Technical White Paper

Dell EMC PowerScale and NVIDIA DGX A100 Systems for Deep Learning

Abstract

This document demonstrates how the Dell EMC Isilon F800 all-flash scale-out NAS and NVIDIA DGX™ A100 systems with NVIDIA® A100 Tensor Core GPUs can be used to accelerate and scale deep learning training workloads

November 2020

H18597

Revisions

Date	Description
November 2020	Initial release

Acknowledgments

Author: Damien Mas

Other: NVIDIA

Table of contents

Revisions....2
Acknowledgements....2
Table of contents....3
Executive summary....4
Audience....4
1 Introduction....5
1.1 Deep learning dataflow....5
2 Solution architecture....7
2.1 Overview....7
2.2 Storage – Dell EMC Isilon F800....8
2.3 Networking....9
2.3.1 Dell EMC PowerSwitch data center switches....9
2.3.2 NVIDIA Mellanox SN3700V Ethernet switch for Storage....10
2.3.3 NVIDIA Mellanox QM8700 InfiniBand switch for GPU Interconnect....10
2.4 Compute: NVIDIA DGX A100 system....10
2.5 NVIDIA NGC....11
2.6 Bill of materials....11
3 Deep learning training performance and analysis....13
3.1 Benchmark methodology....13
3.2 MLPerf Benchmark results....13
3.3 NVIDIA collective communication library (NCCL)....14
3.4 Storage-only performance using FIO....15
4 Solution sizing guidance....17
5 Conclusion....18
6 References....19

Executive summary

Deep learning (DL) techniques have enabled great successes in many fields such as computer vision, natural language processing (NLP), gaming and autonomous driving by enabling a model to learn from existing data and then to make corresponding predictions. The success is due to a combination of improved algorithms, access to larger datasets and increased computational power. To be effective at enterprise scale, the computational intensity of DL requires highly efficient parallel architectures. The choice and design of the system components, carefully selected and tuned for DL use-cases, can have a big impact on the speed, accuracy and business value of implementing artificial intelligence (AI) techniques.

In such a demanding environment, it is critical that organizations be able to rely on vendors that they trust. Over the last few years, Dell Technologies and NVIDIA have established a strong partnership to help organizations fast-track their AI initiatives. Our partnership is built on the philosophy of offering flexibility and informed choice across a broad portfolio which combines best of breed GPU accelerated compute, scale-out storage, and networking.

This paper focuses on how Dell EMC Isilon F800 all-flash scale-out NAS accelerates AI innovation by delivering the performance, scalability and I/O concurrency to complement the requirements of NVIDIA DGX A100 systems for high-performance AI workloads.

Audience

This document is intended for organizations interested in simplifying and accelerating DL solutions with advanced computing and scale-out data management solutions. Solution architects, system administrators and other interested readers within those organizations constitute the target audience.

1 Introduction

DL is an area of AI which uses artificial neural networks to enable accurate pattern recognition of complex real-world patterns by computers. These new levels of innovation have applicability across nearly every industry vertical. Some of the early adopters include advanced research, precision medicine, high tech manufacturing, advanced driver assistance systems (ADAS) and autonomous driving. Building on these initial successes, AI initiatives are springing up in various business units, such as manufacturing, customer support, life sciences, marketing, and sales. Gartner predicts that AI augmentation will generate $2.9 trillion in business value by 2021 alone. Organizations are faced with a multitude of complex choices related to data, analytic skillsets, software stacks, analytic toolkits, and infrastructure components; each with significant implications on the time to market and the value associated with these initiatives.

In such a complex environment, it is critical that organizations be able to rely on vendors that they trust. Over the last few years, Dell Technologies and NVIDIA have established a strong partnership to help organizations accelerate their AI initiatives. Our partnership is built on the philosophy of offering flexibility and informed choice across an extensive portfolio. Together our technologies provide the foundation for successful AI solutions which drive the development of advanced DL software frameworks, deliver massively parallel compute in the form of NVIDIA GPUs for parallel model training and scale-out file systems to support the concurrency, performance, and capacity requirements of unstructured image and video data sets.

This document focuses on the latest step in the Dell Technologies and NVIDIA collaboration, a new AI reference architecture with Dell EMC Isilon F800 storage and DGX A100 systems for DL workloads. This new offer gives customers more flexibility in how they deploy scalable, high performance DL infrastructure. The results of standard image classification training benchmark using MLPerf 0.7 and micro-benchmark utilities, are included.

1.1 Deep learning dataflow

As visualized in Figure 1, DL usually consists of two distinct workflows, model development and inference.

