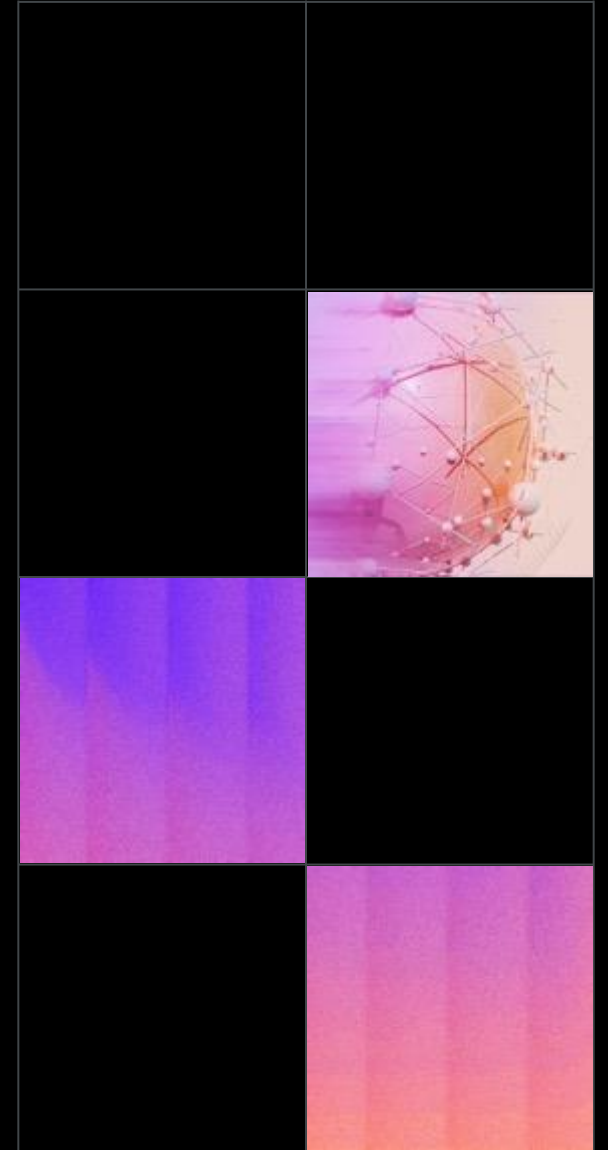


WEKA Update HPC User Forum May 2026

Bob Vassar
Sr. Systems Engineer, Federal



NeuralMesh — one parallel filesystem, every workload

Parallel filesystem for HPC simulation, AI training, inference, and data lake — same software, same namespace.



HPC Simulation

MPI scratch · checkpoint/restart · post-processing

Streamlined read & write paths.

Direct client-to-drive on large I/O.
Parallel parity writes for checkpoint throughput.

Sub-file parallelism, automatic.

Large files distributed across servers at extent boundaries — no manual striping.



AI Training

GPU pipelines · checkpoints at scale

Distributed metadata at scale.

Millions of small-file reads/sec without metadata bottlenecks. Keeps GPU pipelines fed.

Checkpoint throughput at line rate.

Large sequential writes streamed in parallel — out of the critical path.



AI Inference

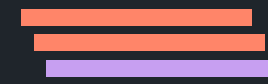
Model serving · RAG retrieval · KV-cache extension

Token-level latency, persistent context.

Microsecond-class data path. KV-cache extension via Augmented Memory Grid.

Cost-efficient TTFT.

Reuse KV-cache across sessions instead of recomputing prefill on every long context.



Data Lake

Analytics · shared data · S3 · mixed POSIX + object

One namespace, every protocol.

POSIX, S3, NFS, SMB — same data, no copies.

Compute, query, archive — concurrent.

Same data accessible to compute, analytics, and archive workflows simultaneously.

NeuralMesh by WEKA — one parallel filesystem, every workload

Userspace kernel bypass client · Distributed metadata · Single namespace across POSIX, NFS, SMB, S3
Same software on enterprise x86 or ARM + NVMe, WEKApod appliance, and natively on AWS, Azure, GCP, OCI

Three deployment options — same software

Pick the one that matches your hardware reality. NeuralMesh architecture is identical across all three.

