WEKA®

SOLUTION BRIEF

Modern Storage for the Exascale Era

NeuralMesh™ by WEKA®, purpose-built for large-scale enterprise and AI infrastructure

Challenges

- Legacy storage can't scale with modern, distributed workloads
- Metadata bottlenecks limit performance at scale
- Manual tiering and upgrades slow down operations
- Monolithic architectures introduce latency in the data path

Solution

- NeuralMesh by WEKA delivers fully containerized, service-oriented storage
- Distributes and rebalances all data and metadata services across every node
- Unified namespace with intelligent, built-in tiering
- Runs seamlessly across on-prem, cloud, and hybrid environments

Benefits

- Linearly scalable performance with no hotspots or idle nodes
- Fast, non-disruptive upgrades and elastic expansion
- Simplified operations—no manual tuning or migration tools required
- Built-in enterprise-grade resiliency, security, and multi-tenancy

As enterprises and research institutions enter the exascale era, their infrastructure must evolve to meet the needs of next-generation workloads. Artificial intelligence, high-performance computing, digital twins, and real-time analytics are increasingly pushing the limits of conventional architectures. These workloads are data-intensive, unpredictable, and inherently distributed. They demand consistent, low-latency access to massive datasets—often spread across cloud, data center, and edge environments.

Meanwhile, the rest of the enterprise tech stack has already moved on. Compute runs in containers. Networking is software-defined. Observability and security are delivered as services. Storage, however, has been slow to evolve—still bound by monolithic designs and legacy architectures that were never meant to support the demands of AI-scale operations.

Traditional storage solutions, even those labeled "high performance," tend to fall short under modern conditions. Many isolate metadata and data functions onto separate nodes, creating resource silos that lead to imbalances as systems grow. Scaling these systems often introduces complexity and cost, while metadata-heavy workloads like AI training and analytics can create bottlenecks that slow everything down.

The rise of generative AI has further exposed the limitations of legacy architectures. Training large models and serving inference at scale requires consistent, low-latency access to massive and rapidly growing datasets. Yet most storage systems were built for pre-AI workloads and are rife with scaling walls, slow metadata handling, and inflexible deployment models. These issues directly contribute to high latency, which not only degrades throughput but also increases the time and cost per token—critical metrics for AI providers. Some teams attempt to manually cache or stage data to alleviate these bottlenecks, but these stopgaps are operationally expensive and unsustainable.

In addition, managing capacity and performance across disjointed storage tiers—flash, disk, and cloud object storage—requires manual intervention, scripting, or external tools. Upgrades and expansions can be disruptive, and achieving true multi-tenancy often requires trade-offs in security or performance. For enterprises and exascale environments, these limitations are unacceptable.

NeuralMesh by WEKA is purpose-built to solve the very challenges that traditional storage architectures have been slow to address. While legacy systems struggle to adapt to the demands of AI-scale operations, NeuralMesh introduces a fundamentally different, service-oriented approach to storage—one that aligns with how the rest of the modern data stack is already designed to operate.

Built on a software-defined, containerized microservices architecture, NeuralMesh is not just capable of scaling—it improves with scale. It distributes all core storage services—data, metadata, telemetry, and tiering—across every node in the system, dynamically balancing workloads in real time and ensuring that performance scales linearly as infrastructure grows. Its fully decoupled services allow upgrades, changes, and scaling to occur with zero disruption, while native support for multiple protocols and deployment models means it can run anywhere—from bare metal to multicloud.

Unlike traditional architectures still tied to physical constraints and monolithic scaling models, NeuralMesh is designed for the emerging world of agentic AI and reasoning models. It supports the rapid delivery and storage of large token streams, reduces time to first token, and serves as a flexible foundation for building scalable, AI-optimized infrastructures—including inference-serving pipelines, model training environments, token warehouses, and data-as-a-service offerings.

NeuralMesh by WEKA delivers a new foundation for AI and exascale environments that require more than just high-performance storage—they need intelligent, scalable, and adaptive infrastructure.

- **Linearly Scalable Performance**
  NeuralMesh grows stronger as it scales, distributing services across all cores for consistent performance—ideal for AI workloads that demand low latency and predictable throughput at any size.
- **Faster Time to First Token**
  With ultra-low-latency access paths and direct GPU integration, NeuralMesh shortens time to first token, accelerating model responsiveness and reducing operational costs.
- **Build Anywhere, Deploy Everywhere**
  Its container-native, service-based design makes NeuralMesh cloud-agnostic and hardware-flexible, allowing organizations to build infrastructures that span on-prem, edge, and public cloud seamlessly.

- **Designed for AI Factories and Agentic Workflows**
  Whether serving inference, orchestrating agents, or training foundation models, NeuralMesh supports emerging AI patterns like token warehouses, context caching, and massive parallelism without hitting legacy bottlenecks.
- **Future-Proof Infrastructur**
  As workloads evolve, NeuralMesh evolves with them—adapting to new compute topologies, scaling to exabytes, and maintaining resiliency and performance without requiring architectural overhauls.

With NeuralMesh, organizations can finally align their storage layer with the rest of their service-oriented infrastructure. Performance scales linearly as the system grows—thanks to full distribution of metadata and data handling across all nodes. Storage resources are efficiently used, with no idle capacity or overloaded hotspots, even under high concurrency or unpredictable access patterns.

NeuralMesh by WEKA marks a fundamental shift in storage architecture—away from monolithic, hardware-constrained designs and toward a dynamic, service-oriented model that matches the rest of the modern data center. For enterprises and exascale environments facing unprecedented scale and complexity, NeuralMesh provides the performance, resilience, and flexibility needed to stay ahead.

Inside the NeuralMesh Architecture

NeuralMesh by WEKA breaks from traditional storage design by treating every storage function—data, metadata, telemetry, protection, and tiering—as a containerized service that can run anywhere in the cluster. These independently orchestrated services work together as a cohesive, self-balancing mesh that dynamically adjusts to workload demands.

This architecture is built around five core components:

- Core: Delivers resilient, self-healing storage with fast rebuilds and improved efficiency at scale.
- Accelerate: Creates low-latency, direct data paths between compute and storage to maximize GPU throughput.
- Deploy: Provides full deployment flexibility across bare metal, public cloud, and hybrid environments.
- Observe: Offers deep, real-time observability through integrated logs, traces, and metrics.
- Enterprise Services: Adds built-in features like encryption, snapshots, tiering, and multitenancy—without added complexity.

Together, these components form a dynamic, distributed storage mesh purpose-built for the speed, scale, and agility of modern AI infrastructure.

 WKA432-01 06/25