WekaFS™

Faster Deep Learning for AI and Analytics

EPIC PERFORMANCE AND LIMITLESS SCALE
Delivers > 10 GB/sec to a single client. Performance scales linearly as system grows.

RUNS ANYWHERE
Runs anywhere your data lives, on any standard server hardware, whether on-premises, in the cloud, or shared across both.

UNIFIES YOUR DATA
One global namespace for your entire data lake; easily access and manage billions of files from one directory.

BEST ECONOMICS
Run WekaFS software on any standard server. Burst workloads to the cloud when more GPU instances are needed.

tu simple

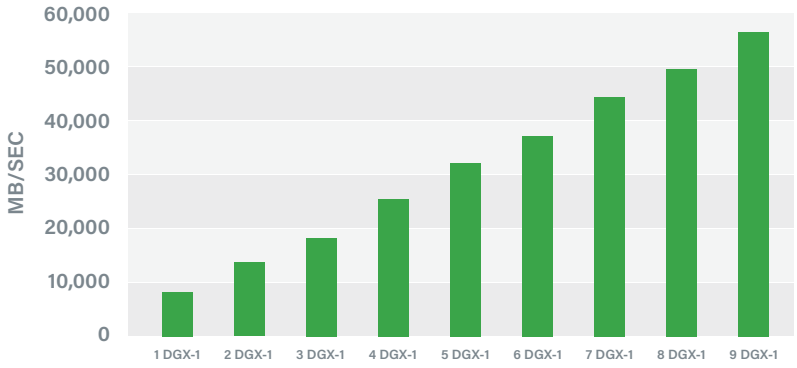
WekaIO was the clear choice for our DNN training...standard NAS would not scale and WekaFS was the most performant of all the parallel file systems we evaluated...we really liked that it was hardware-independent, allowing us better control over our infrastructure costs.

Dr. Xiaodi Hou
Co-founder and CTO, TuSimple

PREVENT GPU STARVATION AND GET TO THE ANSWERS FASTER

Modern analytics platforms are GPU-intensive and require large data sets to deliver the highest levels of accuracy to training, inference, or analytics models. Typical data sets can span from terabytes to tens of petabytes, and the data access pattern for each epoch is unique and unpredictable. GPU servers are expensive and scarce company resources and need to be fully utilized, but legacy networked storage cannot deliver the performance required to keep up with the data ingest demands. Many organizations have resorted to copying data to the local storage found inside a GPU server, but it just wastes time and is cumbersome to administer at scale.

This calls for a modern storage infrastructure that can instantaneously and consistently feed large amounts of random data, all emanating from a single shared data pool, to multiple GPU nodes in real time. A very high-bandwidth, low-latency storage infrastructure is required to ensure that GPU clusters are fully saturated with as much data as applications need. WekaFS is a new file system that is built for those who solve big problems and that easily meets the I/O requirements of the most demanding AI and analytics models. It is the world's fastest and most scalable file system, perfect for data-intensive applications, whether hosted on-premises or in the public cloud. Its performance has been proven to be both outstanding and scalable, delivering over 80 gigabytes per second (GB/sec) bandwidth to a single GPU server . On a single 100 Gbit network link, Weka delivers 10x more data than legacy network protocols such as NFS and 3x more than what is possible with NVMe solid state drives (SSDs) inside the local GPU server. Weka scales performance linearly as the GPU server load grows in size so you don't have to worry about the performance impact of future expansion.



Perfect Linear Scaling from 8 to 72 GPU nodes

ELIMINATE COMPLEX DATA COPY OPERATIONS

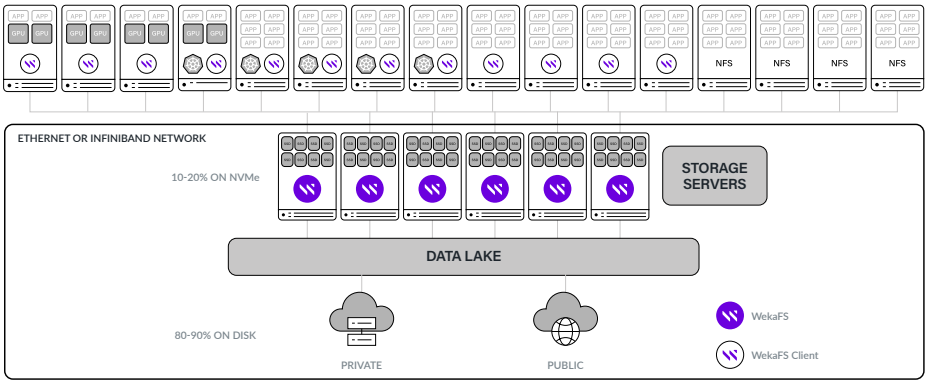
Managing large amounts of data is challenging when AI training models span multiple GPU servers. Local disks deliver predictable performance, but data must first be copied into the server's SSD storage, introducing significant server idle time and adding complexity to the workflows. A shared file system eliminates this cumbersome operation, but legacy file systems cause GPU I/O starvation due to poor performance and latency. WekaFS solves both of these issues, presenting a shared POSIX file system to the GPU servers and delivering extreme performance to keep data-intensive applications compute-bound. Weka customers have seen an 80x increase in the number of AI training epochs completed by increasing performance and eliminating local copy operations that an NFS-based architecture would require.