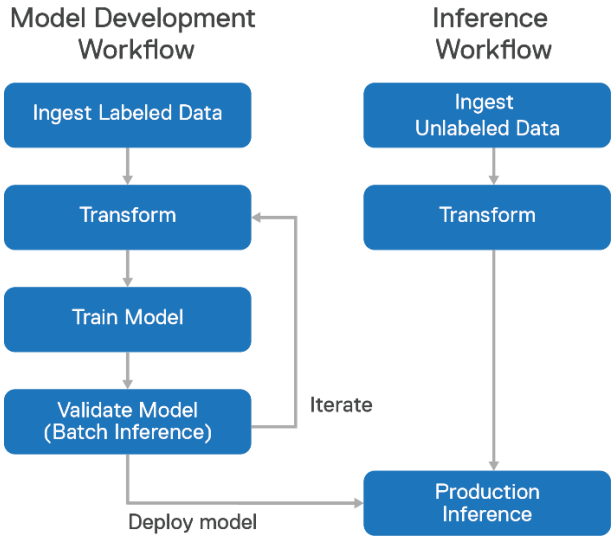

Figure 1 Common DL Workflows - Model development and inference

Note: The Isilon storage and DGX A100 system architecture is optimized for the model development workflow which consists of the model training and the batch inference validation steps. It is not intended for and nor was it benchmarked for production inference.

The workflow steps are defined and detailed below.

1. **Ingest Labeled Data** — The labeled data (e.g. images and their labels which indicate whether the image contains a dog, cat, or horse) are ingested into the DL system.
2. **Transform** — Transformation includes all operations that are applied to the labeled data before they are passed to the DL algorithm. It is sometimes referred to as preprocessing. For images, this often includes file parsing, JPEG decoding, cropping, resizing, rotation, and color adjustments. Transformations can be performed on the entire dataset ahead of time, storing the transformed data on disk. Many transformations can also be applied in a training pipeline, avoiding the need to store the intermediate data.
3. **Train Model** — The model parameters (edge weights) are learned from the labeled data using the stochastic gradient descent optimization method. In the case of image classification, there are several prebuilt structures of neural networks that have been shown to work well.
4. **Validate Model** — Once the model training phase completes with a satisfactory accuracy, you'll want to measure the accuracy of it on validation data – data that the model training process has not seen. This is done by using the trained model to make inferences from the validation data and comparing the result with the correct label. This is often referred to as inference but keep in mind that this is a distinct step from production inference.
5. **Production Inference** — The trained and validated model is then often deployed to a system that can perform real-time inference. It will accept as input a single image and output the predicted class (dog, cat, horse). In some cases, inputs are batched for higher throughput but higher latency.

2 Solution architecture

2.1 Overview

Figure 2 illustrates the reference architecture showing the key components that made up the solution as it was tested and benchmarked. Note that in a customer deployment, the number of DGX A100 systems and F800 storage nodes will vary and can be scaled independently to meet the requirements of the specific DL workloads. Refer to Solution sizing guidance for details.

