WEKA

Top 5

Misconceptions About GPUs for Generative AI

Generative AI has captured our imagination in ways not seen in decades. There seems no limit to the ways generative AI technology is being used, from predictive maintenance, patient diagnosis, enhanced customer support, and much more. Much of the innovation behind Generative AI is enabled by the accelerated compute capabilities offered in Graphical Processing Units (GPUs). GPUs parallelize matrix operations, enabling the large language models behind Generative AI to process vast amounts of data simultaneously, which significantly speeds up training times. This increased computational power allows researchers to train complex language models with millions or billions of parameters efficiently.

However, as transformative as GPUs have been for Generative AI and data science as a whole, it is human nature to simplify and assume that they are all we need to make AI projects successful and they are without caveats. This blind faith has led to surprise issues that delay rolling out data science projects or worse, cause them to fail. Here are 5 misconceptions about GPUs for AI that you need to avoid making when building out your AI projects.

1 My GPU is Giving Me the Fastest Results Possible

GPUs enable massive parallelism where each core is focused on making efficient calculations to substantially reduce infrastructure costs and provide superior performance for end-to-end data science workflows. 12 current NVIDIA GPUs can deliver the deep-learning performance of 2,000 modern CPUs. And adding 8 more GPUs to this same server can provide as many as 55,000 additional cores. While GPUs accelerate your compute processes, studies have shown they can spend half their time waiting for data, which means you end up waiting for results. The increase in computational power they provide requires a more powerful network and storage.

Up to 70% of an epoch happens before the GPU. Lots of time is spent copying data between systems for various stages of the data pipeline – NAS for persistent storage, local file systems or parallel file systems for fast storage, and object storage for archival data. This makes it challenging to keep your GPU fully utilized for lower epoch times and faster time to insights.

WEKA's Data Platform for AI addresses the storage challenges posed by today's enterprise technical computing workloads and other high-performance applications running on-premises, in the cloud or bursting between platforms. The Zero Copy Architecture runs the entire pipeline on the same storage backend and eliminates the cost and stalls of copies. With WEKA, you accelerate every step of your GPU-powered data pipeline – from data ingestion to cleansing to modeling, training validation, and inference – for accelerated business outcomes.

Throughput is King, All Hail Throughput 2

While it's easy to think about throughput as "the" metric you need to focus on to optimize your GPU usage, the throughput does not accurately reflect the full nature of AI workloads. To optimize your data pipeline you need to worry about more than feeding massive amounts of data to your GPUs – IOPs and metadata are important as well.

Each step of a data pipeline usually has a completely different profile for what the data looks like. And when you have different IO demands to work on the data throughout the pipeline, this can cause issues with traditional storage that is tuned to address one data type or throughput performance profile, creating silos of storage and a management problem. Depending on the workload you need performance profiles that deliver IOPS, latency, and or metadata operations in addition to throughput. Some steps need low latency and random small IO. Others need massive streaming throughput. Others need a concurrent mix of the two because of sub-steps within the process. In most environments, multiple pipelines will run concurrently, but at different stages, amplifying the need to handle different IO profiles simultaneously, WEKA provides a data platform that easily handles concurrent high bandwidth and high IOP conditions with ease.

WEKA delivers performance across all dimensions, so you can consolidate many storage platforms into one to eliminate wasted cycles copying data between platforms. Each stage will also run faster on WEKA than any other platform. WEKA exposes each application to the same data sets across all the available protocols and ultimately removes many of the "nerd knobs" for tuning, resulting in a simplified high-performance storage experience.

3 GPU-Powered AI Workloads Will Always Struggle with Small Files

Training the large language models that power most generative AI applications involve 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-types of work normalize the data and then Stochastic Gradient Descent is used to train the model. This presents a massive metadata and random read problem that is dominated by many small IO requests in the first part of an AI deep learning pipeline. that many storage platforms can't handle.

WEKA's architecture however provides a solution to this. By aligning all data requests along the native boundaries in NVMe devices, WEKA is able to handle not just small IO adeptly but also can provide large bandwidth as well by aggregating all of the small IO. On top of this, WEKA automatically scales out virtual metadata servers within a WEKA cluster to ensure that you can handle more and more metadata operations as the cluster scales up. The results are clear: One customer doing deep learning is averaging 4.2M IOPs and 250GB/s where much of the data starts as 10-kilobyte files.

Storage??? GPUs Are All About Compute Power 4

Artificial intelligence workloads have requirements for performance, availability, and flexibility that are not well met by increasingly traditional storage platforms. The storage selected for AI workloads will have a significant impact on the ability to meet business requirements. Successful AI projects tend to grow very rapidly in terms of both compute and storage needs and the implications of this growth on storage choices need to be carefully considered. Product selection.

However, most AI Infrastructure focus and spend is on GPUs and networks – these can consume up to 90% of a project's budget. This leaves whatever small percent of the budegt unspent left for storage to bring up the system. Performance at scale for AI storage is equally as important as "traditional" requirements such as availability, flexibility, and ease of use. It's often only after installation that organizations realize they are woefully short of storage to keep their ever-growing training data sets and are stuck with a far less useful environment.

The WEKA Data Platform offers linear scaling from terabyte to tens of exabytes and redefines scalability in the cloud era. A unified namespace scale allows customers to scale in every dimension possible with no performance impact as their AI workload needs grow. Scale across file and object automatically through intelligent tiering and scale the NVMe tier for increased performance and scale the object tier for increased capacity.

5 The Fastest Storage for a GPU is Local Storage

As AI datasets continue to grow, the time spent loading data begins to impact workload performance. Previously, the best way to keep GPUs fed with data was by fetching it from local NVMe storage. This avoids bottlenecks and latency caused by transferring data from a storage array and across a network. Networks and protocol stacks that run on them add overhead and can't keep the data transfers from the performance requirements of today's modern systems. But GPUs have gotten so fast that server hosts simply can't deliver data fast enough. GPUs are increasingly starved by slow IO.

WEKA can deliver IO faster than local storage for the fastest inferencing and highest images/ secs benchmarks. Local storage is limited by the resources of the local server – for example, the number of PCIE lanes and queues available to serve the IO. WEKA aggregates the resources of multiple servers to any GPU. And by enabling intelligent parallel access to all of the servers load balance requests to avoid any temporary IO stalls that might slow down access. New parallel data-plane and control-plane protocols avoid the issues with legacy network data access to ensure that the fastest way to serve data to a GPU is with the WEKA Data Platform.

WEKA