Hewlett Packard Enterprise

Technical white paper

Accelerate time to value and AI insights

Reducing the AI development cycle with HPE, NVIDIA, WekaIO, and Mellanox

Contents

Introduction

Artificial intelligence (AI), with an emphasis on deep learning (DL), has fueled the growth of innovation across a broad range of use cases including autonomous vehicles (AV), fraud detection, speech recognition, and predictive medicine. NVIDIA® has led the industry with GPU advancements that allow data scientists to build increasingly more complex models with huge data training sets—in the petascale range—putting greater focus on the underlying infrastructure required to create a balanced solution.

AI development is complex and requires the right technology and methodology to be successful. HPE provides a single source for both. Working closely with partners, HPE delivers technology along with guidance from resources such as the HPE Deep Learning Cookbook and consultative services, enabling development of complete AI solutions.

This paper explores the impact of storage I/O on the training portion of the DL workflow and on inferencing for training model validation. AI development is akin to software development; while it is desirable to reduce overall cycle time, the finished product must function correctly, predictably, and reliably. The results shared in this paper show how the current popular neural network models are able to fully utilize GPU resources without saturating storage resources. However, increasing storage performance will help avoid I/O bottlenecks during model validation, reducing overall time for model development.

For real-life training environments, particularly in complex workloads found in AV and fraud detection, data sets can range from hundreds of terabytes to tens of petabytes, making a shared storage solution essential for the DL training process. In this case, local storage is not an option and requires a high-performance, scalable, shared storage solution.

This paper was created in partnership with engineers from HPE, NVIDIA, WekaIO and Mellanox. These engineers worked together to design a DL architecture that would provide high performance for DL training and validation workflows. The results shared in this paper demonstrate that the resulting solution—based on the HPE Apollo 6500 Gen10 System, NVIDIA Tesla® V100 GPUs, WekaIO Matrix flash-optimized parallel file system, and Mellanox InfiniBand networking—delivers a high-performance solution for deep learning.

HPE, together with our partners, provides the tools, hardware, support, and guidance to enable modern AI solutions along with the confidence of understanding how to optimize solutions to provide the desired outcomes.

Intended audience

This paper is intended for data scientists, solution builders, and IT personnel who are interested in the performance levels of an AI solution and how that would affect time to production.

Deep learning dataflow

The deep learning dataflow is complex, with multiple steps required to transform and manipulate data for use with DL models.

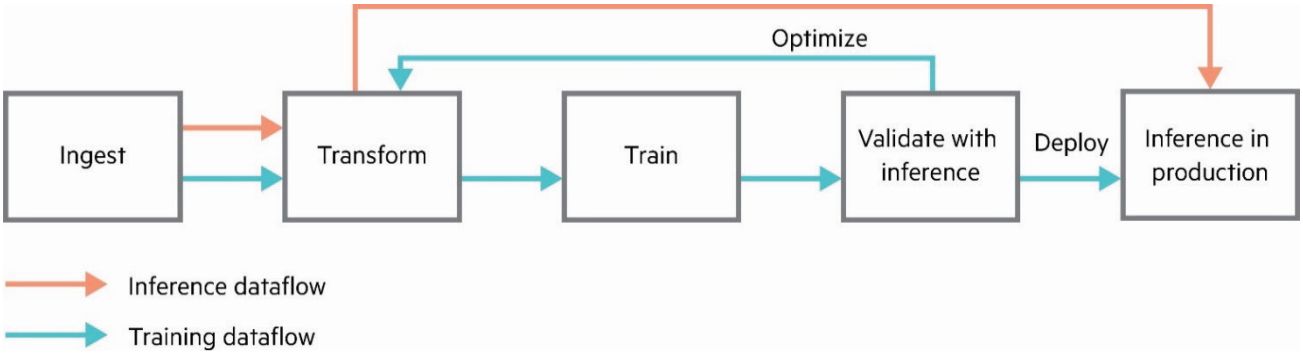


Figure 1. Deep learning dataflow

Data is ingested and transformed through cleansing and pre-processing techniques to be usable as part of a data set to train a DL model. Critically, after a model is trained, it must be validated to ensure it meets production inference requirements before deployment. To ensure the trained model meets accuracy, reliability, and other quality and performance requirements, the model must be tested in a batch inferencing or simulated environment. For example, a model trained for fraud detection must be tested to ensure accuracy and false positives meet requirements. Or, a model for autonomous vehicles designed to recognize pedestrians must be tested to ensure it will recognize them in varying lighting conditions. If the model fails in the validation stage, it must be further trained. Depending on the results of the model validation, the model may need to be retrained many times to improve performance or accuracy, which takes a significant amount of time. This is an iterative phase of model development and is analogous to many continuous integration workflows that exist in other disciplines of software development.

Deep learning infrastructure

Infrastructure choices have a significant impact on the performance and scalability of a deep learning workflow. Model complexity, catalog data size, and input type (such as images and text) will impact key elements of a solution, including the number of GPUs, servers, network interconnects, and storage type (local disk or shared). The more complex the environment, the greater the need to balance components. HPE has integrated best-in-class components into its server infrastructure for deep learning, including GPUs from NVIDIA, 100 Gbps InfiniBand networking from Mellanox, and the Matrix high-performance shared file system software from WekaIO.

Compute: HPE Apollo 6500 Gen10 system

The HPE Apollo 6500 Gen10 system is an ideal DL platform that provides performance and flexibility with industry leading GPUs, fast GPU interconnects, high bandwidth fabric, and a configurable GPU topology to match varied workloads. The HPE Apollo 6500 provides rock-solid reliability, availability, and serviceability (RAS) features; includes up to eight GPUs per server; next generation NVIDIA NVLink™ for fast GPU-to-GPU communication; support for Intel® Xeon® Scalable processors; a choice of high-speed/low-latency fabric; and is workload-enhanced using flexible configuration capabilities.

