HOW AI IS RE-ARCHITECTING (→) THE DATA CENTER OF THE FUTURE

PRAVEEN MIDHA
GOPAL SHARMA

“WHITE PAPER”

SEPTEMBER.2025

HOW AI IS RE-ARCHITECTING (→) THE DATA CENTER OF THE FUTURE

Executive Summary

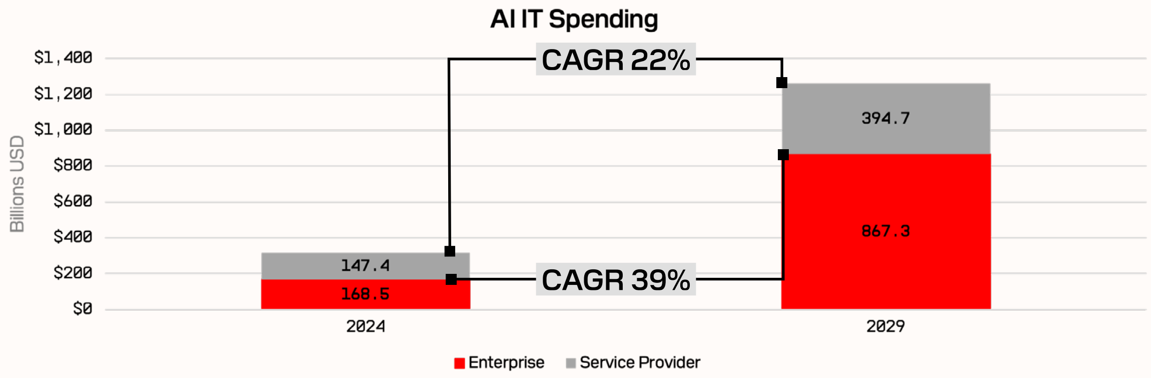
Artificial intelligence (AI) has proven to be a paradigm shift for workloads in the modern data center. It has reshaped industry perspectives on everything from data processing to power consumption. This white paper examines AI's transformative impact, including the rising temperature of data utilization, the growing energy demands of accelerated computing, and the limitations of current storage architectures. We explore how new and innovative technologies are bridging these gaps to meet the needs of tomorrow's AI-driven workloads.

Three Key Trends in the Data Center

1. AI: A Mission Critical Enterprise Workload

AI has long existed in the tech ecosystem in various forms, but when a game changing AI chat came to market in November 2022, it sparked a revolution in compute centered around Large Language Models (LLMs) and other Generative AI (GenAI). This new workload provides exciting new capabilities that in turn provided a surge of interest and investment.

AI IT spending is projected to grow at a compound annual growth rate (CAGR) of 31.92%, from $315.9 billion in 2024 to $1.262 trillion by 2029. By then, AI spending will constitute 16.4% of total IT expenditures.[1]



Service providers and enterprises are heavily investing in relatively nascent and distinct AI infrastructure. According to a recent report, dated August 4th 2025, the 2025 Global Cloud CapEx is now tracking to $445B and 56% Y/Y growth, with upside to +30% growth in 2026.[2]

We are exiting the early stages of what some have called the AI gold rush, and it is already clear that AI's rise signifies a paradigm shift in the entire data center architecture, with implications for storage, compute, and networking. At the end of 2024, Sandisk introduced a new AI Data Cycle storage framework to help customers unlock AI's potential.
Learn more about the framework

2. Rising Global Data Temperatures

Data fuels AI. At its core, AI is fundamentally about data utilization and data generation. LLMs thrive on ever-expanding datasets to enhance accuracy, versatility, and bias mitigation. In the process, AI systems create new data, making existing repositories more valuable for model context and learning.

The current trend for published data points for LLMs shows that training dataset sizes are growing dramatically. AI models operate in a continuous loop of data consumption and generation and the vast majority of this data is unstructured. However, this insatiable demand for data creates challenges for storage architects.

