Whitepaper

# Dell Precision Data Science Workstation with Isilon H400

This whitepaper demonstrates how the Dell Precision 7920 Tower Data Science Workstation with NVIDIA® Quadro® RTX 6000 GPUs and Dell EMC Isilon H400 scale-out NAS can be used to provide an excellent environment for small teams performing data science, AI, and deep learning. The results of industry-standard image classification training benchmarks using TensorFlow are included.

March 2020

Dell Precision Data Science Workstation with Isilon

## Table of Contents

## Revisions

| Date | Description | Author |
|---|---|---|
| March 2020 | Initial release | Claudio Fahey |

## Executive summary

This whitepaper focuses on how the Dell Precision 7920 Tower Workstation with NVIDIA GPUs and Dell EMC Isilon scale-out NAS (Network Attached Storage) accelerates AI innovation and collaboration by providing shared access to very large datasets with high performance and scalability. This is an excellent entry-level solution for small teams of up to 5 members that expect to grow and need a system that can grow with them.

## Audience

This document is intended for organizations and small data science teams that are looking to accelerate AI innovation and collaboration. Solution architects, system administrators, data scientists, data engineers, and other interested readers within those organizations constitute the target audience.

## Introduction

An efficient data science team often requires the ability to share large amounts of data while providing high performance, reliability, and seamless access from multiple operating systems. A new team may often start small with as few as two members, but they will still need access to large amounts of data. Further, as the data science team grows, the compute and storage needs will increase as well.

Dell Technologies and NVIDIA offer multiple GPU-accelerated servers and systems for data center environments including the Dell EMC PowerEdge C4140, Dell EMC DSS 8440, NVIDIA DGX-1™, and NVIDIA DGX-2™. We also have complete solutions that combine these systems with networking and Dell EMC Isilon scale-out NAS,

This document focuses on the latest step in the Dell Technologies and NVIDIA collaboration, a new AI reference architecture that combines the Dell Precision 7920 Tower Data Science Workstation with NVIDIA GPUs and Isilon H400 storage for small data science teams of up to 5 workstations. This is an excellent entry-level solution for small teams that expect to grow and need a system that can grow with them without the need of data migration. In such use cases, Dell Data Science Workstations are ideal enterprise-class high performance development platforms flexible for AI/ML/DL model experimentation and development prior to taking the models to scale in the business's data center containing even larger datasets and more GPUs.

Deep learning (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 skill-sets, 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.

The results of industry standard DL image classification benchmarks using TensorFlow are included in this whitepaper.

## Solution architecture

### OVERVIEW

Figure 1 illustrates the reference architecture showing the key components that make up the solution. Note that in a customer deployment, the number of workstations and Isilon storage nodes will vary and can be scaled independently to meet the requirements of the specific workload. Refer to the Solution sizing guidance section for details.

