WEKA

# Weka AI Reference Architecture with NVIDIA DGX A100 Systems

Scaling Deep Learning Performance with Weka Software and Industry Standard Servers

In partnership with

NVIDIA

## CONTENTS



## LIST OF TABLES



## LIST OF FIGURES

# EXECUTIVE SUMMARY

Organizations of various sizes, use cases, and technical skills are looking for infrastructure solutions to accelerate their artificial intelligence (AI), machine learning (ML), and deep learning (DL) initiatives. WekaIO™ (Weka) and NVIDIA® partnered to architect and validate a high-performance scalable AI solution accessible to everyone. This document contains validation information for the Weka AI™ reference architecture (RA) solution. The design was implemented using up to four NVIDIA DGX™ A100 systems, and NVIDIA® Mellanox® Spectrum™ Ethernet and NVIDIA Mellanox Quantum™ InfiniBand switches. The operation and performance of this system was validated by NVIDIA and Weka using industry-standard benchmark tools. Based on the validation testing results, this architecture delivers excellent linear scaling for training workloads. Organizations can start small and easily and independently scale compute and storage resources to multi-rack configurations with predictable performance to meet any ML workload requirement.

# INTENDED AUDIENCE

This document is for Weka partners, solutions engineers, and customers. It describes the architecture used to determine the specific equipment, cabling, and configurations required to support the validated workload.

# Weka AI SOLUTION

The Weka AI RA, powered by NVIDIA DGX A100 systems and Weka's industry leading file system WekaFS™, was developed and verified by Weka and NVIDIA. It gives IT organizations an architectural framework that significantly reduces the time to productivity by eliminating the integration complexity of multiple infrastructure components. Weka AI simplifies DL and AI deployments by combining WekaFS based-storage systems with DGX A100 systems and NVIDIA networking, to create a tightly integrated solution that minimizes time to production. Weka AI can start as small as 50 TB of capacity and scale seamlessly to exascale in a single namespace, while effortlessly managing data across the edge to the core to the cloud.

The WekaFS system performance has currently been verified with up to four DGX A100 systems. By adding additional storage nodes, the architecture can grow to support many more DGX A100 systems while maintaining linear performance scalability. With the addition of any Amazon Simple Storage Service (S3) compliant object storage, WekaFS can expand the global namespace to support a massive data lake. The Weka AI design improves the flexibility to tune for performance scalability and/or capacity scalability with independent levers for compute and storage based on the size of the data lake, and the performance needs of the DL training models.

# DL DATA PIPELINE

DL data pipelines consist of two overlapping paths, one for model training and the other for production inference. The training workflow, which is the focus of this paper follows a series of steps as shown in Figure 1. DL model accuracy is highly influenced by the size of the data set and the depth of the neural network, more complex models lead to better outcomes.

Training is the most computationally intensive path, involving many computations to train a model from specified input data, once trained, the model must undergo regression testing via large-scale inference against a known data set, comparing the model to the desired outcome. In this final step, the production inference path uses ingested normalized input data to generate a conclusion, a much lighter computational requirement than the iterative back-propagation of training.

Many of the DL models generate massive data sets with a continuous feedback loop to improve the model accuracy. The infrastructure has to cater to the full end to end data training pipeline as outlined below. Data is ingested at the edge from devices such as motor vehicles, manufacturing lines or other IoT devices, it is fed to the core training models and ultimately to the cloud for data reproducibility and responsible AI.

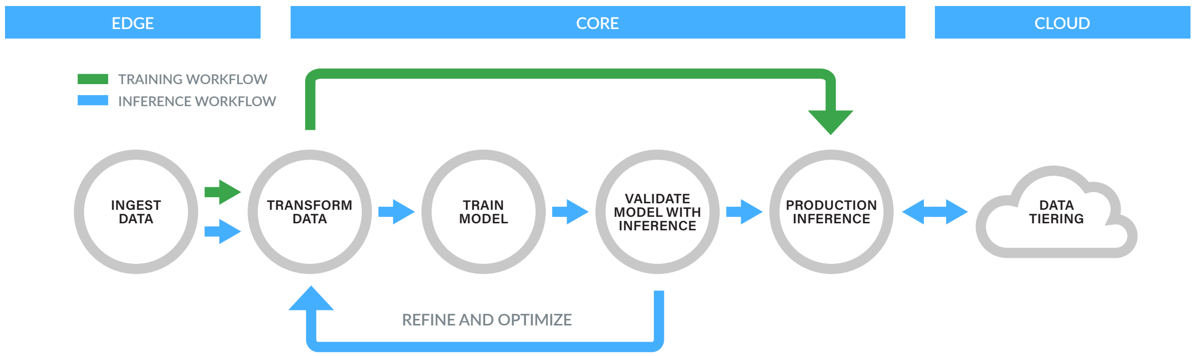


*Figure 1 – DL edge-core-cloud data pipeline*

The following list describes some of the activities that occur in one or more of these areas.