NeuralMesh Axon

Converged on GPU servers

WHAT IT IS

NeuralMesh runs in containers on Compute servers themselves — local NVMe, CPU cores, and DRAM contributed into a single namespace. No separate storage tier.

WHEN TO USE

GPU-dense deployments where storage and compute can share resources. Eliminates the storage rack, recovers stranded NVMe inside GPU nodes.

Neoclouds · AI factories

WEKApod

WEKA-built appliance

WHAT IT IS

Pre-validated, performance-optimized appliance. Current-generation Prime and Nitro configurations. Nitro is NVIDIA SuperPOD and NCP certified.

WHEN TO USE

Predictable appliance experience. Required for NVIDIA SuperPOD / BasePOD certification. Faster procurement, no custom config debate.

NVIDIA-certified · turnkey

WQP

WEKA Qualified Configurations

WHAT IT IS

Same WEKApod hardware spec, on partner servers — Dell, HPE, Supermicro, Lenovo. Performance-equivalent, mirrored components.

WHEN TO USE

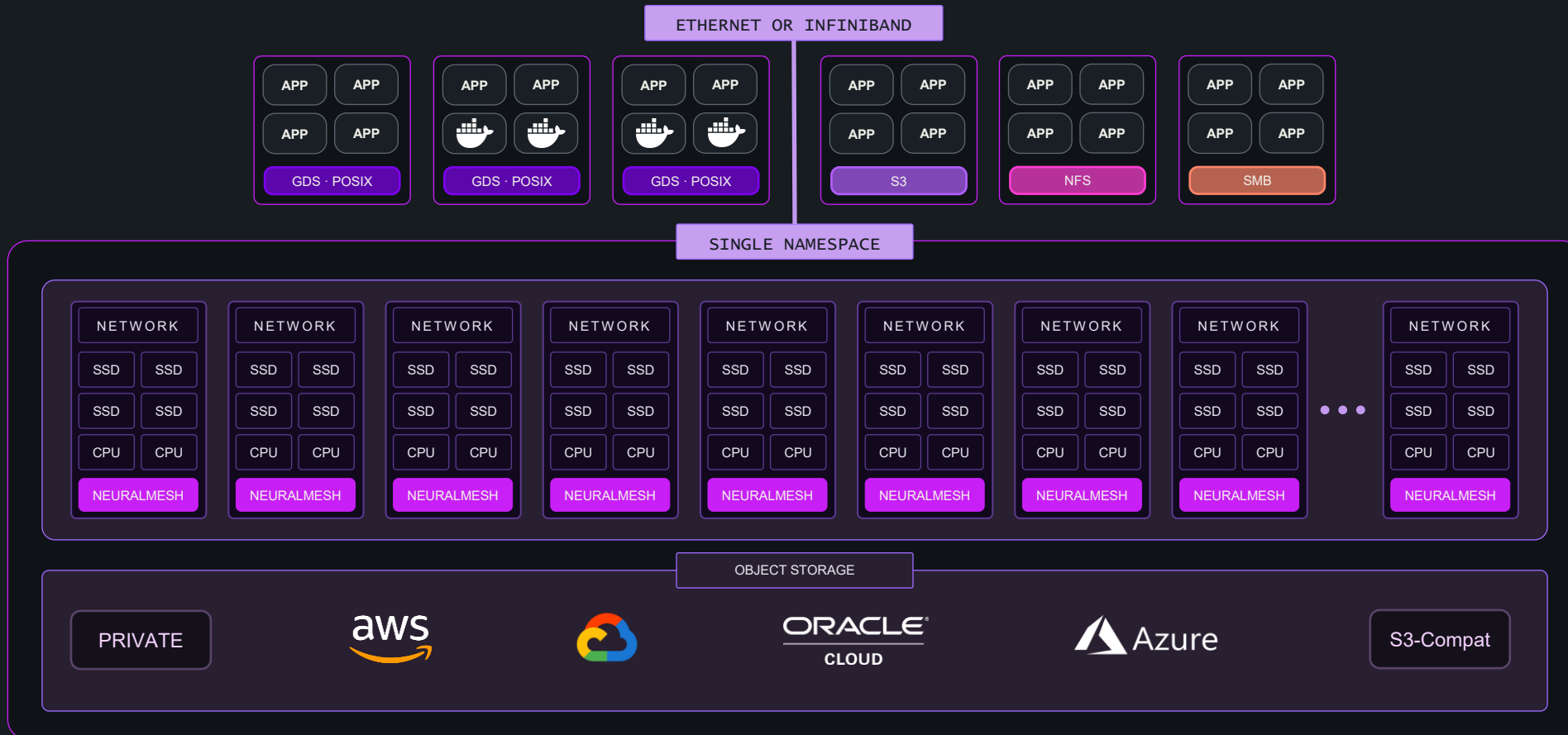
Existing partner relationships, custom requirements, or broader hardware choice. Same software image, same performance envelope.