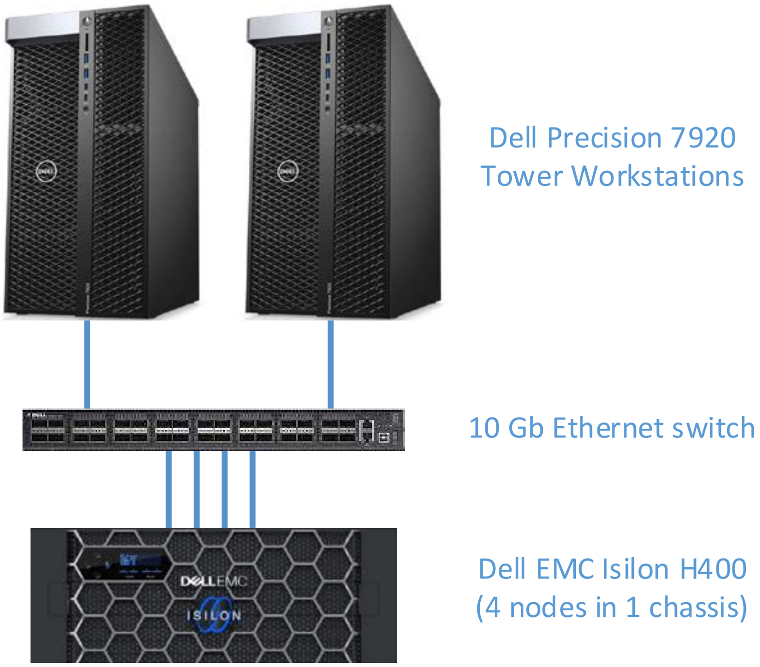

*Figure 1: Reference Architecture*

DELL PRECISION 7920 TOWER DATA SCIENCE WORKSTATION

The world's most powerful workstation, the Dell Precision 7920 Tower Workstation, provides ultimate performance and scalability to grow alongside your AI initiatives and data. The Dell Data Science Workstation as tested for this document includes an Intel® Xeon® Gold 6134 8 core CPU, 128 GB CPU RAM, two NVIDIA Quadro RTX 6000 GPUs with 24 GB of GPU RAM each, a Samsung PM951 NVMe drive, and a Mellanox ConnectX4LX NIC. Software includes Ubuntu Linux and the NVIDIA Quadro RTX GPU accelerated NVIDIA Data Science Software stack

DELL EMC ISILON H400 SCALE-OUT NAS

An efficient data science team often requires the ability to share massive amounts of data while providing high performance, reliability, and seamless access from multiple operating systems. The Dell EMC Isilon scale-out NAS provides this critical capability. Dell EMC Isilon hybrid storage platforms, powered by the OneFS operating system, use a highly versatile yet simple scale-out storage architecture to speed access to massive amounts of data, while dramatically reducing cost and complexity. The hybrid storage platforms are highly flexible and strike the balance between large capacity and high-performance storage to provide support for a broad range of enterprise file workloads. The H400 provides a balance of performance, capacity and value to support a wide range of file workloads. And it delivers up to 3 GB/s bandwidth per chassis and provides capacity options ranging from 120 TB to 720 TB per chassis.

Each H400 chassis, shown in Figure 2, contains four storage nodes, 60 SATA HDD drives and eight 10 GbE network connections. OneFS combines up to 252 nodes in 63 chassis into a single high-performance file system designed to handle the most intense I/O workloads such as DL. As performance and capacity demands increase, both can be scaled-out simply and non-disruptively, allowing applications and users to continue working.

*Figure 2: Isilon H400 chassis, containing four storage nodes*

In the solution tested for this document, four H400 nodes, in one chassis, were used.

Dell EMC Isilon H400 has the following features.

**High capacity with the ability to grow as needed:**

- 120 TB to 720 TB raw HDD capacity per chassis; up to 45 PB per cluster
- Up to 3 GB/s throughput per chassis

**The ability to run AI in-place on data using multi-protocol access:**

- Multi-protocol support such as SMB, NFS, HTTP, and native HDFS to maximize operational flexibility

This eliminates the costly need to migrate/copy data and results over to a separate AI stack.

**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 SmartPools
- Grow-as-you-go scalability with up to 45 PB HDD 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
- 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 Isilon OneFS that make it an excellent storage system for DL workloads that require performance, concurrency, and scale. These features are detailed below.

Storage tiering

Dell EMC Isilon 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.

SmartPools allows you to define the value of the data within your workflows based on policies and automatically aligns data to the appropriate price/performance tier over time. Data movement is seamless and with file-level granularity and control via automated policies, manual control or API, you can tune performance and layout, storage tier alignment and protection settings – all with minimal impact to your end-users.

Storage tiering has a very convincing value proposition, namely segregating data according to its business value and aligning it with the appropriate class of storage and levels of performance and protection. Information Lifecycle Management techniques have been around for several years, but have typically suffered from the following inefficiencies: complex to install and manage, involves changes to the file system, requires the use of stub files, etc.

