WekaFS™ for Weka AI™

Accelerated DataOps for Edge to Core to Cloud Data Pipelines

FASTER TIME-TO-MARKET, TO VALUE
Reduced epoch times, while delivering lowest inference times

EPIC PERFORMANCE, LIMITLESS SCALE
Industry's best GPUDirect performance: more than 80GB/sec of bandwidth to a single GPU client system

UNIFIES YOUR DATA
Single storage platform for entire data pipeline; easily access and manage billions of files from one directory

OPERATIONAL AGILITY
Best agility for data management across the Edge, Core, and Cloud

BEST ECONOMICS
Best TCO leveraging NVMe for performance and HDD-based object storage for capacity and archiving; 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*

WEKA AI EXPEDITES TIME-TO-MARKET, AND DELIVERS AGILITY AND SECURITY AT SCALE

Artificial Intelligence (AI) data pipelines are inherently different from traditional file-based IO applications. Each stage within AI data pipelines has distinct storage IO requirements: massive bandwidth for ingest and training; mixed read/write handling for extract, transform, load (ETL); ultra-low latency for inference; and a single namespace for entire data pipeline visibility. Furthermore, AI at the edge is driving the need for edge-to-core-to-cloud data pipelines. Hence, an ideal data storage solution must meet all these varied requirements and deliver timely insights at scale. Traditional solutions lack these capabilities and often fall short in meeting performance, data shareability, and mobility requirements. DataOps is data management for the AI era and delivers operational agility, governance, and actionable intelligence for these data pipelines by breaking silos.

Weka AI is a framework of customizable reference architectures (RAs) and software development kits (SDKs) with leading technology alliances that is underpinned by the Weka File System (WekaFS). Weka AI enables chief data officers, data scientists, and data engineers to accelerate genomics, medical imaging, the financial services industry (FSI), and advanced driver-assistance systems (ADAS) deep learning (DL) pipelines. Weka AI is available to easily scale from small to medium and large integrated bundled solutions. It is architected to accelerate DataOps by solving the storage challenges common with IO-intensive workloads and to deliver production-ready solutions. Weka AI provides operational agility with versioning, explainability and reproducibility, and governance and compliance with in-line encryption and data protection. Engineered solutions with partners in the Weka Innovation Network™ (WIN) program ensure that Weka AI will provide data collection, workspace and deep neural network (DNN) training, simulation, inference, and lifecycle management for the entire data pipeline.
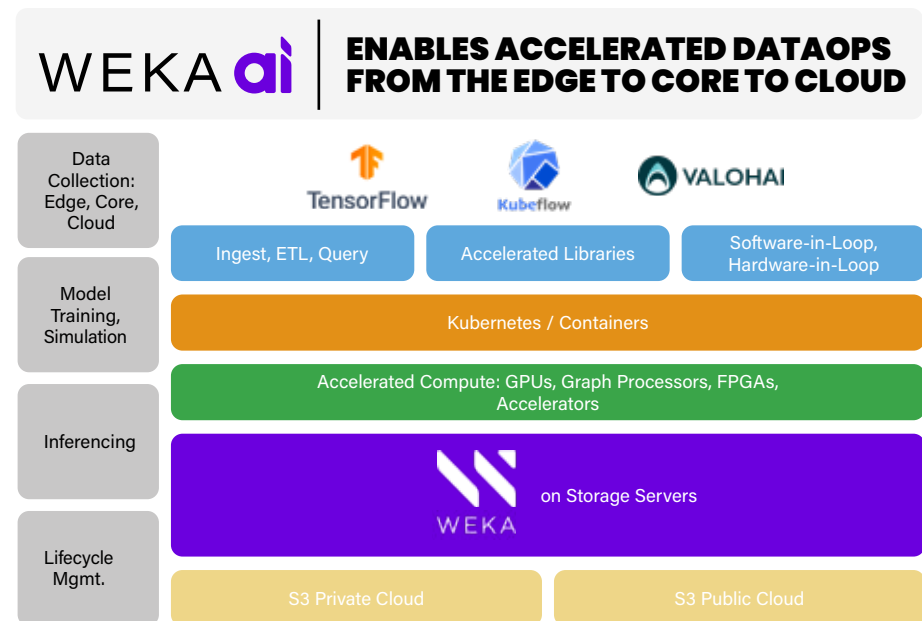
Weka AI: a solutions framework that enables Accelerated DataOps from the Edge to Core to Cloud

FASTEST FILE STORAGE FOR AI

WekaFS (Weka) is a modern file system that is built for those who solve big problems and that easily meets the IO 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

---

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

in the public cloud. Its performance has been proven to be both outstanding and scalable, delivering over 80GB/sec of bandwidth to a single GPU client. 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 so that you do not have to worry about any negative performance impact from future expansion.

ResNet-50 Model Training

Perfect Linear Scaling from 8 to 72 GPU nodes
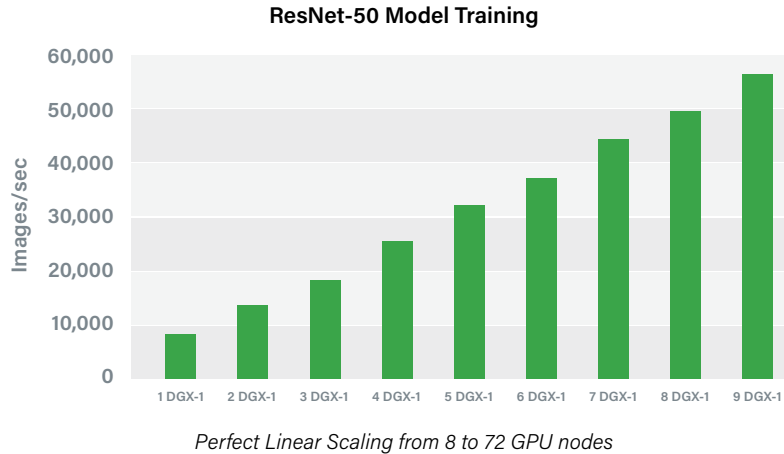
Managing large amounts of data is challenging when AI training models span multiple GPU servers. Local disks deliver predictable performance, but data must first be copied into the server's SSD storage, introducing significant server idle time and adding complexity to the workflows. A shared file system eliminates this cumbersome operation, but legacy file systems cause GPU IO starvation from poor performance and latency. WekaFS solves both 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 with superior performance and elimination of local copy operations that an NFS-based architecture would require.

SCALE PERFORMANCE ACROSS THE GPU CLUSTER

The performance needs of modern data analytics demand a complete departure from legacy file structures and hard-disk-based architectures. A single high-performance GPU server can experience IO 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 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.

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.

> " 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

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

For more information on Weka AI, read the Weka AI Reference Architecture document.

TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008   T: +1 844.392.0665   E: info@weka.io   www.weka.io