



NVIDIA DGX A100 Systems (BasePOD)

Scaling Deep Learning Performance with WEKA
Software and Industry Standard Servers.

REFERENCE ARCHITECTURE

Contents

Tested Solution Details	3
1. Introduction	4
Audience	4
Purpose	4
Document Version History	4
2. WEKA Data Platform	5
Performance at Scale	5
Multi-Protocol Ready	5
Expandable Global Namespace over S3 Object Store	5
Advanced Durability and Security	5
Cloud Bursting and Data Mobility	6
Container Support	6
3. NVIDIA DGX Architecture	6
NVIDIA A100 GPU	7
Multi-Instance GPU (MIG)	7
Third Generation NVLink and NVSwitch	8
Mellanox ConnectX-6	9
Base Command	9
4. Using NVIDIA DGX BasePOD with WEKA	10
5. Solution Design	11
Design Decisions	11
Network	12
6. Validation and Benchmarking	13
7. Conclusion (Getting started)	17
8. Appendix	18

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.

This design was implemented using [NVIDIA DGX BasePOD](#) reference architecture with NVIDIA DGX™ A100 systems. The operation and performance of this system were validated by NVIDIA and WEKA using industry-standard benchmark tools.

This architecture delivers excellent linear scaling based on the validation testing results for training workloads. Organizations can start small and quickly and independently scale compute and storage resources to multi-rack configurations with predictable performance to meet any ML workload requirement.

Tested Solution Details

Product Name	Product Version	WEKA Version	Source	WEKA OS
NVIDIA DGX	3.99.9	3.x	On-Premises	CentOS

1. Introduction

Audience

We wrote this reference architecture for those who design, manage, and support WEKA. Consumers of this document should already be familiar with NVIDIA DGX and the WEKA® Data Platform.

We have organized this document to address critical items for enabling successful design, implementation, and transition to operation.

Purpose

This document covers the following subject areas:

- Overview of the WEKA solution.
- Overview of NVIDIA DGX.
- The benefits of NVIDIA DGX with WEKA.
- Recommendations for architecting a complete NVIDIA DGX solution on the WEKA platform.
- Benchmarks for NVIDIA DGX performance with WEKA.

Document Version History

Version Number	Published	Notes
1.0	September 2021	Original publication.
1.1	September 2022	Updated format and BasePOD

2. WEKA Data Platform

The WEKA Data Platform is deployed on commercially available NVMe servers. Performance testing was carried out on the HPE ProLiant DL325 server platform with WEKA. An entry starting cluster size requires eight server nodes for full availability with the ability to survive up to a two-node failure. Each server has a 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

At the core of the WEKA Data Platform is WekaFS is the world's fastest and most scalable POSIX-compliant parallel file system. It is 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 100GbE and 400GbE or InfiniBand into an NVMe-over-Fabrics architecture, creating a highly 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 supports all the standard 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 the 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 develop 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 explainability. 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. With crucial management integration, Weka's integrated snapshots and end-to-end encryption features ensure that 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 offers 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 the best storage performance and latency. Figure X provides an overview of WekaFS in a typical production deployment.

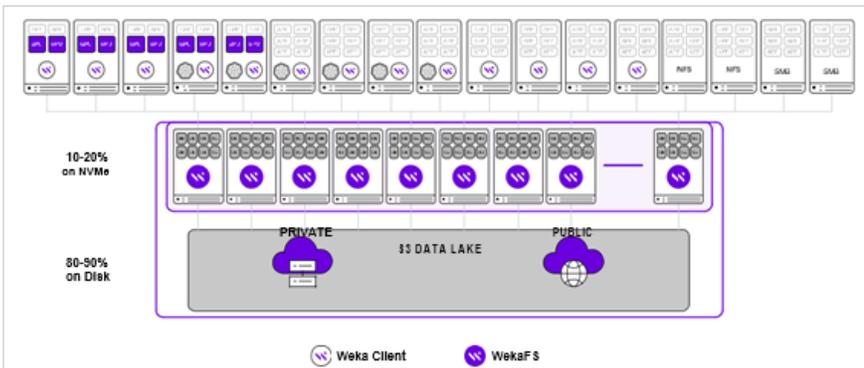


FIG. 1 WekaFS

3. NVIDIA BasePOD Architecture

Figure X shows an exploded view of the significant components in the NVIDIA DGX A100 system.