1. **Data Ingest** into the DL pipeline. An example would be images and associated metadata. A typical I/O profile consists of small file sequential reads, often followed by a large number of small or large file sequential writes that are aggregated for increased throughput.

2. **Transform Data** to clean and store input data in a normalized form. For image data, this can consist of image decoding and various conversions in size and/or color, often performance is limited by server host CPUs.

   There is a secondary stage of data transformation during the 'refine and optimize' loop, when some noise/jitter is added to the images to increase the versatility and resiliency of the trained models. For example, if training for autonomous driving, the transformation phase may add rain or fog over the images.

3. **Train Model** by passing each dataset member through the model. Each data item perturbs coefficients set at each layer, biasing it towards a future conclusion. Supervised training is the most common method, with each input accompanied by a desired output value. Unsupervised training examples lack a provided output value, requiring the model to organize the data based on common characteristics e.g. grouping similar items into a cluster. For example, when training for autonomous driving, the supervised training is going to be using labeled images that mark the different key elements in each frame (other cars, pedestrians, traffic lights, etc).

4. **Model Validation with Inference** using input data that the model hasn't yet seen. Inference based on this data helps to ensure accuracy and avoid false positives. High error rates indicate the need for further training. I/O in this stage can be characterized as large file sequential reads. The file sizes are like the ones used during the training phase. The I/O pattern during this stage can be characterized as very high amount of sequential file reads. For example, when training for autonomous driving, at this phase the infrastructure will try to simulate driving through as many miles as possible. This step is the most tasking on the storage infrastructure, as more well driven miles equates higher quality of the model. This is also the phase with the highest level of parallelism as more independent "virtual cars" can be simulated. It becomes critical for the storage system to support more concurrent I/O workloads, resulting in direct improvement of the reliability of the model.

5. **Production Inference** occurs when a trained and validated model is used in a production environment. Input data is often aggregated into larger files containing many items reduced to short tensor values to maximize throughput. For the autonomous driving example, the GPU inferencing occurs in the car itself on live image feed from the surrounding car cameras.

6. **Data Tiering** and bursting to the cloud provides data mobility and provisioning, by allowing users to create another file namespace instantly in the public clouds. This can be used for data and DL model backup, DR and cloud bursting for on-demand GPU compute. Long-term persistent copy of data models may be required for regulation or responsible AI.

Throughout the DL data pipeline, the I/O pattern from the GPU infrastructure to the storage system varies significantly, placing significant demands on the file storage. Data sets can range from billions of tiny files to thousands of very large files (e.g. video streams). The storage system must be capable of handling any of the demands thrown at it including:

- Simultaneous support for tens to hundreds of GPU systems
- Ability to read millions of files concurrently at lowest latency
- Scale to billions of files in a single directory
- Massively parallel data access
- Seamless scaling of namespace to Petascale and beyond
- End-to-end encryption for data security, while maintaining optimal performance
- Data management including integration with containers for data mobility
- Cloud integration for data lifecycle management and on-demand hybrid cloud

Weka AI delivers a software-defined storage solution that leverages industry-standard server architectures and full cloud integration to support the demands of a DL pipeline, from edge to core to cloud. This RA documents WekaFS as an on-premises solution utilizing dedicated servers to deliver best of breed performance. The same software is capable of delivering the solution in the public cloud on NVMe™-enabled instances, or it can be deployed in a converged model across a multiple of DGX A100 systems.

## SOLUTION OVERVIEW

The Weka AI RA leverages the WekaFS file system software, accelerated computing hardware and software from NVIDIA, networking from NVIDIA, and shared servers from HPE. This solution was tested and benchmarked scaling from one to four DGX A100 systems with the Weka storage solution. For production environments, the number of servers managed by the Weka file system can scale significantly beyond the entry RA configuration of eight nodes to meet the specific needs of the DL workload. Additional storage nodes can be dynamically added to scale performance. This RA was validated with an eight node WekaFS storge solution, four DGX A100 systems, and two NVIDIA Mellanox Spectrum SN3700 200 GbE switches. Figure 2 shows the basic solution architecture.

