NeuralMesh Data Reduction

Technical Brief

AI and HPC workloads have made storage both a cost bottleneck and a performance dependency. Data growth from training pipelines, simulations, and multi-modal datasets is driving petabyte-scale demand, while NVMe costs continue to rise. Traditional data reduction techniques cannot be applied in these environments without introducing latency, throughput variability, and operational instability, making them unsuitable for GPU-based infrastructure.

NeuralMesh Data Reduction combines similarity-based compression with a background-first processing model that removes reduction from the write path. By capturing structural similarity rather than exact duplication and applying reduction asynchronously across the cluster, the system achieves significant capacity savings without compromising deterministic performance. The result is a storage platform that improves both cost efficiency and performance density, enabling higher GPU utilization, reduced data movement, and more scalable infrastructure.

NeuralMesh Data Reduction carries a contractual 3:1 reduction guarantee. The projection from a Data Reduction Estimation scan on a representative sample of your own data becomes the committed number, and where realized reduction falls short WEKA covers the gap in software licenses.

Executive Overview

Data reduction in modern high-performance environments must be fundamentally rethought. Traditional approaches were designed for static, structured datasets and rely on inline compression or exact-match deduplication. These techniques break down in AI and HPC environments, where datasets are not only larger, but structurally different—composed of model checkpoints, multi-modal data, simulations, and continuously evolving datasets with high degrees of similarity rather than exact duplication.

Enterprises are experiencing rapid data growth driven by AI training and inference pipelines, large-scale simulation workloads such as EDA and CFD, genomics research, and multi-modal data processing. These workloads generate petabyte-scale datasets with extreme throughput

© 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.

requirements. Without a new approach to data reduction, storage infrastructure costs scale linearly with data growth, quickly becoming economically unsustainable.

At the same time, modern AI infrastructure operates at unprecedented scale. Clusters span tens of thousands of GPUs and require multi-petabyte to exabyte-scale storage systems. In this environment, data reduction must operate not just within a node or filesystem, but across the entire cluster, leveraging massive parallelism to identify similarity globally without introducing bottlenecks.

NeuralMesh Data Reduction is designed for this new reality. It combines similarity-based compression with cluster-wide reduction and a background-first architecture that removes reduction from the write path. By distributing reduction work across the cluster and avoiding centralized coordination, the system scales linearly with workload size while preserving deterministic performance.

In production environments, these benefits are already being realized. Field deployments have delivered data reduction of approximately 2.6:1, saving hundreds of terabytes within a single environment. On the strength of those results and the reduction efficiency improvements delivered since, WEKA now stands behind a contractual 3:1 data reduction guarantee.

Crucially, this efficiency is achieved without compromising performance. Because reduction processing is decoupled from the write path, write performance remains effectively unchanged, with under 5% overhead. Sustained throughput is maintained, ensuring that GPUs remain fully utilized and infrastructure efficiency is preserved.

The Storage Constraint in AI Infrastructure

The economics of flash storage have shifted significantly. Enterprise NVMe prices have increased sharply since 2025, with supply constraints expected to persist into 2028. At the same time, AI infrastructure has expanded rapidly, with clusters scaling to tens of thousands of GPUs and requiring multiple petabytes of high performance storage.

At current pricing levels, storage is no longer a secondary component of infrastructure cost. A multi-petabyte flash deployment represents a multi-million dollar investment before accounting for compute, networking, and facility costs. As clusters scale, storage becomes a primary budget constraint.

Moreover, the nature of AI workloads amplifies storage demand. Model training pipelines generate large volumes of checkpoint data, intermediate outputs, and replicated datasets. These workloads are characterized by high throughput requirements and frequent write operations, which place sustained pressure on the storage system.

© 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.

Therefore, any reduction in required physical capacity directly translates into meaningful savings. But, reduction must be achieved without compromising performance. Any degradation in throughput or latency reduces effective GPU utilization and increases overall training time.

Limitations of Traditional Approaches

Traditional storage platforms implement data reduction using inline compression, global deduplication, or a combination of both. These techniques are quite effective in environments such as backup or archival storage, where throughput requirements are moderate and latency variability is acceptable.

In performance sensitive environments, these approaches introduce several limitations.

