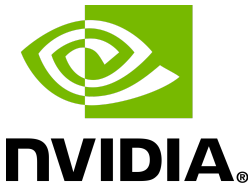

White Paper

NVIDIA DGX Station

AI Workstation for Data Science and Research Teams

Table of Contents

Abstract

NVIDIA® DGX Station™ (Figure 1) is the world's the *world's fastest workstation for leading-edge AI development. DGX Station features four NVIDIA® Tesla® V100 Tensor Core GPUs fully connected through NVIDIA® NVLink™, the NVIDIA high-performance GPU interconnect, and is powered by a complete DGX software stack, including NVIDIA GPU-optimized containers from NVIDIA GPU Cloud (NGC). The Tesla V100 GPU accelerator features the new Tensor Core architecture to help deliver unprecedented levels of performance not seen in any AI workstation prior. Offering whisper quiet, breakthrough performance, DGX Station gives computational scientists, data scientists and AI researchers the fastest start in deep learning, HPC, and data science from the convenience of their desks. Unlike other platforms, DGX Station software provides highly optimized libraries and containers designed for maximized deep learning performance leveraging its four Tesla V100 GPUs, providing a flexible, versatile and scalable platform for running deep learning workloads for research and in production.*

Figure 1 NVIDIA DGX Station

1.0 Introduction

Deep learning is quickly changing virtually every industry and is having a large impact on the economics of businesses:

- The number of GPU deep learning developers has leapt 10 times in the last 5 years.. NVIDIA's CUDA has been downloaded 8M times, increased 5x in 5 years.
- The number of active US startups developing AI systems has increased 14x since 2000. [AI Index, Page 16]
- Facebook CTO Mike Schroepfer noted that they have deployed more than 40 PFLOPs of GPU capability in house to support deep learning across their organization [Schroepfer 2016:6:54].
- Organizations are looking to the new-generation of HPC systems to provide the performance, reliability, and flexibility that they need. IDC estimates that for every $43 spent on HPC, users can generate $515 in revenue, making HPC adoption a highly lucrative investment. [HPCwire]

The ever-increasing computational power required for running deep learning workloads are driving the need for advanced workstations that make use of hardware accelerators to turbocharge the parallelized algorithms. According to IDC, worldwide accelerated computing infrastructure revenue will grow from $2.5 billion in 2016 to $6.8 billion in 2021, a CAGR of 21.9%. The on-premises portion of this market is forecast to grow from $1.6 billion in 2016 to $3.4 billion in 2021, at a CAGR of 16.3%. IDC expects an increasing portion of this market to be served by new form factors, including desktops and workstations. Workstations were designed for those professionals who don't want to wait in a queue for IT to approve the compute cycles in a server. They are meant to be placed in professional's office and used as a deskside-based supercomputer at the fingertips.

To satisfy this insatiable need for high performance, GPU accelerated computing, NVIDIA designed DGX Station™ (shown in Figure 1), which is the world's fastest workstation for deep learning training and data science workloads. NVIDIA's goal with DGX Station™ was to create the world's fastest platform for training deep neural networks and optimizing data science workloads that can be deployed quickly and run quietly by deep learning researchers and data scientists in their office. The architecture of DGX Station™ draws on NVIDIA's experience in the field of high-performance computing and knowledge gained from optimizing deep learning frameworks on NVIDIA GPUs.

2.0 NVIDIA DGX Station Architecture

DGX Station is a deep learning and data science workstation architected for high performance, multi-GPU neural network training equivalent or better than what's traditionally found in a data center, now placed at the developer's fingertips. The core of the system is a complex of four NVIDIA® Tesla® V100 GPUs (with 32 GB memory each) in a fully connected NVLink™ topology, described in Section 2.1.2. Using mixed-precision multiply and accumulate operations, the new Tensor Core architecture enables Volta V100 to deliver the performance required to train large neural networks. In addition to the four GPUs, DGX Station™ includes one 20-core CPU, fast local storage (3 SSDs configured in RAID 0), a water-cooling system for the GPUs, dual 10 GbE networking, all balanced to optimize throughput and deep learning training time.

As mentioned above, NVIDIA® Tesla® V100 now offers a 32 GB high bandwidth memory configuration. Providing 2X the memory capacity improves deep learning training performance for next-generation AI models like language translations and ResNet 1K models by over 50%, by training more data in parallel with these larger models. Supporting larger AI and data science models also improves AI developers' and data scientists' productivity, allowing them to deliver more AI and data science breakthroughs in less time. This higher memory configuration allows HPC applications to run larger simulations more efficiently than ever before. For example, large 3D FFT calculations which are commonly used in seismic, climate, and signal processing applications are up to 50% faster on the V100 GPU.

Figure 2 shows the DGX Station™ system components.