FIG. 2 DGX A100 components

NVIDIA A100 Tensor Core GPU

At the core, the NVIDIA DGX A100 system leverages the NVIDIA A100 GPU (Figure X), designed to accelerate large, complex, efficient AI workloads and several small workloads, including enhancements and new features for increased performance over the NVIDIA V100 GPU. The A100 GPU incorporates up to 80 gigabytes (GB) of high-bandwidth HBM2 memory and more extensive and faster caches and is designed to reduce AI and HPC software and programming complexity.

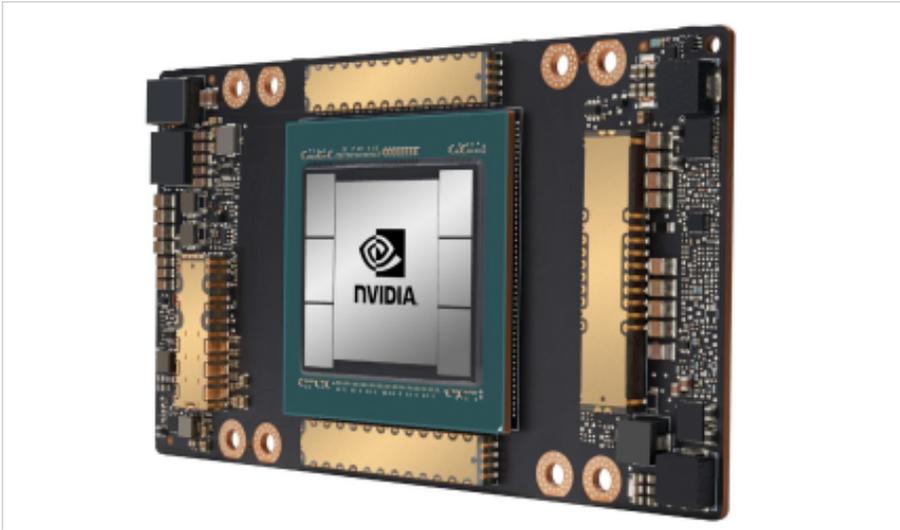


FIG. 3 NVIDIA A100 GPU

The NVIDIA A100 GPU includes new features to accelerate AI workload and HPC application performance further.

- Third-generation Tensor Cores
- Fine-grained Structured Sparsity
- Multi-Instance GPU

You can find additional [product information](#) online.

Multi-Instance GPU (MIG)

The NVIDIA A100 GPU incorporates a new partitioning capability called Multi-Instance GPU (MIG) for increased GPU utilization. MIG uses spatial partitioning to carve the physical resources of a single A100 GPU into as many as seven independent GPU instances. With MIG, the NVIDIA A100 GPU can deliver the guaranteed quality of service at up to 7 times higher throughput than V100 with simultaneous instances per GPU.

On an NVIDIA A100 GPU with MIG enabled, parallel compute workloads can access isolated GPU memory and physical GPU resources as each GPU instance has its memory, cache, and streaming multiprocessor. This allows multiple users to share the same GPU and run all instances simultaneously, maximizing GPU efficiency.

MIG can be enabled selectively on any number of GPUs in the DGX A100 system - not all GPUs need to be MIG-enabled. However, if all GPUs in a DGX A100 system are MIG enabled, up to 56 users can simultaneously and independently take advantage of GPU acceleration.

Typical uses cases that can benefit from MIG are

- Multiple inference jobs with batch sizes of one that involves small, low-latency models and that doesn't require all the performance of a full GPU
- Jupyter notebooks for model exploration
- Resource sharing of the GPU among multiple users

Third Generation NVLink and Second Generation NVSwitch

The DGX A100 system contains six second-generation NVIDIA® NVSwitch™ fabrics interconnecting the A100 GPUs using third-generation NVIDIA NVLink® high-speed interconnects. Each A100 GPU uses twelve NVLink interconnects to communicate with all six NVSwitches, which means there are two links from each GPU to each switch. This provides maximum bandwidth to communicate across GPUs over the links.

The second-generation NVSwitch [Figure 4] is two times faster than the previous version, which was first introduced in the NVIDIA DGX-2 system. The combination of six NVSwitches and third-generation NVLinks enables individual GPU to GPU communication to peak at 600 GB/s, which means that if all GPUs communicate with each other, the total amount of data transferred peaks at 4.8 TB/s for both directions.