SCALE PERFORMANCE ACROSS THE GPU CLUSTER

The performance needs of modern data analytics require a complete departure from legacy file structures and hard-disk-based architectures. A single high-performance GPU server can experience I/O demands in excess of 80GB/sec of data processing. Predictable and seamless performance scaling is impossible with traditional NAS filers due to file system protocol limitations, resulting in data starvation and poor

[1] WekaFS supports Magnum IO, which has 8 lanes of 100 Gbit InfiniBand to a single DGX-2

utilization of expensive GPU server resources. The clean-sheet design of WekaFS leverages the performance of NVMe flash technology and fast networking — Ethernet or InfiniBand — ensuring the highest performance and lowest latency for the most demanding and unpredictable workloads generated by AI systems. WekaFS is uniquely able to meet the performance needs of these workloads, which have a highly randomized access pattern to both small and large files.



The Weka File System in a Typical Production Environment

WekaFS is a fully parallel and distributed file system that spreads both data and metadata across the entire storage infrastructure to ensure massively parallel access to files. The software supports servers running on InfiniBand or Ethernet (10 Gbit and above) networks, and Weka's network stack delivers performance at sub-100 microsecond latency to the applications. Data locality is no longer a pre-requisite for performance, and WekaFS can easily handle the most demanding data and metadata intensive operations.

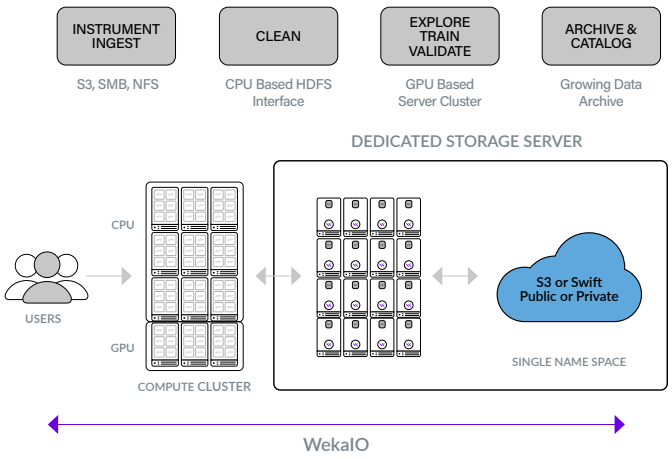
SCALE CAPACITY WITH ADVANCED DATA SECURITY

WekaFS is the only shared storage solution that provides end-to-end data management for data-intensive AI and analytics workloads. A single global namespace spans high-performance flash for fast data reads to the AI models, while an integrated hard disk-based layer provides exabyte scaling for long-term data growth and preservation. The entire data set is presented to the applications and users have complete visibility and instant access to the corporate-wide data set under management. WekaFS offers instantaneous backup and DR features to any S3 compatible object store through a unique feature called snap-to-object. In addition, the software has advanced data security with authentication, end-to-end data encryption, and key management integration. Weka's encryption has been shown to have no impact on I/O performance.

SCALE WORKLOADS TO THE CLOUD

WekaFS runs anywhere your data lives, and can be deployed on pre-tested platforms from all the major server vendors or in the public cloud for on-demand provisioning of storage to GPU instances. WekaFS also supports a hybrid model with its cloud-bursting feature; users with on-premises GPU clusters can elastically grow their environment in response to peak workload periods. When the cloud workload is complete, data can be returned to on-premises storage, resulting in faster time to results without the need to buy additional GPUs.

To find out more or to arrange for a free trial, contact us at info@weka.io.



Weka's storage scalability and ability to grow the infrastructure without losing performance was a key factor in the decision to select the Weka file system.

Oren Ben Ibghei
IT Manager, Innoviz

INNOVIZ™ TECHNOLOGIES

WEKA

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