1. GPUs
4X NVIDIA Tesla® V100 32 GB/GPU
500 TFLOPS (Mixed Precision)
20,480 Total NVIDIA CUDA® Cores
2,560 Tensor Cores

2. SYSTEM MEMORY
256 GB RDIMM DDR4

3. GPU INTERCONNECT
NVIDIA NVLink™,
Fully Connected 4-Way

4. STORAGE
Data: 3 x 1.92 TB SSD RAID 0
OS: 1 x 1.92 TB SSD

5. CPU
Intel Xeon E5-2698 v4
2.2 GHz 20-Core

6. NETWORKING
2X 10 GbE

7. DISPLAYS
3X DisplayPort,
4K Resolution

8. COOLING
Water-Cooled

9. POWER
1500 W

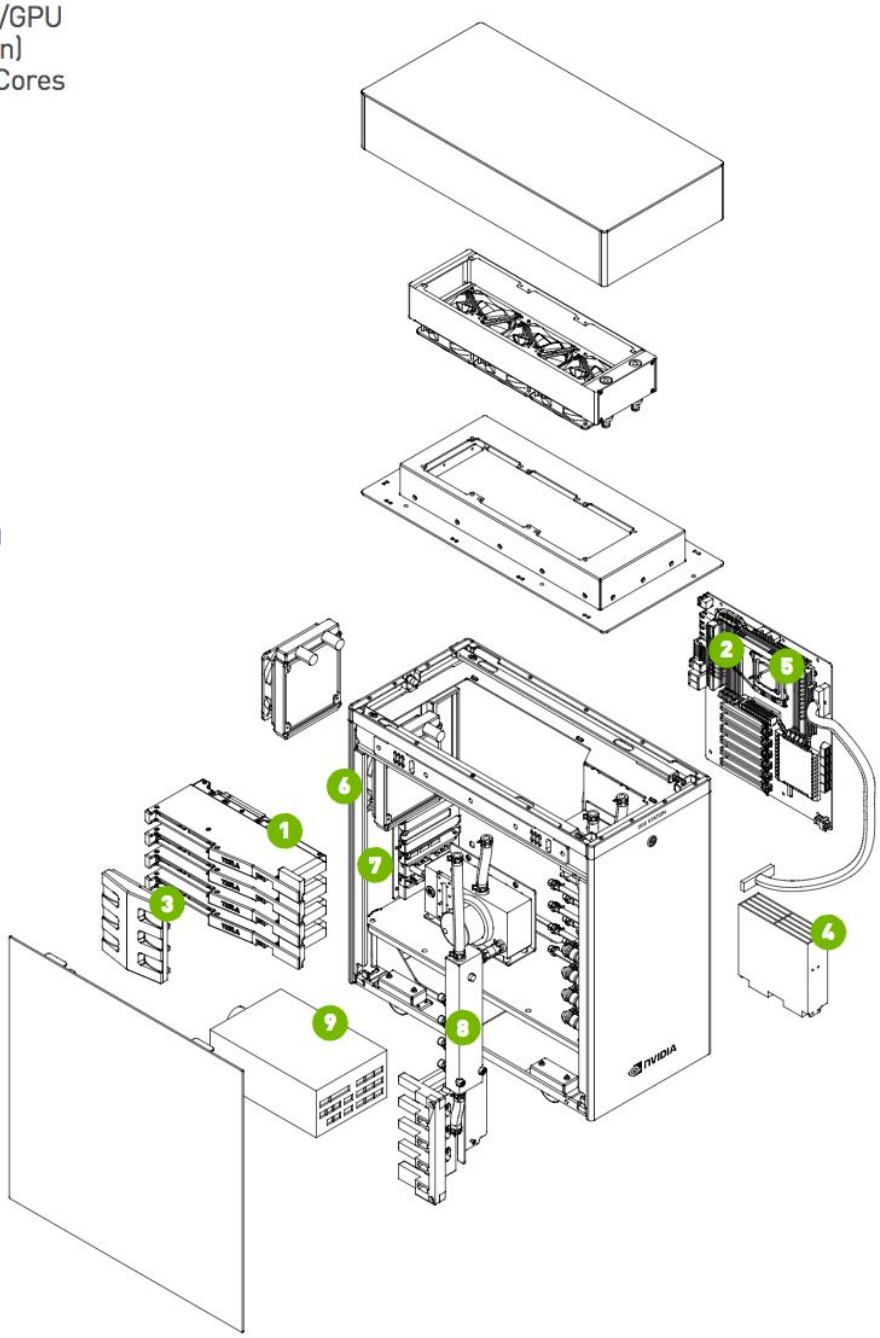

Figure 2 DGX Station™ components

2.1 NVIDIA Tesla V100

Figure 3 NVIDIA Tesla V100

Tesla V100 (Figure 3) is the latest NVIDIA accelerator, designed for high performance computing and deep learning applications [NVIDIA Corporation 2017c]. The Tesla V100 accelerator features the GV100 GPU, which incorporates 80 streaming multiprocessors (SMs), each with:

- 8 Tensor Cores;
- 64 single-precision (FP32) cores;
- 64 integer (INT32) cores;
- 32 double-precision (FP64) cores;
- 256KB of register file (RF);
- Up to 96KB of shared memory (configurable).

Tesla V100 peak[1] computational throughput is:

- 125 Tensor TFLOP/s [NVIDIA Corporation 2017c];
- 15.7 TFLOP/s for FP32;
- 7.8 TFLOP/s for FP64.

To support this high computational throughput, Tesla V100 incorporates HBM2 (High Bandwidth Memory version 2). V100 includes 32 GB of HBM2 stacked memory with 900 GB/s of bandwidth; significantly higher than the bandwidth of GDDR5 RAM. Because HBM2 memory is stacked memory located on the same physical package as the GPU, it provides considerable space savings compared to traditional GDDR5, which enables high-density GPU workstations like DGX Station, including its water-cooling blocks.

Each Tesla V100 in DGX Station has 4 NVLink connections each capable of 50 GB/s of bidirectional bandwidth, for an aggregate of up to 200 GB/s bidirectional bandwidth. NVLink and the DGX Station interconnect topology and its implications are discussed in detail in Section 3.

[1] Based on GPU Boost Clock.

The PCIe links between the GPUs and CPUs provide access to system memory to enable working set and dataset streaming to and from the GPUs. The system memory capacity is twice the GPU memory capacity to enable simplified buffer management and balance for deep learning workloads. In addition to the 256 GB of system memory, the four Tesla V100 GPUs have a total of 128 GB HBM2 memory with net GPU memory bandwidth of 4 × 900 GB/s = 3.6 TB/s.

Following are some key compute features of Tesla V100 (see Figure 4):

New Streaming Multiprocessor (SM) Architecture Optimized for Deep Learning
Volta features a major new redesign of the SM processor architecture that is at the center of the GPU. The new Volta SM is 50% more energy efficient than the previous generation Pascal design, enabling major boosts in FP32 and FP64 performance in the same power envelope. New Tensor Cores designed specifically for deep learning deliver up to 12x higher peak TFLOPS for training and 6x higher peak TFLOPS for inference. With independent parallel integer and floating-point data paths, the Volta SM is also much more efficient on workloads with a mix of computation and addressing calculations. Volta's new independent thread scheduling capability enables finer-grain synchronization and cooperation between parallel threads. Finally, a new combined L1 data cache and shared memory unit significantly improves performance while also simplifying programming.

Second-Generation NVIDIA NVLink™
The second generation of NVIDIA's NVLink high-speed interconnect delivers higher bandwidth, more links, and improved scalability for multi-GPU and multi-GPU/CPU system configurations. DGX Station uses four NVLink links with a total bandwidth of 200 GB/sec. to deliver greater scalability for ultra-fast deep learning training.

HBM2 Memory: Faster, Higher Efficiency
Volta's highly tuned 32 GB HBM2 memory subsystem delivers 900 GB/sec peak memory bandwidth. The combination of both a new generation HBM2 memory from Samsung, and a new generation memory controller in Volta, provides 1.5x delivered memory bandwidth versus Pascal GP100, with up to 95% memory bandwidth utilization running many workloads.

Volta Multi-Process Service
Volta Multi-Process Service (MPS) is a new feature of the Volta GV100 architecture providing hardware acceleration of critical components of the CUDA MPS server, enabling improved performance, isolation, and better quality of service (QoS) for multiple compute applications sharing the GPU. Volta MPS also triples the maximum number of MPS clients from 16 on Pascal to 48 on Volta.

Enhanced Unified Memory and Address Translation Services
GV100 Unified Memory technology includes new access counters to allow more accurate migration of memory pages to the processor that accesses them most frequently, improving efficiency for memory ranges shared between processors.

Cooperative Groups and New Cooperative Launch APIs
Cooperative Groups is a new programming model introduced in CUDA 9 for organizing groups of communicating threads. Cooperative Groups allows developers to express the granularity at which threads are communicating, helping them to express richer, more efficient parallel decompositions. Basic Cooperative Groups functionality is supported on all NVIDIA GPUs since Kepler. Pascal and Volta include support for new cooperative launch APIs that support synchronization amongst CUDA thread blocks. Volta adds support for new synchronization patterns.

Volta Optimized Software
New versions of deep learning frameworks such as Caffe2, MXNet, CNTK, TensorFlow, and others harness the performance of Volta to deliver dramatically faster training times and higher multi-node training performance. Volta-optimized versions of GPU accelerated libraries such as cuDNN, cuBLAS, and TensorRT leverage the new features of the Volta GV100 architecture to deliver higher performance for both deep learning inference and High Performance Computing (HPC) applications. The NVIDIA CUDA Toolkit version 9.0 includes new APIs and support for Volta features to provide even easier programmability.

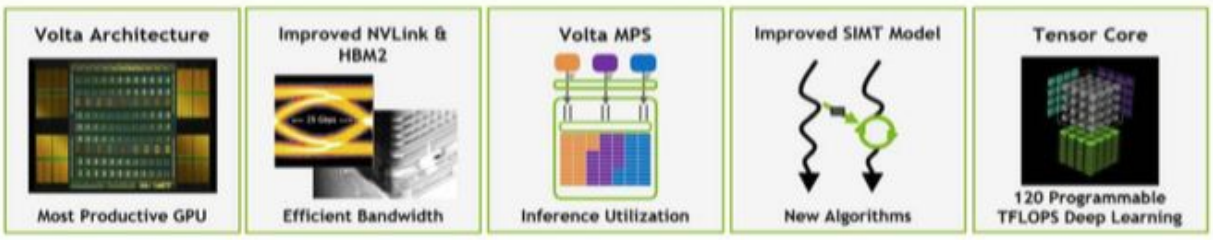

Figure 4 New Technologies in Tesla V100

2.2 Second-Generation NVIDIA NVLink™

NVLink is NVIDIA's high-speed interconnect technology first introduced in 2016 with the Tesla P100 accelerator. NVLink provides significantly more performance for both GPU-to-GPU system configurations compared to using PCIe interconnects. Tesla V100 introduces the second generation of NVLink, which provides higher link speeds, more links per GPU, CPU mastering, cache coherence, and scalability improvements.

