Technology Insight Paper

Architectural Considerations to Reduce AI Training Time and Accelerate Intelligence

By Frederic Van Haren

Sr. Analyst and HPC / AI Practice Lead

April 20, 2018

Evaluator Group

Enabling you to make the best technology decisions

Executive Summary

The applications for Big Data, Artificial Intelligence (AI) and cognitive computing driven by high-performance computing (HPC) are the workloads of the future and will transform the industry. The tools and platforms that are required for these new applications rely on large quantities of data to produce predictions, recommendations, and other intelligent feedback.

In a fast-moving world where data is the business driver, there is a need for new cloud-scale high-performing architectures that not only can ingest data quickly but also have the capability to process streaming data as fast as possible, in some cases this must be in real-time. The major challenge for existing data-center architectures is that they were never designed to handle the amount of data nor the speed of processing that is common today.

In this White Paper the architectural challenges for Big Data, AI, and HPC are outlined and how they can be solved with an innovative, software-only, high-performance file-based storage solution. Two common AI applications – speech recognition and image recognition – are reviewed to identify storage challenges and solutions associated with such AI problems.

Challenges of modern workloads

Modern applications driven by Internet of Things (IoT), Big Data, Artificial Intelligence, and cognitive computing are heavily dependent on the ability to ingest and process streaming data as fast as possible with the lowest latency. In most industries, applying AI is becoming a must to understand the underlying business, how to react quickly to an ever-changing environment, and to stay ahead of competition. HPC and AI workloads rely on cloud-scale distributed architectures with fast storage, a highly parallelized compute platform, a High-Performance File System (HPFS) that exposes POSIX compliant file system semantics, and a network (interconnects) that acts as a fast, low latency transport fabric between the storage and the compute platform.

The modern cloud-scale application architectures that are required to process that data are characterized by high volumes of data streaming in from numerous sources and a high dependency on complex real-time analytics based on AI to quickly process those streams. The distributed nature of these new applications requires that the data must be visible and equally accessible across the distributed architecture with high throughput, and it must be equally adept at handling a large amount of relatively small files with low latency to ensure the shortest time to results. As file count increases, metadata also increases, making it more difficult to maintain high performance. Current architectures have a strong focus on processing capabilities, but many storage and network interconnect technology innovations have not yet made it to mainstream designs yet. Compute solutions must deal with storage bottlenecks that cause processing units to become idle as they wait for data to be delivered, resulting in wasted compute resources.

Data has been growing at a faster rate than Moore's Law for processing power can accommodate, so there is a need for increases in processing power that keep pace with this data growth. In the Machine

Learning world, a popular new form of processing capability is the GPGPU (General Purpose Graphic Processing Unit) that is well suited for Artificial Intelligence workloads because the related Neural Network algorithms require a highly parallelized compute platform. GPGPUs deliver much better overall performance than their CPU counterparts for AI workloads and can handle highly parallel workloads at a level of scalability that can't be achieved with CPUs. GPGPUs are essentially turning a compute bound problem into a storage bound problem because they have dramatically increased the processing power of each server while the amount of data that needs processing is increasing at an even higher rate. Does that mean that CPUs are outdated and will be replaced? In short no, GPGPUs are there to assist the CPU not to replace them. The GPGPU is often referred to as a Compute Accelerator, and there are other products that also aim at the AI market. Examples are Google's Tensor Processing Unit (TPU) or Intel Xeon Phi.

Current storage solutions struggle to deliver data fast enough to CPU based infrastructure, and the same solutions will only perform worse with GPGPU based infrastructure where there are thousands of cores. GPGPUs represent a large investment and having them idle results in inefficient use of resources (low ROI), both capital and human. It requires a new storage solution that can deliver the needed throughput with low latency at cloud-scale to solve this problem.