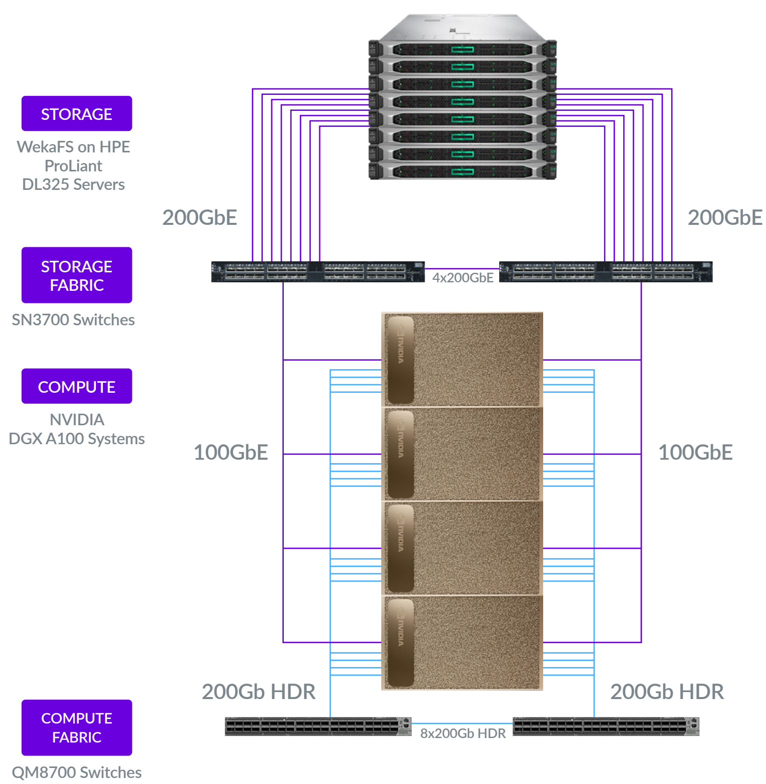


*Figure 2 – Weka AI validated architecture with DGX A100 systems*

# NVIDIA DGX A100 SYSTEMS

The DGX A100 system is universal system for AI workloads—from analytics to training to inference and HPC applications. A DGX A100 system contains eight NVIDIA A100 Tensor Core GPUs, with each system delivering over 5 petaFLOPS of DL training performance. The eight GPUs within a DGX system A100 are interconnected in a hybrid cube-mesh topology using the next generation NVIDIA NVLink™ technology which doubles the GPU-to-GPU direct bandwidth to 600 gigabytes per second (GB/s), and a new NVIDIA NVSwitch™ chip that is 2X faster than the last generation. The DGX A100 system also features eight single-port Mellanox ConnectX®-6 VPI HDR InfiniBand adapters for clustering and one dual-port ConnectX-6 VPI Ethernet adapter for storage.

The DGX A100 system is configured with the DGX software stack including the system operating system and NVIDIA-optimized DL containers for maximum GPU acceleration. NVIDIA provides access to its NVIDIA NGC™ library of containers offers fully optimized versions of popular DL frameworks including TensorFlow™, PyTorch and Caffe2, for maximized GPU-accelerated performance. Each container has everything necessary to run supported applications in a highly performant yet portable fashion and includes the necessary drivers, libraries, communication primitives and other software, enabling users to avoid lost productivity associated with engineering their own AI software stack.

# Weka FILE SYSTEM (WekaFS)

This RA leverages WekaFS on commercially available NVMe servers. Performance testing was carried out on the HPE ProLiant DL325 server platform with WekaFS. An entry starting cluster size requires eight server nodes for full availability with the ability to survive up to a two-node failure. Each node has CPU, NVMe storage and high bandwidth networking. The exact configuration for the RA is detailed in the Technology Requirements section. The cluster can be easily scaled to thousands of nodes.

**Performance at Scale**
WekaFS is the world's fastest and most scalable POSIX compliant parallel file system, designed to transcend the limitations of legacy file systems that leverage local storage, NFS, or block storage making it ideal for data-intensive AI and HPC workloads. WekaFS is a clean sheet design integrating NVMe-based flash storage for the performance tier to the GPU servers, object storage and ultra-low latency interconnect fabrics such as 200 GbE or InfiniBand into an NVMe-over-Fabrics architecture, creating an extremely high-performance scale-out storage system. WekaFS performance scales linearly as more servers are added to the storage cluster allowing the infrastructure to scale with the increasing demands of the business.

**Multi-Protocol Ready**
In addition to POSIX access, WekaFS also supports all the common file access protocols including NFS, SMB and S3 for maximum compatibility and interoperability. Hadoop and Spark environments also benefit from the performance of a shared file system through a fully integrated connector that allows WekaFS to replace HDFS and function as a single, easy to manage data lake for all forms of analytics.