Dell EMC Isilon SmartPools is a next generation approach to tiering that facilitates the management of heterogeneous clusters. The SmartPools capability is native to the Isilon OneFS scale-out file system, which allows for unprecedented flexibility, granularity, and ease of management. In order to achieve this, SmartPools leverages many of the components and attributes of OneFS, including data layout and mobility, protection, performance, scheduling and impact management.

A typical Isilon cluster will store multiple datasets with different performance, protection, and price requirements. Generally, files that have been recently created and accessed should be stored in a hot tier while files that have not been accessed recently should be stored in a cold tier. Because Isilon supports tiering based on a file’s access time, this can be performed automatically. For storage administrators that want more control, complex rules can be defined to set the storage tier based on a file’s path, size, or other attributes.

All files on Isilon are always immediately accessible (read and write) regardless of their storage tier and even while being moved between tiers. The file system path to a file is not changed by tiering. Storage tiering policies are applied, and files are moved by the Isilon SmartPools job, which runs daily at 22:00 by default.

For more details, see Storage Tiering with Dell EMC Isilon SmartPools.

### OneFS caching

The 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. Remote memory is accessed over an internal interconnect and has lower latency than accessing hard disk drives and SSDs.

For files marked with an access pattern of concurrent or streaming, 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.

OneFS uses up to three levels of read cache, plus an NVRAM-backed write cache. L1 and L2 read caches use RAM while L3 uses the SSDs that are available on all Isilon hybrid nodes.

For more details, see OneFS SmartFlash.

### Locks and concurrency

OneFS has a fully distributed lock manager that coordinates locks on data across all nodes in a storage cluster. The lock manager is highly extensible and allows for multiple lock personalities to support both file system locks as well as cluster-coherent protocol-level locks such as SMB share mode locks or NFS advisory-mode locks. Every node in a cluster is a coordinator for locking resources and a coordinator is assigned to lockable resources based upon an advanced hashing algorithm.

Efficient locking is critical to support the efficient parallel I/O profile demanded by many iterative AI and DL workloads enabling concurrent file read access up into the millions.

For more details, see the OneFS Technical Overview.

## KEY HARDWARE COMPONENTS

Table 1 shows the key hardware components as tested for this document.

| Component | Purpose | Quantity |
|---|---|---|
| Dell EMC Isilon H400<br>120 TB HDD<br>12.8 TB SSD<br>256 GB RAM<br>Four 1 GbE, eight 10 GbE interfaces | Shared storage | 1 4U chassis<br>(4 nodes) |
| Dell Precision 7920 Tower Data Science Workstation<br>Intel(R) Xeon(R) Gold 6134 8-core CPU @ 3.20GHz<br>128 GB RAM | Workstation | 1 |

| |
|---|
| 2 NVIDIA Quadro RTX 6000 with 24 GB of RAM each<br>PM951 NVMe SAMSUNG 1024GB<br>Mellanox ConnectX4LX NIC |

*Table 1: Hardware Components*

SOFTWARE VERSIONS

Table 2 shows the software versions that were tested for this document.

| Component | Version |
|---|---|
| AI Benchmark Util | https://github.com/claudiofahey/ai-benchmark-util/commit/ca7f5d2 |
| Dell EMC Isilon – OneFS | 8.2.0.0 |
| NVIDIA Driver | 435.21 |
| NVIDIA CUDA | 10.0 |
| Mellanox OFED Driver | 4.7-3.2.9.0-ubuntu19.10-x86_64 |
| Ubuntu | 19.10 |
| Docker Engine | 19.03.3 |
| NVIDIA GPU Cloud TensorFlow Image | nvcr.io/nvidia/tensorflow:19.09-py3 |
| TensorFlow | 1.14.0 |
| TensorFlow Benchmarks | https://github.com/claudiofahey/benchmarks/commit/31ea13f |

*Table 2: Software Versions*

## Deep learning training performance and analysis

### BENCHMARK METHODOLOGY

In order to measure the performance of the solution, various benchmarks from the TensorFlow Benchmarks repository were executed. This suite of benchmarks performs training of an image classification convolutional neural network (CNN) on labeled images. 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 1024 TFRecord files. The TFRecord file format is a Protocol Buffers binary format that combines multiple JPEG image files together with their metadata (bounding box for cropping and label) into one binary file. 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.

