

By leveraging flash data storage, companies can efficiently unlock valuable insights, uncover hidden patterns, and extract actionable intelligence from their data, enabling them to stay ahead of the curve and differentiate themselves in the market.

The Value of Data Storage for the AI Era

May 2024

Written by: Jeff Janukowicz, Research Vice President, Solid State Drives and Enabling Technologies

Introduction

The dawn of the artificial intelligence (AI) era has ushered in a transformative wave of innovation, catalyzing unprecedented changes and opportunities in the realm of data storage. With artificial intelligence applications evolving at an unprecedented pace, the demand for robust data storage solutions has never been more pressing. AI applications are becoming not only more sophisticated but also increasingly pervasive across various sectors, from core datacenters to the edge and from healthcare and finance to manufacturing and entertainment. These applications rely heavily on vast amounts of data, necessitating scalable and efficient storage solutions to support their computational demands. As AI algorithms become more complex and data intensive, the sheer volume of generated and processed data continues to escalate. This surge in data creation underscores the critical importance of storage systems capable of accommodating massive data sets while maintaining high levels of performance and accessibility to ensure that computational (i.e., GPU) investments are fully maximized.

Data storage serves as the foundation of the AI revolution, enabling organizations to harness the full potential of artificial intelligence to drive innovation, enhance decision-making, and create transformative value across diverse industries. As the AI landscape continues to evolve, the role of data storage will remain indispensable in shaping the future of technology and enterprise infrastructure. The AI era presents both extraordinary opportunities and notable challenges for data storage. To meet the demands of this transformative era, flash data storage emerges as an invaluable asset. The combination of speed, scalability, and efficiency makes flash data storage ideally suited to address the dynamic requirements of increasingly data-hungry AI applications. As organizations navigate the complexities of the AI landscape, leveraging flash storage solutions will be pivotal in unlocking the full potential of artificial intelligence.

The Top 4 AI Storage Imperatives

AI infrastructure is complex. While the compute infrastructure of CPUs and GPUs is a vital component of AI solutions and has been the focus infrastructure investments up until now, data storage is also a foundational pillar of the hardware infrastructure for the AI era. The widespread adoption of artificial intelligence and machine learning (ML) workloads means that architectures will evolve. This underscores newfound opportunities in data storage as use cases and massive data sets only continue to grow.

AT A GLANCE

WHAT'S IMPORTANT

Flash data storage is emerging as an invaluable asset. Its combination of speed, scalability, and efficiency makes it ideally suited to address the dynamic requirements of AI applications.

These requirements can be understood by examining the demands for data and the role of storage within AI infrastructure solutions:

- » **Volume and location of data:** The AI workflow or pipeline often requires massive amounts of data for ingestion, preparation, training, validation, and inference. With the advent of big data technologies, organizations are collecting and storing vast data sets from various sources, such as sensors, social media, and IoT devices. The ability to manage, store, and archive these large volumes of data efficiently is a key requirement in the AI era. In addition, the ability to store and act on more and more data closer to or at the source of data generation is vital to maintain a high level of efficiency.
- » **Speed of data:** In addition to the sheer volume of data, the speed at which data is generated and processed has increased tremendously. Real-time analytics and decision-making are becoming more critical in domains such as finance, healthcare, the cloud, and edge. Data storage systems need to support high-speed data ingestion, processing, and retrieval to enable real-time AI applications.
- » **High performance:** AI workloads often require high-performance data storage to meet the computational demands of training complex models and processing large data sets. Data storage systems should provide low-latency access to data and high throughput to support intensive AI workloads efficiently and maximize compute infrastructure investments.
- » **Data efficiency:** Efficiency plays an essential role in ensuring that AI solutions attain optimal performance, control costs, enhance energy conservation for accelerated training and inference, facilitate scalability, and bolster reliability. AI servers consume large amounts of energy, and AI solutions often struggle with delivering enough power to the datacenter. Thus leveraging more power- and space-efficient storage can deliver huge benefits to overall infrastructure efficiency, including freeing up space and power needed to accommodate more compute. Through the integration of data-efficient techniques, solutions can deliver more effective and impactful AI data storage systems, thereby generating value across a broad spectrum of AI applications.