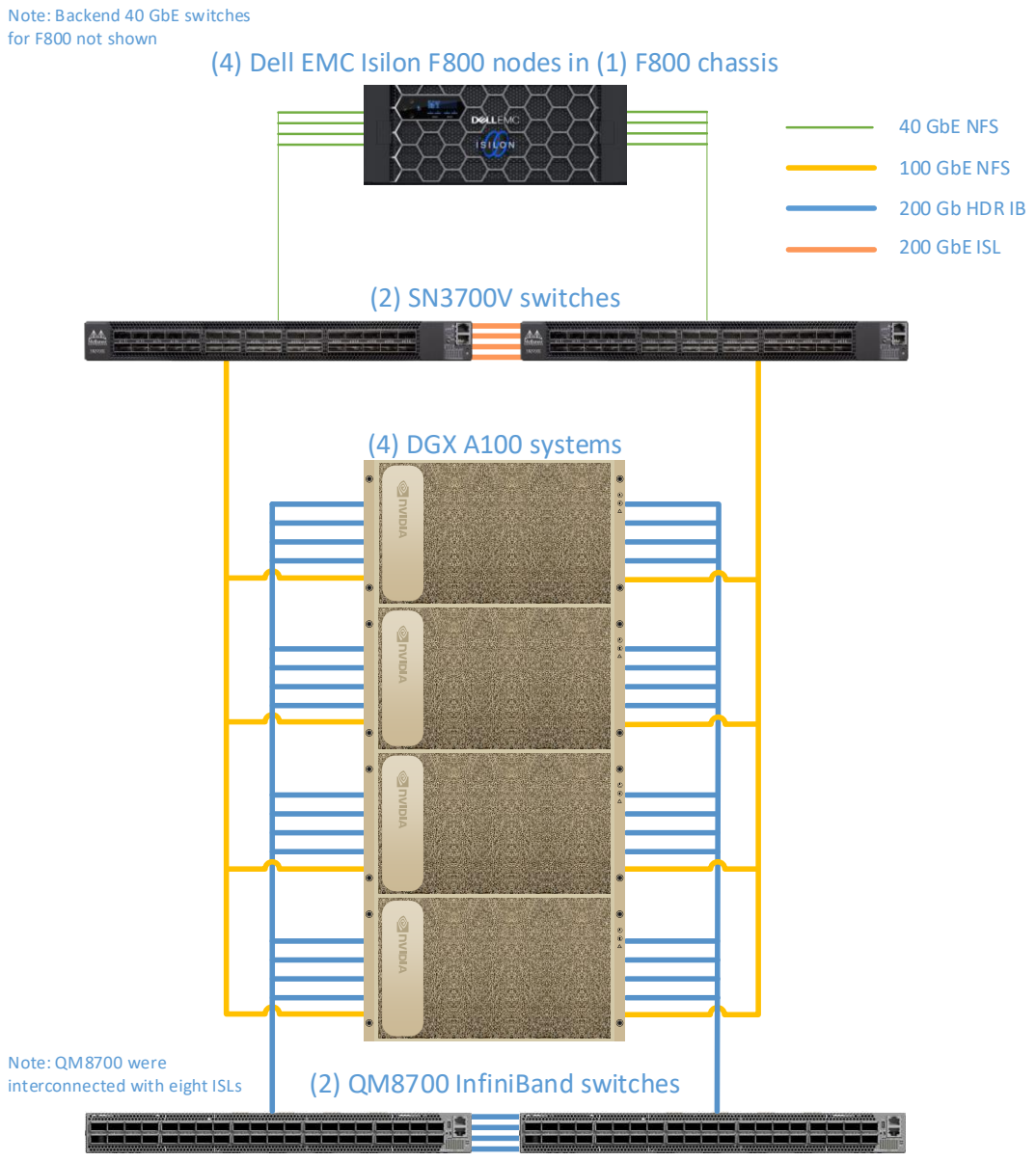

Figure 2 Reference Architecture

2.2 Storage – Dell EMC Isilon F800

Dell EMC Isilon F800 represents the sixth generation of hardware built to run the well-proven and massively scalable Dell EMC PowerScale OneFS operating system. Each Dell EMC Isilon F800 chassis, shown in Figure 3, contains four storage nodes, 60 high-performance solid state drives (SSDs) and eight 40 GbE network connections. OneFS combines up to 252 nodes in 63 chassis into a single high-performance file system designed to handle the most I/O-intense workloads such as DL. As performance and capacity demands increase, the platform can be scaled-out simply and non-disruptively, allowing applications and users to continue working.

Figure 3 Dell EMC Isilon F800 chassis, containing four storage nodes

In the solution tested in this document, four F800 nodes, in one chassis, were used.

Dell EMC Isilon F800 has the following features.

- Low latency, high throughput, and massively parallel I/O for AI
  - Up to 250,000 file IOPS per chassis, up to 15.75 million IOPS per cluster
  - Up to 15 GB/s throughput per chassis, up to 945 GB/s per cluster
  - 96 TB to 924 TB raw flash capacity per chassis; up to 58 PB per cluster (all-flash)

  This shortens time for training and testing analytical models for data sets from tens of TBs to tens of PBs on AI platforms such as RAPIDS, TensorFlow, SparkML, Caffe, or proprietary AI platforms.

- The ability to run AI in-place on data using multi-protocol access
  - Multi-protocol support such as SMB, NFS, HTTP, S3, and native HDFS to maximize operational flexibility

  This eliminates the need to migrate/copy data and results over to a separate AI stack. Organizations can perform DL and run other IT apps on the same data already on Isilon by adding additional Isilon nodes to an existing cluster.

- Enterprise grade features out-of-box
  - Enterprise data protection and resiliency
  - Robust security options

  This enables organizations to manage AI data lifecycle with minimal cost and risk, while protecting data and meeting regulatory requirements.

- Extreme scale

- Seamlessly tier between All Flash, Hybrid, and Archive nodes via Dell EMC PowerScale SmartPools
- Grow-as-you-go scalability with up to 58 PB flash capacity per cluster
- New nodes can be added to a cluster simply by connecting power, back-end Ethernet and front-end Ethernet
- As new nodes are added, storage capacity, throughput, IOPS, cache, and CPU grow
- Up to 63 chassis (252 nodes) may be connected to form a single cluster with a single namespace and a single coherent cache
- Up to 85% storage efficiency to reduce costs with Dell EMC PowerScale SmartDedupe software
- Optional data de-dup and compression enabling up to a 3:1 data reduction

Organizations can achieve AI at scale in a cost-effective manner, enabling them to handle multi-petabyte datasets with high resolution content without re-architecture and/or performance degradation.

