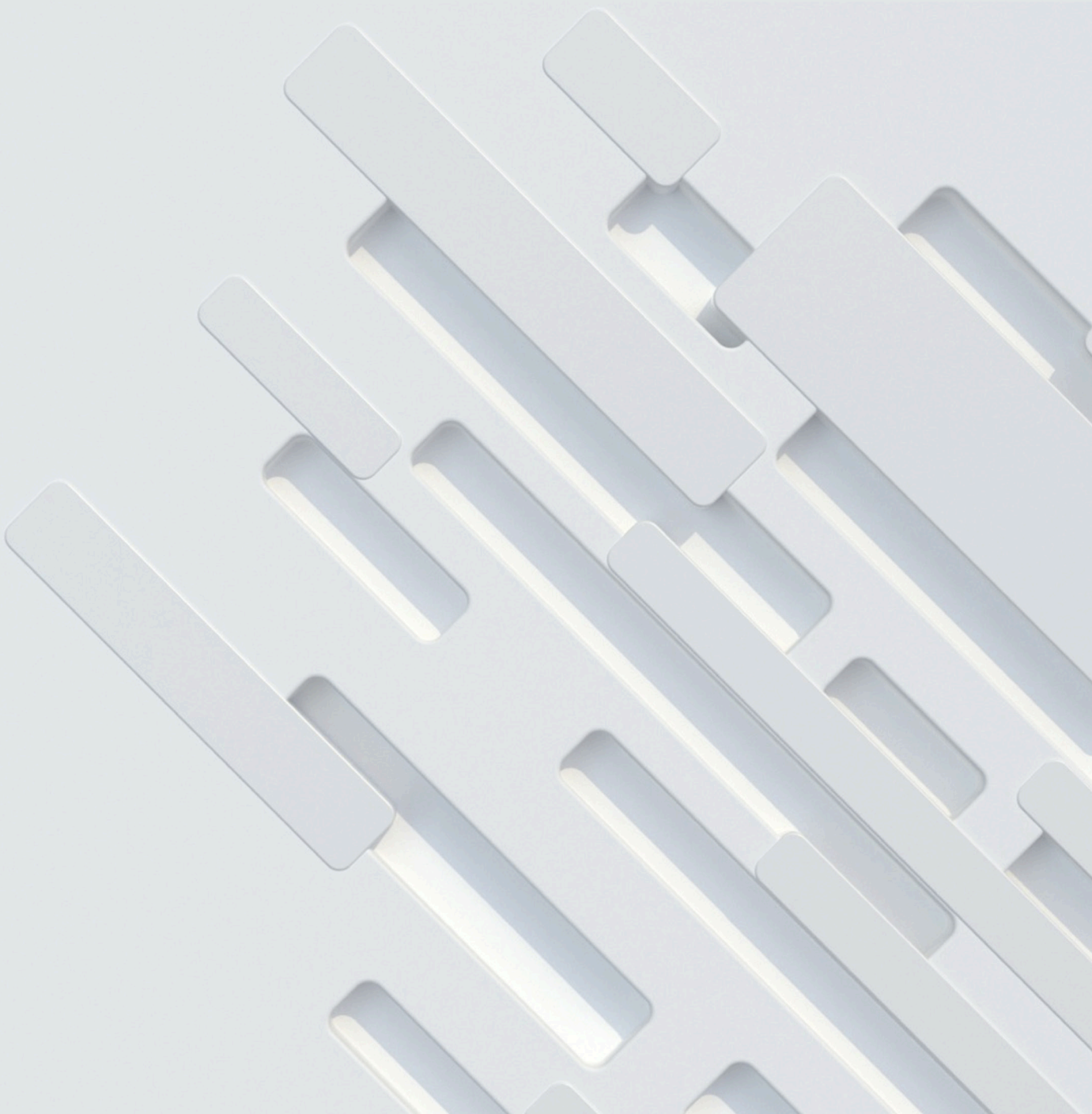
Modernizing Enterprise Storage for Machine Learning Analytics Workloads

TechTarget | Custom Media



CONTENT

INTRODUCTION

As analytics workloads have taken hold in IT organizations as potential game-changing applications, storage requirements have changed dramatically. New kinds of solutions are necessary to meet the performance, agility and economics challenges associated with machine learning analytics workloads, and selecting the right storage infrastructure is critical.