- Inline deduplication requires global coordination and metadata lookups during write operations. This increases latency and introduces variability in write performance. As systems scale, the metadata structures required to support deduplication grow significantly, creating additional pressure on memory and rebuild workflows.
- Fixed block compression operates on predefined chunk sizes and cannot effectively capture redundancy across block boundaries. While larger compression dictionaries can improve efficiency, they also increase decompression cost, which directly impacts read performance.
- Exact match deduplication relies on cryptographic hashes and is effective only when data is identical. Modern workloads such as AI training, simulations, and analytics produce data that is structurally similar but not identical. Small variations in offsets or values prevent traditional deduplication from identifying redundancy.
- Post-process reduction techniques avoid write path penalties but introduce unpredictable background activity that competes with foreground I/O. This can result in performance variability under sustained load.

These limitations make traditional approaches poorly suited for AI and HPC environments where sustained throughput, low latency, and predictable performance are critical.

NeuralMesh Data Reduction

WEKA treats data reduction as a capacity multiplier that must never cost performance. The system is built on the assumption that foreground I/O stays deterministic regardless of what reduction is doing in the background.

The architecture is guided by several principles:

- Reduction operations must not introduce global coordination on the write path.

© 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.

- The system must scale linearly with cluster size and workload intensity.
- Metadata growth must remain bounded to avoid long term scalability issues.
- Background processing must be isolated in a way that prevents interference with application performance.

These principles lead to a separation between data ingestion and reduction processing. Writes are handled at full device speed without waiting for compression or deduplication decisions. Reduction work is performed asynchronously, allowing the system to amortize computational cost over time.

The result is an architecture that delivers capacity efficiency while preserving the performance characteristics required for demanding workloads.

Similarity-Based Compression and Data Layout

At the core of NeuralMesh Data Reduction is a similarity-based compression model that extends beyond traditional deduplication techniques. Instead of relying on exact matches, the system analyzes structural characteristics of data blocks and identifies groups of blocks that can be compressed efficiently together.

Multiple similarity hashes are derived for each block. These hashes describe different aspects of the data and allow the system to identify relationships between blocks that are not byte identical. By grouping similar blocks, the system constructs shared compression dictionaries that capture common patterns across the dataset.

Compressed data is stored using a reference and delta model. Reference blocks serve as the basis for compression dictionaries, while delta blocks encode differences relative to these references. The system places references and associated deltas in close proximity to minimize read amplification and maintain high throughput during decompression.

During read operations, this model requires reconstruction of data by retrieving both reference blocks and associated delta blocks, followed by decompression. This can increase the number of I/O operations and CPU utilization compared to non-reduced reads.

However, this overhead is offset by the reduction in physical data that must be read from storage, often resulting in comparable or improved effective throughput depending on workload characteristics.

This approach enables significantly higher reduction ratios on machine generated and semi structured data compared to traditional methods, while keeping metadata overhead manageable.

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.

Techniques to Eliminate Redundancy at Scale

The architectural choices described above are realized through a set of complementary techniques that operate across the cluster. NeuralMesh Data Reduction combines block-variable differential compression with cross-filesystem deduplication to identify and eliminate redundancy at scale. These techniques are not independent features, but direct consequences of the system's similarity-based and background-first design.

- Block-variable differential compression operates at fine granularity, analyzing data in fixed 4K blocks and encoding each block as a variable-length delta against a reference block that serves as its compression dictionary. Rather than compressing fixed chunks in isolation, the system groups structurally similar blocks and compresses them against a shared reference. This captures redundancy that fixed-block compression misses, both across block boundaries and across datasets, and it is particularly effective on the machine-generated data common in AI and HPC environments.
- Cross-filesystem deduplication extends this model beyond individual datasets. During ingest, the system computes similarity fingerprints that describe the structure of each block. These fingerprints are later used to identify related data across all data reduction enabled filesystems in the cluster. As a result, redundancy is eliminated not only within a single workload, but across model checkpoints, replicated training datasets, container layers, and environment snapshots.

Together, these techniques enable the system to capture similarity rather than exact duplication, which is essential for achieving high reduction efficiency on modern machine-generated data.

Separation of Control Plane and Data Plane

WEKA maintains a clear separation between control plane and data plane responsibilities within the data reduction architecture.

- The data plane is responsible for performing inline lightweight operations such as zero block detection and ensuring that read and write paths remain deterministic. It also handles decompression during read operations, maintaining high throughput through parallel processing across drive cores.
- The control plane manages capacity accounting, snapshot policies, and observability. It tracks logical and physical usage independently and provides accurate reporting for operational planning. Importantly, control plane activities do not interfere with data path operations, ensuring that management functions do not impact application performance.

This separation contributes to system stability and simplifies operational management at scale.

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.

Performance Characteristics

NeuralMesh Data Reduction is designed to operate without introducing measurable impact on foreground write performance. Because compression and deduplication decisions are deferred to background processes, writes proceed at full NVMe speed. In practice, write performance impact is negligible, typically observed at less than ~5% compared to non–data-reduced operation, and often effectively identical from an application perspective.

