Get the Most Out of Your GPUs

Make Your GPU Run 20X Faster

Challenges

- AI is transforming industries but the growing shortage of GPUs is hindering its adoption
- Data stalls can keep GPUs idle 50% of the time or more
- Feeding GPUs with lots of small files is incredibly inefficient with legacy data storage on-premises or in the cloud

Solution

- The WEKA® Data Platform unlocks the full power of your GPUs to accelerate every step of your training lifecycle for the maximum acceleration of model development.
- WEKA eliminates data stalls and feeds GPUs with enough data–no matter the file size or number–to keep GPUs from idling.

Benefits

- Faster GPU-based AI learning model solutions with greater accuracy
- Increased time-to-value for GPU-driven life sciences, media editing, and a variety of custom HPC applications
- More efficient use of GPUs lower costs and enables projects to start more quickly

AI is experiencing exponential growth and remarkable diversification across various sectors. From healthcare and finance to manufacturing and entertainment, AI technologies are rapidly evolving, transforming industries in profound ways. Machine learning, natural language processing, computer vision, and robotics are becoming integral parts of business operations, enhancing decision-making, automating processes, and driving efficiency. As AI algorithms become more advanced and accessible, organizations are harnessing their power to analyze enormous datasets, personalize user experiences, optimize supply chains, and even make strides in scientific research.

As the Cambrian explosion in generative AI (GenAI) gains momentum, so do the underlying demands on core enabling technologies - data, networks, and most importantly accelerated compute capabilities offered from GPUs. However, a rapidly growing shortage of GPUs is putting companies of every size under increasing pressure. Today, most organizations are focused on the limited availability and consequently high cost to gain access to the most powerful GPUs useful for training Large Language Models (LLMs). However, leading organizations are starting to realize they are missing a trick: the GPUs they do have are sitting idle, starved for data.

The Root of the Problem

The key to unlocking AI and GenAI's full potential lies in making efficient use of high-quality, diverse training data. The large amount of data needed to train models effectively, as well as acquiring, storing, and processing this data can be a challenge. During an AI pipeline of operations, including ingestion of data, transformation/cleaning/pre-processing, model development, training, and then recursive validation/backtesting, IO patterns are widely varied.

According to reports from Google, Microsoft and organizations around the world, 70% of model training time is taken up by data staging operations. Put another way, your GPUs spend up to 70% of their time sitting idle, starved of data to train the model. It's no surprise when you start to look at the typical generative AI data pipeline.

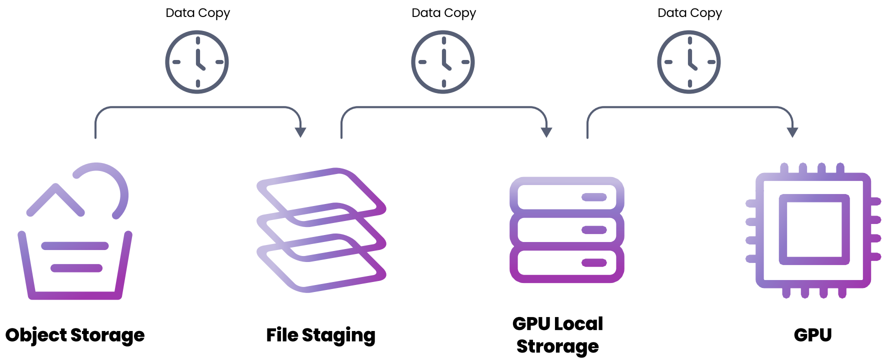

FIG. 1 Generative AI Pipeline Workflow

As shown in the diagram above, at the beginning of each training epoch, training data kept on high-capacity object storage is typically moved to a file staging tier and then moved again to GPU local storage which is used as scratch space for GPU calculations. Each "hop" introduces data copying time latency and management intervention, slowing each training epoch considerably. Valuable GPU processing resources are kept idle waiting for data, and vital training time is needlessly extended.