Here are some key insights from IDC's 2025 Global Datasphere[3]:

- **Data generation growth:** 24% CAGR from 2023 to 2028
- **Enterprise data outpacing consumer data:** 30% CAGR
- **Unstructured data dominance:** 92.9% of all data in 2023

LLMs have become multimodal, processing diverse data types such as text, video, and images. The evolution of data-rich generative AI chat models impacts the capacity requirements of underlying storage, and the speed, latency, retrievability, and scalability.

To optimize training time and GPU usage, storage architectures must quickly deliver relevant datasets to GPUs. IDC categorizes data into three latency types[4]:

- Ultra Real-Time (<40ms transfer latency)
- Real-Time (40ms–200ms latency)
- Nominal-Time (>200ms latency)

As data latency trends towards Real-Time and Ultra Real-Time in GPU centric workloads there is a growing need for faster storage and retrieval in data centers. Service providers are responding to this with innovations like a major cloud provider's S3 optimization (Nov 2023), offering up significantly better performance for AI and Machine Learning (AI/ML) training with single-digit millisecond latency.[5]

3. Power: The New Metric for Data Center Capacity

Data center capacity has traditionally been measured by floor space size. However, AI-driven workloads are inherently more power intensive than mainstream workloads. While GPUs are more power efficient than CPUs for AI model training, the amount of GPUs required to train the latest models is measured in the tens to hundreds of thousands of units operating constantly at full capacity for months. Likewise, AI inferencing workloads leverage large clusters of GPUs to generate responses, and consume more power than traditional analytics or database workloads.

AI data center power consumption is growing:

- AI data center energy consumption is projected to grow at a 19.5%% CAGR, between 2023 and 2028.[6]
- The share of U.S. power demand consumed by data centers is expected to increase from 4.4% today to between 6.7 and 12% by 2028.[7]

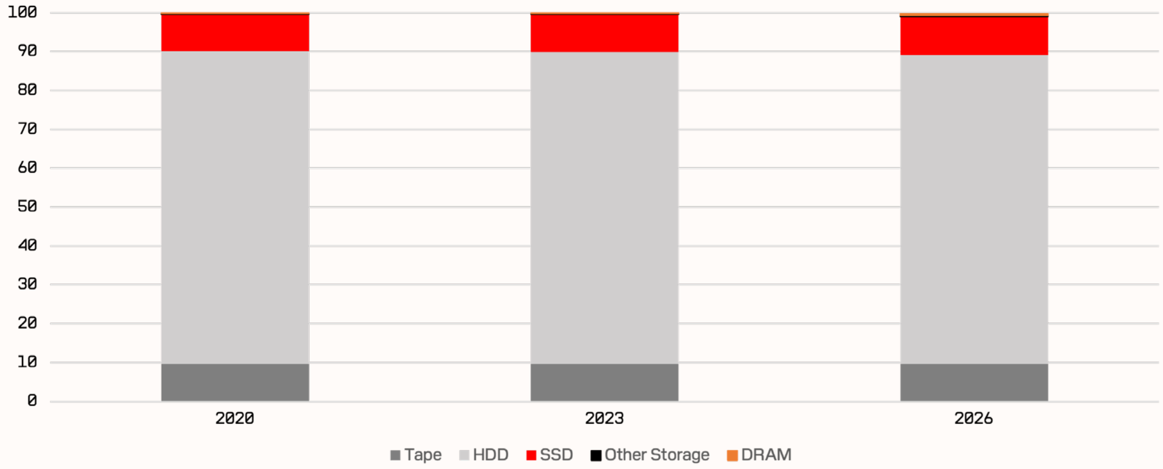
Power has become a critical metric for data centers, incurring the largest outgoing expense of data center operation and accounting for 60% of total spending for service providers.[6] Rising electricity prices and growing demand underscore the importance of energy-efficient architectures, and a complex interplay of factors will make supply more unpredictable and costly, such as environmental regulation, geopolitical factors and extreme weather events. Power has emerged as the newly important unit of measure as it better represents data center capacity than floor space.