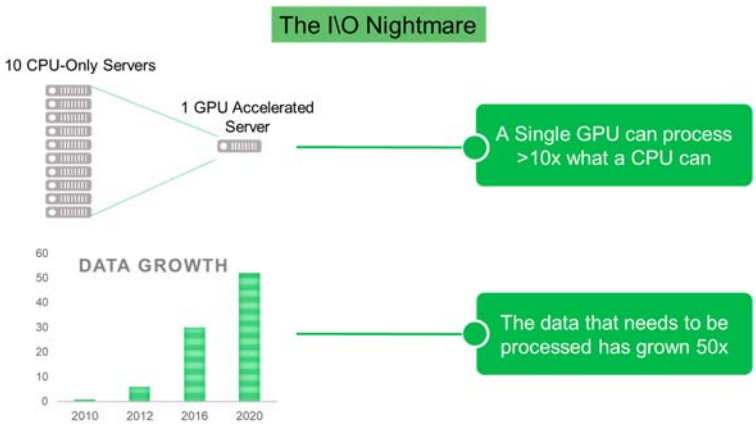

Figure 1

Despite innovative advancements in storage media such as Flash, NVMe (Non-Volatile Memory Express) and Intel 3D-XPoint, most scale-out storage architectures still rely on designs built around old hard disk drive concepts and utilize a Network File System (NFS) transport layer that has an upward limit around 1.5GByte/second bandwidth. A common storage design channels data flow through dedicated controllers that manage a back-end pool of drives, and these controllers eventually become bottlenecks at scale.

New solid-state drives (SSDs) based on NVMe perform significantly better and have much lower latency (~100 µseconds) than previous generations of SSDs, but NVMe drives need new cloud-scale storage architectures to take full advantage of their performance capabilities for maximum efficiency. NVMe over Fabrics (NVMeoF) is a specification that extends the benefits of NVMe to larger fabrics beyond the reach and scalability of a single server. It enables NVMe message-based commands to transfer data between a host computer and a target NVMe SSD over a high-performance Ethernet, Fibre Channel (FC),

or InfiniBand (IB) network[1]. An end-to-end NVMe solution reduces access latency and improves performance, particularly when paired with a low-latency, high-efficiency transport such as Remote Direct Memory Access (RDMA).

It is all well and good to have fast storage with low latency and a scalable compute platform but if the network (interconnects) adds a lot of overhead, that is problematic for HPC and AI workloads. Luckily there are a few options regarding low latency networks that solve the problem. They feature very high throughput, low latency and require very few CPU cycles from the host. First, we have IB High Data Rate (HDR) that delivers data packets with a latency of around 0.5 µseconds. Another option is RDMA over Converged Ethernet (RoCE) delivering packets with latency of around 1.3 µseconds. These modern networks are 100x faster than an NVMe SSD (~100 µseconds) therefore moving data over a low latency network (~1 µseconds) isn't a big deal and data locality is no longer a concern.

Traditional High-Performance File Systems (HPFS) were designed in the 1990's to handle high bandwidth, large sequential writes to spinning hard disk drives. While they are delivered as a generic software component they tend to require a complex setup and are time-consuming tuning to integrate with the actual underlying physical storage devices. The additional latency for metadata services also means that the overhead generated for each file access impacts the performance and scalability and is amplified with large quantities of small files and metadata intensive applications.

Practical Application - Speech Recognition

Speech Recognition technology has matured in the last decades to the point where most people are comfortable using it. Think of IoT devices such as Amazon Alexa in your home that can play music, provide weather reports and answer basic questions or Apple Siri on your iPhone. Market acceptance for innovative technology only exists if the technology is reliable enough and delivers a useful service. In the case of Speech Recognition, that means the need for a very high degree of accuracy in understanding someone's speech. This can only be achieved by collecting spoken data from people covering many different languages, dialects, and pronunciations. Improved accuracy is obtained by continuously adding and processing new incoming data from all kinds of IoT devices. The stored data consists of small files that never get deleted. This means that the underlying storage system must scale in terms of capacity, performance, and reliability. Even though only a fraction of the data can be processed on any given day, it should be maintained within a high-performance storage environment. A flexible tiering strategy is also required to move data between hot and cold storage for efficiency and cost. Sound familiar? It really is a general problem description that can be applied to many of the modern HPC / AI workloads.

[1] 100Gbit Ethernet, 100Gbit EDR InfiniBand and 128Gbit Fibre Channel

The algorithms used to process data have been adapted over time to handle the increase in data. Traditional HPC for Speech Recognition used CPUs to do the heavy lifting while the new AI and neural network algorithms require a much higher level of parallelization and are well-suited for GPGPUs. Why is that important? A server with eight (8) GPGPUs delivers more than 40,000 cores and needs significant storage bandwidth to keep those cores saturated. It is common to see configurations composed of sixteen (16) of those servers clustered together. Keeping the data as local as possible is almost impossible to achieve due to the dynamic nature of the data and doing so is not scalable. Also consider that each generation of GPGPU performs twice as fast as the previous generation, making performance scalability a growing problem.