Complicated by Lots of Small Files

Training large language models that power most generative AI applications involves a significant amount of small files. Everything from millions of small images to IoT per-device logs for analysis, and more. Once pulled into the data pipeline, ETL type of work normalizes the data, and then Stochastic Gradient Descent is used to train the model. This presents massive metadata and a random read problem dominated by many small IO requests in the first part of an AI deep learning pipeline that many storage platforms are incapable of handling efficiently.

KEY INSIGHT

A 20X Reduction in Epoch Time By Switching to WEKA

One example of an organization using WEKA to accelerate its AI data pipeline is Atomwise. Atomwise is a pharmaceutical research company that uses artificial intelligence for structure-based drug discovery. They use 3D structural analysis to train their model, which is then used to identify a pipeline of small-molecule drug candidates that advance into preclinical trials. Model training typically relies on millions of structures, tens of millions of individual small files, and 30 to 50 epochs requiring as many as 12 data scientists managing an AI pipeline that could take as long as 4 days to complete. A deep dive into the data pipeline showed a major bottle next in I/O that, once cleared could result in a dramatic improvement in the data pipeline. That's when Atomwise adopted the WEKA Data Platform, running in AWS for the model training and drug discovery workflows. With the new solution, Atomwise was able to shift from a traditional multi-copy data pipeline, where it would take them 80 hours to do each training cycle, to WEKA's zero-copy data pipeline. They reduced their epoch time to 4 hours - 20X improvement in model training times. This allowed them to do in 12 days what would take a year on their old infrastructure, drastically speeding their final product to market.

"We wanted to train a model on the 30 million files we had, but the models are fairly large, with 30-50 epochs, a timeline of up to four days, and a lot of random-access-file lookups. GPUs are quite fast and hungry for data—you want to feed them as much data as you can," says Jon Sorenson, VP of Technology Development at Atomsie. With WEKA, "We could now consider experiments that earlier—because of all these headaches—might take us three months to figure out how to run. Now we can do this same experiment in less than a week."

Current file systems that originally focused on the HPC space, such as Lustre, GlusterFS, GPFS, and HDFS were architected for a problem of a different era and not for low-latency access to small files. Originally designed for slower hard disk storage of the past, they were architected to deliver high aggregate bandwidth for large files, including how they dealt with metadata management, data layout, striping design, and cache management. Because of this, the performance and storage efficiency of these file systems on-premises or in the cloud will be significantly reduced for massive small file applications.

WEKA Has a Better Way: The Data Platform for AI

Born in the cloud, the WEKA Data Platform is a software solution that ensures you can constantly saturate your GPUs doing the model training by providing the highest throughput at the lowest latency. The more data a deep learning model can consume and learn from, the faster it can converge on a solution, and the greater its accuracy will be.

WEKA collapses the typical GPU-starving “multi-hop” AI data pipeline using a single namespace where your entire data set is stored. This zero-copy architecture eliminates the multiple steps needed to stage data before training. Your GPUs gain fast access to data needed for training, while WEKA automatically manages tiering of data between high-performance, NVMe-based storage, and low-cost object storage. Incorporating the WEKA Data Platform for AI into deep learning data pipelines saturates data transfer rates to NVIDIA GPU systems. It eliminates wasteful data copying and transfer times between storage silos to geometrically increase the number of training data sets that can be analyzed per day.

The WEKA Data Platform efficiently handles large numbers of files creating virtual metadata servers that scale on the fly with every server that is added to the cluster. WEKA’s patented data layout algorithms distribute and parallelize all metadata and data across the cluster in small 4k chunks, this creates incredibly low latency and high performance whether the IO size is small, large, or a mixture of both. Because the WEKA Data Platform is software-defined, the same technology is used on-premises or in the cloud, providing you the same benefits and user experience.

For More Information or to Arrange a Free Trial

Visit us at https://www.weka.io/get-started or email us at info@weka.io.