

GPUs are Fast. WEKA Makes Them Faster. 20x Faster.

GPUs have become the gold standard for intensive workloads because they are fast. But what happens when your GPUs spend half their day sitting idle? To be truly fast and efficiently train AI models, GPUs must be fed data continuously.

Data-hungry GPUs sitting idle at the end of a congested data pipeline cost time, money, and insights.

Does your data spend more time standing in line than it does training your model/feeding your GPUs?

GPUs Are Starved for Data

Why? The current AI pipeline was not designed to optimize AI workloads.



70% of model training time is taken up by data staging operations.

Data must "hop" between stages in the pipeline requiring data copying time.

Data Hops

Increase Latency

Each "hop" increases latency and management intervention.

Slower epochs waste valuable GPU resources because they are kept idle waiting for data.

GPUs waiting for data are GPUs that are not training models.

Most AI pipelines spend more time managing data than leveraging it.

GenAl model training requires lots of small files.

Lots of small files create massive amounts of metadata.

More metadata leads to a random read problem dominated by many small IO requests in the first part of the pipeline. Lots of Small Files Slow the Pipeline

> Traditional storage has not been modernized to tackle the current structure challenges of AI model training.

Get the Most Out of Your GPUs

With the WEKA Data Platform, get your data moving through the pipeline, feed your GPUs continuously, and solve more problems, faster.

How does WEKA make the GPU 20X faster?

WEKA collapses the typical GPU-starving "multi-hop" AI data pipeline.

Using a single namespace and zero-copy architecture, data is written once and transparently accessed by all resources in the deep learning data flow.

> No more multiple steps to stage data before training. Remove the Bottlenecks in the Data Pipeline

Efficiently Manage Lots of Small Files WEKA's data layout algorithms distribute and parallelize all metadata.

The result is incredibly low latency and high performance whether the IO size is small, large, or a mixture of both.

 No more early pipeline congestion traditionally caused by lots of small files.

The WEKA Data Platform supports NVIDIA's GPUDirect Storage protocol.

Bypassing the GPU server CPU & memory allows the GPUs to communicate directly with WEKA storage, accelerating throughput to the fastest possible performance.

> No more waiting for your GPUs to get all the data they need to drive your intensive workload. Accelerate Throughput and Drive Performance

Checkpointing

The highly efficient write performance in WEKA is crucial for AI model checkpointing, ensuring training efficiency, scalability, resource optimization, and support for real-time applications, where latency plays a critical role in allowing jobs to continue their training without unnecessary delays.

> > No more unnecessary delays that hold up your training model.

With a modern data pipeline engineered for intensive Al workloads, you can get your data moving, optimize your GPUs, and train your model faster.

And all of that means you get quicker answers to more of the problems that matter most to your business.

Learn more about how <u>WEKA makes the GPU 20X faster</u> and <u>contact us</u> to talk to us about how WEKA can help you get the most out of your GPUs.