Current storage solutions and HPFSs are not optimized for millions of small files and large volumes of data and can't deliver data fast enough to the compute platform, resulting in a condition commonly referred to as I/O starvation. Very often the data becomes isolated in data silos created by storage systems with different performance characteristics. It is almost impossible to manage such an environment. So, if there is one thing we know for sure it is that AI use cases, and their data sets, will keep on growing and the problem will only get worse.

WekaIO Matrix™

What is it?

WekaIO's Matrix is an innovative, high-performance distributed parallel file system that is elastic, highly-scalable and designed from the ground up to be a high-performance solution for HPC and AI workloads. By combining the performance of all-flash arrays with the scalability and economics of the cloud, Matrix prevents I/O starvation and keeps GPGPUs fully utilized. See Figure 2.

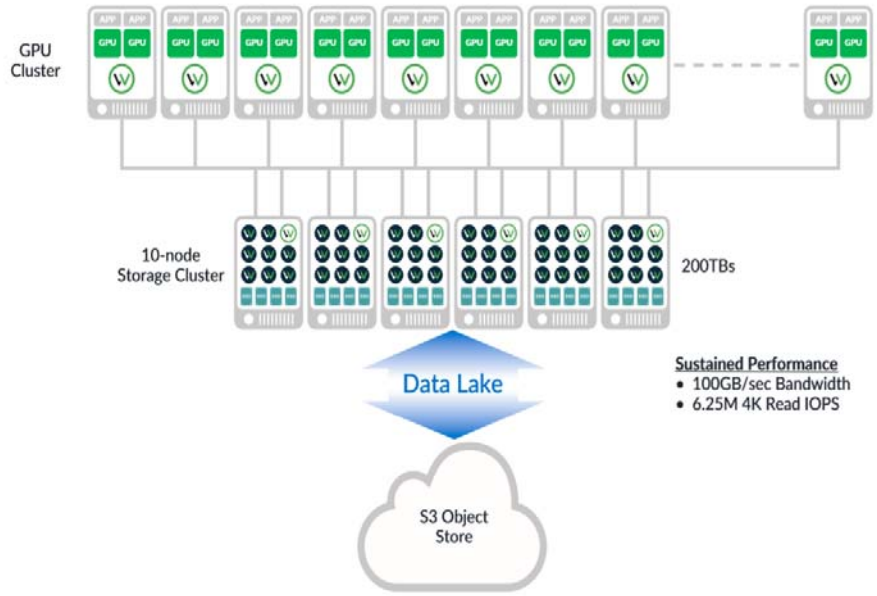

Figure 2

The software does not rely on specialized hardware and can be provisioned on standard servers as a dedicated storage layer, or hyperconverged on the GPU cluster. It presents a fully POSIX compliant file system presents and also supports other protocols including NFS v3 (Network File System), SMB 3.1 (Server Message Block), and HDFS (Hadoop Distributed File System). Additionally, Matrix file system can be deployed on-premises or public clouds, can interact with any AWS S3 compatible object store or NAS filer for low cost tiered storage.

Architecture

Key architectural features important for HPC/AI workloads include:

- A clean-sheet design with native NVMe support for higher performance, and lowest latency. With its own network stack, Matrix essentially delivers a parallel file system over NVMeoF.
- High system throughput, small-file IOPS, and metadata performance is achieved by a distributed design that automatically balances I/O activity to prevent a single node from becoming a performance bottleneck. HPC/AI workloads are notorious for having many small files and this design keeps throughput high and latency low to saturate the GPGPUs in a cluster for greater server utilization.
- Performance scales by adding more nodes (up to 64K) for increased I/O parallelism. Parallelism allows simultaneously reads and writes of multiple small chunks of data, effectively reducing the time it takes to read the entire file. See Figure 3.