The Economics of "State of the Art" Storage

HDDs Dominate Today, but Change Is Coming

In 2024, Tape and SSDs represented a combined 20% while HDDs represented the remaining 80% of the hyperscale storage media share of installed base of capacity. [8]

EB Installed Base Share[6]



Layered on this storage media, Cloud Service Providers (CSPs) offer access to all three storage protocols: Block, File & Object.

While Block and Cloud File storage have smaller footprints, Object storage is the largest and fastest-growing segment in data storage. Block storage workloads are performance-demanding, latency-sensitive, and have random access patterns (IOPS). Cloud File storage is designed for shared file access, with both Block and File using a mix of SSDs and HDDs.

Object storage emphasizes scale and economics, with common use cases like backup & restore and archiving. As a hypothetical example, a major cloud provider boasts 350 trillion objects, amounting to exabytes (possibly zettabytes) of data. Object storage workloads are mostly sequential (bandwidth), favoring HDDs.

Given its focus on scale and cost-efficiency, it's no surprise that most cloud data has traditionally been stored on HDDs.

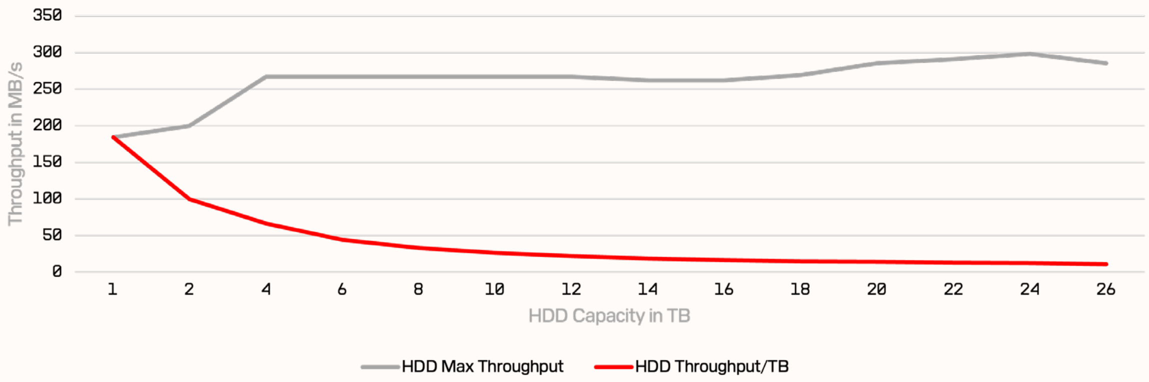
The Changing Workload Mix

AI is driving the adoption of data lakes as an emerging use case, besides the traditional backup & restore and archive for object storage. Data lakes are centralized repositories that store and process large amounts of unstructured data, and they support both training and inference tasks. Unlike traditional workloads, data lakes require high performance and low latency. HDDs are struggling to meet these demands, paving the way for SSDs that can check the boxes for both performance and capacity in performance-critical scenarios.

Service providers are responding to this growing AI workload by introducing a new storage class for object storage that is optimized to deliver a significant performance uplift needed by these workloads. This requires a solution that can check both the boxes of higher performance and larger capacity.

As the following graph shows, HDD performance has stagnated over the years, while performance per terabyte has steadily decreased as capacity has increased.

HDD Throughput and Capacity Growth[9]



CSPs have typically offered several storage classes, which can be broadly divided into 3 buckets: Standard, Infrequent Access and Archival. These have traditionally been served from HDD, and with some architectural changes HDDs could perform as needed with over-provisioning for capacity or performance. With the recent demands of AI, CSPs have started to offer a performance-optimized tier that is difficult to achieve simply by adding more capacity - this is where high capacity SSDs are fit for purpose.

Total Cost of Ownership (TCO) Is Key

Historically, storage buildouts have been driven by device acquisition costs, given the availability of abundant space, low-cost power, and reasonable performance metrics. On average, the price-per-gigabyte (ASP $/GB) for flash storage has been approximately six times higher than HDDs, making HDDs the clear economic choice.