The explosion of data represents both a blessing and a curse for organizations. On the plus side, it offers up a world of opportunity to parse the data for hidden treasures—data insights that enable businesses and industries to solve their greatest challenges like manufacturing yields, cancer diagnosis and treatments, financial trading trends, and pharma drug discovery. But with that dramatic data growth comes a big challenge: ensuring that data-rich workloads can be properly supported and accelerated with next-generation IT infrastructure—especially storage.

In particular, the opportunities presented by artificial intelligence for powerful data insights are causing IT professionals to think long and hard about choosing the correct storage infrastructure. High-performance machine learning (ML) workloads require substantially greater bandwidth, performance and storage capacity in order to process enormous amounts of data in real time so it can be transformed into actionable insights and accelerate time to value.

Choosing the wrong storage infrastructure would be costly, both in financial terms and in terms of not meeting critical business objectives. Imagine how C-suite executives and other business stakeholders will react when they learn that the increased Capex funding they allocated to new compute and storage infrastructure for those workloads will not meet the needs of the organization.

For years, network-attached storage (NAS) was an attractive and effective way to deal with storage demands for enterprise workloads. NAS has played a valuable role in helping organizations share and access data across multiple users and diverse client devices from a central disk repository. The proliferation of local-area networks made NAS appliances a cost-effective option for sharing storage across standard Ethernet connectivity, resulting in easy access for relatively low capital expense.

However, in the modern, high-performance computing era, enterprise application performance requirements have evolved dramatically. In particular, the pivot toward new storage architectures has been accelerated by the dramatic impact of artificial intelligence (AI), inclusive of deep learning and machine learning, workloads that demand not only much higher storage capacity, but also much higher performance to the compute nodes housing the applications. This class of performance can only be delivered with a highly parallel and scalable file system.

Adding to these new workload demands are such issues as shortening backup windows, more demanding recovery time/recovery point objectives (RTOs/RPOs), the need to share data across applications and workloads, the proliferation of "cold" data and the demand for instantaneous access for analytics and other business-critical decision-making.

This paper examines a modernized approach for enterprise storage that delivers enhanced data center agility, accelerates data transformation and optimizes data center investments.

WHAT'S CAUSING THE PROBLEM

Today's modern, analytics-driven workloads offer exciting business potential—but they represent an operational hurdle for IT organizations. These workloads are marked by massive data sets that require far more than higher-capacity storage or cost-efficient scaling; today's workloads are sharing critical data and need the kind of performance delivered by parallel scale-out storage systems. Adding to the problem, massively parallel GPU-based servers have concentrated the I/O demands into one single machine, placing even more demands on the storage infrastructure. Figure 1 contrasts CPU-based architectures with new GPU-based infrastructure.

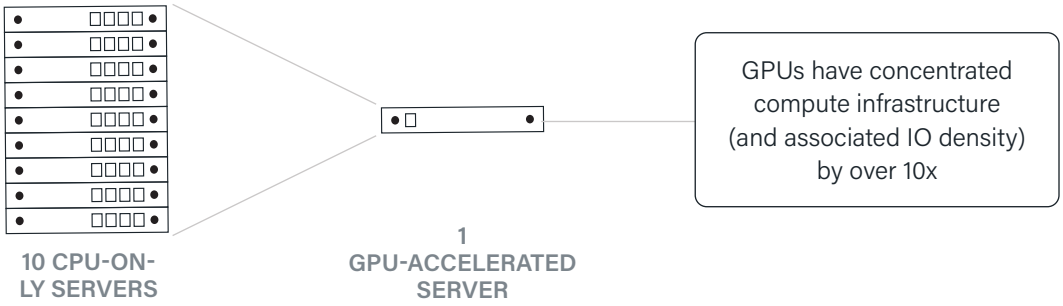


Figure 1: GPU accelerated workloads have increased I/O demands by over 10x per server node

A new approach is necessary because of the data I/O bottleneck that has resulted in a "performance tax" that is inhibiting organizations' ability to surface data insights in a timely manner without having to invest more Capex and deal with greater management complexity.

