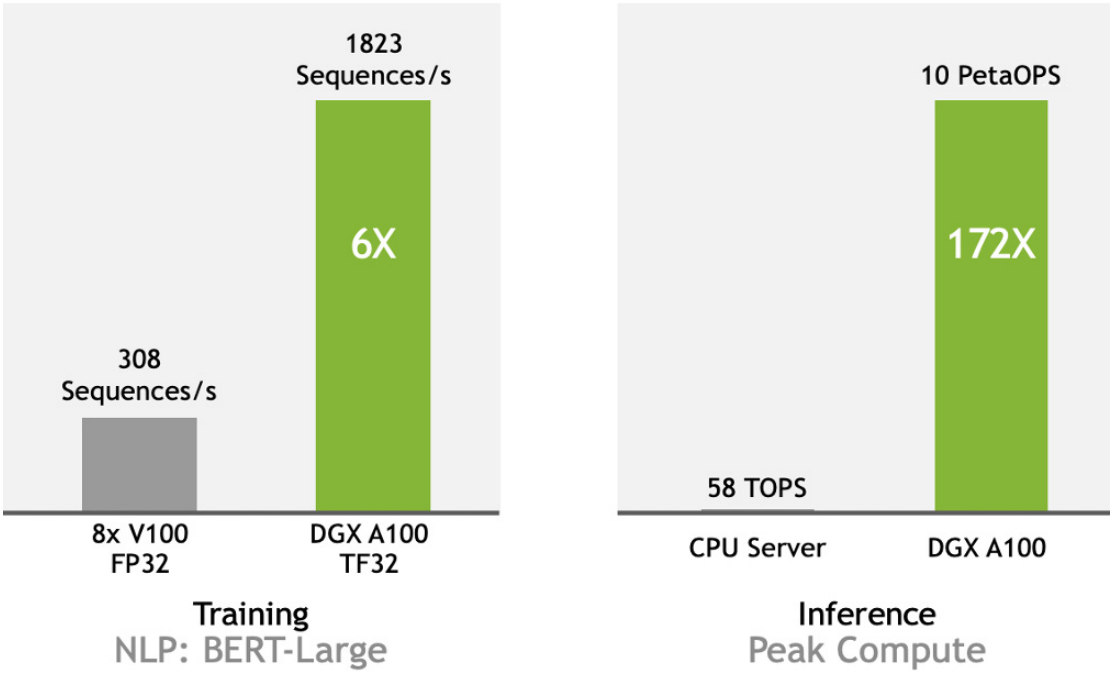
13.1 Details for Figure 7: Inference Throughput with MIG



Results on 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)

13.2 Details for Figure 12: DGX A100 AI Training and Inference Performance



- Training:
  - > DGX A100 system with 8x NVIDIA A100 GPUs, TF32 precision vs. DGX-1 system with 8x NVIDIA V100/16GB GPUs, FP32 precision.
  - > Deep learning language model: the large version of one of the world's most advanced AI language models–Bidirectional Encoder Representations from Transformers (BERT) on the popular PyTorch framework.
  - > Pre-training throughput using PyTorch NGC Container 20.06, sequence length 128
- Inference:
  - > DGX A100 system with 8x NVIDIA A100 GPUs using INT8 with Structured Sparsity vs. a CPU server with 2x Intel Platinum 8280 using INT8.

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, CUDA, NVIDIA POD, and NVIDIA SuperPOD 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