There are several key features of OneFS that make it an excellent storage system for DL workloads that require performance, concurrency, and scale. These features are listed below.

- Storage Tiering using Dell EMC PowerScale SmartPools software enables multiple levels of performance, protection, and storage density to co-exist within the same file system and unlocks the ability to aggregate and consolidate a wide range of applications within a single extensible, ubiquitous storage resource pool. This helps provide granular performance optimization, workflow isolation, higher utilization, and independent scalability – all with a single point of management. For more details, see Storage Tiering with Dell EMC Isilon SmartPools.
- OneFS caching infrastructure design is predicated on aggregating the cache present on each node in a cluster into one globally accessible pool of memory. This allows all the memory cache in a node to be available to every node in the cluster . OneFS can take advantage of prefetching of data based on heuristics used by the Isilon SmartRead component. This greatly improves sequential-read performance across all protocols and means that reads come directly from RAM within milliseconds. For high-sequential cases, SmartRead can very aggressively prefetch ahead, allowing reads of individual files at very high data rates. For more details, see OneFS SmartFlash.
- OneFS has a fully distributed lock manager that coordinates locks on data across all nodes in a storage cluster. Efficient locking is critical to support the efficient parallel I/O profile demanded by many iterative DL workloads enabling concurrent file read access up into the millions. For more details, see the OneFS Technical Overview.

2.3 Networking

2.3.1 Dell EMC PowerSwitch data center switches

The benchmark testing in this brief was performed in NVIDIA's partner facility and the networking materials mentioned represent the equipment they used during the testing. Dell Technologies offers top-of-rack switches built for building high-capacity network fabrics, and core/aggregation switches designed for building optimized data center leaf/spine fabrics of virtually any size. Dell EMC PowerSwitch S- and Z-Series are tested and proven in Dell Technologies' performance labs, top ranked in industry tests (Tolly and IT Brand Pulse), and are currently deployed in customer data centers around the world.

Learn more about Dell EMC PowerSwitch S- and Z-Series

Dell EMC PowerSwitch Data Center Quick Reference Guide

2.3.2 NVIDIA Mellanox SN3700V Ethernet switch for Storage

The NVIDIA Mellanox SN3700V Ethernet switches provide the high speed “front-end” Ethernet connectivity between the Isilon F800 cluster nodes and NVIDIA DGX A100 systems. The F800 nodes connect with 25GbE or 40GbE connections, the DGX A100 systems connect with 100GbE or 200GbE connections, and the SN3700 switches automatically forward traffic across the different speed connections with minimal latency. Based on the NVIDIA Spectrum-2 switch ASIC and purpose built for the modern datacenter, the SN3000 switches combine high performance packet processing, rich datacenter features, cloud network scale and visibility. A flexible unified buffer to ensure fair and predictable performance across any combination of ports and speeds from 10Gb/s to 200Gb/s, and an Open Ethernet design supports multiple network OS choices including NVIDIA Cumulus Linux, NVIDIA Onyx, and SONiC.

Learn more about the NVIDIA Mellanox Spectrum SN3000 series switches.

2.3.3 NVIDIA Mellanox QM8700 InfiniBand switch for GPU Interconnect

The NVIDIA Mellanox QM8700 InfiniBand switches provide high-throughput, low-latency networking between the DGX A100 systems. Designed for both EDR 100Gb/s and HDR 200 Gb/s InfiniBand links, they minimize latency and maximize throughput for all GPU-to-GPU communication between systems. The QM8700 switches support Remote Direct Memory Access (RDMA) and in-network computing offloads for AI and data analytics to enable faster and more efficient data transfers. They support NVIDIA GPUDirect, Mellanox SHARP for network-based AI and analytics offloads (such as MPI AllReduce), and Mellanox SHIELD for maximum resiliency in a self-healing network.

Learn more about the NVIDIA Mellanox Quantum QM8700 InfiniBand switches,

2.4 Compute: NVIDIA DGX A100 system

The DGX A100 system (Figure 4) is a fully integrated, turnkey hardware and software system that is purpose-built for DL workflows. Each DGX A100 system is powered by eight NVIDIA A100 Tensor Core GPUs that are interconnected using NVIDIA NVSwitch® technology, which provides an ultra-high bandwidth low-latency fabric for inter-GPU communication. This topology is essential for multi-GPU training, eliminating the bottleneck that is associated with PCIe-based interconnects that cannot deliver linearity of performance as GPU count increases. The DGX A100 system is also equipped with eight single-port NVIDIA Mellanox ConnectX-6 VPI HDR InfiniBand adapters for clustering and two dual-port ConnectX-6 VPI Ethernet adapter for storage and networking, all capable of 200Gb/s.

