Faster, More Affordable AI Training in the Cloud

Stability AI is a visionary open-source generative AI company on a mission to build an intelligent foundation to activate humanity's potential. The company delivers breakthrough, open-access AI models with minimal resource requirements for imaging, video, 3D, language, code, and audio.

Delivering fast time to market for features and new versions of the Stable Diffusion model requires access to the fastest GPU infrastructure. However, Stability found the scale and performance limitations of their legacy Lustre file systems deployment in the cloud limited their ability to utilize their GPU infrastructure fully and drove big cost surprises.

Stability AI turned to WEKA—the leading AI-native data platform- to develop a new approach to high-performance data for AI model training and tuning designed to improve resource efficiency for the AI model training environment— WEKA Converged Mode for AWS. Unlike traditional architectures that run data on a separate infrastructure set, WEKA Converged Mode runs data and AI model training on the same compute, network, and storage resources. Using this novel approach, Stability AI was able to dramatically reduce data infrastructure costs, increase GPU utilization, and accelerate time to market for new capabilities.

Stability AI—Leading the AI Model Revolution

Stability AI, the leading open-source generative AI model company, aims to deliver the AI building blocks that will unlock human potential. By providing modular, open models in every modality, Stability AI can deliver new AI models to market faster, more sustainably, and with greater safety controls. Stability AI customers can bring pre-trained models into their domain and fine-tune them using their own IP-rich datasets in a secure and controlled manner. This approach improves AI researchers' ability to deeply understand model dynamics, reduce the risk of model hallucination, and build AI applications that are safer and more responsible. By creating multiple modular open-source AI models, enterprises have the building blocks they need for next-generation applications. Further, the smaller model sizes developed by Stability AI are typically more resource-efficient, lending itself to a greener approach to AI development.

stability.ai

Challenges

- Low utilization of GPU infrastructure
- High costs for Lustre-based data environment
- Unexpected spikes in storage costs due to over-provisioning Lustre

Solution

- WEKA Data Platform Converged Mode on AWS
- WEKA cluster running on ~400 Amazon EC2 P4d instances
- 2.2 PB NVMe Flash and 5 PB object storage, all in AWS drive Stable Diffusion model training

Benefits

- Achieve 93% GPU Utilization efficiency during model training runs
- 95% reduction in cost per TB for data infrastructure
- 35% improvement in model training times (eliminated three weeks out of a 60-day training cycle)

The Business Challenge:
Winning the race in Generative AI for the People by the People

Stability AI seeks to win the generative AI race based on market speed, model accuracy, responsiveness, and accessibility. In the Generative AI space, the pace of innovation, already off the charts, is expected to remain high for years as foundation model providers create the AI building blocks that will drive next-generation applications. "The most important thing right now is speed to market," says Richard Vencu, MLOPs Lead at Stability AI. "We're shipping a new model every week, challenging the current state of the art for open-source and even closed-source models simultaneously." However, the fastest GPU infrastructure needed to train models is expensive and scarce. Central to managing this dynamic is resource efficiency. Efficient infrastructure use helps Stability AI control costs, manage explosive company growth, and deliver more capabilities to market faster.

> " We had a lot of NVMe drives not being utilized fully, and we felt we could do better. There's no point in paying for other NVMe drives until you've used what you already have."

The Technical Challenge:
Squeeze Everything from the Infrastructure

To deliver on this mission, Stability AI is innovating at every layer of its technology stack, down to the infrastructure. Most AI companies today focus on raw infrastructure horsepower in terms of fast networks and massive clusters of the fastest GPUs available. However, raw speed is just the starting point. "What we do is make the models smaller. They require less compute and infrastructure, which helps us control costs," says Vencu. "The models are faster to train, which helps us be more nimble and give the customer a better experience."

However, the fastest networking, compute, and GPUs required to support model training and tuning have become scarce commodities, commanding high prices. "We live in a scarcity era where accelerators are not readily available for everyone," says Vencu. "I had to develop strategies and build cluster infrastructure, so we're grabbing every piece of GPU available to minimize waste." By implementing resource sharing, backfilling, and job scheduling across multiple projects on the same cluster at the same time, the Stability AI infrastructure team was able to make significant steps toward achieving that goal.

Meanwhile, Stability AI realized that fast model training times require fast and resilient architecture. "The researchers just want their models to converge and predict the best output possible in the shortest time, " says Vencu. "If you're training a model and don't hit that goal, then it fails, and you have to go back and retrain the model." The Stability AI team built automation tools to manage network health, memory, and GPUs in the training environment and started to see progress. However, a crucial sticking point in this journey was in data storage.