**Expandable Global Namespace over S3 Object Store**
WekaFS delivers best of breed performance from NVMe flash tier, and the namespace can expand to any S3 object store, on premises or in the cloud. This optional hybrid storage model with the ability to expand the global namespace to lower cost hard disk drives in an object store delivers a cost-effective data lake without compromising performance. The integrated tiering to multiple S3 targets enables the cost-effective data lifecycle management for older or less used training data.

**Advanced Durability and Security**
Large and agile datasets in AI/ML frequently require a data versioning capability. This is achieved using Weka's instant and space-efficient snapshots capability for experiment reproducibility and explain-ability. The snap-to-object feature captures a point-in-time copy of the entire, unified [flash and object store] file namespace that can be presented as another file namespace instance in a private or public cloud. Weka's integrated snapshots and end-to-end encryption features, with key management integration, ensure data is always backed up and secure throughout its lifecycle. WekaFS also provides immutability and data mobility for these datasets with instant recovery. WekaFS can seamlessly back up to multiple cloud targets providing backup, DR and data governance capability.

**Cloud Bursting and Data Mobility**
In addition to providing versioning, Weka's snap-to-object feature provide additional benefits beyond backup and DR to the public

cloud, it enables secure data portability from on-premises to the public cloud for organizations that require access to on-demand GPU resources in the public cloud.

**Container Support**

Organizations are increasingly adopting containers deployed on Kubernetes (K8s) for AI workloads. Using the WekaFS K8s CSI plug-in, organizations now have flexibility in how and where they deploy containerized applications. It provides easy data mobility from on-premises to the cloud and back, while delivering best storage performance and latency. Figure 3 provides an overview of WekaFS in a typical production deployment.

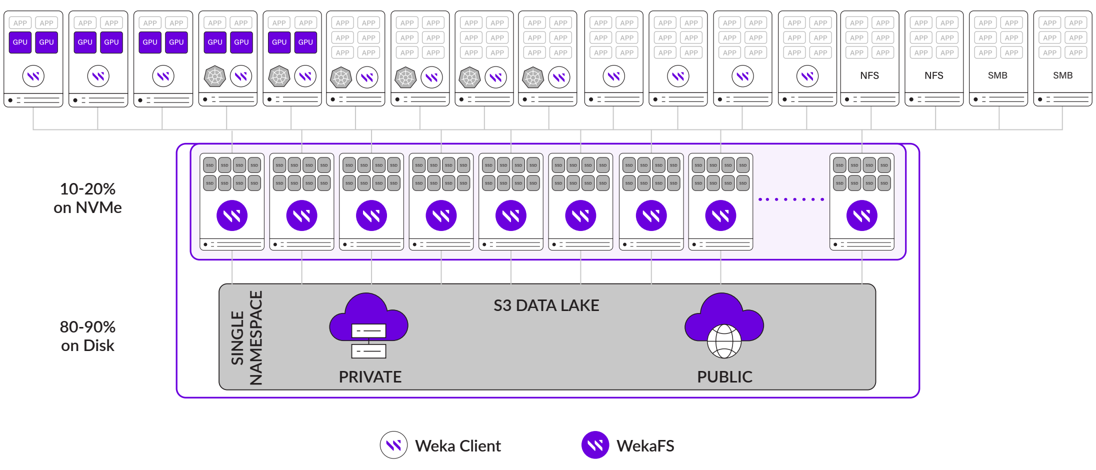


*Figure 3 – Typical WekaFS production deployment*

# NVIDIA NETWORKING

## NVIDIA MELLANOX SPECTRUM AND QUANTUM SWITCHES

Networking is a critical part of the DL infrastructure that is responsible for moving massive amounts of data between the end points efficiently and effectively. In any good system design, it is important to maximize the performance of the most critical (and often the most expensive) component. In the case of DL infrastructures, the performance of the specialized compute elements such as GPUs must be maximized. The Network is the critical part of the infrastructure that determines the overall system performance. With consistent performance, intelligent load balancing and comprehensive telemetry are an ideal network element for DL workloads.

**Consistent Performance**

The NVIDIA Mellanox Spectrum SN3700 switches coupled with ConnectX adapters leverage hardware-accelerated end-to-end congestion management to provide a robust data path for RDMA over Converged Ethernet (RoCE) based NVIDIA GPUDirect® traffic. Mellanox Spectrum Ethernet switches support high bandwidth storage traffic with fair and predictable performance.

**Intelligent Load Balancing**

Mellanox Spectrum switches support Adaptive Routing (AR) to maximize cross-sectional bandwidth in data center fabrics. AR leverages end-to-end congestion notification mechanisms to maximize the cross-sectional bandwidth of fabric. Maximizing cross-sectional bandwidth reduces the remote access-related performance penalty that often limits scale-out system performance.