2.3 Water-Cooling System for the GPUs

The water-cooling system for the GPUs in the DGX Station captures 90% of the GPUs' Thermal Design Power (TDP). This level of efficiency allows for whisper-quiet operation with better performance at higher TDP and more than twice as much noise abatement compared to air cooling. Its design enables both the radiator and other components to be cooled by a single set of fans.

The water-cooled design (see Figure 5) of DGX Station allows each Tesla V100 GPU board to match the performance of a V100 SXM2 module in a NVIDIA DGX-1 rack-mountable server. This means NVIDIA DGX systems provide higher performance out-of-the-box than air-cooled solutions.

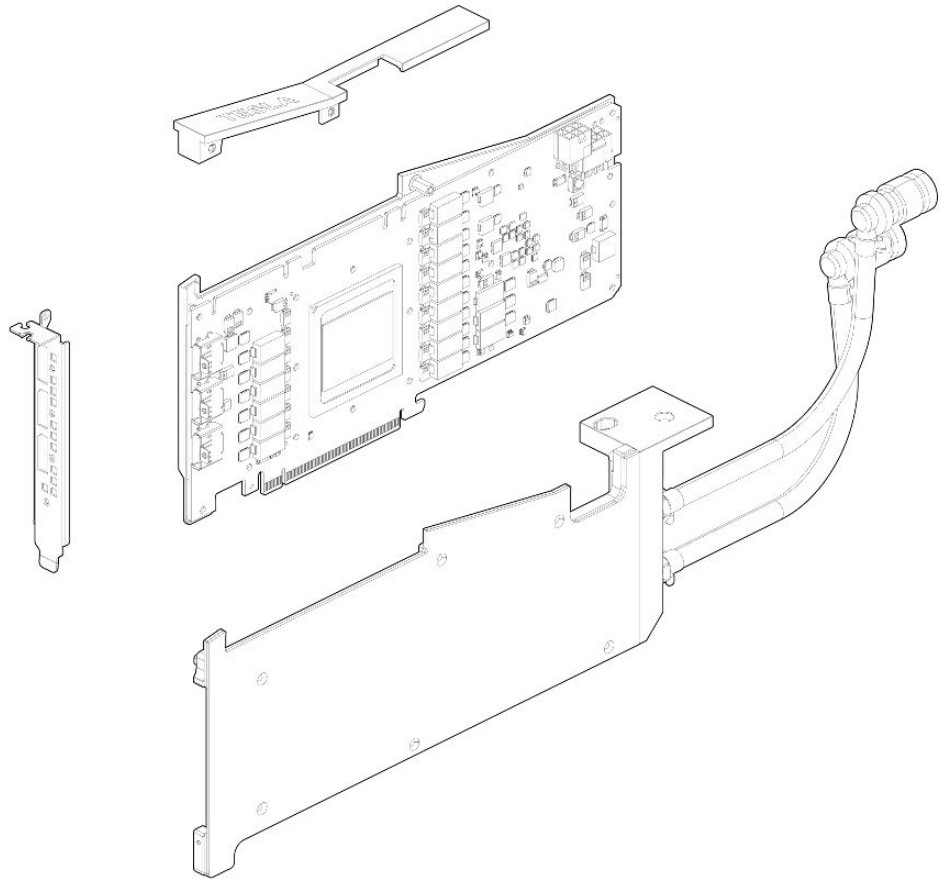

Figure 5 NVIDIA Tesla V100 for DGX Station, with the water-block assembly, exploded view

2.4 GPU and System Memory

The PCIe links between the GPUs and the CPU enable access to the CPU's bulk DRAM to enable working set and dataset streaming to and from the GPUs. The CPU memory capacity is configured with two times the GPU memory capacity, to enable simplified buffer management and balance for deep learning workloads. In addition to the 256 GB of system RDIMM DDR4, each Tesla V100 GPU includes 32 GB of HBM2 co-packaged memory with memory bandwidth 900 GB/s, yielding a total of 128 GB HBM2 memory with net GPU memory bandwidth of 3600 GB/s.

NVIDIA GPU copy engines transfer data between multiple GPUs or between GPUs and CPUs. In previous generation GPUs, performing copy engine transfers (which are like DMA transfers) could cause fatal faults if either the source or destination memory addresses were not mapped in the GPU page tables. The prior copy engines required both source and destination memory regions to be pinned (non-pageable).

The new Volta GV100 GPU copy engines can generate page faults for addresses that are not mapped into the page tables. The memory subsystem can then service the page faults, mapping the addresses into the page table, after which the copy engine can perform the transfer. This is an important enhancement, because pinning memory for multiple copy engine operations between multiple processors can substantially reduce available memory. With hardware page faulting, addresses can be passed to the copy engines without worrying if they are resident, and the copy process just works.

2.5 Other Workstation Components

Efficient, high-bandwidth streaming of training data is critical to the performance of DGX Station™ as a deep learning system, as is reliable, low failure rate storage. Each system comes configured with a single 1.92 TB boot OS SSD, and three 1.92 TB SSDs (5.76 TB total) configured as a RAID 0 striped volume for performance. External storage can be added to DGX Station via two eSATA ports as well as two USB 3.1 ports. Alternatively, a Network Attached Storage (NAS) device can be connected to a 10 GbE LAN port.

In addition to the four GPUs, DGX Station includes:

- A server-class Intel Xeon CPU (Intel Xeon E5-2698 v4 2.2 GHz) for overall system throughput and performance.
- One of the NVIDIA Tesla V100 GPU cards in the DGX Station provides three DisplayPort connectors, enabling up to three displays at 4K resolution to be connected to the DGX Station.
- Two 10GBASE-T (RJ45) ethernet ports are included for fast networking access to and from DGX Station.

DGX Station's power consumption is dynamic based on workload. The TDP of DGX Station is 1,500 W, but is equipped with a power supply rated at 1,600 W. For ease of deployment, DGX Station is designed to be operated at temperatures between 10°C and 30°C (50°F and 86°F).

3.0 Multi-GPU with NVLink

Workstations with two or more GPUs per CPU are becoming common as developers increasingly expose and leverage the available parallelism in their applications. While dense GPU systems provide a great vehicle for scaling single-node performance, multi-GPU application efficiency can be constrained by the performance of the PCIe (Peripheral Component Interconnect Express) bus connections between GPUs. Similarly, data center applications are growing outside the box, requiring efficient scaling across multiple interconnected systems. To address both of these needs, DGX Station incorporates the new NVLink high-speed GPU interconnect for multi-GPU scalability within a system.

Given that communication is an expensive operation, developers must overlap data transfers with computation or carefully orchestrate GPU accesses over PCIe interconnect to maximize performance. As GPUs get faster and GPU-to-CPU ratios climb, a higher-performance GPU interconnect is warranted.

This challenge motivated the creation of the NVLink high-speed interconnect, which enables NVIDIA GPUs to connect to peer GPUs and/or to NVLink-enabled CPUs or other devices within a node. NVLink supports the GPU ISA, which means that programs running on NVLink-connected GPUs can execute directly on data in the memory of another GPU as well as on local memory. GPUs can also perform atomic memory operations on remote GPU memory addresses, enabling much tighter data sharing and improved application scaling.