Dell · HPE · Supermicro · Lenovo

AlloyFlash — capacity flash, no tiering

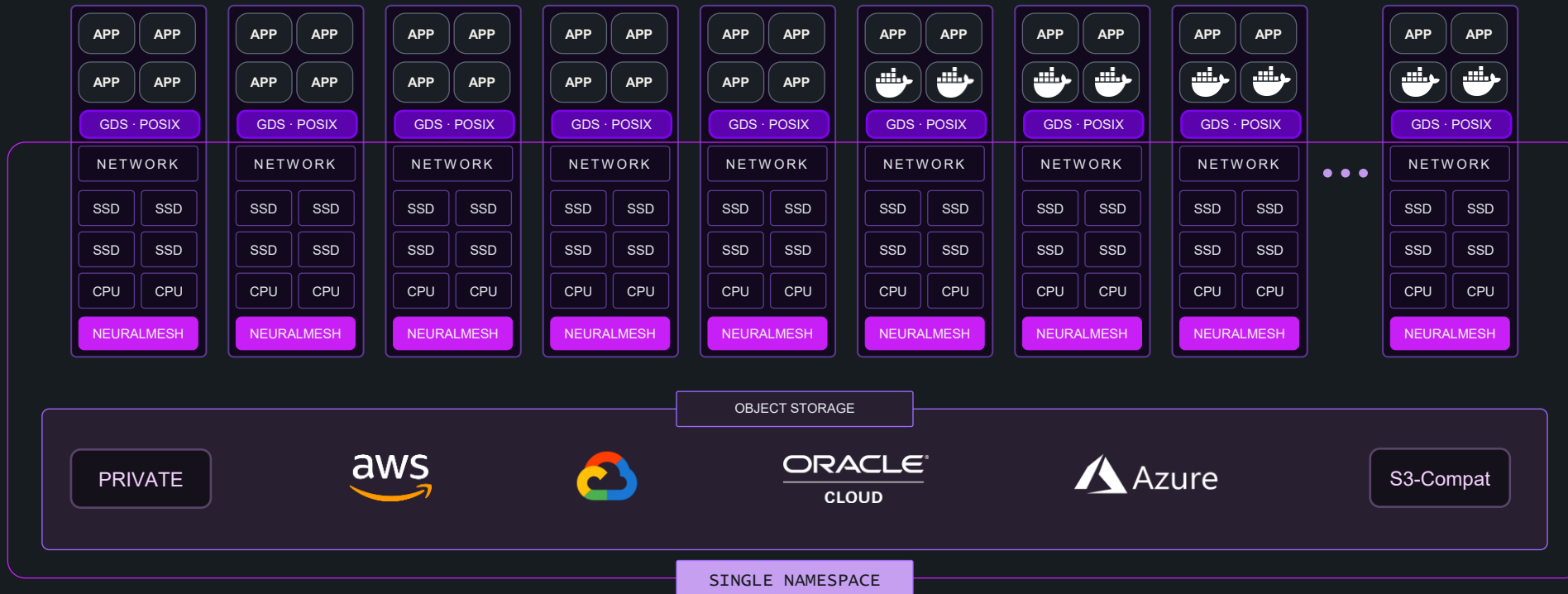
Mixed TLC + eTLC in one namespace. Available on WEKApod and WQP. WEKApod Prime delivers 65% better price-performance.