FIG. 4 NVIDIA NVSwitches

NVIDIA ConnectX-6

Multi-system scaling of AI deep learning and HPC computational workloads requires strong communications between GPUs in multiple systems to match the significant GPU performance of each design. In addition to NVLink for high-speed communication internally between GPUs, the DGX A100 is purpose-built for multi-system AI scaling with eight single-port NVIDIA ConnectX-6 200Gb/s HDR InfiniBand ports (also configurable as 200Gb/s Ethernet ports), providing 3.2 Tb/s of peak bandwidth from a single system that can be used to immediately build a high-speed cluster of DGX A100 systems such as NVIDIA DGX SuperPOD.

The most common methods of moving data to and from the GPU involve leveraging the onboard storage and using the NVIDIA ConnectX-6 network adapters through Remote Direct Memory Access (RDMA). The DGX A100 incorporates a one-to-one relationship between the IO cards and the GPUs, which means each GPU can communicate directly with external sources without blocking other GPUs' access to the network.

The NVIDIA ConnectX-6 I/O cards offer flexible connectivity as they can be configured as HDR InfiniBand or 200Gb/s Ethernet. This allows the NVIDIA DGX A100 to be clustered with other nodes to run HPC and AI workloads using low latency, high bandwidth InfiniBand, or RDMA over Converged Ethernet (RoCE).

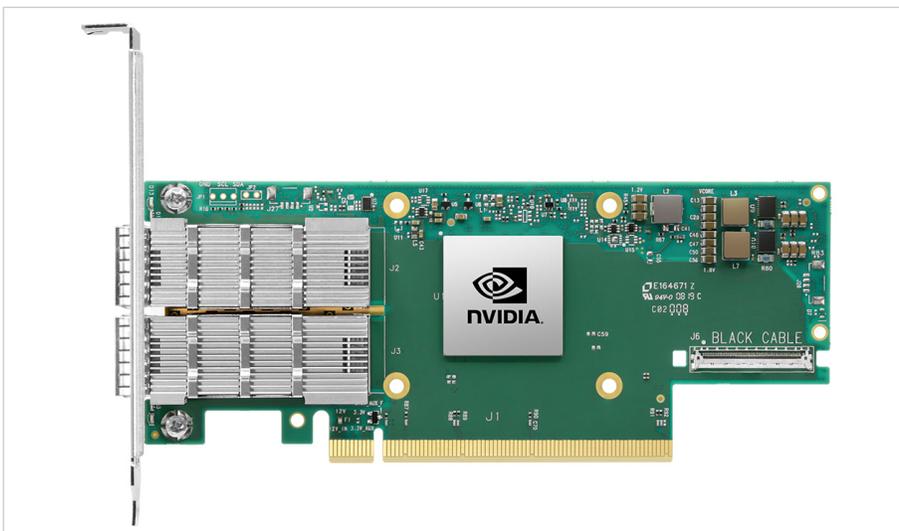


FIG. 5 Mellanox Single-port ConnectX-6

NVIDIA Base Command

NVIDIA Base Command provides enterprise-grade orchestration and cluster management, and it now features a complete software stack for maximizing AI developer productivity, IT manageability, and workload performance. The workflow management features of Base Command enable centralized control of AI development projects with simplified collaboration for project teams and integrated monitoring and reporting dashboards.

Base Command works with the NVIDIA AI Enterprise software suite, which is now included with every DGX system. The NVIDIA AI Enterprise enables end-to-end AI development and deployment with supported AI and data science tools, optimized frameworks, and pretrained models.

Base Command also offers enterprise-workflow management and MLOps integrations with DGX-Ready Software providers like Domino Data Lab, [Run.ai](#), and Weights & Biases. It also includes libraries that optimize and accelerate compute, storage, and network infrastructure — while ensuring maximized system uptime, security, and reliability.

4. WEKA Solutions built on DGX BasePOD

The WEKA Data Platform powered by DGX BasePOD will enable us to extend these benefits to our customers at a much larger scale. Key highlights of the new solution include:

- The WEKA Data Platform and DGX BasePOD are now directly applicable to mission-critical enterprise AI workflows, including natural language processing and larger-scale workloads for customers in the life sciences, healthcare, and financial services industries, among many others. WEKA can efficiently serve large and small files across various workload types.
- WEKA continued innovation and support of Magnum IO GPUDirect Storage technology provides low-latency, direct access between GPU memory and storage. This frees CPU cycles from servicing the I/O operations and delivers higher performance for other workloads.