The second generation of NVLink allows direct load/store/atomic access from the CPU to each GPU's HBM2 memory. Coupled with a new CPU mastering capability, NVLink supports coherency operations allowing data reads from graphics memory to be stored in the CPU's cache hierarchy. The lower latency of access from the CPU's cache is key for CPU performance. While P100 supported peer GPU atomics, sending GPU atomics across NVLink and completed at the target CPU was not supported. NVLink adds support for atomics initiated by either the GPU or the CPU. Support for Address Translation Services (ATS) has been added allowing the GPU to access the CPU's page tables directly. A low-power mode of operation for the link has been added allowing for significant power savings when the link is not being heavily used.

The increased number of links, faster link speed, and enhanced functionality of second-generation NVLink, combined with Volta's new Tensor Cores, results in significant increases in deep learning performance in multi-GPU Tesla V100 systems.

3.1 DGX NVLink Network Topology for Efficient Application Scaling

High-performance applications typically scale their computations in one of two ways, known as strong scaling and weak scaling[2]. Strong scaling measures the improvement in time to solution when increasing the number of parallel processors applied to a fixed total problem size. With perfect strong scaling, the speedup achieved would be equal to the number of processors used.

Weak scaling, on the other hand, measures the improvement in time to solution when increasing the number of parallel processors applied to a problem of fixed size per processor. In other words, the problem size is increased along with the number of processors. Here the execution time tends to remain fairly constant as the problem size (and the number of processors) increases. Perfect weak scaling, then, implies that the time to solution did not increase by scaling up the problem linearly with the number of processors.

As individual processors and clusters of processors get ever wider (having ever more parallel processing elements), the benefit of weak scaling for some problems diminishes—eventually these problems may run out of parallelism. It is at this point that these problems are forced into the realm of strong scaling. But in reality, while most applications do exhibit some degree of strong scaling, it is usually not perfectly linear.

A key reason for this is the cost of communication. Strong-scaling a problem onto an increasing number of processors gives each processor progressively less work to do, and increases the relative cost of communicating among those processors. In the strong-scaling regime, fast interconnects and communication primitives tuned for those interconnects are essential.

To provide the highest possible computational density, DGX Station™ includes four NVIDIA Tesla V100 accelerators. Application scaling on this many highly parallel GPUs is hampered by today's PCIe interconnect. NVLink (see Figure 6) provides the communications performance needed to achieve good (weak and strong) scaling on deep learning and other applications. Each Tesla V100 GPU in DGX Station has four NVLink connection points, each providing a point-to-point connection to another GPU at a peak bandwidth of 25GB/s. Multiple NVLink connections can be bonded together, multiplying the available interconnection bandwidth between a given pair of GPUs. The result is that NVLink provides a flexible interconnect that can be used to build a variety of network topologies among multiple GPUs. V100 also supports 16 lanes of PCIe 3.0. In DGX Station™, these are used for connecting between the CPUs and GPUs. PCIe is also used for accessing the high-speed networking interface cards.

[2] Note that "weak" is not an inferior form of scaling to "strong" scaling. Both are important metrics in practice.

The design of the NVLink network topology for DGX Station™ aims to optimize a number of factors, including the bandwidth achievable for a variety of point-to-point and collective communications primitives, the flexibility of the topology, and its performance with a subset of the GPUs. During the design, NVIDIA engineers modeled projected strong and weak scaling of a variety of applications, such as deep learning, sorting, Fast Fourier Transforms (FFT), molecular dynamics, graph analytics, computational fluid dynamics, seismic imaging, ray tracing, and others. This paper focuses on the analysis of scaling of deep learning training.

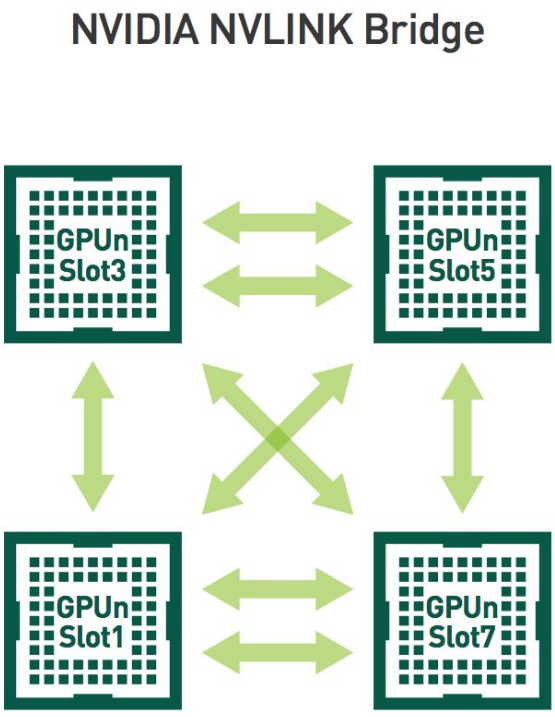

Figure 6 NVIDIA NVLink Bridge

3.2 Scaling Deep Learning Training on NVLink

Deep neural networks learn many layers of abstraction, in a hierarchy of simple to complex concepts. The strength of deep models is the ability to learn complex distributed representations from massive amounts of training data. A deep neural network is trained by feeding it input and letting it compute layer-by-layer to generate output for comparison with a known correct answer. After computing the error at the output, this error flows backward through the net by back-propagation. At each step backward the model parameters are tuned in a direction that

tries to reduce the error using one of many numerical optimization methods, such as stochastic gradient descent (SGD). This process sweeps over the data, improving the model as it goes.

Training deep neural networks in parallel across multiple GPUs and/or multiple nodes requires distributing either:

- the input data ("data parallel"): In data-parallel approaches, separate workers must periodically resynchronize the gradients with respect to the model that are calculated during back-propagation such that the model parameters are kept in sync across workers. This amounts to an all-reduce operation.
- the model being trained ("model parallel"): Model-parallel approaches may either elect one worker at a time to broadcast its gradients with respect to the input data, or they may use an all-gather of (successive subsets of) the data gradients so that all workers' outbound bandwidths are utilized concurrently.
- a hybrid of the two [Wu et al. 2015][Krizhevsky 2014].

In weak-scaling the training of a deep neural network, the global effective SGD minibatch size increases as the number of GPUs increases. Perhaps unsurprisingly, weak-scaled approaches have high parallel efficiency, even with relatively slow interconnections among GPUs. However, the minibatch size can only be scaled to a certain point before the convergence of the SGD optimization is negatively impacted. The relative tolerance of various networks to increased amounts of weak scaling varies with the network.

To demonstrate scaling of training performance on DGX Station from 1, to 2, to 4 Tesla V100, the bars in Figure 7 represent training performance in images per second for the ResNet-50 deep neural network architecture using the 18.11py3 (Python 3) DGX optimized containers for each of the three framework displayed (TensorFlow, MXNet, and PyTorch). These benchmark numbers were achieved using mixed precision with Tensor Cores available in V100, and show the linear scalability of V100s connected via NVLink.

