WekaIO MATRIX™
FASTER DEEP LEARNING FOR AI AND ANALYTICS

>10GB/SECOND TO A SINGLE GPU CLIENT
Ensure applications never have to wait for data

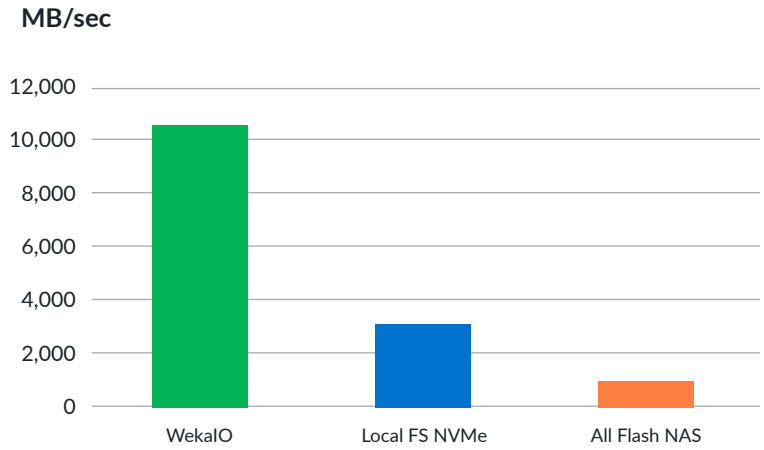
PERFORMANCE SCALES ACROSS THE CLUSTER
Performance scales linearly with each client node for maximum GPU utilization

VIRTUALLY UNLIMITED CAPACITY SCALING
Automatic tiering to object storage for massive scale

BEST ECONOMICS
Run software on any white-box server. Burst workloads to the cloud when more GPU instances are needed.

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 the training or analytics systems. They also require a high bandwidth, low latency storage infrastructure to ensure a GPU cluster is fully saturated with as much data as the application needs. Typical data sets can span from terabytes to tens of petabytes, and the data access pattern for each epoch is unique and unpredictable. This calls for a data infrastructure that can instantaneously and consistently feed large amounts of random data to multiple GPU nodes in real-time, all emanating from a single shared data pool. WekaIO Matrix is the world's fastest and most scalable file system for these data intensive applications, whether hosted on premises or in the public cloud. It has proven scalable performance of over 10GBytes per second bandwidth to a single GPU node, delivering 10x more data than NFS and 3x more than a local NVMe SSD.



1MB Read Performance to a Single GPU Client

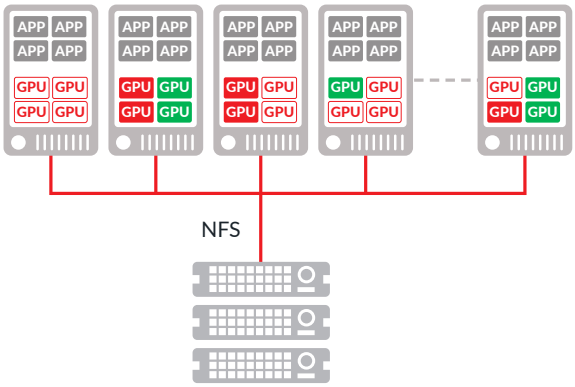
ELIMINATE COMPLEX DATA COPY OPERATIONS

Managing large amounts of data is challenging when the AI training system spans multiple GPU nodes. Local disk delivers predictable performance but data has to be copied into the local SSD, adding complexity to the workflows. A shared file system eliminates this operation, but legacy hard-disk optimized file systems cause GPU starvation. WekaIO Matrix solves both of these issues, presenting a shared POSIX file system to the GPU servers and delivering sufficient performance to keep data intensive applications compute bound.

