Multitenancy Built for AI Scale: True Isolation, Zero Waste

Isolation every tenant demands. Economics every AI cloud needs.

Challenges

- Dedicated clusters create cost that never converts to revenue
- Tenant onboarding takes days
- Serving all tenant types profitably requires two solutions

Solution

NeuralMesh delivers the isolation every tenant requires — dedicated or shared — from one platform.

Benefits

- Cost per tenant drops as tenant count grows
- New tenants online in minutes, on demand
- Every tenant type served profitably on one system

AI infrastructure economics come down to one thing: how efficiently you turn hardware into a productive system. Every tenant that requires a dedicated cluster, reserved compute, or idle resources is siloed infrastructure cost that doesn't convert to revenue. NeuralMesh™ by WEKA® eliminates that waste without compromising on isolation. Whether your tenants need hardware-dedicated clusters or network-virtualized environments, NeuralMesh delivers the architecture each customer demands — all on the same platform, managed from a single control plane. More tenants, more efficient infrastructure, better economics at every layer of the stack.

Your anchor customers get the dedicated infrastructure they need to justify their commitment. Your smaller and shared-infrastructure tenants get the agility and economics of logical isolation. NeuralMesh delivers both models on one architecture: composable clusters for tenants whose workloads, contracts, or compliance posture require dedicated resources; native multitenancy for tenants who share infrastructure with full tenant-level boundaries. The same control plane, APIs, and operational tooling cover both. Most AI storage architectures require a separate platform for each isolation model. NeuralMesh handles both.

When tenants share infrastructure, they operate as if they don't. Network spaces give each tenant dedicated VLANs and IP ranges, with mount enforcement at the network layer. Overlapping IP address ranges across tenants are fully supported — onboarding a new customer doesn't require reworking network design. Per-tenant QoS ceilings at the tenant and filesystem level keep performance predictable across the shared cluster. Each tenant brings its own KMS, its own LDAP, and its own access policies.

More tenants. Better margins. Flat operational cost.

The economic case for NeuralMesh multitenancy is straightforward: every tenant added to a shared platform drives down the cost of the ones already on it. Storage capacity and compute are shared dynamically — tenants consume what they need and release resources immediately when done. Idle capacity is eliminated by design, not managed by policy. A new tenant comes online in minutes through the standard control plane, with capacity and performance boundaries set on demand. No hardware reconfiguration. No maintenance windows. No per-tenant provisioning ceremony. The speed of onboarding is itself a revenue motion — time between contract and first workload is time a service provider isn't billing.

Operations scale the same way. Cluster administrators provision, monitor, and manage every tenant through one API, one GUI, and one observability stack. Tenant administrators get full self-service control of their own filesystems, object buckets, snapshots, and users — with zero visibility into other tenants or the underlying cluster. As tenant count grows from ten to hundreds, operational overhead stays flat. The result is a more profitable AI cloud business: more customers served, lower cost per customer, and no additional headcount required to support the growth.

Key NeuralMesh multitenancy capabilities:

- **Physical and logical isolation on one solution:** Composable clusters deliver dedicated hardware per tenant. Native multitenancy delivers shared-infrastructure isolation with full tenant boundaries. One deployment, one control plane.
- **Elastic resource sharing:** Storage capacity and compute shared dynamically across tenants — no reserved idle capacity, no per-tenant cluster footprint
- **Tenant onboarding in minutes:** New tenants provisioned through the standard control plane — no hardware reconfiguration, no maintenance windows
- **Network spaces:** Per-tenant VLANs, IP isolation, and mount enforcement at the network layer — VPC-like isolation on shared hardware
- **Dual administrator model:** Cluster admins own infrastructure; tenant admins own their own environment — separation of duties by design.

Built for every tenant, at every scale

Two isolation models run on the same platform from one control plane. Platform teams match each tenant to the model it requires without operating multiple platforms, maintaining duplicate infrastructure, or splitting operational tooling. AI clouds and service providers run their largest anchor tenants alongside hundreds of smaller ones, each fully isolated and right-sized. Every token generated, every model trained, every inference served costs less as the platform grows — because the infrastructure doing the work isn't sitting idle between jobs.

Physical isolation	Logical isolation
Dedicated CPU, memory, and storage per tenant	Network spaces with private VLANs and IP ranges
Hardware-level resource carve-out — no sharing, no contention	Per-tenant QoS ceilings at tenant and filesystem level
Ideal for anchor tenants with dedicated-infrastructure contracts or compliance requirements	Overlapping IP address ranges supported — onboard tenants without redesigning your network
Same control plane, APIs, and operational tooling as logical isolation	Independent KMS, LDAP, and filesystem authentication per tenant

NeuralMesh multitenancy is available now. Talk to your WEKA account team to see how it fits your environment, or request a personalized demo below.

See NeuralMesh multitenancy in action. **Click here →**