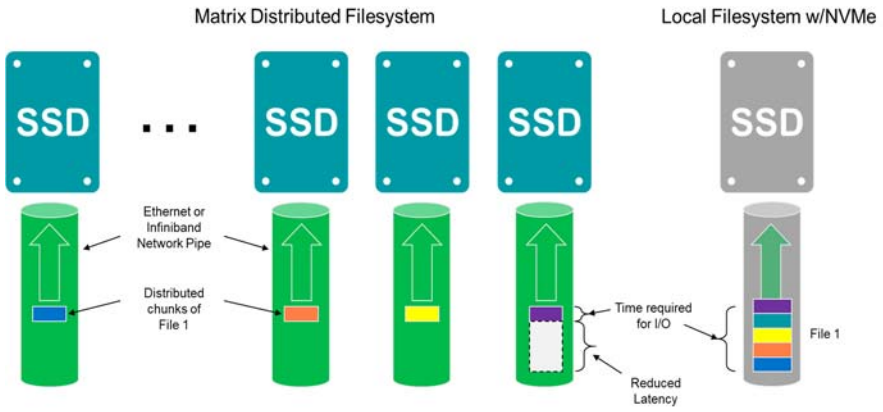

Figure 3

- Policy-based automatic tiering is supported between a high-performance storage tier and cold storage tiers, both of which are seamlessly managed within a global namespace. The global namespace means shared access to very large training data sets, even for clusters tapping into public cloud resources.
- Clones and snapshots can be moved automatically to the Cloud for cloud bursting, remote backup, and disaster recovery. Cloud bursting is especially useful to support periods of peak needs or to create an HPC/AI dev/test configuration in the Public Cloud. It also means that on-premises GPGPU clusters can be smaller, using cloud-based clusters on-demand.
- Low latency interconnects (Ethernet or InfiniBand) means that Matrix can provide faster access to data from a shared resource than direct attached storage. This allows large HPC/AI data sets to scale to Exabyte size without the need to continuously rebalance data placement for performance.
- Distributed data protection comes in the form of N+2 and N+4 protection schemes and prioritized rebuilds means that in the event of multiple failures, data common to the failures is rebuilt first to ensure your data is returned to the next higher level of protection as quickly as possible.
- Matrix supports bare-metal deployments for maximum performance, virtual environments for flexibility, containers for application portability, or in a cloud (private, public or hybrid) for on-

demand scalability. HPC/AI workloads vary greatly, which means that the supporting infrastructure must be both adaptable and easily re-configurable.

Autonomous Vehicle (AV) Weka.IO use case

An AV manufacturer was building a production training system that needed to ingest massive amounts of data streaming from an array of IoT sensors. The initial solution, an all-flash NAS system, worked well for the machine learning algorithm development that utilized a limited set of synthetic data. However, the size of the new GPGPU cluster and larger data volumes required performance to scale 10x that of the pilot system to keep the GPGPUs fully utilized. System designers benchmarked several storage solutions, comparing the latest iteration of a blade-based all-flash-array (NAS) from an established vendor, the GPGPU local file system with direct attached SSDs, and a small cluster running WekaIO's Matrix™ running on commodity servers. As can be seen in Figure 4, the WekaIO solution was able to hit the extremely high single client bandwidth requirements to fully utilize the processing power of the GPGPUs. The all-flash NAS suffered from the limitations of NFS based access while the local file system could only deliver the performance of a few locally attached NVMe drives. As a distributed parallel file system, Matrix allows sharing of storage resources, meaning that larger data sets can be analyzed than would fit on a single GPU server and with lower latency and improved data resiliency.

Figure 4

Customer results for WekaIO:

- 2x faster than local disk
- 7.1x faster than the blade-based Flash-Array
- GPGPUs are fully saturated!

Conclusion

Traditional storage architectures are being challenged by new workloads driven by HPC, Big Data, and AI. As we continue to analyze bigger data sets and shrink processing into ever more powerful GPGPUs, storage bottlenecks and particularly fast data access will become the biggest limiter to data insights. An opportunity for innovative solutions to address the old and future problems is clearly here.

WekaIO Matrix™ is a pioneering solution that promises to deliver a fast, efficient, and resilient distributed parallel file system. It addresses the performance and the scale needs for demanding storage workloads in verticals such as Machine Learning, real-time analytics, genomics, media rendering and manufacturing. Its integrated architecture provides the performance of NVMe, the simplicity of file storage, the scalability of the cloud and an economic model that puts the buyer in control of hardware purchasing power.

About Evaluator Group

*Evaluator Group Inc. is dedicated to helping **IT professionals** and vendors create and implement strategies that make the most of the value of their storage and digital information. Evaluator Group services deliver **in-depth, unbiased analysis** on storage architectures, infrastructures and management for IT professionals. Since 1997, Evaluator Group has provided services for thousands of end users and vendor professionals through product and market evaluations, competitive analysis and **education**. **www.evaluatorgroup.com** Follow us on Twitter @evaluator_group*