Figure 4 NVIDIA DGX A100 system with eight NVIDIA A100 Tensor Core GPUs

2.5 NVIDIA NGC

The NVIDIA NGC™ container registry provides researchers, data scientists and developers with simple access to a comprehensive catalog of GPU-accelerated software for AI, DL, machine learning (ML) and HPC that take full advantage of NVIDIA DGX A100 systems. NGC provides containers for today's most popular AI frameworks such as RAPIDS, Caffe2, TensorFlow, PyTorch, MXNet and TensorRT, which are optimized for NVIDIA GPUs. The containers integrate the framework or application, necessary drivers, libraries and communications primitives and they are optimized across the stack by NVIDIA for maximum GPU-accelerated performance. NGC containers incorporate the NVIDIA CUDA® Toolkit, which provides the NVIDIA CUDA Basic Linear Algebra Subroutines Library (cuBLAS), the NVIDIA CUDA Deep Neural Network Library (cuDNN), and much more. The NGC containers also include the NVIDIA Collective Communications Library (NCCL) for multi-GPU and multi-node collective communication primitives, enabling topology awareness for DL training. NCCL enables communication between GPUs inside a single DGX A100 system and across multiple DGX A100 systems.

2.6 Bill of materials

Table 1 Bill of materials

Component	Purpose	Quantity
• Dell EMC Isilon F800 • 96 TB SSD • 1 TB RAM • Four 1 GbE, eight 40 GbE interfaces	Shared storage	1 chassis (4 nodes)
• NVIDIA Mellanox SN3700V 200Gb Ethernet Switch	Storage Fabric Switch	2
• NVIDIA Mellanox QM8700 InfiniBand HDR Switch	Compute Fabric Switch	2
• NVIDIA DGX A100 system • 8 NVIDIA A100 Tensor Core GPUs with 40GB	Compute Server	4

Component	Purpose	Quantity
• Two 64-Core AMD EPYC 7742 @3.3 GHz • 1TB RAM • 2x Dual-Port NVIDIA Mellanox ConnectX-6 VPI 200 Gb/s Ethernet • 8x Single-Port NVIDIA Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand		

- SOFTWARE VERSIONS

Table 2 Software versions that were tested for this document

Component	Version
• Dell EMC Isilon – OneFS	• 8.2.1.0 • Patches: 8.2.1_KGA-RUP_2020-04_268538, 8.2.1_UGA-PATCH-INFRA_2019-11_263088, 8.2.1_UGA-RUP_2020-04_268536
• NVIDIA Mellanox SN3700V – NCLU Version	• 1.0-cl4.2.1u1
• NVIDIA Mellanox SN3700V – Distribution Release	• 4.2.1
• NVIDIA Mellanox QM8700 Product Release	• 3.9.0606
• DGX A100 – Base OS	• 4.99.11
• DGX A100 – Linux kernel	• 5.3.0-59-generic
• DGX A100 – NVIDIA Driver	• 450.51.06
• DGX A100 – Ubuntu	• 18.04.5 LTS
• NVIDIA NGC MXNet Image	• nvcr.io/nvidia/mxnet:20.06-py3
• MLPerf Benchmarks	• https://github.com/mlperf/training_results_v0.7/tree/master/NVIDIA/benchmarks/resnet/implementations/mxnet

3 Deep learning training performance and analysis

3.1 Benchmark methodology

In order to measure the performance of the solution, the image classification benchmark from the MLPerf Benchmark Suite repository was executed. This benchmark performs training of an image classification convolutional neural network (CNN) on labeled images using MXNet. Essentially, the system learns whether an image contains a cat, dog, car, train, etc. The well-known ILSVRC2012 image dataset (often referred to as ImageNet) was used. This dataset contains 1,281,167 training images in 144.8 GB[1]. All images are grouped into 1000 categories or classes. This dataset is commonly used by DL researchers for benchmarking and comparison studies.

The individual JPEG images in the ImageNet dataset were converted to RecordIO format. The dataset was not resized, not normalized and no preprocessing was performed on the raw ImageNet JPEG images. It maintains the image compression offered by the JPEG format and the total size of the dataset remained roughly the same (148 GB). The average image size was 115 KB.

The benchmark results in this section were obtained with four F800 nodes in the cluster. Each result is the average of five executions.

3.2 MLPerf Benchmark results

There are a few conclusions that we can make from the benchmarks represented in Figure 5.