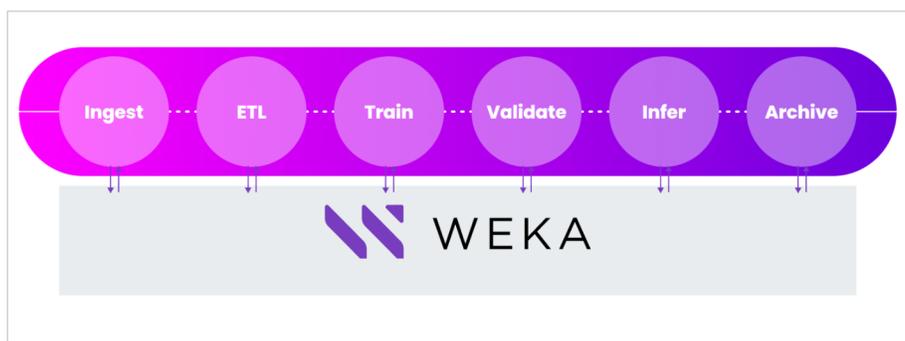


FIG. 6 WEKA Data Platform

5. Solution Design

Design Decisions

The following tables cover design decisions and rationale for DGX BasePOD with WEKA.

General Design Decisions: DGX BasePOD

Item	Details	Rationale
Minimum Size	6 WEKA Servers 2 NVIDIA DGX systems	Minimum size requirements
Scale approach	Incremental, modular scale	Allow for growth from PoC to a massive scale
Scale Units	Servers	Granular scale to precisely meets the capacity demands.

General Design Decisions: Networking

Item	Details	Rationale
Compute/DGX	InfiniBand is recommended	The GPU-to-GPU network needs to be fast
Storage	InfiniBand or 200GB	WEKA recommends 200GB

NVIDIA DGX BasePOD Sizing

AI is powering mission-critical use cases in every industry—from healthcare to manufacturing to financial services. The DGX BasePOD™ reference architecture provides the critical foundation on which business transformation is realized and AI applications are born.

DGX BasePOD is an all-inclusive system that includes the following components:

- NVIDIA – 2 or more DGX A100
- Networking for the DGX compute layer and the storage layer
- Accelerated Storage Fabric

NOTE: *It's always a good practice to add a buffer for contingency and growth.*

Network

The solution tested for this RA consisted of four 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 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 200Gb/s Ethernet NICs. Each NIC was connected to a separate NVIDIA Mellanox SN3700 switch for 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.