Background reduction activity consumes CPU resources, but it is scheduled with lower priority relative to user I/O. This allows the system to adapt to available compute capacity without affecting application workloads. In environments operating near peak utilization, careful consideration of CPU headroom is recommended to ensure that background processing does not contend with foreground tasks.

Decompression performance is optimized for parallel execution. Each drive core is capable of delivering high throughput during decompression, enabling the system to sustain aggregate read bandwidth at cluster scale. The placement of reference and delta blocks minimizes cross node dependencies and ensures efficient data access.

Read performance reflects a balance between reduced physical data transfer and additional reconstruction work. Typical real-world workloads sustain approximately 3 GB/s per drive. Worst-case per-drive throughput is 1.6 to 1.8 GB/s, depending on data characteristics and reduction ratios.

NeuralMesh Data Reduction Processing Pipeline

Data reduction is a post-process, background activity. This is an intentional architectural choice that preserves write performance:

`Ingest (real-time)`

Data enters the system through a high performance ingest path. During this phase, writes are stored uncompressed to maintain maximum throughput at full NVMe speed. At the same time, the system computes similarity hashes for each block. These fingerprints capture structural characteristics that will later be used to identify related data.

`Clusterization (background)`

The system groups similar 4K blocks across all DR-enabled filesystems, running at lower priority than user I/O. Background tasks are capped at 5% of overall cluster CPU, and only one reduction or defragmentation task runs cluster-wide at a time.

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.

Compression & write-back

Grouped blocks are compressed, and the results written back to the drive tier. Original uncompressed blocks are marked for deletion.

Defragmentation

Invalidated uncompressed blocks are permanently reclaimed when sufficient quantities accumulate, returning physical capacity.

Decompression on read

Drive containers decompress data inline, transparently, before serving it to clients. Applications require no modification.

Applications are not aware of the reduction process and do not require modification. This pipeline ensures that write performance remains unaffected while enabling continuous optimization of stored data.

Data Reduction in AI Workflows

AI training environments present a workload profile that is particularly favorable for data reduction. Checkpoint cadence is the single largest driver of AI storage growth, and it happens to be almost perfectly suited to similarity-based reduction, because consecutive checkpoints are near-identical by construction. By reducing the physical footprint of these workloads, data reduction allows more data to reside within the high performance storage tier. This supports faster iteration cycles, improved collaboration across teams, and more efficient use of infrastructure resources.

NeuralMesh Data Reduction for AI Training	
Model checkpointing cadence	Frequent checkpointing produces multiple versions of model weights with minimal variation. Training runs save checkpoints every N steps; adjacent checkpoints differ by only a small fraction of total weights. A 70B parameter model checkpoint at BF16 precision occupies ~140 GB — and consecutive saves may differ by less than 5%.
Shared base models	Multiple teams fine-tuning from the same foundation model (Llama, Mistral, Falcon) store near-identical base weight tensors. Cross-filesystem deduplication eliminates this redundancy automatically.

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.

Dataset repetition	Training datasets are frequently duplicated across experiments, versions, and preprocessing pipelines. Deduplication captures these savings transparently.
Container and environment layers	Containerized environments introduce further duplication through repeated storage of software layers and dependencies. These patterns create a high degree of similarity across the dataset, enabling efficient compression and deduplication. Docker images, Python environments, and CUDA toolkit files stored on the cluster are identical across nodes and jobs.

Field-Proven Results

Deployments of NeuralMesh Data Reduction have demonstrated measurable results across real-world environments. In one representative production deployment spanning multiple filesystems, the system achieved a 2.6:1 data reduction, resulting in 765 TB of saved physical capacity. These results, together with the reduction efficiency improvements delivered in subsequent releases, are the basis for WEKA's contractual 3:1 guarantee. Where realized reduction falls short of the committed ratio, WEKA covers the difference in software licenses. These results highlight the practical impact of similarity-based compression in enterprise environments.

