Meet NeuralMesh™

AI workloads are evolving fast—larger models, longer context windows, real-time inference, and distributed agents are pushing infrastructure to its limits. Traditional storage architectures weren't built to handle the scale, latency sensitivity, and concurrency demands of today's AI environments.

NeuralMesh™ by WEKA was engineered to eliminate these bottlenecks, delivering consistent microsecond latency, scalable resilience, and container-native flexibility across any deployment model. One global service provider, for example, used NeuralMesh to build a multitenant AI platform serving hundreds of customers—each with isolated workloads and QoS controls—all on a single shared, high-performance infrastructure. This overview breaks down the five core components of NeuralMesh, each purpose-built to support AI infrastructure that scales smarter, not harder.

Core – Scalable Resilience and Intelligent Performance

- How it Works: Core distributes data and metadata across all nodes, balancing I/O dynamically with built-in auto-healing, auto-scaling, and fast rebuild capabilities. Performance and resilience improve as the system scales.
- Strategic Advantage: Unlike legacy systems that degrade under load, Core gets stronger with growth—ideal for petabyte-scale AI workloads that demand always-on availability and deterministic performance.
- Outcome: AI environments stay online, performant, and self-optimizing even as data volumes and concurrency increase.
- Example: A genomics research institution scaled from 2PB to 12PB without downtime or rebalancing, achieving consistent I/O latency and eliminating weekend maintenance windows.

Accelerate – High-Speed Data Access for AI Performance

- How it Works: Accelerate establishes direct, parallelized paths between applications and data by distributing data and metadata intelligently. It includes support for multiple protocols, containerized delivery, and optimized network paths.
- Strategic Advantage: Delivers consistent microsecond latency and linearly scalable throughput that keeps GPUs fully utilized—even under mixed or bursty workloads.
- Outcome: Maximizes GPU investment, supports concurrent model training and inference, and eliminates performance bottlenecks from traditional storage layers.
- Example: An AI startup running multimodal inference pipelines on H100s reduced job queuing and increased throughput 4x after switching to WEKA from legacy NAS.

Deploy – Flexible, Composable Infrastructure Anywhere

- How it Works: Deploy uses a containerized microservices architecture that runs across bare metal, VMs, and all major clouds and neoclouds. Supports converged and disaggregated setups with full API control for orchestration.
- Strategic Advantage: Enables consistent, cloud-native deployment across on-prem, hybrid, and multicloud environments—simplifying scaling and lifecycle management.
- Outcome: Customers can deploy NeuralMesh once and run it anywhere, with no architectural redesign as workloads evolve.
- Example: A global enterprise deployed NeuralMesh bridging on-prem Dell PowerEdge nodes and Azure cloud regions, maintaining a unified namespace with identical performance and no replatforming.

Observe – Intelligent Visibility at Scale

- How it Works: Observe provides real-time, petabyte-scale observability across data paths, providing insights into performance metrics, and infrastructure health, ntegrated with dashboards, alerts, and telemetry APIs.
- Strategic Advantage: Delivers deep insight into system behavior, enables proactive remediation, and simplifies capacity planning in complex AI environments.
- Outcome: Reduces operational burden while improving SLA compliance, efficiency, and user experience.
- Example: A large media company used Observe to detect and resolve a data skew condition in minutes—preventing GPU underutilization during a critical training run.

Services – Secure, Feature-Rich, and Production-Ready

- How it Works: Offers enterprise-grade capabilities including global namespace, encryption in-flight/at-rest, snapshots, Snap-to-Object, RBAC, tiering between TLC/QLC NVMe and object stores, and container storage integration.
- Strategic Advantage: Enables secure, zero-tuning performance across shared or regulated environments with rich data protection and compliance features.
- Outcome: Customers can confidently run mission-critical AI workloads with minimal operational complexity and no need for traditional tiering policies.
- Example: A financial services firm used NeuralMesh's enterprise snapshot and tiering capabilities to reduce recovery times by 80% while maintaining compliance with data retention policies.

Built to Keep Up with AI

NeuralMesh brings together the speed, flexibility, and enterprise features required to support AI infrastructure at scale. Whether you're building training clusters, deploying inference services, or enabling multitenant AI platforms, NeuralMesh ensures the data layer is never the constraint. Its modular architecture empowers teams to deliver faster outcomes, higher efficiency, and a future-ready foundation for what comes next.

 WKA430-01 06/2025