- Image throughput and therefore storage throughput scale linearly from 8 to 32 GPUs.
- The difference between Epoch 0 (when the data is pulled from storage and cached) and Overall is minor, so the storage is not a bottleneck.

[1] All unit prefixes in this document use the SI standard (base 10) where 1 GB is 1 billion bytes.

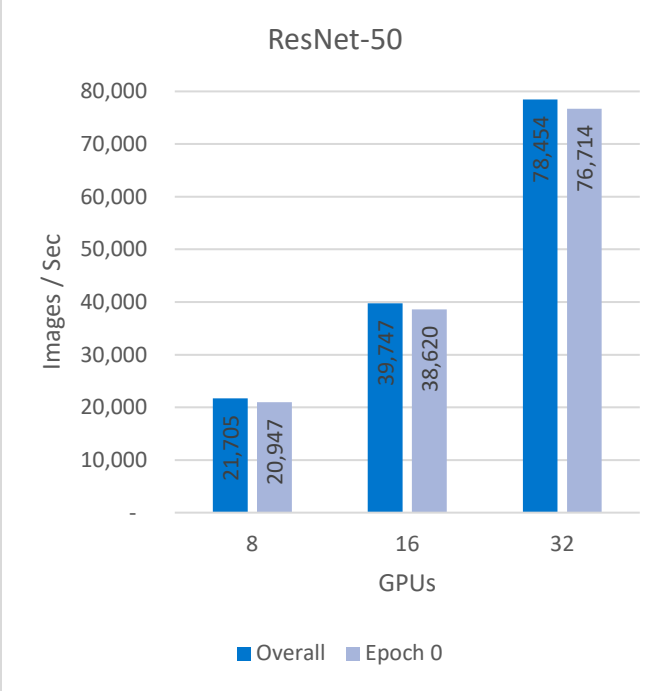

Figure 5 Model Development – Training Benchmark Results

3.3 NVIDIA collective communication library (NCCL)

There are two significant types of network traffic when performing training. First, there is the NFS traffic required to read the images (RecordIO files). This uses standard TCP/IP on Ethernet. Second, the gradient (the derivative of the loss function with respect to each parameter) calculated on each GPU must be averaged with the gradient from all other GPUs and the resulting average gradient must be distributed to all GPUs. This is performed optimally using the MPI All Reduce algorithm. See Figure 6 below, which optimally calculates the sum of values from all nodes (top) and stores the sum on all nodes (bottom).

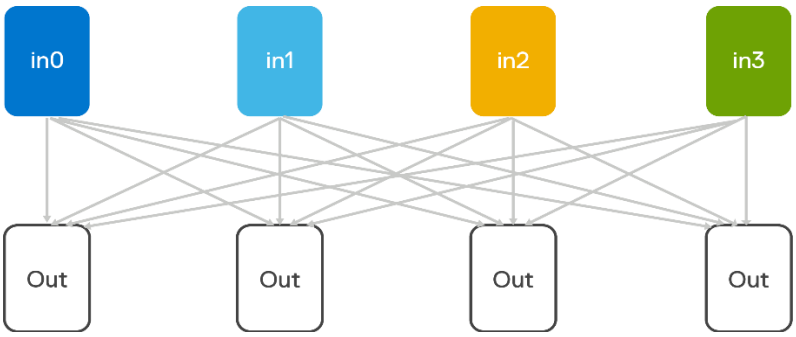

Figure 6 The All Reduce algorithm

NCCL provides fast collectives (such as All Reduce) over multiple GPUs both within and across nodes. It supports a variety of interconnect technologies including PCIe, NVSwitch technology, InfiniBand Verbs, and IP sockets. NCCL also automatically patterns its communication strategy to match the system's underlying GPU interconnect topology. When configured properly on the DGX A100 system, NCCL allows GPUs in different nodes to communicate with each other through the PCIe switch, NIC, and the InfiniBand switch, bypassing the CPU and the main system memory. Figure 7 demonstrates the performance of the interconnects between GPUs. We can see that with eight GPUs the bandwidth is higher than 16 and 32. This is an expected behavior because with eight GPUs, only internal NVLink communications are used.