Rapid growth in researcher and customer interest drove explosive data growth in the Lustre environment. However, file system management and clean-up associated with this growth increased performance overhead levels, impacting the training cluster and slowing down model training times. Over-provisioning the Luster environment helped alleviate the situation initially but led to unexpected cost overruns in the Lustre system. Says Vencu, "You'd push a button, and the cost would just blow up like $100 grand per month in a matter of hours." So, the Stability AI team embarked on a series of initiatives to get data costs back under control without impacting the AI researchers' work—starting with a mandate from the CTO to trim down data sets. However, as Vencu points out, top-down mandates for data clean-up rarely go according to plan. "Scanning Lustre was a big headache. We were consuming tens of thousands of CPU hours every month just to maintain the content of the file system."

> " We got 15 times more capacity at about 80% of the cost when we switched to WEKA."

The Journey:
WEKA Converged Mode on AWS

In early 2023, Stability AI started looking for a new data storage solution that could help them get control over their data costs. The team conducted a successful initial POC to establish WEKA's baseline performance. This step showed that a small 6-node WEKA cluster could deliver 1.3 million IOPs and 18 GBps second of bandwidth. Based on this minimum spec size test, the Stability AI team confirmed that a WEKA solution would meet or exceed their performance requirements.

With a baseline comfort level in WEKA's ability to support their needs, the Stability AI team wanted to go much further with a mindset toward their mission to get hyper-efficient in resource use, control costs, and increase environment stability. The partnership between WEKA and Stability AI engineers was evident from the start. "We knew when we started the POC that WEKA would be performant," says Vencu. "We were most impressed by how closely the WEKA engineering team was willing to work with us." The Stability AI infrastructure team had already considered novel ways to increase utilization of their GPU infrastructure. Says Vencu, "We had a lot of NVMe drives not being utilized fully, and we felt we could do better. There's no point in paying for other NVMe drives until you've used what you already have."

So, working together during the spring and summer of 2023, the WEKA and Stability AI engineering teams developed WEKA Converged Mode for AWS. In Converged Mode (see Figure 1), the data storage environment resides on the same infrastructure resources as the model training environment. Contrast this with a traditional architecture where the model training (or applications) and data storage reside on separate silos of infrastructure. This approach aimed to drive significant cost savings across their data infrastructure, increase GPU utilization to squeeze even more usage out of those expensive and hard-to-acquire resources and improve the resilience of the model training environment.

WEKA Converged

Lustre Dedicated Storage
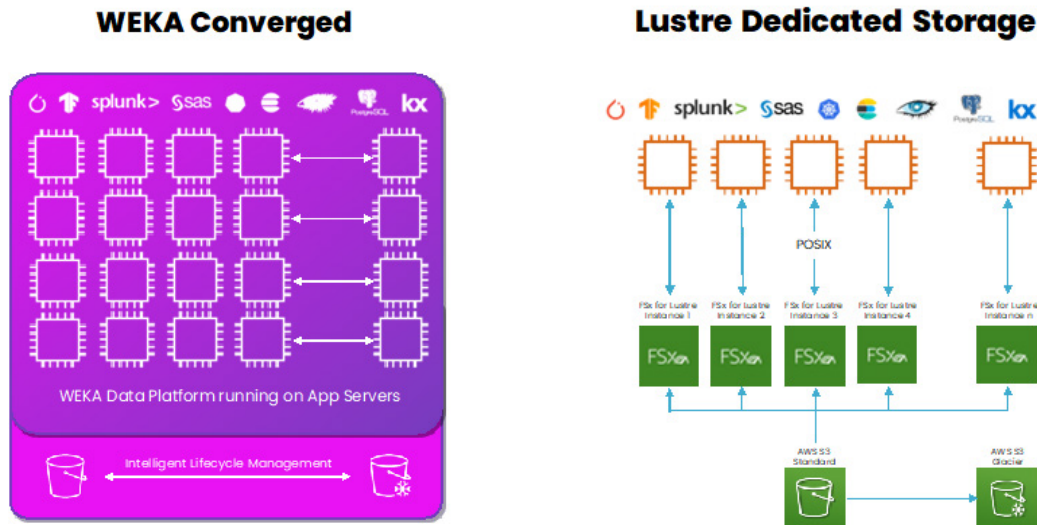
FIG. 3 WEKA Converged Mode for AWS versus Traditional Lustre

The Deployment

Today, Stability AI uses WEKA Converged Mode on AWS for its generative AI model training and tuning environment. The Stability AI environment consists of a fleet of GPU-accelerated instances in AWS using a mix of P4d instances with NVIDIA A100s and AWS P5 instances with NVIDIA H100s. The WEKA data environment consists of 2.2 PB NVMe flash and 5 PB Object storage. The NVMe drives attached to the P4d and P5 instances provide the 2.2 PBs of usable flash storage that comprise the high-performance layer. The single WEKA namespace extends to 5 PB of Amazon S3 object storage for massive, affordable capacity. Stability AI relies on the zero-tuning, zero-copy architecture in WEKA to eliminate the need for MLOps teams to tune the performance of the storage environment or move data between silos tuned for the various stages of their model training process—a common occurrence in the previous deployment.

