WEKA

SOLUTION BRIEF

WEKA® Data Platform Converged Mode

Challenges

- Inefficient GPU resource utilization and poor data pipeline performance drive
- Pockets of idle resources drive up AI infrastructure costs
- Large GPU farms that drive model training/tuning/re-tuning consume large quantities of energy and drive carbon emissions. The CO_2 emissions for an NLP pipeline is estimated to be 78,468 CO2e compared to that of the lifetime of a car to be 126,000 CO_2e[1].

Solution

- WEKA® Converged mode is the first scale-out storage system deployed for on-premises GPU server farms or GPU-accelerated instances in the cloud. Customers achieve "zero footprint storage" by leveraging available local flash memory in next-generation GPU systems, increasing GPU resource utilization, and eliminating high costs associated with over-provisioning legacy filesystems.

Benefits

- Increase GPU resource utilization by 80% or more.
- Reduce data infrastructure costs by 50%
- Accelerate project timelines by 5x
- Build a more sustainable AI business model
- Increase application flexibility and agility in the cloud or on-premises

Introduction

In today's rapidly evolving digital landscape, businesses face the challenge of keeping pace with constant innovation while maintaining cost control and advancing sustainability initiatives. This delicate balancing act requires organizations to continuously innovate and optimize their operations to improve profits without compromising their long-term environmental and social responsibility goals.

AI and Generative AI are prime examples of how organizations juggle these conflicting priorities. According to a recent S&P Global Market Intelligence study, most AI deployments are focused on revenue-focused projects. McKinsey predicts that generative AI could add the equivalent of $4.4 trillion annually to the global economy. At the same time, IT professionals face the challenge of meeting the demand for AI-led innovation while maintaining cost control, forcing many companies to do more with existing infrastructure investments.

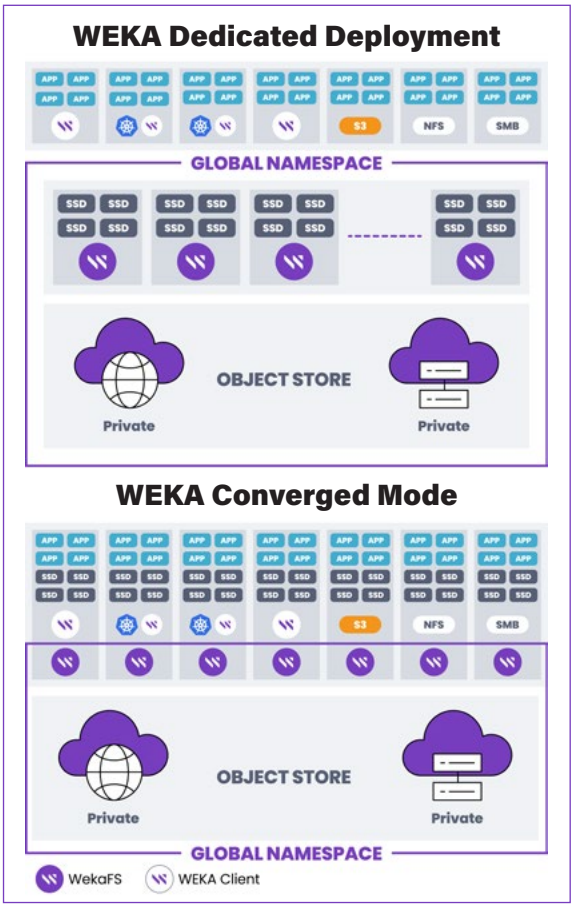
The GPUs that play a pivotal role in AI data pipelines often come bundled with substantial memory, network capabilities, and NVMe storage in a GPU server. These GPU servers or large-scale GPU server farms are often used for AI model training, re-tuning, interference, or tailored to specific application workloads. Traditionally, organizations would deploy external filesystems involving additional servers, rack space, network components, and cooling, incurring extra costs to provide shared and secure file capacity for CPU/GPU application instances. However, organizations with large-scale server farms may create pockets of idle and redundant resources. By maximizing their use of existing infrastructure, they can lower costs and support their sustainability efforts to reduce emissions and carbon footprints.

WEKA Data Platform Converged Mode

The WEKA® Data Platform can be deployed in Converged Mode to achieve a "zero storage footprint" for specific use cases such as AI/GenAI model training, (re)tuning, inferencing, HPC, etc. In this mode, WEKA can be safely and predictably deployed across large-scale application/GPU server farms, sharing server resources, including NVMe storage, alongside applications.

