NeuralMesh Replication

Replication re-imagined for the AI era

Challenges

- GPU capacity is scarce, expensive, and constantly shifting across regions and providers.
- Petabyte-scale AI data cannot keep pace as workloads move to wherever GPU capacity is available.
- Full-copy staging leaves GPUs idle while network, storage, and infrastructure costs continue to accumulate.

Solution

NeuralMesh Replication makes the source filesystem browsable at the destination after metadata sync, then lets teams choose how file contents become local.

Benefits

- Keep GPUs productive instead of waiting on data.
- Run workloads wherever capacity is available and cost-effective.
- Move and store only the data each workload needs.

GPU capacity is scarce, expensive, and increasingly distributed across clouds, regions, colocation facilities, partner sites, and on-premises environments. Organizations need the flexibility to run workloads wherever capacity is available and economically viable. But the data those workloads depend on often cannot move fast enough to meet them there.

AI datasets can span hundreds of terabytes or petabytes. Traditional replication requires the full dataset to be copied before the destination becomes useful, leaving valuable GPU capacity waiting while teams stage and manage data.

NeuralMesh Replication breaks the dependency between where data lives and where AI work runs.

Destination clusters quickly receive a browsable view of the source namespace, including directories, file names, sizes, and permissions, before file contents arrive. File contents then become local according to policy. Teams can retrieve and cache files on demand, proactively hydrate selected files or directories, copy the complete filesystem, or combine these approaches as requirements change.

The result is a more flexible way to put available GPU capacity to work sooner without copying everything everywhere.

A Different Model for Replication

Traditional full-copy replication often treats filesystem visibility, data movement, and data freshness as one operation. NeuralMesh replication separates them. Teams configure how often the destination receives an updated filesystem view independently from which file contents become local and when.

This separation supports a cache, a selective copy, a full replica, or a mix of the three within one architecture.

Architectural distinction	Traditional full-copy model	NeuralMesh Replication
Destination visibility	The destination becomes useful after the copy completes.	Metadata syncs first, making the filesystem browsable before file contents arrive.
Data locality	Every destination is treated as a full copy.	Cache on demand, hydrate selected paths, copy everything, or combine approaches.
Policy control	The copy schedule determines both freshness and data movement.	Metadata update cadence and hydration policy are configured independently.
Data path	Object storage or staging pipelines may be used as handoff layers.	Metadata and file contents move directly between WEKA clusters.

Flexibility is what makes NeuralMesh Replication different. Teams don't have to choose between an on-demand cache and a complete replica. Across destination filesystems, teams can cache data as workloads access it, proactively hydrate selected files or directories, or maintain a full copy, choosing the right approach for each filesystem as requirements change.

One foundation. Every distributed AI use case.

NeuralMesh Replication is built for how AI infrastructure actually runs — not just for DR. The same replication foundation supports workload mobility, collaboration, migration, and recovery.

- **Cloudbursting:** Teams can browse the source filesystem and hydrate only the data the job needs.
- **Multi-site collaboration:** Distributed teams can work from a shared source filesystem without maintaining a full copy at every location.
- **Migration:** Maintain namespace visibility while selectively or fully hydrating data into a new environment before cutover.
- **Recovery and resilience:** Use full hydration when a complete local copy is required. Snapshot cadence determines RPO, while the read-only destination preserves source authority.

What Smarter, Faster Data Movement Provides

Moving only the data that's needed, when it's needed, pays off across GPU utilization, storage footprint, and operations.

- **Put scarce GPU capacity to work sooner.** A destination can become browsable and begin hydrating the required data without waiting for the entire filesystem to arrive. Teams spend less GPU time and money on data staging.
- **Match data locality to the workload.** Cache most data on demand, pre-hydrate the paths a planned job will use, and maintain a full local copy where recovery or operational requirements demand it.
- **Scale storage and network usage with actual demand.** Destinations only need to store the files that workloads access or that policies select, unless a full replica is required."
- **Simplify distributed operations.** Direct cluster-to-cluster movement replaces object-store handoffs and custom transfer pipelines with a native replication workflow.

> "At our scale, data mobility isn't a nice-to-have; it's foundational. NeuralMesh's intelligent replication makes data mobility real at scale: we can make datasets visible across sites and pull exactly the data each job needs to the next GPU allocation as it becomes available. That shifts replication from a back-end protection function to a core part of how our distributed AI infrastructure needs to operate, improving workload mobility, capacity efficiency, and the resiliency our customers depend on."
>
> **Sam Tabar, CEO, WhiteFiber**

Ready to make AI data available wherever GPUs are?

Read the NeuralMesh replication Technical Brief or speak with a WEKA solutions engineer.