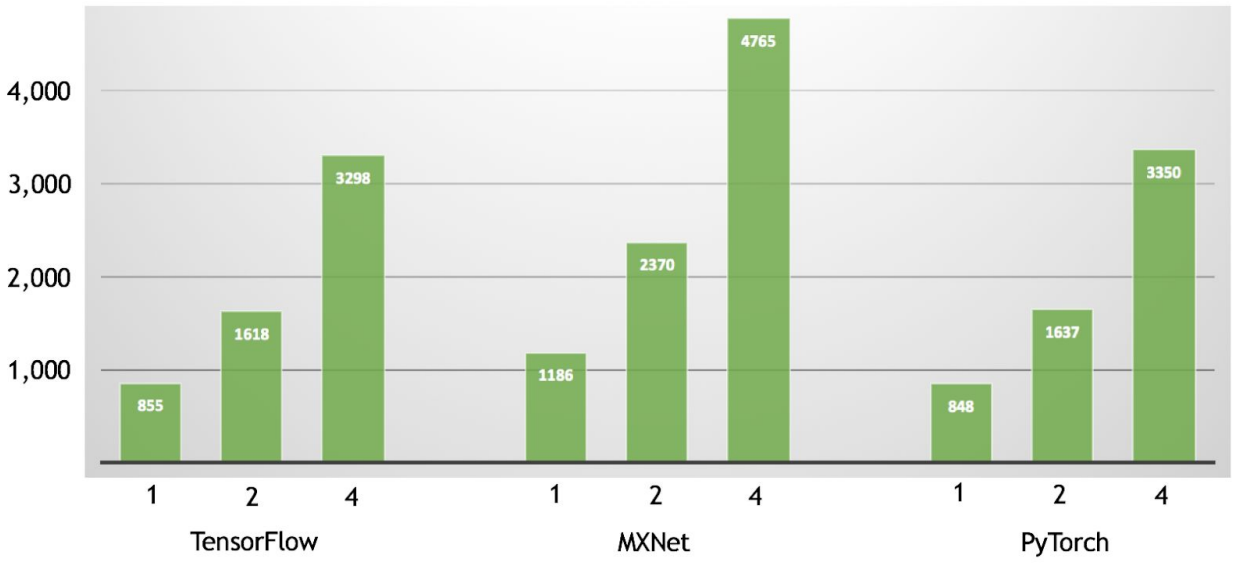

Figure 7: DGX Station scalability with 4x Tesla V100. ResNet-50 Training, Volta with Tensor Core (mixed precision), 18.11py3 DGX optimized containers from NGC. Score: Images per second.

In addition, the similar scores across the three different deep learning frameworks highlight the work the NVIDIA engineering team is doing for all major frameworks.

4.0 DGX Station Software Stack for AI

The DGX Station™ software stack for AI has been built to run deep learning and data science workloads at scale.
Note: DGX Station can also be used for High Performance Computing (HPC) and accelerated analytics workloads. See Section 5 for details.

A key goal is to enable practitioners to deploy deep learning frameworks, data science algorithms, and other applications on DGX Station™ with minimal setup effort. The design of the platform software is centered around an operating system based on Ubuntu Desktop with appropriate developer software and drivers installed on the workstation, and provisioning of all application software and additional SDK software with the NVIDIA Container Runtime for Docker for NVIDIA GPUs.

Deep learning containers optimized to take full advantage of DGX Station are available from the NVIDIA GPU Cloud (NGC) container registry. NGC features ready-to-run GPU-accelerated containers for the top deep learning software to provide multi-GPU performance on DGX Station. NGC containers are portable and can be used across DGX Station and other supported NGC platforms such as NVIDIA GPU-enabled instances on cloud service providers, NVIDIA DGX-1 and DGX-2, PCs with select NVIDIA TITAN and Quadro GPUs, and NGC-Ready systems.

Optimized containers available from NGC for DGX Station™ include all of the top deep learning frameworks, RAPIDS for accelerated data science and machine learning, NVIDIA TensorRT inference accelerator and NVIDIA TensorRT Inference Server, NVIDIA DIGITS deep learning training application, NVIDIA CUDA Toolkit, third-party managed HPC application containers, NVIDIA HPC visualization containers, and partner applications. Figure 8 shows the DGX Station™ software stack.

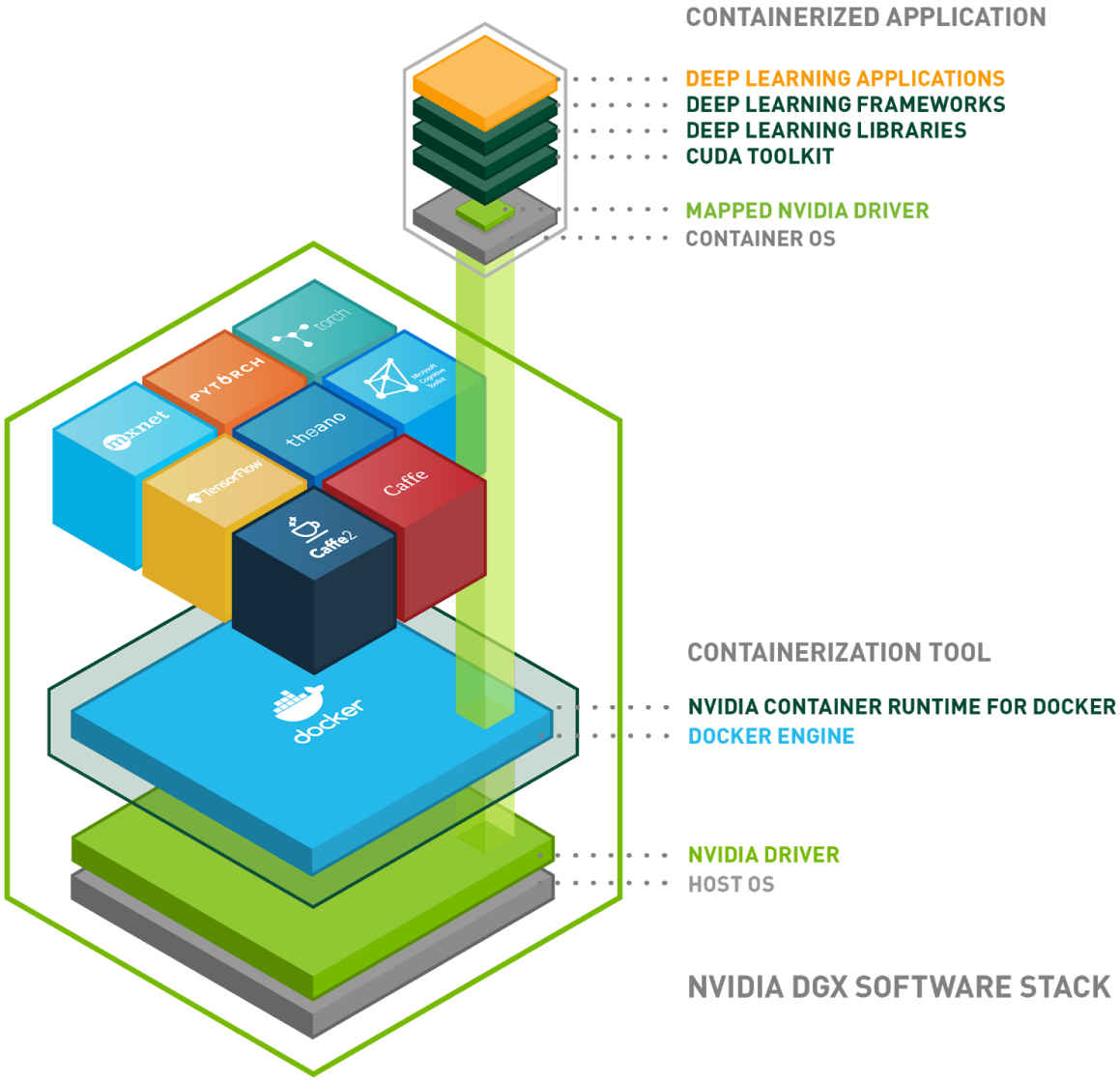

Figure 8 DGX Station™ software stack

This software architecture has many advantages:

- Since each framework, application, or library is in a separate container, each can use different versions of supporting software such as libc, cuDNN, and others, and not interfere with each other.
- This also means that different software with different workloads and test procedures can be run simultaneously on DGX Station by assigning different GPUs to each container.
- As software in the containers is improved for performance or bug fixes, new versions of the containers are made available in the NGC container registry.
- The system is easy to maintain, and the OS image stays clean, since applications are not installed directly on the OS.
- Security updates, driver updates and OS patches can be delivered seamlessly.

