Kubernetes-Native Storage for AI Infrastructure

WEKA manages the complete storage cluster as a Kubernetes workload: one control plane for compute, networking, and storage.

Challenges

- Storage runs outside the Kubernetes control plane
- Config errors fail workloads after GPU resources are committed
- Storage consumption is invisible to cluster cost reporting

Solution

WEKA deploys and manages the full storage cluster as Kubernetes workloads, governed by the same control plane, scheduler, and admission framework as compute.

Benefits

- No more storage outages. Clusters self-heal and are managed entirely through kubectl.
- Catch bad configs before they burn GPU hours. Admission webhooks validate before provisioning begins.
- Run one control plane, not two. The full cluster lifecycle deploys, scales, and recovers inside Kubernetes automatically.

Kubernetes was built to orchestrate the full infrastructure stack. CPU scheduling, GPU allocation, networking, and virtual machine management all have native Kubernetes integration paths. Storage has been the exception. Most storage vendors deliver a CSI-only operator: a deployment wrapper for a CSI driver. The storage cluster runs outside the scheduler, outside admission control, and outside the Kubernetes lifecycle.

For AI infrastructure and CI/CD teams, that gap is felt every day. Storage doesn't deploy from a Helm chart, doesn't reconcile through GitOps, and doesn't appear in kubectl. When a training run stalls on a storage issue, Kubernetes events offer no signal. Engineers switch to a separate console and runbook to diagnose what the control plane cannot see. Node churn breaks mounts with no automated recovery. Storage consumption won't appear in a cluster cost report.

As Kubernetes absorbs more infrastructure responsibility through KubeVirt, Dynamic Resource Allocation, and Cluster API, the gap in storage capabilities compounds.

WEKA closes it.

Two Kinds of K8s Storage Operators

Not all Kubernetes storage operators work the same way. The distinction that matters is whether the operator manages the storage cluster itself, or only manages the bridge to a storage system that runs outside Kubernetes. That classification determines what Kubernetes can see, schedule, govern, and automate.

Classification	Examples	What It Manages
CSI-Only	NetApp Trident, VAST CSI, HPE CSI, Dell CSM	Manages the CSI driver installation and configuration. The storage cluster runs outside Kubernetes, with separate tooling and separate operational runbooks.
Kubernetes-Native	WEKA	Manages the full storage cluster as Kubernetes workloads: deployment, scheduling, admission control, driver distribution, and lifecycle management.

The WEKA Kubernetes Operator ships five **Custom Resource Definitions (CRDs)**, or extensions that allow you to define custom object types, that bring the full storage cluster under Kubernetes governance. Each capability extends the control plane into a domain that CSI-only operators cannot reach.

	CSI-Only Operator	WEKA Kubernetes-Native Operator
Cluster lifecycle	Managed outside Kubernetes via separate tooling	Deployed and managed as Kubernetes workloads via CRDs
Scheduler visibility	Storage invisible to the scheduler	Storage capacity registered as schedulable Kubernetes resources
Admission control	Configuration validated at runtime only	Admission webhooks validate before provisioning begins
Node drain handling	Manual intervention required	Automated through WekaContainer lifecycle management
Multi-tenant isolation	Software isolation only	Independent WekaCluster CRDs with hardware-level partitioning

Why the Architecture Difference Matters

- **Unified control plane at AI factory scale.** At hundreds to thousands of nodes, separate storage monitoring, lifecycle management, and upgrade processes compound with every workload added.
- **Kubernetes adoption is outpacing storage architecture evolution.** 82% of organizations now run Kubernetes in production, up from 66% three years ago (CNCF, 2026 Annual Cloud Native Survey; 54% of AI/ML workloads already run on it (Spectro Cloud, 2025 State of Kubernetes report). Storage vendors have not followed this trend.
- **HPC undergoing a structural transition to Kubernetes.** NVIDIA's acquisition of SchedMD brings Slurm, the GPU Operator, CUDA, and DGX under one ownership. HPC teams now evaluate how their storage architecture can participate in a Kubernetes-native operational model.
- **Storage misconfiguration at AI scale fails workloads after resources are committed.** Without a native storage operator, errors surface only at runtime, after GPU resources are already allocated.

Validated in Production

The WEKA Kubernetes Operator is in production with hyperscalers, government-scale air-gapped infrastructure, and dense GPU clusters where storage compute must be Kubernetes-native.

Production at Scale

Three deployments show what a Kubernetes-native storage operator makes possible.

coupang	together.ai	DCAI
Bringing storage compute fully under Kubernetes resource management, closing the efficiency gap on dense GPU infrastructure.	Co-developed the admission policies feature shipped in Operator v1.12 — a policy-as-code capability built and validated by a hyperscale AI inference customer.	Running composable, multi-tenant storage at government scale through a single Kubernetes control plane.

Ready to manage storage in Kubernetes?

Visit weka.io/kubernetes to schedule a demo or speak with a solutions engineer.