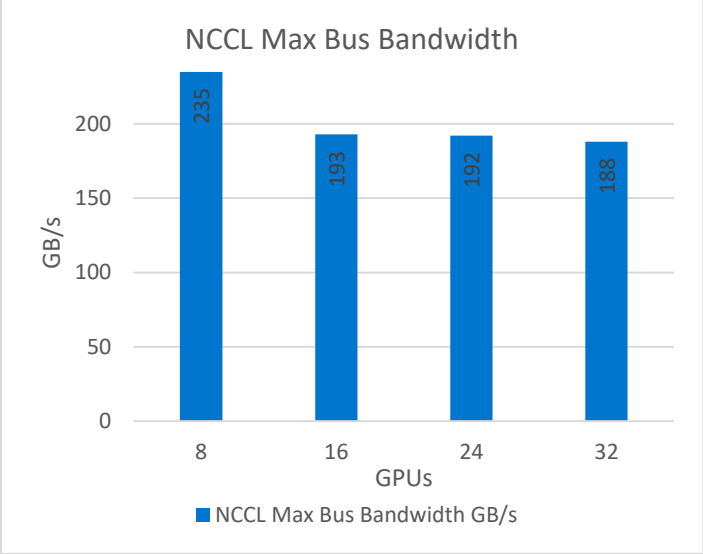

Figure 7 The NCCL Max Bus Bandwidth

3.4 Storage-only performance using FIO

To understand the limits of Isilon storage I/O in this environment, we used the common storage benchmark tool FIO to perform Reads and Writes tests. During those tests, each DGX A100 system was pointing to a different F800 NICs, with a one to one relationship, which means that with four DGX A100 systems, only one F800 chassis with four nodes were used. In a real production workload, it is highly recommended to distribute the load across every available node within the Dell EMC Isilon F800 cluster to increase the performance. A single DGX A100 system was measured to read from Isilon at 4,959 MB/s saturating the storage link (40Gb/s) for this network interface. With four DGX A100 systems, we hit the throughput limit of the four F800 nodes at 18,924 MB/sec, shown in Figure 8. We can also see that the platform scales linearly as shown in Figure 8 and Figure 9.