As many datasets are often significantly larger than ImageNet, we wanted to determine the performance with datasets that are larger than the 256 GB of coherent shared cache available across the four-node Isilon H400 cluster. To accomplish this, we simply made 13 exact copies of each TFRecord file, creating a 2.0 TB dataset. Having 13 copies of the exact same images doesn't improve training accuracy or speed but it does produce the same I/O pattern for the storage, network, and GPUs. Having identical files did not provide an unfair advantage as Isilon deduplication was not enabled and all images are reordered randomly (shuffled) in the input pipeline.

A key critical question one has when trying to size a system is how fast the storage must be so that it is not a bottleneck. To answer this question, we copied the 148 GB dataset to the workstation's NVMe drive and ran the benchmark from this high-speed disk. The image rate (images/sec) measured in this way accounts for the significant preprocessing pipeline as well as the GPU computation. To determine the throughput (bytes/sec) demanded by this workload, we simply multiply the images/sec by the average image size (115 KB). In the next section, results using this method are labeled Local NVMe.

Prior to each execution of the benchmark, the L1, L2, and L3 caches on Isilon were flushed with the command `isi_for_array isi_flush`. In addition, the Linux buffer cache was flushed on all compute systems by running `sync; echo 3 > /proc/sys/vm/drop_caches`. However, note that the training process will read the same files repeatedly and after just several minutes, much of the data will be served from one of these caches.

### BENCHMARK RESULTS

There are a few conclusions that we can make from the benchmarks represented in Figure 3.

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

- Image throughput and therefore storage throughput scale linearly from 1 to 2 GPUs.
- There is no significant difference in image throughput between Local NVMe and Isilon.
- The highest storage throughput demand was 139 MB/sec and occured during ResNet-50 with 2 GPUs.
- The Isilon H400 is easily capable of handling this workload.

