WEKA

SOLUTION BRIEF

Fueling the Future of Cloud Services with AI-Native Infrastructure

Unlock the full potential of your cloud services with WEKA.
Our advanced, AI-native architecture delivers unparalleled performance, scalability, efficient resource management, and simplified deployment and orchestration. For cloud service providers lacking deep storage expertise, WEKA simplifies deployment and maintenance, optimizes GPU utilization, and reduces costs. Enhance your service offerings and operational efficiency today—partner with WEKA and transform your cloud infrastructure to meet the demands of diverse, data-intensive workloads.

Challenges

Cloud service providers face significant challenges in delivering scalable compute, GPU resources, and network services while lacking specialized storage expertise. This gap can lead to inefficiencies and poor performance, especially for data-intensive workloads. Additionally, providers must support diverse GPU-powered workloads with flexible, scalable infrastructure that adapts to varying demands while maintaining cost-efficiency. With thin margins and growing power constraints, optimizing performance per kilowatt-hour is crucial for sustainable, high-performance operations.

Multiple Workloads

Cloud service providers must support diverse and evolving GPU-powered workloads, including AI, HPC, and data analytics. Legacy infrastructure often requires multiple architectures for different data profiles, leading to inefficiencies. Providers need flexible, scalable infrastructure that adapts dynamically to varying workload demands while maintaining high performance and cost-efficiency.

Storage Expertise

Cloud service providers often focus on delivering scalable compute, GPU resources, and network services rather than specialized storage solutions. This lack of storage expertise can lead to inefficiencies and suboptimal performance, especially for data-intensive workloads. Providers need a storage solution that performs optimally out of the box, without requiring fine-tuning, to bridge this expertise gap.

Challenges

- Managing various GPU-powered workloads with unique requirements.
- Maintaining high performance and cost-efficiency to meet evolving customer demands.
- Optimizing storage solutions for optimal performance for data-intensive workloads.
- Addressing power constraints to minimize energy usage for power-intensive applications like AI and HPC.

Solution

- The WEKA® Data Platform delivers unparalleled performance, scalability, and ease of use to transform your cloud infrastructure to meet the demands of diverse, data-intensive workloads.

Benefits

- High performance with low latency, optimizing GPU efficiency for diverse AI, ML, and HPC workloads while ensuring scalability to meet growing customer needs.
- Service provider-oriented multi tenancy features efficient resource management, physical tenant isolation, and automated deployment through Kubernetes.
- Zero-tuning setup, painless Day 2 operations, and industry-leading power efficiency simplifies management, reduces operational complexity, and minimizes energy consumption.

Thin Margins

Cloud service providers need infrastructure that minimizes administration and maximizes efficiency. Efficient cost management and revenue maximization allow providers to offer competitive pricing, attract more customers, and maintain profitability, which is essential for growth and sustainability.

Power Availability

As AI-oriented data centers densify, power availability has become a greater constraint than physical space. Optimizing performance per kilowatt-hour (kWh) is crucial, as data centers must maximize computational output within power capacity limits. This focus not only addresses power constraints but also aligns with sustainability goals, enabling support for high-density, power-intensive applications without exceeding power limits.

> The WEKA Data Platform is crucial in optimizing the performance of Yotta's Shakti Cloud, India's fastest AI supercomputing infrastructure. Shakti Cloud allows us to provide scalable GPU services to enterprises of all sizes, democratizing access to high-performance computing resources and enabling businesses to fully harness AI through our extensive NVIDIA H100 GPU fleet. With this enhancement, our customers can efficiently run real-time generative AI on trillion-parameter language models."
>
> **SUNIL GUPTA, CO-FOUNDER, MANAGING DIRECTOR, AND CEO OF YOTTA DATA SERVICES**

WEKA for Cloud Service Providers

Cloud service providers like AmpZ, Applied Digital, Denvr Data, IREN, NextGen Cloud, Sustainable Metal Cloud, TensorWave, and Yotta trust WEKA for its high-performance, scalable data platform that simplifies workload management and eliminates the need for deep storage expertise. WEKA's AI-native architecture delivers optimal performance out-of-the-box, while its robust multi tenancy and cloud-native integration ensure secure, efficient resource allocation. With a focus on power efficiency and sustainability, WEKA enables providers to maximize output, reduce costs, and confidently scale to meet evolving demands.

Effortless Multi Tenancy

WEKA's multi tenancy capabilities enable cloud service providers to efficiently manage resources across multiple tenants on shared hardware. With physical tenant isolation, automated Kubernetes integration, and independent encryption, WEKA ensures high performance, security, and compliance while reducing costs and complexity.

Simplified Metadata Management

WEKA's integrated architecture eliminates the need for separate metadata and data servers, allowing for seamless scaling and enhanced performance. This approach simplifies management, enabling rapid and efficient infrastructure growth to meet dynamic business needs.

Zero-Tuning, High Performance

WEKA's zero-tuning capabilities deliver optimal performance out-of-the-box, removing the need for manual configurations. This ensures consistent high performance across diverse workloads, reducing operational complexity and maximizing efficiency.

Power Efficiency

WEKA optimizes resource utilization, reducing the physical footprint and power consumption of cloud data centers. By maximizing GPU utilization and eliminating overprovisioning, WEKA delivers top-tier power efficiency and lower energy costs.

WEKA Accelerates AI

WEKA collapses the typical GPU-starving "multi-hop" AI data pipeline using a single namespace where the entire data set is stored. This zero-copy architecture eliminates the multiple steps needed to stage data before training. GPUs gain fast access to data needed for training, while WEKA automatically manages tiering of data between high-performance, NVMe-based storage, and low-cost object storage. Incorporating the WEKA Data Platform for AI into deep learning data pipelines saturates data transfer rates to NVIDIA GPU systems. It eliminates wasteful data copying and transfer times between storage silos to geometrically increase the number of training data sets that can be analyzed per day.

The WEKA Data Platform efficiently handles large numbers of files creating virtual metadata servers that scale on the fly with every server that is added to the cluster. WEKA's patented data layout algorithms distribute and parallelize all metadata and data across the cluster in small 4k chunks, this creates incredibly low latency and high performance whether the IO size is small, large, or a mixture of both.

About the WEKA Data Platform

The WEKA® Data Platform removes the barriers to data-driven innovation through its advanced software architecture optimized to solve complex data challenges and streamline the data pipelines that fuel AI, ML, and other modern performance-intensive workloads.

The design philosophy behind the WEKA® Data Platform was to create a single architecture that runs on-premises or in the public cloud with the performance of all-flash arrays, the simplicity and feature set of network-attached storage (NAS), and the scalability and economics of the cloud. Whether on-premises, in the cloud, at the edge, or bursting between platforms, WEKA accelerates every step of the enterprise AI data pipeline – from data ingestion, cleansing, and modeling to training validation or inference.

Mind-bendingly fast. Seductively simple. Infinitely scalable. Sustainable. Spanning edge, core, hybrid, and cloud. The WEKA Data Platform helps to overcome complex data challenges and accelerate next-generation workloads to unleash your organization's imagination, creativity, and potential.

© 2019-2024 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc. Other trade names used herein may be trademarks of their respective owners.

WKA???? 08/24