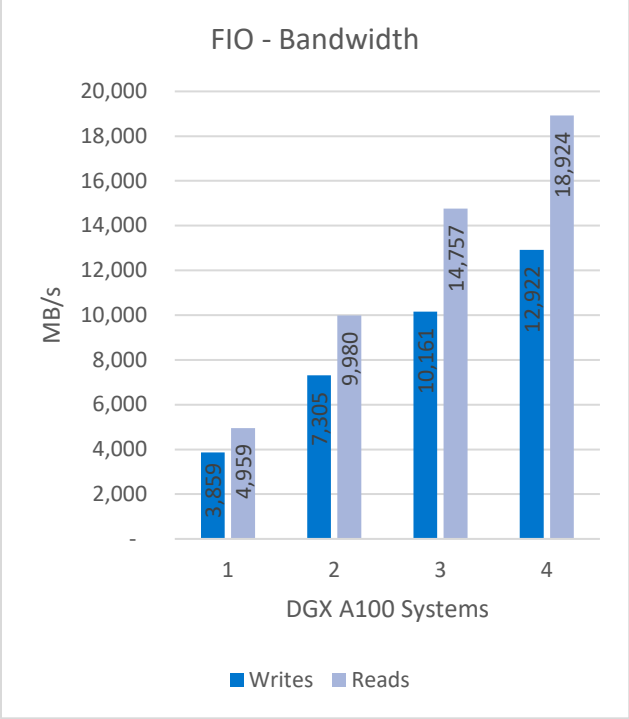

Figure 8 FIO read/write bandwidth from Isilon

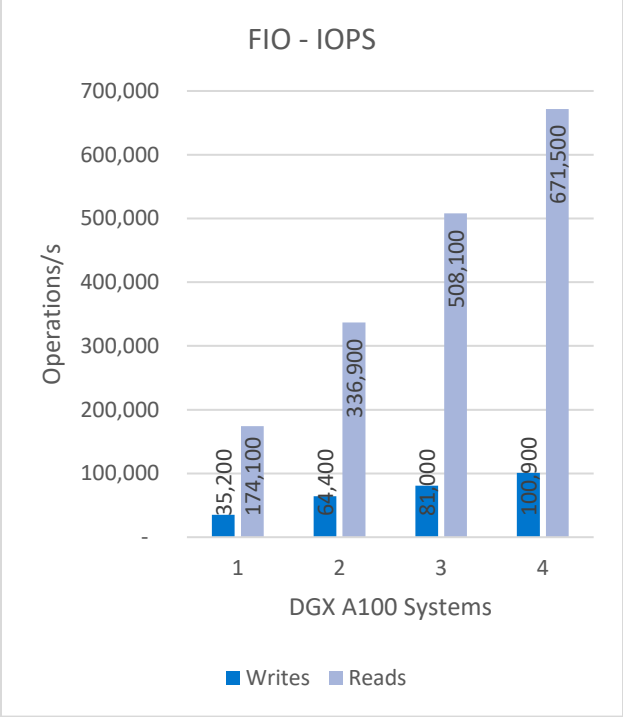

Figure 9 FIO read/write IOPS from Isilon

4 Solution sizing guidance

DL workloads vary significantly with respect to the demand for compute, memory, disk, and I/O profiles, often by orders of magnitude. Sizing guidance for the GPU quantity and configuration of Isilon nodes can only be provided when these resource requirements are known in advance. That said, it's usually beneficial to have a few data points on the ratio of GPUs per Isilon node for common image classification benchmarks. Based on the results from the MLPerf benchmark, we can estimate that a single F800 node could handle eight NVIDIA A100 Tensor Core GPUs. In conclusion, a single Dell EMC Isilon F800 chassis made of four nodes could serve four NVIDIA DGX A100 systems for this type of image classification workload.

5 Conclusion

This document presents a high-performance architecture for DL by combining NVIDIA DGX A100 systems with NVIDIA A100 Tensor Core GPUs, NVIDIA Mellanox SN3700V and QM8700 switches, and Dell EMC Isilon F800 all-flash NAS storage. We discuss key features of Isilon that makes it a powerful persistent storage for DL solutions. This new reference architecture extends the commitment that Dell Technologies and NVIDIA have to make AI simple and accessible to every organization with our unmatched set of joint offerings. Together we offer our customers informed choice and flexibility in how they deploy high-performance DL at scale. Throughout the benchmarks we validated that the Isilon all-flash F800 storage was able to keep pace and linearly scale performance with DGX A100 systems.

It is important to point out that DL algorithms have a diverse set of requirements with various compute, memory, I/O, and disk capacity profiles. That said, the architecture and the performance data points presented in this whitepaper can be utilized as the starting point for building DL solutions tailored to varied sets of resource requirements. More importantly, all the components of this architecture are linearly scalable and can be independently expanded to provide DL solutions that can manage tens of PBs of data.

While the solution presented here provides several performance data points and speaks to the effectiveness of Isilon in handling large scale DL workloads, there are several other operational benefits of persisting data for DL on Isilon:

- The ability to run AI in-place on data using multi-protocol access
- Enterprise grade features out-of-box
- Seamlessly tier to more cost-effective nodes
- Scale up to 58 PB per cluster

In summary, Isilon-based DL solutions deliver the capacity, performance and high concurrency to eliminate the I/O storage bottlenecks for AI. This provides a rock-solid foundation for large scale, enterprise-grade DL solutions with a future proof scale-out architecture that meets your AI needs of today and scales for the future.

6 References

Name	Link
MLPerf Benchmark Suite	https://github.com/mlperf/training_results_v0.7/tree/master/NVIDIA/benchmarks/resnet/implementations/mxnet
DELL EMC PowerScale system	https://www.delltechnologies.com/nl-be/collaterals/unauth/data-sheets/products/storage/h18248-spec-sheet-dell-emc-powerscale.pdf
NVIDIA DGX A100 system	https://www.nvidia.com/en-us/data-center/dgx-a100/
Dell EMC PowerSwitch S- and Z-Series	https://www.delltechnologies.com/en-us/networking/data-center-switches/index.htm
Dell EMC PowerSwitch Reference Guide	https://www.dellemc.com/resources/en-us/asset/quick-reference-guides/products/networking/Dell_EMC_Networking_-_QRG_-_Data_Center.pdf.external
NVIDIA Mellanox SN3700V Ethernet switch	https://www.mellanox.com/products/ethernet-switches/sn3000
NVIDIA Mellanox QM8700 InfiniBand switch	https://www.mellanox.com/products/infiniband-switches/QM8700
NVIDIA NGC MXNet image	https://ngc.nvidia.com/catalog/containers/nvidia:mxnet (Tag 20.06-py3)
Dell EMC Isilon OneFS Best Practices	https://www.emc.com/collateral/white-papers/h16857-wp-onefs-best-practices.pdf
Dell EMC Isilon OneFS SmartFlash	https://www.dellemc.com/resources/en-us/asset/white-papers/products/storage/h13249-isilon-onefs-smartflash-wp.pdf
Dell EMC Isilon OneFS Technical Overview	https://www.dellemc.com/en-tz/collaterals/unauth/technical-guides-support-information/products/storage/h10719-isilon-onefs-technical-overview-wp.pdf
Dell EMC Isilon Storage Tiering	https://www.dellemc.com/resources/en-us/asset/white-papers/products/storage/h8321-wp-smartpools-storage-tiering.pdf
Gartner	https://www.gartner.com/en/newsroom/press-releases/2019-08-05-gartner-says-ai-augmentation-will-create-2point9-trillion-of-business-value-in-2021
ImageNet	http://www.image-net.org/challenges/LSVRC/2012/