**Hardware Accelerated Visibility**

Mellanox Spectrum provides detailed and contextual telemetry to answer the "When, What, Who, Where and Why" questions as soon as an issue happens. Hardware-accelerated histograms track and summarize queue depths at a sub-microsecond granularity. This avoids false-alerts common to simple watermarks/thresholds.

**In-Network Computing**

The NVIDIA Mellanox Quantum QM8700 switch improves the performance of DL operations by processing data as it traverses the network and eliminating the need to send data multiple times between endpoints. The combination of Mellanox In-Network Computing SHARP with NVIDIA A100 Tensor Core GPU technology and Collective Communications Library (NCCL) deliver leading efficiency and scalability to DL and AI applications.

## TECHNOLOGY REQUIREMENTS

This section covers the hardware and software that was used for all the testing described in the Solution Verification Section.

### Hardware Requirements

Table 1 lists the hardware components that were used to verify this solution.

| HARDWARE | QUANTITY |
|---|---|
| DGX A100 systems | 4 |
| HPE ProLiant DL325 Gen10 Plus Servers | 8, includes per server AMD EPYC™ 7402 CPU, 128GB RAM, 7 x 15.3TB Micron 9300 Pro NVMe SSDs per server, 2 x 200Gb/s ConnectX6 NVIDIA Mellanox NICs |
| SN3700 Ethernet switch (storage fabric) | 2 |
| QM8700 InfiniBand switch (compute fabric) | 2 |

*Table 1 – Hardware requirements*

### Software Requirements

Table 2 lists the software components that were used to validate the solution.

| SOFTWARE | VERSION |
|---|---|
| WekaFS | 3.9.0 |
| Server OS | CentOS 8.2 |

| SOFTWARE | VERSION |
|---|---|
| DGX OS | 4.99.9 |
| Docker container platform | 19.03.8 |
| Container version | nvcr.io/nvidia/mxnet:20.06-py3 – Mlperf test tensorflow:20.05-tf2-py3 – other tests |
| OFED version | OFED-internal-5.0-2.1.8 |

*Table 2 – Software requirements*

## SOLUTION ARCHITECTURE

This RA has been tested and validated to meet the I/O demands for DL workloads. The end-to-end solution outlined below, provides insight to the base-line storage and networking requirement to meet to the performance measured in this RA.

**Network Configuration**
The solution tested for this RA consisted of four NVIDIA DGX A100 systems connected to two NVIDIA Mellanox SN3700 Ethernet switches with two 100 Gb/s network connections per DGX A100 system. Each NIC was connected to a separate NVIDIA Mellanox SN3700 switch for a total of eight network links to the storage system. Each DGX A100 system also had eight single-port Mellanox ConnectX- 6 200Gb/s HDR InfiniBand ports (also configurable as 200Gb/s Ethernet ports) for inter-GPU system communication.

The HPE ProLiant DL325 Gen10 Plus servers were each configured with two ConnectX-6 200 Gb/s Ethernet NICs. Each NIC was connected to a separate NVIDIA Mellanox SN3700 switch for a total of 16 100Gb/s network links.

The network switch was configured with a message transmission unit (MTU) size of 9000.

Weka AI RA does not require RoCE to be configured on any portion of the design and does not require priority flow control (PFC) to be configured on the network switch, greatly simplifying the network deployment.

## SOLUTION VERIFICATION

This RA was validated using synthetic benchmark utilities and DL benchmark tests to establish baseline performance and operation of the system. Each of the tests described in this section was performed with the specific equipment and software listed in the Technology Requirements section. The tests outlined in Table 3 were performed with one, two, three, and four DGX A100 systems to validate basic operation, scalability and performance of the deployed infrastructure.

**Tests Performed**

| TEST NAME | SOURCE CODE |
|---|---|
| NVIDIA NCCL all_reduce_perf test | https://github.com/NVIDIA/nccl-tests |
| FIO bandwidth and IOPS test | https://fio.readthedocs.io/en/latest/fio_doc.html#source |
| Mdtest for metadata performance | https://github.com/hpc/ior |
| MLPerf Training – ResNet-50 | https://github.com/mlperf/training |

*Table 3 – Tests completed for this RA*

The following sections describe details and results for each of these tests.

**NVIDIA NCCL all_reduce_perf Test**
The NVIDIA Collective Communications Library (NCCL) tests the maximum scalability across multiple DGX A100 systems. Within a system, the bottleneck should be the bandwidth of the NVIDIA NVLink high-speed interconnect. Across multiple systems, the bottleneck should be from whatever InfiniBand or RoCE-enabled Ethernet adapters are assigned for GPU-to-GPU communication across DGX A100 systems.

