WEKA

ActualTech MEDIA

The WEKA Data Platform

James Sullivan and Dan Sullivan

- Cloud-native Architecture Runs Across Cloud and On-premises Architecture
- Intelligent Tiering and Autoscaling Stores Petabyte Scales of Data While Balancing Performance and Storage Costs
- Support for Self-describing Snapshots Enables Cloud Bursting, Disaster Recovery, Archiving, and Migrations

IN THIS PAPER

The WEKA Data Platform combines high performance storage, auto-scaling, and support for cloud and hybrid deployments to accelerate AI and HPC workloads while reducing costs, improving developer productivity, and improving GPU utilization.

Highlights include:

- Managing petabytes with a unified namespace
- Balancing performance and storage costs with intelligent tiering and autoscaling
- How the Zero copy approach reduces points of failure

CONTENTS

Throughout this series, we've seen how modern AI and other high-performance workloads are pushing the limits of legacy data stacks, particularly with respect to using compute accelerators efficiently. We have also highlighted the full range of capabilities that are needed to support HPC. This extends beyond just having storage and compute resources available to include support for scaling and deployments to various infrastructure platforms. The WEKA Data Platform meets the full range of requirements of the modern data stack and addresses these needs in a coordinated and comprehensive way.

Let's now delve into the key elements of the WEKA Data Platform to understand how it's positioned to support your data-intensive workloads as you shift them to the cloud.

The WEKA Data Platform provides automatic scaling to help reduce costs while ensuring adequate resources are available when needed by your HPC workloads.

Power of a Unified Namespace

The WEKA unified namespace provides a single logical model for organizing data, enabling you to manage petabytes of data stored across a variety of systems as a single logical entity. The namespace can span high-performance flash drives directly attached to servers and massively scalable object storage systems such as Amazon Simple Storage Service (S3).

The unified namespace also brings support for full lifecycle data management, including automatic tiering. WEKA manages tiering to optimize for performance which simplifies complex operations for customers.. The automatic tiering system is tuned for performance, so it minimizes frequent IO operations on tiered storage.

The unified namespace ensures developers, data scientists, and machine learning engineers aren't losing valuable time with low-value and time-consuming data management and movement tasks commonly needed in more siloed environments.

Scale Up and Scale Down Capabilities

The WEKA Data Platform provides automatic scaling to help reduce costs while ensuring adequate resources are available when needed by your HPC workloads. WEKA integrates with cloud provider auto-scaling mechanisms so nodes can be readily added to a cluster when there is additional demand on compute resources. Similarly, when the demand is less for compute resources, nodes can be removed from the cluster. This ensures you do not over-provision storage to support peak periods, and you do not pay for storage resources that are not actually used. Of course, removing nodes is done in a safe manner so data on decommissioned nodes is not lost.

One of the advantages of working with consistent workloads and a single type of storage system is that you can iteratively tune that storage system to best meet the needs of your workloads.

Supports Hybrid and Multi-Cloud Strategies

The WEKA Data Platform is designed to run where you want to run your workloads, including Amazon Web Services, Google Cloud, Microsoft Azure, and Oracle Cloud. Since the platform is designed to support hybrid and multi-cloud deployments, you have the flexibility to choose where your workloads run. WEKA software is fully containerized so it runs in any of the major public cloud providers as well as in on-premises server and storage hardware from Dell, HPE, Lenovo, SMC, and others.

This is especially important when data is distributed across multiple clouds and on-premises storage systems. The cost of moving large volumes of data can be prohibitive so bringing your jobs to your data is an especially effective way to keep costs down without sacrificing

workload performance or functionality. WEKA software also provides for data portability between clouds and from on-premises to cloud platforms when needed. This simplifies hybrid cloud operations, including bursting to cloud, use of the cloud for long-term retention, archiving data, disaster recovery and cloud migrations.

High Performance with Zero Tuning

One of the advantages of working with consistent workloads and a single type of storage system is that you can iteratively tune that storage system to best meet the needs of your workloads. Unfortunately for those responsible for HPC workloads, virtually none of the preconditions for that ideal situation apply to them. Most AI and other HPC projects require manually caching data sets to local storage. Developers waste time and the company spends money on multiple copies of data while the project pipeline is slowed.

HPC storage requirements span the range of needing low latency IO operations to keep compute resources operating with sufficient data to being able to cost-effectively store large volumes of data for years or more. Workloads also vary in how they put demands on storage systems. Some parts of the workload will require sustained, high-volume IO operations while other segments of the workload may have only limited need for storage operations. For example, ingest needs massive concurrency and high write throughput while pre-processing requires fast indexing and the ability to operate on large numbers of small files. During the training and analysis stages, we expect massive read operation. Model validation also needs massive bandwidth and mixed IO operations while inferencing demands low latency operation. Archiving requirements are fundamentally different from other stages with the need for massive write bandwidth as well as support for versioning and reproducibility.

The WEKA Data platform employs automated tiering to reduce IO bottlenecks by optimizing the placement and movement of data.

Another way customers solve the performance limitations in legacy storage is by over-provisioning storage resources, which wastes money. Very few storage resources scale both up and down, so once you add resources to meet peak, you're stuck paying for that. The WEKA Data platform employs automated tiering to reduce IO bottlenecks by optimizing the placement and movement of data. The platform also utilizes the most appropriate and performant storage with attention to small file processing and throughput performance.

Zero Copy Architecture

Copying data can be expensive and time-consuming. Most storage systems are optimized for only one or two steps in a data pipeline. Some customers solve this by copying data from shared storage into a local cache, but managing multiple copies of data is expensive.