However, this equation is shifting. Rising electricity costs, increasing power demands from GPUs, and the need for faster data lakes are reshaping the storage landscape for AI workloads. As a result, Total Cost of Ownership (TCO) is becoming a more relevant metric than ASP ($/GB) alone.

TCO includes both:

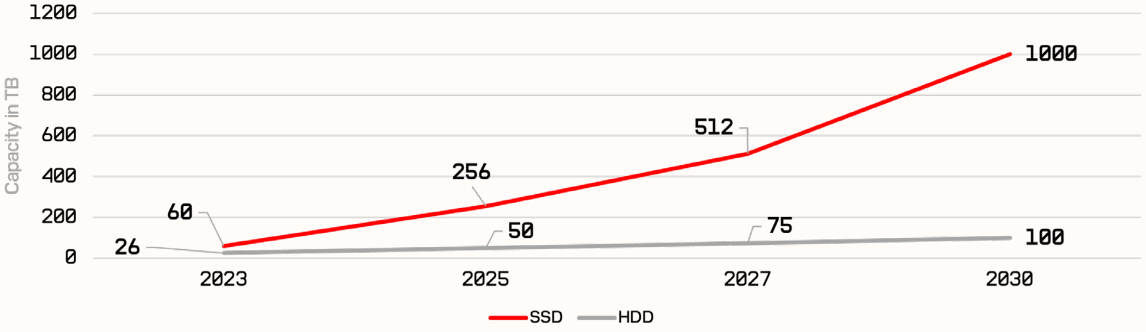
- **Acquisition costs:** Servers, storage, networking, and software.
- **Operational costs:** Infrastructure (floor space, power—both idle and active), labor, support (hardware replacements, data protection, reliability).

Beyond TCO, another critical shift is moving from **raw storage** to **effective storage**, which provides a more accurate measure of storage efficiency. Effective storage considers factors such as:

- Capacity utilization
- Duty cycle (active vs. idle)
- Replication factor
- Performance multiplier
- Data reduction ratios

As the demand for both storage density and performance density grows, evaluating the available HDD and SSD capacity points becomes essential.

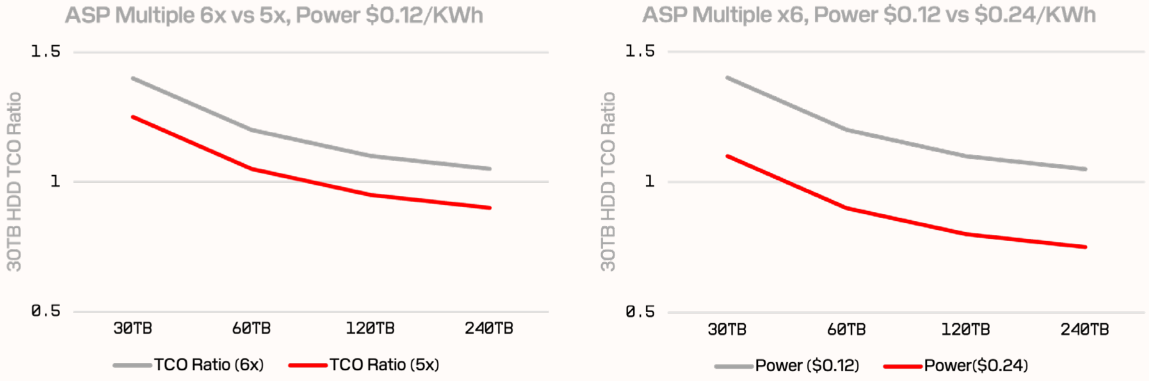
Projected SSD and HDD Capacity Growth[9]



Lets assume two variations of SSD/HDD ASP ($/GB) with multiples of 6x and 5x.