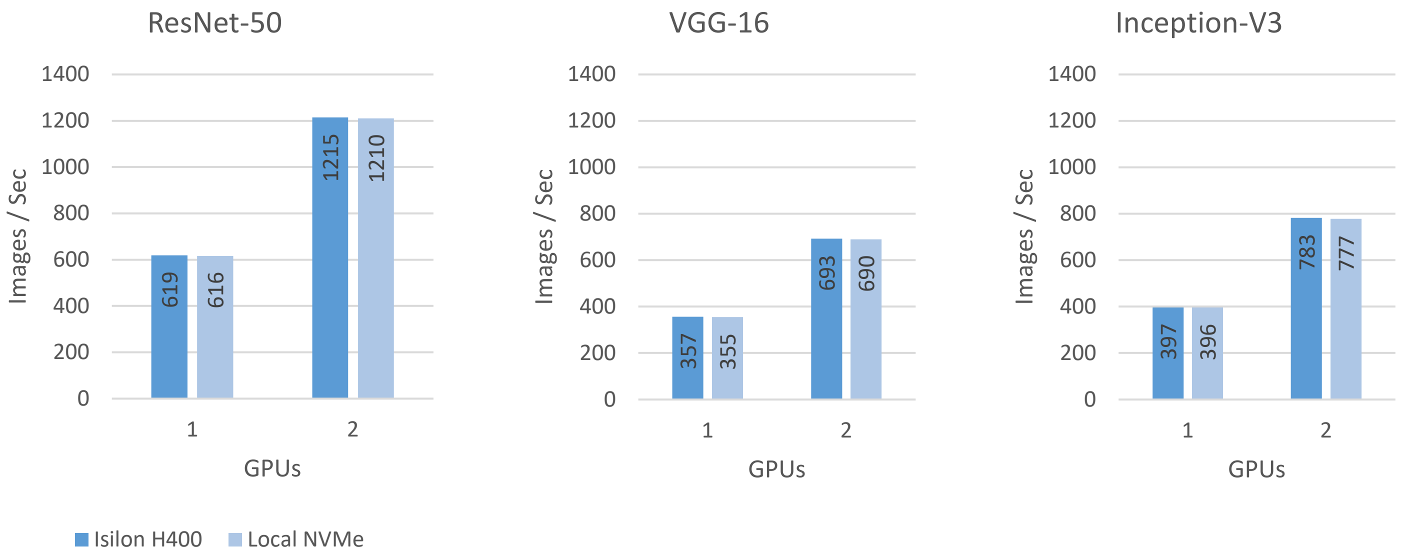

*Figure 3: Model Development – Training Benchmark Results*

## Solution sizing guidance

The Data Science Precision Workstation solution described in this whitepaper is an entry-level solution for small data science teams using up to 5 workstations. The Isilon H400 is expected to provide approximately 3 GB/sec for most read workloads and 2 GB/sec for most write workloads. With up to 5 workstations (10 GPUs), each GPU would be able to use 200-300 MB/sec on average. This is enough for many AI workloads. However, be aware that AI/DL workloads vary significantly with respect to the demand for compute, memory, disk, and I/O profiles, often by orders of magnitude.

As teams grow or demand more storage performance and capacity, Isilon can be easily expanded by adding additional nodes. Additional H400 nodes can be added in increments of two. If a higher performance-to-capacity ratio is needed, four or more nodes of different types can be added, and storage tiering can be utilized. Isilon provides faster (H500) and higher-capacity (H5600) hybrid nodes as well as all-flash nodes (F800 and F810) for extreme performance.

An understanding of the I/O throughput demanded per GPU for the specific workload and the total storage capacity requirements can help provide better guidance on Isilon node count and configuration. It is recommended to reach out to the Dell EMC account and SME teams to provide this guidance for the specific AI workload, throughput, and storage requirements.

## Conclusions

This document presented a scalable architecture for small data science teams by combining Dell Precision 7920 Tower Data Science Workstations with single and dual NVIDIA Quadro RTX 6000 GPUs, and Dell EMC Isilon H400 scale-out NAS. One can expect comparable results from single or dual NVIDIA Quadro RTX 8000 GPU configurations providing larger amounts of GPU RAM. We discussed key features of Isilon that make it a powerful persistent storage system for AI solutions. This new reference architecture extends the commitment of Dell Technologies and NVIDIA to making AI simple and accessible to every organization with our unmatched set of joint offerings. Together we provide our customers with informed choices and flexibility in how they deploy DL at any scale.

It is important to point out that AI 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 AI 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 AI solutions that can manage tens of PBs of data.

While the benchmarks presented here provide several performance data points, there are many other operational benefits of persisting data for AI on Isilon:

- The ability to run AI in-place on data using multi-protocol access without the need for migration
- Enterprise grade features out-of-box
- Seamlessly tier to more cost-effective nodes
- Scale up to 58 PB per cluster (some node models will have lower limits)

In summary, Dell EMC Isilon-based storage solutions deliver the capacity, performance and high concurrency to eliminate the I/O storage bottlenecks for AI. Combined with the Dell Precision 7920 Data Science Workstation, this solution provides a rock-solid foundation for large scale, enterprise-grade data science solutions with a future proof scale-out architecture that meets your AI needs of today and scales for the future.

