WEKA

SOLUTION BRIEF

WEKA Eliminates Costly Checkpoint Bottlenecks

Challenges

- Slow checkpoint writes stall large-scale LLM training
- Frequent checkpointing can overwhelm legacy network file systems (NFS)
- Risk of losing days/weeks of progress if a GPU fails

WEKA Solution

- WEKA's massively parallel architecture accelerates writes
- Concurrency at scale significantly reduces checkpoint latency
- Rapid recovery from partial failures, even in multi-week runs

Benefits

- Shorter training cycles — better GPU efficiency
- Reduced risk of reruns or lost model progress
- Consistent performance for small or large model checkpoints

Transforming LLM Training with WEKA Data Platform's Fast, Resilient Parallel Write Strategy

Frequent checkpointing is critical for artificial intelligence (AI/ML) teams training Large Language Models (LLMs), especially when running multi-week jobs across hundreds or even thousands of graphics processing units (GPUs). Checkpointing involves periodically saving the model's state so that progress isn't lost if a node fails or the job interrupts. However, these periodic writes can quickly turn into major delays if your storage layer isn't engineered for massive parallel throughput, especially for writes.

WEKA addresses this challenge by providing a high-performance data platform purpose-built for AI/ML, delivering ultra-fast checkpoint commits at scale. With WEKA, you can safely checkpoint as often as needed — every 20 minutes or more — without incurring long stalls that idle expensive GPU resources. The result is faster experimentation, tighter risk control, and minimal disruptions to your AI pipeline.

Why Mitigating Risks Matters for AI Development

In modern LLM training, it is common for billions of parameters to span hundreds or thousands of GPU clusters. Conventional advice now suggests checkpointing every 20 minutes or so to minimize loss if a GPU fails mid-training. But these repetitive writes must be completed quickly and simultaneously — standard NFS solutions often choke on such simultaneous data bursts.

When training cycles span two weeks or more, even a moderate failure rate across hundreds of GPUs can cause a total job restart if a usable checkpoint isn't available. This equates to time lost, GPU hours wasted, and a steep opportunity cost in AI research productivity.

Moreover, hardware or software glitches remain a reality in many high-performance GPU clusters, causing random node failures or reboots. The high price of GPUs magnifies the cost of these interruptions to computing throughput. With some HPC environments exceeding $1,000 per hour, losing even a week of training can quickly balloon into tens of thousands of dollars in wasted spend. Frequent checkpointing mitigates this risk by ensuring you can resume almost immediately rather than re-running substantial workloads.

In addition to checkpointing, AI pipelines also typically involve data ingest, pre-processing, on-the-fly validation, and final model commits. Legacy file systems, optimized for one type of workload, may struggle to juggle these concurrent reads, writes, and metadata operations.

50% Faster Checkpoints

WEKA cuts checkpoint latency by ~50% vs. NFS, saving precious GPU hours.

900 Microsecond Write Latency

WEKA consistently delivers sub-900 µs checkpoint writes, so overhead barely registers.

Checkpoint as often as needed for confidence

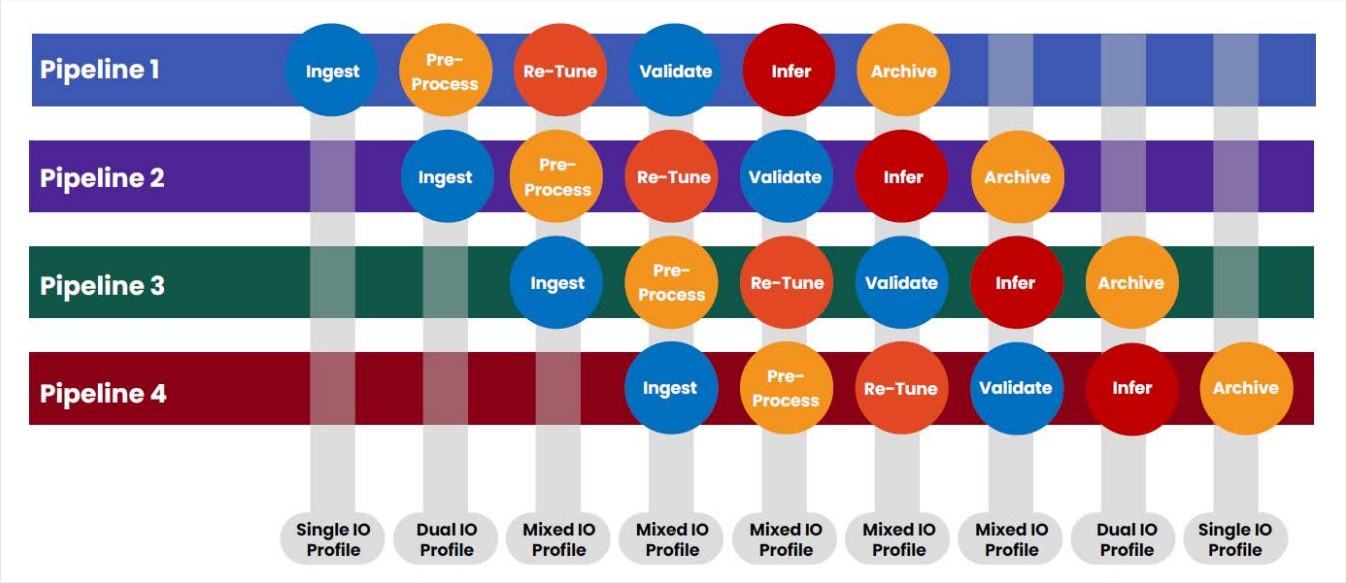
Even if you checkpoint every 20 minutes, WEKA speed keeps GPUs running at full tilt.