The Benefits:

Resource Efficiency that Drives Massive Cost Savings at Scale

Using WEKA, Stability AI transformed its infrastructure strategy, driving increased resource utilization that reduces data infrastructure costs while supporting the mission to deliver open, safe, and sustainable AI to everyone. It's also improved the training environment's stability, helping the research team speed innovations to market faster. After switching to WEKA Converged Mode on AWS, Stability AI estimates they reduced their storage costs by 95% on a cost-per-TB basis while improving bandwidth and latency. Says Vencu, "What I love about WEKA is I don't worry about accidentally overprovisioning the file system and incurring catastrophic costs we didn't account for." With WEKA, most AI training data is retained in low-cost object storage on Amazon S3 while automatically tiering data up into the high-performance NVMe flash tier as needed to complete a model training run—all without intervention by researchers, infrastructure, or MLOps teams. The Stability AI team has also seen those improved economics magnified further by running the WEKA software on the same infrastructure they had already procured for their model training environment. "When we switched to WEKA Converged Mode, we got 15 times more cloud storage capacity at about 80% of the previous cost." says Vencu.

As expected, the Stability AI team is also seeing overall resource utilization of their GPU-driven AI infrastructure increase dramatically. "We can reach 93% GPU utilization while running our model training environment on WEKA," says Vencu about his testing with WEKA. He explains how this is possible: "In AI model training, you have petabytes of data moved for every training run. We want to ensure that the storage sits on the same network where the training cluster sits." This strategy brings the data set used in model training as close to the compute infrastructure as possible—minimizing network latency. Vencu concludes, "Converged mode is all about having data on the same system as the GPUs."

Improvements in resource efficiency and GPU utilization also help support sustainable AI goals. "There's no more wasted memory now. Before, you'd have all this GPU memory we weren't using," according to Vencu. "Our goal is to optimize the infrastructure we already have—which WEKA has helped us do."

Operationally, the Stability AI research teams were also pleased with how much more stable the model training environment was and how that quickly translated into improvements in time to market for new features. "My researchers want to buy me some beers now," says Vencu, who supports the day-to-day operations on behalf of the Stability AI research team. The Stability AI team estimates that the improvements in the reliability of their environment have translated into a significant acceleration in their model training times. "With the improved stability that WEKA enables, we've accelerated our model training times by 2 to 3 weeks out of a 12-week training cycle," says Vencu. "That faster time to market is a huge value for us."

Another way WEKA is supporting the Stability AI mission is through increased flexibility and agility. "What I love about WEKA is that it's completely cloud and can work on-prem. So I can be cloud agnostic and not tied to storage hardware on-prem," says Vencu. "As we try to get AI anywhere, it helps to have WEKA so I can deploy anywhere." Because the same WEKA software runs on any cloud—AWS, Azure, Google, or OCI—and any commodity on-prem hardware, Stability AI recognizes a significant improvement in flexibility and choice.

> ↗ **Learn more at weka.io/customers/stability-ai**

About the WEKA Data Platform

The WEKA® Data Platform removes the barriers to data-driven innovation through its advanced software architecture optimized to solve complex data challenges and streamline the data pipelines that fuel AI, ML, and other modern performance-intensive workloads.

The design philosophy behind the WEKA® Data Platform was to create a single architecture that runs on-premises or in the public cloud with the performance of all-flash arrays, the simplicity and feature set of network-attached storage (NAS), and the scalability and economics of the cloud. Whether on-premises, in the cloud, at the edge, or bursting between platforms, WEKA accelerates every step of the enterprise AI data pipeline – from data ingestion, cleansing, and modeling to training validation or inference.

Mind-bendingly fast. Seductively simple. Infinitely scalable. Sustainable. Spanning edge, core, hybrid, and cloud. The WEKA Data Platform helps to overcome complex data challenges and accelerate next-generation workloads to unleash your organization's imagination, creativity, and potential.

WEKA   weka.io | 844.392.0665 | ✉ in 𝕏 f ▶