The results in Figure 4 shows the single system inter-GPU bandwidth reaches NVLink interconnect capabilities. Multi system inter-GPU bandwidth reaches aggregate bandwidth of all InfiniBand or Ethernet adapters assigned to the test.

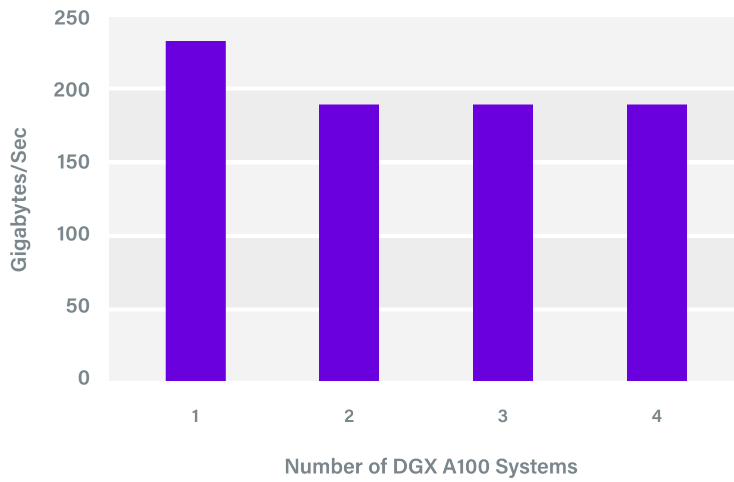


*Figure 4 – NCCL bandwidth test results*

**FIO Bandwidth and IOPS Tests**

The FIO benchmark provides a baseline I/O testing to measure the maximum raw I/O performance capability of the combined RA. The storage system I/O was tested with FIO (Flexible I/O), a versatile I/O workload generator that can simulate different I/O patterns according to use case. Two separate I/O configurations were measured, the first to measure maximum bandwidth from the storage system and the second measures the maximum I/O operations per second (IOPS). Both configurations were run using 100% reads and 100% writes.

Results of the test in Figure 5 demonstrate the linear read performance scalability as more DGX A100 systems were added to the test. Write performance reaches near maximum performance at a single DGX A100 system as the storage system's aggregate NVMe SSD write throughput was quickly reached. Similarly, the results in Figure 6 shows the linear scalability of Read throughput IOPS, while the Write throughput IOPS hit near maximum performance at two DGX A100 systems. This is consistent with what is expected from the limited number of drives in the configuration rather than a limit of WekaFS, the HPE DL325 and DGX A100 systems.