Because AI-driven workloads are a must for all enterprises, investing in the right storage infrastructure is a big decision. But that decision must be made with the critical knowledge that 80% of the time that data scientists devote to analyzing data is spent in transforming and preparing the data for use in analytics workloads.

Traditional storage architectures, while still doing yeoman's work for a variety of enterprise applications, were never envisioned for the kinds of algorithmic-based analytics workloads that no longer can be done by a group of data scientists manually tuning those storage systems. Instead, IT decision-makers need to look for a shared-file system that delivers all the I/O required to avoid waiting for data availability.

THE 3 PILLARS OF STORAGE SOLUTIONS FOR ML-DRIVEN ANALYTICS WORKLOADS

There are three "pillars" that highlight the plight of storage administrators looking to ensure their storage infrastructure can keep up with their ML-driven workloads:

Drive Data Center Agility

While the term "business critical" has been widely used to describe the increasingly valuable array of data being created, stored and shared in enterprise computing environments, the term has taken on real significance with the introduction of analytics workloads, powered by such technologies as artificial intelligence, machine learning and deep learning. The resultant explosion of data has created a big strain on storage systems, many of which lack the bandwidth and management functionality necessary to keep business-critical data available and flowing among key workloads.

Traditional data centers with legacy storage systems have been hamstrung by isolated storage silos that have sprung up over the years in response to mounting demands. In these data centers, data no longer can be ingested or shared efficiently, nor can it be adequately protected in the event of security threats, user error or configuration problems.

These data centers no longer have the requisite agility to manage all this stored data, such as knowing the location or ownership of the data, adhering to governance and compliance requirements and controlling access to the data according to clear, yet flexible, policies and rights management. All organizations need to juggle both on-premises and cloud-based environments to promote a more nimble approach to delivering IT services.

Clearly, a new approach is required—one that eliminates data silos and reduces complexity, supports multiple deployment models and performs as multi-protocol infrastructure for mixed workloads increasingly prevalent in the data center.

Accelerate Data Transformation for ML Analytics

Artificial intelligence—specifically, machine learning and deep learning—has transformed the way enterprises use data for sophisticated analytics. But whichever AI avenue you go down, enterprise storage has to evolve to keep up.

Specifically, new analytics workloads require:

- Massive amounts of data
- Faster, parallel and more efficient access to that data
- Algorithms for training and facilitating the learning process

Traditional NAS systems are unable to take advantage of higher-speed networking and lack the ability to handle the I/O demands of AI-driven data for analytics. Instead, these and other analytics workloads require a parallel file system with substantially higher performance than legacy storage systems, support for multiple types of data (structured, unstructured and semi-structured), and support for hyper-scale performance that streamlines the requisite data preparation phase in AI. Figure 2 outlines the architectural difference between NAS and Parallel file systems.