> *"A single mid-week GPU failure could roll back an entire week of progress, wasting thousands of compute hours. With WEKA's fast checkpoints, we lose at most 20 minutes."*

WEKA High-Performance Parallel Writes

WEKA uses a distributed metadata engine that distributes checkpoint writes across multiple nodes. Rather than funnel data through a single bottleneck, the WEKA data platform leverages concurrency to create large model snapshots in seconds, not minutes. This architecture is illustrated in the figure below.change simulations. WEKA enables HPC teams to achieve greater accuracy and make impactful discoveries faster by accelerating data-processing workflows.

This concurrency approach ensures that ingest, training, checkpointing, and final commits run simultaneously without causing performance degradation. WEKA keeps the entire pipeline flowing by efficiently handling diverse and unpredictable I/O patterns.

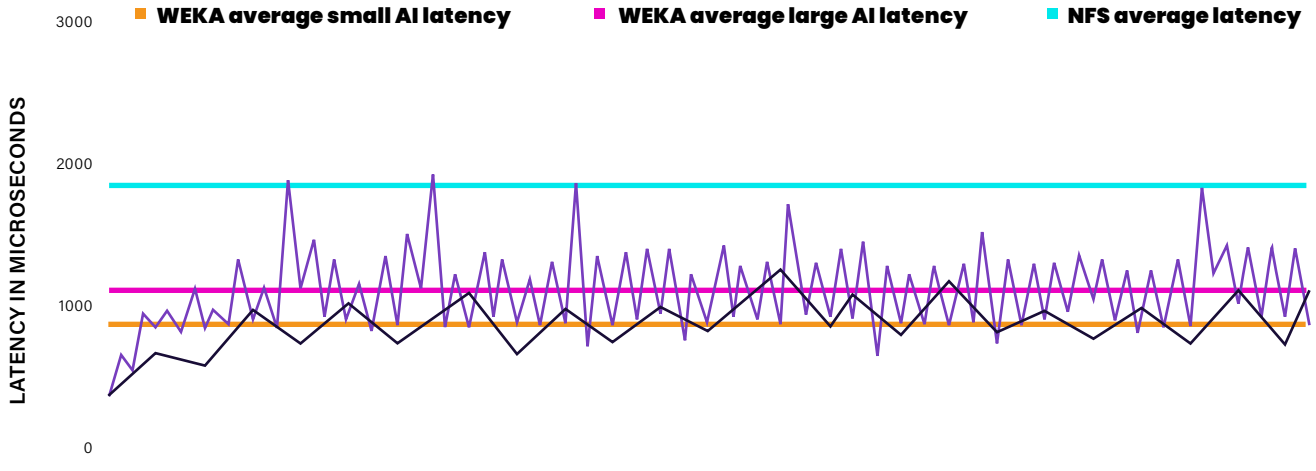


Consistent Latency, Even at Scale

Generative AI workloads: Handle resource-intensive generative AI tasks, including video editing, 3D rendering, and natural language processing, by delivering consistent performance and scaling dynamically to meet demand without bottlenecks.

Demonstrated Performance on Small and Large Models

In testing with two markedly different language models — FastPitch (relatively small) and BioMegatron (very large) — WEKA consistently outperformed a high-end NFS server. The below graph underscores WEKA's dramatic write-speed advantage compared to NFS, providing consistent, low-latency commits regardless of model size.



Checkpoint Latency for a Small AI Model

WEKA processes frequent, smaller checkpoints with sub-millisecond latency.

Checkpoint Latency for a Large AI Model

WEKA handles huge, multi-GB parameter snapshots without slowing

Simplifying Hybrid AI Pipelines with WEKA

Beyond speed, WEKA provides multi-protocol support (POSIX, NFS, SMB, S3) and can run on-prem or in the cloud. Teams can unify their training, validation, and inference pipelines on one data platform — reducing complexity. This is especially valuable for extensive, distributed clusters or multi-tenant environments.

WEKA Zero-Tuning Architecture

WEKA's zero-tuning architecture automatically adapts to changing AI workloads. Whether handling billions of small files or multi-petabyte data sets, you won't have to reconfigure storage layers as your pipelines grow. This capability also eliminates the complexity of manual optimizations, ensuring you can scale capacity and performance on demand without forklift upgrades.

Legacy NFS vs. WEKA Features

Feature	Legacy storage (NFS)	WEKA powered storage
Checkpointing frequency	Limited due to storage performance constraints	No restrictions — checkpoint as often as needed
Parallel write handling	High latency, inconsistent throughput	High-speed, optimized parallel writes
Scalability	Degrades as model and dataset size increases	Seamlessly scales to meet AI demands
AI workflow integration	Requires manual tuning and workarounds	Native support for AI training frameworks

Why WEKA?

Rather than forcing you to design your pipelines around the limitations of legacy NFS, WEKA helps you checkpoint early and often to safeguard progress. Engineered for parallelism at enormous scales, WEKA keeps job stalls to a minimum, no matter how many GPUs are used or how big the model is. By eliminating I/O bottlenecks, organizations can train faster, reduce risk, and confidently scale to multi-week runs on multi-petabyte datasets.

Whether you're fine-tuning a smaller model like FastPitch on a single GPU or orchestrating a BioMegatron run across hundreds of HPC nodes, WEKA consistently delivers sub-millisecond latency. Faster checkpoints mean less idle GPU time and minimal risk of losing days or weeks of work to hardware or software glitches, helping your AI initiatives remain cost-efficient and time-effective.

WKA428-01 04/25