Reduction Efficiency Across Workloads

Achievable ratios depend heavily on workload characteristics: repetition, similarity, and data type. It is critical to understand that not all datasets benefit from data reduction. In some environments, existing datasets may yield little benefit, making upfront workload qualification essential. The effectiveness of data reduction depends on the characteristics of the workload:

- Data that exhibits repetition, structural similarity, or incremental changes over time is particularly well suited for similarity-based compression.
- AI and machine learning workloads often achieve high reduction ratios due to the nature of model checkpointing (snapshots of weights saved frequently during training runs) and dataset reuse. Consecutive checkpoints contain only small differences relative to previous versions, exhibiting high cross-version similarity, and making them excellent candidates for deduplication. A training run saving checkpoints every 100 steps may write dozens of

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.

nearly-identical weight tensors. Across teams sharing a cluster, the same base models, datasets, and Docker layers are duplicated further.
- Electronic design automation (EDA) workloads also benefit from high levels of reduction due to repetitive simulation outputs and structured data patterns.
- Other domains such as genomics, analytics, and time series data exhibit moderate to strong reduction depending on data structure and processing pipelines.

Conversely, data that is already compressed, encrypted, or inherently random does not benefit significantly from reduction and is better suited for separate storage configurations.

Typical Data Reduction Ratios by Workload

The table below reflects expected ratios based on WEKA's field experience across various workload categories:

Workload Type	Published Reduction Ratio (up to)	Notes
AI / ML model training data	**~6X**	Checkpoints, intermediate activations, embeddings
EDA (electronic design auto.)	**~6X**	Highly repetitive simulation outputs & netlists
Genomics & life sciences	**2–3X**	FASTQ, VCF, BAM files; redundancy across samples
Databases & code repositories	**3–5X**	Structured data, source trees, build artifacts
Sensor data / IoT / telemetry	**4–6X**	Time-series logs with high temporal correlation

Economic Impact and Infrastructure Efficiency

Data reduction directly influences the economics of high performance storage deployments. By reducing the physical capacity required to store a given dataset, the system lowers capital expenditure on storage infrastructure like flash.

Economic Impact and Infrastructure Efficiency of NeuralMesh Data Reduction	
Significant cost savings	In large scale deployments, even modest reduction ratios translate into substantial savings. A reduction factor of three effectively reduces required raw capacity to one third of the

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.

	original footprint. At current NVMe pricing levels, this represents significant cost avoidance in multi-petabyte environments.
Lower power consumption	Beyond direct cost savings, reduced physical capacity leads to lower power consumption, decreased rack space requirements, and simplified operational management. Fewer drives reduce the number of potential failure domains and improve overall system reliability.
Faster job completion times and improved utilization	Data reduction also contributes to performance efficiency. Smaller physical datasets reduce the volume of data movement during checkpointing, replication, and rebuild operations. This results in faster job completion times and improved utilization of compute resources.

Example Economic Case

At the guaranteed 3:1 ratio, one-third of the raw NVMe capacity serves the same logical footprint. With enterprise NVMe at $500 to $1,000 per TB, that lands directly on capital expenditure as fewer drives purchased.

Scenario	Raw Capacity	Raw NVMe Cost[1]	With 3× DR	Savings
Single AI training pod (500 TB logical)	500 TB	~$500K	~167 TB raw	**~$330K**
Mid-size GPU cluster (2 PB logical)	2 PB	~$2.0M	~667 TB raw	**~$1.3M**
AI Gigafactory tier (10 PB logical)	10 PB	~$10.0M	~3.3 PB raw	**~$6.7M**

[1] *Based on ~$1,000/TB enterprise NVMe (current 2026 market rate for high-capacity data center NVMe)*

Beyond raw capital savings, data reduction delivers operational benefits that compound over time: fewer drives means fewer failure domains to manage, lower power draw, reduced rack space, and smaller physical footprint, all of which matter at the scale of a 100,000-GPU Gigafactory.

Summary

Three things separate NeuralMesh Data Reduction from conventional approaches. It matches blocks by structural similarity rather than exact content, which captures redundancy that

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.

exact-match deduplication cannot see. It works cluster-wide, spanning every data-reduction-enabled filesystem, so redundancy is eliminated across teams, projects and checkpoints rather than within a single dataset. Lastly, it runs entirely in the background, so writes commit at native NVMe speed with under 5% overhead and reads sustain approximately 3 GB/s per drive.

The results hold in production. A multi-filesystem deployment delivered 2.6:1, saving 765 TB of physical capacity. AI and ML workloads reduce by up to 6x, driven by checkpoint cadence and dataset reuse. WEKA now stands behind a contractual 3:1 data reduction guarantee. Where realized reduction falls short of the committed ratio, WEKA covers the difference in software licenses. Start with a Data Reduction Estimation scan. It runs the production algorithm against a sample of your own data and returns the ratio you can expect, before you buy a single drive.

Take the Next Step with NeuralMesh

Talk to a WEKA solutions architect about qualifying your datasets, and run a Data Reduction Estimation scan on a representative sample.

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.