NeuralMesh Data Efficiency

Technical Brief

NeuralMesh Data Efficiency

NeuralMesh delivers data efficiency as a system of six engineered pillars, not a single feature. Data Reduction is the newest pillar: 2.6:1 reduction in production today across mixed workloads, with up to 6× reduction on AI/ML data and under 5% write overhead. Combined with thin provisioning, snapshots, single-hop writes, drive sharing, and AlloyFlash hybrid flash, the platform delivers up to ~1.9× higher read throughput, ~13× higher write throughput, and 3× lower latency than alternative architectures. NeuralMesh has become the most economically defensible all-flash architecture for AI and HPC.

Overview

Modern AI and HPC environments require a fundamentally different approach to storage efficiency. Data is not only growing rapidly—it is structurally similar, continuously evolving, and generated at massive scale. At the same time, infrastructure must sustain deterministic performance to keep GPUs fully utilized.

WEKA addresses this challenge through a system-level efficiency model. Rather than relying on a single mechanism, the platform combines multiple capabilities—data reduction, provisioning, write path optimization, and media innovation—into a unified architecture that improves both cost and performance simultaneously.

This guide expands on the core efficiency pillars within WEKA and the measurable impact they deliver in real-world environments.

Data Reduction (Similarity-Based)

WEKA's NeuralMesh Data Reduction uses similarity-based compression to identify structurally related data and encode it using shared reference blocks and deltas. Unlike traditional deduplication, which only eliminates exact matches, this approach captures similarity across datasets, checkpoints, and environments—where small changes occur between versions.

© 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.

Reduction is performed asynchronously in the background and fully removed from the write path. Writes complete at full NVMe speed, while reduction work is distributed across the cluster and applied without impacting foreground I/O. Because similarity detection operates across filesystems, redundancy is eliminated globally rather than within isolated volumes.

Field-Proven Impact

- 2.6:1 reduction observed in production deployments (hundreds of TB saved per environment)
- AI/ML workloads commonly achieve ~4–6X reduction due to checkpoint and dataset similarity
- No write performance penalty (typically <5% overhead)
- Reduced physical I/O volume improves effective throughput for checkpoint-heavy workloads

Reduction Ratios by Workload

Reduction ratios vary by workload type based on how structurally similar the underlying data is. The ranges below reflect telemetry from production deployments and reference testing.

Workload	Typical Ratio	Notes
AI / ML training data	~6×	Checkpoints, intermediate activations, embeddings
EDA (electronic design automation)	~8×	Repetitive simulation outputs and netlists
Databases & code repositories	3–5×	Structured data, source trees, build artifacts
Genomics & life sciences	2–3×	FASTQ, VCF, BAM files; raw FASTQ/BAM/CRAM typically close to 1:1
Media & entertainment	~2×	Image sequences; video frames with scene similarity
Text, mixed, and executable files	1.4–1.9× per type	Reference telemetry across file types
Already-compressed files	~1:1	JPEG, MP4, ZIP, gzip, compressed Parquet, pre-compressed weights
Sensor data, IoT, and telemetry	4–6×	Time-series logs with high temporal correlation

WEKA® © 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.

Thin Provisioning

Thin provisioning is implemented at the filesystem level, decoupling logical capacity from physical allocation. Storage is allocated only when data is written, while the system continuously tracks logical and physical usage independently.

This allows environments to safely oversubscribe NVMe capacity based on expected data reduction ratios and workload behavior. Instead of sizing infrastructure for peak logical capacity, customers can provision against actual physical usage, improving efficiency and reducing upfront cost.

Field-Proven Impact

- Enables 2–4× logical capacity expansion when combined with data reduction
- Eliminates overprovisioning of NVMe capacity during initial deployment
- Improves flash utilization and reduces stranded or unused capacity

Snapshots

Snapshots are implemented using a shared block architecture, where point-in-time copies reference existing data blocks and only incremental changes consume additional capacity. This allows snapshots to be created instantly without copying data.

Because AI workflows generate frequent checkpoints and iterative dataset versions, snapshots preserve multiple states of data with minimal overhead. Combined with similarity-based reduction, repeated structures across snapshots are further compressed, compounding efficiency gains.