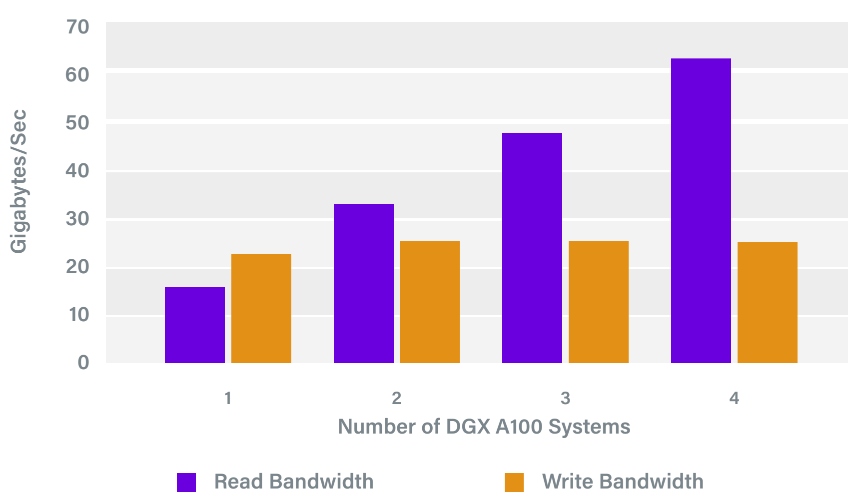


*Figure 5 – FIO bandwidth test results*

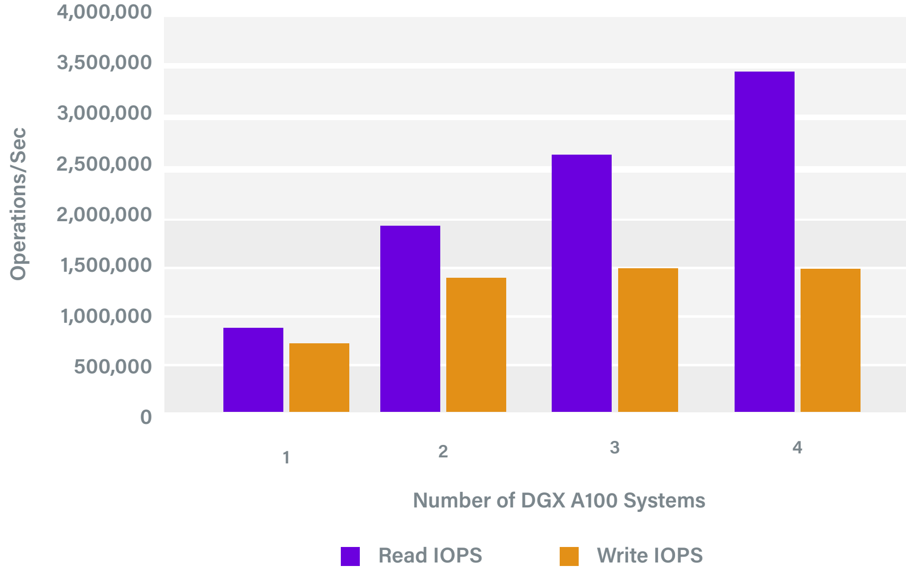


*Figure 6 – FIO IOPS test results*

**mdtest**

mdtest is a file system metadata performance test designed to run in a message passing (MPI) cluster environment with a parallel file system. In each iteration of the test, each MPI task creates, stats, and removes the specified number of directories and/or files and measures the performance in ops/second. After all the iterations complete, the maximum, minimum, mean ops/sec and the std. deviation are reported for each operation. The results in Figure 7 again, demonstrates the performance scaling of WekaFS as the number of DGX A100 systems increase. These results were measured on the minimum Weka configuration, of eight storage nodes, as more nodes are added to the storage cluster, metadata will scale proportionally.

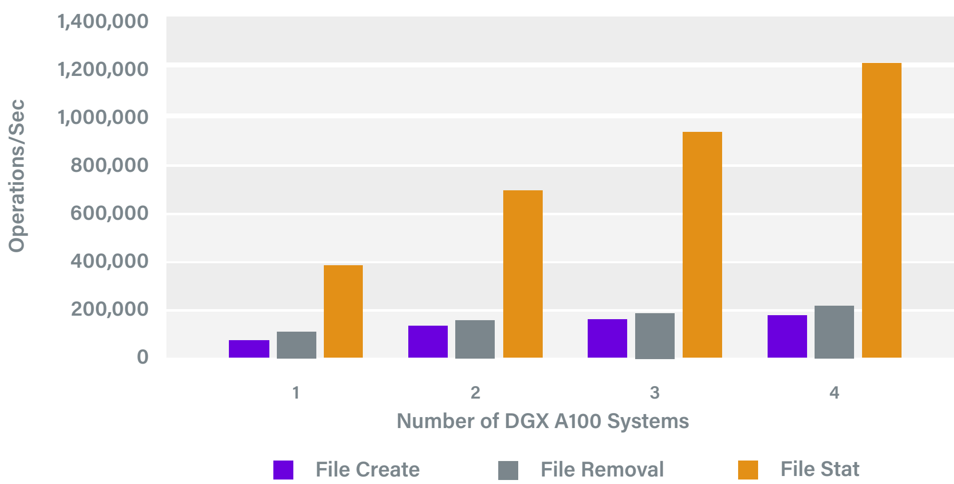


*Figure 7 – mdtest results for File Creation, File Removal and File Stat*

**MLPerf ResNet-50 Test**

MLPerf is the industry standard set of benchmark implementations of neural networks. Focus is on the ResNet-50 neural network as it is a well-known image classification network that can be used with the ImageNet dataset. It is sufficiently computationally intensive, yet capable of driving meaningful storage I/O.

Configurations were tested with one, two, and four DGX A100 systems. Slurm, NVIDIA Pyxis, and Enroot software were used to coordinate the work across all DGX A100 systems involved in the task. The MLPerf v0.7 implementation of ResNet-50 provided by NVIDIA for this RA was built with the MXNet framework. The training data is ImageNet, formatted using RecordIO. The results presented maintain a consistent batch size per system of 408 images as the workload is scaled (weak scaling). DALI is configured to not use mmap.

The training benchmark was measured at Epoch 0 and compared to the overall run average time, which provides insight into the ability of the storage system to keep up with the read bandwidth demands of a complete training job. Epoch 0 is the most I/O intensive portion of the MLPerf benchmark run and will have the biggest impact on time to insight. The results shown in Figure 8 demonstrate that the storage system was able to provide the same images/second for Epoch 0 as overall average, validating that the storage system is not the bottleneck for the workload. The results also demonstrate the linear scalability of WekaFS; as more DGX A100 systems are added, performance scales linearly with the additional compute power.