GPU: NVIDIA Tesla V100 GPU accelerator

The HPE Apollo 6500 Gen10 system supports up to eight NVIDIA Tesla V100 SXM2 32 GB GPU modules, (however, 16 GB GPUs were used in these benchmarks.) Powered by NVIDIA Volta architecture, the Tesla V100 is the world's most advanced data center GPU, designed to accelerate AI, HPC, and graphics. Each Tesla V100 GPU processor offers the performance of up to 100 CPUs in a single GPU and can deliver 15.7 TFLOPS of single-precision performance and 125 TFLOPS of deep learning performance, for a total of one EFLOP when fully populated with eight Tesla V100 GPUs.[1] The tested architecture leverages NVIDIA NVLink technology to provide higher bandwidth and scalability for multi-GPU configurations. A single V100 GPU supports up to six NVIDIA NVLink connections for GPU-to-GPU communication, for a total of 300 GBps.[2]

Networking: Mellanox 100 Gb EDR InfiniBand

When GPU workloads and data sets scale beyond a single HPE Apollo 6500 system, a high-performance network fabric is critical for maintaining high-performance, inter-node communication, as well as enabling the external storage system to deliver full bandwidth to the GPU servers. For networking, Mellanox switches, cables, and network adapters provide industry-leading performance and flexibility for an HPE Apollo 6500 system in a DL solution. Mellanox is an industry-leading supplier of high-performance Ethernet and InfiniBand interconnects for high-performance GPU clusters used for DL workloads and for storage interconnect.

With technologies such as remote direct memory access (RDMA) and GPUDirect, Mellanox enables excellent machine learning scalability and efficiency at network speeds from 10 to 100 Gbps. The InfiniBand network provides a high-performance interconnect between multiple GPU servers as well as providing network connectivity to the shared storage solution.

Storage: WekaIO Matrix on HPE ProLiant DL360 Gen10 Servers

To minimize idle time for compute clients, HPE partners with WekaIO for its high-performance shared storage. WekaIO Matrix[3] includes the MatrixFS flash-optimized parallel file system, qualified on industry-leading HPE Apollo 2000 Gen10 systems and HPE ProLiant DL360 Gen10 Servers, and utilizing advanced Mellanox interconnect features. Matrix is a radically simple storage solution that delivers the performance of all-flash arrays with the scalability and economics of the cloud. Matrix transforms NVMe-based flash storage, compute nodes, and interconnect fabrics into a high-performance, scale-out parallel storage system that is well suited for I/O-bound use cases.

WekaIO MatrixFS meets or exceeds the requirements of AI architectures. MatrixFS was purpose-built with distributed data and metadata support to avoid hotspots or bottlenecks encountered by traditional scale-out storage solutions, exceeding the performance capabilities of even local NVMe storage. It supports distributed data protection (MatrixDDP) for data resiliency with minimal overhead and reliability that increases as the storage cluster scales.

Eight of the HPE ProLiant DL360 Servers, interconnected with Mellanox 100 Gbps EDR networking and running WekaIO Matrix File system software, are capable of delivering 30 GBps for sequential 1 MB reads and over 2.5 million IOPS for small 4K random reads.[4] The infrastructure is capable of scaling to hundreds of storage nodes in a single namespace.

Guidance: HPE deep learning resources

HPE provides multiple resources for designing and benchmarking AI architectures:

- The HPE Deep Learning Cookbook delivers benchmarking standardization and insights from deep learning workloads.

[1] NVIDIA data sheet, "NVIDIA Tesla V100 GPU Accelerator," March 2018.
[2] NVIDIA NVLink, "NVLink Fabric, A Faster, More Scalable Interconnect," December 2017.
[3] HPE architecture guide, "Architecture guide for HPE servers and WekaIO Matrix," June 2018.
[4] Testing was performed with WekaIO Matrix v3.1.6.

- The HPE Deep Learning Benchmarking Suite is an automated benchmarking tool used to collect performance measurements on various solution configurations in a unified, consistent way.
- The HPE Deep Learning Performance Guide is a knowledgebase of benchmarking results. It enables querying and analysis of measured results as well as performance prediction based on analytical performance models. Reference solution configurations are also available for selected workloads.

Benchmark architecture

Hardware

A single HPE Apollo 6500 Gen10 system with eight NVIDIA Tesla V100 SXM2 16 GB GPUs was used as the testbed for running training and inference workloads. Two storage configurations were tested to highlight the contrast between storage requirements under training and inference scenarios, both using TFRecords of ImageNet for the data set.

1. A single NVMe SSD local to the HPE Apollo 6500 Gen10 system, using the XFS file system.

> **Note**
> A single drive was used to provide a comparison of external storage and local storage behavior. It is not meant to illustrate a production-ready local storage configuration, or to show a recommended DL storage solution.

2. A cluster of eight HPE ProLiant DL360 Gen10 Servers running WekaIO Matrix and containing a total of 32 NVMe SSDs, using the Matrix POSIX client. The HPE Apollo 6500 is connected to this cluster via Mellanox 100 Gbps EDR InfiniBand.

The NVMe SSDs within the WekaIO cluster are the same as the NVMe drive used locally within the Apollo 6500. More details of hardware under test are covered in Appendix A.