Field-Proven Impact

- Near-zero capacity overhead for unchanged data
- Efficient retention of dozens to hundreds of checkpoints
- Example: consecutive model checkpoints (~140GB each) differ by <5%, minimizing incremental storage consumption

Single-Hop Writes

WEKA uses a single-hop write architecture, where data and parity are written directly to drives in parallel streams. This avoids intermediate compute nodes, data redistribution stages, or multi-hop write pipelines common in traditional architectures.

WEKA® © 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.

By eliminating extra network hops and write amplification, the system reduces latency and improves throughput. Data flows directly from client to storage targets in a distributed manner, maximizing parallelism and minimizing overhead.

Field-Proven Impact

- Fewer network hops → lower latency and higher sustained write throughput
- High parallelism (multiple concurrent streams per write)
- Reduced CPU and network overhead compared to multi-hop storage architectures

Drive Sharing (SSD Proxy)

WEKA virtualizes NVMe devices using an SSD proxy architecture, allowing multiple I/O containers to share physical drives through multi-queue access. This removes the traditional one-to-one mapping between compute and storage resources.

By enabling many cores to access the same drive concurrently, the system increases parallelism and improves utilization of available SSD bandwidth. This is particularly important for decompression and read operations, where additional CPU resources can directly improve throughput.

Field-Proven Impact

- Higher SSD utilization through parallel access across multiple containers
- Supports many concurrent I/O contexts per device
- Enables more CPU cores to participate in decompression → higher read throughput

AlloyFlash (Hybrid Flash)

WEKA AlloyFlash combines TLC (high-performance) and QLC (high-density) flash within a single namespace. Data placement is handled automatically, with hot data served from performance media and warm data stored on higher-density capacity media.

Unlike traditional tiering, this approach does not require manual policies or data movement between tiers. Data remains accessible through a unified namespace, preserving performance while optimizing cost and density.

Field-Proven Impact

- Lower cost per TB compared to all-TLC configurations
- Increased storage density without sacrificing performance

WEKA® © 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.

- No tiering penalties, rebalancing overhead, or operational complexity

Data Reduction Guarantee

WEKA pairs a contractual data reduction commitment with a throughput guarantee on comparable hardware. Both figures are grounded in a pre-sales scan of the customer's actual data rather than a vendor average, and both are written into the same agreement.

Every WEKA quote in a Guarantee-eligible deal includes four numbers: usable capacity, expected data reduction ratio, effective capacity, and guaranteed throughput with Data Reduction enabled. If realized reduction falls short over a 90-day measurement window, WEKA covers the gap with free WEKA software licenses.

Field-Proven Impact

- Contractual data reduction commitment paired with a throughput guarantee on comparable hardware
- Grounded in the customer's own data rather than a vendor-average claim
- Shortfalls remedied with free WEKA software licenses over a 90-day measurement window

DRET: Grounding the Guarantee in Real Data

The Data Reduction Estimation Tool (DRET) is a non-destructive static binary that analyzes a representative sample of the customer's actual data using the same similarity-hashing and grouping algorithm that runs in production. The resulting projection becomes the contractual data reduction number, and because the same engine runs in production, realized reduction tracks the projection closely.

Field-Proven Impact

- Runs against a representative sample of the customer's own data, not synthetic benchmarks
- Uses the identical similarity-hashing and grouping algorithm that runs in production
- Produces the data reduction number that goes directly into the contract

System-Level Efficiency, Not Feature-Level Optimization

WEKA delivers data efficiency through a unified, distributed architecture where multiple capabilities work together—not in isolation—to maximize both performance and capacity. Similarity-based data reduction eliminates redundancy across the cluster, while thin provisioning and snapshots extend logical capacity and preserve multiple data versions with minimal overhead.

WEKA® © 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.

At the same time, single-hop writes, drive sharing, and hybrid flash optimize how data is written, accessed, and stored—ensuring that efficiency gains do not come at the expense of performance. The result is a system that improves both economics and performance simultaneously.

Combined System Impact

- Up to ~1.9X higher read throughput
- Up to ~13X higher write throughput
- 3X lower latency compared to alternative architectures

WEKA® © 2026 All rights reserved. WEKA and the WEKA logo are registered trademarks of WekaIO, Inc.
Other trade names used herein may be trademarks of their respective owners.