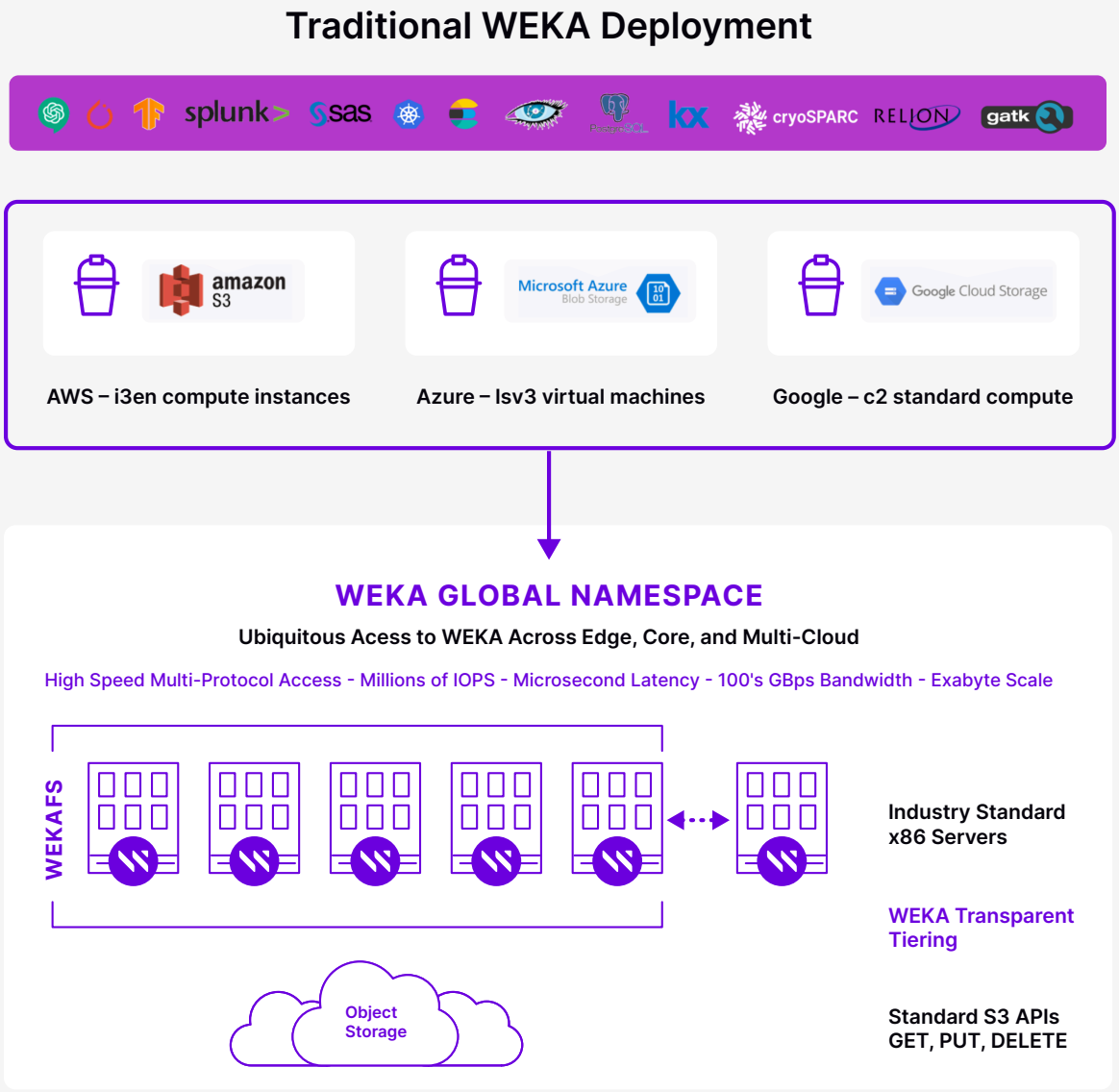

FIGURE 1: The WEKA architecture addresses key problem areas in traditional data architectures to create an architecture designed specifically for the demands of AI and other HPC workloads

The WEKA Data Platform helps to avoid unnecessary data movement or copy operations by delivering performance optimized for every performance profile in a single storage system. There is no need for extra copies in a workflow taking up resources. This zero copy approach also helps to reduce points of failure because there is no need for multiple copy operations in the workflow. A further benefit of using a single snapshot of storage during a workload is that a single copy ensures all operations on the data set are working with a consistent view of the data.

How WEKA Works

The WEKA platform uses a combination of cloud-native architecture, intelligent data tiering, auto-scaling, support for hybrid and multi-cloud, as well as edge computing support (see **FIGURE 1**). This combination of capabilities enables:

- Running in clouds or on-premises using dense storage virtual machines and local NVMe for high-performance storage
- Extended namespace to include object storage providing a more streamlined management experience
- Policy-driven data tiering
- Auto-scaling that matches infrastructure resources to workload demands
- Cloud bursting, disaster recovery, archiving, and migrations between clouds

WRAPPING UP

The WEKA Data Platform is a comprehensive solution for AI and other HPC building on a unified namespace, autoscaling capabilities, and flexible deployment options across cloud and on-premises environments. With the WEKA Data Platform, you get high performance without manual tuning, automation to optimize data placement, and support for a wide variety of workloads with varying requirements for compute and storage resources. The platform's Zero Copy architecture brings efficiency to workload processing by maintaining a single storage snapshot, which reduces resource consumption and ensures data consistency throughout a run of a workflow.

The WEKA Data Platform helps to avoid unnecessary data movement or copy operations by delivering performance optimized for every performance profile in a single storage system.

LEARN MORE

Learn more about the WEKA Data Platform by visiting WEKA for HPC and by downloading the NAND Research paper, "The Impact of Storage Architecture on the AI Lifecycle."

WEKA