While it may be convenient to think of all AI solutions as being the same, the reality is that data in the AI era comes in various formats, including structured, semi-structured, and unstructured data. This includes texts, images, videos, sensor data, and time series data. Data storage systems must be capable of handling diverse data types efficiently and support data models of various sizes to be compatible with AI algorithms and solutions.

For AI solutions, one of the most important considerations is the increased value of data given that large and high-quality data sets result in the most accurate and complete AI models. Extracting actionable insights and value from data is the primary goal of AI applications. Data storage systems should provide mechanisms for data analysis, visualization, and integration with AI algorithms to deliver meaningful insights and drive business decisions.

Overview of the Flash Data Storage Landscape

AI workloads often require high-performance storage systems and servers to meet the computational demands of training complex models and processing large data sets. Enterprise storage systems and servers should provide low-latency access to data and high throughput to support intensive AI workloads efficiently. Flash storage can be highly beneficial for AI workloads due to its unique characteristics that align well with the demands of AI applications.

What Is Flash Storage?

Flash storage is a type of storage device that uses solid state storage technology (typically NAND flash memory) to store data, unlike traditional mechanical hard disk drives (HDDs), which use spinning disks and moving read/write heads to access data. Solid state drives (SSDs) are the most common type of flash storage and have no moving parts, making them faster, more durable, and more energy efficient.

SSDs consist of NAND flash memory chips, a controller, and a cache. The controller manages data storage and retrieval operations, optimizing performance and ensuring data integrity. The NAND flash memory chips store data in a nonvolatile manner, meaning the data persists even when the power is turned off. There are several types of NAND flash memory used in SSDs including single-level cell (SLC), multilevel cell (MLC), triple-level cell (TLC), and quad-level cell (QLC), with each offering different levels of performance, endurance, and cost. In enterprise environments, SSDs provide high performance, reliability, and endurance to meet the demanding requirements of enterprise and datacenter applications and workloads.

Advantages of SSD Storage for AI

SSDs offer several advantages for AI applications, especially when compared with legacy HDDs, due to their characteristics and performance capabilities:

» Speed and performance:

- SSDs provide significantly faster data access times and high throughput. This high-speed access is crucial for AI applications, especially for tasks such as data preprocessing, model training, and inference, where large data sets need to be read and processed quickly.
- SSDs offer low latency, meaning they can quickly respond to read and write requests. SSDs also help mitigate the I/O blender effect when simultaneously random read and write requests to the storage system result in mixed and unpredictable I/O patterns, leading to decreased overall storage performance. These characteristics are essential for AI workloads that require real-time processing, such as natural language processing, recommendation engines, and real-time analytics.

» Scalability:

- SSDs are highly scalable in terms of both capacity and performance. They can be easily scaled up by adding more SSDs or deploying higher-capacity configurations (such as QLC) to meet the growing storage demands of AI applications. This scalability ensures that the data storage infrastructure can keep pace with the increasing volume of data generated by AI workloads.
- SSDs in AI solutions enable costly GPUs to scale and improve utilization by providing fast and efficient storage for data preprocessing, loading, and transfer, as well as reliable storage for model checkpointing and recovery. By leveraging the capabilities of SSDs, AI systems can maximize the efficiency of GPU resources, accelerate model training, and achieve better performance outcomes.

» Operational efficiency:

- The operational efficiency of SSDs in datacenters is driven by their speed, reliability, scalability, energy efficiency, and density improvements. At scale, SSDs have lower power consumption compared with over-provisioned HDDs, resulting in reduced cooling requirements within the datacenter and helping

address datacenter power bottlenecks due to their more energy-efficient profile. Newer SSD form factors also enable higher density as solutions move away from traditional HDD-based ones.

- AI solutions can be expensive due to the high cost of CPUs, GPUs, and memory. SSDs provide efficient data storage to support the vast amount of AI data to deliver a cost-effective AI solution. While SSDs typically have a higher initial cost (in terms of dollars per gigabyte) compared with HDDs, they can be more cost effective when factoring in performance, reliability, and the total cost of ownership, particularly in high-performance workloads such as AI.

Considering Solidigm

Data storage requirements are evolving rapidly with the explosion of the AI era, and it's important to find the right partner that can provide the flexibility and breadth for each specific AI application. Solidigm is a global leader in innovative NAND flash storage solutions with a comprehensive portfolio of SSD products based on SLC, TLC, and QLC technologies. Headquartered in Rancho Cordova, California, Solidigm operates as a standalone U.S. subsidiary of SK hynix with offices across 13 locations worldwide.