## Appendix – Benchmark details

The command below was used to perform the ResNet-50 training with 2 GPUs.

```
mpirun \
--n 2 \
-allow-run-as-root \
--host localhost:2 \
--report-bindings \
-bind-to none \
-map-by slot \
-x LD_LIBRARY_PATH \
-x PATH \
-mca plm_rsh_agent ssh \
-mca plm_rsh_args "-p 2222" \
-mca pml ob1 \
-mca btl ^openib \
-mca btl_tcp_if_include enp53s0 \
-x NCCL_DEBUG=INFO \
-x NCCL_IB_HCA=mlx5 \
-x NCCL_IB_SL=4 \
-x NCCL_IB_GID_INDEX=3 \
-x NCCL_NET_GDR_READ=1 \
-x NCCL_SOCKET_IFNAME=^docker0,lo \
./round_robin_mpi.py \
python \
-u \
/mnt/isilon/data/tensorflow-benchmarks/scripts/tf_cnn_benchmarks/\
tf_cnn_benchmarks.py \
--model=resnet50 \
--batch_size=192 \
--batch_group_size=20 \
--num_batches=1000 \
--nodistortions \
--num_gpus=1 \
--device=gpu \
--force_gpu_compatible=True \
--data_format=NCHW \
--use_fp16=True \
--use_tf_layers=True \
--data_name=imagenet \
--use_datasets=True \
--num_intra_threads=1 \
--num_inter_threads=40 \
--datasets_prefetch_buffer_size=40 \
--datasets_num_private_threads=4 \
--train_dir=/mnt/isilon/data/train_dir/2019-10-24-14-53-59-resnet50 \
--sync_on_finish=True \
--summary_verbosity=1 \
--save_summaries_steps=100 \
--save_model_secs=600 \
--variable_update=horovod \
--horovod_device=gpu \
    --data_dir=/mnt/isilon1/data/imagenet-scratch/tfrecords-13x
```

For the other models, only the `--model` parameter was changed. Each result is the average of three executions.

## References

| Name | Link |
|---|---|
| AI Benchmark Utilities | https://github.com/claudiofahey/ai-benchmark-util/tree/dsws |
| Deep Learning with Dell EMC Isilon | https://www.dellemc.com/resources/en-us/asset/white-papers/products/storage/h17361_wp_deep_learning_and_dell_emc_isilon.pdf |
| Dell EMC Isilon and Dell EMC DSS 8440 Servers for Deep Learning | https://www.dellemc.com/resources/en-us/asset/white-papers/products/storage/h17843_wp_dell_emc_isilon_and_dss_8440_servers_for_deep_learning.pdf |
| Dell EMC Isilon H400 | https://www.dellemc.com/en-ca/collaterals/unauth/data-sheets/products/storage/h16071-ss-isilon-hybrid.pdf |
| 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 |
| Dell EMC Isilon and NVIDIA DGX-1 servers for deep learning | https://www.dellemc.com/resources/en-us/asset/white-papers/products/storage/Dell_EMC_Isilon_and_NVIDIA_DGX_1_servers_for_deep_learning.pdf |
| Dell EMC Isilon, PowerSwitch and NVIDIA DGX-2 Systems for Deep Learning | https://www.dellemc.com/resources/en-us/asset/white-papers/partner/h18079-dell-emc-isilon-powerswitch-and-nvidia-dgx-2-systems-for-deep-learning.pdf |
| Dell EMC Ready Solutions for AI, Machine and Deep Learning | https://www.dellemc.com/content/dam/uwaem/production-design-assets/en-gb/solutions/assets/pdf/dell-emc-ready-solutions-for-ai-and-dl.pdf |
| Dell Precision 7920 Tower | https://i.dell.com/sites/csdocuments/Shared-Content_data-Sheets_Documents/en/us/Precision-7920-Tower-Spec-Sheet.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/ |
| TensorFlow | https://github.com/tensorflow/tensorflow |
| TensorFlow Benchmarks | https://github.com/tensorflow/benchmarks |