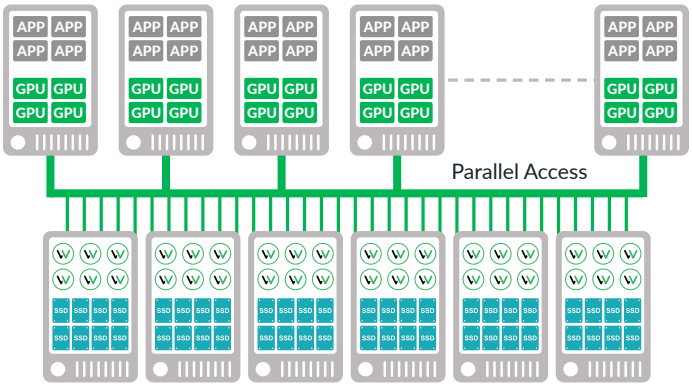
SCALE PERFORMANCE ACROSS THE GPU CLUSTER

The performance needs of modern data analytics require a complete departure from legacy file structures and hard disk based architectures. A single GPU node can experience I/O demands in excess of 10GBytes/second of data processing. Predictable and seamless

performance scaling is a challenge with traditional NAS filers that results in data starvation and poor utilization of expensive GPU resources. Matrix was written from scratch to leverage the performance of NVMe flash technology, ensuring the highest performance and lowest latency for the most demanding and unpredictable workloads generated by AI systems—a highly random access pattern consisting of both small and large files.



NFS Shared Storage Results in GPU Starvation

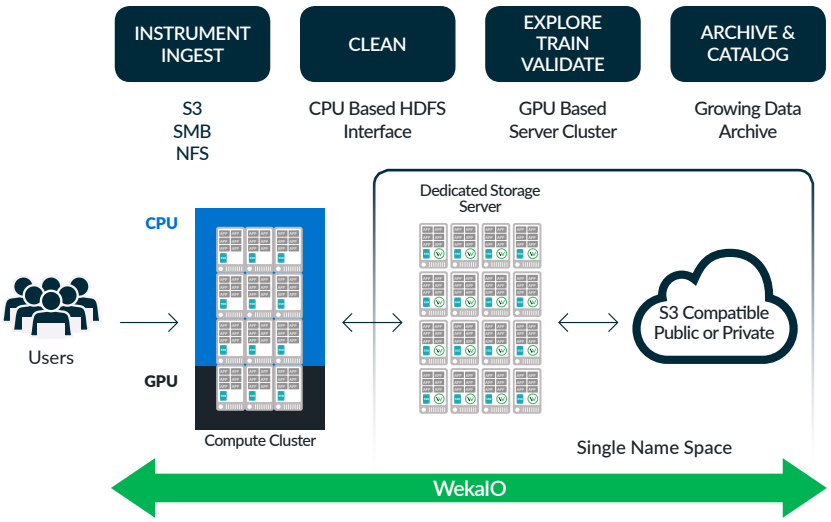


WekaIO Delivers Full Bandwidth to the GPU Cluster

WekaIO Matrix is a fully parallel and distributed file system, both data and metadata are distributed across the entire storage infrastructure to ensure massively parallel access. The software has an optimized network stack that runs on InfiniBand or Ethernet (10Gbit and above), so data locality is no longer a necessary factor for performance, resulting in a solution that can handle the most demanding data and metadata intensive operations.

SCALE CAPACITY WITH ADVANCED DATA PROTECTION

WekaIO Matrix is the only solution that provides end-to-end data management for data intensive AI and analytics workloads. A single global namespace spans from high performance flash for ingest and inference, to petabyte scaling to any S3 compatible object store (on-prem or public cloud) for long term data growth and preservation. Administrators and users have instant access to, and complete visibility of, the corporate-wide data set under management. A patented data protection scheme distributes data across the entire file system, and overall system reliability increases as the system scales.



SCALE WORKLOADS TO THE CLOUD

WekaIO Matrix can be deployed on pre-tested platforms from several server vendors or in the Amazon cloud as self-provisionable storage for P3 GPU instances. Matrix 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 on-premises, resulting in fast 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.IO
2001 Gateway Place, Suite 400W, San Jose, CA 95110 USA T 408.335.0085 E info@weka.io www.weka.io

W07r2DS201807