NeuralMesh deployment



Scales linearly · thousands of clients · hundreds of storage servers · exabyte-scale namespace across SSD and object

NeuralMesh Axon deployment



Scales linearly · thousands of clients · hundreds of storage servers · exabyte-scale namespace across SSD and object

Augmented Memory Grid (AMG)

Memory wall problem: GPU HBM is the bottleneck, not compute. AMG extends GPU memory with persistent shared NVMe.

DEDICATED · NeuralMesh on its own hardware

GPU SERVER · 8× GPU · 2× 400Gb NIC



NIXL · WEKA AMG CONNECTOR · RDMA + GDS

WEKA AUGMENTED MEMORY GRID

NEURALMESH STORAGE CLUSTER · NVMe pool
Dedicated infrastructure

CONVERGED · AXON · NeuralMesh runs on the GPU servers

GPU SERVER · 8× GPU · 2× 400Gb NIC



NIXL · WEKA AMG CONNECTOR · RDMA + GDS

WEKA AUGMENTED MEMORY GRID

NVMe **INSIDE GPU SERVERS** · same chassis
No dedicated infrastructure layer

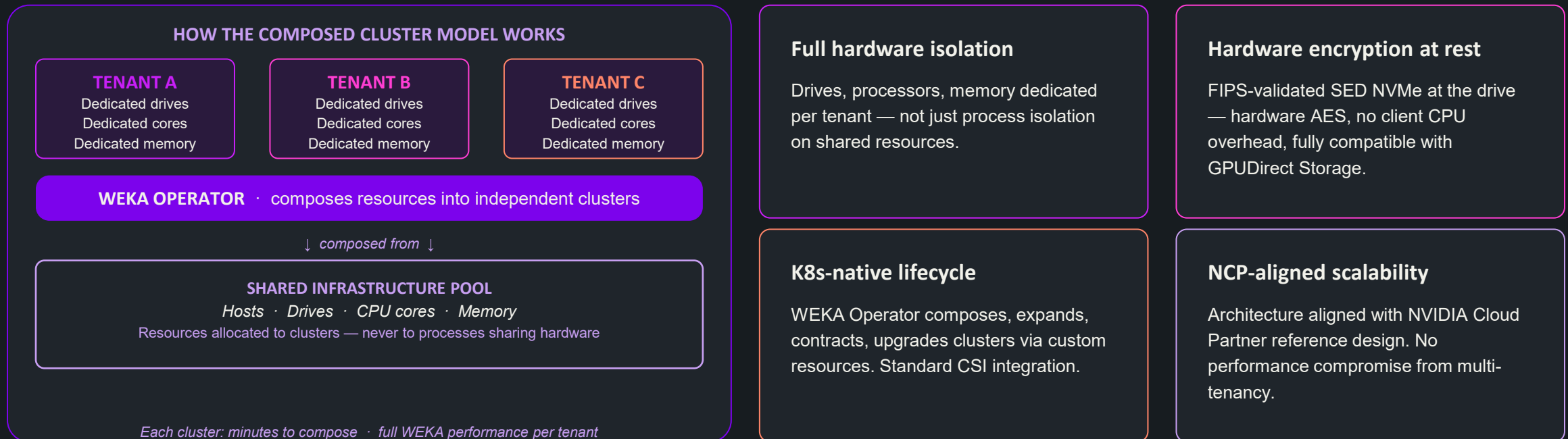
What it does at scale — Qwen3-Coder-480B FP8, 11 GPU hosts, real customer POC

GPU Hosts	Working Set	TTFT (DRAM)	TTFT (AMG)	Tok/s (DRAM)	Tok/s (AMG)	Advantage
11	44M tokens	2.85 s	2.02 s	798,728	1,173,795	1.47×
11	88M tokens	7.42 s	1.95 s	175,991	1,142,341	6.49×