A concern of many infrastructure architects is that system utilization will deteriorate as more GPU systems are utilized to accelerate the training pipeline. The results in figure 8 also demonstrate that WekaFS delivers linear scaling of time to insights as more DGX A100 systems are added to the workload. The time to insights almost halved from 41 minutes to 22 minutes going from one to two DGX A100 systems, and again to 12 minutes when scaled to 4 systems. In summary, with no I/O bottleneck, time to insight improved linearly with almost 4 times faster results, going from a single DGX A100 system to four DGX A100 systems.

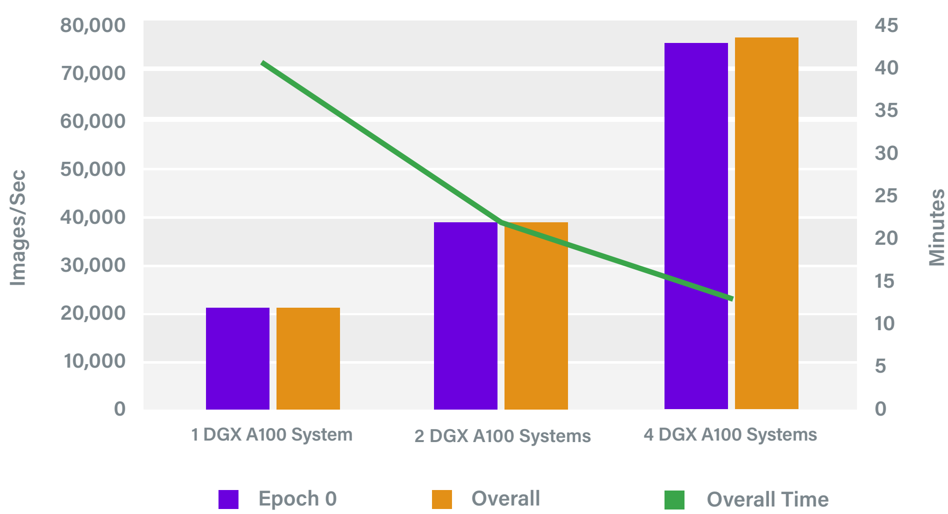


*Figure 8 – MLPerf ResNet-50 images per second and overall time*

# Weka AI SOLUTION SCALING

The testing conducted in this RA leverage HPE Proliant DL325 storage servers. The minimum configuration of eight storage nodes was used to perform the testing, and the servers were partially populated, each with seven NVMe drives. The results demonstrate that this configuration is capable of delivering 64 GB/second read performance and 3.54 million 4K IOPS. The Weka software can scale to hundreds of storage nodes and performance scales linearly as additional nodes are added to the system.

# CONCLUSION

AI, fueled by rapid innovation in DL solutions, is becoming common place in a wide range of industries. Organizations who invest in AI and turn their data into intelligence and new products will lead their competition. While many organizations want to kickstart their AI initiatives, challenges building a scalable and AI-optimized infrastructure often hold them back. Traditional compute infrastructures are not suitable for demanding AI workloads due to slow legacy CPU architectures and varying system requirements. This drives up complexity, increases cost, and limits scale. Engineers at Weka and NVIDIA partnered to architect a scalable and powerful infrastructure that pushes the boundaries of AI innovation and performance.

The results show robust linear performance scalability from one to four DGX A100 systems, allowing organizations to start small and grow seamlessly as AI projects ramp. The results demonstrate that scaling GPU infrastructure to accelerate time to insights will be well supported by Weka AI. The validated Weka AI RA makes it easy for teams to focus on developing new products and gain new faster insights with AI/ML.

# ADDITIONAL INFORMATION

To learn more about the information that is described in this document, review the following resources:

| PRODUCT AND TECHNOLOGY | |
|---|---|
| **Weka File System** | |
| WekaFS Datasheet | https://bit.ly/2K8s44W |
| WekaFS Architecture White Paper | https://bit.ly/3f9uOdH |
| **NVIDIA DGX A100** | |
| NVIDIA DGX A100 System | https://www.nvidia.com/en-us/data-center/dgx-a100/ |
| NVIDIA A100 Tensor core GPU | https://www.nvidia.com/en-us/data-center/a100/ |
| **NVIDIA Mellanox Networking** | |
| NVIDIA Mellanox Spectrum SN3000 | https://www.mellanox.com/products/ethernet-switches/sn3000 |
| NVIDIA Mellanox Quantum QM8700 | https://www.mellanox.com/products/infiniband-switches/QM8700 |
| **Machine Learning Frameworks** | |
| TensorFlow | https://www.tensorflow.org/ |
| Horovod | https://eng.uber.com/horovod/ |

WEKA 910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io



W01R3RA201911