Solidigm has a long legacy of solid state innovation and leadership and is positioning itself with solutions for the AI era. Solidigm currently offers a storage portfolio that optimizes and accelerates AI infrastructure, scales efficiently, and improves operation efficiency from core datacenters to the edge. The products that power AI solutions include:

- » For high read/write and endurance requirements, the Solidigm D7-P5810 is a fast PCIe 4.0 SSD based on 144-layer SLC 3D NAND, delivering up to 1.6TB capacity with up to 65DWPD for sequential workloads and up to 50DWPD for random workloads with extreme write-intensity and high-endurance needs.
- » For many mainstream and read-intensive workloads, the Solidigm D5-P5430 provides a blend of hyperdense, affordable storage and high throughput. The Solidigm D5-P5430 is designed to deliver TLC-like read performance with massive lifetime writes, on QLC economics, leading to reduced total cost of ownership and improved sustainability of datacenter and edge infrastructures. Available in a broad range of form factors and capacities up to 30.72TB, the Solidigm D5-P5430 can be deployed in a wide array of 1U and 2U configurations.
- » For the lowest cost with read optimization, the Solidigm D5-P5336 is part of the company's fourth-generation QLC SSDs (using 192-layer QLC NAND flash) for the datacenter, delivering a blend of high capacities up to 61.44TB with read-optimized performance and high value to read and data-intensive workloads. The Solidigm D5-P5336 has been designed to efficiently accelerate and scale with increasingly massive data sets found in data-hungry workloads, such as data pipelines and data lakes for AI, ML, and big data and analytics.

Challenges

Storage requirements in the AI era are rapidly evolving, driven by the exponential growth of data in next-generation AI systems. The volume of data stored and data storage location and retention duration are critical factors that will continue to shape storage infrastructure as AI becomes increasingly ubiquitous worldwide. Storage solutions offered by vendors such as Solidigm must not only meet current demand but also remain adaptable and flexible as enterprise datacenter infrastructure expands to support AI at scale. This necessitates a focus on high-speed access, scalability, and efficiency to accommodate the dynamic nature of AI workloads. Advancements in storage technologies tailored for AI are expected to play a pivotal role in shaping the future of data storage in AI-driven environments.

Conclusion

The onset of the AI era has triggered a significant revolution in data storage, leading to groundbreaking innovations and new opportunities. As the landscape of AI evolves, the significance of data storage remains paramount in shaping the trajectory of technology and enterprise infrastructure. With the growing value of data in the AI ecosystem, there arises a heightened need to prioritize data storage. By leveraging flash data storage, companies can efficiently unlock valuable insights, uncover hidden patterns, and derive actionable intelligence from their data, enabling them to stay ahead of the curve and differentiate themselves in the market.

With the growing value of data in the AI ecosystem, there arises a heightened need to prioritize data storage.

About the Analyst



Jeff Janukowicz, Research Vice President, Solid State Drives and Enabling Technologies

Jeff Janukowicz is a research vice president at IDC where he provides insight and analysis on the SSD market for the client PC, enterprise datacenter, and cloud market segments. In this role, Jeff provides expert opinion, in-depth market research, and strategic analysis on the dynamics, trends, and opportunities facing the industry. His research includes market forecasts, market share reports, and technology trends of clients, investors, suppliers, and manufacturers.

MESSAGE FROM THE SPONSOR

Solidigm is a leader in the data storage industry, bringing to market an unmatched technology portfolio. Solidigm was formed in 2021 from the sale of the Intel NAND and SSD business to SK hynix Inc. Headquartered in Rancho Cordova, California, Solidigm has over 1,700 employees in 13 locations across the globe, comprised of pioneering inventors, leaders, and problem solvers in the data storage industry. The company's product teams have been focused on delivering QLC NAND designed to offer the best value proposition for performance, reliability, and cost for the data workloads of today and into the future.



The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
 140 Kendrick Street
 Building B
 Needham, MA 02494, USA
 T 508.872.8200
 F 508.935.4015
 Twitter @IDC
idc-insights-community.com
www.idc.com

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2024 IDC. Reproduction without written permission is completely forbidden.