The remainder of this section covers the key components of the DGX Station software stack (above the GPU Compute Software Driver) in detail. Section 5 provides details of the ready-to-run software available for DGX Station from NVIDIA GPU Cloud (NGC)..

4.1 NVIDIA CUDA Toolkit

CUDA is a parallel computing platform and programming model created by NVIDIA to give application developers access to the massive parallel processing capability of GPUs. CUDA is the foundation for GPU acceleration of deep learning, data science, as well as a wide range of other computation- and memory-intensive applications ranging from astronomy, to molecular dynamics simulation, to computational finance. Today there are over 600 GPU-accelerated applications that leverage the CUDA parallel computing platform [NVIDIA 2017b]. DGX Station is not only the fastest platform for deep learning and data science in a workstation form factor, but the most advanced CUDA platform for a wide variety of GPU-accelerated applications that your team can run in your office.

The NVIDIA CUDA Toolkit provides a comprehensive environment for C and C++ developers building GPU-accelerated applications. The CUDA Toolkit includes NVCC, the CUDA C++ compiler for NVIDIA GPUs, a suite of libraries of GPU-accelerated algorithms, debugging and profiling tools, examples, and comprehensive programming guides and documentation. While the CUDA Toolkit comes directly installed on DGX Station as part of the Ubuntu-based operating system, it is also provided as an NVIDIA Docker container image on NVIDIA GPU Cloud which can be used as the base layer for any containerized CUDA application (as Figure 8 shows). In addition, the full CUDA Toolkit is embedded in every deep learning framework container image.

4.2 NVIDIA Libraries

NVIDIA provides a complete suite of GPU-accelerated libraries built on top of the CUDA parallel computing platform. The following libraries provide GPU-accelerated primitives for deep neural networks and data science:

- **CUDA Basic Linear Algebra Subroutines library (cuBLAS)**: cuBLAS is a GPU-accelerated version of the complete standard BLAS library that delivers significant speedup running on GPUs. The cuBLAS generalized matrix-matrix multiplication (GEMM) routine is a key computation used in deep neural networks, for example in computing fully connected layers.
- **CUDA Deep Neural Network library (cuDNN)**: cuDNN is a GPU-accelerated library of primitives for deep neural networks. cuDNN provides highly tuned implementations of standard routines such as forward and backward convolution, pooling, normalization, and activation layers.
- **DataFrame (cuDF):** This is a GPU accelerated DataFrame-manipulation library based on GPU Apache Arrow. It's designed to enable data wrangling data for model training. The Python bindings of the core-accelerated, low-level CUDA C++ kernels mirror the pandas API for seamless onboarding and transition from pandas.
- **Machine Learning Libraries (cuML)**: This collection of GPU-accelerated machine learning libraries will eventually provide GPU versions of all machine learning algorithms available in Scikit-Learn.

When deployed using the NGC containers for DGX Station, deep learning frameworks and data science libraries are automatically configured to use parallel routines optimized for the Tesla V100 architecture in DGX Station.

The NVIDIA Collective Communication Library (NCCL, pronounced "Nickel") is a library of multi-GPU MPI-compatible collective communication primitives that are topology-aware and can be easily integrated into applications. NCCL is designed to be light-weight, depending only on common C++ and CUDA libraries. NCCL can be deployed in single-process or multi-process applications, handling required inter-process communication transparently. The NCCL API is designed to be familiar to anyone with experience using MPI collectives such as broadcast, reduce, gather, scatter, all-gather, all-reduce, or all-to-all that are optimized to achieve high bandwidth over PCIe and NVLink high-speed interconnect.

NGC containers for DGX Station include a version of NCCL that optimizes these collectives for the DGX Station architecture's four-GPU second generation NVLink. When deployed using these containers, deep learning frameworks such as Caffe2, PyTorch, Microsoft Cognitive Toolkit, and TensorFlow automatically use this version of NCCL when run on multiple GPUs.

4.3 NVIDIA Container Runtime for Docker

Over the last few years there has been a dramatic rise in the use of software containers for simplifying deployment of data center applications at scale. Containers encapsulate an application's dependencies to provide reproducible and reliable execution of applications and services without the overhead of a full virtual machine.

A Docker container is a mechanism for bundling a Linux application with all of its libraries, configuration files, and environment variables so that the execution environment is always the same, on whatever Linux system it runs and between instances on the same host (see Figure 8). Docker containers are user-mode only, so all kernel calls from the container are handled by the host system kernel. DGX Station uses Docker containers from NVIDIA GPU Cloud as the mechanism for deploying deep learning frameworks and other application software.

Docker containers are platform- and hardware-agnostic, and achieve this with separation of user mode code (in the container) from kernel mode code (accessed on the host). This separation presents a problem when using specialized hardware such as NVIDIA GPUs, since GPU drivers consist of a matched set of user mode and kernel mode modules. An early work-around to this problem was to fully install the NVIDIA drivers inside the container and map in the character devices corresponding to the NVIDIA GPUs (for example, /dev/nvidia0) on launch. This solution is brittle because the version of the host driver must exactly match the version of the driver installed in the container. This requirement drastically reduced the portability of these early containers, undermining one of Docker's more important features.

To enable portability in Docker images that leverage GPUs, NVIDIA developed NVIDIA Container Runtime for Docker [NVIDIA Corporation 2015], an open-source project that provides the nvidia-docker command-line tool to mount the user-mode components of the NVIDIA driver and the GPUs into the Docker container at launch, as Figure 8 shows. For this to work, it is essential that the developer does not install an NVIDIA driver into the Docker image at docker build time.

The nvidia-docker tool is essentially a wrapper around Docker that transparently provisions a container with the necessary components to execute code on the GPU.

4.4 NVIDIA GPU Cloud Container Access

NVIDIA provides a library of pre-integrated Docker containers for use with DGX Station and other compatible systems on NVIDIA GPU Cloud (NGC). This library of GPU-optimized software includes NVIDIA tuned, tested, certified, and maintained containers for the top deep learning software, and GPU-accelerated containers for third-party managed HPC applications, NVIDIA

HPC visualization tools, and partner applications. This eliminates the complexity typically associated with deploying these applications, making it simple to get up and running quickly.

At a high level, there are two steps to use these containers on DGX Station. First, you will use a web browser and access NGC GPU Cloud. to browse the container registry. As a DGX Station user, your organization was registered with NGC when your DGX Station was delivered. Please check with your system administrator if you need access details, though you can browse the containers and use many of them even without registering. The NGC user interface provides detailed instructions and the exact command needed to pull each container on to your system.

Once you have accessed NGC and chosen which container to pull, simply open a shell on your DGX Station, login to the NGC container registry with the provided credentials, and enter the appropriate docker pull command (this command is provided on the container's page in the NGC browser user interface, so you can easily copy and paste it). The container will download to your DGX Station. After the download is complete, the container is ready to run.

For more details, please see the NGC Container User Guide at:
http://docs.nvidia.com/ngc/ngc-user-guide/

5.0 Ready-to-Run Software and Tools for DGX Station

DGX Station includes a complete library of ready-to-run software for AI, machine learning, and HPC from NVIDIA GPU Cloud (NGC). The following sections give details on the software available from NGC.

5.1 Deep Learning Frameworks

The NVIDIA Deep Learning SDK accelerates the most popular deep learning frameworks such as DIGITS, MXNet, NVCaffe, TensorFlow, and PyTorch. The following sections describe the deep learning frameworks and tools NVIDIA has optimized for DGX systems.

The DGX Station software stack provides containerized versions of these frameworks optimized for the system. These frameworks, including all necessary dependencies, are pre-built, tested, and ready-to-run. For users who need more flexibility to build custom deep learning solutions, each framework container image also includes the framework source code to enable custom modifications and enhancements, along with the complete software development stack described in Section 4.