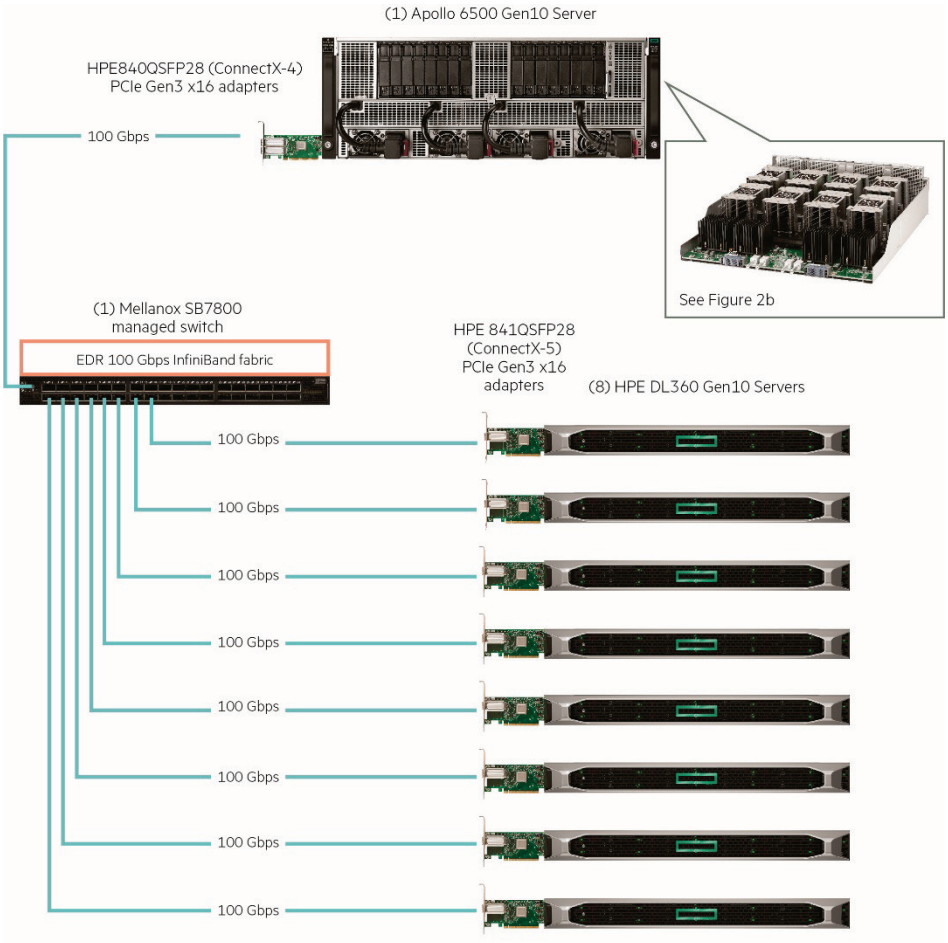


Figure 2a. Benchmark architectural diagram

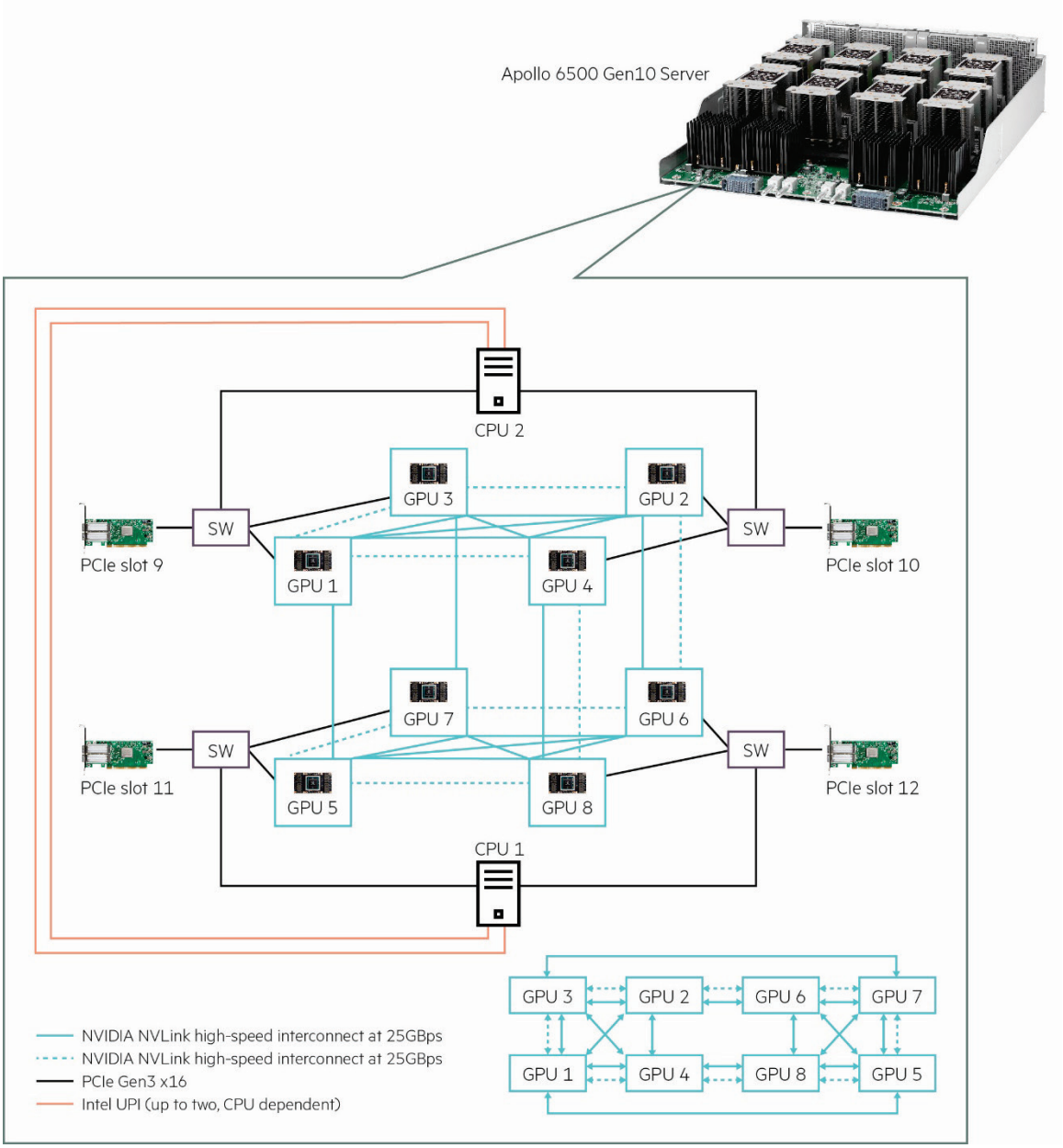


Figure 2b. NVIDIA Tesla V100 GPU topology with NVLink

Software

A modified ImageNet data set was used for inference tests, where the images were decomposed to tensors of 16-bit floating point format, with each tensor file containing 20,000 images in an attempt to maximize throughput.

TensorRT 3.0.4 was used as the inference runtime within the HPE Deep Learning Benchmarking Suite, with some custom code that enables the use of pinned memory. Pinning memory enables better performance by minimizing extra data copying to enable the data to be sent to the GPUs, and enables testing to be NUMA-aware, further optimizing performance.

To reproduce the training tests, the Docker container from the NVIDIA GPU Cloud Deep Learning Platform must be built. See the documentation provided by NVIDIA on their platform for detailed instructions on how to use this tool. See Appendix B for more detailed versioning documentation.

To reproduce the inference tests, the Deep Learning Benchmarking Suite would need to be cloned from GitHub, and both Docker and NVIDIA Docker must be installed. See the Deep Learning Benchmarking Suite GitHub page for a more detailed explanation of how to run the tool, and see Appendix B for more detailed versioning documentation.

Performance testing

Significant work has been done at HPE to document expected training performance for common DL models, such as GoogleNet, ResNet, and VGG.[5] The HPE Deep Learning Performance Guide is a central location for HPE DL performance results. Since the models are convolutional neural networks, which are optimized for image recognition, the data set commonly tested against these models is a database of images called ImageNet.

The "Training results" and "Training analysis" sections of this paper provide a more detailed explanation of data that is already available on the HPE Deep Learning Performance Guide webpage.