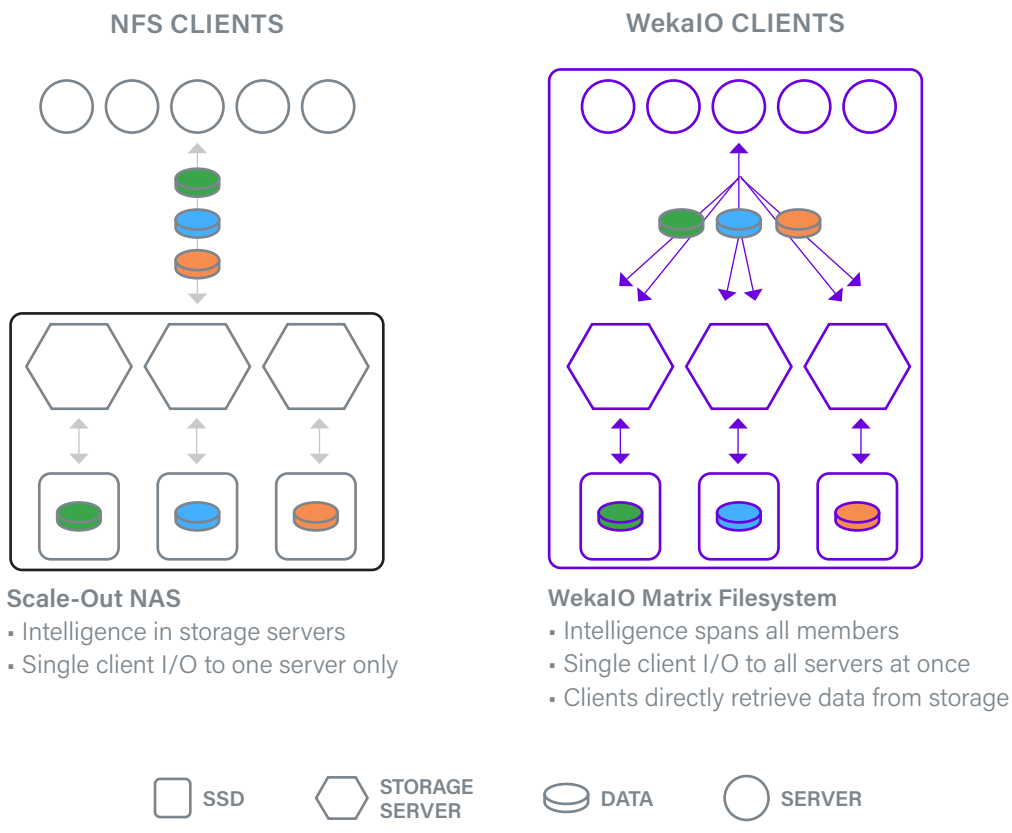


Figure 2: Architectural difference between NAS and parallel file systems

Accelerating data transformation for ML workloads not only means substantially improved performance, but also far better cost efficiency, simplified management, massive scalability and a more agile, sustainable approach to future data transformation. A new class of parallel file storage systems can speed local file system performance, reduce data copies and support cloud bursting for load-leveling under demand peaks.

Optimize Infrastructure Investment

Finally, since cost efficiency is a critical mantra for IT and storage administrators in a data-rich environment, organizations have put significantly greater emphasis on increasing "bang for the buck" when it comes to storage infrastructure investments. New storage systems must optimize CPU and GPU utilization without requiring organizations to make massive new investments in Capex, while also leveraging object storage to improve the economics of storing, protecting and backing up masses of data. Not to be overlooked is the need for a new storage management approach that reduces management costs through integrated tiering, remote backup and recovery, and encryption both at rest and in flight.

Making the wrong choice for optimized storage in ML-driven workloads can be a costly mistake, not only resulting in higher Capex and Opex costs, but also in delaying the time of critical data and insights getting to the right person in the right place at the right time. And poor storage management controls can result in costly misconfigurations that further limit data availability and result in performance bottlenecks.

Avoid the temptation to turn an array of legacy systems into a solution for next-generation, ML workloads. Build-it-yourself scale-out file systems built around open source software is a time-consuming, people-dependent exercise that can easily turn into a science project.

THE WekaIO SOLUTION

WekaIO has made its mark with a new approach to enterprise storage solutions that works equally well in on-premises environments, in cloud deployments or in hybrid scenarios. The core product—WekaIO Matrix™—is a software-based, scale-out storage solution optimized for ML-driven analytics and performance-intensive workloads.

WekaIO Matrix offers the simplicity of NAS, but it delivers all-flash-level performance, cloud scalability and simplified management never envisioned by NAS product vendors. In production environments, Matrix has shown to deliver 10 times the performance of a traditional NAS system with linear scaling as the infrastructure grows.

The MatrixFS file system is a distributed, scalable, high-performance software platform that connects multiple servers with locally attached solid-state drives into a POSIX-compliant global namespace for webscale performance and simplified management. The software is deployed on standard commercially available servers providing true hardware independence and best cost. The software supports internal tiering to any commercially available S3 object storage solution delivering massive scalability and great economics for an ever growing data catalog. Figure 3 provides an architectural overview of a typical deployment for deep learning environments.