Using these assumptions and what we have learned about the new deployments supporting AI workloads in the data center, we built a Total Cost of owernship (TCO) model for a hypothetical greenfield data center with 1EB (1000PB) of storage. The media options are all-HDD with the highest available 30TB HDD vs all-SSD with various capacity points from 30T to 240T.

SSD/HDD TCO Ratio[10]



It is clear from the chart that as SSD capacities go up, their TCOs start to look more competitive vs HDD. In the second chart, we kept the ASP multiple at 6x but doubled the cost of power. At higher power cost, the all-SSD TCO starts to look more attractive at lower capacity points.

NOTE: For SAS/SATA, the conversion from HDD to SSD was realized at 3x multiple

Performance Considerations

The latest SSD capacity points leverage Quad-Level Cell (QLC) NAND technology, which increases bit density by 33% compared to Triple-Level Cell (TLC). While QLC has lower random read/write performance, lower sequential write speeds, and lower endurance than TLC—along with slightly higher write power consumption—its increasing layer count and compact packaging continue to drive capacity advancements.

Moreover, QLC SSDs are well-suited for **read-heavy, large-block sequential workloads**, aligning with object storage use cases. Despite QLC's trade-offs, SSDs **significantly outperform HDDs**, offering:

- **Massive IOPS advantages** in random workloads.
- **2–3x higher sequential throughput** (bandwidth) compared to HDDs.[11]
- **Superior scalability** through the NVMe™ interface, unlike HDDs, which remain constrained by legacy SAS/SATA interfaces.

Closing Thoughts on the Growing Role of AI in Data Centers

AI is rapidly emerging as a strategic workload in the data center. The race to develop next-generation LLMs is well underway, with these models expected to be more compute-intensive on average. As they train on larger datasets with faster GPUs, the demands on storage architecture are evolving. Future storage solutions must not only scale in capacity and speed but also optimize power management more intelligently than legacy alternatives.

IDC's view is that GenAI will have a greater positive impact on SSD and memory spending than HDDs.[12]

At Sandisk, we're excited about the future of AI-driven storage innovation. Our next-generation SSDs are being designed with higher capacities, continued performance scaling through PCIe™ Gen transitions, and configurable power profiles that align with the performance and cooling requirements of modern data centers.

As SSD capacities grow, new technical challenges and opportunities arise—such as blast radius management, data retention vs. power-performance tradeoffs, indirection unit (IU) size, and data placement strategies. At Sandisk, we recognize these challenges and are actively innovating to address them, ensuring that our storage solutions meet the evolving needs of AI-powered data centers.

[1] IDC, Worldwide Artificial Intelligence IT Spending Forecast 2024–2028, #US52635424, October 2024.
[2] Morgan Stanley Research, August 4 2025, Cloud Capex.
[3] IDC, Worldwide Global DataSphere Structured and Unstructured Data Forecast, 2024–2028 #US52554824 September 2024.
[4] IDC, Streaming and Real-Time Data in IDC's Global DataSphere, #US53122225, 2025.
[5] AWS Blog, November 2023, https://aws.amazon.com/blogs/aws/new-amazon-s3-express-one-zone-high-performance-storage-class/
[6] IDC, The Financial Impact of Increased Consumption and Rising Electricity Rates in Datacenter Facilities Spending, #US52548324, September 2024.
[7] Lawrence Berkley National Laboratory, Dec 2024, 2024 United States Data Center Energy Usage Report.
[8] IDC, The Evolving Datacenter Memory and Storage Media Hierarchy: A Deep Dive into the Installed Base, #US51902124, March 2024.
[9] Based on Internal SANDISK and TRENDFOCUS Projections.
[10] Using SNIA TCO Calculator, https://www.snia.org/forums/cmsi/programs/TCOcalc as of September 2025.
[11] Based on internal testing of SANDISK products.
[12] IDC, Worldwide Hard Disk Drive Forecast 2025-2029, #US53465525, June 2025

Product specifications subject to change without notice. Pictures shown may vary from actual products. References to Sandisk products do not imply they will be made available in all regions.