AMG stays consistent under load. KV-cache hit rates hold as concurrent users grow — DRAM-only falls off as the working set exceeds memory. Run the same benchmarks: github.com/callanjfox/kv-cache-tester

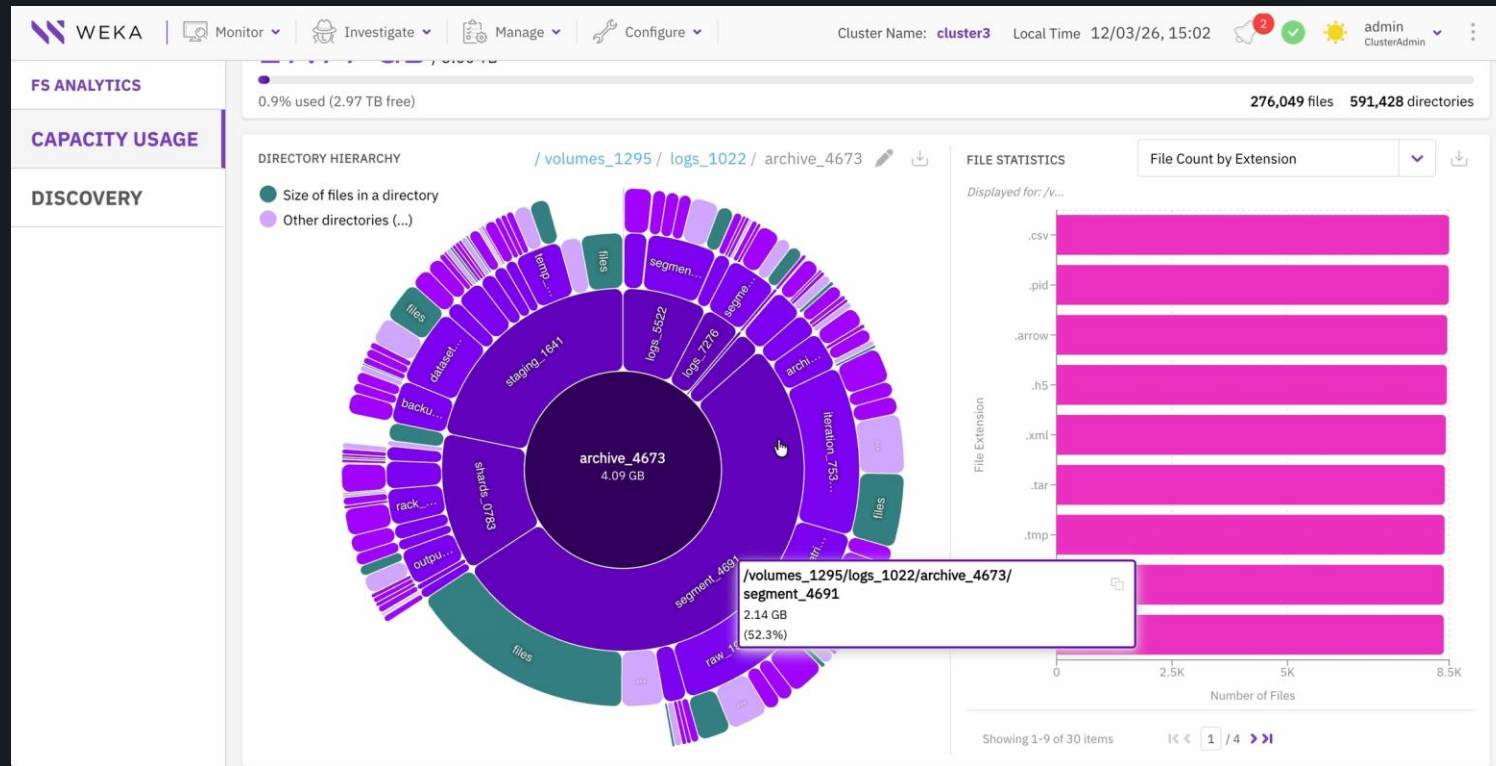
Multi-tenancy — Composable Clusters

Hardware-level isolation per tenant. Multiple WEKA clusters share infrastructure with dedicated drives, cores, and memory each.