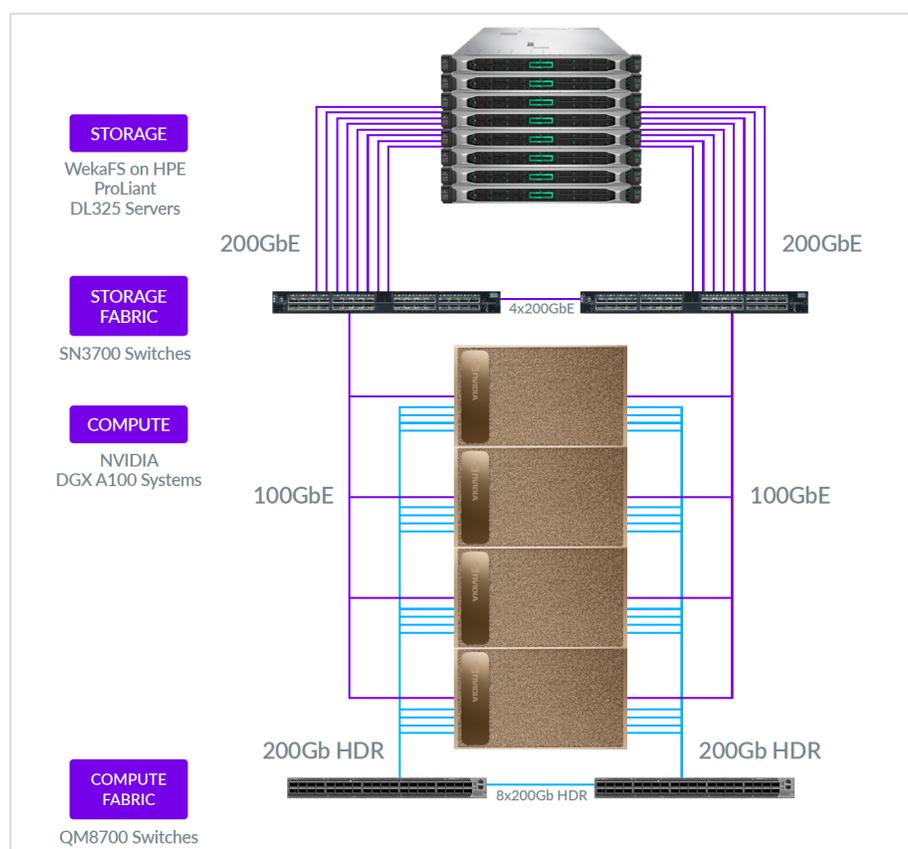


FIG. 7 Network Architecture

6. Validation and Benchmarking

We conducted this validation using synthetic benchmark utilities and DL benchmark tests to establish the baseline performance and operation of the system. Each test 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 the essential operation, scalability, and performance of the deployed infrastructure.

Hardware Requirements

Hardware	Quantity
DGX A100 systems	4
HPE ProLiant DL325 Gen10 Plus Server	8, includes per server AMD EPYC™ 7402 CPU, 132GB RAM, 7 x 15.3TB Micron 9300 Pro NVMe SSDs per server, 2 x 200Gb/s ConnectX6 NVIVIA NICs
SN3700C Ethernet switch (storage fabric)	2
QM8700 InfiniBand switch (compute fabric)	2

Hardware Requirements

Software	Version
WekaFS	3.9.0
Server OS	CentOS 8.2
DGX OS	4.99.9
Docker container platform	
Container version	
OFED version	19.03.8
	nvcv.io/nvidia/mxnet :20.06-py3 – MLPerf test tensorflow:20.05-tf2-py3 – other tests
	OFED-internal-5.0-2.1.8

Benchmarks

Tests Performed

Test Name	Source Code
NVIDIA NCCL all_reduce_perf test	https://github.com/NVIDIA/ncccl-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

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 scenario, the bottleneck should be the NVIDIA NVLink high-speed interconnect bandwidth. Across various methods, 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 show that the single system inter-GPU bandwidth reaches NVLink interconnect capabilities. Multi-system inter-GPU bandwidth goes to the aggregate bandwidth of all IB or Ethernet adapters assigned to the test.

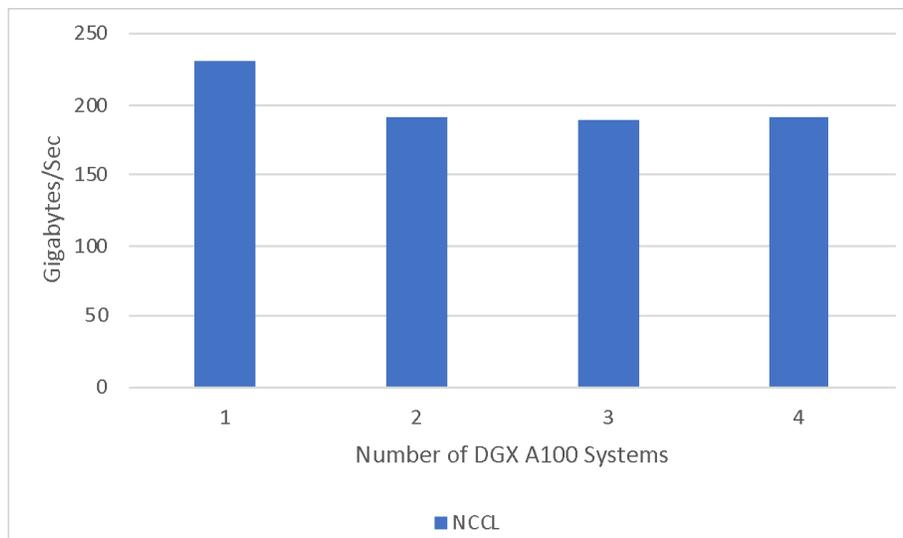


FIG. 8 NCCL bandwidth test results (GB/s)

FIO Bandwidth and IOPS Tests

The FIO benchmark provides a baseline I/O testing to measure the combined RA's maximum raw I/O performance capability. 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 the use case. Two separate I/O configurations were measured the first measured maximum bandwidth from the storage system, and the second measured the maximum I/O operations per second (IOPS). Both designs 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 a maximum at a single DGX A100 system as the storage system's aggregate NVMe SSD write throughput was quickly reached. Similarly, Figure 6 shows the linear scalability of Read throughput IOPS, while the Write throughput IOPS hit a limit at three DGX A100 systems. This is consistent with what is expected from the limited number of drives in the configuration rather than a limitation of WekaFS, the HPE DL325, and DGX A110 systems.