As a software-only solution, WEKA can be installed on server hardware from various hardware vendors or standardized compute instances from public cloud providers. Customers deploy WEKA in a dedicated or converged mode without requiring specialized server hardware or hardware offload.

When deploying WEKA in Converged Mode, the WEKA software will spin up and run inside Linux containers and always have a predefined amount of server resources (CPU, RAM, network, and NVMe capacity) it can use and can never exceed. These resources will always be pre-allocated to the WEKA containers on startup to avoid any scenario of just-in-time resource allocation that might fail due to resource unavailability. WEKA will use the existing NVMe resources within the application servers to create a high-performance shared storage pool.



Unlike legacy solutions, WEKA employs parallel data distribution across multiple servers/instances, maintaining performance without costly short-term write cache buffers. Rather, the data is persisted directly to NVMe without ephemeral write cache, ensuring speed and data integrity. The WEKA design distinguishes it from other solutions. It protects your application instances by pinning resources in a predictable and controlled manner, avoiding the random elasticity that can impact other processes on the same server or instance.

WEKA Converged Mode Benefits

- **Lower Costs:** Eliminate additional storage system purchases, networking infrastructure, rack space, and cooling expenses. There is no need to pay for externally provisioned filesystems, whether on-premises or from cloud providers.
- **Faster Project Deployment:** Accelerated project timelines result in quicker outcomes, eliminating the need to wait for additional equipment or resource provisioning.
- **Lower Energy Usage:** Enhanced compute instance utilization and elimination of idle resources lowers power consumption and for a greener data center footprint.
- **Greater Flexibility & Agility:** Converged solutions allow for budget-sensitive deployment, enabling gradual expansion or transitioning into a dedicated WEKA system.

Considerations and Best Practices for WEKA Converged Mode

Root Access

- Limiting root access in a shared resource environment to only admin roles in IT will reduce the risk of users indiscriminately rebooting servers or upgrading software components.

Orchestration

- Using workload managers such as Slurm helps to schedule exclusive resource reservations for compute and arbitrate contention within HPC clusters.

Kubernetes

- Containerized workloads provide load balancing and simplify application management on multiple shared hosts.

Structured Failure Domain Strategy

- Minimize the risk of shared resource failure by grouping WEKA instances across multiple failure domains to increase system availability at scale.

When to Choose WEKA Converged

One may choose to deploy WEKA in Converged Mode for several reasons, but we have highlighted three scenarios where this may be the optimal solution for your environment.

1. **Hardware is underutilized.** While this is a good reason to deploy WEKA in general, when you are using a large GPU farm or GPU in the cloud, it may be a good idea to explore the benefits of this deployment model for your particular workload requirements to get the most out of your hardware or instance environment.
2. **Budget Sensitivity.** You may be nearing the end of your fiscal year, presented with a new unbudgeted project, or simply need to maximize your overall investment. Whatever the case that may drive you to be budget-sensitive, this is one of the key benefits of converged solutions, and WEKA Converged Mode is the right option to jumpstart your project results.
3. **Cloud Infrastructure.** You can deploy WEKA Converged Mode on GPU-accelerated cloud instances as an option since WEKA is a software-only solution. For example, Amazon EC2's highest performance GPU-based instances that are suited to deep learning and HPC applications. Such instances have considerable and underutilized resources, fully capable of harnessing WEKA Converged Mode.

A WEKA converged solution offers organizations a more efficient, cost-effective, and manageable approach to expanding or introducing an AI infrastructure to the IT ecosystem. It is an attractive option for businesses looking to achieve greater utilization from their large-scale GPU investments, maximize existing resources to meet budget requirements and reduce the environmental impact of their AI and Generative AI deployments.

About WEKA

WEKA addresses the complex and diverse requirements of modern hybrid cloud environments. With WEKA, organizations can optimize data storage costs, get cloud scalability for every workload, enjoy consistent and straightforward data management for every application, and improve data resilience and portability.

For More Information or to Arrange a Free Trial

Visit us at https://www.weka.io/get-started or email us at info@weka.io.

1 "Energy and Policy Considerations for Deep Learning NLP", UMASS Amherst Research