Where this fits · multi-PI labs · GPU-as-a-service · HPC service providers · regulated environments needing hardware separation

Data Catalog — your namespace, indexed



"What hasn't been touched in 6 months?"

File age & access-time analysis. Surface candidates for tiering, archival, or deletion.

"Where did all the capacity go?"

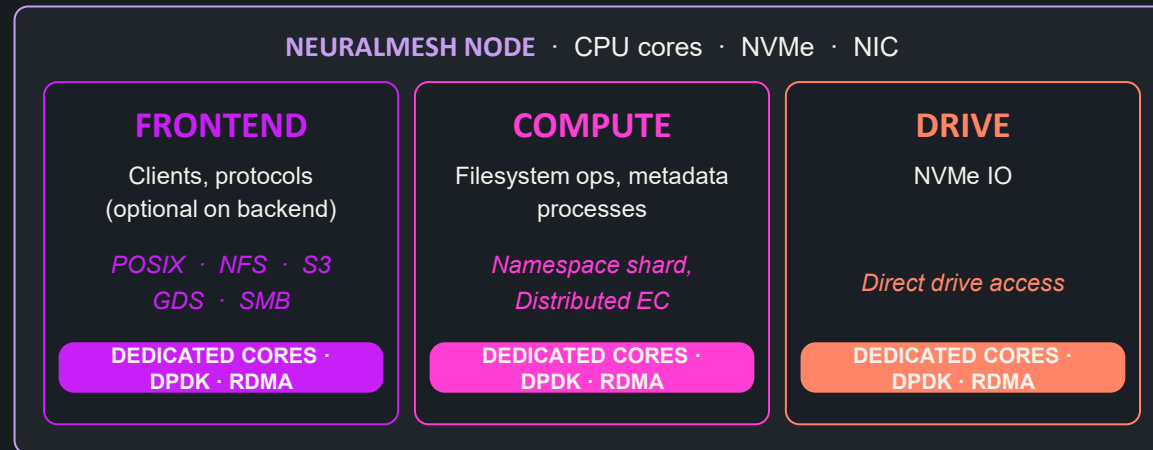
Capacity by directory, extension, and file type. One query, no external scan.

"Who's burning my scratch?"

Top users and groups by storage consumed. Quota conversations grounded in data.

Process-oriented architecture

Predictable performance, linear scale. Userspace containers, no kernel context switches.



Linear scale

Add servers, add bandwidth.

Adding a server contributes proportional compute and drive capacity. Frontend scales when you need it. Verified at 2K+ backend nodes per cluster.

Resilience by design

Failure is local, not global.

Metadata distributed data protection — and distributed striping across failure domains. Losing one node affects a fraction of each stripe. The system keeps running.

Same software. Same architecture. From a single rack to multi-exabyte deployments.

NeuralMesh AIDP — NVIDIA reference design

The AI-factory landing zone — integration risk pre-solved.

NVIDIA
compute, fabric, AI software

GPU systems · Spectrum-X / InfiniBand · BlueField · NIM · NeMo · GPUDirect Storage

WEKA
data layer, end-to-end

NeuralMesh — single namespace, every protocol, microsecond latency

+ Augmented Memory Grid · NeuralMesh Axon (converged) · Data Catalog · Multi-tenancy · Multi-cloud portability

Same software image — on-prem (Dell, HPE, Supermicro, Lenovo, WEKApod) and AWS, Azure, GCP, OCI

CUSTOMER
workload, data, identity

Models · data · MLOps · identity (SAML / OAuth federation) · governance

WHAT WEKA ADDS ABOVE THE REFERENCE DESIGN

Augmented Memory Grid

KV-cache extension at memory-class latency

NeuralMesh Axon (converged)

Run on the GPU nodes — no separate storage tier

Multi-cloud portability

Same software, on-prem and cloud

The reference design gets you started. The data layer is what scales.

Thank you.

Same software. Same architecture. Where compute, network, and storage scale in balance.

Find us at the WEKA table — happy to go deeper on any of this.

Bob Vassar · Senior Federal Systems Engineer · bvassar@weka.io