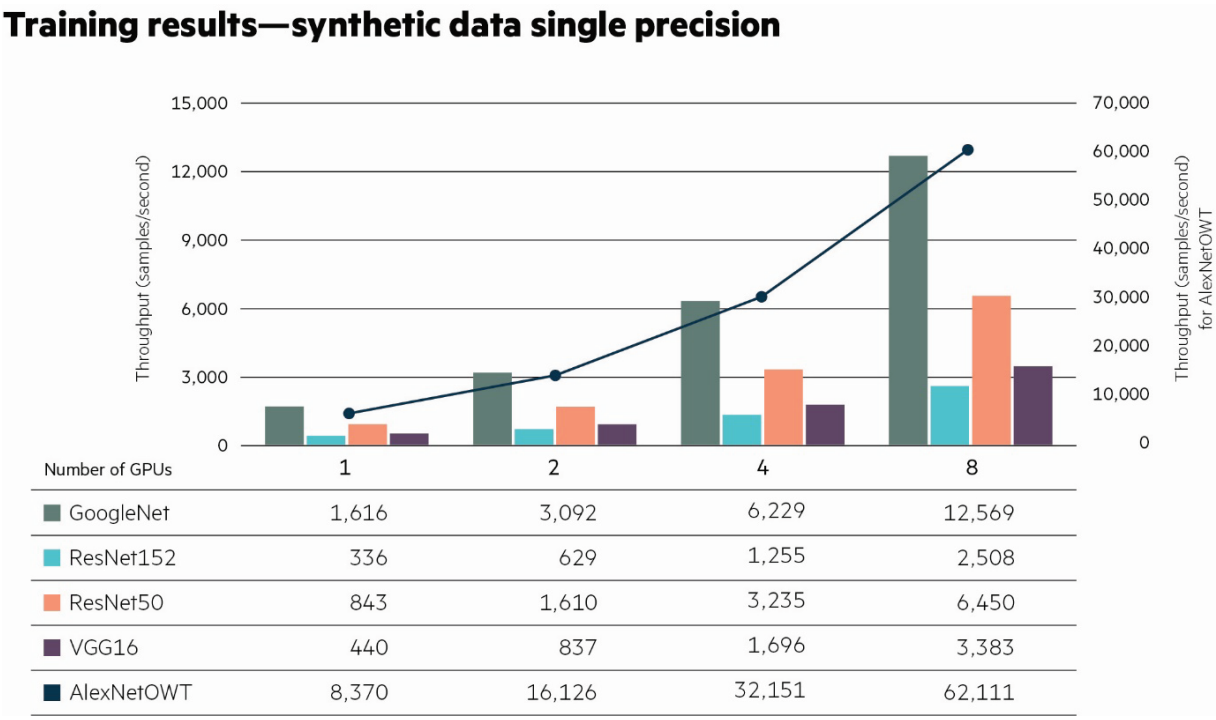
This testing was designed to examine and quantify the differences in performance between local storage running on an NVMe drive in the HPE Apollo 6500, and external network attached storage (NAS) utilizing the WekaIO Matrix file system. A suite of benchmarks was completed on the local NVMe drive inside the HPE Apollo 6500 and repeated utilizing the external storage system. Tests were conducted for both the training portion of the machine learning workload as well as inference validation. The benchmark tests were conducted on one, two, four, and eight NVIDIA V100 GPUs to understand how storage performance was impacted as the workload scaled. The results were gathered and compared between local NVMe SSD performance and the external shared storage.

Training results

To determine if storage can be a bottleneck for training a single HPE Apollo 6500 with NVIDIA GPUs, the NVIDIA GPU Cloud Deep Learning Platform was used with real data in an attempt to differentiate between local storage and the WekaIO Matrix shared storage solution. (Refer to Appendix B for more details.)

However, synthetic benchmarks were also run to gauge the best possible performance of the system. With synthetic data being randomly generated to eliminate non-GPU bottlenecks, it is often used in the DL community as an upper bound of performance since it doesn't rely on data preprocessing or fetching. The results below show a nice linear scaling in performance as the number of GPUs are increased.

Note
The AlexNetOWT results are shown as a curve and refer to the scale on the right so the other results can be shown visibly.



Number of GPUs	1	2	4	8
GoogleNet	1,616	3,092	6,229	12,569
ResNet152	336	629	1,255	2,508
ResNet50	843	1,610	3,235	6,450
VGG16	440	837	1,696	3,383
AlexNetOWT	8,370	16,126	32,151	62,111

Figure 3. Synthetic data half precision training results using NVIDIA NGC containers

[5] See hewlettpackard.github.io/dlcookbook-dlbs/#/models/models?id=supported-models for a full list of supported models.

In the real world, performance generally doesn't scale linearly due to a variety of factors, including data movement between CPU and GPU, the limitations of various models and frameworks, and data preprocessing. Data preprocessing, in particular, can be very heavy, and often becomes a bottleneck for training across multiple GPU-enabled servers. WekaIO Matrix presents a shared POSIX file system to the GPU servers to reduce the performance overhead of copying data between multiple nodes, delivering the performance to utilize all GPU resources without straining I/O capacity.

HPE chose TFRecords of the popular ImageNet data set to enable reproducible results. Numerous tests were performed with various batch sizes, but only the batch sizes yielding the highest performance numbers for each of the five tested models are included in the results presented in Table 1. The following results were achieved using mixed precision.