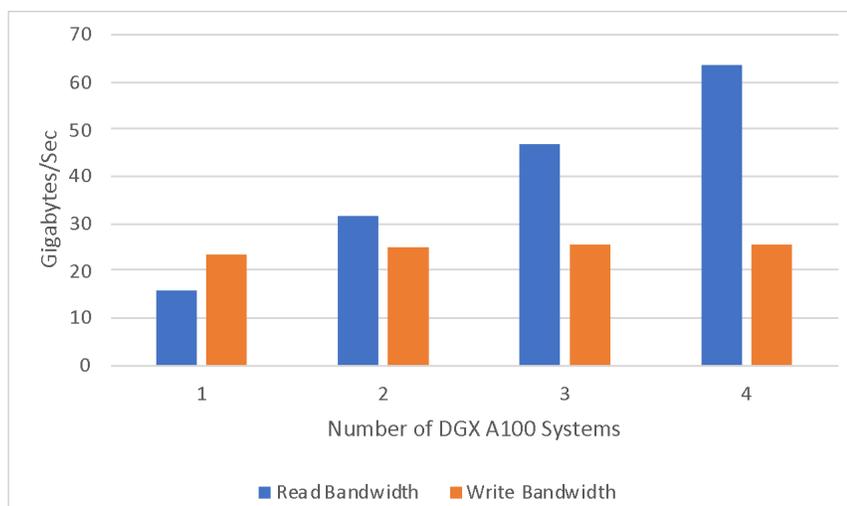


FIG. 9 FIO bandwidth test results (GB/s)

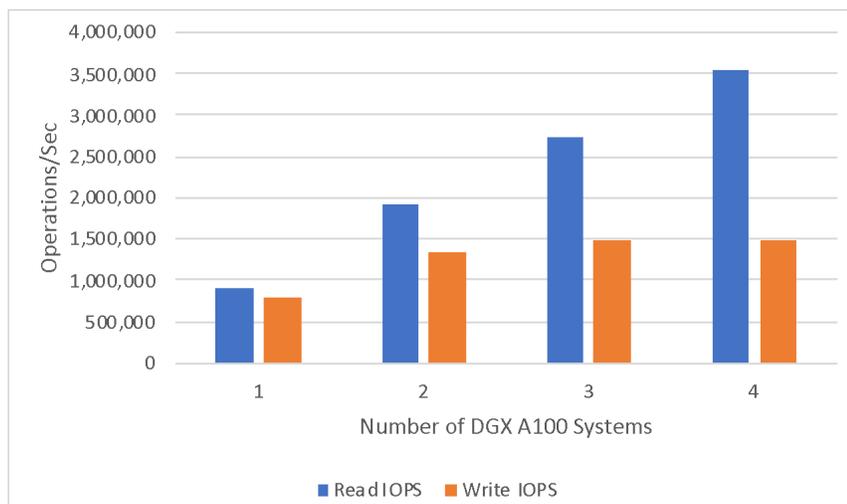


FIG. 10 FIO IOPS test results (operations/s)

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 trial, each MPI task creates stats, removes the specified number of directories and files, and measures the performance in ops/second. After all the iterations are complete, the maximum, minimum, mean ops/sec, and the std. Deviations are reported for each operation. The results in Figure 7 again demonstrate the linear performance scaling of WekaFS as the number of DGX A100 systems increases.