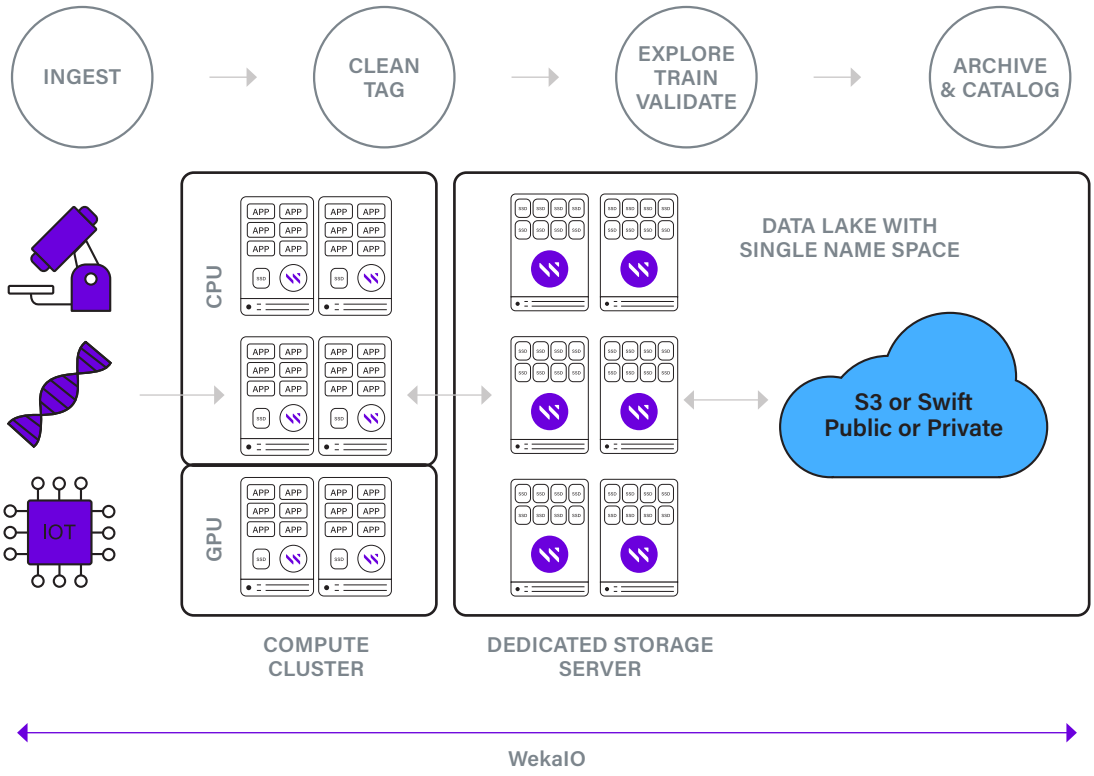


Figure 3: WekaIO Matrix deployed in a deep learning environments

WekaIO Matrix offers enterprise storage and IT decision-makers important deployment flexibility. For instance, Matrix can be implemented in converged infrastructure, as a dedicated storage server or as a public cloud storage resource. It is designed as an integrated, easily configured and quickly deployed storage resource that flexibly adapts to changes in enterprise IT requirements, including support for workload migration to and from multiple cloud environments.

"WekaIO has an excellent understanding of the needs for delivering information storage in environments where arriving at answers has become a business-critical operation," according to Randy Kerns, senior strategist and analyst at Evaluator Group. "Their solution provides an advanced approach to massive scaling with a parallel file system for those demanding applications."

CONCLUSION

Today's ML workloads demand a new class of storage that delivers the performance, manageability, scalability and cost efficiency demanded in the era of digital transformation.

WekaIO's parallel file storage system was designed and optimized for modern workloads, and is ideally engineered to take storage performance and data availability to the next level as performance demands further intensify. Instead of suffering from I/O starvation, IT and storage professionals can break through NAS' boundaries with WekaIO's shared file system that ensures data is available where and when it is needed, regardless of its format, file structure or location.

The WekaIO solution helps storage become an enabler of digital agility, rather than an inhibitor.

Contact the WekaIO team today to start your 30 day free trial of WekaIO Matrix™ on-premises or on AWS cloud.

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io

W04R1WP201904