Most deep learning frameworks have begun to merge support for half-precision training techniques that exploit Tensor Core calculations in Volta. Some frameworks include support for FP16 storage and Tensor Core math. To achieve optimum performance, you can train a model

using Tensor Core math and FP16 mode on some frameworks. For information about which frameworks are optimized for Volta, see http://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html .

Containers are integrated with the following NVIDIA Deep Learning SDK components:

- cuDNN.
- CUDA.
- cuBLAS.
- NCCL with NVLink support for improved multi-GPU scaling.

The containers are tuned for single and limited multi-GPU execution.

Each of the containers for your DGX Station have detailed release notes available in NGC that provide the versions used for each of these components. You can read about the latest updates to the top deep learning containers in the Deep Learning DGX Documentation Release Notes and view a matrix of component versions in the deep learning containers in the NVIDIA Optimized Frameworks Support Matrix.

5.2 RAPIDS

The DGX Station software stack supports RAPIDS, which is available as a container on NGC. RAPIDS is a set of open-source libraries for GPU-accelerated data science and data analytics. These libraries accelerate end-to-end workflows from data preparation (cuDF), model training (cuML), to visualization. It's based on Python, has pandas-like and scikit-learn-like interfaces, built on Apache Arrow in-memory data format.

5.3 HPC

A large number of HPC and visualization containers are available from NGC. The HPC containers make application deployment much faster and easier, providing access to the latest features and optimizing performance. HPC and visualization containers include NAMD, LAMMPS, GROMACS, RELION, GAMESS, ParaView, NVIDIA IndeX™ volume renderer, NVIDIA OptiX™ ray tracing tools, NVIDIA Engine Bridge artistic renderer, and many more. For a complete list, view the HPC and visualization categories on NGC.

5.4 Partner Containers

NGC also includes a wide variety of GPU-accelerated software from NVIDIA partners, with new content being added frequently. These containers are tested and optimized, and are ready to

run on DGX Station. Browse through the categories on NGC to see the latest partner container offerings.

6.0 Sharing a DGX Station

NVIDIA DGX Station was designed to be one's personal AI supercomputer under the desk. It provides the best computational performance in such a form factor, and connections to drive one or more displays, keyboard and mouse, so anyone can work directly on it.

The DGX Station, with its integrated software stack and four of the world's most advanced data center GPUs in one system, can be shared among several data scientists or deep learning practitioners. There are many situations where a user, or groups of users, would want to run and manage multiple jobs at the same time on the system. Example use cases include:

- Multiple users sharing the same system
- Multiple simultaneous projects
- Multiple jobs with interdependencies
- Maximizing GPU utilization via queueing or scheduling workloads

IDC has explored the benefits and use cases of the DGX Station in "Accelerated Workstation: Run Deep Learning Workloads at Your Desk". They offer the following guidance:

> "An additional use case for DGX Station is its role as a workgroup server. While some organizations will have individual developers who can sufficiently consume the entire compute capacity available in a dedicated AI workstation, some environments will have multiple team members, each running their own experiments in a workgroup setting either simultaneously or at varying schedules. DGX Station can provide a more-than-adequate solution that can multiplex its GPU computing power to several users, thereby improving the utilization and overall economic benefit of the platform."

7.0 Results: DGX Station for Highest Deep Learning Performance at Your Desk

7.1 Volta Architecture Performance

As general computing performance increases are flattening (“The End of Moore’s Law”, [NVIDIA Blog]), NVIDIA is able to dramatically increase performance of its GPUs with each generation. The delivery of the Volta architecture, only within a year of the launch of the previous Pascal architecture, was especially impressive. Although impressive performance gains could be achieved for different kind of computations, such as single or double precision calculations often needed for HPC workloads, the biggest performance gains can be seen in mixed precision (half and single precision) operations thanks to the Tensor Core architecture.

The bars in Figure 9 represent training performance in images per second for the ResNet-50 deep neural network architecture using the 18.11py3 (Python 3) DGX optimized containers from NGC for each of the three framework displayed (TensorFlow, MXNet, and PyTorch). It compares the performance of using 4x P100 PCIe GPUs with 4x V100 GPUs available in DGX Station.

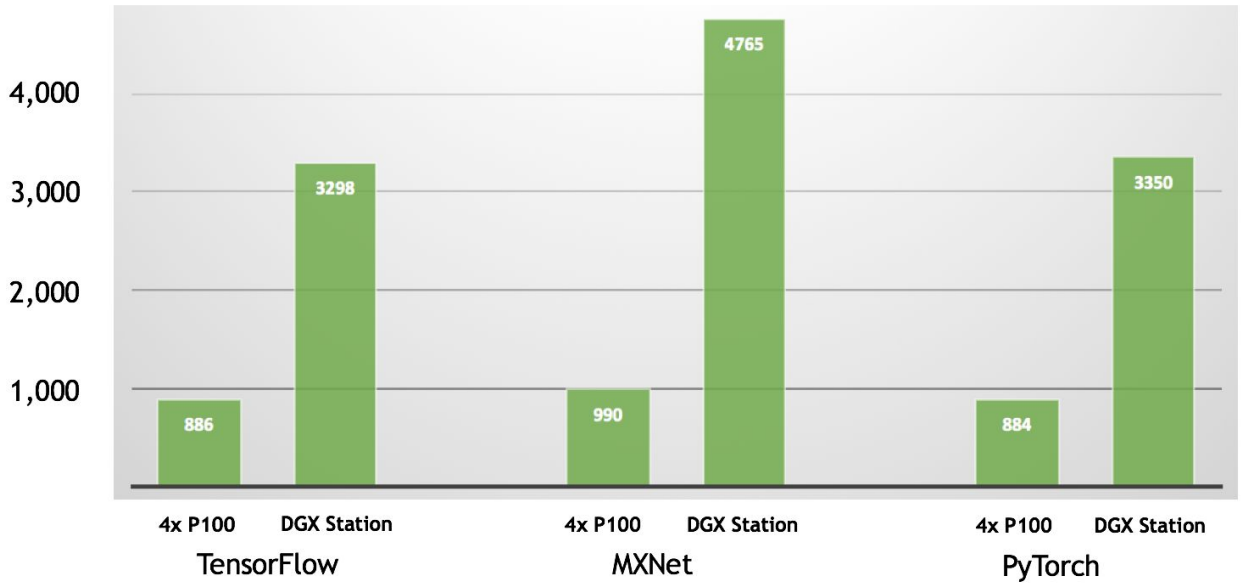

Figure 9: DGX Station with 4x Tesla V100. ResNet-50 Training, Volta with Tensor Core (mixed precision), 18.11py3 DGX optimized container. 4x Tesla P100 PCIe. ResNet-50 Training, FP32, 17.09 DGX Optimized container. Score: Images per second.

7.2 Scalability

As already shown in Section 3.2, it can be demonstrated nicely that scaling of training performance on DGX Station from 1, to 2, to 4 Tesla V100 GPUs is close to linear.
The bars in Figure 10 represent training performance in images per second for the ResNet-50 deep neural network architecture using the 18.11py3 (Python 3) DGX optimized containers for each of the three framework displayed (TensorFlow, MXNet, and PyTorch). These benchmark numbers were achieved using mixed precision with Tensor Cores available in V100, and show the linear scalability of V100s connected via NVLink.