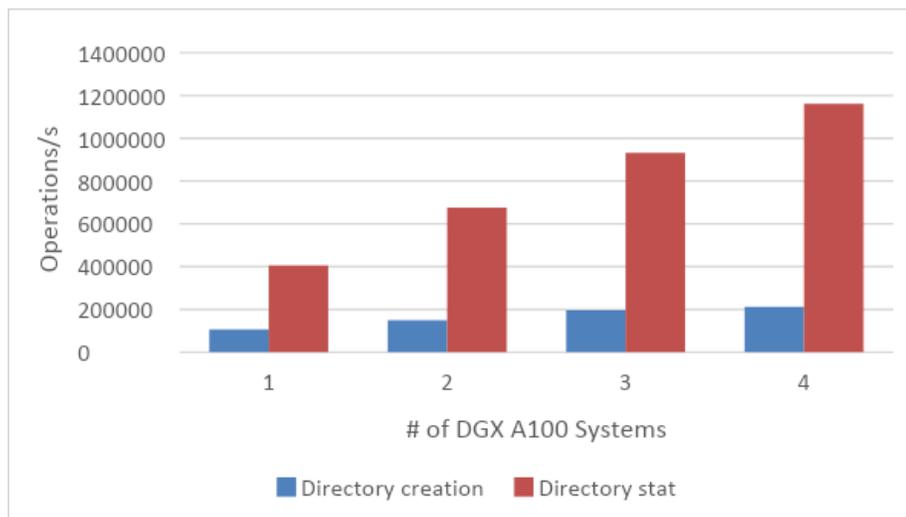


FIG. 11 mdtest results for directory creation and directory stat

MLPerf ResNet-50 Test

MLPerf is the industry standard set of benchmark implementations of neural networks. The focus is on the ResNet-50 neural network, a well-known image classification network that can be used with the ImageNet dataset. It is 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 to maintain a uniform batch size per system of 408 images as the workload is scaled (weak scaling). DALI is configured not to 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 IO-intensive portion of the MLPerf benchmark run and will significantly impact insights on time. The results shown in figure 8 demonstrate that the storage system could provide the exact images/second for Epoch 0 as an 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.

Many infrastructure architects are concerned 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.

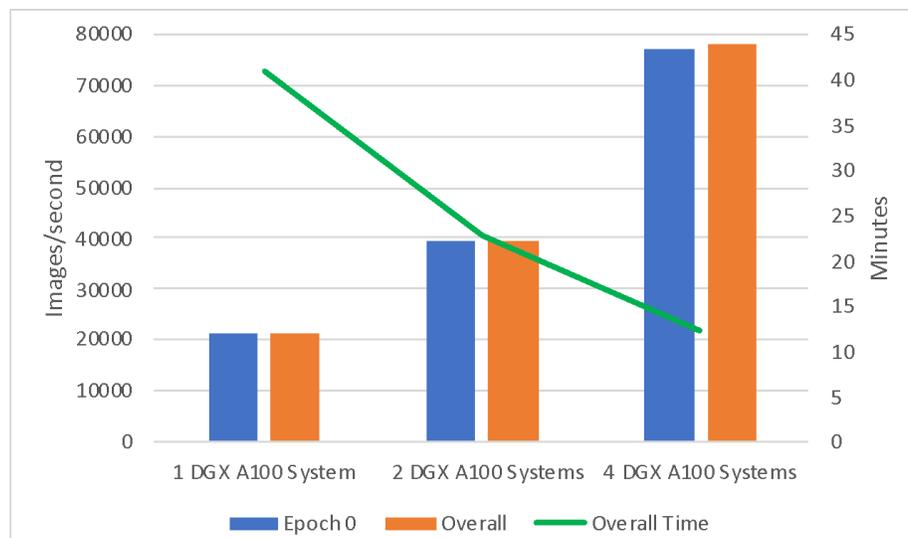


FIG. 12 MLPerf ResNet-50 Images per Second and Overall Time

7. Conclusion

AI, fueled by rapid innovation in DL solutions, is becoming common in many industries. Organizations that 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 in building a scalable and AI-optimized infrastructure often hold them back. Traditional compute infrastructures are unsuitable 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 robust 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.

8. Appendix

References

Product & Technology	
WEKA Data Platform	
WekaFS Datasheet	https://www.weka.io/wp-content/uploads/files/2020/03/WekaFS-DS-W01R14DS201808.pdf
WEKA Architecture White Paper	https://www.weka.io/wp-content/uploads/files/2017/12/Architectural_WhitePaper-W02R6WP201812-1.pdf
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/

9. About WEKA

WEKA offers a modern subscription software-based data platform delivering 10x+ performance and scale demanded by today's cloud and AI workloads. With the simplicity of NAS, the performance of SAN or DAS, and the scale of object storage, no more compromises between Simplicity, Speed, or Scale. Learn more at www.weka.io or follow us on Twitter @WekaIO.

