WEKA

CASE STUDY

How Synthesia Maximizes GPU Efficiency and Accelerates AI Model Training with WEKA on AWS

About Synthesia

Synthesia was founded in 2017 by a team of AI researchers and entrepreneurs from UCL, Stanford, TUM, and Cambridge. In just four years, the company has become the world's leading enterprise AI video communications platform – making video creation, collaboration, and sharing simple and intuitive without the need for cameras, studios, or expensive editing software. More than 55,000 businesses, including more than half of the Fortune 100, use Synthesia to create engaging video content with AI avatars and voice-overs virtually indistinguishable from human actors.

The Business Challenge:

Build Lifelike Video Content With Generative AI

Synthesia provides text-to-video and text-to-speech capabilities that are intuitive to use and simple to get started with, providing AI-generated avatars to create lifelike video content without requiring camera equipment or human actors. The Synthesia Studio provides an intuitive interface to craft compelling video content. Organizations can match avatars with voices, display graphics, and apply animations to create engaging and polished videos.

Artificial intelligence research is central to Synthesia's service. Synthesia's AI researchers continue to push the boundaries of what's possible, constantly training their models to generate AI avatars that produce more realistic renditions of human movement and speech. New model versions will provide even more lifelike avatars, while continuous model tuning expands the catalog of avatars and graphics content available to Synthesia customers. Synthesia researchers constantly push to improve their models through continued model training cycles and new data sets to support this effort.

synthesia

Challenges

- Manual data management processes slowed AI research
- Legacy storage performance limitations bogged down AI model training
- Storage bottlenecks resulted in low GPU utilization

Solution

- WEKA Data Platform on AWS

Benefits

- Eliminated manual data management processes and storage bottlenecks for fast, efficient AI model training in the cloud
- Rapid, easy data migration between AWS Regions
- Increased researcher productivity

The Technical Challenge:

Manual Data Management Slows AI Research

In the text-to-video space, fast time to market is critical to success, with newer, more powerful models constantly coming to market. Synthesia AI researchers rely on the flexibility and agility of AWS GPU infrastructure to enable rapid model training. The compute foundation consists of clusters of NVIDIA H100 Tensor Core GPUs included in Amazon EC2 P5 instances. AWS Parallel Cluster, an open-source cluster management tool, provides resource orchestration for the Synthesia environment. This approach enables the Synthesia infrastructure team to scale resources quickly based on AI researcher project needs and timelines.

downloaded correctly from S3 and the state of the systems was as expected."" says Alex Balan, Technical Lead, ML Platform at Synthesia. This wasted many hours in manual data copy movement and management and introduced data consistency challenges that required constant infrastructure monitoring during model training epochs.

Early in the journey, the team investigated a Lustre-based system in AWS to simplify data operations. However, that approach's poor performance and scale limitations quickly created data bottlenecks in feeding model data

> "During AI model training, WEKA can fully saturate our GPUs, helping us get more performance out of our AI infrastructure and scale when needed."

However, from day one, AI research experienced delays due to manual data operations required in the legacy storage system. AI researchers preferred to store model training data locally on the GPU-enabled EC2 instances to accelerate model training. This provided researchers with the high-performance storage they needed, but it came with serious side effects. First, with no tool available to move data sets between their S3 object store and locally attached NVMe within their P5 instances, the infrastructure team spent hours manually copying the correct data sets from Amazon S3 directly into the training cluster. "There was a lot of manual babysitting. I had to constantly check whether the datasets were

into the GPU cluster, resulting in idle GPU resources. As the team dug in, they realized storage IO throttling and slow metadata look-ups across millions of small files in the AI training data set were slowing down the entire environment. This resulted in low GPU utilization and longer AI model training times than the researchers considered acceptable. "All those GPUs had to sit and wait for the data to be loaded into the cluster," explains Sash Stasyk, ML Platform Engineering Manager at Synthesia. The infrastructure team needed a solution that continued to provide the performance of local instance-attached storage while eliminating the need for manual data management operations.

The Solution:

The WEKA Data Platform Running on AWS

The search for a data platform that could deliver Synthesia's needs quickly led them to WEKA. Synthesia conducted an initial POC running WEKA in their AWS environment, using a roughly 1 TB training data set consisting of approximately 15 million small files. The goal of the POC was two-fold: first, to show the researchers that WEKA could deliver the same level of high-performance data for fast AI model training they were used to with locally attached NVMe storage within their GPU instances. Second, WEKA's zero-copy, zero-tuning architecture could indeed eliminate the need for time-consuming, manual data operations. Explains Sash: "WEKA provided a superior architecture for handling millions of small files, which is what we found."

Today, Synthesia relies on the WEKA Data Platform to drive high-performance data operations for the entire AI model training and tuning environment. WEKA's software runs on a cluster of Amazon EC2 i3en instances; the aggregate NVMe flash attached to the instances—now at 80 TB with the ability to linearly scale with developer needs— provides the high-performance storage layer that delivers "local" performance over a shared file system directly in AWS. The WEKA environment extends to hundreds of TBs of object storage in Amazon S3 for massive capacity at a low cost. Synthesia relies on seamless scaling from WEKA to meet the needs of the rapidly growing research team, which has grown from a team of five to over 30 people working across multiple teams. Researchers are now empowered to load their datasets into the WEKA environment with minimal intervention. WEKA software provides automatic tiering between the S3 object and NVMe flash tiers, dramatically simplifying data operations and accelerating AI training and tuning times.

> "With WEKA, we can migrate our entire dataset with the push of a button."

Outcomes:

Maximum GPU Utilization for Fast, Flexible AI Model Training

With the WEKA Data Platform, Synthesia can run AI model training workloads faster and achieve high GPU infrastructure utilization. "During AI model training, WEKA can fully saturate our GPUs, helping us get more performance out of our AI infrastructure and scale when needed," said Sash.

The WEKA Data Platform has helped the Synthesia team avoid manual data management processes, simplifying data operations and reducing costs. WEKA's zero-copy, zero-tuning architecture eliminates the need for infrastructure teams to manually copy data from Amazon S3 into the training cluster and maintains data consistency across every node. Synthesia is experiencing improved researcher productivity thanks to the quantum leap in performance provided by WEKA. With automated storage tiering, WEKA loads datasets into the local NVME storage resources attached to instances in the model training cluster. "WEKA eliminated our storage bottlenecks. 10 out of 10," said Sash.

> WEKA helped us with engineering peace of mind. We never have to worry about creating a solution ourselves and it saved us a massive deployment effort that would have delayed any delivery of getting the product to market."

Synthesia recently migrated AWS Regions to support continued growth in their GPU-accelerated compute cluster. During the migration, the Synthesia infrastructure team took advantage of the flexibility built into WEKA to make the move simple and easy from a data perspective. Synthesia was able to rapidly stand up an entire new WEKA environment in the new region thanks to WEKA deployment automation via Terraform. Next, the Synthesia team leveraged industry-unique WEKA Snap to Object capability for the data migration. Snap to Object creates a fully usable snapshot of production data - including all data and metadata - and saves that snapshot to an object store in Amazon S3. Synthesia could generate a fully usable copy of their AI training data in just a few minutes and move it from one region to another. Synthesia could run AI model training side by side in both regions to continue running experiments while migrating the compute stack. Once ready, the shift to the new region was seamless, enabling a zero downtime migration between AWS Regions. "With WEKA, we can migrate our entire data set virtually with the push of a button. We continued doing model training across the two regions simultaneously over two weeks while we migrated the compute," says Alex. "WEKA gives us increased flexibility that simply wasn't possible with any other solution we tried."

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.

WKA397-01 07/24