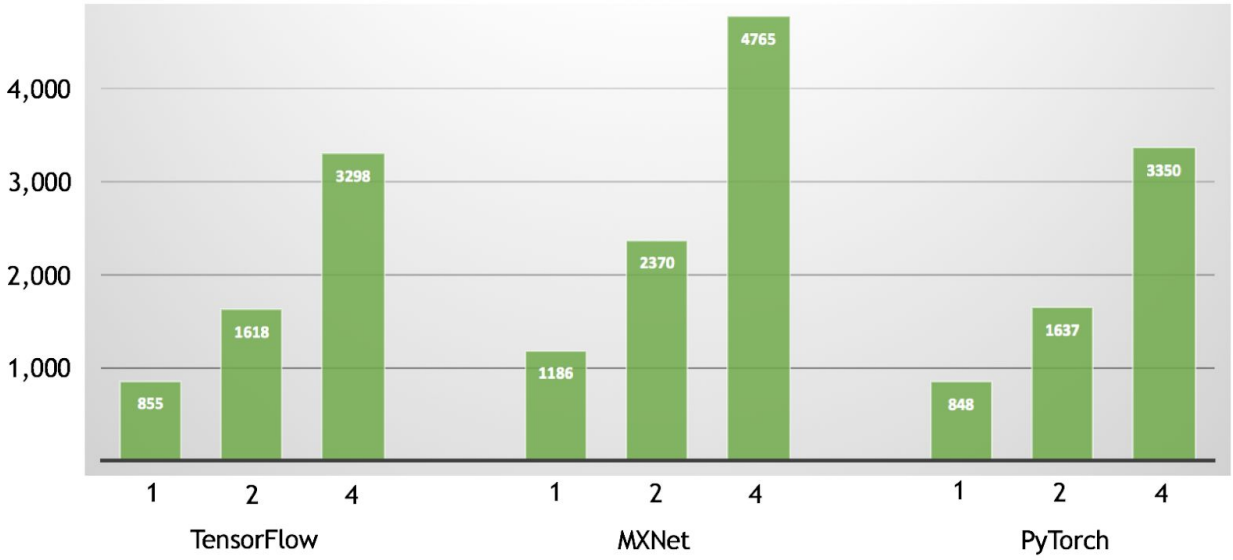

Figure 10: DGX Station with 4x Tesla V100. ResNet-50 Training, Volta with Tensor Core (mixed precision), 18.11py3 DGX optimized containers from NGC. Score: Images per second

7.3 Continuous Optimizations

Years of technology interlock between NVIDIA Engineering and the developers of the leading deep learning frameworks has resulted in steady, incremental progress, advancing the art and science of deep learning performance.
Figure 11 highlights the type of improvements that were achieved in the course of only one year.

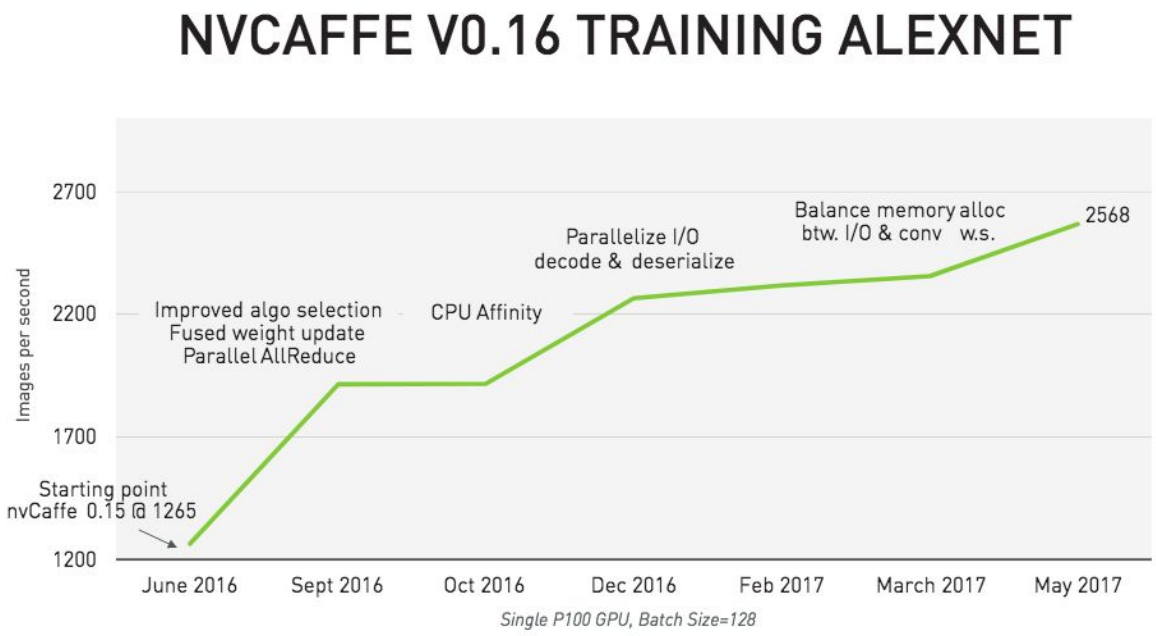

Figure 11: Performance improvements of NVCaffee, AlexNet, from 2016 to 2017

With each new framework release comes enhancements to DGX software, that have been planned early on in the development cycle, ahead of the actual framework release to the general public. These improvements typically fall into several pillars of targeted development that NVIDIA engineers have perfected over time, performance gains being one of the most important goals.

The bars in Figure 12 represent training performance in images per second for the ResNet-50 deep neural network architecture comparing the 17.09 DGX optimized containers for each of the three framework displayed (MXNet, TensorFlow, PyTorch) with the 18.03py3 and 18.11py3 versions. As you can see, in only thirteen months, NVIDIA engineers managed to achieve performance gains between 31% and up to 49%, purley by optimizing the software.

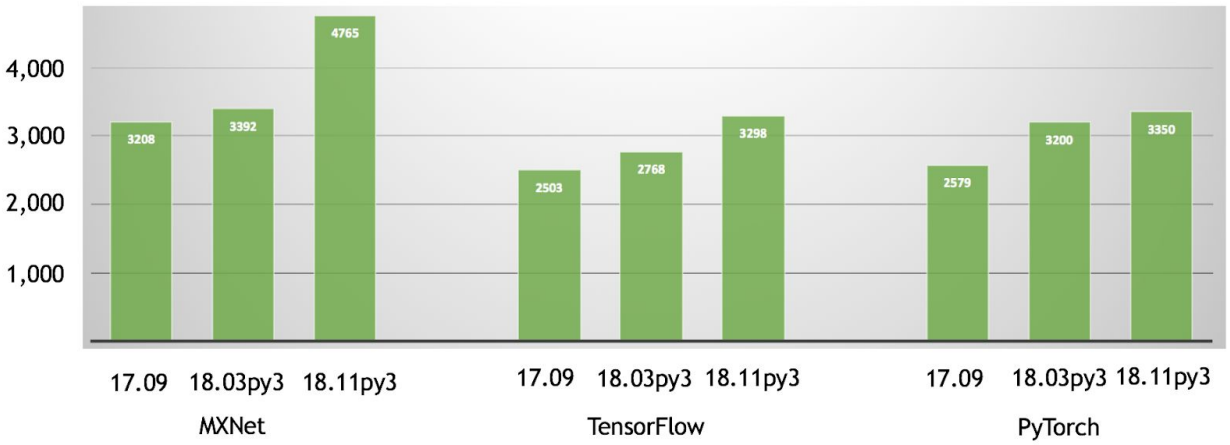

Figure 12: DGX Station with 4x Tesla V100. ResNet-50 Training, Volta with Tensor Core (mixed precision), 17.09, 18.03py3, and 18.11py3 DGX optimized containers from NGC. Score: Images per second.

Notice

ALL INFORMATION PROVIDED IN THIS WHITE PAPER, INCLUDING COMMENTARY, OPINION, 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 MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification 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. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

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 these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do 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 specification. NVIDIA does not accept any 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 specification, 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 specification. 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 specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

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. 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 NVIDIA terms and conditions of sale for the product

Trademarks

NVIDIA, the NVIDIA logo, CUDA, Pascal, Tesla, NVLink, and DGX-1 are trademarks or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright