The Secret to Speeding Up Inferencing in Large Language Models

Large Language Models (LLMs) are the foundation of many AI applications that we use in the world today. What sets these complex, sophisticated AI models apart is their ability to understand and generate human language. Using deep learning and neural networks, these models process and generate language-based tasks. Everything from text generation to translation, summarization, and answering questions is powered by LLMs.

In order to deploy an LLM, it must be trained on enormous text data sets. After the model is trained, it enters the inferencing phase. Model inferencing is the process of using a trained, deep learning model to make predictions or generate outputs based on new input data. Essentially, inferencing allows the model to apply its knowledge to practical, real-world scenarios. After a model has been trained on a dataset, it can apply the patterns and relationships it has learned to new, unseen data to produce results. This is a critical step in deploying AI systems, enabling them to perform tasks such as image recognition, language translation, and recommendation systems in real-time applications.

A completed model's file size is typically in the tens or hundreds of gigabytes (GBs). Each model is trained to meet user requirements or a specific operation, such as embedding data or understanding and answering textual context. This process is typically done on servers or cloud instances with GPUs (although other accelerator alternatives exist in the market, such as IPUs, TPUs, WSE, and even CPUs). Examples of well-known models that inference at scale are ChatGPT by OpenAI, Command-R by Cohere, Megatron by NVIDIA, Llama by Meta. Many other models are self-trained or open-sourced from locations such as Hugging Face or other model repositories to publish, compare, and share open-source models.

WEKA has a rich history of driving performance and scalability in the training phase of AI data pipelines, but as more focus is moving toward the inferencing phase, WEKA is meeting the key challenges in this phase of the AI data pipeline as well.

Inferencing Challenges

As inferencing is typically associated with running a model in GPU memory, a common misconception exists that storage does not play a part in the inferencing model life cycle. The result is that these expensive GPU instances are underutilized. They must allow for unpredictable bursts from API-prompt activity while delivering low-latency reply times to end users. The business impact is that organizations are often paying for expensive, underutilized GPU instances. The faster new inference instances can be created and the relevant models loaded into their GPU memory, the faster the inference farm can handle the additional inferencing load, thereby increasing GPU utilization and saving money on GPU instance time.

In this short example, we explore some of the storage-related inferencing pipeline challenges that a WEKA customer experienced and how the WEKA Data Platform allowed them to accelerate their inference pipeline and provide better service to their customers and internal users, reduce costs, and simplify their overall environment.

Enhancing GPU Utilization and Efficiency with the WEKA Data Platform

At WEKA, we work with a customer that is a well-known LLM provider operating large-scale LLMs in a cloud environment, herein referred to as “LinguaModel Labs.”

LinguaModel Labs came to WEKA because they were having challenges with inferencing efficiency and performance. They have an inferencing farm consisting of multiple GPU instances that load the models into GPU memory and execute them to utilize all these models. This inferencing farm is built to handle the operations of loading new models into the GPU memory and scale dynamically to allocate new, unpredictable inference instances that come from API requests.

The ability to quickly create new inference instances and load relevant models into GPU memory is crucial for managing inferencing loads. This not only increases GPU utilization but also saves money on GPU instance time.

Let’s explore the challenges faced by LinguaModel Labs and how the WEKA Data Platform provided effective solutions.

Challenges Faced by LinguaModel Labs

LinguaModel faced several challenges in optimizing their inferencing operations:

Model Loading

Loading the relevant models into GPU memory swiftly and handling switching between loaded models as needed.

Scaling GPU Instances

Quickly spinning up additional GPU instances as the load increases.

Maximizing GPU Utilization

Fully utilizing existing GPU instances to maximize their value and efficiency.

The WEKA Data Platform Optimizes Inferencing

By implementing the WEKA Data Platform, LinguaModel Labs achieved significant improvements in their inferencing capabilities.

Faster Model Load Times

Switching from an S3 repository to a high-performance file system mount point significantly boosted model load times.

½

Less than ½ the time
to dynamically spin up GPU inferencing instances shortening the model load time to the GPU memory.

35%

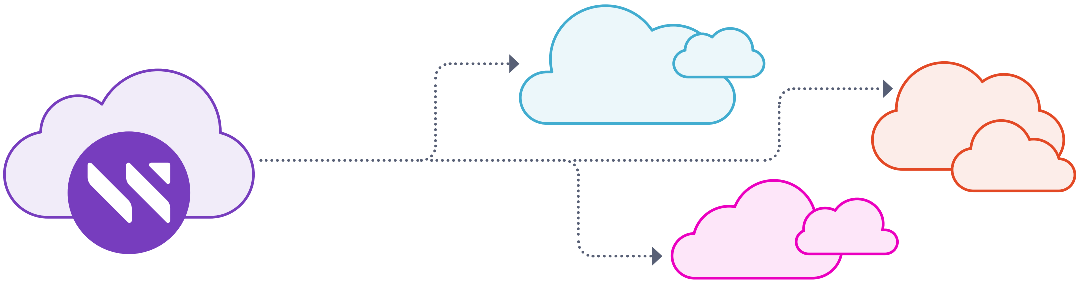
35% reduction of load times
from 265 seconds to 195 seconds for a 13GB model file size.

100GB+

100GB+ model sizes
loaded in a similar timeframe as much smaller models previously.

Enhanced Cloud Environment Interoperability

WEKA's snapshot and replication capabilities enabled LinguaModel Labs to share models between different cloud environments seamlessly.



Train in one cloud environment with WEKA

Distribute and use the data In different cloud environments

Future-Proofing with GPU Direct Storage (GDS)

Using GDS further shortened model load times and took advantage of future GPU memory increases.

80GB/s

80 GB/s loading data
into GPU memory in the cloud

1 SEC

1 second
to saturate GPU memory

190GB/s

190GB/s load times
with GDS

Additional Benefits of WEKA in Inferencing Environments

Beyond performance enhancements, the WEKA Data Platform offers several additional advantages.

Efficient Downloading of Inferencing Artifacts

WEKA efficiently downloads the LLM inferencing artifacts (texts, audio, video) to clear GPU and CPU memory, maximizing GPU utilization and value.

Rapid Load and Unload of GPU Memory

The ability to load and unload GPU memory in one second or less allows inference GPUs to save their current session, status, and tokens to stable storage. This frees the GPU for other inference activities while the previous session can be loaded to another GPU to continue from the same point when needed.

Increased Embedding Frequency

WEKA enables more frequent embedding by utilizing retrieval augmentation to keep the model updated with the latest data. This reduces model "hallucinations" and allows models to return the most up-to-date answers. Organizations can now run embedded jobs much more frequently, benefiting from a more accurate "source of truth" for their queries.

The WEKA Data Platform has transformed LinguaModel Labs's inferencing operations, making them faster, more efficient, and future-proof. This solution not only addresses the immediate challenges but also sets the stage for continued innovation and growth in AI and machine learning applications.

☑ **Learn more about WEKA for GPU Acceleration**