Table 1. Training benchmark batch sizes

	ResNet152	ResNet50	GoogleNet	VGG16	AlexNet
Batch size	128	256	128	128	1024

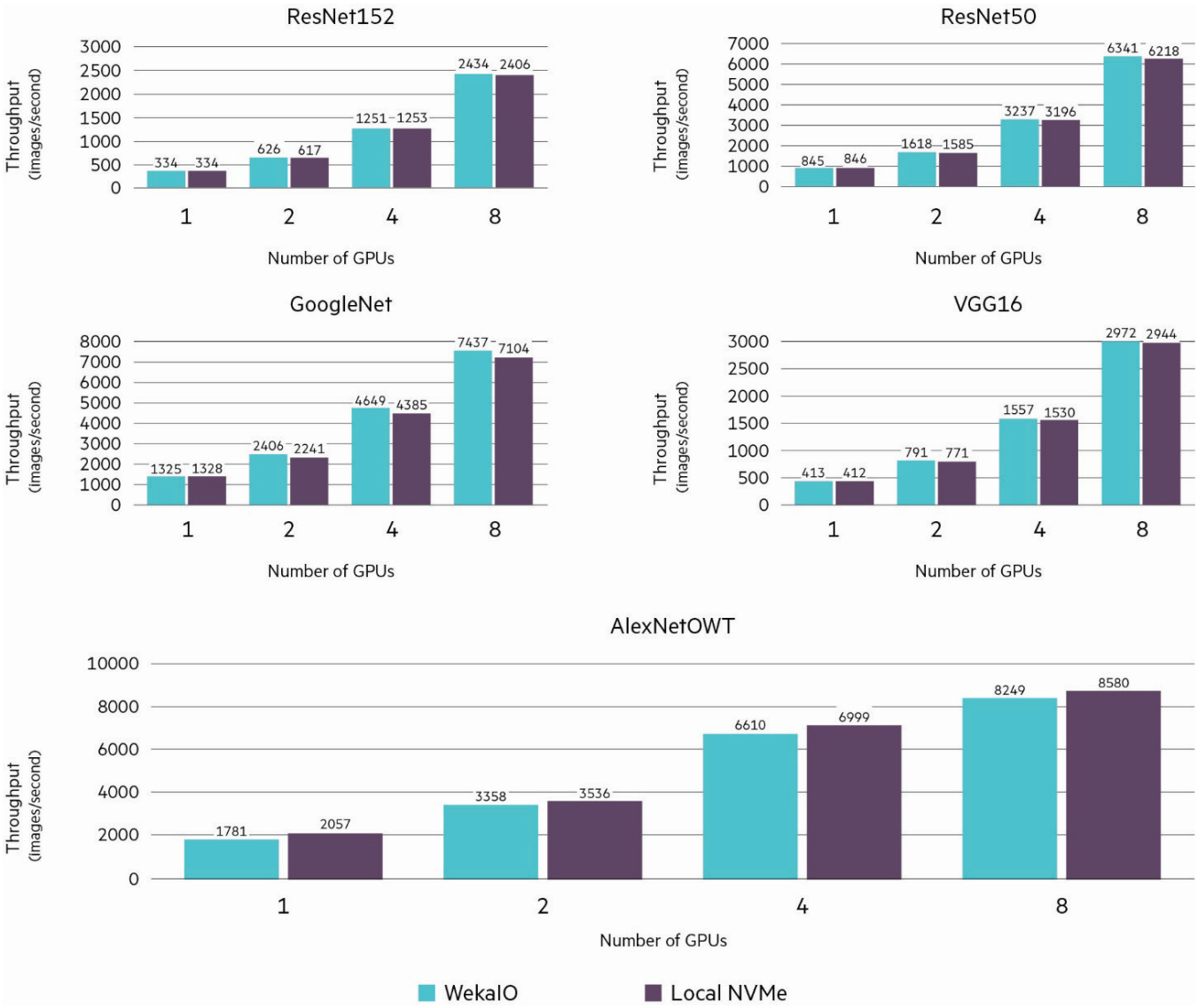
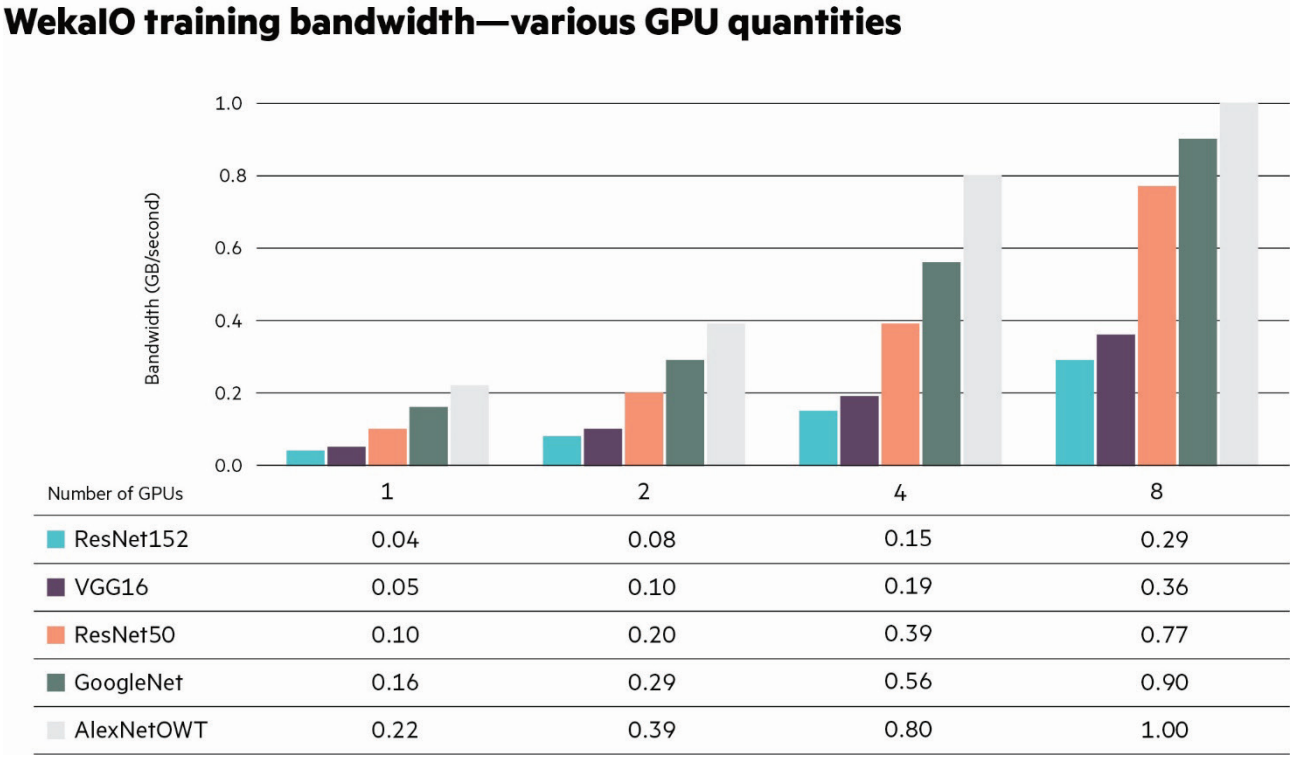


Figure 4. Training benchmark results using NVIDIA NGC containers—WekaIO versus local NVMe drives



Number of GPUs	1	2	4	8
ResNet152	0.04	0.08	0.15	0.29
VGG16	0.05	0.10	0.19	0.36
ResNet50	0.10	0.20	0.39	0.77
GoogleNet	0.16	0.29	0.56	0.90
AlexNetOWT	0.22	0.39	0.80	1.00

Figure 5. Training model bandwidth results with WekaIO

Training analysis

The training data shows that the HPE Apollo 6500 with eight NVIDIA SXM2 16 GB Tesla V100 GPUs was able to fully utilize the available GPU resources while easily addressing the I/O demands of the various benchmarks. With the average 121 KB size of a single JPEG image in the ImageNet data set, the maximal throughput of 8,249 images per second for AlexNetOWT using eight GPUs equates to a read bandwidth of approximately 1.0 GBps—outlined in Figure 5—which is easily achievable by the underlying storage systems. With more complex models, such as the popular ResNet50, the bandwidth is limited to just 0.77 GBps, and in extreme cases, such as with ResNet152, bandwidth is less than 0.29 GBps.

Comparisons between the local NVMe drive and WekaIO system, shown in Figure 4, demonstrate that the WekaIO shared file system delivers comparable performance to a local NVMe drive and, in all but one test, WekaIO is faster than the local file system when scaled to four or eight GPUs. There has been a lot of emphasis on storage I/O requirements for training, but as our results show, for current standard benchmark models and benchmark data sets to a single HPE Apollo 6500 with NVIDIA Tesla GPUs, storage I/O requirements are relatively low on a per-client basis.

However, while the benchmark system was not I/O bound, it should be noted that this local NVMe configuration is not recommended as a production-level DL solution, as it does not provide any data protection. It is intended to provide a comparison between a simple, local storage setup and the WekaIO external storage solution in a single node configuration. In real production environments, it is reasonable to assume that data sets will grow larger than can be stored on a local drive, model sizes will increase in complexity, and GPU servers will be clustered and scaled out instead of just scaling up as demonstrated in this report.

What these results show is that WekaIO and Mellanox networking should minimize the dependency on data locality as a way to improve storage performance. The combination of the WekaIO flash-optimized architecture and the Mellanox InfiniBand EDR network demonstrate performance as fast—or faster than—local NVMe storage and should satisfy production-level machine learning environments without the need for a local data copy.

Inference

Inference is a process that typically occurs at the edge after a trained model has been deployed into production. As such, it does not require communication with the storage infrastructure used to train the neural network. However, inference is also used to validate the model during training, enabling more informed tuning of the neural network to improve performance or accuracy. In the latter use case, storage and computational devices have a bigger impact on the overall performance of a DL application. For that reason, the validation use case was tested using the HPE Deep Learning Cookbook to understand the impact of I/O on the overall model training time.

Inferencing results

HPE tested five different DL models against both local NVMe and MatrixFS. A consistent batch size was used for both the local NVMe and WekaIO MatrixFS testing, as presented in Table 2. Ten warmup batches were used, followed by 400 batches, the results of which were averaged into the data presented in Figure 7. The testing framework enables specifying the number of threads that prefetch data from storage, as well as the inference queue depth that the test framework works to maintain. These tests specify 13 prefetch threads and an inference queue depth of 32 with mixed precision.

Since inference is driven by data processing and movement, synthetic benchmarks were not used in this round of testing, as real-data benchmarks were expected to be more indicative of real-world results. With ongoing testing and tuning, the benchmarks shown here should continue to show improvements in the future.

Table 2. Inference benchmark batch sizes

	ResNet152	ResNet50	GoogleNet	VGG16	AlexNet
Batch size	128	256	128	128	1024

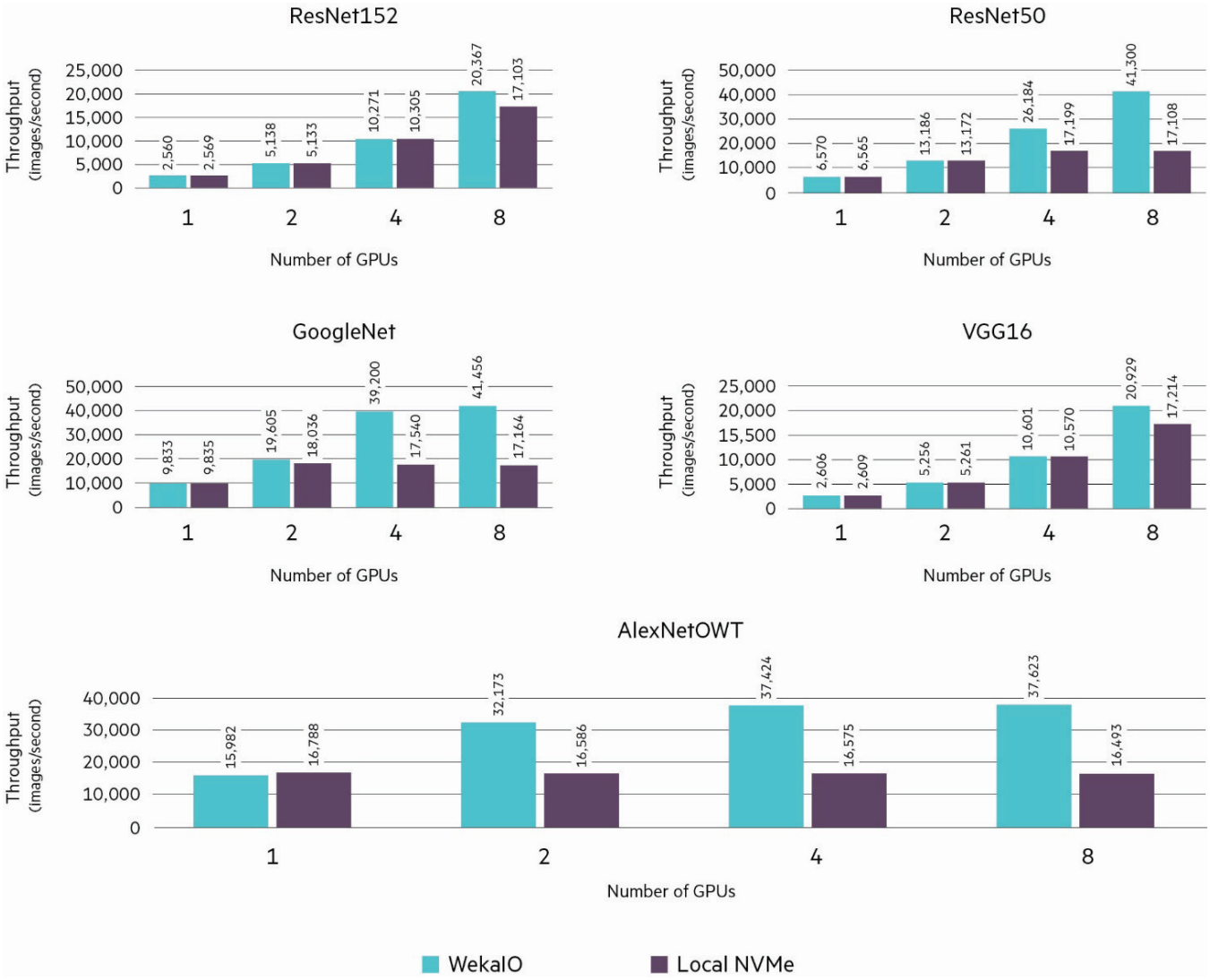
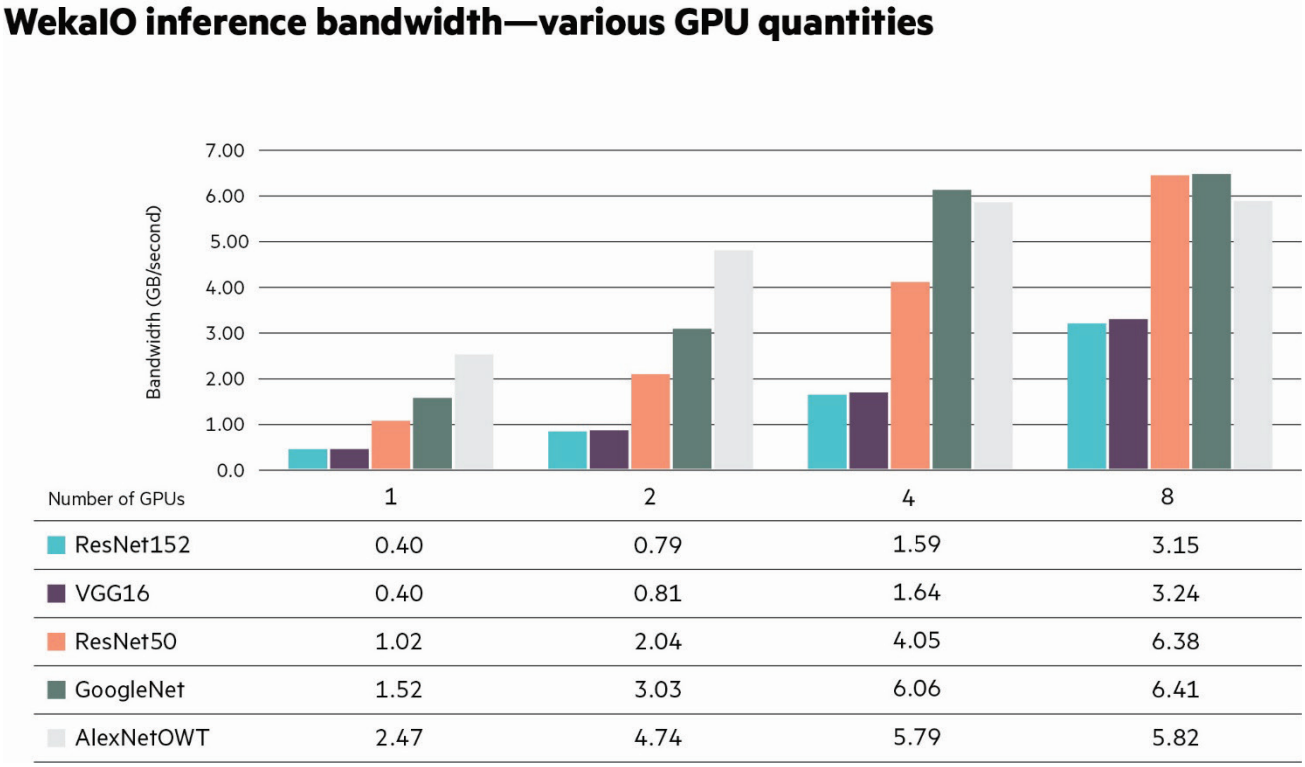


Figure 6. Inference benchmark results—WekaIO versus local NVMe drive



Number of GPUs	1	2	4	8
ResNet152	0.40	0.79	1.59	3.15
VGG16	0.40	0.81	1.64	3.24
ResNet50	1.02	2.04	4.05	6.38
GoogleNet	1.52	3.03	6.06	6.41
AlexNetOWT	2.47	4.74	5.79	5.82

Figure 7. Inference model bandwidth results with WekaIO

Inference analysis

The inference benchmark results show that the I/O demands to a single HPE Apollo 6500 with NVIDIA Tesla GPUs can extend well beyond the capability of the local NVMe drive and that I/O can definitively be a bottleneck. Comparing local NVMe drive performance with shared external storage, as seen in Figure 6, the local NVMe drive became I/O-bound when the benchmarks scale to four GPUs and beyond. For very high image-per-second throughput targets, WekaIO clearly outstrips the local NVMe drive, with both GoogleNet and ResNet50 exceeding 40,000 images per second, over twice as fast as the local drive could deliver. For deeper networks like VGG16 and ResNet152, WekaIO still provided performance superior to the local NVMe device, peaking at over 20,000 images per second when using eight GPUs for inference.

The results show that WekaIO Matrix with HPE ProLiant DL360 Gen10 Servers is capable of providing extremely high bandwidth data movement to support optimized inference pipelines. Figure 6 provides an overview of the model bandwidth measurements between the WekaIO enabled shared storage system and the HPE Apollo 6500. At its peak, the benchmark results for eight GPUs demonstrate read I/O of 6.4 GBps to the single HPE Apollo 6500 system. The network interconnect is critical to achieving the high bandwidth measured during the testing. The Mellanox 100 Gbps InfiniBand EDR network can deliver over 12 GBps bandwidth, allowing the solution to take advantage of the parallelism from the external storage system, while in contrast, the internal drives were I/O limited.

With the images processed and transformed prior to being used for inference, the CPU no longer has a significant role in the data pipeline, maximizing the time spent on both the GPUs and the storage devices, and increasing the time spent making inferences on the input data. Once storage performance requirements dramatically increase in the inference case, the single local NVMe device is—as expected—no longer able to keep up with the Matrix cluster. This contrast further highlights that more inference work can be done when a data pipeline is setup to take advantage of higher storage performance.

Conclusion: Reduce overall AI development time

The benchmarking results shared in this paper show that the current AI benchmark models are compute bound versus I/O bound for training, and that a shared storage solution can easily service the I/O requirements to these models. When it comes to inference benchmarks, the I/O requirements were significantly more demanding, resulting in an I/O bottleneck when benchmarks were run on local NVMe drives, while the shared storage solution delivered over twice the performance at scale. Real-world training models are significantly more demanding, with far larger data sets and more complex models, and would benefit from a shared storage solution that can deliver equivalent to or better than local file system performance.

The HPE Apollo 6500 Gen10 System shows linear performance scaling from one to eight NVIDIA Tesla V100 GPU processors when I/O to the server is not bottlenecked. When the solution became I/O bound during the inference training stage, the local NVMe drive solution peaked, while I/O continued to scale from the storage cluster. The combination of WekaIO Matrix with parallel I/O to the cluster of NVMe drives and Mellanox 100 Gbps InfiniBand interconnect provided the network bandwidth to service the I/O demands of the inference workload.

WekaIO Matrix offers a level of performance that meets or exceeds a local file system as the number of GPUs scale, and quite a bit higher than traditional NFS file systems. HPE expects this will deliver even greater advantage when real training data sets increase in capacity and clusters scale-out beyond a single GPU client. Given that every training cycle needs to be followed by a validation cycle, similar to a software quality assurance (QA) cycle, the advantages of WekaIO are multiplied many times over, and can significantly reduce the overall time to create a production-ready DL model.

HPE provides the tools and expertise to guide the creation of DL solutions, including GPU-enabled servers, shared storage and networking, and DL application expertise. The HPE Deep Learning Cookbook enables clear, reproducible benchmarking for the AI solution space, and guidance with neural network models, data format, and solution architectures to quickly create effective DL applications.

Appendix A: Hardware configuration

This section contains a detailed description of the hardware components, SKUs, and quantities used for the benchmarks. It does not cover all components in a full solution order—such as service and support, or factory configuration options—and is intended only as an accurate representation of the testbed hardware.

Figure 8. HPE Apollo 6500 Gen10 system

Table 3. HPE Apollo 6500 Gen10 system configuration

Component name	Quantity	SKU
HPE XL270d Gen10 Node CTO Server	1	P00392-B21
HPE XL270d Gen10 Xeon-Gold 6150 FIO Processor Kit	1	P01278-L21
HPE XL270d Gen10 Xeon-Gold 6150 Processor Kit	1	P01278-B21
HPE 16 GB 2Rx8 PC4-2666V-R Smart Memory Kit	12	835955-B21
HPE 1.6 TB NVMe x4 Lanes Mixed Use SFF (2.5"), SCN 3-year warranty, digitally signed firmware SSD (extended)	1	877994-B21
HPE DL38X Gen10 Premium 8 SFF/SATA Bay Kit	1	826690-B21
HPE XL270d Gen10 NVMe FIO Enable Kit	1	P01056-B22
HPE 6+2 NVMe Instr Spec FIO	1	878192-B21
HPE Apollo PCIe/SATA M.2 FIO Riser Kit	1	863661-B21
HPE InfiniBand EDR/Ethernet 100 Gbps 2-port 841QSFP28 Adapter	1	872726-B21
HPE 2200 W Platinum Hot Plug Power Supply Kit	4	P01062-B21
HPE 2.0 m 250 V 16 A C19-C20 WW Single IPD Enabled Jumper Cord	4	TK738A
HPE XL270d Gen10 8 SXM2 GPU FIO Module	1	P01786-B22
HPE XL270d Gen10 SXM2 Heat Sink FIO Kit	2	P02939-B22
HPE NVIDIA Tesla V100 SXM2 16 GB Computational Accelerator	8	Q2N66A

Table 4. HPE ProLiant DL360 Gen10 Server configuration

Component name	Quantity	SKU
HPE DL360 Gen10 Premium 10 NVMe CTO Server	8	867960-B21
HPE DL360 Gen10 Intel Xeon-Gold 6134 (3.2 GHz/8-core/130 W) FIO Processor Kit	8	860683-L21
HPE DL360 Gen10 Intel Xeon-Gold 6134 (3.2 GHz/8-core/130 W) Processor Kit	8	860683-B21
HPE 8 GB (1 x 8 GB) Single Rank x8 DDR4-2666 CAS-19-19-19 Registered Smart Memory Kit	96	815097-B21
HPE 800 W Flex Slot Titanium Hot Plug Low Halogen Power Supply Kit	16	865438-B21
HPE InfiniBand EDR 100 Gbps 1-port 841QSFP28 Adapter	8	872725-B21
HPE DL360 Gen10 SATA M.2 2280 Riser Kit	8	867978-B21
HPE 240 GB SATA 6G Mixed Use M.2 2280, 3-year warranty, digitally signed firmware SSD	16	875488-B21
HPE 1.6 TB NVMe x4 Lanes Mixed Use SFF (2.5") SCN, 3-year warranty, digitally signed firmware SSD	32	877994-B21

Appendix B: Benchmarking software documentation

Table 5. Software documentation for training tests. See ngc.nvidia.com for access to this and later containers

Category	Value
Operating system	Ubuntu 16.04.4 LTS
WekaIO MatrixFS version	3.1.6
Framework	TensorFlow 1.8.0
Container	nvcr.io/nvidia/tensorflow:18.07-py3

Table 6. Software documentation for inference tests

Category	Value
Operating system	Ubuntu 16.04.4 LTS
WekaIO MatrixFS version	3.1.6
Framework	TensorRT 3.0.4 GA
Container	dlbs/tensorrt:18.10

Resources

HPE WekaIO Matrix product page
hpe.com/storage/wekaio

HPE Deep Learning solutions
hpe.com/info/deep-learning

HPE Deep Learning Cookbook
developer.hpe.com/platform/hpe-deep-learning-cookbook/home

WekaIO Matrix product page
weka.io

HPE/NVIDIA Alliance page
hpe.com/us/en/solutions/hpc-high-performance-computing/nvidia-collaboration.html

NVIDIA Volta architecture
nvidia.com/en-us/data-center/volta-gpu-architecture/

NVIDIA Tesla
nvidia.com/en-us/data-center/tesla/

NVIDIA GPU Cloud
nvidia.com/en-us/gpu-cloud/

Mellanox Technologies
mellanox.com

Learn more at
hpe.com/storage/wekaio

Our solution partners

NVIDIA

Mellanox TECHNOLOGIES

WEKA.IO

Sign up for updates

a00056398ENW, October 2018